Statistical Discovery.™ From SAS.

Version 13

Basic Analysis

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

Marcel Proust

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

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2016. *JMP® 13 Basic Analysis*. Cary, NC: SAS Institute Inc.

JMP® 13 Basic Analysis

Copyright © 2016, SAS Institute Inc., Cary, NC, USA

ISBN 978-1-62960-471-8 (Hardcopy)
ISBN 978-1-62960-566-1 (EPUB)
ISBN 978-1-62960-567-8 (MOBI)

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 License Rights; Restricted Rights: 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.

September 2016

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.

Technology License Notices

- Scintilla - Copyright © 1998-2014 by Neil Hodgson <neilh@scintilla.org>.

  All Rights Reserved.

  Permission to use, copy, modify, and distribute this software and its documentation for any purpose and without fee is hereby granted, provided that the above copyright notice appear in all copies and that both that copyright notice and this permission notice appear in supporting documentation.

  NEIL HODGSON DISCLAIMS ALL WARRANTIES WITH REGARD TO THIS SOFTWARE, INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS, IN NO EVENT SHALL NEIL HODGSON BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.

- Telerik RadControls: Copyright © 2002-2012, Telerik. Usage of the included Telerik RadControls outside of JMP is not permitted.
- ZLIB Compression Library - Copyright © 1995-2005, Jean-Loup Gailly and Mark Adler.
- Made with Natural Earth. Free vector and raster map data @ naturalearthdata.com.
- Packages - Copyright © 2009-2010, Stéphane Sudre (s.sudre.free.fr). All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  Neither the name of the WhiteBox nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED

WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- iODBC software - Copyright © 1995-2006, OpenLink Software Inc and Ke Jin (www.iodbc.org). All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  – Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  – Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  – Neither the name of OpenLink Software Inc. nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL OPENLINK OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- bzip2, the associated library "libbzip2", and all documentation, are Copyright © 1996-2010, Julian R Seward. All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

The origin of this software must not be misrepresented; you must not claim that you wrote the original software. If you use this software in a product, an acknowledgment in the product documentation would be appreciated but is not required.

Altered source versions must be plainly marked as such, and must not be misrepresented as being the original software.

The name of the author may not be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE AUTHOR "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- R software is Copyright © 1999-2012, R Foundation for Statistical Computing.
- MATLAB software is Copyright © 1984-2012, The MathWorks, Inc. Protected by U.S. and international patents. See www.mathworks.com/patents. MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See www.mathworks.com/trademarks for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.
- libopc is Copyright © 2011, Florian Reuter. All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  – Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  – Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and / or other materials provided with the distribution.

– Neither the name of Florian Reuter nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- libxml2 - Except where otherwise noted in the source code (e.g. the files hash.c, list.c and the trio files, which are covered by a similar licence but with different Copyright notices) all the files are:

  Copyright © 1998 - 2003 Daniel Veillard. All Rights Reserved.

  Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

  The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

  THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL DANIEL VEILLARD BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

  Except as contained in this notice, the name of Daniel Veillard shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization from him.

- Regarding the decompression algorithm used for UNIX files:

  Copyright © 1985, 1986, 1992, 1993

The Regents of the University of California. All rights reserved.

THIS SOFTWARE IS PROVIDED BY THE REGENTS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE REGENTS OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

3. Neither the name of the University nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

- Snowball - Copyright © 2001, Dr Martin Porter, Copyright © 2002, Richard Boulton.

  All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and / or other materials provided with the distribution.

  3. Neither the name of the copyright holder nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS \"AS IS\" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE

DISCLAIMED.IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES(INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT(INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Get the Most from JMP®

Whether you are a first-time or a long-time user, there is always something to learn about JMP.

Visit JMP.com to find the following:

- live and recorded webcasts about how to get started with JMP
- video demos and webcasts of new features and advanced techniques
- details on registering for JMP training
- schedules for seminars being held in your area
- success stories showing how others use JMP
- a blog with tips, tricks, and stories from JMP staff
- a forum to discuss JMP with other users

http://www.jmp.com/getstarted/

Contents

Basic Analysis

1 Learn about JMP

Documentation and Additional Resources . . . 19

Formatting Conventions . . . 20

JMP Documentation . . . 21

JMP Documentation Library . . . 21

JMP Help . . . 27

Additional Resources for Learning JMP . . . 27

Tutorials . . . 28

Sample Data Tables . . . 28

Learn about Statistical and JSL Terms . . . 28

Learn JMP Tips and Tricks . . . 28

Tooltips . . . 29

JMP User Community . . . 29

JMPer Cable . . . 29

JMP Books by Users . . . 30

The JMP Starter Window . . . 30

Technical Support . . . 30

2 Introduction to Basic Analysis

Overview of Fundamental Analysis Methods . . . 31

3 Distributions

Using the Distribution Platform . . . 33

Overview of the Distribution Platform . . . 34

Categorical Variables . . . 34

Continuous Variables . . . 34

Example of the Distribution Platform . . . 34

Launch the Distribution Platform . . . 36

The Distribution Report . . . 37

Histograms . . . 39

The Frequencies Report . . . 42

The Quantiles Report . . . 42

The Summary Statistics Report . . . 42

Distribution Platform Options . . . 45

Options for Categorical Variables . . . 46

Display Options Submenu for Categorical Variables 46
Histogram Options Submenu for Categorical Variables 46
Save Submenu for Categorical Variables 47
Options for Continuous Variables 47
Display Options Submenu for Continuous Variables 47
Histogram Options Submenu for Continuous Variables 48
Normal Quantile Plot 50
Outlier Box Plot 51
Quantile Box Plot 52
Stem and Leaf 52
CDF Plot 53
Test Mean 53
Test Std Dev 54
Test Equivalence 55
Confidence Intervals for Continuous Variables 56
Save Commands for Continuous Variables 57
Prediction Intervals 58
Tolerance Intervals 59
Capability Analysis 59
Fit Distributions 61
Continuous Fit 61
Discrete Fit 62
Fit Distribution Options 63
Additional Examples of the Distribution Platform 66
Example of Selecting Data in Multiple Histograms 66
Examples of the Test Probabilities Option 67
Example of Prediction Intervals 69
Example of Tolerance Intervals 70
Example of Capability Analysis 71
Statistical Details for the Distribution Platform 72
Standard Error Bars 72
Quantiles 73
Summary Statistics 73
Normal Quantile Plot 74
Wilcoxon Signed Rank Test 75
Standard Deviation Test 77
Normal Quantiles 77
Saving Standardized Data 77
Prediction Intervals 78
Tolerance Intervals 78
Capability Analysis 79
Continuous Fit Distributions 83
Discrete Fit Distributions 89

Fitted Quantiles . . . 92
Fit Distribution Options . . . 92

4 Introduction to Fit Y by X
Examine Relationships Between Two Variables . . . 95
Overview of the Fit Y by X Platform . . . 96
Launch the Fit Y by X Platform . . . 96
Launch Specific Analyses from the JMP Starter Window . . . 97

5 Bivariate Analysis
Examine Relationships between Two Continuous Variables . . . 99
Example of Bivariate Analysis . . . 100
Launch the Bivariate Platform . . . 100
The Bivariate Plot . . . 101
Fitting Commands and Options . . . 102
Fit Mean Options . . . 103
Fitting Command Categories . . . 104
Fit the Same Command Multiple Times . . . 105
Histogram Borders . . . 105
Fit Mean . . . 105
Fit Mean Report . . . 106
Fit Line and Fit Polynomial . . . 106
Linear Fit and Polynomial Fit Reports . . . 107
Fit Special . . . 113
Fit Special Reports and Menus . . . 113
Flexible . . . 114
Fit Spline . . . 115
Kernel Smoother . . . 116
Fit Each Value . . . 116
Fit Orthogonal . . . 117
Fit Orthogonal Options . . . 117
Orthogonal Regression Report . . . 117
Robust . . . 118
Fit Robust . . . 118
Fit Cauchy . . . 119
Density Ellipse . . . 119
Correlation Report . . . 120
Nonpar Density . . . 120
Nonparametric Bivariate Density Report . . . 121
Group By . . . 121
Fitting Menus . . . 122
Fitting Menu Options . . . 122
Diagnostics Plots . . . 125

Additional Examples of the Bivariate Platform 126
Example of the Fit Special Command 126
Example Using the Fit Orthogonal Command 127
Example Using the Fit Robust Command 129
Example of Group By Using Density Ellipses 131
Example of Group By Using Regression Lines 132
Statistical Details for the Bivariate Platform 133
Fit Line 133
Fit Spline 133
Fit Orthogonal 133
Summary of Fit Report 134
Lack of Fit Report 135
Parameter Estimates Report 135
Smoothing Fit Reports 136
Correlation Report 136

6 Oneway Analysis
Examine Relationships between a Continuous Y and a Categorical X Variable 137
Overview of Oneway Analysis 138
Example of Oneway Analysis 138
Launch the Oneway Platform 140
The Oneway Plot 140
Oneway Platform Options 141
Display Options 144
Quantiles 146
Outlier Box Plots 146
Means/Anova and Means/Anova/Pooled t 147
The Summary of Fit Report 147
The t-test Report 148
The Analysis of Variance Report 149
The Means for Oneway Anova Report 150
The Block Means Report 150
Mean Diamonds and X-Axis Proportional 150
Mean Lines, Error Bars, and Standard Deviation Lines 151
Analysis of Means Methods 152
Analysis of Means for Location 152
Analysis of Means for Scale 153
Analysis of Means Charts 154
Analysis of Means Options 155
Compare Means 155
Using Comparison Circles 157
Each Pair, Student's t 158
All Pairs, Tukey HSD 158

With Best, Hsu MCB . . . 159
With Control, Dunnett's . . . 160
Compare Means Options . . . 160
Nonparametric Tests . . . 161
The Wilcoxon, Median, and Van der Waerden Test Reports . . . 162
Kolmogorov-Smirnov Two-Sample Test Report . . . 163
Nonparametric Multiple Comparisons . . . 164
Unequal Variances . . . 167
Tests That the Variances Are Equal Report . . . 168
Equivalence Test . . . 170
Robust . . . 170
Robust Fit . . . 170
Cauchy Fit . . . 170
Power . . . 171
Power Details Window and Reports . . . 171
Normal Quantile Plot . . . 172
CDF Plot . . . 173
Densities . . . 173
Matching Column . . . 173
Additional Examples of the Oneway Platform . . . 174
Example of an Analysis of Means Chart . . . 174
Example of an Analysis of Means for Variances Chart . . . 175
Example of the Each Pair, Student's t Test . . . 176
Example of the All Pairs, Tukey HSD Test . . . 178
Example of the With Best, Hsu MCB Test . . . 180
Example of the With Control, Dunnett's Test . . . 181
Example Contrasting All of the Compare Means Tests . . . 182
Example of the Nonparametric Wilcoxon Test . . . 183
Example of the Unequal Variances Option . . . 186
Example of an Equivalence Test . . . 187
Example of the Robust Fit Option . . . 188
Example of the Power Option . . . 190
Example of a Normal Quantile Plot . . . 191
Example of a CDF Plot . . . 192
Example of the Densities Options . . . 193
Example of the Matching Column Option . . . 194
Example of Stacking Data for a Oneway Analysis . . . 195
Statistical Details for the Oneway Platform . . . 202
Comparison Circles . . . 202
Power . . . 203
Summary of Fit Report . . . 204
Tests That the Variances Are Equal . . . 204
Nonparametric Test Statistics . . . 206

7 Contingency Analysis

Examine Relationships between Two Categorical Variables 209
Example of Contingency Analysis 210
Launch the Contingency Platform 211
The Contingency Report 211
Contingency Platform Options 212
Mosaic Plot 214
Context Menu 215
Contingency Table 216
Description of the Contingency Table 217
Tests 218
Description of the Tests Report 218
Fisher's Exact Test 219
Analysis of Means for Proportions 219
Correspondence Analysis 219
Understanding Correspondence Analysis Plots 220
Correspondence Analysis Options 220
The Details Report 220
Cochran-Mantel-Haenszel Test 221
Agreement Statistic 221
Relative Risk 222
Two Sample Test for Proportions 222
Measures of Association 223
Cochran Armitage Trend Test 224
Exact Test 225
Additional Examples of the Contingency Platform 225
Example of Analysis of Means for Proportions 225
Example of Correspondence Analysis 226
Example of a Cochran Mantel Haenszel Test 229
Example of the Agreement Statistic Option 230
Example of the Relative Risk Option 231
Example of a Two Sample Test for Proportions 233
Example of the Measures of Association Option 234
Example of the Cochran Armitage Trend Test 234
Statistical Details for the Contingency Platform 235
Agreement Statistic Option 235
Odds Ratio Option 236
Tests Report 236
Details Report in Correspondence Analysis 237

8 Logistic Analysis

Examine Relationships between a Categorical Y and a Continuous X Variable 239
Overview of Logistic Regression 240

Nominal Logistic Regression . . . 240
Ordinal Logistic Regression . . . 240
Example of Nominal Logistic Regression . . . 241
Launch the Logistic Platform . . . 242
Data Structure . . . 243
The Logistic Report . . . 243
Logistic Plot . . . 244
Iterations . . . 245
Whole Model Test . . . 245
Parameter Estimates . . . 247
Logistic Platform Options . . . 247
ROC Curves . . . 249
Save Probability Formula . . . 250
Inverse Prediction . . . 250
Additional Examples of Logistic Regression . . . 250
Example of Ordinal Logistic Regression . . . 250
Additional Example of a Logistic Plot . . . 252
Example of ROC Curves . . . 254
Example of Inverse Prediction Using the Crosshair Tool . . . 255
Example of Inverse Prediction Using the Inverse Prediction Option . . . 256
Statistical Details for the Logistic Platform . . . 258
Whole Model Test Report . . . 258

9 Tabulate
Create Summary Tables Interactively . . . 259
Example of the Tabulate Platform . . . 260
Launch the Tabulate Platform . . . 263
Use the Dialog . . . 265
Add Statistics . . . 266
The Tabulate Output . . . 269
Analysis Columns . . . 271
Grouping Columns . . . 271
Column and Row Tables . . . 272
Edit Tables . . . 273
Tabulate Platform Options . . . 273
Show Test Build Panel . . . 274
Right-Click Menu for Columns . . . 275
Additional Examples of the Tabulate Platform . . . 275
Example of Combining Columns into a Single Table . . . 279
Example Using a Page Column . . . 281

10 Simulate

Answer Challenging Questions with Parametric Resampling 283
Overview of Simulate 284
Examples That Use Simulate 284
Construct an Accurate Confidence Interval for Variance Components 285
Conduct a Permutation Test 290
Explore Retaining a Factor in Generalized Regression 293
Conduct Prospective Power Analysis for a Nonlinear Model 298
Launch the Simulate Window 306
The Simulate Window 307
The Simulate Table 307
Simulated Power Report 308

11 Bootstrapping

Approximate the Distribution of a Statistic through Resampling 309
Overview of Bootstrapping 310
Example of Bootstrapping 311
Bootstrapping Window Options 313
Stacked Results Table 314
Unstacked Bootstrap Results Table 315
Analysis of Bootstrap Results 316
Additional Example of Bootstrapping 317
Statistical Details for Bootstrapping 321
Calculation of Fractional Weights 322
Bias-Corrected Percentile Intervals 322

12 Text Explorer

Explore Unstructured Text in Your Data 325
Text Explorer Platform Overview 326
Text Processing Steps 327
Example of the Text Explorer Platform 328
Launch the Text Explorer Platform 331
Customize Regex: Regular Expression Editor 333
The Text Explorer Report 338
Summary Counts Report 338
Term and Phrase Lists 339
Text Explorer Platform Options 341
Text Preparation Options 341
Text Analysis Options 346
Save Options 349
Report Options 351
Latent Class Analysis 352
Latent Semantic Analysis (SVD) 353

SVD Plots Report 354
Topic Analysis 354
Topic Words Report 354
Topic Scores Report 355
Additional Example of the Text Explorer Platform 355

A Statistical Details
Basic Analysis 359
Platforms That Support Validation 360

B References

Index
Basic Analysis 365

Chapter 1

Learn about JMP

Documentation and Additional Resources

This chapter includes the following information:

- book conventions
- JMP documentation
- JMP Help
- additional resources, such as the following:
  - other JMP documentation
  - tutorials
  - indexes
  - Web resources
  - technical support options

Formatting Conventions

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

- Sample data table names, column names, pathnames, filenames, file extensions, and folders appear in Helvetica font.
- Code appears in `Lucida Sans Typewriter` font.
- Code output appears in *`Lucida Sans Typewriter`* italic font and is indented farther than the preceding code.
- **Helvetica bold** formatting indicates items that you select to complete a task:
  - buttons
  - check boxes
  - commands
  - list names that are selectable
  - menus
  - options
  - tab names
  - text boxes
- The following items appear in italics:
  - words or phrases that are important or have definitions specific to JMP
  - book titles
  - variables
  - script output
- Features that are for JMP Pro only are noted with the JMP Pro icon JMP PRO. For an overview of JMP Pro features, visit http://www.jmp.com/software/pro/.

Note: Special information and limitations appear within a Note.

Tip: Helpful information appears within a Tip.

JMP Documentation

JMP offers documentation in various formats, from print books and Portable Document Format (PDF) to electronic books (e-books).

- Open the PDF versions from the **Help > Books** menu.
- All books are also combined into one PDF file, called *JMP Documentation Library*, for convenient searching. Open the *JMP Documentation Library* PDF file from the **Help > Books** menu.
- You can also purchase printed documentation and e-books on the SAS website:

  http://www.sas.com/store/search.ep?keyWords=JMP

JMP Documentation Library

The following table describes the purpose and content of each book in the JMP library.

Document Title	Document Purpose	Document Content
Discovering JMP	If you are not familiar with JMP, start here.	Introduces you to JMP and gets you started creating and analyzing data.
Using JMP	Learn about JMP data tables and how to perform basic operations.	Covers general JMP concepts and features that span across all of JMP, including importing data, modifying columns properties, sorting data, and connecting to SAS.
Basic Analysis	Perform basic analysis using this document.	Describes these Analyze menu platforms: • Distribution • Fit Y by X • Tabulate • Text Explorer Covers how to perform bivariate, one-way ANOVA, and contingency analyses through Analyze > Fit Y by X. How to approximate sampling distributions using bootstrapping and how to perform parametric resampling with the Simulate platform are also included.

Document Title	Document Purpose	Document Content
Essential Graphing	Find the ideal graph for your data.	Describes these Graph menu platforms: • Graph Builder • Overlay Plot • Scatterplot 3D • Contour Plot • Bubble Plot • Parallel Plot • Cell Plot • Treemap • Scatterplot Matrix • Ternary Plot • Chart The book also covers how to create background and custom maps.
Profilers	Learn how to use interactive profiling tools, which enable you to view cross-sections of any response surface.	Covers all profilers listed in the Graph menu. Analyzing noise factors is included along with running simulations using random inputs.
Design of Experiments Guide	Learn how to design experiments and determine appropriate sample sizes.	Covers all topics in the DOE menu and the Specialized DOE Models menu item in the Analyze > Specialized Modeling menu.

Document Title	Document Purpose	Document Content
Fitting Linear Models	Learn about Fit Model platform and many of its personalities.	Describes these personalities, all available within the Analyze menu Fit Model platform: • Standard Least Squares • Stepwise • Generalized Regression • Mixed Model • MANOVA • Loglinear Variance • Nominal Logistic • Ordinal Logistic • Generalized Linear Model

Document Title	Document Purpose	Document Content
Predictive and Specialized Modeling	Learn about additional modeling techniques.	Describes these Analyze > Predictive Modeling menu platforms: • Modeling Utilities • Neural • Partition • Bootstrap Forest • Boosted Tree • K Nearest Neighbors • Naive Bayes • Model Comparison • Formula Depot Describes these Analyze > Specialized Modeling menu platforms: • Fit Curve • Nonlinear • Gaussian Process • Time Series • Matched Pairs Describes these Analyze > Screening menu platforms: • Response Screening • Process Screening • Predictor Screening • Association Analysis The platforms in the Analyze > Specialized Modeling > Specialized DOE Models menu are described in *Design of Experiments Guide*.

Document Title	Document Purpose	Document Content
Multivariate Methods	Read about techniques for analyzing several variables simultaneously.	Describes these Analyze > Multivariate Methods menu platforms: • Multivariate • Principal Components • Discriminant • Partial Least Squares Describes these Analyze > Clustering menu platforms: • Hierarchical Cluster • K Means Cluster • Normal Mixtures • Latent Class Analysis • Cluster Variables
Quality and Process Methods	Read about tools for evaluating and improving processes.	Describes these Analyze > Quality and Process menu platforms: • Control Chart Builder and individual control charts • Measurement Systems Analysis • Variability / Attribute Gauge Charts • Process Capability • Pareto Plot • Diagram

Document Title	Document Purpose	Document Content
Reliability and Survival Methods	Learn to evaluate and improve reliability in a product or system and analyze survival data for people and products.	Describes these Analyze > Reliability and Survival menu platforms: • Life Distribution • Fit Life by X • Cumulative Damage • Recurrence Analysis • Degradation and Destructive Degradation • Reliability Forecast • Reliability Growth • Reliability Block Diagram • Repairable Systems Simulation • Survival • Fit Parametric Survival • Fit Proportional Hazards
Consumer Research	Learn about methods for studying consumer preferences and using that insight to create better products and services.	Describes these Analyze > Consumer Research menu platforms: • Categorical • Multiple Correspondence Analysis • Multidimensional Scaling • Factor Analysis • Choice • MaxDiff • Uplift • Item Analysis
Scripting Guide	Learn about taking advantage of the powerful JMP Scripting Language (JSL).	Covers a variety of topics, such as writing and debugging scripts, manipulating data tables, constructing display boxes, and creating JMP applications.

Document Title	Document Purpose	Document Content
JSL Syntax Reference	Read about many JSL functions on functions and their arguments, and messages that you send to objects and display boxes.	Includes syntax, examples, and notes for JSL commands.

Note: The **Books** menu also contains two reference cards that can be printed: The *Menu Card* describes JMP menus, and the *Quick Reference* describes JMP keyboard shortcuts.

JMP Help

JMP Help is an abbreviated version of the documentation library that provides targeted information. You can open JMP Help in several ways:

- On Windows, press the F1 key to open the Help system window.
- Get help on a specific part of a data table or report window. Select the Help tool ? from the **Tools** menu and then click anywhere in a data table or report window to see the Help for that area.
- Within a JMP window, click the **Help** button.
- Search and view JMP Help on Windows using the **Help > Help Contents**, **Search Help**, and **Help Index** options. On Mac, select **Help > JMP Help**.
- Search the Help at http://jmp.com/support/help/ (English only).

Additional Resources for Learning JMP

In addition to JMP documentation and JMP Help, you can also learn about JMP using the following resources:

- Tutorials (see "Tutorials" on page 28)
- Sample data (see "Sample Data Tables" on page 28)
- Indexes (see "Learn about Statistical and JSL Terms" on page 28)
- Tip of the Day (see "Learn JMP Tips and Tricks" on page 28)
- Web resources (see "JMP User Community" on page 29)
- JMPer Cable technical publication (see "JMPer Cable" on page 29)
- Books about JMP (see "JMP Books by Users" on page 30)
- JMP Starter (see "The JMP Starter Window" on page 30)

- Teaching Resources (see "Sample Data Tables" on page 28)

Tutorials

You can access JMP tutorials by selecting **Help > Tutorials**. The first item on the **Tutorials** menu is **Tutorials Directory**. This opens a new window with all the tutorials grouped by category.

If you are not familiar with JMP, then start with the **Beginners Tutorial**. It steps you through the JMP interface and explains the basics of using JMP.

The rest of the tutorials help you with specific aspects of JMP, such as designing an experiment and comparing a sample mean to a constant.

Sample Data Tables

All of the examples in the JMP documentation suite use sample data. Select **Help > Sample Data Library** to open the sample data directory.

To view an alphabetized list of sample data tables or view sample data within categories, select **Help > Sample Data**.

Sample data tables are installed in the following directory:

On Windows: C:\Program Files\SAS\JMP\13\Samples\Data

On Macintosh: \Library\Application Support\JMP\13\Samples\Data

In JMP Pro, sample data is installed in the JMPPRO (rather than JMP) directory. In JMP Shrinkwrap, sample data is installed in the JMPSW directory.

To view examples using sample data, select **Help > Sample Data** and navigate to the Teaching Resources section. To learn more about the teaching resources, visit http://jmp.com/tools.

Learn about Statistical and JSL Terms

The **Help** menu contains the following indexes:

Statistics Index Provides definitions of statistical terms.

Scripting Index Lets you search for information about JSL functions, objects, and display boxes. You can also edit and run sample scripts from the Scripting Index.

Learn JMP Tips and Tricks

When you first start JMP, you see the Tip of the Day window. This window provides tips for using JMP.

To turn off the Tip of the Day, clear the **Show tips at startup** check box. To view it again, select **Help > Tip of the Day**. Or, you can turn it off using the Preferences window. See the *Using JMP* book for details.

Tooltips

JMP provides descriptive tooltips when you place your cursor over items, such as the following:

- Menu or toolbar options
- Labels in graphs
- Text results in the report window (move your cursor in a circle to reveal)
- Files or windows in the Home Window
- Code in the Script Editor

Tip: On Windows, you can hide tooltips in the JMP Preferences. Select **File > Preferences > General** and then deselect **Show menu tips**. This option is not available on Macintosh.

JMP User Community

The JMP User Community provides a range of options to help you learn more about JMP and connect with other JMP users. The learning library of one-page guides, tutorials, and demos is a good place to start. And you can continue your education by registering for a variety of JMP training courses.

Other resources include a discussion forum, sample data and script file exchange, webcasts, and social networking groups.

To access JMP resources on the website, select **Help > JMP User Community** or visit https://community.jmp.com/.

JMPer Cable

The JMPer Cable is a yearly technical publication targeted to users of JMP. The JMPer Cable is available on the JMP website:

http://www.jmp.com/about/newsletters/jmpercable/

JMP Books by Users

Additional books about using JMP that are written by JMP users are available on the JMP website:

http://www.jmp.com/en_us/software/books.html

The JMP Starter Window

The JMP Starter window is a good place to begin if you are not familiar with JMP or data analysis. Options are categorized and described, and you launch them by clicking a button. The JMP Starter window covers many of the options found in the Analyze, Graph, Tables, and File menus. The window also lists JMP Pro features and platforms.

- To open the JMP Starter window, select **View** (**Window** on the Macintosh) > **JMP Starter**.
- To display the JMP Starter automatically when you open JMP on Windows, select **File > Preferences > General**, and then select **JMP Starter** from the Initial JMP Window list. On Macintosh, select **JMP > Preferences > Initial JMP Starter Window**.

Technical Support

JMP technical support is provided by statisticians and engineers educated in SAS and JMP, many of whom have graduate degrees in statistics or other technical disciplines.

Many technical support options are provided at http://www.jmp.com/support, including the technical support phone number.

Chapter 2

Introduction to Basic Analysis

Overview of Fundamental Analysis Methods

This book describes the initial types of analyses that you often perform in JMP:

- The Distribution platform illustrates the distribution of a single variable using histograms, additional graphs, and reports. Once you know how your data is distributed, you can plan the appropriate type of analysis going forward. See Chapter 3, "Distributions".
- The Fit Y by X platform analyzes the pair of X and Y variables that you specify, by context, based on modeling type. See Chapter 4, "Introduction to Fit Y by X". The four types of analyses include:
  - The Bivariate platform, which analyzes the relationship between two continuous X variables. See Chapter 5, "Bivariate Analysis".
  - The Oneway platform, which analyzes how the distribution of a continuous Y variable differs across groups defined by a categorical X variable. See Chapter 6, "Oneway Analysis".
  - The Contingency platform, which analyzes the distribution of a categorical response variable Y as conditioned by the values of a categorical X factor. See Chapter 7, "Contingency Analysis".
  - The Logistic platform, which fits the probabilities for response categories (Y) to a continuous X predictor. See Chapter 8, "Logistic Analysis".
- The Tabulate platform interactively constructs tables of descriptive statistics. See Chapter 9, "Tabulate".
- The Simulate feature provides parametric and nonparametric simulation capability. See Chapter 10, "Simulate".
- Bootstrap analysis approximates the sampling distribution of a statistic. The data is re-sampled with replacement and the statistic is computed. This process is repeated to produce a distribution of values for the statistic. See Chapter 11, "Bootstrapping".
- The Text Explorer platform enables you to categorize and analyze unformatted text data. You can use regular expressions to clean up the data before you proceed to analysis. See Chapter 12, "Text Explorer".

Chapter 3

Distributions

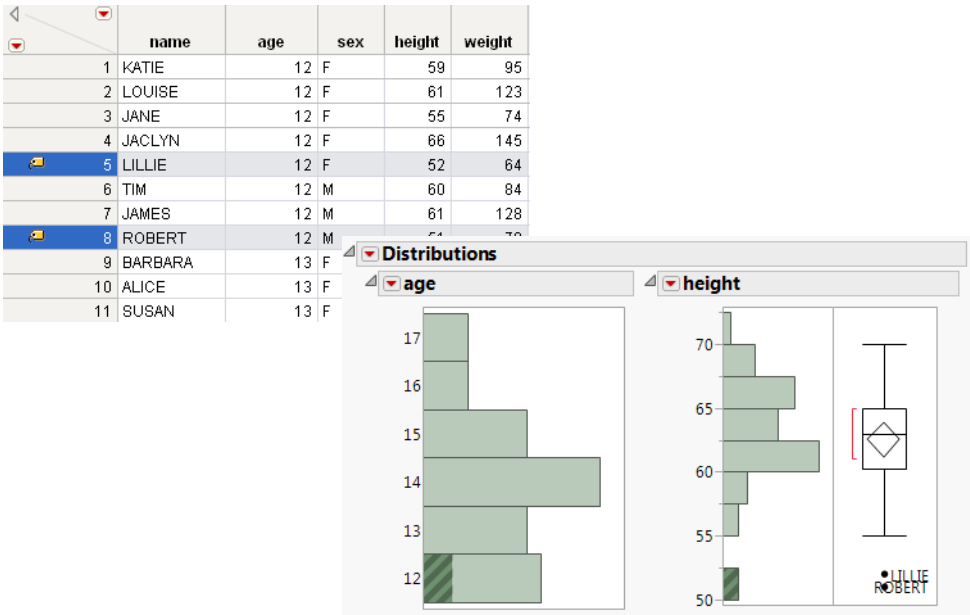
Using the Distribution Platform

The Distribution platform illustrates the distribution of a single variable using histograms, additional graphs, and reports. The word *univariate* simply means involving one variable instead of two (bivariate) or many (multivariate). However, you can examine the distribution of several individual variables within a report. The report content for each variable changes depending on whether the variable is categorical (nominal or ordinal) or continuous.

Once you know how your data is distributed, you can plan the appropriate type of analysis going forward.

The Distribution report window is interactive. Clicking on a histogram bar highlights the corresponding data in any other histograms and in the data table. See Figure 3.1.

Figure 3.1 Example of the Distribution Platform



Overview of the Distribution Platform

The treatment of variables in the Distribution platform is different, depending on the modeling type of variable, which can be categorical (nominal or ordinal) or continuous.

Categorical Variables

For categorical variables, the initial graph that appears is a histogram. The histogram shows a bar for each level of the ordinal or nominal variable. You can also add a divided (mosaic) bar chart.

The reports show counts and proportions. You can add confidence intervals and test the probabilities.

Continuous Variables

For numeric continuous variables, the initial graphs show a histogram and an outlier box plot. The histogram shows a bar for grouped values of the continuous variable. The following options are also available:

- quantile box plot
- normal quantile plot
- stem and leaf plot
- CDF plot

The reports show selected quantiles and summary statistics. Report options are available for the following:

- saving ranks, probability scores, normal quantile values, and so on, as new columns in the data table
- testing the mean and standard deviation of the column against a constant you specify
- fitting various distributions and nonparametric smoothing curves
- performing a capability analysis for a quality control application
- confidence intervals, prediction intervals, and tolerance intervals

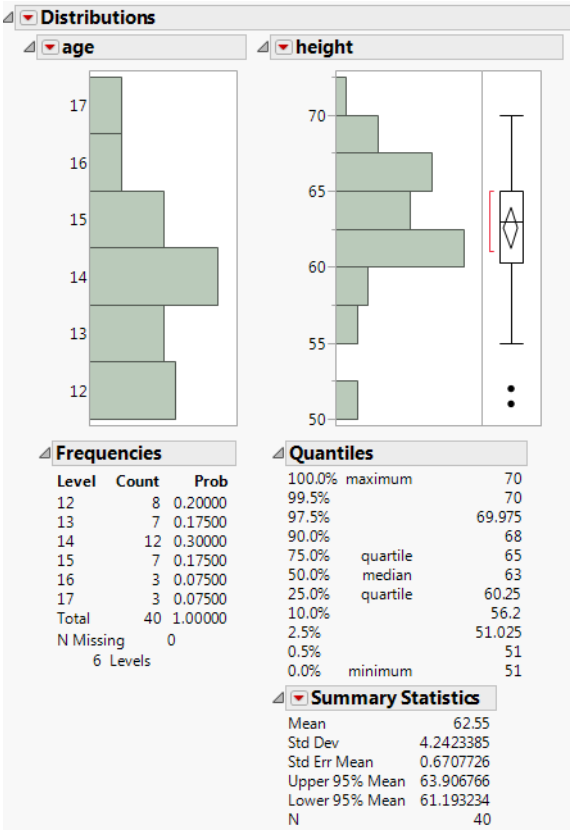
Example of the Distribution Platform

Suppose that you have data on 40 students, and you want to see the distribution of age and height among the students.

1. Select **Help > Sample Data Library** and open Big Class.jmp.

2. Select **Analyze > Distribution**.
3. Select age and height and click **Y, Columns**.
4. Click **OK**.

Figure 3.2 Example of the Distribution Platform



From the histograms, you notice the following:

- The ages are not uniformly distributed.
- For height, there are two points with extreme values (that might be outliers).

Click on the bar for 50 in the height histogram to take a closer look at the potential outliers.

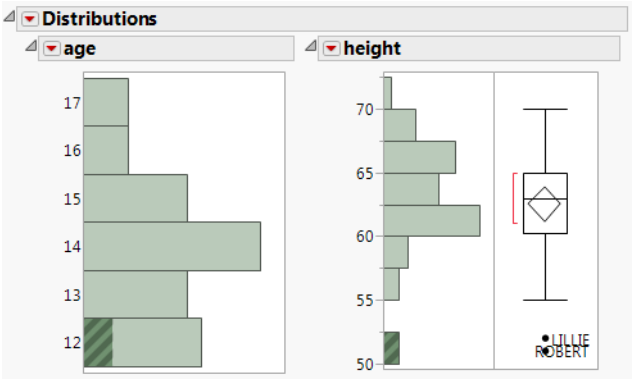
- The corresponding ages are highlighted in the age histogram. The potential outliers are age 12.
- The corresponding rows are highlighted in the data table. The names of the potential outliers are Lillie and Robert.

Add labels to the potential outliers in the height histogram.

1. Select both outliers.
2. Right-click on one of the outliers and select **Row Label**.

   Label icons are added to these rows in the data table.
3. Resize the box plot wider to see the full labels.

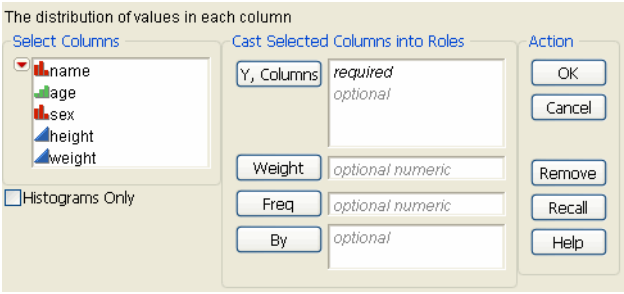
Figure 3.3 Potential Outliers Labeled



Launch the Distribution Platform

Launch the Distribution platform by selecting **Analyze > Distribution**.

Figure 3.4 The Distribution Launch Window



Y, Columns Assigns the variables that you want to analyze. A histogram and associated reports appear for each variable.

Weight Assigns a variable that specifies weights for observations on continuous Ys. For categorical Ys, the Weight column is ignored. Any statistic that is based on the sum of the weights is affected by weights.

Freq Assigns a frequency variable to this role. This is useful if you have summarized data. In this instance, you have one column for the Y values and another column for the frequency of occurrence of the Y values. The sum of this variable is included in the overall count appearing in the Summary Statistics report (represented by N). All other moment statistics (mean, standard deviation, and so on) are also affected by the **Freq** variable.

By Produces a separate report for each level of the **By** variable. If more than one **By** variable is assigned, a separate report is produced for each possible combination of the levels of the **By** variables.

Histograms Only Removes everything except the histograms from the report window.

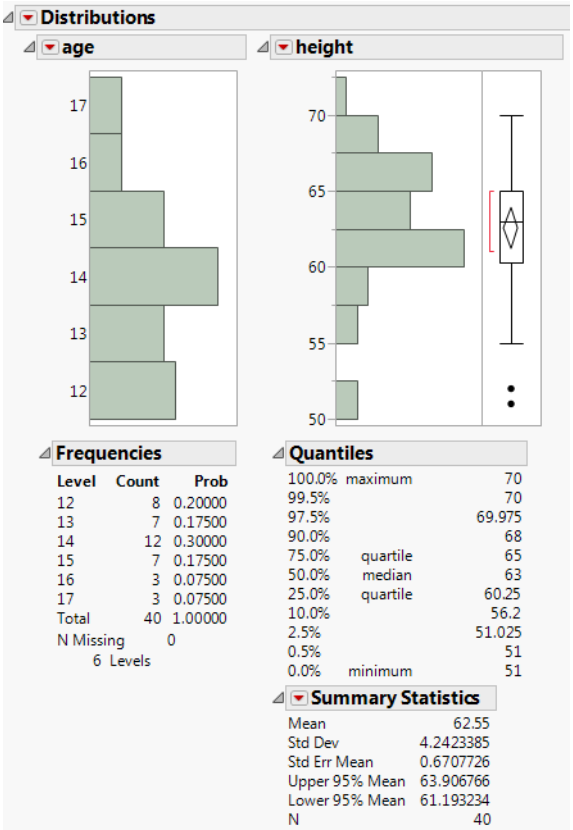
For general information about launch windows, see the Get Started chapter in the *Using JMP* book.

After you click **OK**, the Distribution report window appears. See "The Distribution Report" on page 37.

The Distribution Report

Follow the instructions in "Example of the Distribution Platform" on page 34 to produce the report shown in Figure 3.5.

Figure 3.5 The Initial Distribution Report Window



Note: When you apply only the Hidden row state to rows in the data table, the corresponding points do not appear in plots that show points. However, histograms are constructed using the hidden rows. If you want to exclude rows from the construction of the histograms and from analysis results, apply the Exclude row state and select **Redo > Redo Analysis** from the Bivariate red triangle menu.

The initial Distribution report contains a histogram and reports for each variable. Note the following:

- To replace a variable in a report, from the Columns panel of the associated data table, drag and drop the variable into the axis of the histogram.
- To insert a new variable into a report, creating a new histogram, drag and drop the variable outside of an existing histogram. The new variable can be placed before, between, or after the existing histograms.

Note: To remove a variable, select **Remove** from the red triangle menu.

- The red triangle menu next to Distributions contains options that affect all of the variables. See "Distribution Platform Options" on page 45.
- The red triangle menu next to each variable contains options that affect only that variable. See "Options for Categorical Variables" on page 46 or "Options for Continuous Variables" on page 47. If you hold down the Control key and select a variable option, the option applies to all of the variables that have the same modeling type.
- Histograms visually display your data. See "Histograms" on page 39.
- The initial report for a categorical variable contains a Frequencies report. See "The Frequencies Report" on page 42.
- The initial report for a continuous variable contains a Quantiles and a Summary Statistics report. See "The Quantiles Report" on page 42 and "The Summary Statistics Report" on page 42.

Histograms

Histograms visually display your data. For categorical (nominal or ordinal) variables, the histogram shows a bar for each level of the ordinal or nominal variable. For continuous variables, the histogram shows a bar for grouped values of the continuous variable.

Highlighting data Click on a histogram bar or an outlying point in the graph. The corresponding rows are highlighted in the data table, and corresponding sections of other histograms are also highlighted, if applicable. See "Highlight Bars and Select Rows" on page 41.

Creating a subset Double-click on a histogram bar, or right-click on a histogram bar and select **Subset**. A new data table is created that contains only the selected data.

Resizing the entire histogram Hover over the histogram borders until you see a double-sided arrow. Then click and drag the borders.

Rescaling the axis (Continuous variables only) Click and drag on an axis to rescale it.

Alternatively, hover over the axis until you see a hand. Then, double-click on the axis and set the parameters in the Axis Specification window.

Resizing histogram bars (Continuous variables only) There are multiple options to resize histogram bars. See "Resize Histogram Bars for Continuous Variables" on page 40.

Specifying your selection Specify the data that you select in multiple histograms. See "Specify Your Selection in Multiple Histograms" on page 41.

To see additional options for the histogram or the associated data table:

- Right-click on a histogram. For details, see the *Using JMP* book.
- Right-click on an axis. You can add a label or modify the axis. For details, see the JMP Reports chapter in the *Using JMP* book.

- Click on the red triangle next to the variable, and select **Histogram Options**. Options are slightly different depending on the variable modeling type. See "Options for Categorical Variables" on page 46 or "Options for Continuous Variables" on page 47.

Resize Histogram Bars for Continuous Variables

Resize histogram bars for continuous variables by using the following:

- the Grabber (hand) tool
- the **Set Bin Width** option
- the **Increment** option

Use the Grabber Tool

The Grabber tool is a quick way to explore your data.

1. Select **Tools > Grabber**.

   Note: (Windows only) To see the menu bar, you might need to hover over the bar below the window title. You can also change this setting in **File > Preferences > Windows Specific**.

2. Place the grabber tool anywhere in the histogram.
3. Click and drag the histogram bars.

Think of each bar as a bin that holds a number of observations:

- Moving the hand to the left increases the bin width and combines intervals. The number of bars decreases as the bar size increases.
- Moving the hand to the right decreases the bin width, producing more bars.
- Moving the hand up or down shifts the bin locations on the axis, which changes the contents and size of each bin.

Use the Set Bin Width Option

The **Set Bin Width** option is a more precise way to set the width for all bars in a histogram. To use the Set Bin Width option, from the red triangle menu for the variable, select **Histogram Options > Set Bin Width**. Change the bin width value.

Use the Increment Option

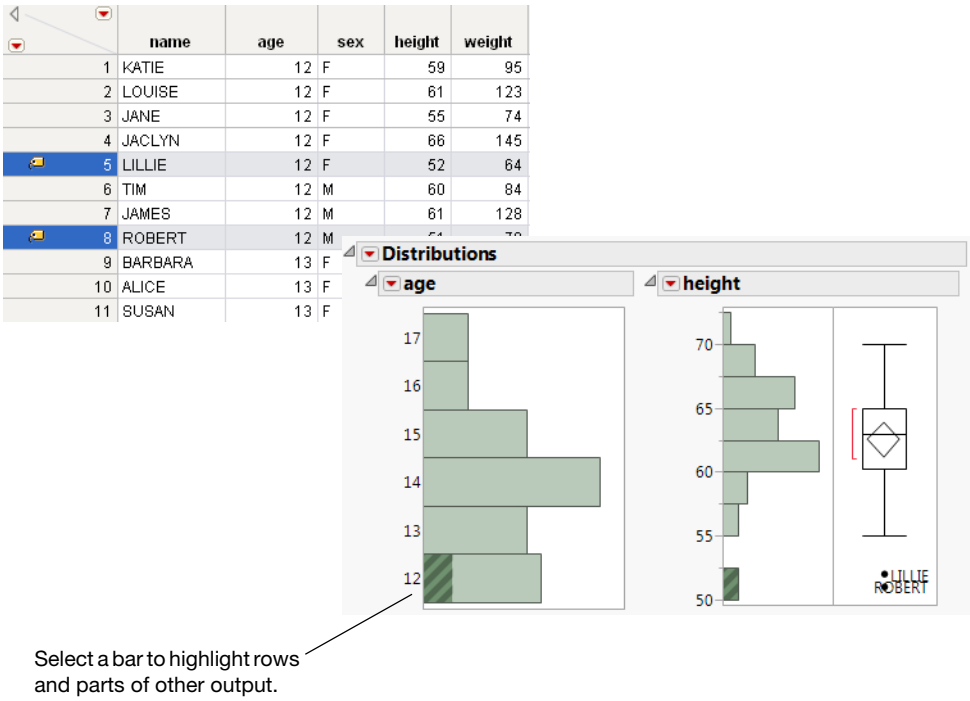
The **Increment** option is another precise way to set the bar width. To use the **Increment** option, double-click on the axis, and change the Increment value.

Highlight Bars and Select Rows

Clicking on a histogram bar highlights the bar and selects the corresponding rows in the data table. The appropriate portions of all other graphical displays also highlight the selection. Figure 3.6 shows the results of highlighting a bar in the height histogram. The corresponding rows are selected in the data table.

Tip: To deselect specific histogram bars, press the Control key and click the highlighted bars.

Figure 3.6 Highlighting Bars and Rows



Specify Your Selection in Multiple Histograms

Extend or narrow your selection in histograms as follows:

- To extend your selection, hold down the Shift key and select another bar. This is the equivalent of using an *or* operator.
- To narrow your selection, hold down the Control and Alt keys (Windows) or Command and Alt keys (Macintosh) and select another bar. This is the equivalent of using an *and* operator.

Related Information

- "Example of Selecting Data in Multiple Histograms" on page 66

The Frequencies Report

For nominal and ordinal variables, the Frequencies report lists the levels of the variables, along with the associated frequency of occurrence and probabilities.

For each level of a categorical (nominal or ordinal) variable, the Frequencies report contains the information described in the following list. Missing values are omitted from the analysis.

Tip: Click a value in the Frequencies report to select the corresponding data in the histogram and data table.

Level Lists each value found for a response variable.

Count Lists the number of rows found for each level of a response variable. If you use a Freq variable, the Count is the sum of the Freq variables for each level of the response variable.

Prob Lists the probability (or proportion) of occurrence for each level of a response variable. The probability is computed as the count divided by the total frequency of the variable, shown at the bottom of the table.

StdErr Prob Lists the standard error of the probabilities. This column might be hidden. To show the column, right-click in the table and select **Columns > StdErr Prob**.

Cum Prob Contains the cumulative sum of the column of probabilities. This column might be hidden. To show the column, right-click in the table and select **Columns > Cum Prob**.

The Quantiles Report

For continuous variables, the Quantiles report lists the values of selected quantiles (sometimes called percentiles).

Related Information

- "Quantiles" on page 73

The Summary Statistics Report

For continuous variables, the Summary Statistics report displays the mean, standard deviation, and other summary statistics. You can control which statistics appear in this report by selecting **Customize Summary Statistics** from the red triangle menu next to Summary Statistics.

Tip: To specify which summary statistics show in the report each time you run a Distribution analysis for a continuous variable, select **File > Preferences > Platforms > Distribution Summary Statistics**, and select the ones you want to appear.

- "Description of the Summary Statistics Report" describes the statistics that appear by default.
- "Additional Summary Statistics" describes additional statistics that you can add to the report using the **Customize Summary Statistics** window.

Description of the Summary Statistics Report

Mean Estimates the expected value of the underlying distribution for the response variable, which is the arithmetic average of the column's values. It is the sum of the non-missing values divided by the number of non-missing values.

Std Dev The normal distribution is mainly defined by the mean and standard deviation. These parameters provide an easy way to summarize data as the sample becomes large:

- 68% of the values are within one standard deviation of the mean
- 95% of the values are within two standard deviations of the mean
- 99.7% of the values are within three standard deviations of the mean

Std Err Mean The standard error of the mean, which estimates the standard deviation of the distribution of the mean.

Upper 95% Mean and Lower 95% Mean Are 95% confidence limits about the mean. They define an interval that is very likely to contain the true population mean.

N Is the total number of nonmissing values.

Additional Summary Statistics

Sum Weight The sum of a column assigned to the role of Weight (in the launch window). Sum Wgt is used in the denominator for computations of the mean instead of *N*.

Sum The sum of the response values.

Variance The sample variance, and the square of the sample standard deviation.

Skewness Measures sidedness or symmetry.

Kurtosis Measures peakedness or heaviness of tails.

CV The percent coefficient of variation. It is computed as the standard deviation divided by the mean and multiplied by 100. The coefficient of variation can be used to assess relative variation, for example when comparing the variation in data measured in different units or with different magnitudes.

N Missing The number of missing observations.

N Zero The number of zero values.

N Unique The number of unique values.

Uncorrected SS The uncorrected sum of squares or sum of values squared.

Corrected SS The corrected sum of squares or sum of squares of deviations from the mean.

Autocorrelation (Appears only if you have not specified a Frequency variable.) First autocorrelation that tests if the residuals are correlated across the rows. This test helps detect non-randomness in the data.

Minimum Represents the 0 percentile of the data.

Maximum Represents the 100 percentile of the data.

Median Represents the 50th percentile of the data.

Mode The value that occurs most often in the data. If there are multiple modes, the smallest mode appears.

Trimmed Mean The mean calculated after removing the smallest p% and the largest p% of the data. The value of p is entered in the **Enter trimmed mean percent** text box at the bottom of the window. The Trimmed Mean option is not available if you have specified a Weight variable.

Geometric Mean The *n*th root of the product of the data. Zero and negative numbers are treated like missing values. For example, you might want to compare two companies based on varying metrics that come from different ranges. The statistic is also helpful when the data contains a large value in a skewed distribution.

Range The difference between the maximum and minimum of the data.

Interquartile Range The difference between the 3rd and 1st quartiles.

Median Absolute Deviation (Does not appear if you have specified a Weight variable.) The median of the absolute deviations from the median.

Robust Mean The robust mean, calculated in a way that is resistant to outliers, using Huber's M-estimation. See Huber and Ronchetti (2009).

Robust Std Dev The robust standard deviation, calculated in a way that is resistant to outliers, using Huber's M-estimation. See Huber and Ronchetti (2009).

Enter (1-alpha) for mean confidence interval Specify the alpha level for the mean confidence interval.

Enter trimmed mean percent Specify the trimmed mean percentage. The percentage is trimmed off each side of the data.

Summary Statistics Options

The red triangle menu next to Summary Statistics contains these options:

Customize Summary Statistics Select which statistics you want to appear from the list. You can select or deselect all summary statistics.

Show All Modes Shows all of the modes if there are multiple modes.

Related Information

- "Summary Statistics" on page 73

Distribution Platform Options

The red triangle menu next to Distributions contains options that affect all of the reports and graphs in the Distribution platform.

Uniform Scaling Scales all axes with the same minimum, maximum, and intervals so that the distributions can be easily compared.

Stack Changes the orientation of the histogram and the reports to horizontal and stacks the individual distribution reports vertically. Deselect this option to return the report window to its original layout.

Arrange in Rows Enter the number of plots that appear in a row. This option helps you view plots vertically rather than in one wide row.

Save for Adobe Flash platform (.SWF) Saves the histograms as .swf files that are Adobe Flash player compatible. Use these files in presentations and in Web pages. An HTML page is also saved that shows you the correct code for using the resulting .swf file.

For more information about this option, go to http://www.jmp.com/support/swfhelp/en.

See the JMP Reports chapter in the *Using JMP* book for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Options for Categorical Variables

The red triangle menus next to each variable in the report window contain additional options that apply to the variable. This section describes the options that are available for categorical (nominal or ordinal) variables.

To see the options that are available for continuous variables, see "Options for Continuous Variables" on page 47.

Display Options Submenu for Categorical Variables

Frequencies Shows or hides the Frequencies report. See "The Frequencies Report" on page 42.

Horizontal Layout Changes the orientation of the histogram and the reports to vertical or horizontal.

Axes on Left Moves the **Count**, **Prob**, and **Density** axes to the left instead of the right.

This option is applicable only if **Horizontal Layout** is selected.

Histogram Options Submenu for Categorical Variables

Histogram Shows or hides the histogram. See "Histograms" on page 39.

Vertical Changes the orientation of the histogram from a vertical to a horizontal orientation.

Std Error Bars Draws the standard error bar on each level of the histogram.

Separate Bars Separates the histogram bars.

Histogram Color Changes the color of the histogram bars.

Count Axis Adds an axis that shows the frequency of column values represented by the histogram bars.

Prob Axis Adds an axis that shows the proportion of column values represented by histogram bars.

Density Axis Adds an axis that shows the length of the bars in the histogram.

The count and probability axes are based on the following calculations:

prob = (bar width)*density

count = (bar width)*density*(total count)

Show Percents Labels the percent of column values represented by each histogram bar.

Show Counts Labels the frequency of column values represented by each histogram bar.

Mosaic Plot Displays a mosaic bar chart for each nominal or ordinal response variable. A mosaic plot is a stacked bar chart where each segment is proportional to its group's frequency count.

Order By Reorders the histogram, mosaic plot, and Frequencies report in ascending or descending order, by count. To save the new order as a column property, use the **Save > Value Ordering** option.

Test Probabilities Displays a report that tests hypothesized probabilities. See "Examples of the Test Probabilities Option" on page 67 for more details.

Confidence Interval This menu contains confidence levels. Select a value that is listed, or select **Other** to enter your own. JMP computes score confidence intervals.

Save Submenu for Categorical Variables

Level Numbers Creates a new column in the data table called Level <colname>. The level number of each observation corresponds to the histogram bar that contains the observation.

Value Ordering (Use with the **Order By** option) Creates a new value ordering column property in the data table, reflecting the new order.

Script to log Displays the script commands to generate the current report in the log window. Select **View > Log** to see the log window.

Remove Permanently removes the variable and all its reports from the Distribution report.

Options for Continuous Variables

The red triangle menus next to each variable in the report window contain additional options that apply to the variable. This section describes the options that are available for continuous variables.

To see the options that are available for categorical (nominal and ordinal) variables, see "Options for Categorical Variables" on page 46.

Display Options Submenu for Continuous Variables

Quantiles Shows or hides the Quantiles report. See "The Quantiles Report" on page 42.

Set Quantile Increment Changes the quantile increment or revert back to the default quantile increment.

Custom Quantiles Sets custom quantiles by values or by increments. You can specify the confidence level and choose whether to compute smoothed empirical likelihood quantiles (for large data sets, this can take some time).

- For details about how the weighted average quantiles are estimated, see "Quantiles" on page 73.
- For details about distribution-free confidence limits for the weighted average quantiles, see section 5.2 in Hahn and Meeker (1991).
- Smoothed empirical likelihood quantiles are based on a kernel density estimate. For more details about how these quantiles and their confidence limits are estimated, see Chen and Hall (1993).

Summary Statistics Shows or hides the Summary Statistics report. See "The Summary Statistics Report" on page 42.

Customize Summary Statistics Adds or removes statistics from the Summary Statistics report. See "The Summary Statistics Report" on page 42.

Horizontal Layout Changes the orientation of the histogram and the reports to vertical or horizontal.

Axes on Left Moves the **Count**, **Prob**, **Density**, and **Normal Quantile Plot** axes to the left instead of the right.

This option is applicable only if **Horizontal Layout** is selected.

Histogram Options Submenu for Continuous Variables

Histogram Shows or hides the histogram. See "Histograms" on page 39.

Shadowgram Replaces the histogram with a shadowgram. To understand a shadowgram, consider that if the bin width of a histogram is changed, the appearance of the histogram changes. A shadowgram overlays histograms with different bin widths. Dominant features of a distribution are less transparent on the shadowgram.

Note that the following options are not available for shadowgrams:

- Std Error Bars
- Show Counts
- Show Percents

Vertical Changes the orientation of the histogram from a vertical to a horizontal orientation.

Std Error Bars Draws the standard error bar on each level of the histogram using the standard error. The standard error bar adjusts automatically when you adjust the number of bars with the hand tool. See "Resize Histogram Bars for Continuous Variables" on page 40, and "Standard Error Bars" on page 72.

Set Bin Width Changes the bin width of the histogram bars. See "Resize Histogram Bars for Continuous Variables" on page 40.

Histogram Color Changes the color of the histogram bars.

Count Axis Adds an axis that shows the frequency of column values represented by the histogram bars.

Note: If you resize the histogram bars, the count axis also resizes.

Prob Axis Adds an axis that shows the proportion of column values represented by histogram bars.

Note: If you resize the histogram bars, the probability axis also resizes.

Density Axis The density is the length of the bars in the histogram. Both the count and probability are based on the following calculations:

prob = (bar width)*density

count = (bar width)*density*(total count)

When looking at density curves that are added by the Fit Distribution option, the density axis shows the point estimates of the curves.

Note: If you resize the histogram bars, the density axis also resizes.

Show Percents Labels the proportion of column values represented by each histogram bar.

Show Counts Labels the frequency of column values represented by each histogram bar.

Normal Quantile Plot Adds a normal quantile plot that shows the extent to which the variable is normally distributed. See "Normal Quantile Plot" on page 50.

Outlier Box Plot Adds an outlier box plot that shows the outliers in your data. See "Outlier Box Plot" on page 51.

Quantile Box Plot Adds a quantile box plot that shows specific quantiles from the Quantiles report. See "Quantile Box Plot" on page 52.

Stem and Leaf Adds a stem and leaf report, which is a variation of the histogram. See "Stem and Leaf" on page 52.

CDF Plot Adds a plot of the empirical cumulative distribution function. See "CDF Plot" on page 53.

Test Mean Performs a one-sample test for the mean. See "Test Mean" on page 53.

Test Std Dev Performs a one-sample test for the standard deviation. See "Test Std Dev" on page 54.

Test Equivalence Performs an equivalence test. See "Test Equivalence" on page 55.

Confidence Interval Choose confidence intervals for the mean and standard deviation. See "Confidence Intervals for Continuous Variables" on page 56.

Prediction Interval Choose prediction intervals for a single observation, or for the mean and standard deviation of the next randomly selected sample. See "Prediction Intervals" on page 58.

Tolerance Interval Computes an interval to contain at least a specified proportion of the population. See "Tolerance Intervals" on page 59.

Capability Analysis Measures the conformance of a process to given specification limits. See "Capability Analysis" on page 59.

Continuous Fit Fits distributions to continuous variables. See "Fit Distributions" on page 61.

Discrete Fit Fits distributions to discrete variables. See "Fit Distributions" on page 61.

Save Saves information about continuous or categorical variables. See "Save Commands for Continuous Variables" on page 57.

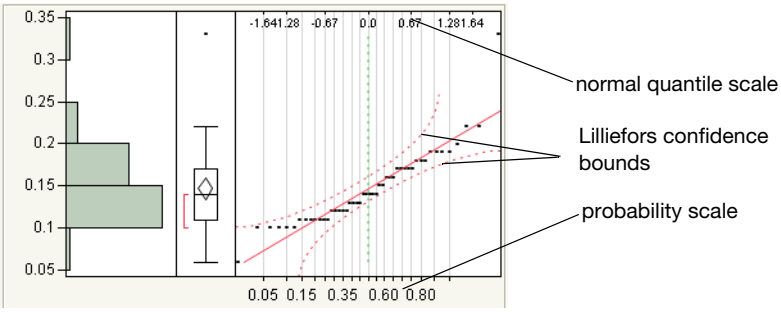
Remove Permanently removes the variable and all its reports from the Distribution report.

Normal Quantile Plot

Use the **Normal Quantile Plot** option to visualize the extent to which the variable is normally distributed. If a variable is normally distributed, the normal quantile plot approximates a diagonal straight line. This type of plot is also called a quantile-quantile plot, or Q-Q plot.

The normal quantile plot also shows Lilliefors confidence bounds (Conover 1980) and probability and normal quantile scales.

Figure 3.7 Normal Quantile Plot



Note the following information:

- The *y*-axis shows the column values.
- The *x*-axis shows the empirical cumulative probability for each value.

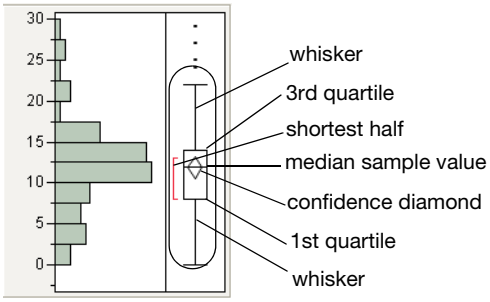
Related Information

- "Normal Quantile Plot" on page 74

Outlier Box Plot

Use the outlier box plot (also called a Tukey outlier box plot) to see the distribution and identify possible outliers. Generally, box plots show selected quantiles of continuous distributions.

Figure 3.8 Outlier Box Plot



Note the following aspects about outlier box plots:

- The horizontal line within the box represents the median sample value.
- The confidence diamond contains the mean and the upper and lower 95% of the mean. If you drew a line through the middle of the diamond, you would have the mean. The top and bottom points of the diamond represent the upper and lower 95% of the mean.
- The ends of the box represent the 25th and 75th quantiles, also expressed as the 1st and 3rd *quartile*, respectively.
- The difference between the 1st and 3rd quartiles is called the *interquartile range*.
- The box has lines that extend from each end, sometimes called *whiskers*. The whiskers extend from the ends of the box to the outermost data point that falls within the distances computed as follows:

  1st quartile - 1.5*(interquartile range)

  3rd quartile + 1.5*(interquartile range)

  If the data points do not reach the computed ranges, then the whiskers are determined by the upper and lower data point values (not including outliers).
- The bracket outside of the box identifies the *shortest half*, which is the most dense 50% of the observations (Rousseuw and Leroy 1987).

Remove Objects from the Outlier Box Plot

To remove the confidence diamond or the shortest half, proceed as follows:

1. Right-click on the outlier box plot and select **Customize**.
2. Click **Box Plot**.

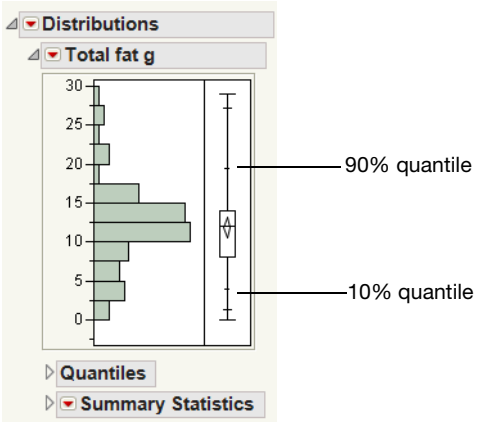
3. Deselect the check box next to **Confidence Diamond** or **Shortest Half**.

For more details about the Customize Graph window, see the JMP Reports chapter in the *Using JMP* book.

Quantile Box Plot

The Quantile Box Plot displays specific quantiles from the Quantiles report. If the distribution is symmetric, the quantiles in the box plot are approximately equidistant from each other. At a glance, you can see whether the distribution is symmetric. For example, if the quantile marks are grouped closely at one end, but have greater spacing at the other end, the distribution is skewed toward the end with more spacing. See Figure 3.9.

Figure 3.9 Quantile Box Plot



Quantiles are values where the *p*th quantile is larger than *p*% of the values. For example, 10% of the data lies below the 10th quantile, and 90% of the data lies below the 90th quantile.

Stem and Leaf

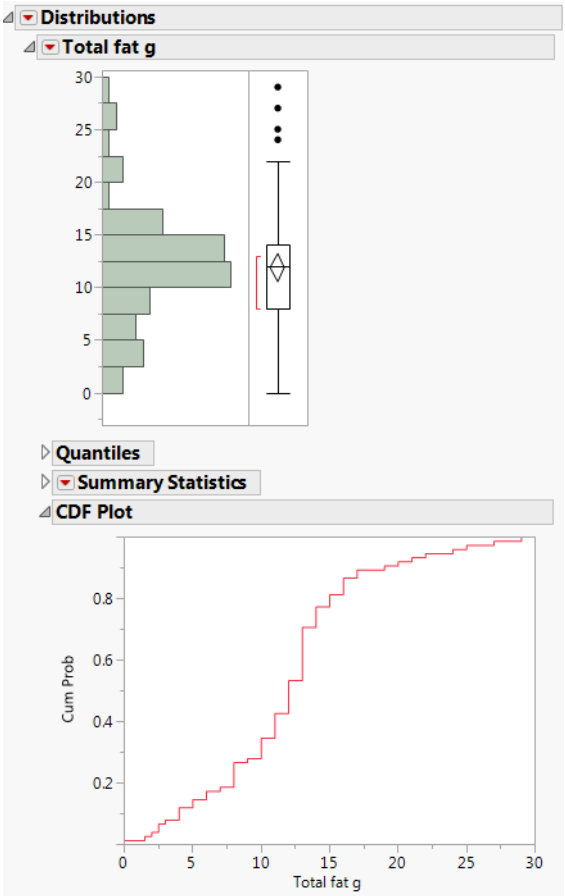
Each line of the plot has a Stem value that is the leading digit of a range of column values. The Leaf values are made from the next-in-line digits of the values. You can see the data point by joining the stem and leaf. In some cases, the numbers on the stem and leaf plot are rounded versions of the actual data in the table. The stem-and-leaf plot actively responds to clicking and the brush tool.

Note: The stem-and-leaf plot does not support fractional frequencies.

CDF Plot

The CDF plot creates a plot of the empirical cumulative distribution function. Use the CDF plot to determine the percent of data that is at or below a given value on the *x*-axis.

Figure 3.10 CDF Plot



For example, in this CDF plot, approximately 30% of the data is less than a total fat value of 10 grams.

Test Mean

Use the **Test Mean** window to specify options for and perform a one-sample test for the mean. If you specify a value for the standard deviation, a *z*-test is performed. Otherwise, the sample standard deviation is used to perform a *t*-test. You can also request the nonparametric Wilcoxon Signed-Rank test.

Use the **Test Mean** option repeatedly to test different values. Each time you test the mean, a new Test Mean report appears.

Description of the Test Mean Report

Statistics that are calculated for Test Mean

t Test (or z Test) Lists the value of the test statistic and the *p*-values for the two-sided and one-sided alternatives.

Signed-Rank (Only appears if the Wilcoxon Signed-Rank test is selected.) Lists the value of the Wilcoxon signed-rank statistic followed by *p*-values for the two-sided and one-sided alternatives. The test uses the Pratt method to address zero values. This is a nonparametric test whose null hypothesis is that the median equals the postulated value. It assumes that the distribution is symmetric. See "Wilcoxon Signed Rank Test" on page 75.

Probability values

Prob>|t| The probability of obtaining an absolute *t*-value by chance alone that is greater than the observed *t*-value when the population mean is equal to the hypothesized value. This is the *p*-value for observed significance of the two-tailed *t*-test.

Prob>t The probability of obtaining a *t*-value greater than the computed sample *t* ratio by chance alone when the population mean is not different from the hypothesized value. This is the *p*-value for an upper-tailed test.

Prob<t The probability of obtaining a *t*-value less than the computed sample *t* ratio by chance alone when the population mean is not different from the hypothesized value. This is the *p*-value for a lower-tailed test.

Descriptions of the Test Mean Options

PValue animation Starts an interactive visual representation of the p-value. Enables you to change the hypothesized mean value while watching how the change affects the p-value.

Power animation Starts an interactive visual representation of power and beta. You can change the hypothesized mean and sample mean while watching how the changes affect power and beta.

Remove Test Removes the mean test.

Test Std Dev

Use the **Test Std Dev** option to perform a one-sample test for the standard deviation (details in Neter, Wasserman, and Kutner 1990). Use the **Test Std Dev** option repeatedly to test different values. Each time you test the standard deviation, a new Test Standard Deviation report appears.

Test Statistic Provides the value of the Chi-square test statistic. See "Standard Deviation Test" on page 77.

Min PValue The probability of obtaining a more extreme Chi-square value by chance alone when the population standard deviation does not differ from the hypothesized value. See "Standard Deviation Test" on page 77.

Prob>ChiSq The probability of obtaining a Chi-square value greater than the computed sample Chi-square by chance alone when the population standard deviation is not different from the hypothesized value. This is the *p*-value for observed significance of a one-tailed *t*-test.

Prob<ChiSq The probability of obtaining a Chi-square value less than the computed sample Chi-square by chance alone when the population standard deviation is not different from the hypothesized value. This is the *p*-value for observed significance of a one-tailed *t*-test.

Test Equivalence

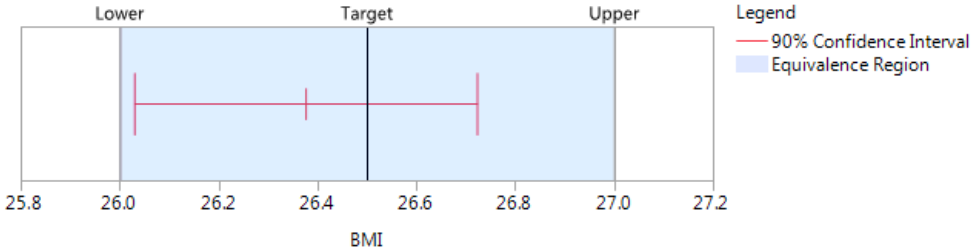
The equivalence test assesses whether a population mean is equivalent to a hypothesized value. You must define a threshold difference that is considered equivalent to no difference. The Test Equivalence option uses the Two One-Sided Tests (TOST) approach. Two one-sided *t*-tests are constructed for the null hypotheses that the difference between the true mean and the hypothesized value exceeds the threshold. If both null hypotheses are rejected, this implies that the true difference does not exceed the threshold. You conclude that the mean can be considered practically equivalent to the hypothesized value.

When you select the Test Equivalence option, you specify the Hypothesized Mean, the threshold difference (Difference Considered Practically Zero), and the Confidence Level. The Confidence Level is 1 - alpha, where alpha is the significance level for each one-sided test.

The Test Equivalence report in Figure 3.11 is for the variable BMI in the Diabetes.jmp sample data table. The Hypothesized Mean is 26.5 and the Difference Considered Practically Zero is specified as 0.5.

Figure 3.11 Equivalence Test Report

Test Equivalence



Mean	Lower Bound	Upper Bound	Lower 90%	Upper 90%
26.37579	26	27	26.0294	26.72218

Test	t Ratio	p-Value	Null Hypothesis
Upper Threshold	-2.97	0.0016*	Mean ≤ 26
Lower Threshold	1.7882	0.0372*	Mean ≥ 27
Max over both		0.0372*	

Mean is equivalent to 26.5.

The report shows the following:

- A plot of your defined equivalence region that shows the Target and boundaries, defined by vertical lines labeled Lower and Upper.
- A confidence interval for the hypothesized mean. This confidence interval is a 1 - 2*alpha level interval.
- A table that shows the calculated mean, the specified lower and upper bounds, and a (1 - 2*alpha) level confidence interval for the mean.
- A table that shows the results of the two one-sided tests.
- A note that summarizes the results, and states whether the mean can be considered equivalent to the Target value.

Confidence Intervals for Continuous Variables

The **Confidence Interval** options display confidence intervals for the mean and standard deviation. The **0.90**, **0.95**, and **0.99** options compute two-sided confidence intervals for the mean and standard deviation. Use the **Confidence Interval > Other** option to select a confidence level, and select one-sided or two-sided confidence intervals. You can also type a known sigma. If you use a known sigma, the confidence interval for the mean is based on *z*-values rather than *t*-values.

The Confidence Intervals report shows the mean and standard deviation parameter estimates with upper and lower confidence limits for 1 - α.

Save Commands for Continuous Variables

Use the Save menu commands to save information about continuous variables. Each Save command generates a new column in the current data table. The new column is named by appending the variable name (denoted <colname> in the following definitions) to the Save command name. See Table 3.1.

Select the Save commands repeatedly to save the same information multiple times under different circumstances, such as before and after combining histogram bars. If you use a Save command multiple times, the column name is numbered (name1, name2, and so on) to ensure unique column names.

Table 3.1 Descriptions of Save Commands

Command	Column Added to Data Table	Description
Level Numbers	Level <colname>	The level number of each observation corresponds to the histogram bar that contains the observation. The histogram bars are numbered from low to high, beginning with 1. **Note:** To maintain source information, value labels are added to the new column, but they are turned off by default.
Level Midpoints	Midpoint <colname>	The midpoint value for each observation is computed by adding half the level width to the lower level bound. **Note:** To maintain source information, value labels are added to the new column, but they are turned off by default.
Ranks	Ranked <colname>	Provides a ranking for each of the corresponding column's values starting at 1. Duplicate response values are assigned consecutive ranks in order of their occurrence in the data table.
Ranks Averaged	RankAvgd <colname>	If a value is unique, then the averaged rank is the same as the rank. If a value occurs *k* times, the average rank is computed as the sum of the value's ranks divided by *k*.

Table 3.1 Descriptions of Save Commands *(Continued)*

Command	Column Added to Data Table	Description
Prob Scores	Prob <colname>	For *N* nonmissing scores, the probability score of a value is computed as the averaged rank of that value divided by *N* + 1. This column is similar to the empirical cumulative distribution function.
Normal Quantiles	N-Quantile <colname>	Saves the Normal quantiles to the data table. See "Normal Quantile Plot" on page 74.
Standardized	Std <colname>	Saves standardized values to the data table. See "Saving Standardized Data" on page 77.
Centered	Centered <colname>	Saves values for centering on zero.
Spec Limits	(none)	Stores the specification limits applied in a capability analysis as a column property of the corresponding column in the current data table. Automatically retrieves and displays the specification limits when you repeat the capability analysis.
Script to Log	(none)	Prints the script to the log window. Run the script to recreate the analysis.

Prediction Intervals

Prediction intervals concern a single observation, or the mean and standard deviation of the next randomly selected sample. The calculations assume that the given sample is selected randomly from a normal distribution. Select one-sided or two-sided prediction intervals.

When you select the **Prediction Interval** option for a variable, the Prediction Intervals window appears. Use the window to specify the confidence level, the number of future samples, and either a one-sided or two-sided limit.

Related Information

- "Prediction Intervals" on page 78
- "Example of Prediction Intervals" on page 69

Tolerance Intervals

A tolerance interval contains at least a specified proportion of the population. It is a confidence interval for a specified proportion of the population, not the mean, or standard deviation. Complete discussions of tolerance intervals are found in Hahn and Meeker (1991) and in Tamhane and Dunlop (2000).

When you select the **Tolerance Interval** option for a variable, the Tolerance Intervals window appears. Use the window to specify the confidence level, the proportion to cover, and either a one-sided or two-sided limit. The calculations are based on the assumption that the given sample is selected randomly from a normal distribution.

Related Information

- "Tolerance Intervals" on page 78
- "Example of Tolerance Intervals" on page 70

Capability Analysis

The **Capability Analysis** option measures the conformance of a process to given specification limits. By default, capability is calculated based on a normal distribution. However, if you fit a distribution (continuous or discrete), then the capability report is produced based on that fit. To use a fit distribution, select **Spec Limits** or set the **Spec Limits** column property. When you select the **Capability Analysis** option for a variable, the Capability Analysis window appears. Use the window to enter specification limits, distribution type, and information about sigma.

Note: To save the specification limits to the data table as a column property, select **Save > Spec Limits**. When you repeat the capability analysis, the saved specification limits are automatically retrieved.

The Capability Analysis report is organized into two sections: Capability Analysis and the distribution type (Long Term Sigma, Specified Sigma, and so on).

Description of the Capability Analysis Window

<Distribution type> By default, the normal distribution is assumed when calculating the capability statistics and the percent out of the specification limits. To perform a capability analysis on non-normal distributions, see the description of Spec Limits under "Capability Analysis" on page 79.

<Sigma type> Estimates sigma (σ) using the selected methods. See "Capability Analysis" on page 79.

Description of the Capability Analysis Report

Specification Lists the specification limits.

Value Lists the values that you specified for each specification limit and the target.

Portion and % Actual Portion labels describe the numbers in the % Actual column, as follows:

- Below LSL gives the percentage of the data that is below the lower specification limit.
- Above USL gives the percentage of the data that is above the upper specification limit.
- Total Outside gives the total percentage of the data that is either below LSL or above USL.

Capability Type of process capability indices. See "Description of the Capability Analysis Options" on page 60.

Note: There is a preference for Capability called **Ppk Capability Labeling** that labels the long-term capability output with P_{pk} labels. Open the Preference window (**File > Preferences**), then select **Platforms > Distribution** to see this preference.

Index Process capability index values.

Upper CI Upper confidence interval.

Lower CI Lower confidence interval.

Portion and Percent Portion labels describe the numbers in the Percent column, as follows:

- Below LSL gives the percentage of the fitted distribution that is below the lower specification limit.
- Above USL gives the percentage of the fitted distribution that is above the upper specification limit.
- Total Outside gives the total percentage of the fitted distribution that is either below LSL or above USL.

PPM (parts per million) The PPM value is the Percent column multiplied by 10,000.

Sigma Quality Sigma Quality is frequently used in Six Sigma methods, and is also referred to as the *process sigma*. See "Capability Analysis" on page 79.

Description of the Capability Analysis Options

Z Bench Shows the values (represented by Index) of the Benchmark Z statistics. According to the AIAG Statistical Process Control manual, Z represents the number of standard deviation units from the process average to a value of interest such as an engineering specification. When used in capability assessment, Z USL is the distance to the upper specification limit and Z LSL is the distance to the lower specification limit. See "Capability Analysis" on page 79.

Capability Animation Interactively change the specification limits and the process mean to see the effects on the capability statistics. This option is available only for capability analyses based on the Normal distribution.

Related Information

- "Capability Analysis" on page 79
- "Example of Capability Analysis" on page 71

Fit Distributions

Use the **Continuous** or **Discrete Fit** options to fit a distribution to a continuous or discrete variable.

A curve is overlaid on the histogram, and a Parameter Estimates report is added to the report window. A red triangle menu contains additional options. See "Fit Distribution Options" on page 63.

Note: The Life Distribution platform also contains options for distribution fitting that might use different parameterizations and allow for censoring. See the Life Distribution chapter in the *Reliability and Survival Methods* book.

Continuous Fit

Use the **Continuous Fit** options to fit the following distributions to a continuous variable.

- The Normal distribution is often used to model measures that are symmetric with most of the values falling in the middle of the curve. JMP uses the unbiased estimate when determining the parameters for the Normal distribution.
- The LogNormal distribution is often used to model values that are constrained by zero but have a few very large values. The LogNormal distribution can be obtained by exponentiating the Normal distribution. JMP uses the maximum likelihood estimation when determining the parameters for the LogNormal distribution.
- The Weibull, Weibull with threshold, and Extreme Value distributions often provide a good model for estimating the length of life, especially for mechanical devices and in biology.
- The Exponential distribution is especially useful for describing events that randomly occur over time, such as survival data. The exponential distribution might also be useful for modeling elapsed time between the occurrence of non-overlapping events, such as the time between a user's computer query and response of the server, the arrival of customers at a service desk, or calls coming in at a switchboard.

- The Gamma distribution is bound by zero and has a flexible shape.
- The Beta distribution is useful for modeling the behavior of random variables that are constrained to fall in the interval 0,1. For example, proportions always fall between 0 and 1.
- The Normal Mixtures distribution fits a mixture of normal distributions. This flexible distribution is capable of fitting multi-modal data. You can also fit two or more distributions by selecting the **Normal 2 Mixture**, **Normal 3 Mixture**, or **Other** options.
- The Smooth Curve distribution... A smooth curve is fit using nonparametric density estimation (kernel density estimation). The smooth curve is overlaid on the histogram and a slider appears beneath the plot. Control the amount of smoothing by changing the kernel standard deviation with the slider. The initial Kernel Std estimate is calculated from the standard deviation of the data.
- The Johnson Su, Johnson Sb, and Johnson Sl Distributions are useful for its data-fitting capabilities because it supports every possible combination of skewness and kurtosis.
- The Generalized Log (Glog) distribution is useful for fitting data that are rarely normally distributed and often have non-constant variance, like biological assay data.

Comparing All Distributions

The **All** option fits all applicable continuous distributions to a variable. The Compare Distributions report contains statistics about each fitted distribution. Use the check boxes to show or hide a fit report and overlay curve for the selected distribution. By default, the best fit distribution is selected.

The Show Distribution list is sorted by AICc in ascending order.

If your variable contains negative values, the Show Distribution list does not include those distributions that require data with positive values. Only continuous distributions are fitted by this command. Distributions with threshold parameters, like Beta and Johnson Sb, are not included in the list of possible distributions.

Related Information

- "Continuous Fit Distributions" on page 83
- "Fitted Quantiles" on page 92
- "Fit Distribution Options" on page 92

Discrete Fit

Use the Discrete Fit options to fit a distribution (such as Poisson or Binomial) to a discrete variable. The available distributions are as follows:

- Poisson

- Binomial
- Gamma Poisson
- Beta Binomial

Related Information

- "Discrete Fit Distributions" on page 89
- "Fitted Quantiles" on page 92
- "Fit Distribution Options" on page 92

Fit Distribution Options

Each fitted distribution report has a red triangle menu that contains additional options.

Diagnostic Plot Creates a quantile or a probability plot. See "Diagnostic Plot" on page 64.

Density Curve Uses the estimated parameters of the distribution to overlay a density curve on the histogram.

Goodness of Fit Computes the goodness of fit test for the fitted distribution. See "Goodness of Fit" on page 65.

Fix Parameters Enables you to fix parameters and re-estimate the non-fixed parameters. An Adequacy LR (likelihood ratio) Test report also appears, which tests your new parameters to determine whether they fit the data.

Quantiles Returns the un-scaled and un-centered quantiles for the specific lower probability values that you specify.

Set Spec Limits for K Sigma Use this option when you do not know the specification limits for a process and you want to use its distribution as a guideline for setting specification limits.

Usually specification limits are derived using engineering considerations. If there are no engineering considerations, and if the data represents a trusted benchmark (well behaved process), then quantiles from a fitted distribution are often used to help set specification limits. See "Fit Distribution Options" on page 92.

Spec Limits Computes generalizations of the standard capability indices, based on the specification limits and target you specify. See "Spec Limits" on page 65.

Save Fitted Quantiles Saves the fitted quantile values as a new column in the current data table. See "Fitted Quantiles" on page 92.

Save Density Formula Creates a new column in the current data table that contains fitted values that have been computed by the density formula. The density formula uses the estimated parameter values.

Save Spec Limits Saves the specification limits as a column property. See "Fit Distribution Options" on page 92.

Save Transformed Creates a new column and saves a formula. The formula can transform the column to normality using the fitted distribution. This option is available only when one of the Johnson distributions or the Glog distribution is fit.

Remove Fit Removes the distribution fit from the report window.

Diagnostic Plot

The **Diagnostic Plot** option creates a quantile or a probability plot. Depending on the fitted distribution, the plot is one of four formats.

The fitted quantiles versus the data

- Weibull with threshold
- Gamma
- Beta
- Poisson
- GammaPoisson
- Binomial
- BetaBinomial

The fitted probability versus the data

- Normal
- Normal Mixtures
- Exponential

The fitted probability versus the data on log scale

- Weibull
- LogNormal
- Extreme Value

The fitted probability versus the standard normal quantile

- Johnson Sl
- Johnson Sb
- Johnson Su
- Glog

The following options are available in the Diagnostic Plot red triangle menu:

Rotate Reverses the *x*- and *y*-axes.

Confidence Limits Draws Lilliefors 95% confidence limits for the Normal Quantile plot, and 95% equal precision bands with $a = 0.001$ and $b = 0.99$ for all other quantile plots (Meeker and Escobar (1998)).

Line of Fit Draws the straight diagonal reference line. If a variable fits the selected distribution, the values fall approximately on the reference line.

Median Reference Line Draws a horizontal line at the median of the response.

Goodness of Fit

The **Goodness of Fit** option computes the goodness of fit test for the fitted distribution. The goodness of fit tests are not Chi-square tests, but are EDF (Empirical Distribution Function) tests. EDF tests offer advantages over the Chi-square tests, including improved power and invariance with respect to histogram midpoints.

- For Normal distributions, the Shapiro-Wilk test for normality is reported when the sample size is less than or equal to 2000, and the KSL test is computed for samples that are greater than 2000.
- For discrete distributions (such as Poisson distributions) that have sample sizes less than or equal to 30, the Goodness of Fit test is formed using two one-sided exact Kolmogorov tests combined to form a near exact test. For details, see Conover 1972. For sample sizes greater than 30, a Pearson Chi-squared goodness of fit test is performed.

Related Information

- "Fit Distribution Options" on page 92

Spec Limits

The **Spec Limits** option launches a window requesting specification limits and target, and then computes generalizations of the standard capability indices. This is done using the fact that for the normal distribution, 3σ is both the distance from the lower 0.135 percentile to median (or mean) and the distance from the median (or mean) to the upper 99.865 percentile. These percentiles are estimated from the fitted distribution, and the appropriate percentile-to-median distances are substituted for 3σ in the standard formulas.

Related Information

- "Fit Distribution Options" on page 92

Additional Examples of the Distribution Platform

This section contains additional examples using the Distribution platform.

Example of Selecting Data in Multiple Histograms

1. Select **Help > Sample Data Library** and open Companies.jmp.
2. Select **Analyze > Distribution**.
3. Select Type and Size Co and click **Y, Columns.**
4. Click **OK**.

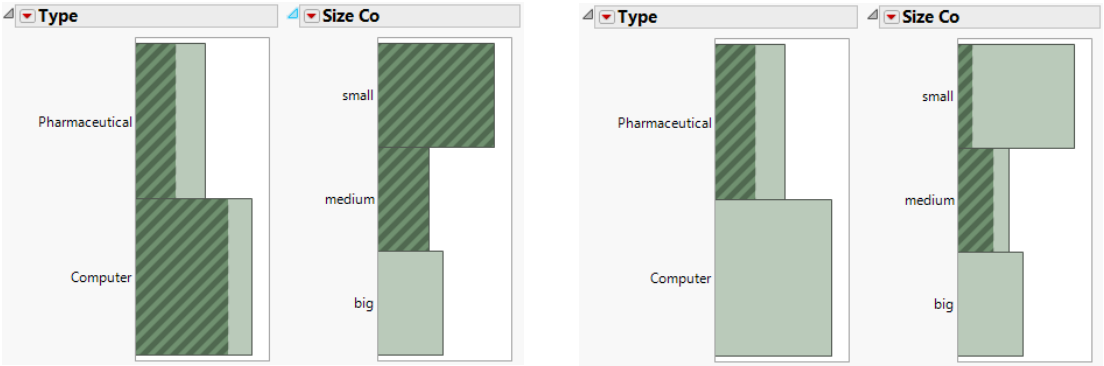
   You want to see the type distribution of companies that are small.
5. Click on the bar next to small.

   You can see that there are more small computer companies than there are pharmaceutical companies. To broaden your selection, add medium companies.
6. Hold down the Shift key. In the Size Co histogram, click on the bar next to medium.

   You can see the type distribution of small and medium sized companies. See Figure 3.12 at left. To narrow your selection, you want to see the small and medium pharmaceutical companies only.
7. Hold down the Control and Shift keys (on Windows) or the Command and Shift keys (on Macintosh). In the Type histogram, click in the Computer bar to deselect it.

   You can see how many of the small and medium companies are pharmaceutical companies. See Figure 3.12 at right.

Figure 3.12 Selecting Data in Multiple Histograms



Broaden the selection using the Shift key.

Narrow the selection using Control + Shift (Windows) or Command + Shift (Macintosh).

Examples of the Test Probabilities Option

Initiate a test probability report for a variable with more than two levels:

1. Select **Help > Sample Data Library** and open VA Lung Cancer.jmp.
2. Select **Analyze > Distribution**.
3. Select Cell Type and click **Y, Columns**.
4. Click **OK**.
5. From the red triangle menu next to Cell Type, select **Test Probabilities**.

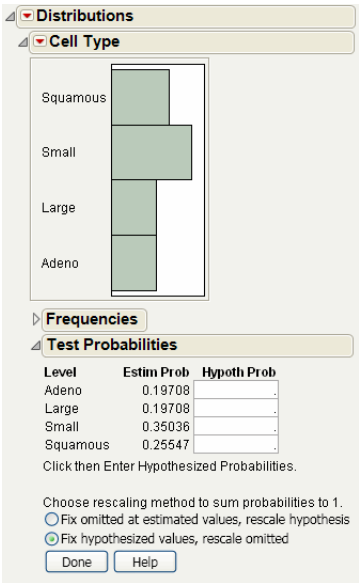
   See Figure 3.13 at left.

Initiate a test probability report for a variable with exactly two levels:

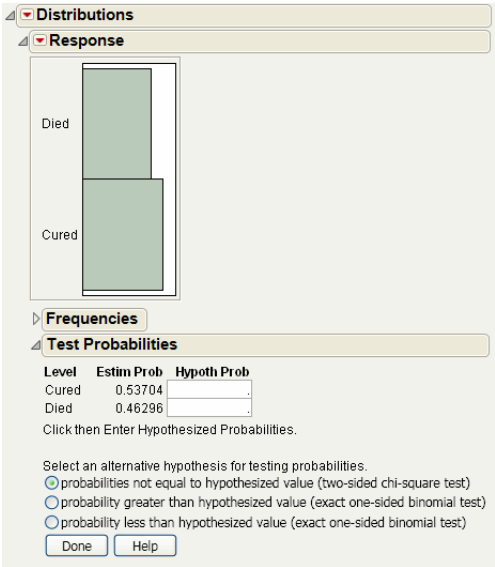
1. Select **Help > Sample Data Library** and open Penicillin.jmp.
2. Select **Analyze > Distribution**.
3. Select Response and click **Y, Columns**.
4. Click **OK**.
5. From the red triangle menu next to Response, select **Test Probabilities**.

   See Figure 3.13 at right.

Figure 3.13 Examples of Test Probabilities Options



report options for a variable with more than two levels



report options for a variable with exactly two levels

Example of Generating the Test Probabilities Report

To generate a test probabilities report for a variable with more than two levels:

1. Refer to Figure 3.13 at left. Type 0.25 in all four Hypoth Prob fields.
2. Click the **Fix hypothesized values, rescale omitted** button.
3. Click **Done**.

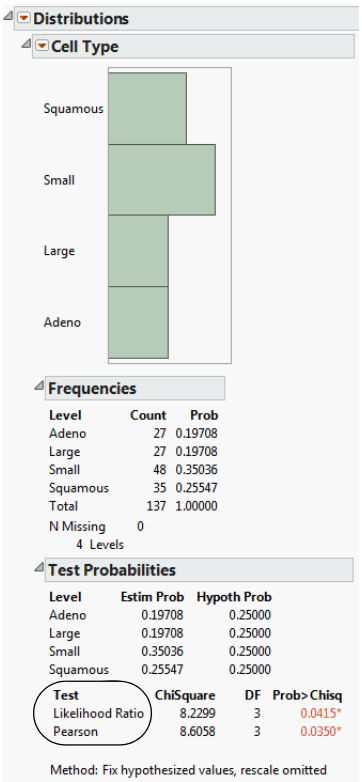
Likelihood Ratio and Pearson Chi-square tests are calculated. See Figure 3.14 at left.

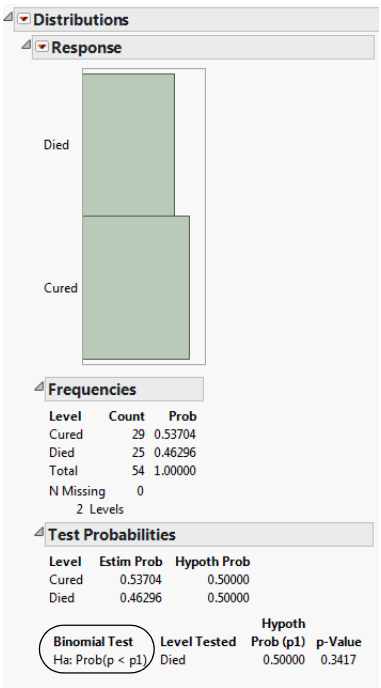
To generate a test probabilities report for a variable with exactly two levels:

1. Refer to Figure 3.13 at right. Type 0.5 in both Hypoth Prob fields.
2. Click the **probability less than hypothesized value** button.
3. Click **Done**.

Exact probabilities are calculated for the binomial test. See Figure 3.14 at right.

Figure 3.14 Examples of Test Probabilities Reports



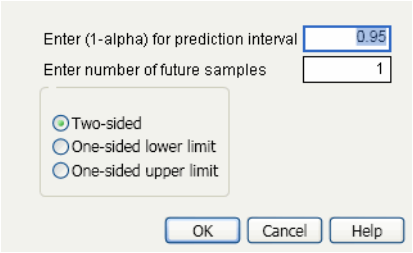


Example of Prediction Intervals

Suppose you are interested in computing prediction intervals for the next 10 observations of ozone level.

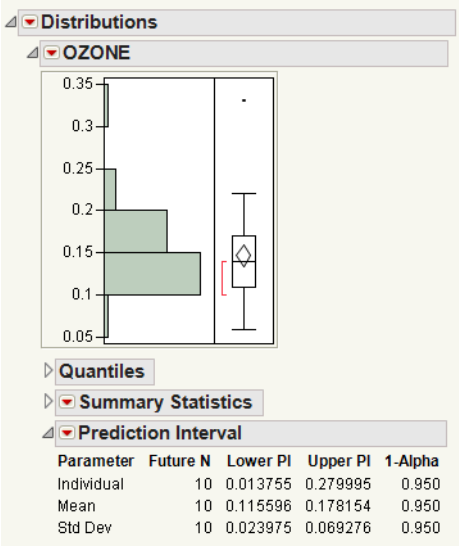
1. Select **Help > Sample Data Library** and open Cities.jmp.
2. Select **Analyze > Distribution**.
3. Select OZONE and click **Y, Columns**.
4. Click **OK**.
5. From the red triangle next to OZONE, select **Prediction Interval**.

Figure 3.15 The Prediction Intervals Window



6. In the Prediction Intervals window, type 10 next to **Enter number of future samples**.
7. Click **OK**.

Figure 3.16 Example of a Prediction Interval Report



Parameter	Future N	Lower PI	Upper PI	1-Alpha
Individual	10	0.013755	0.279995	0.950
Mean	10	0.115596	0.178154	0.950
Std Dev	10	0.023975	0.069276	0.950

In this example, you can be 95% confident about the following:

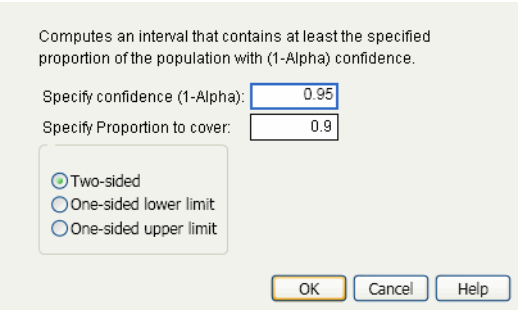
- Each of the next 10 observations will be between 0.013755 and 0.279995.
- The mean of the next 10 observations will be between 0.115596 and 0.178154.
- The standard deviation of the next 10 observations will be between 0.023975 and 0.069276.

Example of Tolerance Intervals

Suppose you want to estimate an interval that contains 90% of ozone level measurements.

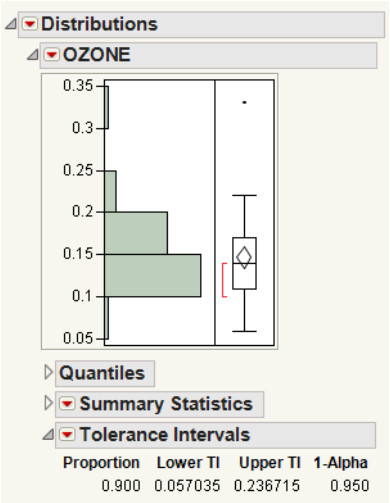
1. Select **Help > Sample Data Library** and open Cities.jmp.
2. Select **Analyze > Distribution**.
3. Select OZONE and click **Y, Columns**.
4. Click **OK**.
5. From the red triangle menu next to OZONE, select **Tolerance Interval**.

Figure 3.17 The Tolerance Intervals Window



6. Keep the default selections, and click **OK**.

Figure 3.18 Example of a Tolerance Interval Report



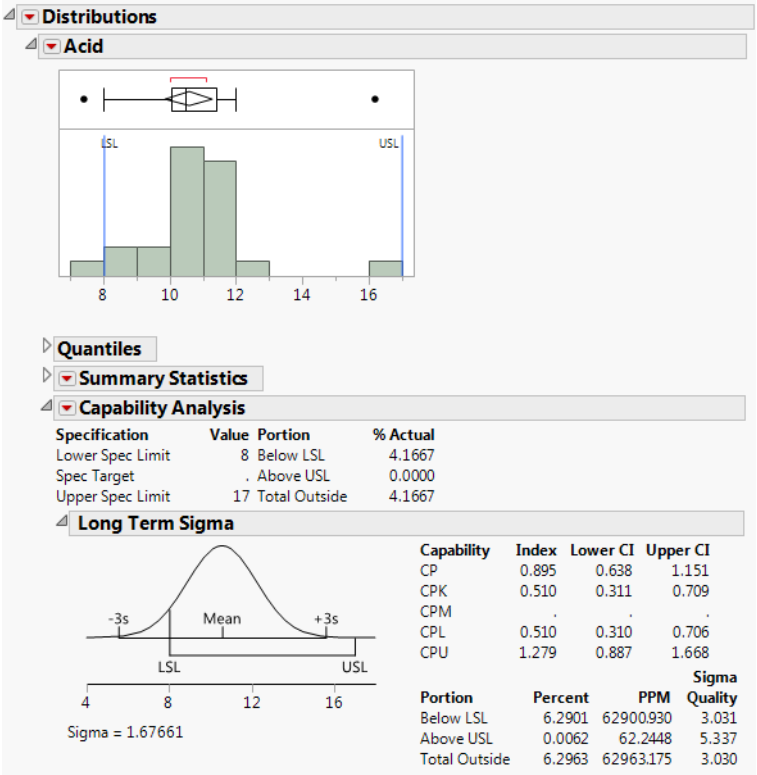
In this example, you can be 95% confident that at least 90% of the population lie between 0.057035 and 0.236715, based on the Lower TI (tolerance interval) and Upper TI values.

Example of Capability Analysis

Suppose you want to characterize the abrasion levels of the tires your company manufactures. The lower and upper specification limits are 100 and 200, respectively.

1. Select **Help > Sample Data Library** and open Quality Control/Pickles.jmp.
2. Select **Analyze > Distribution**.
3. Select Acid and click **Y, Columns**.
4. Click **OK**.
5. From the red triangle menu next to Acid, select **Capability Analysis**.
6. Type 8 for the **Lower Spec Limit**.
7. Type 17 for the **Upper Spec Limit**.
8. Keep the rest of the default selections, and click **OK**.
9. From the red triangle menu next to Acid, select **Histogram Options > Vertical**.

Figure 3.19 Example of the Capability Analysis Report



The spec limits are added to the histogram so that the data can be visually compared to the limits. As you can see, some of the acidity levels are below the lower spec limit, and some are very close to the upper spec limit. The Capability Analysis results are added to the report. The CPK value is 0.510, indicating a process that is not capable, relative to the given specification limits.

Statistical Details for the Distribution Platform

This section contains statistical details for Distribution options and reports.

Standard Error Bars

Standard errors bars are calculated using the standard error $\sqrt{np_i(1-p_i)}$ where $p_i = n_i/n$.

Quantiles

This section describes how quantiles are computed.

To compute the pth quantile of n non-missing values in a column, arrange the n values in ascending order and call these column values $y_1, y_2, ..., y_n$. Compute the rank number for the pth quantile as $p / 100(n + 1)$.

- If the result is an integer, the pth quantile is that rank's corresponding value.
- If the result is not an integer, the pth quantile is found by interpolation. The pth quantile, denoted q_p, is computed as follows:

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

  where:
  - n is the number of non-missing values for a variable
  - $y_1, y_2, ..., y_n$ represents the ordered values of the variable
  - y_{n+1} is taken to be y_n
  - i is the integer part and f is the fractional part of $(n+1)p$.
  - $(n + 1)p = i + f$

For example, suppose a data table has 15 rows and you want to find the 75th and 90th quantile values of a continuous column. After the column is arranged in ascending order, the ranks that contain these quantiles are computed as follows:

$$\frac{75}{100}(15 + 1) = 12 \text{ and } \frac{90}{100}(15 + 1) = 14.4$$

The value y_{12} is the 75th quantile. The 90th quantile is interpolated by computing a weighted average of the 14th and 15th ranked values as $y_{90} = 0.6y_{14} + 0.4y_{15}$.

Summary Statistics

This section contains statistical details for specific statistics in the Summary Statistics report.

Mean

The mean is the sum of the non-missing values divided by the number of non-missing values. If you assigned a **Weight** or **Freq** variable, the mean is computed by JMP as follows:

1. Each column value is multiplied by its corresponding weight or frequency.
2. These values are added and divided by the sum of the weights or frequencies.

Std Dev

The standard deviation measures the spread of a distribution around the mean. It is often denoted as s and is the square root of the sample variance, denoted s^2.

$$s = \sqrt{s^2} \text{ where}$$

$$s^2 = \sum_{i=1}^{N} \frac{w_i(y_i - \bar{y}_w)^2}{N-1}$$

$$\bar{y}_w = \text{weighted mean}$$

Std Err Mean

The standard error means is computed by dividing the sample standard deviation, s, by the square root of N. In the launch window, if you specified a column for Weight or Freq, then the denominator is the square root of the sum of the weights or frequencies.

Skewness

Skewness is based on the third moment about the mean and is computed as follows:

$$\sum w_i^{\frac{3}{2}} z_i^3 \frac{N}{(N-1)(N-2)} \text{ where } z_i = \frac{x_i - \bar{x}}{\mathrm{s}}$$

and w_i is a weight term (= 1 for equally weighted items).

Kurtosis

Kurtosis is based on the fourth moment about the mean and is computed as follows:

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} w_i^2 \left(\frac{x_i - \bar{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

where w_i is a weight term (= 1 for equally weighted items). Using this formula, the Normal distribution has a kurtosis of 0.

Normal Quantile Plot

The empirical cumulative probability for each value is computed as follows:

$$\frac{r_i}{N+1}$$

where r_i is the rank of the ith observation, and N is the number of non-missing (and nonexcluded) observations.

The normal quantile values are computed as follows:

$$\Phi^{-1}\left(\frac{r_i}{N+1}\right)$$

where Φ is the cumulative probability distribution function for the normal distribution.

These normal quantile values are Van Der Waerden approximations to the order statistics that are expected for the normal distribution.

Wilcoxon Signed Rank Test

The Wilcoxon signed-rank test can be used to test for the median of a single population or to test matched-pairs data for a common median. In the case of matched pairs, the test reduces to testing the single population of paired differences for a median of 0. The test assumes that the underlying population is symmetric.

The Wilcoxon test allows tied values. The test statistic is adjusted for differences of zero using a method suggested by Pratt. See Lehman (2006), Pratt (1959), and Cureton (1967).

Testing for the Median of a Single Population

- There are N observations:

  $X_1, X_2, ..., X_N$
- The null hypothesis is:

  H_0: median = m
- The differences between observations and the hypothesized value m are calculated as follows:

  $D_j = X_j - m$

Testing for the Equality of Two Population Medians with Matched Pairs Data

A special case of the Wilcoxon signed-rank test is applied to matched-pairs data.

- There are N pairs of observations from two populations:

  $X_1, X_2, ..., X_N$ and $Y_1, Y_2, ..., Y_N$
- The null hypothesis is:

  H_0: $\text{median}_{X-Y} = 0$
- The differences between pairs of observations are calculated as follows:

  $D_j = X_j - Y_j$

Wilcoxon Signed-Rank Test Statistic

The test statistic is based on the sum of the signed ranks. Signed ranks are defined as follows:

- The absolute values of the differences, $|D_j|$, are ranked from smallest to largest.
- The ranks start with the value 1, even if there are differences of zero.
- When there are tied absolute differences, they are assigned the average, or *midrank*, of the ranks of the observations.

Denote the rank or midrank for a difference D_j by R_j. Define the signed rank for D_j as follows:

- If the difference D_j is positive, the signed rank is R_j.
- If the difference D_j is zero, the signed rank is 0.
- If the difference D_j is negative, the signed rank is $-R_j$.

The signed-rank statistic is computed as follows:

$$W = \frac{1}{2}\sum_{j=1}^{N} \text{signed ranks}$$

Define the following:

d_0 is the number of signed ranks that equal zero

R^+ is the sum of the positive signed ranks

Then the following holds:

$$W = R^+ - \frac{1}{4}[N(N+1) - d_0(d_0+1)]$$

Wilcoxon Signed-Rank Test P-Values

For $N \le 20$, exact *p*-values are calculated.

For $N > 20$, a Student's *t* approximation to the statistic defined below is used. Note that a correction for ties is applied. See Iman (1974) and Lehmann (1998).

Under the null hypothesis, the mean of *W* is zero. The variance of *W* is given by the following:

$$Var(W) = \frac{1}{24}\left[N(N+1)(2N+1) - d_0(d_0+1)(2d_0+1) - \frac{1}{2}\sum_{i>0} d_i(d_i+1)(d_i-1)\right]$$

The last summation in the expression for *Var(W)* is a correction for ties. The notation d_i for i > 0 represents the number of values in the ith group of non-zero signed ranks. (If there are no ties for a given signed rank, then $d_i = 1$ and the summand is 0.)

The statistic t given by the following has an approximate t distribution with N - 1 degrees of freedom:

$$t = W / \sqrt{Var(W)}$$

Standard Deviation Test

Here is the formula for calculating the Test Statistic:

$$\frac{(n-1)s^2}{\sigma^2}$$

The Test Statistic is distributed as a Chi-square variable with n - 1 degrees of freedom when the population is normal.

The Min PValue is the p-value of the two-tailed test, and is calculated as follows:

2*min(p1,p2)

where p1 is the lower one-tail p-value and p2 is the upper one-tail p-value.

Normal Quantiles

The normal quantile values are computed as follows:

$$\Phi^{-1}\left(\frac{r_i}{N+1}\right) \text{ where:}$$

- Φ is the cumulative probability distribution function for the normal distribution
- r_i is the rank of the ith observation
- N is the number of non-missing observations

Saving Standardized Data

The standardized values are computed using the following formula:

$$\frac{X - \overline{X}}{S_X} \text{ where:}$$

- X is the original column
- $\overline{X}$ is the mean of column X
- S_X is the standard deviation of column X

Prediction Intervals

The formulas that JMP uses for computing prediction intervals are as follows:

- For m future observations:

$$[\underset{\sim}{y}_m, \tilde{y}_m] = \overline{X} \pm t_{(1-\alpha/2m;n-1)} \times \sqrt{1+\frac{1}{n}} \times s \quad \text{for} \quad m \geq 1$$

- For the mean of m future observations:

$$[Y_l, Y_u] = \overline{X} \pm t_{(1-\alpha/2,\, n-1)} \times \sqrt{\frac{1}{m}+\frac{1}{n}} \times s \quad \text{for} \quad m \geq 1 \cdot$$

- For the standard deviation of m future observations:

$$[s_l, s_u] = \left[s \times \sqrt{\frac{1}{F_{(1-\alpha/2;(n-1,\, m-1))}}},\, s \times \sqrt{F_{(1-\alpha/2;(m-1,\, n-1))}} \right] \quad \text{for} \quad m \geq 2$$

where m = number of future observations, and n = number of points in current analysis sample.

- The one-sided intervals are formed by using 1-α in the quantile functions.

For references, see Hahn and Meeker (1991), pages 61-64.

Tolerance Intervals

This section contains statistical details for one-sided and two-sided tolerance intervals.

One-Sided Interval

The one-sided interval is computed as follows:

Upper Limit $= \bar{x} + g's$

Lower Limit $= \bar{x} - g's$

where

$g' = t(1-\alpha, n-1, \Phi^{-1}(p) \cdot \sqrt{n}) / \sqrt{n}$ from Table 1 of Odeh and Owen (1980).

t is the quantile from the non-central t-distribution, and Φ^{-1} is the standard normal quantile.

Two-Sided Interval

The two-sided interval is computed as follows:

$$[\tilde{T}_{p_L}, \tilde{T}_{p_U}] = [\bar{x} - g_{(1-\alpha/2;p_L,n)}s\ ,\ \bar{x} + g_{(1-\alpha/2;p_U,n)}s]$$

where

s = standard deviation and $g_{(1-\alpha/2;p,n)}$ is a constant that can be found in Table 4 of Odeh and Owen 1980).

To determine g, consider the fraction of the population captured by the tolerance interval. Tamhane and Dunlop (2000) give this fraction as follows:

$$\Phi\left(\frac{\bar{x} + gs - \mu}{\sigma}\right) - \Phi\left(\frac{\bar{x} - gs - \mu}{\sigma}\right)$$

where Φ denotes the standard normal c.d.f. (cumulative distribution function). Therefore, g solves the following equation:

$$P\left\{\Phi\left(\frac{\bar{X} + gs - \mu}{\sigma}\right) - \Phi\left(\frac{\bar{X} - gs - \mu}{\sigma}\right) \geq 1 - \gamma\right\} = 1 - \alpha$$

where 1-γ is the fraction of all future observations contained in the tolerance interval.

More information is given in Tables A.1a, A.1b, A.11a, and A.11b of Hahn and Meeker (1991).

Capability Analysis

All capability analyses use the same formulas. Options differ in how sigma (σ) is computed:

- **Long-term** uses the overall sigma. This option is used for P_{pk} statistics, and computes sigma as follows:

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

Note: There is a preference for Distribution called **Ppk Capability Labeling** that labels the long-term capability output with P_{pk} labels. This option is found using **File > Preferences**, then select **Platforms > Distribution**.

- **Specified Sigma** enables you to type a specific, known sigma used for computing capability analyses. Sigma is user-specified, and is therefore not computed.
- **Moving Range** enables you to enter a range span, which computes sigma as follows:

$$\sigma = \frac{\bar{R}}{d_2(n)} \text{ where}$$

$\bar{R}$ is the average of the moving ranges

$d_2(n)$ is the expected value of the range of n independent normally distributed variables with unit standard deviation.

- **Short Term Sigma, Group by Fixed Subgroup Size** if r is the number of subgroups of size nj and each ith subgroup is defined by the order of the data, sigma is computed as follows:

$$\sigma = \sqrt{\frac{1}{r}\sum_{i=1}^{r}\sum_{j=1}^{n_j}\frac{(X_{ij}-\bar{X}_i)^2}{n_j-1}} \text{ where } \bar{X}_i = \frac{1}{n_j}\sum_{j=1}^{nj} X_{ij}$$

- This formula is commonly referred to as the Root Mean Square Error, or RMSE.

Capability Index Definitions

Writing T for the target, LSL, and USL for the lower and upper specification limits, and P_α for the α*100th percentile, the generalized capability indices are as follows:

$$C_{pl} = \frac{P_{0.5} - LSL}{P_{0.5} - P_{0.00135}}$$

$$C_{pu} = \frac{USL - P_{0.5}}{P_{0.99865} - P_{0.5}}$$

$$C_p = \frac{USL - LSL}{P_{0.99865} - P_{0.00135}}$$

$$C_{pk} = \min\left(\frac{P_{0.5} - LSL}{P_{0.5} - P_{0.00135}}, \frac{USL - P_{0.5}}{P_{0.99865} - P_{0.5}}\right)$$

$$K = 2 \times \frac{\left|\frac{1}{2}(USL + LSL) - P_{0.5}\right|}{USL - LSL}$$

$$C_{pm} = \frac{\min\left(\frac{T - LSL}{P_{0.5} - P_{0.00135}}, \frac{USL - T}{P_{0.99865} - P_{0.5}}\right)}{\sqrt{1 + \left(\frac{\mu - T}{\sigma}\right)^2}}$$

If the data are normally distributed, these formulas reduce to the formulas for standard capability indices. See Table 3.2.

Note: The confidence intervals in the following table are computed using an alpha level of 0.05.

Table 3.2 Descriptions of Capability Indices and Computational Formulas

Index	Index Name	Formula
CP	process capability ratio, C_p	(USL - LSL)/6*s* where: • USL is the upper spec limit • LSL is the lower spec limit
CIs for CP	Lower CI on CP	$\mathrm{CP}\sqrt{\frac{\chi^2_{\alpha/2,\,n-1}}{n-1}}$
	Upper CI on CP	$\mathrm{CP}\sqrt{\frac{\chi^2_{1-\alpha/2,\,n-1}}{n-1}}$
CPK (PPK for AIAG)	process capability index, C_{pk}	min(CPL, CPU)
CIs for CPK See Bissell (1990)	Lower CI	$\hat{C}_{pk}\left[1-\Phi^{-1}(1-\alpha/2)\sqrt{\frac{1}{9n\hat{C}^2_{pk}}+\frac{1}{2(n-1)}}\right]$
	Upper CI	$\hat{C}_{pk}\left[1+\Phi^{-1}(1-\alpha/2)\sqrt{\frac{1}{9n\hat{C}^2_{pk}}+\frac{1}{2(n-1)}}\right]$
CPM	process capability index, C_{pm}	$\frac{\min(\mathrm{Target}-\mathrm{LSL},\ \mathrm{USL}-\mathrm{Target})}{3\sqrt{s^2+(\bar{x}-\mathrm{Target})^2}}$ **Note:** CPM confidence intervals are not reported when the target is not within the Lower and Upper Spec Limits range. CPM intervals are only reported when the target is within this range. JMP writes a message to the log to note why the CPM confidence intervals are missing.

Table 3.2 Descriptions of Capability Indices and Computational Formulas *(Continued)*

Index	Index Name	Formula
CIs for CPM	Lower CI on CPM	$\text{CPM}\sqrt{\frac{\chi^2_{\alpha/2,\gamma}}{\gamma}}$, where $\gamma = \frac{n\left(1+\left(\frac{\bar{x}-\text{Target}}{s}\right)^2\right)^2}{1+2\left(\frac{\bar{x}-\text{Target}}{s}\right)^2}$
	Upper CI on CPM	$\text{CPM}\sqrt{\frac{\chi^2_{1-\alpha/2,\gamma}}{\gamma}}$ where γ = same as above.
CPL	process capability ratio of one-sided lower spec	(mean - LSL)/3*s*
CPU	process capability ratio of one-sided upper spec	(USL - mean)/3*s*

- A capability index of 1.33 is considered to be the minimum acceptable. For a normal distribution, this gives an expected number of nonconforming units of about 6 per 100,000.
- Exact 100(1 - α)% lower and upper confidence limits for CPL are computed using a generalization of the method of Chou et al. (1990), who point out that the 100(1 - α) lower confidence limit for CPL (denoted by CPLLCL) satisfies the following equation:

  $$\Pr\{T_{n-1}(\delta = 3\sqrt{n})\text{CPLLCL} \le 3\text{CPL}\sqrt{n}\} = 1-\alpha$$

  where $T_{n\text{-}1}(\delta)$ has a non-central *t*-distribution with *n* - 1 degrees of freedom and noncentrality parameter δ.
- Exact 100(1 - α)% lower and upper confidence limits for CPU are also computed using a generalization of the method of Chou et al. (1990), who point out that the 100(1 - α) lower confidence limit for CPU (denoted CPULCL) satisfies the following equation:

  $$\Pr\{T_{n-1}(\delta = 3\sqrt{n})\text{CPULCL} \ge 3\text{CPU}\sqrt{n}\} = 1-\alpha$$

  where $T_{n\text{-}1}(\delta)$ has a non-central *t*-distribution with *n* - 1 degrees of freedom and noncentrality parameter δ.

Note: Because of a lack of supporting research at the time of this writing, computing confidence intervals for capability indices is not recommended, except for cases when the capability indices are based on the standard deviation.

- Sigma Quality is defined as the following

$$\text{Sigma Quality} = \text{Normal Quantile}\left(1 - \frac{\%\ \text{outside}}{100}\right) + 1.5$$

$$\text{Sigma Quality Above} = \text{Normal Quantile}\left(1 - \frac{\%\ \text{above}}{100}\right) + 1.5$$

$$\text{Sigma Quality Below} = \text{Normal Quantile}\left(1 - \frac{\%\ \text{below}}{100}\right) + 1.5$$

  For example, if there are 3 defects in n=1,000,000 observations, the formula yields 6.03, or a 6.03 sigma process. The results of the computations of the Sigma Quality Above USL and Sigma Quality Below LSL column values do not sum to the Sigma Quality Total Outside column value because calculating Sigma Quality involves finding normal distribution quantiles, and is therefore not additive.

- Here are the Benchmark Z formulas:

  Z USL = (USL-Xbar)/sigma = 3 * CPU

  Z LSL = (Xbar-LSL)/sigma = 3 * CPL

  Z Bench = Inverse Cumulative Prob(1 - P(LSL) - P(USL))

  where:

  P(LSL) = Prob(X < LSL) = 1 - Cum Prob(Z LSL)

  P(USL) = Prob(X > USL) = 1 - Cum Prob(Z USL).

Continuous Fit Distributions

This section contains statistical details for the options in the Continuous Fit menu.

Normal

The **Normal** fitting option estimates the parameters of the normal distribution. The normal distribution is often used to model measures that are symmetric with most of the values falling in the middle of the curve. Select the **Normal** fitting for any set of data and test how well a normal distribution fits your data.

The parameters for the normal distribution are as follows:

- μ (the mean) defines the location of the distribution on the x-axis
- σ (standard deviation) defines the dispersion or spread of the distribution

The standard normal distribution occurs when $\mu = 0$ and $\sigma = 1$. The Parameter Estimates table shows estimates of μ and σ, with upper and lower 95% confidence limits.

$$\text{pdf:} \quad \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad \text{for} \quad -\infty < x < \infty; \quad -\infty < \mu < \infty; \quad 0 < \sigma$$

$$E(x) = \mu$$

$$\mathrm{Var}(x) = \sigma^2$$

LogNormal

The **LogNormal** fitting option estimates the parameters μ (scale) and σ (shape) for the two-parameter lognormal distribution. A variable Y is lognormal if and only if $X = \ln(Y)$ is normal. The data must be greater than zero.

$$\text{pdf:} \quad \frac{1}{\sigma\sqrt{2\pi}}\frac{\exp\left[\frac{-(\log(x)-\mu)^2}{2\sigma^2}\right]}{x} \quad \text{for} \quad 0 \le x; \quad -\infty < \mu < \infty; \quad 0 < \sigma$$

$$E(x) = \exp(\mu + \sigma^2/2)$$

$$\mathrm{Var}(x) = \exp(2(\mu + \sigma^2)) - \exp(2\mu + \sigma^2)$$

Weibull, Weibull with Threshold, and Extreme Value

The Weibull distribution has different shapes depending on the values of α (scale) and β (shape). It often provides a good model for estimating the length of life, especially for mechanical devices and in biology. The **Weibull** option is the same as the Weibull with threshold option, with a threshold (θ) parameter of zero. For the Weibull with threshold option, JMP estimates the threshold as the minimum value. If you know what the threshold should be, set it by using the **Fix Parameters** option. See "Fit Distribution Options" on page 63.

The pdf for the Weibull with threshold is as follows:

$$\text{pdf:} \quad \frac{\beta}{\alpha^\beta}(x-\theta)^{\beta-1}\exp\left[-\left(\frac{x-\theta}{\alpha}\right)^\beta\right] \quad \text{for} \quad \alpha,\beta > 0; \quad \theta < x$$

$$E(x) = \theta + \alpha\Gamma\left(1 + \frac{1}{\beta}\right)$$

$$\mathrm{Var}(x) = \alpha^2\left\{\Gamma\left(1 + \frac{2}{\beta}\right) - \Gamma^2\left(1 + \frac{1}{\beta}\right)\right\}$$

where $\Gamma(\cdot)$ is the Gamma function.

The Extreme Value distribution is a two parameter Weibull (α, β) distribution with the transformed parameters $\delta = 1 / \beta$ and $\lambda = \ln(\alpha)$.

Exponential

The exponential distribution is especially useful for describing events that randomly occur over time, such as survival data. The exponential distribution might also be useful for modeling elapsed time between the occurrence of non-overlapping events, such as the time between a user's computer query and response of the server, the arrival of customers at a service desk, or calls coming in at a switchboard.

The Exponential distribution is a special case of the two-parameter Weibull when $\beta = 1$ and $\alpha = \sigma$, and also a special case of the Gamma distribution when $\alpha = 1$.

pdf: $\frac{1}{\sigma}\exp(-x/\sigma)$ for $0 < \sigma;\ 0 \le x$

$E(x) = \sigma$

$Var(x) = \sigma^2$

Devore (1995) notes that an exponential distribution is *memoryless*. Memoryless means that if you check a component after *t* hours and it is still working, the distribution of additional lifetime (the conditional probability of additional life given that the component has lived until *t*) is the same as the original distribution.

Gamma

The **Gamma** fitting option estimates the gamma distribution parameters, $\alpha > 0$ and $\sigma > 0$. The parameter α, called alpha in the fitted gamma report, describes shape or curvature. The parameter σ, called sigma, is the scale parameter of the distribution. A third parameter, θ, called the Threshold, is the lower endpoint parameter. It is set to zero by default, unless there are negative values. You can also set its value by using the **Fix Parameters** option. See "Fit Distribution Options" on page 63.

pdf: $\frac{1}{\Gamma(\alpha)\sigma^{\alpha}}(x-\theta)^{\alpha-1}\exp(-(x-\theta)/\sigma)$ for $\theta \le x;\ 0 < \alpha,\sigma$

$E(x) = \alpha\sigma + \theta$

$Var(x) = \alpha\sigma^2$

- The *standard* gamma distribution has $\sigma = 1$. Sigma is called the scale parameter because values other than 1 stretch or compress the distribution along the *x*-axis.
- The Chi-square $\chi^2_{(\nu)}$ distribution occurs when $\sigma = 2$, $\alpha = \nu/2$, and $\theta = 0$.

- The exponential distribution is the family of gamma curves that occur when $\alpha = 1$ and $\theta = 0$.

The standard gamma density function is strictly decreasing when $\alpha \leq 1$. When $\alpha > 1$, the density function begins at zero, increases to a maximum, and then decreases.

Beta

The standard beta distribution is useful for modeling the behavior of random variables that are constrained to fall in the interval 0,1. For example, proportions always fall between 0 and 1. The **Beta** fitting option estimates two shape parameters, $\alpha > 0$ and $\beta > 0$. There are also θ and σ, which are used to define the lower threshold as θ, and the upper threshold as $\theta + \sigma$. The beta distribution has values only for the interval defined by $\theta \leq x \leq (\theta + \sigma)$. The θ is estimated as the minimum value, and σ is estimated as the range. The standard beta distribution occurs when $\theta = 0$ and $\sigma = 1$.

Set parameters to fixed values by using the **Fix Parameters** option. The upper threshold must be greater than or equal to the maximum data value, and the lower threshold must be less than or equal to the minimum data value. For details about the Fix Parameters option, see "Fit Distribution Options" on page 63.

pdf: $$\frac{1}{B(\alpha, \beta)\sigma^{\alpha+\beta-1}}(x-\theta)^{\alpha-1}(\theta+\sigma-x)^{\beta-1} \quad \text{for} \quad \theta \leq x \leq \theta+\sigma; \quad 0 < \sigma, \alpha, \beta$$

$$E(x) = \theta + \sigma\frac{\alpha}{\alpha+\beta}$$

$$\mathrm{Var}(x) = \frac{\sigma^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

where $B(\cdot)$ is the Beta function.

Normal Mixtures

The **Normal Mixtures** option fits a mixture of normal distributions. This flexible distribution is capable of fitting multi-modal data.

Fit a mixture of two or three normal distributions by selecting the **Normal 2 Mixture** or **Normal 3 Mixture** options. Alternatively, you can fit a mixture of *k* normal distributions by selecting the **Other** option. A separate mean, standard deviation, and proportion of the whole is estimated for each group.

pdf: $$\sum_{i=1}^{k} \frac{\pi_i}{\sigma_i}\phi\left(\frac{x-\mu_i}{\sigma_i}\right)$$

$$E(x) = \sum_{i=1}^{k} \pi_i \mu_i$$

$$Var(x) = \sum_{i=1}^{k} \pi_i(\mu_i^2 + \sigma_i^2) - \left(\sum_{i=1}^{k} \pi_i \mu_i \right)^2$$

where μ_i, σ_i, and π_i are the respective mean, standard deviation, and proportion for the ith group, and $\phi(\cdot)$ is the standard normal pdf.

Smooth Curve

The **Smooth Curve** option fits a smooth curve using nonparametric density estimation (kernel density estimation). The smooth curve is overlaid on the histogram and a slider appears beneath the plot. Control the amount of smoothing by changing the kernel standard deviation with the slider. The initial Kernel Std estimate is calculated from the standard deviation of the data.

Johnson Su, Johnson Sb, Johnson Sl

The Johnson system of distributions contains three distributions that are all based on a transformed normal distribution. These three distributions are the following:

- Johnson Su, which is unbounded.
- Johnson Sb, which has bounds on both tails defined by parameters that can be estimated.
- Johnson Sl, which is bounded in one tail by a parameter that can be estimated. The Johnson Sl family contains the family of lognormal distributions.

The S refers to system, the subscript of the range. Although we implement a different method, information about selection criteria for a particular Johnson system can be found in Slifker and Shapiro (1980).

Johnson distributions are popular because of their flexibility. In particular, the Johnson distribution system is noted for its data-fitting capabilities because it supports every possible combination of skewness and kurtosis.

If Z is a standard normal variate, then the system is defined as follows:

$$Z = \gamma + \delta f(Y)$$

where, for the Johnson Su:

$$f(Y) = \ln\left(Y + \sqrt{1 + Y^2}\right) = \sinh^{-1} Y$$

$$Y = \frac{X-\theta}{\sigma} \qquad -\infty < X < \infty$$

where, for the Johnson Sb:

$$f(Y) = \ln\left(\frac{Y}{1-Y}\right)$$

$$Y = \frac{X-\theta}{\sigma} \qquad \theta < X < \theta + \sigma$$

and for the Johnson Sl, where $\sigma = \pm 1$.

$$f(Y) = \ln(Y)$$

$$Y = \frac{X-\theta}{\sigma} \qquad \begin{array}{ll} \theta < X < \infty & \text{if } \sigma = 1 \\ -\infty < X < \theta & \text{if } \sigma = -1 \end{array}$$

Johnson Su

$$\text{pdf:} \quad \frac{\delta}{\sigma}\left[1+\left(\frac{x-\theta}{\sigma}\right)^2\right]^{-1/2}\phi\left[\gamma+\delta\sinh^{-1}\left(\frac{x-\theta}{\sigma}\right)\right] \quad \text{for} \quad -\infty < x, \theta, \gamma < \infty; \quad 0 < \theta, \delta$$

Johnson Sb

$$\text{pdf:} \quad \phi\left[\gamma+\delta\ln\left(\frac{x-\theta}{\sigma-(x-\theta)}\right)\right]\left(\frac{\delta\sigma}{(x-\theta)(\sigma-(x-\theta))}\right) \quad \text{for} \quad \theta < x < \theta+\sigma; \quad 0 < \sigma$$

Johnson Sl

$$\text{pdf:} \quad \frac{\delta}{|x-\theta|}\phi\left[\gamma+\delta\ln\left(\frac{x-\theta}{\sigma}\right)\right] \quad \text{for} \quad \theta < x \text{ if } \sigma = 1; \quad \theta > x \text{ if } \sigma = -1$$

where $\phi(\cdot)$ is the standard normal pdf.

Note: The parameter confidence intervals are hidden in the default report. Parameter confidence intervals are not very meaningful for Johnson distributions, because they are transformations to normality. To show parameter confidence intervals, right-click in the report and select **Columns > Lower 95%** and **Upper 95%**.

Generalized Log (Glog)

This distribution is useful for fitting data that are rarely normally distributed and often have non-constant variance, like biological assay data. The Glog distribution is described with the parameters μ (location), σ (scale), and λ (shape).

$$\text{pdf:} \quad \phi\left\{\frac{1}{\sigma}\left[\log\left(\frac{x+\sqrt{x^2+\lambda^2}}{2}\right)-\mu\right]\right\}\frac{x+\sqrt{x^2+\lambda^2}}{\sigma(x^2+\lambda^2+x\sqrt{x^2+\lambda^2})}$$

for $0 \le \lambda$; $0 < \sigma$; $-\infty < \mu < \infty$

The Glog distribution is a transformation to normality, and comes from the following relationship:

$$\text{If } z = \frac{1}{\sigma}\left[\log\left(\frac{x+\sqrt{x^2+\lambda^2}}{2}\right)-\mu\right] \sim N(0,1), \text{ then } x \sim \text{Glog}(\mu,\sigma,\lambda).$$

When $\lambda = 0$, the Glog reduces to the LogNormal (μ,σ).

Note: The parameter confidence intervals are hidden in the default report. Parameter confidence intervals are not very meaningful for the GLog distribution, because it is a transformation to normality. To show parameter confidence intervals, right-click in the report and select **Columns > Lower 95%** and **Upper 95%**.

All

In the Compare Distributions report, the ShowDistribution list is sorted by AICc in ascending order.

The formula for AICc is as follows:

$$\text{AICc} = -2\log L + 2\nu + \frac{2\nu(\nu+1)}{n-(\nu+1)}$$

where:

- logL is the log-likelihood
- n is the sample size
- ν is the number of parameters

If the column contains negative values, the Distribution list does not include those distributions that require data with positive values. Only continuous distributions are listed. Distributions with threshold parameters, such as Beta and Johnson Sb, are not included in the list of possible distributions.

Discrete Fit Distributions

This section contains statistical details for the options in the Discrete Fit menu.

Poisson

The Poisson distribution has a single scale parameter $\lambda > 0$.

$$\text{pmf:} \quad \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for} \quad 0 \le \lambda < \infty; \quad x = 0,1,2,...$$

$E(x) = \lambda$

$Var(x) = \lambda$

Since the Poisson distribution is a discrete distribution, the overlaid curve is a step function, with jumps occurring at every integer.

Gamma Poisson

This distribution is useful when the data is a combination of several Poisson(μ) distributions, each with a different μ. One example is the overall number of accidents combined from multiple intersections, when the mean number of accidents (μ) varies between the intersections.

The Gamma Poisson distribution results from assuming that $x \mid \mu$ follows a Poisson distribution and μ follows a Gamma(α,τ). The Gamma Poisson has parameters $\lambda = \alpha\tau$ and $\sigma = \tau+1$. The parameter σ is a dispersion parameter. If $\sigma > 1$, there is over dispersion, meaning there is more variation in x than explained by the Poisson alone. If $\sigma = 1$, x reduces to Poisson(λ).

$$\text{pmf:} \quad \frac{\Gamma\left(x + \frac{\lambda}{\sigma - 1}\right)}{\Gamma(x+1)\Gamma\left(\frac{\lambda}{\sigma - 1}\right)}\left(\frac{\sigma - 1}{\sigma}\right)^x \sigma^{-\frac{\lambda}{\sigma - 1}} \quad \text{for} \quad 0 < \lambda; \quad 1 \le \sigma; \quad x = 0,1,2,...$$

$E(x) = \lambda$

$Var(x) = \lambda\sigma$

where $\Gamma(\cdot)$ is the Gamma function.

Remember that $x \mid \mu \sim$ Poisson(μ), while $\mu \sim$ Gamma(α,τ). The platform estimates $\lambda = \alpha\tau$ and $\sigma = \tau+1$. To obtain estimates for α and τ, use the following formulas:

$$\hat{\tau} = \hat{\sigma} - 1$$

$$\hat{\alpha} = \frac{\hat{\lambda}}{\hat{\tau}}$$

If the estimate of σ is 1, the formulas do not work. In that case, the Gamma Poisson has reduced to the Poisson(λ), and $\hat{\lambda}$ is the estimate of λ.

If the estimate for α is an integer, the Gamma Poisson is equivalent to a Negative Binomial with the following pmf:

$$p(y) = \binom{y+r-1}{y} p^r (1-p)^y \quad \text{for} \quad 0 \le y$$

with r = α and (1-*p*)/*p* = τ.

Run demoGammaPoisson.jsl in the JMP Samples/Scripts folder to compare a Gamma Poisson distribution with parameters λ and σ to a Poisson distribution with parameter λ.

Binomial

The **Binomial** option accepts data in two formats: a constant sample size, or a column containing sample sizes.

$$\text{pmf:} \quad \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for} \quad 0 \le p \le 1\,; \quad x = 0,1,2,\ldots,n$$

E(*x*) = *np*

Var(*x*) = *np*(1-*p*)

where n is the number of independent trials.

Note: The confidence interval for the binomial parameter is a Score interval. See Agresti (1998).

Beta Binomial

This distribution is useful when the data is a combination of several Binomial(p) distributions, each with a different p. One example is the overall number of defects combined from multiple manufacturing lines, when the mean number of defects (p) varies between the lines.

The Beta Binomial distribution results from assuming that x|π follows a Binomial(n,π) distribution and π follows a Beta(α,β). The Beta Binomial has parameters p = α/(α+β) and δ = 1/(α+β+1). The parameter δ is a dispersion parameter. When δ > 0, there is over dispersion, meaning there is more variation in *x* than explained by the Binomial alone. When δ < 0, there is under dispersion. When δ = 0, *x* is distributed as Binomial(n,p). The Beta Binomial only exists when $n \ge 2$.

$$\text{pmf:} \quad \binom{n}{x} \frac{\Gamma\left(\frac{1}{\delta}-1\right)\Gamma\left[x+p\left(\frac{1}{\delta}-1\right)\right]\Gamma\left[n-x+(1-p)\left(\frac{1}{\delta}-1\right)\right]}{\Gamma\left[p\left(\frac{1}{\delta}-1\right)\right]\Gamma\left[(1-p)\left(\frac{1}{\delta}-1\right)\right]\Gamma\left(n+\frac{1}{\delta}-1\right)}$$

for $0 \le p \le 1$; $\max(-\frac{p}{n-p-1}, -\frac{1-p}{n-2+p}) \le \delta \le 1$; $x = 0,1,2,...,n$

$E(x) = np$

$Var(x) = np(1-p)[1+(n-1)\delta]$

where $\Gamma(\cdot)$ is the Gamma function.

Remember that $x \mid \pi \sim \text{Binomial}(n,\pi)$, while $\pi \sim \text{Beta}(\alpha,\beta)$. The parameters $p = \alpha/(\alpha+\beta)$ and $\delta = 1/(\alpha+\beta+1)$ are estimated by the platform. To obtain estimates of α and β, use the following formulas:

$$\hat{\alpha} = \hat{p}\left(\frac{1-\hat{\delta}}{\hat{\delta}}\right)$$

$$\hat{\beta} = (1-\hat{p})\left(\frac{1-\hat{\delta}}{\hat{\delta}}\right)$$

If the estimate of δ is 0, the formulas do not work. In that case, the Beta Binomial has reduced to the Binomial(n,p), and $\hat{p}$ is the estimate of p.

The confidence intervals for the Beta Binomial parameters are profile likelihood intervals.

Run demoBetaBinomial.jsl in the JMP Samples/Scripts folder to compare a Beta Binomial distribution with dispersion parameter δ to a Binomial distribution with parameters p and n = 20.

Fitted Quantiles

The fitted quantiles in the Diagnostic Plot and the fitted quantiles saved with the **Save Fitted Quantiles** command are formed using the following method:

1. The data are sorted and ranked. Ties are assigned different ranks.
2. Compute the $p_{[i]} = \text{rank}_{[i]}/(n+1)$.
3. Compute the $\text{quantile}_{[i]} = \text{Quantile}_d(p_{[i]})$ where Quantile_d is the quantile function for the specific fitted distribution, and i = 1,2,...,n.

Fit Distribution Options

This section describes Goodness of Fit tests for fitting distributions and statistical details for specification limits pertaining to fitted distributions.

Goodness of Fit

Table 3.3 Descriptions of JMP Goodness of Fit Tests

Distribution	Parameters	Goodness of Fit Test
Normal[a]	μ and σ are unknown	Shapiro-Wilk (for n ≤ 2000) Kolmogorov-Smirnov-Lillefors (for n > 2000)
	μ and σ are both known	Kolmogorov-Smirnov-Lillefors
	either μ or σ is known	(none)
LogNormal	μ and σ are known or unknown	Kolmogorov's *D*
Weibull	α and β known or unknown	Cramér-von Mises W^2
Weibull with threshold	α, β and θ known or unknown	Cramér-von Mises W^2
Extreme Value	α and β known or unknown	Cramér-von Mises W^2
Exponential	σ is known or unknown	Kolmogorov's *D*
Gamma	α and σ are known	Cramér-von Mises W^2
	either α or σ is unknown	(none)
Beta	α and β are known	Kolmogorov's *D*
	either α or β is unknown	(none)
Binomial	ρ is known or unknown and *n* is known	Kolmogorov's *D* (for n ≤ 30) Pearson χ^2 (for n > 30)
Beta Binomial	ρ and δ known or unknown	Kolmogorov's *D* (for n ≤ 30) Pearson χ^2 (for n > 30)
Poisson	λ known or unknown	Kolmogorov's *D* (for n ≤ 30) Pearson χ^2 (for n > 30)
Gamma Poisson	λ or σ known or unknown	Kolmogorov's *D* (for n ≤ 30) Pearson χ^2 (for n > 30)

a. For the three Johnson distributions and the Glog distribution, the data are transformed to Normal, then the appropriate test of normality is performed.

Set Spec Limits for K Sigma

Type a K value and select one-sided or two-sided for your capability analysis. Tail probabilities corresponding to K standard deviations are computed from the Normal distribution. The probabilities are converted to quantiles for the specific distribution that you have fitted. The resulting quantiles are used for specification limits in the capability analysis. This option is similar to the **Quantiles** option, but you provide K instead of probabilities. K corresponds to the number of standard deviations that the specification limits are away from the mean.

For example, for a Normal distribution, where K=3, the 3 standard deviations below and above the mean correspond to the 0.00135^{th} quantile and 0.99865^{th} quantile, respectively. The lower specification limit is set at the 0.00135^{th} quantile, and the upper specification limit is set at the 0.99865^{th} quantile of the fitted distribution. A capability analysis is returned based on those specification limits.

Chapter 4

Introduction to Fit Y by X

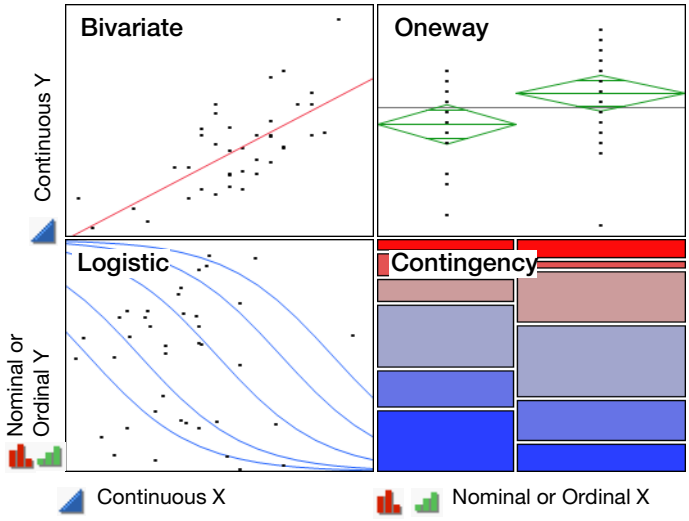
Examine Relationships Between Two Variables

The Fit Y by X platform analyzes the pair of X and Y variables that you specify, by context, based on modeling type.

Here are the four types of analyses:

- Bivariate fitting
- One-way analysis of variance
- Logistic regression
- Contingency table analysis

Figure 4.1 Examples of Four Types of Analyses



Overview of the Fit Y by X Platform

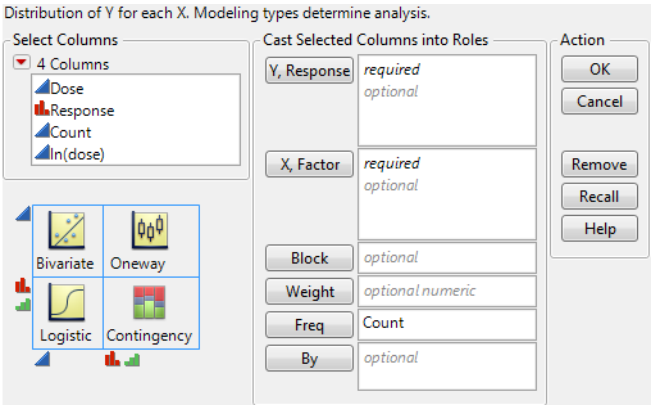
The Fit Y by X platform is a collection of four specific platforms (or types of analyses).

Specific Platform	Modeling Types	Description
Bivariate	Continuous Y by continuous X	Analyzes the relationship between two continuous variables. See "Bivariate Analysis".
Oneway	Continuous Y by nominal or ordinal X	Analyzes how the distribution of a continuous Y variable differs across groups defined by a categorical X variable. See "Oneway Analysis".
Logistic	Nominal or ordinal Y by continuous X	Fits the probabilities for response categories to a continuous X predictor. See "Logistic Analysis".
Contingency	Nominal or ordinal Y by nominal or ordinal X	Analyzes the distribution of a categorical response variable Y as conditioned by the values of a categorical X factor. See "Contingency Analysis".

Launch the Fit Y by X Platform

Launch the **Fit Y by X** platform by selecting **Analyze** > **Fit Y by X**.

Figure 4.2 The Fit Y by X Launch Window



Bivariate, Oneway, Logistic, Contingency This grid shows which analysis results from the different combinations of data types. Once you have assigned your columns, the applicable platform appears as a label above the grid.

Block (Optional, for Oneway and Contingency only):

- For the Oneway platform, identifies a second factor, which forms a two-way analysis without interaction. The data should be balanced and have equal counts in each block by group cell. If you specify a Block variable, the data should be balanced and have equal counts in each block by group cell. In the plot, the values of the Y variable are centered by the Block variable.
- For the Contingency platform, identifies a second factor and performs a Cochran-Mantel-Haenszel test.

For more information about launch windows, see the Get Started chapter in the *Using JMP* book.

Launch Specific Analyses from the JMP Starter Window

From the JMP Starter window, you can launch a specific analysis (**Bivariate**, **Oneway**, **Logistic**, or **Contingency**). If you select this option, specify the correct modeling types (Y and X variables) for the analysis. See Table 4.1.

To launch a specific analysis from the JMP Starter Window, click the **Basic** category, and select a specific analysis.

Most of the platform launch options are the same. However, the naming for some of the Y and X platform buttons is tailored for the specific analysis that you are performing.

Table 4.1 Platforms and Buttons

Platform or Analysis	Y Button	X Button
Fit Y by X	Y, Response	X, Factor
Bivariate	Y, Response	X, Regressor
Oneway	Y, Response	X, Grouping
Logistic	Y, Categorical Response	X, Continuous Regressor
Contingency	Y, Response Category	X, Grouping Category

Chapter 5

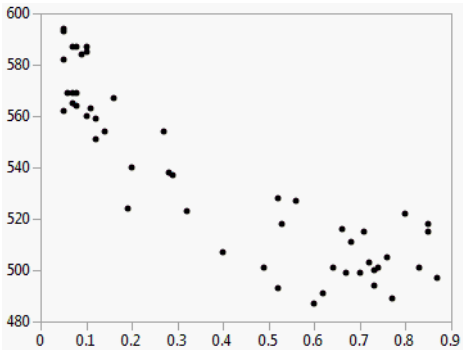
Bivariate Analysis

Examine Relationships between Two Continuous Variables

The Bivariate platform shows the relationship between two continuous variables. It is the *continuous by continuous* personality of the Fit Y by X platform. The word bivariate simply means involving two variables instead of one (univariate) or many (multivariate).

The Bivariate analysis results appear in a scatterplot. Each point on the plot represents the X and Y scores for a single subject; in other words, each point represents two variables. Using the scatterplot, you can see at a glance the degree and pattern of the relationship between the two variables. You can interactively add other types of fits, such as simple linear regression, polynomial regression, and so on.

Figure 5.1 Example of Bivariate Analysis

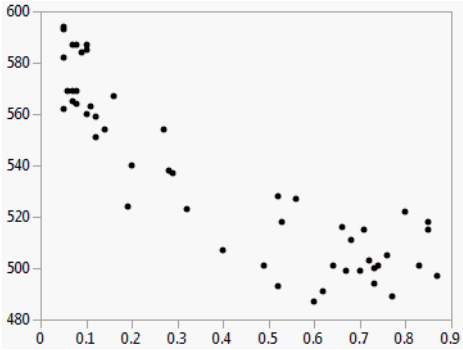


Example of Bivariate Analysis

This example uses the SAT.jmp sample data table. SAT test scores for students in the 50 U.S. states, plus the District of Columbia, are divided into two areas: verbal and math. You want to find out how the percentage of students taking the SAT tests is related to verbal test scores for 2004.

1. Select **Help > Sample Data Library** and open SAT.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select 2004 Verbal and click **Y, Response**.
4. Select % Taking (2004) and click **X, Factor.**
5. Click **OK**.

Figure 5.2 Example of SAT Scores by Percent Taking



You can see that the verbal scores were higher when a smaller percentage of the population took the test.

Launch the Bivariate Platform

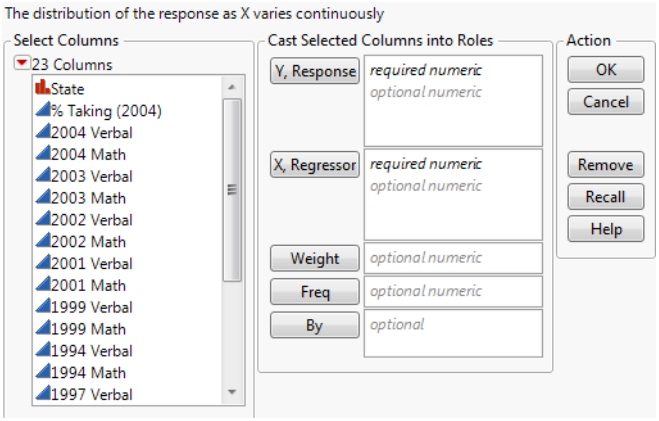
You can perform a bivariate analysis using either the Fit Y by X platform or the Bivariate platform. The two approaches give equivalent results.

- To launch the Fit Y by X platform, select **Analyze > Fit Y by X**.

  or

- To launch the Bivariate platform, from the JMP Starter window, click on the **Basic** category and click **Bivariate**.

Figure 5.3 The Bivariate Launch Window



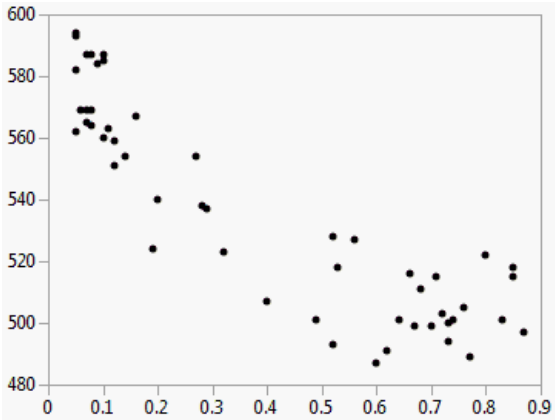
For information about this launch window, see "Introduction to Fit Y by X" chapter on page 95.

After you click **OK**, the Bivariate plot appears. See "The Bivariate Plot" on page 101.

The Bivariate Plot

To produce the plot shown in Figure 5.4, follow the instructions in "Example of Bivariate Analysis" on page 100.

Figure 5.4 The Bivariate Plot



Note: Any rows that are excluded in the data table are also hidden in the Bivariate plot.

The Bivariate report begins with a plot for each pair of *X* and *Y* variables. Replace variables in the plot by dragging and dropping a variable, in one of two ways: swap existing variables by dragging and dropping a variable from one axis to the other axis; or, click on a variable in the Columns panel of the associated data table and drag it onto an axis.

You can interact with this plot just as you can with other JMP plots (for example, resizing the plot, highlighting points with the arrow or brush tool, and labeling points). For details about these features, see the JMP Reports chapter in the *Using JMP* book.

You can fit curves on the plot and view statistical reports and additional menus using the fitting commands that are located within the red triangle menu. See "Fitting Commands and Options" on page 102.

Fitting Commands and Options

Note: The Fit Group menu appears if you have specified multiple Y or multiple X variables. Menu options allow you to arrange reports or order them by RSquare. See the Standard Least Squares chapter in the *Fitting Linear Models* book for more information.

The Bivariate Fit red triangle menu contains display options, fitting options, and control options.

Show Points Hides or shows the points in the scatterplot. A check mark indicates that points are shown.

Histogram Borders Attaches histograms to the *x*- and *y*-axes of the scatterplot. A check mark indicates that histogram borders are turned on. See "Histogram Borders" on page 105.

Note: When you apply only the Hidden row state to rows in the data table, the corresponding points do not appear in the scatterplot. However, the histograms are constructed using the hidden rows. If you want to exclude rows from the construction of the histograms and from analysis results, apply the Exclude row state and select **Redo > Redo Analysis** from the Bivariate red triangle menu.

Group By Lets you select a classification (or grouping) variable. A separate analysis is computed for each level of the grouping variable, and regression curves or ellipses are overlaid on the scatterplot. See "Group By" on page 121.

See the JMP Reports chapter in the *Using JMP* book for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

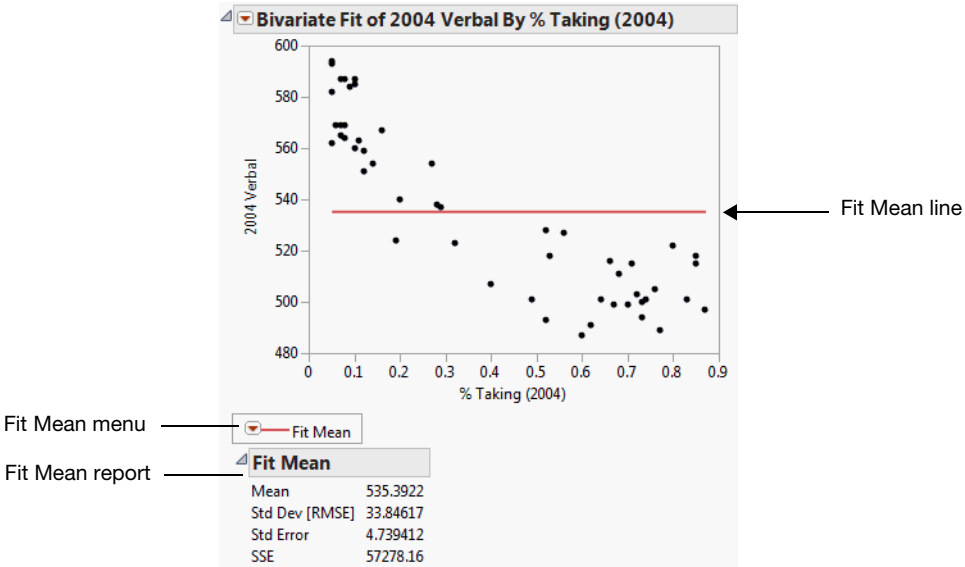
Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Fit Mean Options

Each fitting command adds the following:

- a line, curve, or distribution to the scatterplot
- a red triangle menu to the report window
- a specific report to the report window

Figure 5.5 Example of the Fit Mean Fitting Command



The following Fit Mean options are available:

Fit Mean Adds a horizontal line to the scatterplot that represents the mean of the Y response variable.See "Fit Mean" on page 105.

Fit Line Adds straight line fits to your scatterplot using least squares regression. See "Fit Line and Fit Polynomial" on page 106.

Fit Polynomial Fits polynomial curves of a certain degree using least squares regression. See "Fit Line and Fit Polynomial" on page 106.

Fit Special Transforms *Y* and *X*. Transformations include: log, square root, square, reciprocal, and exponential. You can also turn off center polynomials, constrain the intercept and the slope, and fit polynomial models. See "Fit Special" on page 113.

Flexible Provides options that enable you to control the smoothness of the estimated regression curve. See "Flexible" on page 114.

Fit Orthogonal Provides options for orthogonal regression fits, which are useful when X is assumed to vary. This option provides sub-options that reflect various assumptions about the variances of X and Y. See "Fit Orthogonal" on page 117.

Robust Provides options that reduce the influence of outliers in your data set on the fitted model. See "Robust" on page 118.

Density Ellipse Plots density ellipsoids for the bivariate normal distribution fit to the *X* and *Y* variables. See "Density Ellipse" on page 119.

Nonpar Density Plots density contours based on a smoothed surface. The contours describe the density of data points. See "Nonpar Density" on page 120.

Note: You can remove a fit using the **Remove Fit** command. For details, see "Fitting Menu Options" on page 122.

Fitting Command Categories

Fitting command categories include regression fits and density estimation.

Category	Description	Fitting Commands
Regression Fits	Regression methods fit a curve to the observed data points. The fitting methods include least squares fits as well as spline fits, kernel smoothing, orthogonal fits, and robust fits.	**Fit Mean** **Fit Line** **Fit Polynomial** **Fit Special** **Flexible** **Fit Orthogonal** **Robust**
Density Estimation	Density estimation fits a bivariate distribution to the points. You can either select a bivariate normal density, characterized by elliptical contours, or a general nonparametric density.	**Density Ellipse** **Nonpar Density**

Fit the Same Command Multiple Times

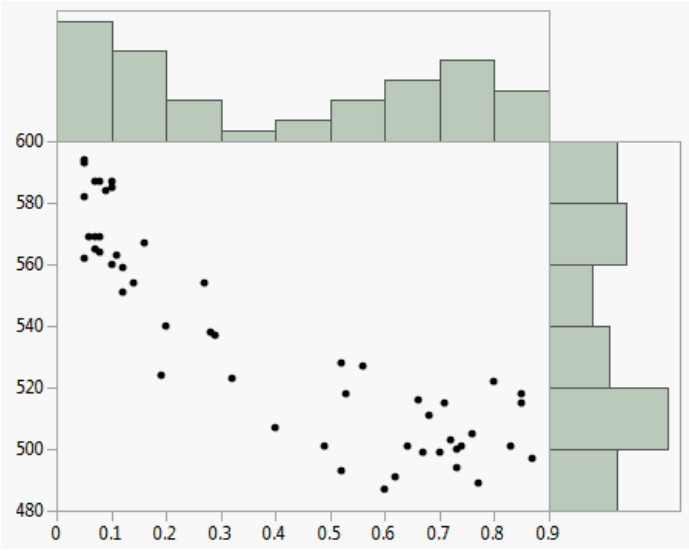
You can select the same fitting command multiple times, and each new fit is overlaid on the scatterplot. You can try fits, exclude points and refit, and you can compare them on the same scatterplot.

To apply a fitting command to multiple analyses in your report window, hold down the Ctrl key and select a fitting option.

Histogram Borders

The **Histogram Borders** option appends histograms to the *x*- and *y*-axes of the scatterplot. You can use the histograms to visualize the marginal distributions of the X and Y variables.

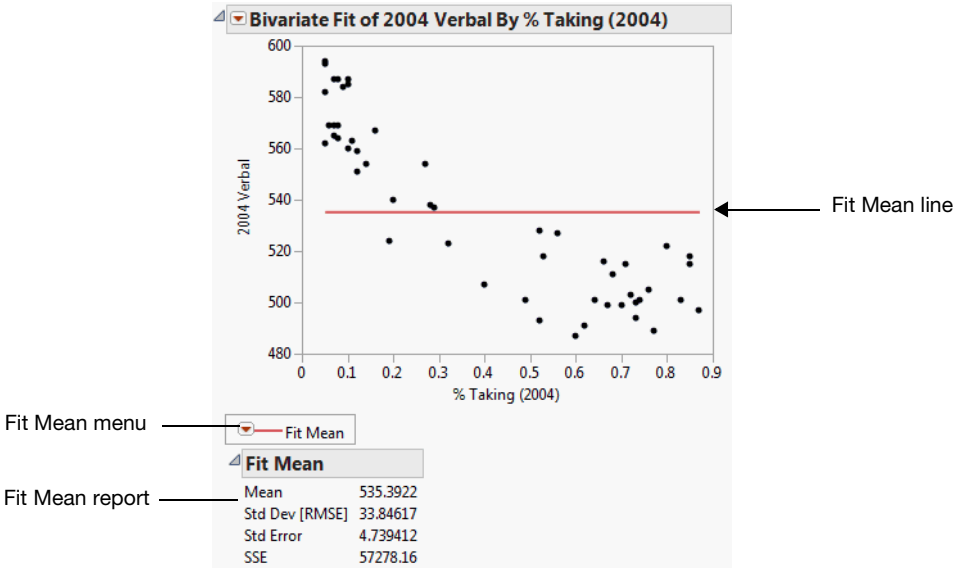
Figure 5.6 Example of Histogram Borders



Fit Mean

Using the **Fit Mean** command, you can add a horizontal line to the scatterplot that represents the mean of the *Y* response variable. You can start by fitting the mean and then use the mean line as a reference for other fits (such as straight lines, confidence curves, polynomial curves, and so on).

Figure 5.7 Example of Fit Mean



Fit Mean Report

The Fit Mean report shows summary statistics about the fit of the mean.

Mean Mean of the response variable. The predicted response when there are no specified effects in the model.

Std Dev [RMSE] Standard deviation of the response variable. Square root of the mean square error, also called the root mean square error (or RMSE).

Std Error Standard deviation of the response mean. Calculated by dividing the RMSE by the square root of the number of values.

SSE Error sum of squares for the simple mean model. Appears as the sum of squares for Error in the analysis of variance tables for each model fit.

Related Information

- "Fitting Menus" on page 122

Fit Line and Fit Polynomial

Using the **Fit Line** command, you can add straight line fits to your scatterplot using least squares regression. Using the **Fit Polynomial** command, you can fit polynomial curves of a certain degree using least squares regression.

Figure 5.8 Example of Fit Line and Fit Polynomial

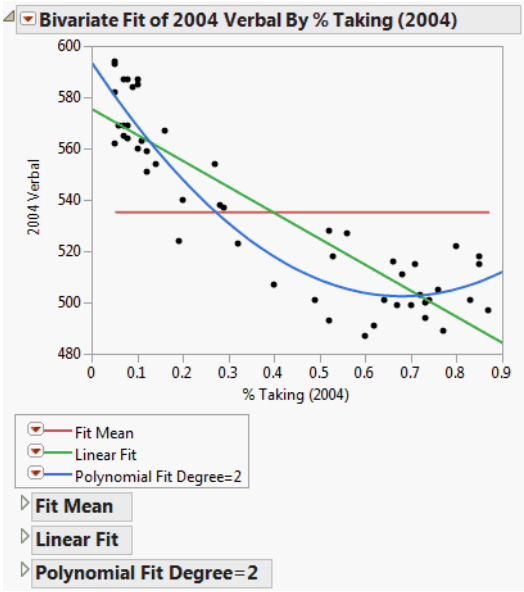


Figure 5.8 shows an example that compares a linear fit to the mean line and to a degree 2 polynomial fit.

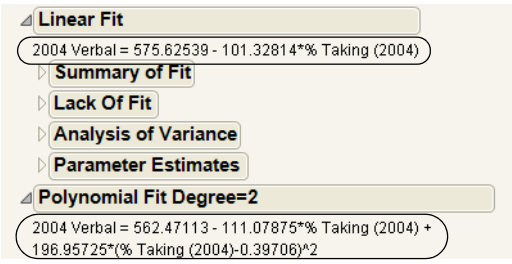
Note the following information:

- The **Fit Line** output is equivalent to a polynomial fit of degree 1.
- The **Fit Mean** output is equivalent to a polynomial fit of degree 0.

Linear Fit and Polynomial Fit Reports

The Linear Fit and Polynomial Fit reports begin with the equation of fit.

Figure 5.9 Example of Equations of Fit



Note: You can edit the equation by clicking on it.

Each Linear and Polynomial Fit Degree report contains at least three reports. A fourth report, Lack of Fit, appears only if there are *X* replicates in your data.

Summary of Fit Report

The Summary of Fit reports show the numeric summaries of the response for the linear fit and polynomial fit of degree 2 for the same data. You can compare multiple Summary of Fit reports to see the improvement of one model over another, indicated by a larger RSquare value and smaller Root Mean Square Error.

Figure 5.10 Summary of Fit Reports for Linear and Polynomial Fits

Linear Fit

2004 Verbal = 575.62539 - 101.32814*% Taking (2004)

Summary of Fit

RSquare	0.790274
RSquare Adj	0.785994
Root Mean Square Error	15.65752
Mean of Response	535.3922
Observations (or Sum Wgts)	51

Polynomial Fit Degree=2

2004 Verbal = 562.47113 - 111.07875*% Taking (2004) + 196.95725*(% Taking (2004)-0.39706)^2

Summary of Fit

RSquare	0.879413
RSquare Adj	0.874389
Root Mean Square Error	11.99566
Mean of Response	535.3922
Observations (or Sum Wgts)	51

The Summary of Fit report contains the following columns:

RSquare Measures the proportion of the variation explained by the model. The remaining variation is not explained by the model and attributed to random error. The RSquare is 1 if the model fits perfectly.

Note: A low RSquare value suggests that there may be variables not in the model that account for the unexplained variation. However, if your data are subject to a large amount of inherent variation, even a useful regression model may have a low RSquare value. Read the literature in your research area to learn about typical RSquare values.

The RSquare values in Figure 5.10 indicate that the polynomial fit of degree 2 gives a small improvement over the linear fit. See "Summary of Fit Report" on page 134.

RSquare Adj Adjusts the RSquare value to make it more comparable over models with different numbers of parameters by using the degrees of freedom in its computation. See "Summary of Fit Report" on page 134.

Root Mean Square Error Estimates the standard deviation of the random error. It is the square root of the mean square for Error in the Analysis of Variance report. See Figure 5.12.

Mean of Response Provides the sample mean (arithmetic average) of the response variable. This is the predicted response when no model effects are specified.

Observations Provides the number of observations used to estimate the fit. If there is a weight variable, this is the sum of the weights.

Lack of Fit Report

Note: The Lack of Fit report appears only if there are multiple rows that have the same *x* value.

Using the Lack of Fit report, you can estimate the error, regardless of whether you have the right form of the model. This occurs when multiple observations occur at the same *x* value. The error that you measure for these exact replicates is called *pure error*. This is the portion of the sample error that cannot be explained or predicted no matter what form of model is used. However, a lack of fit test might not be of much use if it has only a few degrees of freedom for it (few replicated *x* values).

Figure 5.11 Examples of Lack of Fit Reports for Linear and Polynomial Fits

Linear Fit

2004 Verbal = 575.62539 - 101.32814*% Taking (2004)

Summary of Fit

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	36	9,662.983	268.416	1.4850
Pure Error	13	2,349.750	180.750	Prob > F
Total Error	49	12,012.733		0.2252
				Max RSq
				0.9590

Polynomial Fit Degree=2

2004 Verbal = 562.47113 - 111.07875*% Taking (2004) + 196.95725*(% Taking (2004)-0.39706)^2

Summary of Fit

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	35	4,557.2473	130.207	0.7204
Pure Error	13	2,349.7500	180.750	Prob > F
Total Error	48	6,906.9973		0.7862
				Max RSq
				0.9590

The difference between the residual error from the model and the pure error is called the *lack of fit error*. The lack of fit error can be significantly greater than the pure error if you have the wrong functional form of the regressor. In that case, you should try a different type of model fit. The Lack of Fit report tests whether the lack of fit error is zero.

The Lack of Fit report contains the following columns:

Source The three sources of variation: Lack of Fit, Pure Error, and Total Error.

DF The *degrees of freedom* (DF) for each source of error.

- The **Total Error** DF is the degrees of freedom found on the Error line of the Analysis of Variance table (shown under the "Analysis of Variance Report" on page 110). It is the difference between the **Total** DF and the **Model** DF found in that table. The **Error** DF is partitioned into degrees of freedom for lack of fit and for pure error.
- The **Pure Error** DF is pooled from each group where there are multiple rows with the same values for each effect. See "Lack of Fit Report" on page 135.
- The **Lack of Fit** DF is the difference between the **Total Error** and **Pure Error** DF.

Sum of Squares The sum of squares (SS for short) for each source of error.

- The **Total Error** SS is the sum of squares found on the **Error** line of the corresponding Analysis of Variance table, shown under "Analysis of Variance Report" on page 110.

– The **Pure Error** SS is pooled from each group where there are multiple rows with the same value for the *x* variable. This estimates the portion of the true random error that is not explained by model *x* effect. See "Lack of Fit Report" on page 135.

– The **Lack of Fit** SS is the difference between the **Total Error** and **Pure Error** sum of squares. If the lack of fit SS is large, the model might not be appropriate for the data. The *F*-ratio described below tests whether the variation due to lack of fit is small enough to be accepted as a negligible portion of the pure error.

Mean Square The sum of squares divided by its associated degrees of freedom. This computation converts the sum of squares to an average (mean square). *F*-ratios for statistical tests are the ratios of mean squares.

F Ratio The ratio of mean square for lack of fit to mean square for Pure Error. It tests the hypothesis that the lack of fit error is zero.

Prob > F The probability of obtaining a greater *F*-value by chance alone if the variation due to lack of fit variance and the pure error variance are the same. A high *p* value means that there is not a significant lack of fit.

Max RSq The maximum R^2 that can be achieved by a model using only the variables in the model. See "Lack of Fit Report" on page 135.

Analysis of Variance Report

Analysis of variance (ANOVA) for a regression partitions the total variation of a sample into components. These components are used to compute an *F*-ratio that evaluates the effectiveness of the model. If the probability associated with the *F*-ratio is small, then the model is considered a better statistical fit for the data than the response mean alone.

The Analysis of Variance reports in Figure 5.12 compare a linear fit (**Fit Line**) and a second degree (**Fit Polynomial**). Both fits are statistically better from a horizontal line at the mean.

Figure 5.12 Examples of Analysis of Variance Reports for Linear and Polynomial Fits

Linear Fit

2004 Verbal = 575.62539 - 101.32814*% Taking (2004)

Summary of Fit

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	45,265.424	45,265.4	184.6379
Error	49	12,012.733	245.2	Prob > F
C. Total	50	57,278.157		<.0001*

Polynomial Fit Degree=2

2004 Verbal = 562.47113 - 111.07875*% Taking (2004) + 196.95725*(% Taking (2004)-0.39706)^2

Summary of Fit

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	50,371.160	25,185.6	175.0265
Error	48	6,906.997	143.9	Prob > F
C. Total	50	57,278.157		<.0001*

The Analysis of Variance Report contains the following columns:

Source The three sources of variation: **Model**, **Error**, and **C. Total**.

DF The degrees of freedom (DF) for each source of variation:

- A degree of freedom is subtracted from the total number of non missing values (*N*) for each parameter estimate used in the computation. The computation of the total sample variation uses an estimate of the mean. Therefore, one degree of freedom is subtracted from the total, leaving 49. The total corrected degrees of freedom are partitioned into the Model and Error terms.
- One degree of freedom from the total (shown on the **Model** line) is used to estimate a single regression parameter (the slope) for the linear fit. Two degrees of freedom are used to estimate the parameters (β_1 and β_2) for a polynomial fit of degree 2.
- The **Error** degrees of freedom is the difference between **C. Total** df and **Model** df.

Sum of Squares The sum of squares (SS for short) for each source of variation:

- In this example, the total (**C. Total**) sum of squared distances of each response from the sample mean is 57,278.157, as shown in Figure 5.12. That is the sum of squares for the base model (or simple mean model) used for comparison with all other models.
- For the linear regression, the sum of squared distances from each point to the line of fit reduces from 12,012.733. This is the residual or unexplained **(Error)** SS after fitting the model. The residual SS for a second degree polynomial fit is 6,906.997, accounting for slightly more variation than the linear fit. That is, the model accounts for more variation because the model SS are higher for the second degree polynomial than the linear fit. The **C. total** SS less the **Error** SS gives the sum of squares attributed to the model.

Mean Square The sum of squares divided by its associated degrees of freedom. The *F*-ratio for a statistical test is the ratio of the following mean squares:

- The **Model** mean square for the linear fit is 45,265.4. This value estimates the error variance, but only under the hypothesis that the model parameters are zero.
- The **Error** mean square is 245.2. This value estimates the error variance.

F Ratio The model mean square divided by the error mean square. The underlying hypothesis of the fit is that all the regression parameters (except the intercept) are zero. If this hypothesis is true, then both the mean square for error and the mean square for model estimate the error variance, and their ratio has an *F*-distribution. If a parameter is a significant model effect, the *F*-ratio is usually higher than expected by chance alone.

Prob > F The observed significance probability (*p*-value) of obtaining a greater *F*-value by chance alone if the specified model fits no better than the overall response mean. Observed significance probabilities of 0.05 or less are often considered evidence of a regression effect.

Parameter Estimates Report

The terms in the Parameter Estimates report for a linear fit are the intercept and the single x variable.

For a polynomial fit of order k, there is an estimate for the model intercept and a parameter estimate for each of the k powers of the X variable.

Figure 5.13 Examples of Parameter Estimates Reports for Linear and Polynomial Fits

Linear Fit

2004 Verbal = 575.62539 - 101.32814*% Taking (2004)

Summary of Fit

Lack Of Fit

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	575.62539	3.684288	156.24	<.0001*
% Taking (2004)	-101.3281	7.457094	-13.59	<.0001*

Polynomial Fit Degree=2

2004 Verbal = 562.47113 - 111.07875*% Taking (2004) + 196.95725*(% Taking (2004)-0.39706)^2

Summary of Fit

Lack Of Fit

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	562.47113	3.583843	156.95	<.0001*
% Taking (2004)	-111.0788	5.942967	-18.69	<.0001*
(% Taking (2004)-0.39706)^2	196.95725	33.06487	5.96	<.0001*

The Parameter Estimates report contains the following columns:

Term Lists the name of each parameter in the requested model. The intercept is a constant term in all models.

Estimate Lists the parameter estimates of the linear model. The prediction formula is the linear combination of these estimates with the values of their corresponding variables.

Std Error Lists the estimates of the standard errors of the parameter estimates. They are used in constructing tests and confidence intervals.

t Ratio Lists the test statistics for the hypothesis that each parameter is zero. It is the ratio of the parameter estimate to its standard error. If the hypothesis is true, then this statistic has a Student's t-distribution.

Prob>|t| Lists the observed significance probability calculated from each t-ratio. It is the probability of getting, by chance alone, a t-ratio greater (in absolute value) than the computed value, given a true null hypothesis. Often, a value below 0.05 (or sometimes 0.01) is interpreted as evidence that the parameter is significantly different from zero.

To reveal additional statistics, right-click in the report and select the **Columns** menu. Statistics not shown by default are as follows:

Lower 95% The lower endpoint of the 95% confidence interval for the parameter estimate.

Upper 95% The upper endpoint of the 95% confidence interval for the parameter estimate.

Std Beta The standardized parameter estimate. It is useful for comparing the effect of X variables that are measured on different scales. See "Parameter Estimates Report" on page 135.

VIF The variance inflation factor.

Design Std Error The design standard error for the parameter estimate. See "Parameter Estimates Report" on page 135.

Related Information

- "Fit Line" on page 133
- "Fitting Menus" on page 122

Fit Special

Using the **Fit Special** command, you can transform Y and X. Transformations include the following: log, square root, square, reciprocal, and exponential. You can also constrain the slope and intercept, fit a polynomial of specific degree, and center the polynomial.

The Specify Transformation or Constraint Window contains the following options:

Y Transformation Use these options to transform the Y variable.

X Transformation Use these options to transform the X variable.

Degree Use this option to fit a polynomial of the specified degree.

Centered Polynomial To turn off polynomial centering, deselect the **Centered Polynomial** check box. See Figure 5.19. Note that for transformations of the *X* variable, polynomial centering is not performed. Centering polynomials stabilizes the regression coefficients and reduces multicollinearity.

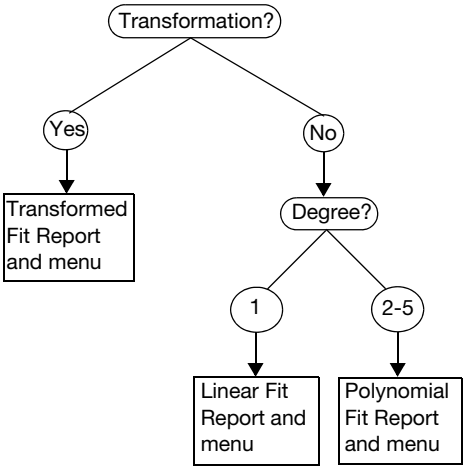
Constrain Intercept to Select this check box to constrain the model intercept to be the specified value.

Constrain Slope to Select this check box to constrain the model slope to be the specified value.

Fit Special Reports and Menus

Depending on your selections in the Fit Special window, you see different reports and menus. The flowchart in Figure 5.14 shows you what reports and menus you see depending on your choices.

Figure 5.14 Example of Fit Special Flowchart



Transformed Fit Report

The Transformed Fit report contains the reports described in "Linear Fit and Polynomial Fit Reports" on page 107. However, if you transformed *Y*, the Fit Measured on Original Scale report appears. This shows the measures of fit based on the original *Y* variables, and the fitted model transformed back to the original scale.

Related Information

- "Example of the Fit Special Command" on page 126
- "Linear Fit and Polynomial Fit Reports" on page 107
- "Fitting Menus" on page 122

Flexible

Use the options in the Flexible menu to control the smoothness of the estimated regression curve.

- Fit Spline uses a penalized least squares approach. Adjust the degree of smoothness using the parameter lambda.
- Kernel Smoother is based on locally weighted fits. Control the influence of local behavior using the parameter alpha.
- Fit Each Value calculates the mean response at each X value.

Fit Spline

Using the **Fit Spline** command, you can fit a smoothing spline that varies in smoothness (or flexibility) according to the lambda (λ) value. The lambda value is a tuning parameter in the spline formula. As the value of λ decreases, the error term of the spline model has more weight and the fit becomes more flexible and curved. As the value of λ increases, the fit becomes stiff (less curved), approaching a straight line.

Note the following information:

- The smoothing spline can help you see the expected value of the distribution of *Y* across *X*.
- The points closest to each piece of the fitted curve have the most influence on it. The influence increases as you lower the value of λ, producing a highly flexible curve.
- If you want to use a lambda value that is not listed on the menu, select **Fit Spline > Other**. If the scaling of the *X* variable changes, the fitted model also changes. To prevent this from happening, select the **Standardize X** option. Note that the fitted model remains the same for either the original *X* variable or the scaled *X* variable.
- You might find it helpful to try several λ values. You can use the **Lambda** slider beneath the Smoothing Spline report to experiment with different λ values. However, λ is not invariant to the scaling of the data. For example, the λ value for an *X* measured in inches, is not the same as the λ value for an *X* measured in centimeters.

Smoothing Spline Fit Report

The Smoothing Spline Fit report contains the R-Square for the spline fit and the Sum of Squares Error. You can use these values to compare the spline fit to other fits, or to compare different spline fits to each other.

R-Square Measures the proportion of variation accounted for by the smoothing spline model. For more information, see "Smoothing Fit Reports" on page 136.

Sum of Squares Error Sum of squared distances from each point to the fitted spline. It is the unexplained error (*residual*) after fitting the spline model.

Change Lambda Enables you to change the λ value, either by entering a number, or by moving the slider.

Related Information

- "Fitting Menus" on page 122
- "Fit Spline" on page 133

Kernel Smoother

The **Kernel Smoother** command produces a curve formed by repeatedly finding a locally weighted fit of a simple curve (a line or a quadratic) at sampled points in the domain. The many local fits (128 in total) are combined to produce the smooth curve over the entire domain. This method is also called *Loess* or *Lowess*, which was originally an acronym for Locally Weighted Scatterplot Smoother. See Cleveland (1979).

Use this method to quickly see the relationship between variables and to help you determine the type of analysis or fit to perform.

Local Smoother Report

The Local Smoother report contains the R-Square for the smoother fit and the Sum of Squares Error. You can use these values to compare the smoother fit to other fits, or to compare different smoother fits to each other.

R-Square Measures the proportion of variation accounted for by the smoother model. For more information, see "Smoothing Fit Reports" on page 136.

Sum of Squares Error Sum of squared distances from each point to the fitted smoother. It is the unexplained error (*residual*) after fitting the smoother model.

Local Fit (lambda) Select the polynomial degree for each local fit. Quadratic polynomials can track local bumpiness more smoothly. Lambda is the degree of certain polynomials that are fitted by the method. Lambda can be 0, 1 or 2.

Weight Function Specify how to weight the data in the neighborhood of each local fit. Loess uses tri-cube. The weight function determines the influence that each xi and yi has on the fitting of the line. The influence decreases as xi increases in distance from x and finally becomes zero.

Smoothness (alpha) Controls how many points are part of each local fit. Use the slider or type in a value directly. Alpha is a smoothing parameter. It can be any positive number, but typical values are 1/4 to 1. As alpha increases, the curve becomes smoother.

Robustness Re-weights the points to de-emphasize points that are farther from the fitted curve. Specify the number of times to repeat the process (number of passes). The goal is to converge the curve and automatically filter out outliers by giving them small weights.

Related Information

- "Fitting Menus" on page 122

Fit Each Value

The **Fit Each Value** command fits a value to each unique X value. The fitted values are the means of the response for each unique X value.

Fit Each Value Report

The Fit Each Value report shows summary statistics about the model fit.

Number of Observations Gives the total number of observations.

Number of Unique Values Gives the number of unique X values.

Degrees of Freedom Gives the pure error degrees of freedom.

Sum of Squares Gives the pure error sum of squares.

Mean Square Gives the pure error mean square.

Related Information

- "Fitting Menus" on page 122

Fit Orthogonal

The **Fit Orthogonal** command fits linear models that account for variability in *X* as well as *Y*.

Fit Orthogonal Options

Select one of the following options to specify a variance ratio.

Univariate Variances, Prin Comp Uses the univariate variance estimates computed from the samples of *X* and *Y*. This turns out to be the standardized first principal component. This option is not a good choice in a measurement systems application since the error variances are not likely to be proportional to the population variances.

Equal Variances Uses 1 as the variance ratio, which assumes that the error variances are the same. Using equal variances is equivalent to the non-standardized first principal component line. Suppose that the scatterplot is scaled the same in the *X* and *Y* directions. When you show a normal density ellipse, you see that this line is the longest axis of the ellipse.

Fit X to Y Uses a variance ratio of zero, which indicates that *Y* effectively has no variance.

Specified Variance Ratio Lets you enter any ratio that you want, giving you the ability to make use of known information about the measurement error in *X* and response error in *Y*.

Orthogonal Regression Report

The Orthogonal Regression report shows summary statistics about the orthogonal regression model.

Variable Gives the names of the variables used to fit the line.

Mean Gives the mean of each variable.

Std Dev Gives the standard deviation of each variable.

Variance Ratio Gives the variance ratio used to fit the line.

Correlation Gives the correlation between the two variables.

Intercept Gives the intercept of the fitted line.

Slope Gives the slope of the fitted line.

LowerCL Gives the lower confidence limit for the slope.

UpperCL Gives the upper confidence limit for the slope.

Alpha Enter the alpha level used in computing the confidence interval.

Related Information

- "Fitting Menus" on page 122
- "Fit Orthogonal" on page 133
- "Example Using the Fit Orthogonal Command" on page 127

Robust

The **Robust** option provides two methods to reduce the influence of outliers in your data set. Outliers can lead to incorrect estimates and decisions.

Fit Robust

The Fit Robust option reduces the influence of outliers in the response variable. The Huber M-estimation method is used. Huber M-estimation finds parameter estimates that minimize the Huber loss function, which penalizes outliers. The Huber loss function increases as a quadratic for small errors and linearly for large errors. For more details about robust fitting, see Huber (1973) and Huber and Ronchetti (2009).

Related Information

- "Fitting Menus" on page 122
- "Example Using the Fit Robust Command" on page 129

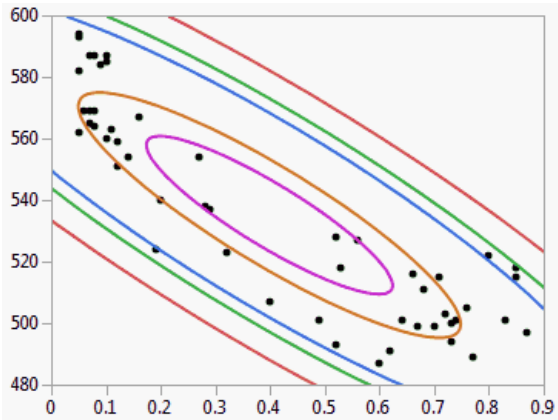
Fit Cauchy

Assumes that the errors have a Cauchy distribution. A Cauchy distribution has fatter tails than the normal distribution, resulting in a reduced emphasis on outliers. This option can be useful if you have a large proportion of outliers in your data. However, if your data are close to normal with only a few outliers, this option can lead to incorrect inferences. The Cauchy option estimates parameters using maximum likelihood and a Cauchy link function.

Density Ellipse

Using the **Density Ellipse** option, you can draw an ellipse (or ellipses) that contains the specified mass of points. The number of points is determined by the probability that you select from the **Density Ellipse** menu).

Figure 5.15 Example of Density Ellipses



The density ellipsoid is computed from the bivariate normal distribution fit to the *X* and *Y* variables. The bivariate normal density is a function of the means and standard deviations of the *X* and *Y* variables and the correlation between them. The **Other** selection lets you specify any probability greater than zero and less than or equal to one.

These ellipses are both density contours and confidence curves. As confidence curves, they show where a given percentage of the data is expected to lie, assuming the bivariate normal distribution.

The density ellipsoid is a good graphical indicator of the correlation between two variables. The ellipsoid collapses diagonally as the correlation between the two variables approaches either 1 or –1. The ellipsoid is more circular (less diagonally oriented) if the two variables are less correlated.

Correlation Report

The Correlation report that accompanies each **Density Ellipse** fit shows the correlation coefficient for the *X* and *Y* variables.

Note: To see a matrix of ellipses and correlations for many pairs of variables, use the **Multivariate** command in the **Analyze > Multivariate Methods** menu.

Variable Gives the names of the variables used in creating the ellipse

Mean Gives the average of both the *X* and *Y* variable.

Std Dev Gives the standard deviation of both the *X* and *Y* variable.

A discussion of the mean and standard deviation are in the section "The Summary Statistics Report" on page 42 in the "Distributions" chapter.

Correlation The Pearson correlation coefficient. If there is an exact linear relationship between two variables, the correlation is 1 or –1 depending on whether the variables are positively or negatively related. If there is no relationship, the correlation tends toward zero.

For more information, see "Correlation Report" on page 136.

Signif. Prob Probability of obtaining, by chance alone, a correlation with greater absolute value than the computed value if no linear relationship exists between the *X* and *Y* variables.

Number Gives the number of observations used in the calculations.

Related Information

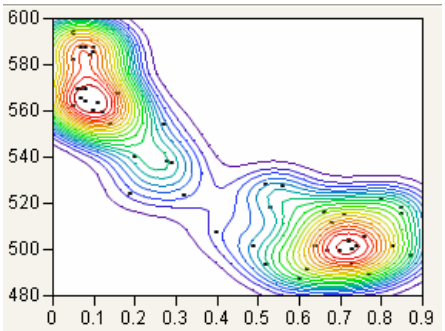
- "Fitting Menus" on page 122
- "Example of Group By Using Density Ellipses" on page 131

Nonpar Density

When a plot shows thousands of points, the mass of points can be too dark to show patterns in density. Using the Nonpar Density (nonparametric density) option makes it easier to see the patterns.

Nonpar Density estimates a smooth nonparametric bivariate surface that describes the density of data points. The plot adds a set of contour lines showing the density (Figure 5.16). The contour lines are quantile contours in 5% intervals. This means that about 5% of the points generated from the estimated nonparametric distribution are below the lowest contour, 10% are below the next contour, and so on. The highest contour has about 95% of the points below it.

Figure 5.16 Example of Nonpar Density



You can change the size of a nonparametric density contour grid to create smoother contours. The default value is 51 points, which can create jagged contours around dense points.

Press Shift and select **Nonpar Density** from the Bivariate red triangle menu. Enter a larger value than the default 51 points.

Nonparametric Bivariate Density Report

The nonparametric bivariate density report shows the standard deviations used in creating the nonparametric density.

Related Information

- "Fitting Menus" on page 122

Group By

Using the **Group By** option, you can select a classification (grouping) variable. When a grouping variable is in effect, the Bivariate platform computes a separate analysis for each level of the grouping variable. Regression curves or ellipses then appear on the scatterplot. The fit for each level of the grouping variable is identified beneath the scatterplot, with individual popup menus to save or remove fitting information.

The **Group By** option is checked in the Fitting menu when a grouping variable is in effect. You can change the grouping variable by first selecting the **Group By** option to remove (uncheck) the existing variable. Then, select the **Group By** option again and respond to its window as before.

You might use the Group By option in these different ways:

- An overlay of linear regression lines lets you compare slopes visually.
- An overlay of density ellipses can show clusters of points by levels of a grouping variable.

Related Information

- "Example of Group By Using Density Ellipses" on page 131
- "Example of Group By Using Regression Lines" on page 132

Fitting Menus

In addition to a report, each fitting command adds a fitting menu to the report window. The following table shows the fitting menus that correspond to each fitting command.

Fitting Command	Fitting Menu
Fit Mean	Fit Mean
Fit Line	Linear Fit
Fit Polynomial	Polynomial Fit Degree=X*
Fit Special	Linear Fit Polynomial Fit Degree=X* Transformed Fit X* Constrained Fits
Fit Spline	Smoothing Spline Fit, lambda=X*
Smoother	Local Smoother
Fit Each Value	Fit Each Value
Fit Orthogonal	Orthogonal Fit Ratio=X*
Fit Robust	Robust Fit
Fit Cauchy	Cauchy Fit
Density Ellipse	Bivariate Normal Ellipse P=X*
Nonpar Density	Quantile Density Colors

*X=variable character or number

Fitting Menu Options

The Fitting menu for the option that you have selected contains options that apply to that fit.

- "Options That Apply to Most Fits" on page 123.

- "Options That Apply to Multiple Fits" on page 123.
- "Options That Apply to Bivariate Normal Ellipse" on page 124.
- "Options That Apply to Quantile Density Contours" on page 124.

Options That Apply to Most Fits

Line of Fit Displays or hides the line or curve describing the model fit. For the Bivariate Normal Ellipse report, this option shows or hides the ellipse representing the contour border. Not applicable for Quantile Density Colors.

Line Color Lets you select from a palette of colors for assigning a color to each fit. Not applicable for Quantile Density Colors.

Line Style Lets you select from the palette of line styles for each fit. Not applicable for Quantile Density Colors.

Line Width Gives three line widths for the line of fit. The default line width is the thinnest line. Not applicable for Quantile Density Colors.

Report Turns the fit's report on and off. Does not modify the Bivariate plot.

Remove Fit Removes the fit from the graph and removes its report.

Options That Apply to Multiple Fits

Confid Curves Fit Displays or hides the confidence limits for the expected value (mean). This option is not available for the Fit Spline, Density Ellipse, Fit Each Value, and Fit Orthogonal fits and is dimmed on those menus.

Confid Curves Indiv Displays or hides the confidence limits for an individual predicted value. The confidence limits reflect variation in the error and variation in the parameter estimates. This option is not available for the Fit Mean, Fit Spline, Density Ellipse, Fit Each Value, and Fit Orthogonal fits and is dimmed on those menus.

Save Predicteds Creates a new column in the current data table called Predicted colname where colname is the name of the *Y* variable. This column includes the prediction formula and the computed sample predicted values. The prediction formula computes values automatically for rows that you add to the table. This option is not available for the Fit Each Value and Density Ellipse fits and is dimmed on those menus.

You can use the **Save Predicteds** and **Save Residuals** commands for each fit. If you use these commands multiple times or with a grouping variable, it is best to rename the resulting columns in the data table to reflect each fit.

Save Residuals Creates a new column in the current data table called Residuals colname where colname is the name of the *Y* variable. Each value is the difference between the actual (observed) value and its predicted value. Unlike the **Save Predicteds** command, this

command does not create a formula in the new column. This option is not available for the Fit Each Value and Density Ellipse fits and is dimmed on those menus.

You can use the **Save Predicteds** and **Save Residuals** commands for each fit. If you use these commands multiple times or with a grouping variable, it is best to rename the resulting columns in the data table to reflect each fit.

Mean Confidence Limit Formula Creates a new column in the data table containing a formula for the mean confidence intervals.

Indiv Confidence Limit Formula Creates a new column in the data table containing a formula for the individual confidence intervals.

Plot Residuals (Linear, Polynomial, and Fit Special Only) Produces four diagnostic plots: residual by predicted, actual by predicted, residual by row, and a normal quantile plot of the residuals. See "Diagnostics Plots" on page 125.

Set a Level Enables you to set the alpha level used in computing confidence bands for various fits.

Confid Shaded Fit Draws the same curves as the **Confid Curves Fit** command and shades the area between the curves.

Confid Shaded Indiv Draws the same curves as the **Confid Curves Indiv** command and shades the area between the curves.

Save Coefficients Saves the spline coefficients as a new data table, with columns called X, A, B, C, and D. The X column gives the knot points. A, B, C, and D are the intercept, linear, quadratic, and cubic coefficients of the third-degree polynomial. These coefficients span from the corresponding value in the X column to the next highest value.

Options That Apply to Bivariate Normal Ellipse

Shaded Contour Shades the area inside the density ellipse.

Select Points Inside Selects the points inside the ellipse.

Select Points Outside Selects the points outside the ellipse.

Options That Apply to Quantile Density Contours

Kernel Control Displays a slider for each variable, where you can change the standard deviation that defines the range of X and Y values for determining the density of contour lines.

5% Contours Shows or hides the 5% contour lines.

Contour Lines Shows or hides the 10% contour lines.

Contour Fill Fills the areas between the contour lines.

Select Points by Density Selects points that fall in a user-specified quantile range.

Color by Density Quantile Colors the points according to density.

Save Density Quantile Creates a new column containing the density quantile each point is in.

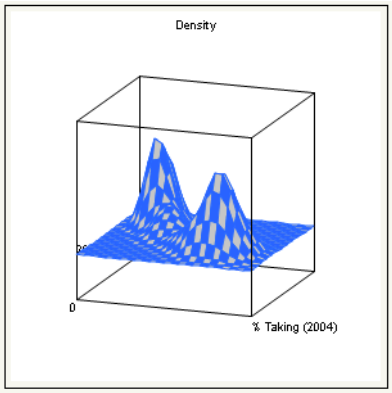
Mesh Plot Is a three-dimensional plot of the density over a grid of the two analysis variables. See Figure 5.17.

Modal Clustering Creates a new column in the current data table and fills it with cluster values.

Note: If you save the modal clustering values first and then save the density grid, the grid table also contains the cluster values. The cluster values are useful for coloring and marking points in plots.

Save Density Grid Saves the density estimates and the quantiles associated with them in a new data table. The grid data can be used to visualize the density in other ways, such as with the Scatterplot 3D or the Contour Plot platforms.

Figure 5.17 Example of a Mesh Plot



Diagnostics Plots

The **Plot Residuals** option creates residual plots and other plots to diagnose the model fit. The following plots are available:

Residual by Predicted Plot is a plot of the residuals vs. the predicted values. A histogram of the residuals is also created.

Actual by Predicted Plot is a plot of the actual values vs. the predicted values.

Residual by Row Plot is a plot of the residual values vs. the row number.

Residual by X Plot is a plot of the residual values vs. the X variable.

Residual Normal Quantile Plot is a Normal quantile plot of the residuals.

Additional Examples of the Bivariate Platform

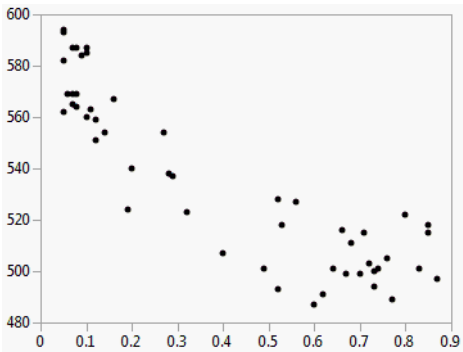
This section contains additional examples using the fitting commands in the Bivariate platform.

Example of the Fit Special Command

To transform Y as log and X as square root, proceed as follows:

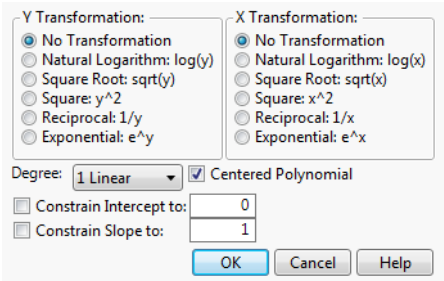
1. Select **Help > Sample Data Library** and open SAT.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select 2004 Verbal and click **Y, Response**.
4. Select % Taking (2004) and click **X, Factor.**
5. Click **OK**.

Figure 5.18 Example of SAT Scores by Percent Taking



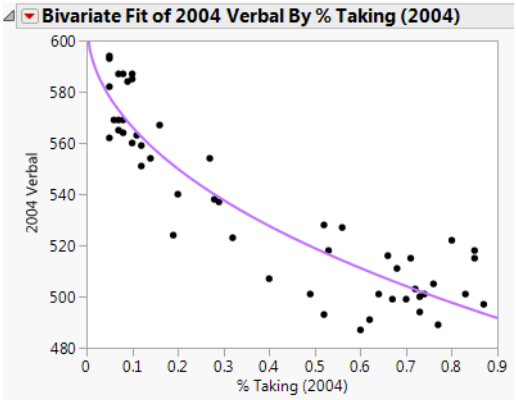
6. From the red triangle menu for Bivariate Fit, select **Fit Special**. The Specify Transformation or Constraint window appears. For a description of this window, see "Fit Special" on page 113.

Figure 5.19 The Specify Transformation or Constraint Window



7. Within Y Transformation, select Natural Logarithm: log(y).
8. Within X Transformation, select Square Root: sqrt(x).
9. Click **OK**.

Figure 5.20 Example of Fit Special Report



Transformed Fit Log to Sqrt

Transformed Fit Log to Sqrt

Log(2004 Verbal) = 6.4095983 - 0.223274*Sqrt(% Taking (2004))

Summary of Fit

RSquare	0.843694
RSquare Adj	0.840504
Root Mean Square Error	0.025112
Mean of Response	6.281054
Observations (or Sum Wgts)	51

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.16678913	0.166789	264.4873
Error	49	0.03090004	0.000631	Prob > F
C. Total	50	0.19768916		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6.4095983	0.008651	740.91	<.0001*
Sqrt(% Taking (2004))	-0.223274	0.013729	-16.26	<.0001*

Fit Measured on Original Scale

Sum of Squared Error	8684.2196
Root Mean Square Error	13.312737
RSquare	0.8483851
Sum of Residuals	8.4357902

Figure 5.20 shows the fitted line plotted on the original scale. The model appears to fit the data well, as the plotted line goes through the cloud of points.

Example Using the Fit Orthogonal Command

This example involves two parts. First, standardize the variables using the Distribution platform. Then, use the standardized variables to fit the orthogonal model.

Standardize the Variables

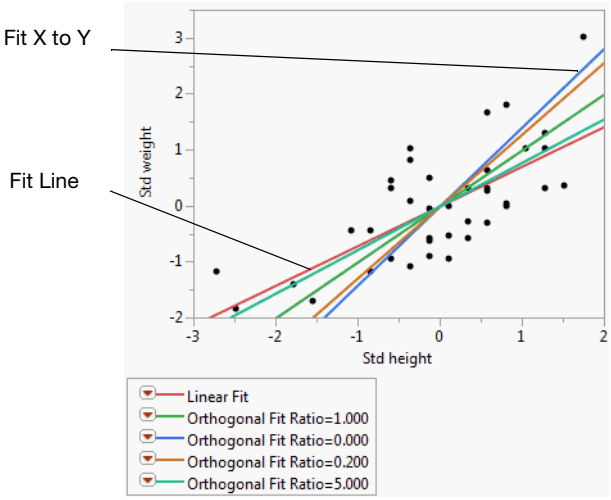
1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Distribution**.
3. Select height and weight and click **Y, Columns**.
4. Click **OK**.
5. Hold down the Ctrl key. On the red triangle menu next to height, select **Save > Standardized**.

   Holding down the Ctrl key broadcasts the operation to all variables in the report window. Notice that in the Big Class.jmp sample data table, two new columns have been added.
6. Close the Distribution report window.

Use the Standardized Variables to Fit the Orthogonal Model

1. From the Big Class.jmp sample data table, select **Analyze > Fit Y by X**.
2. Select Std weight and click **Y, Response**.
3. Select Std height and click **X, Factor**.
4. Click **OK**.
5. From the red triangle menu, select **Fit Line**.
6. From the red triangle menu, select **Fit Orthogonal**. Then select each of the following:
   - **Equal Variances**
   - **Fit X to Y**
   - **Specified Variance Ratio** and type 0.2.
   - **Specified Variance Ratio** and type 5.

Figure 5.21 Example of Orthogonal Fitting Options



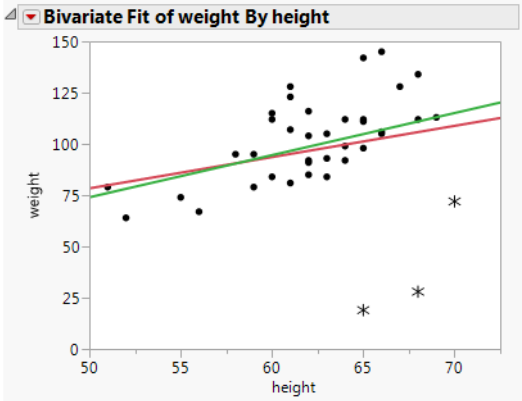
The scatterplot in Figure 5.21 shows the standardized height and weight values with various line fits that illustrate the behavior of the orthogonal variance ratio selections. The standard linear regression (**Fit Line**) occurs when the variance of the *X* variable is considered to be very small. **Fit X to Y** is the opposite extreme, when the variation of the *Y* variable is ignored. All other lines fall between these two extremes and shift as the variance ratio changes. As the variance ratio increases, the variation in the *Y* response dominates and the slope of the fitted line shifts closer to the *Y* by *X* fit. Likewise, when you decrease the ratio, the slope of the line shifts closer to the *X* by *Y* fit.

Example Using the Fit Robust Command

The data in the Weight Measurements.jmp sample data table shows the height and weight measurements taken by 40 students.

1. Select **Help > Sample Data Library** and open Weight Measurements.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response**.
4. Select height and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **Fit Line**.
7. From the red triangle menu, select **Robust > Fit Robust**.

Figure 5.22 Example of Robust Fit



Linear Fit

weight = 2.1466733 + 1.5244337*height

Summary of Fit

RSquare	0.062319
RSquare Adj	0.037643
Root Mean Square Error	25.41387
Mean of Response	97.5
Observations (or Sum Wgts)	40

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1631.144	1631.14	2.5255
Error	38	24542.856	645.86	Prob > F
C. Total	39	26174.000		0.1203

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.1466733	60.13567	0.04	0.9717
height	1.5244337	0.959253	1.59	0.1203

Robust Fit

Sigma	ChiSquare	PValue	LogWorth
19.9537	3.88052	0.0488*	1.31114

Parameter	Robust Estimate	Std Error
Intercept	-28.62	62.5811
height	2.05316	1.04226

If you look at the standard Analysis of Variance report, you might wrongly conclude that height and weight do not have a linear relationship, since the p-value is 0.1203. However, when you look at the Robust Fit report, you would probably conclude that they do have a linear relationship, because the p-value there is 0.0489. It appears that some of the measurements are unusually low, perhaps due to incorrect user input. These measurements were unduly influencing the analysis.

Example of Group By Using Density Ellipses

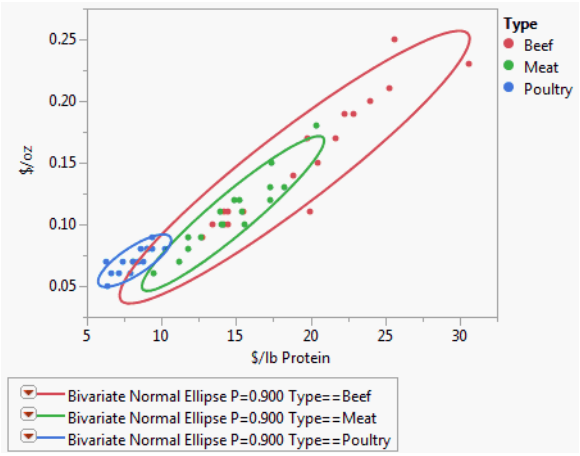
This example uses the Hot Dogs.jmp sample data table. The Type column identifies three different types of hot dogs: beef, meat, or poultry. You want to group the three types of hot dogs according to their cost variables.

1. Select **Help > Sample Data Library** and open Hot Dogs.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select $/oz and click **Y, Response**.
4. Select $/lb Protein and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **Group By**.
7. From the list, select Type.
8. Click **OK**. If you look at the **Group By** option again, you see it has a check mark next to it.
9. From the red triangle menu, select **Density Ellipse > 0.90**.

To color the points according to Type, proceed as follows:

10. Right-click on the scatterplot and select **Row Legend**.
11. Select Type in the column list and click **OK**.

Figure 5.23 Example of Group By



The ellipses in Figure 5.23 show clearly how the different types of hot dogs cluster with respect to the cost variables.

Example of Group By Using Regression Lines

Another use for grouped regression is overlaying lines to compare slopes of different groups.

1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response**.
4. Select height and click **X, Factor**.
5. Click **OK**.

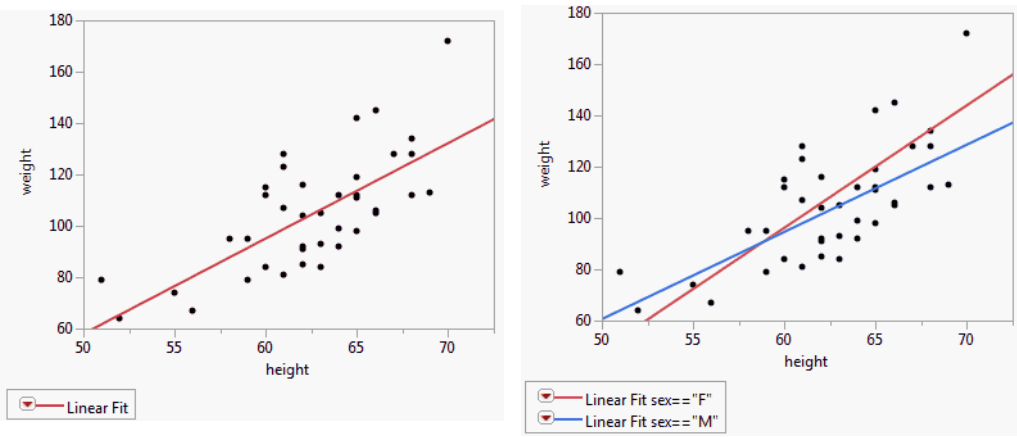
To create the example on the left in Figure 5.24:

6. Select **Fit Line** from the red triangle menu.

To create the example on the right in Figure 5.24:

7. From the Linear Fit menu, select **Remove Fit**.
8. From the red triangle menu, select **Group By**.
9. From the list, select sex.
10. Click **OK**.
11. Select **Fit Line** from the red triangle menu.

Figure 5.24 Example of Regression Analysis for Whole Sample and Grouped Sample



The scatterplot to the left in Figure 5.24 has a single regression line that relates weight to height. The scatterplot to the right shows separate regression lines for males and females.

Statistical Details for the Bivariate Platform

This section contains statistical details for selected commands and reports.

Fit Line

The **Fit Line** command finds the parameters β_0 and β_1 for the straight line that fits the points to minimize the residual sum of squares. The model for the *i*th row is written $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

A polynomial of degree 2 is a parabola; a polynomial of degree 3 is a cubic curve. For degree *k*, the model for the *i*th observation is as follows:

$$y_i = \sum_{j=0}^{k} \beta_j x_i^j + \varepsilon_i$$

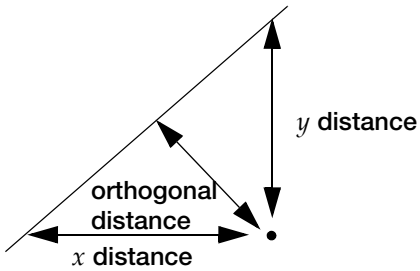
Fit Spline

The cubic spline method uses a set of third-degree polynomials spliced together such that the resulting curve is continuous and smooth at the splices (knot points). The estimation is done by minimizing an objective function that is a combination of the sum of squares error and a penalty for curvature integrated over the curve extent. See the paper by Reinsch (1967) or the text by Eubank (1988) for a description of this method.

Fit Orthogonal

Standard least square fitting assumes that the *X* variable is fixed and the *Y* variable is a function of *X* plus error. If there is random variation in the measurement of *X*, you should fit a line that minimizes the sum of the squared perpendicular differences. See Figure 5.25. However, the perpendicular distance depends on how *X* and *Y* are scaled, and the scaling for the perpendicular is reserved as a statistical issue, not a graphical one.

Figure 5.25 Line Perpendicular to the Line of Fit



The fit requires that you specify the ratio of the variance of the error in Y to the error in X. This is the variance of the error, not the variance of the sample points, so you must choose carefully. The ratio $(\sigma_y^2)/(\sigma_x^2)$ is infinite in standard least squares because σ_x^2 is zero. If you do an orthogonal fit with a large error ratio, the fitted line approaches the standard least squares line of fit. If you specify a ratio of zero, the fit is equivalent to the regression of X on Y, instead of Y on X.

The most common use of this technique is in comparing two measurement systems that both have errors in measuring the same value. Thus, the Y response error and the X measurement error are both the same type of measurement error. Where do you get the measurement error variances? You cannot get them from bivariate data because you cannot tell which measurement system produces what proportion of the error. So, you either must blindly assume some ratio like 1, or you must rely on separate repeated measurements of the same unit by the two measurement systems.

An advantage to this approach is that the computations give you predicted values for both Y and X; the predicted values are the point on the line that is closest to the data point, where closeness is relative to the variance ratio.

Confidence limits are calculated as described in Tan and Iglewicz (1999).

Summary of Fit Report

RSquare

Using quantities from the corresponding analysis of variance table, the RSquare for any continuous response fit is calculated as follows:

$$\frac{\text{Sum of Squares for Model}}{\text{Sum of Squares for C. Total}}$$

RSquare Adj

The RSquare Adj is a ratio of mean squares instead of sums of squares and is calculated as follows:

$$1 - \frac{\text{Mean Square for Error}}{\text{Mean Square for C. Total}}$$

The mean square for Error is in the Analysis of Variance report. See Figure 5.12. You can compute the mean square for C. Total as the Sum of Squares for C. Total divided by its respective degrees of freedom.

Lack of Fit Report

Pure Error DF

For the Pure Error DF, consider the multiple instances in the Big Class.jmp sample data table where more than one subject has the same value of height. In general, if there are g groups having multiple rows with identical values for each effect, the pooled DF, denoted DF_p, is as follows:

$$DF_p = \sum_{i=1}^{g} (n_i - 1)$$

n_i is the number of subjects in the ith group.

Pure Error SS

For the Pure Error SS, in general, if there are g groups having multiple rows with the same x value, the pooled SS, denoted SS_p, is written as follows:

$$SS_p = \sum_{i=1}^{g} SS_i$$

where SS_i is the sum of squares for the ith group corrected for its mean.

Max RSq

Because Pure Error is invariant to the form of the model and is the minimum possible variance, Max RSq is calculated as follows:

$$1 - \frac{\text{SS(Pure error)}}{\text{SS(Total for whole model)}}$$

Parameter Estimates Report

Std Beta

Std Beta is calculated as follows:

$$\hat{\beta}(s_x / s_y)$$

where $\hat{\beta}$ is the estimated parameter, sx and sy are the standard deviations of the X and Y variables.

Design Std Error

Design Std Error is calculated as the standard error of the parameter estimate divided by the RMSE.

Smoothing Fit Reports

R-Square is equal to 1-(SSE/C.Total SS), where C.Total SS is available in the Fit Line ANOVA report.

Correlation Report

The Pearson correlation coefficient is denoted *r*, and is computed as follows:

$$r_{xy} = \frac{s_{xy}^2}{\sqrt{s_x^2 s_y^2}} \text{ where } s_{xy}^2 = \frac{\sum w_i (x_i - \overline{x}_i)(y_i - \overline{y}_i)}{df}$$

Where w_i is either the weight of the *i*th observation if a weight column is specified, or 1 if no weight column is assigned.

Chapter 6

Oneway Analysis

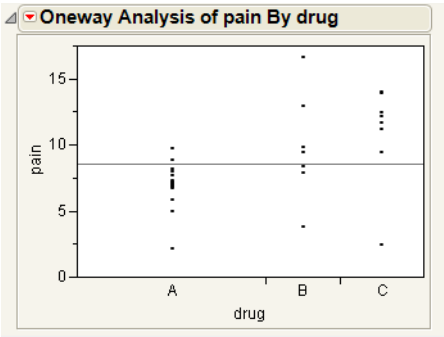
Examine Relationships between a Continuous Y and a Categorical X Variable

Using the Oneway or Fit Y by X platform, you can explore how the distribution of a continuous Y variable differs across groups defined by a single categorical X variable. For example, you might want to find out how different categories of the same type of drug (X) affect patient pain levels on a numbered scale (Y).

The Oneway platform is the *continuous by nominal or ordinal* personality of the Fit Y by X platform. The analysis results appear in a plot, and you can interactively add additional analyses, such as the following:

- a one-way analysis of variance to fit means and to test that they are equal
- nonparametric tests
- a test for homogeneity of variance
- multiple-comparison tests on means, with means comparison circles
- outlier box plots overlaid on each group
- power details for the one-way layout

Figure 6.1 Oneway Analysis



Overview of Oneway Analysis

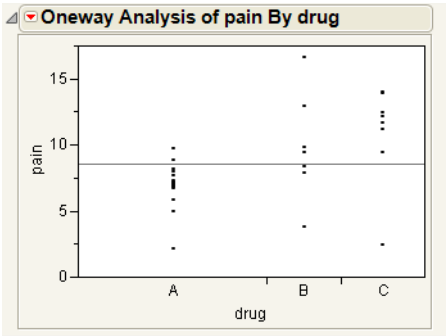
A one-way analysis of variance tests for differences between group means. The total variability in the response is partitioned into two parts: within-group variability and between-group variability. If the between-group variability is large relative to the within-group variability, then the differences between the group means are considered to be significant.

Example of Oneway Analysis

This example uses the Analgesics.jmp sample data table. Thirty-three subjects were administered three different types of analgesics (A, B, and C). The subjects were asked to rate their pain levels on a sliding scale. You want to find out if the means for A, B, and C are significantly different.

1. Select **Help > Sample Data Library** and open Analgesics.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select pain and click **Y, Response**.
4. Select drug and click **X, Factor**.
5. Click **OK**.

Figure 6.2 Example of Oneway Analysis



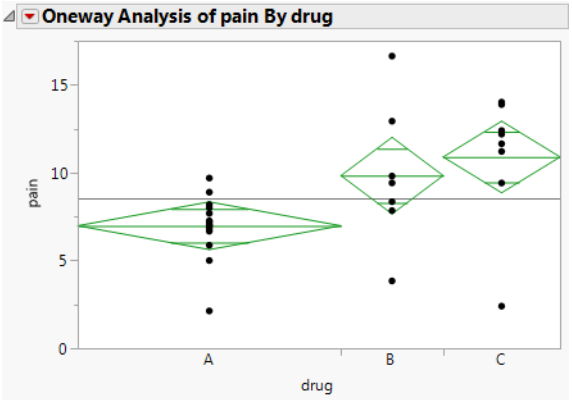
You notice that one drug (A) has consistently lower scores than the other drugs. You also notice that the x-axis ticks are unequally spaced. The length between the ticks is proportional to the number of scores (observations) for each drug.

Perform an analysis of variance on the data.

6. From the red triangle menu for Oneway Analysis, select **Means/Anova**.

Note: If the X factor has only two levels, the **Means/Anova** option appears as **Means/Anova/Pooled t,** and adds a pooled *t*-test report to the report window.

Figure 6.3 Example of the Means/Anova Option



Oneway Anova

Summary of Fit

Rsquare	0.295046
Adj Rsquare	0.248049
Root Mean Square Error	2.820631
Mean of Response	8.533515
Observations (or Sum Wgts)	33

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
drug	2	99.89459	49.9473	6.2780	0.0053*
Error	30	238.67877	7.9560		
C. Total	32	338.57335			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
A	18	6.9791	0.6648	5.6214	8.337
B	7	9.8318	1.0661	7.6545	12.009
C	8	10.8948	0.9972	8.8582	12.931

Std Error uses a pooled estimate of error variance

Welch's Test

F Ratio	DFNum	DFDen	Prob > F
5.0557	2	9.7386	0.0312*

Welch Anova testing Means Equal, allowing Std Devs Not Equal

Note the following observations:

- Mean diamonds representing confidence intervals appear.
  - The line near the center of each diamond represents the group mean. At a glance, you can see that the mean for each drug looks significantly different.
  - The vertical span of each diamond represents the 95% confidence interval for the mean of each group.

  See "Mean Diamonds and X-Axis Proportional" on page 150.
- The Summary of Fit table provides overall summary information about the analysis.

- The Analysis of Variance report shows the standard ANOVA information. You notice that the Prob > F (the *p*-value) is 0.0053, which supports your visual conclusion that there are significant differences between the drugs.
- The Means for Oneway Anova report shows the mean, sample size, and standard error for each level of the categorical factor.

Launch the Oneway Platform

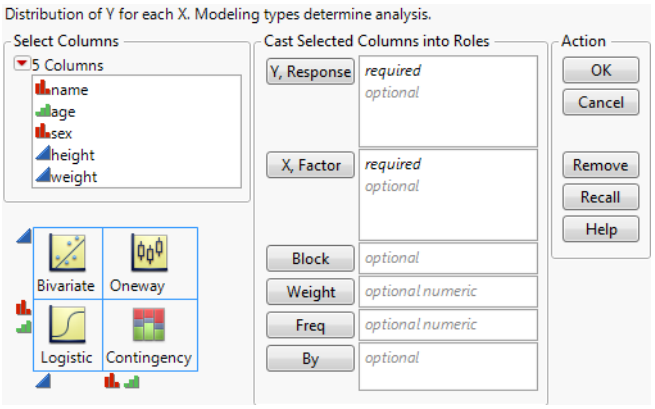
You can perform a Oneway analysis using either the Fit Y by X platform or the Oneway platform. The two approaches are equivalent.

- To launch the Fit Y by X platform, select **Analyze > Fit Y by X**.

  or

- To launch the Oneway platform, from the JMP Starter window, click on the **Basic** category and click **Oneway**.

Figure 6.4 The Oneway Launch Window



For more information about this launch window, see "Introduction to Fit Y by X" chapter on page 95.

After you click **OK**, the Oneway report window appears. See "The Oneway Plot" on page 140.

The Oneway Plot

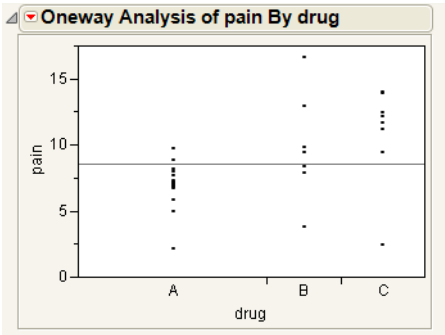
The Oneway plot shows the response points for each X factor value. You can compare the distribution of the response across the levels of the X factor. The distinct values of X are sometimes called levels.

Replace variables in the plot in one of two ways: swap existing variables by dragging and dropping a variable from one axis to the other axis; or, click on a variable in the Columns panel of the associated data table and drag it onto an axis.

You can add reports, additional plots, and tests to the report window using the options in the red triangle menu for Oneway Analysis. See "Oneway Platform Options" on page 141.

To produce the plot shown in Figure 6.5, follow the instructions in "Example of Oneway Analysis" on page 138.

Figure 6.5 The Oneway Plot



Oneway Platform Options

Note: The Fit Group menu appears if you have specified multiple Y or X variables. Menu options allow you to arrange reports or order them by RSquare. See the Standard Least Squares chapter in the *Fitting Linear Models* book for more information.

When you select a platform option, objects might be added to the plot, and a report is added to the report window.

Table 6.1 Examples of Options and Elements

Platform Option	Object Added to Plot	Report Added to Report Window
Quantiles	Box plots	Quantiles report
Means/Anova	Mean diamonds	Oneway ANOVA reports
Means and Std Dev	Mean lines, error bars, and standard deviation lines	Means and Std Deviations report

Table 6.1 Examples of Options and Elements *(Continued)*

Platform Option	Object Added to Plot	Report Added to Report Window
Compare Means	Comparison circles (except Nonparametric Multiple Comparisons option)	Means Comparison reports

The red triangle menu for Oneway Analysis provides the following options. Some options might not appear unless specific conditions are met.

Quantiles Lists the following quantiles for each group:

- 0% (Minimum)
- 10%
- 25%
- 50% (Median)
- 75%
- 90%
- 100% (Maximum)

Activates **Box Plots** from the **Display Options** menu. See "Quantiles" on page 146.

Means/Anova Fits means for each group and performs a one-way analysis of variance to test if there are differences among the means. Also gives the Welch test, which is an ANOVA test for comparing means when the variances within groups are not equal. See "Means/Anova and Means/Anova/Pooled t" on page 147.

If the X factor has two levels, the menu option changes to **Means/Anova/Pooled t**.

Means and Std Dev Gives summary statistics for each group. The standard errors for the means use individual group standard deviations rather than the pooled estimate of the standard deviation.

The plot now contains mean lines, error bars, and standard deviation lines. For a brief description of these elements, see "Display Options" on page 144. For more details about these elements, see "Mean Lines, Error Bars, and Standard Deviation Lines" on page 151.

t test Produces a *t*-test report assuming that the variances are not equal. See "The t-test Report" on page 148.

This option appears only if the X factor has two levels.

Analysis of Means Methods Provides five commands for performing Analysis of Means (ANOM) procedures. There are commands for comparing means, variances, and ranges. See "Analysis of Means Methods" on page 152.

Compare Means Provides multiple-comparison methods for comparing sets of group means. See "Compare Means" on page 155.

Nonparametric Provides nonparametric comparisons of group locations. See "Nonparametric Tests" on page 161.

Unequal Variances Performs four tests for equality of group variances. See "Unequal Variances" on page 167.

Equivalence Test Tests that a difference is less than a threshold value. See "Equivalence Test" on page 170.

Robust Provides two methods for reducing the influence of outliers on your data. See "Robust" on page 170.

Power Provides calculations of statistical power and other details about a given hypothesis test. See "Power" on page 171.

The Power Details window and reports also appear within the Fit Model platform. For further discussion and examples of power calculations, see the Statistical Details appendix in the *Fitting Linear Models* book.

Set α Level You can select an option from the most common alpha levels or specify any level with the **Other** selection. Changing the alpha level results in the following actions:

- recalculates confidence limits
- adjusts the mean diamonds on the plot (if they are showing)
- modifies the upper and lower confidence level values in reports
- changes the critical number and comparison circles for all Compare Means reports
- changes the critical number for all Nonparametric Multiple Comparison reports

Normal Quantile Plot Provides the following options for plotting the quantiles of the data in each group:

- **Plot Actual by Quantile** generates a quantile plot with the response variable on the y-axis and quantiles on the x-axis. The plot shows quantiles computed within each level of the categorical X factor.
- **Plot Quantile by Actual** reverses the x- and y-axes.
- **Line of Fit** draws straight diagonal reference lines on the plot for each level of the X variable. This option is available only once you have created a plot (Actual by Quantile or Quantile by Actual).

CDF Plot Plots the cumulative distribution function for all of the groups in the Oneway report. See "CDF Plot" on page 173.

Densities Compares densities across groups. See "Densities" on page 173.

Matching Column Specify a matching variable to perform a matching model analysis. Use this option when the data in your Oneway analysis comes from matched (paired) data, such as when observations in different groups come from the same subject.

The plot now contains matching lines that connect the matching points. See "Matching Column" on page 173.

Save Saves the following quantities as new columns in the current data table:

- **Save Residuals** saves values computed as the response variable minus the mean of the response variable within each level of the factor variable.
- **Save Standardized** saves standardized values of the response variable computed within each level of the factor variable. This is the centered response divided by the standard deviation within each level.
- **Save Normal Quantiles** saves normal quantile values computed within each level of the categorical factor variable.
- **Save Predicted** saves the predicted mean of the response variable for each level of the factor variable.

Display Options Adds or removes elements from the plot. See "Display Options" on page 144.

See the JMP Reports chapter in the *Using JMP* book for more information about the following options:

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Display Options

Using Display Options, you can add or remove elements from a plot. Some options might not appear unless they are relevant.

All Graphs Shows or hides all graphs.

Points Shows or hides data points on the plot.

Box Plots Shows or hides outlier box plots for each group. For an example, see "Conduct the Oneway Analysis" on page 200.

Mean Diamonds Draws a horizontal line through the mean of each group proportional to its *x*-axis. The top and bottom points of the mean diamond show the upper and lower 95% confidence points for each group. See "Mean Diamonds and X-Axis Proportional" on page 150.

Mean Lines Draws a line at the mean of each group. See "Mean Lines, Error Bars, and Standard Deviation Lines" on page 151.

Mean CI Lines Draws lines at the upper and lower 95% confidence levels for each group.

Mean Error Bars Identifies the mean of each group and shows error bars one standard error above and below the mean. See "Mean Lines, Error Bars, and Standard Deviation Lines" on page 151.

Grand Mean Draws the overall mean of the *Y* variable on the plot.

Std Dev Lines Shows lines one standard deviation above and below the mean of each group. See "Mean Lines, Error Bars, and Standard Deviation Lines" on page 151.

Comparison Circles Shows or hides comparison circles. This option is available only when one of the **Compare Means** options is selected. See "Comparison Circles" on page 202. For an example, see "Conduct the Oneway Analysis" on page 200.

Connect Means Connects the group means with a straight line.

Mean of Means Draws a line at the mean of the group means.

X-Axis proportional Makes the spacing on the *x*-axis proportional to the sample size of each level. See "Mean Diamonds and X-Axis Proportional" on page 150.

Points Spread Spreads points over the width of the interval

Points Jittered Adds small spaces between points that overlay on the same *y* value. The horizontal adjustment of points varies from 0.375 to 0.625 with a $4*(\text{Uniform-}0.5)^5$ distribution.

Matching Lines (Only appears when the **Matching Column** option is selected.) Connects matching points.

Matching Dotted Lines (Only appears when the **Matching Column** option is selected.) Draws dotted lines to connect cell means from missing cells in the table. The values used as the endpoints of the lines are obtained using a two-way ANOVA model.

Histograms Draws side-by-side histograms to the right of the original plot.

Robust Mean Lines (Appears only when a Robust option is selected.) Draws a line at the robust mean of each group.

Legend Displays a legend for the Normal Quantile Plot, CDF Plot, and Densities options.

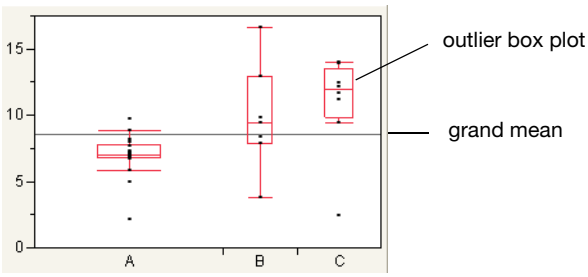
Quantiles

The Quantiles report lists selected percentiles for each level of the X factor variable. The median is the 50th percentile, and the 25th and 75th percentiles are called the *quartiles*.

The **Quantiles** option adds the following elements to the plot:

- the grand mean representing the overall mean of the Y variable
- outlier box plots summarizing the distribution of points at each factor level

Figure 6.6 Outlier Box Plot and Grand Mean



Note: To hide these elements, click the red triangle next to Oneway Analysis and select **Display Options > Box Plots** or **Grand Mean**.

Outlier Box Plots

The outlier box plot is a graphical summary of the distribution of data. Note the following aspects about outlier box plots (see Figure 6.7):

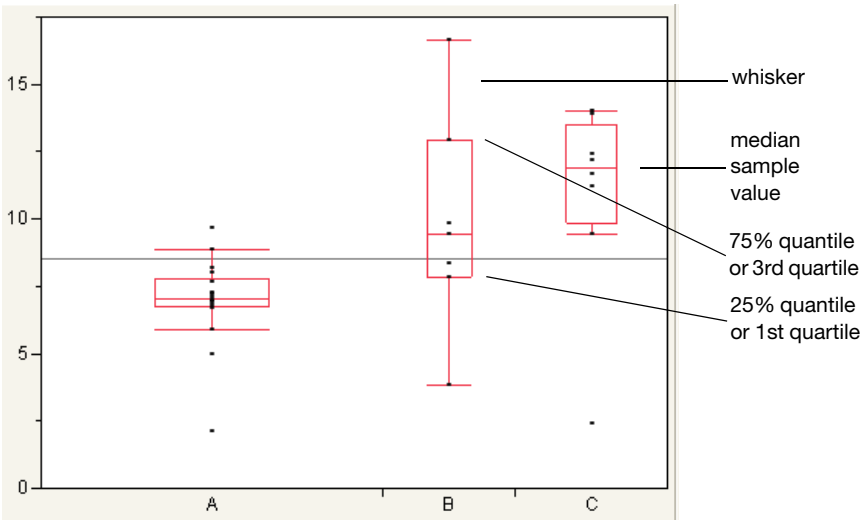
- The horizontal line within the box represents the median sample value.
- The ends of the box represent the 75th and 25th quantiles, also expressed as the 3rd and 1st quartile, respectively.
- The difference between the 1st and 3rd quartiles is called the *interquartile range*.
- Each box has lines, sometimes called *whiskers*, that extend from each end. The whiskers extend from the ends of the box to the outermost data point that falls within the distances computed as follows:

  ```
  3rd quartile + 1.5*(interquartile range)
  ```

  ```
  1st quartile - 1.5*(interquartile range)
  ```

  If the data points do not reach the computed ranges, then the whiskers are determined by the upper and lower data point values (not including outliers).

Figure 6.7 Examples of Outlier Box Plots



Means/Anova and Means/Anova/Pooled t

The **Means/Anova** option performs an analysis of variance. If the X factor contains exactly two levels, this option appears as **Means/Anova/Pooled t**. In addition to the other reports, a *t*-test report assuming pooled (or equal) variances appears.

Mean diamonds are added to the Oneway plot See "Display Options" on page 144 and "Mean Diamonds and X-Axis Proportional" on page 150.

Reports See "The Summary of Fit Report" on page 147, "The Analysis of Variance Report" on page 149, "The Means for Oneway Anova Report" on page 150., "The t-test Report" on page 148, and "The Block Means Report" on page 150.

- The t-test report appears only if the **Means/Anova/Pooled t** option is selected.
- The Block Means report appears only if you have specified a Block variable in the launch window.

The Summary of Fit Report

The Summary of Fit report shows a summary for a one-way analysis of variance.

Rsquare Measures the proportion of the variation accounted for by fitting means to each factor level. The remaining variation is attributed to random error. The R^2 value is 1 if fitting the group means account for all the variation with no error. An R^2 of 0 indicates that

the fit serves no better as a prediction model than the overall response mean. For more information, see "Summary of Fit Report" on page 204.

R^2 is also called the *coefficient of determination*.

Note: A low RSquare value suggests that there may be variables not in the model that account for the unexplained variation. However, if your data are subject to a large amount of inherent variation, even a useful ANOVA model may have a low RSquare value. Read the literature in your research area to learn about typical RSquare values.

Adj Rsquare Adjusts R^2 to make it more comparable over models with different numbers of parameters by using the degrees of freedom in its computation. For more information, see "Summary of Fit Report" on page 204.

Root Mean Square Error Estimates the standard deviation of the random error. It is the square root of the mean square for Error found in the Analysis of Variance report.

Mean of Response Overall mean (arithmetic average) of the response variable.

Observations (or Sum Wgts) Number of observations used in estimating the fit. If weights are used, this is the sum of the weights. See "Summary of Fit Report" on page 204.

The t-test Report

Note: This option is applicable only for the **Means/Anova/Pooled t** option.

There are two types of *t*-Tests:

- Equal variances. If you select the **Means/Anova/Pooled t** option, a *t*-Test report appears. This *t*-Test assumes equal variances.
- Unequal variances. If you select the **t-Test** option from the red triangle menu, a *t*-Test report appears. This *t*-Test assumes unequal variances.

The t-test report contains the following columns:

t Test plot Shows the sampling distribution of the difference in the means, assuming the null hypothesis is true. The vertical red line is the actual difference in the means. The shaded areas correspond to the p-values.

Difference Shows the estimated difference between the two X levels. In the plots, the Difference value appears as a red line that compares the two levels.

Std Err Dif Shows the standard error of the difference.

Upper CL Dif Shows the upper confidence limit for the difference.

Lower CL Dif Shows the lower confidence limit for the difference.

Confidence Shows the level of confidence (1-alpha). To change the level of confidence, select a new alpha level from the **Set** α **Level** command from the platform red triangle menu.

t Ratio Value of the *t*-statistic.

DF The degrees of freedom used in the *t*-test.

Prob > |t| The *p*-value associated with a two-tailed test.

Prob > t The *p*-value associated with a lower-tailed test.

Prob < t The *p*-value associated with an upper-tailed test.

The Analysis of Variance Report

The Analysis of Variance report partitions the total variation of a sample into two components. The ratio of the two mean squares forms the *F* ratio. If the probability associated with the *F* ratio is small, then the model is a better fit statistically than the overall response mean.

Note: If you specified a **Block** column, then the Analysis of Variance report includes the **Block** variable.

Source Lists the three sources of variation, which are the model source, **Error**, and **C. Total** (corrected total).

DF Records an associated degrees of freedom (DF for short) for each source of variation:

- The degrees of freedom for **C. Total** are N - 1, where N is the total number of observations used in the analysis.
- If the X factor has k levels, then the model has k - 1 degrees of freedom.

The **Error** degrees of freedom is the difference between the **C. Total** degrees of freedom and the **Model** degrees of freedom (in other words, N - k).

Sum of Squares Records a sum of squares (SS for short) for each source of variation:

- The total (**C. Total**) sum of squares of each response from the overall response mean. The **C. Total** sum of squares is the base model used for comparison with all other models.
- The sum of squared distances from each point to its respective group mean. This is the remaining unexplained **Error** (residual) SS after fitting the analysis of variance model.

The total SS minus the error SS gives the sum of squares attributed to the model. This tells you how much of the total variation is explained by the model.

Mean Square Is a sum of squares divided by its associated degrees of freedom:

- The **Model** mean square estimates the variance of the error, but only under the hypothesis that the group means are equal.

– The **Error** mean square estimates the variance of the error term independently of the model mean square and is unconditioned by any model hypothesis.

F Ratio Model mean square divided by the error mean square. If the hypothesis that the group means are equal (there is no real difference between them) is true, then both the mean square for error and the mean square for model estimate the error variance. Their ratio has an *F* distribution. If the analysis of variance model results in a significant reduction of variation from the total, the *F* ratio is higher than expected.

Prob>F Probability of obtaining (by chance alone) an *F* value greater than the one calculated if, in reality, there is no difference in the population group means. Observed significance probabilities of 0.05 or less are often considered evidence that there are differences in the group means.

The Means for Oneway Anova Report

The Means for Oneway Anova report summarizes response information for each level of the nominal or ordinal factor.

Level Lists the levels of the X variable.

Number Lists the number of observations in each group.

Mean Lists the mean of each group.

Std Error Lists the estimates of the standard deviations for the group means. This standard error is estimated assuming that the variance of the response is the same in each level. It is the root mean square error found in the Summary of Fit report divided by the square root of the number of values used to compute the group mean.

Lower 95% and Upper 95% Lists the lower and upper 95% confidence interval for the group means.

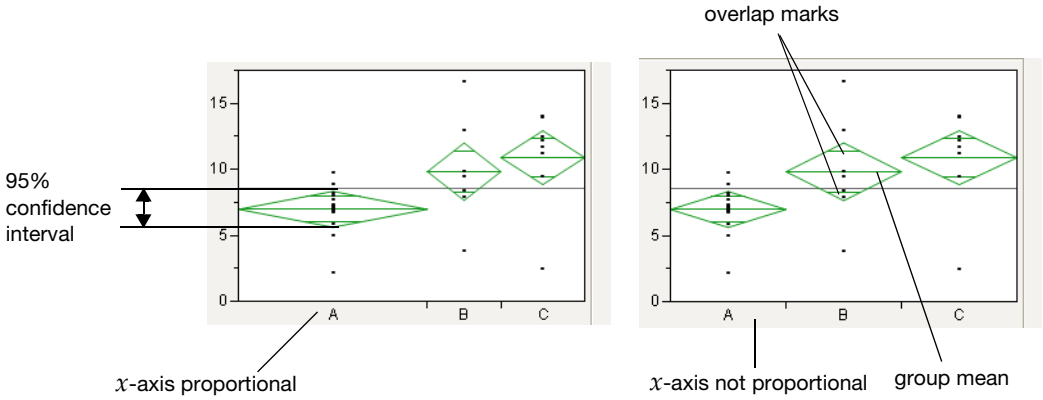
The Block Means Report

If you have specified a Block variable on the launch window, the **Means/Anova** and **Means/Anova/Pooled t** commands produce a Block Means report. This report shows the means for each block and the number of observations in each block.

Mean Diamonds and X-Axis Proportional

A mean diamond illustrates a sample mean and confidence interval.

Figure 6.8 Examples of Mean Diamonds and X-Axis Proportional Options



Note the following observations:

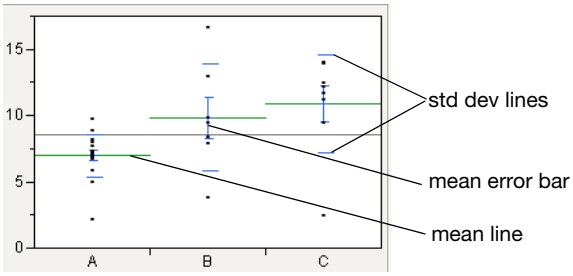
- The top and bottom of each diamond represent the (1-alpha)x100 confidence interval for each group. The confidence interval computation assumes that the variances are equal across observations. Therefore, the height of the diamond is proportional to the reciprocal of the square root of the number of observations in the group.
- If the **X-Axis proportional** option is selected, the horizontal extent of each group along the x-axis (the horizontal size of the diamond) is proportional to the sample size for each level of the X variable. Therefore, the narrower diamonds are usually taller, because fewer data points results in a wider confidence interval.
- The mean line across the middle of each diamond represents the group mean.
- Overlap marks appear as lines above and below the group mean. For groups with equal sample sizes, overlapping marks indicate that the two group means are not significantly different at the given confidence level. Overlap marks are computed as group mean $\pm\ (\sqrt{2})/2 \times \mathrm{CI}/2$. Overlap marks in one diamond that are closer to the mean of another diamond than that diamond's overlap marks indicate that those two groups are not different at the given confidence level.
- The mean diamonds automatically appear when you select the **Means/Anova/Pooled t** or **Means/Anova** option from the platform menu. However, you can show or hide them at any time by selecting **Display Options > Mean Diamonds** from the red triangle menu.

Mean Lines, Error Bars, and Standard Deviation Lines

Show mean lines by selecting **Display Options > Mean Lines**. Mean lines indicate the mean of the response for each level of the X variable.

Mean error bars and standard deviation lines appear when you select the **Means and Std Dev** option from the red triangle menu. See Figure 6.9. To turn each option on or off singly, select **Display Options > Mean Error Bars** or **Std Dev Lines**.

Figure 6.9 Mean Lines, Mean Error Bars, and Std Dev Lines



Analysis of Means Methods

Analysis of means (ANOM) methods compare means and variances and other measures of location and scale across several groups. You might want to use these methods under these circumstances:

- to test whether any of the group means are statistically different from the overall (sample) mean
- to test whether any of the group standard deviations are statistically different from the root mean square error (RMSE)
- to test whether any of the group ranges are statistically different from the overall mean of the ranges

Note: Within the Contingency platform, you can use the **Analysis of Means for Proportions** when the response has two categories. For details, see the "Contingency Analysis" chapter on page 209.

For a description of ANOM methods and to see how JMP implements ANOM, see the book by Nelson et al. (2005).

Analysis of Means for Location

You can test whether groups have a common mean or center value using the following options:

- ANOM
- ANOM with Transformed Ranks

ANOM

Use ANOM to compare group means to the overall mean. This method assumes that your data are approximately normally distributed. See "Example of an Analysis of Means Chart" on page 174.

ANOM with Transformed Ranks

This is the nonparametric version of the **ANOM** analysis. Use this method if your data is clearly non-normal and cannot be transformed to normality. ANOM with Transformed Ranks compares each group's mean transformed rank to the overall mean transformed rank. The ANOM test involves applying the usual ANOM procedure and critical values to the transformed observations.

Transformed Ranks

Suppose that there are n observations. The transformed observations are computed as follows:

- Rank all observations from smallest to largest, accounting for ties. For tied observations, assign each one the average of the block of ranks that they share.
- Denote the ranks by $R_1, R_2, ..., R_n$.
- The transformed rank corresponding to the ith observations is:

$$\text{Transformed } R_i = Normal\ Quantile\left[\left(\frac{R_i}{2n+1}\right) + 0.5\right]$$

The ANOM procedure is applied to the values Transformed R_i. Since the ranks have a uniform distribution, the transformed ranks have a folded normal distribution. For details, see Nelson et al. (2005).

Analysis of Means for Scale

You can test for homogeneity of variation within groups using the following options:

- ANOM for Variances
- ANOM for Variances with Levene (ADM)
- ANOM for Ranges

ANOM for Variances

Use this method to compare group standard deviations (or variances) to the root mean square error (or mean square error). This method assumes that your data is approximately normally distributed. To use this method, each group must have at least four observations. For details about the ANOM for Variances test, see Wludyka and Nelson (1997) and Nelson et al. (2005). For an example, see "Example of an Analysis of Means for Variances Chart" on page 175.

ANOM for Variances with Levene (ADM)

This method provides a robust test that compares the group means of the *absolute deviations from the median* (ADM) to the overall mean ADM. Use ANOM for Variances with Levene (ADM) if you suspect that your data is non-normal and cannot be transformed to normality. ANOM for Variances with Levene (ADM) is a nonparametric analog of the ANOM for Variances analysis. For details about the ANOM for Variances with Levene (ADM) test, see Levene (1960) or Brown and Forsythe (1974).

ANOM for Ranges

Use this test to compare group ranges to the mean of the group ranges. This is a test for scale differences based on the range as the measure of spread. For details, see Wheeler (2003).

Note: ANOM for Ranges is available only for balanced designs and specific group sizes. See "Restrictions for ANOM for Ranges Test" on page 154.

Restrictions for ANOM for Ranges Test

Unlike the other ANOM decision limits, the decision limits for the ANOM for Ranges chart uses only tabled critical values. For this reason, ANOM for Ranges is available only for the following:

- groups of equal sizes
- groups specifically of the following sizes: 2–10, 12, 15, and 20
- number of groups between 2 and 30
- alpha levels of 0.10, 0.05, and 0.01

Analysis of Means Charts

Each Analysis of Means Methods option adds a chart to the report window that shows the following:

- an upper decision limit (UDL)
- a lower decision limit (LDL)
- a horizontal (center) line that falls between the decision limits and is positioned as follows:
  - ANOM: the overall mean
  - ANOM with Transformed Ranks: the overall mean of the transformed ranks
  - ANOM for Variances: the root mean square error (or MSE when in variance scale)
  - ANOM for Variances with Levene (ADM): the overall absolute deviation from the mean
  - ANOM for Ranges: the mean of the group ranges

If a group's plotted statistic falls outside of the decision limits, then the test indicates that there is a statistical difference between that group's statistic and the overall average of the statistic for all the groups.

Analysis of Means Options

Each Analysis of Means Methods option adds an Analysis of Means red triangle menu to the report window.

Set Alpha Level Select an option from the most common alpha levels or specify any level with the **Other** selection. Changing the alpha level modifies the upper and lower decision limits.

Note: For ANOM for Ranges, only the selections 0.10, 0.05, and 0.01 are available.

Show Summary Report The reports are based on the Analysis of Means method:

- For ANOM, creates a report showing group means and decision limits.
- For ANOM with Transformed Ranks, creates a report showing group mean transformed ranks and decision limits.
- For ANOM for Variances, creates a report showing group standard deviations (or variances) and decision limits.
- For ANOM for Variances with Levene (ADM), creates a report showing group mean ADMs and decision limits.
- For ANOM for Ranges, creates a report showing group ranges and decision limits.

Graph in Variance Scale (Only for **ANOM for Variances**) Changes the scale of the *y*-axis from standard deviations to variances.

Display Options Display options include the following:

- **Show Decision Limits** shows or hides decision limit lines.
- **Show Decision Limit Shading** shows or hides decision limit shading.
- **Show Center Line** shows or hides the center line statistic.
- **Point Options: Show Needles** shows the needles. This is the default option. **Show Connected Points** shows a line connecting the means for each group. **Show Only Points** shows only the points representing the means for each group.

Compare Means

Note: Another method for comparing means is ANOM. See "Analysis of Means Methods" on page 152.

Use the Compare Means options to perform multiple comparisons of group means. All of these methods use pooled variance estimates for the means. Each Compare Means option adds comparison circles next to the plot and specific reports to the report window. For details about comparison circles, see "Using Comparison Circles" on page 157.

Option	Description	Nonparametric Menu Option
Each Pair, Student's t	Computes individual pairwise comparisons using Student's *t*-tests. If you make many pairwise tests, there is no protection across the inferences. Therefore, the alpha-size (Type I error rate) across the hypothesis tests is higher than that for individual tests. See "Each Pair, Student's t" on page 158.	**Nonparametric > Nonparametric Multiple Comparisons > Wilcoxon Each Pair**
All Pairs, Tukey HSD	Shows a test that is sized for all differences among the means. This is the *Tukey* or *Tukey-Kramer* HSD (honestly significant difference) test. (Tukey 1953, Kramer 1956). This test is an exact alpha-level test if the sample sizes are the same, and conservative if the sample sizes are different (Hayter 1984). See "All Pairs, Tukey HSD" on page 158.	**Nonparametric > Nonparametric Multiple Comparisons > Steel-Dwass All Pairs**
With Best, Hsu MCB	Tests whether the means are less than the unknown maximum or greater than the unknown minimum. This is the Hsu MCB test (Hsu, 1996 and Hsu, 1981). See "With Best, Hsu MCB" on page 159.	none
With Control, Dunnett's	Tests whether the means are different from the mean of a control group. This is Dunnett's test (Dunnett 1955). See "With Control, Dunnett's" on page 160.	**Nonparametric > Nonparametric Multiple Comparisons > Steel With Control**

Note: If you have specified a **Block** column, then the multiple comparison methods are performed on data that has been adjusted for the Block means.

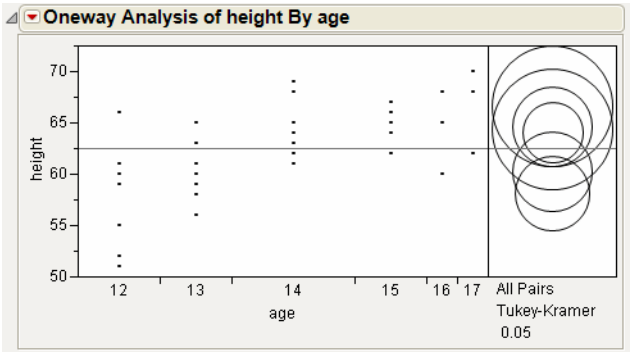
Related Information

- "Example Contrasting All of the Compare Means Tests" on page 182

Using Comparison Circles

Each multiple comparison test begins with a *comparison circles* plot, which is a visual representation of group mean comparisons. Figure 6.10 shows the comparison circles for the All Pairs, Tukey HSD method. Other comparison tests lengthen or shorten the radii of the circles.

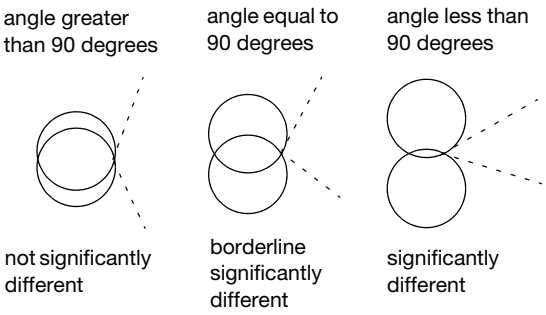
Figure 6.10 Visual Comparison of Group Means



Compare each pair of group means visually by examining the intersection of the comparison circles. The outside angle of intersection tells you whether the group means are significantly different. See Figure 6.11.

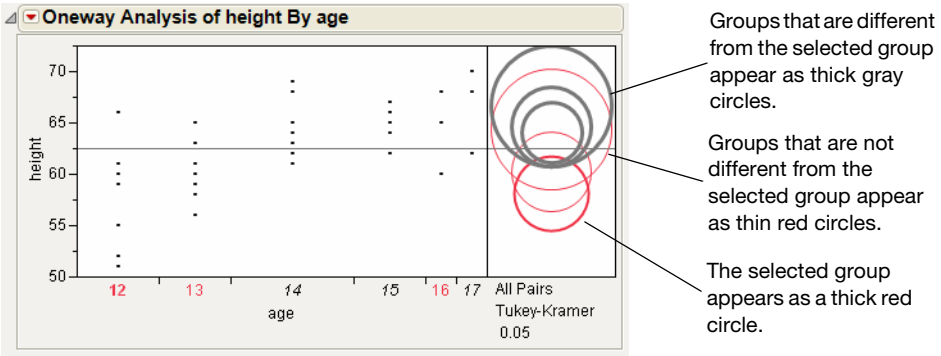
- Circles for means that are significantly different either do not intersect, or intersect slightly, so that the outside angle of intersection is less than 90 degrees.
- If the circles intersect by an angle of more than 90 degrees, or if they are nested, the means are not significantly different.

Figure 6.11 Angles of Intersection and Significance



If the intersection angle is close to 90 degrees, you can verify whether the means are significantly different by clicking on the comparison circle to select it. See Figure 6.12. To deselect circles, click in the white space outside the circles.

Figure 6.12 Highlighting Comparison Circles



Related Information

- "Comparison Circles" on page 202

Each Pair, Student's t

The **Each Pair, Student's t** test shows the Student's *t*-test for each pair of group levels and tests only individual comparisons.

Related Information

- "Example of the Each Pair, Student's t Test" on page 176

All Pairs, Tukey HSD

The **All Pairs, Tukey HSD** test (also called Tukey-Kramer) protects the significance tests of all combinations of pairs, and the HSD intervals become greater than the Student's *t* pairwise LSDs. Graphically, the comparison circles become larger and differences are less significant.

The q statistic is calculated as follows: q* = (1/sqrt(2)) * q where q is the required percentile of the studentized range distribution. For more details, see the description of the T statistic by Neter, Wasserman, and Kutner (1990).

Related Information

- "Example of the All Pairs, Tukey HSD Test" on page 178

With Best, Hsu MCB

The **With Best, Hsu MCB** test determines whether the mean for a given level exceeds the maximum mean of the remaining levels, or is smaller than the minimum mean of the remaining levels. See Hsu, 1996.

The quantiles for the Hsu MCB test vary by the level of the categorical variable. Unless the sample sizes are equal across levels, the comparison circle technique is not exact. The radius of a comparison circle is given by the standard error of the level multiplied by the largest quantile value. Use the p-values of the tests to obtain precise assessments of significant differences. See "Comparison with Max and Min" on page 159.

Note: Means that are not regarded as the maximum or the minimum by MCB are also the means that are not contained in the selected subset of Gupta (1965) of potential maximums or minimum means.

Confidence Quantile

This report gives the quantiles for each level of the categorical variable. These correspond to the specified value of Alpha.

Comparison with Max and Min

The report shows *p*-values for one-sided Dunnett tests. For each level other than the best, the *p*-value given is for a test that compares the mean of the sample best level to the mean of each remaining level treated as a control (potentially best) level. The *p*-value for the sample best level is obtained by comparing the mean of the second sample best level to the mean of the sample best level treated as a control.

The report shows three columns.

Level The level of the categorical variable.

with Max p-Value For each level of the categorical variable, this column gives a *p*-value for a test that the mean of that level exceeds the maximum mean of the remaining levels. Use the tests in this column to screen out levels whose means are significantly smaller than the (unknown) largest true mean.

with Min p-Value For each level of the categorical variable, this column gives a *p*-value for a test that the mean of that level is smaller than the minimum mean of the remaining levels. Use the tests in this column to screen out levels whose means are significantly greater than the (unknown) smallest true mean.

LSD Threshold Matrix

The first report shown is for the maximum and the second is for the minimum.

For the *maximum* report, a column shows the row mean minus the column mean minus the LSD. If a value is positive, the row mean is significantly higher than the mean for the column, and the mean for the column is not the maximum.

For the *minimum* report, a column shows the row mean minus the column mean plus the LSD. If a value is negative, the row mean is significantly less than the mean for the column, and the mean for the column is not the minimum.

Related Information

- "Example of the With Best, Hsu MCB Test" on page 180

With Control, Dunnett's

The **With Control, Dunnett's** test compares a set of means against the mean of a control group. The LSDs that it produces are between the Student's *t* and Tukey-Kramer LSDs, because they are sized to refrain from an intermediate number of comparisons.

In the Dunnett's report, the $|d|$ quantile appears, and can be used in a manner similar to a Student's *t*-statistic. The LSD threshold matrix shows the absolute value of the difference minus the LSD. If a value is positive, its mean is more than the LSD apart from the control group mean and is therefore significantly different.

Related Information

- "Example of the With Control, Dunnett's Test" on page 181

Compare Means Options

The Means Comparisons reports for all four tests contain a red triangle menu with customization options.

Difference Matrix Shows a table of all differences of means.

Confidence Quantile Shows the *t*-value or other corresponding quantiles used for confidence intervals.

LSD Threshold Matrix Shows a matrix showing if a difference exceeds the least significant difference for all comparisons.

Connecting Letters Report Shows the traditional letter-coded report where means that are not sharing a letter are significantly different.

> **Note:** Not available for With Best, Hsu MCB and With Control, Dunnett's.

Ordered Differences Report Shows all the positive-side differences with their confidence interconfidence interval band overlaid on the plot. Confidence intervals that do not fully contain their corresponding bar are significantly different from each other.

Note: Not available for With Best, Hsu MCB and With Control, Dunnett's.

Detailed Comparisons Report Shows a detailed report for each comparison. Each section shows the difference between the levels, standard error and confidence intervals, *t*-ratios, *p*-values, and degrees of freedom. A plot illustrating the comparison appears on the right of each report.

Note: Not available for All Pairs, Tukey HSD, With Best, Hsu MCB, and With Control, Dunnett's.

Nonparametric Tests

Nonparametric tests are useful when the usual analysis of variance assumption of normality is not viable. The Nonparametric option provides several methods for testing the hypothesis of equal means or medians across groups. Nonparametric multiple comparison procedures are also available to control the overall error rate for pairwise comparisons. Nonparametric tests use functions of the response ranks, called rank scores. See Hajek (1969) and SAS Institute (2008).

Note: If you specify a Block column, the nonparametric tests are conducted on data values that are centered using the block means.

Wilcoxon Test Performs a test based on Wilcoxon rank scores. The Wilcoxon rank scores are the simple ranks of the data. The Wilcoxon test is the most powerful rank test for errors with logistic distributions. If the factor has more than two levels, the Kruskal-Wallis test is performed. For information about the report, see "The Wilcoxon, Median, and Van der Waerden Test Reports" on page 162. For an example, see "Example of the Nonparametric Wilcoxon Test" on page 183.

The Wilcoxon test is also called the Mann-Whitney test.

Median Test Performs a test based on Median rank scores. The Median rank scores are either 1 or 0, depending on whether a rank is above or below the median rank. The Median test is the most powerful rank test for errors with double-exponential distributions. For information about the report, see "The Wilcoxon, Median, and Van der Waerden Test Reports" on page 162.

van der Waerden Test Performs a test based on Van der Waerden rank scores. The Van der Waerden rank scores are the ranks of the data divided by one plus the number of observations transformed to a normal score by applying the inverse of the normal distribution function. The Van der Waerden test is the most powerful rank test for errors with normal distributions. For information about the report, see "The Wilcoxon, Median, and Van der Waerden Test Reports" on page 162.

Kolmogorov Smirnov Test Performs a test based on the empirical distribution function, which tests whether the distribution of the response is the same across the groups. Both an approximate and an exact test are given. This test is available only when the X factor has two levels. For information about the report, see "Kolmogorov-Smirnov Two-Sample Test Report" on page 163.

JMP PRO **Exact Test** Provides options for performing exact versions of the Wilcoxon, Median, van der Waerden, and Kolmogorov-Smirnov tests. These options are available only when the X factor has two levels. Results for both the approximate and the exact test are given.

For information about the report, see " 2-Sample, Exact Test" on page 163. For an example involving the Wilcoxon Exact Test, see "Example of the Nonparametric Wilcoxon Test" on page 183.

The Wilcoxon, Median, and Van der Waerden Test Reports

For each test, the report shows the descriptive statistics followed by the test results. Test results appear in the 1-Way Test, ChiSquare Approximation report and, if the X variable has exactly two levels, a 2-Sample Test, Normal Approximation report also appears. The descriptive statistics are the following:

Level The levels of X.

Count The frequencies of each level.

Score Sum The sum of the rank score for each level.

Expected Score The expected score under the null hypothesis that there is no difference among class levels.

Score Mean The mean rank score for each level.

(Mean-Mean0)/Std0 The standardized score. Mean0 is the mean score expected under the null hypothesis. Std0 is the standard deviation of the score sum expected under the null hypothesis. The null hypothesis is that the group means or medians are in the same location across groups.

2-Sample Test, Normal Approximation

When you have exactly two levels of X, a 2-Sample Test, Normal Approximation report appears. This report gives the following:

S Gives the sum of the rank scores for the level with the smaller number of observations.

Z Gives the test statistic for the normal approximation test. For details, see "Two-Sample Normal Approximations" on page 207.

Prob>|Z| Gives the *p*-value, based on a standard normal distribution, for the normal approximation test.

1-Way Test, ChiSquare Approximation

This report gives results for a chi-square test for location. For details, see Conover (1999).

ChiSquare Gives the values of the chi-square test statistic. For details, see "One-Way ChiSquare Approximations" on page 207.

DF Gives the degrees of freedom for the test.

Prob>ChiSq Gives the *p*-value for the test. The *p*-value is based on a ChiSquare distribution with degrees of freedom equal to the number of levels of X minus 1.

JMP PRO 2-Sample, Exact Test

If your data are sparse, skewed, or heavily tied, exact tests might be more suitable than approximations based on asymptotic behavior. When you have exactly two levels of X, JMP computes test statistics for exact tests. Select **Nonparametric > Exact Test** and select the test of your choice. A 2-Sample: Exact Test report appears. This report gives the following:

S Gives the sum of the rank scores for the observations in the smaller group. If the two levels of X have the same numbers of observations, then the value of S corresponds to the last level of X in the value ordering.

Prob£ S Gives a one-sided *p*-value for the test.

Prob ≥ |S-Mean| Gives a two-sided *p*-value for the test.

Kolmogorov-Smirnov Two-Sample Test Report

The Kolmogorov-Smirnov test is available only when X has exactly two levels. The report shows descriptive statistics followed by test results. The descriptive statistics are the following:

Level The two levels of X.

Count The frequencies of each level.

EDF at Maximum For a level of X, gives the value of the empirical cumulative distribution function (EDF) for that level at the value of X for which the difference between the two EDFs is a maximum. For the row named Total, gives the value of the pooled EDF (the EDF for the entire data set) at the value of X for which the difference between the two EDFs is a maximum.

Deviation from Mean at Maximum For each level, gives the value obtained as follows:

- Compute the difference between the EDF at Maximum for the given level and the EDF at maximum for the pooled data set (Total).
- Multiply this difference by the square root of the number of observations in that level, given as Count.

Kolmogorov-Smirnov Asymptotic Test

This report gives the details for the test.

KS A Kolmogorov-Smirnov statistic computed as follows:

$$\mathrm{KS} = \max_j \sqrt{\frac{1}{n}\sum_i n_i (F_i(x_j) - F(x_j))^2}$$

The formula uses the following notation:

- x_j, $j = 1, \ldots, n$ are the observations
- n_i is the number of observations in the ith level of X
- F is the pooled cumulative empirical distribution function
- F_i is the cumulative empirical distribution function for the ith level of X

This version of the Kolmogorov-Smirnov statistic applies even when there are more than two levels of X. Note, however, that JMP only performs the Kolmogorov-Smirnov analysis when X has only two levels of X.

KSa An asymptotic Kolmogorov-Smirnov statistic computed as $\mathrm{KS}\sqrt{n}$, where n is the total number of observations.

D=max|F1-F2| The maximum absolute deviation between the EDFs for the two levels. This is the version of the Kolmogorov-Smirnov statistic typically used to compare two samples.

Prob > D The p-value for the test. This is the probability that D exceeds the computed value under the null hypothesis of no difference between the levels.

D+ = max(F1-F2) A one-sided test statistic for the alternative hypothesis that the level of the first group exceeds the level of the second group.

Prob > D+ The p-value for the test of D+.

D- = max(F2-F1) A one-sided test statistic for the alternative hypothesis that the level of the second group exceeds the level of the first group

Prob > D- The p-value for the test of D-.

JMP PRO Kolmogorov-Smirnov Exact Test

For the Kolmogorov-Smirnov exact test, the report gives the same statistics as does the asymptotic test, but the p-values are computed to be exact.

Nonparametric Multiple Comparisons

This option provides several methods for performing nonparametric multiple comparisons. These tests are based on ranks and, except for the Wilcoxon Each Pair test, control for the

overall experimentwise error rate. For details about these tests, see See Dunn (1964) and Hsu (1996). For information about the reports, see "Nonparametric Multiple Comparisons Procedures" on page 165.

For the Wilcoxon, Median, and Van der Waerden tests, if the X factor has more than two levels, a chi-square approximation to the one-way test is performed. If the X factor has two levels, a normal approximation to the two-sample test is performed, in addition to the chi-square approximation to the one-way test.

Tip: While Friedman's test for nonparametric repeated measures ANOVA is not directly supported in JMP, it can be performed as follows: Calculate the ranks within each block. Define this new column to have an ordinal modeling type. Enter the ranks as Y in the Fit Y by X platform. Enter one of the effects as a blocking variable. Obtain Cochran-Mantel-Haenszel statistics.

Nonparametric Multiple Comparisons Procedures

Wilcoxon Each Pair Performs the Wilcoxon test on each pair. This procedure does not control for the overall alpha level. This is the nonparametric version of the **Each Pair, Student's t** option found on the Compare Means menu. See "Wilcoxon Each Pair, Steel-Dwass All Pairs, and Steel with Control" on page 166.

Steel-Dwass All Pairs Performs the Steel-Dwass test on each pair. This is the nonparametric version of the **All Pairs, Tukey HSD** option found on the Compare Means menu. See "Wilcoxon Each Pair, Steel-Dwass All Pairs, and Steel with Control" on page 166.

Steel With Control Compares each level to a control level. This is the nonparametric version of the **With Control, Dunnett's** option found on the Compare Means menu. See "Wilcoxon Each Pair, Steel-Dwass All Pairs, and Steel with Control" on page 166.

Dunn With Control for Joint Ranks Compares each level to a control level, similar to the Steel With Control option. The Dunn method computes ranks for all the data, not just the pair being compared. The reported p-Value reflects a Bonferroni adjustment. It is the unadjusted *p*-value multiplied by the number of comparisons. If the adjusted *p*-value exceeds 1, it is reported as 1. See "Dunn All Pairs for Joint Ranks and Dunn with Control for Joint Ranks" on page 167.

Dunn All Pairs for Joint Ranks Performs a comparison of each pair, similar to the Steel-Dwass All Pairs option. The Dunn method computes ranks for all the data, not just the pair being compared. The reported *p*-value reflects a Bonferroni adjustment. It is the unadjusted *p*-value multiplied by the number of comparisons. If the adjusted *p*-value exceeds 1, it is reported as 1. See "Dunn All Pairs for Joint Ranks and Dunn with Control for Joint Ranks" on page 167.

Wilcoxon Each Pair, Steel-Dwass All Pairs, and Steel with Control

The reports for these multiple comparison procedures give test results and confidence intervals. For these tests, observations are ranked within the sample obtained by combining only the two levels used in a given comparison.

q* The quantile used in computing the confidence intervals.

Alpha The alpha level used in computing the confidence interval. You can change the confidence level by selecting the Set α Level option from the Oneway menu.

Level The first level of the X variable used in the pairwise comparison.

- Level The second level of the X variable used in the pairwise comparison.

Score Mean Difference The mean of the rank score of the observations in the first level (Level) minus the mean of the rank scores of the observations in the second level (-Level), where a continuity correction is applied.

Denote the number of observations in the first level by n_1 and the number in the second level by n_2. The observations are ranked within the sample consisting of these two levels. Tied ranks are averaged. Denote the sum of the ranks for the first level by ScoreSum_1 and for the second level by ScoreSum_2.

If the difference in mean scores is positive, then the Score Mean Difference is given as follows:

$$\text{Score Mean Difference} = (\text{ScoreSum}_1 - 0.5)/n_1 - (\text{ScoreSum}_2 + 0.5)/n_2$$

If the difference in mean scores is negative, then the Score Mean Difference is given as follows:

$$\text{Score Mean Difference} = (\text{ScoreSum}_1 + 0.5)/n_1 - (\text{ScoreSum}_2 - 0.5)/n_2$$

Std Error Dif The standard error of the Score Mean Difference.

Z The standardized test statistic, which has an asymptotic standard normal distribution under the null hypothesis of no difference in means.

p-Value The *p*-value for the asymptotic test based on Z.

Hodges-Lehmann The Hodges-Lehmann estimator of the location shift. All paired differences consisting of observations in the first level minus observations in the second level are constructed. The Hodges-Lehmann estimator is the median of these differences. The bar graph to the right of Upper CL shows the size of the Hodges-Lehmann estimate.

Lower CL The lower confidence limit for the Hodges-Lehmann statistic.

Note: Not computed if group sample sizes are large enough to cause memory issues.

Upper CL The upper confidence limit for the Hodges-Lehmann statistic.

Note: Not computed if group sample sizes are large enough to cause memory issues.

~Difference A bar graph showing the size of the Hodges-Lehmann estimate for each comparison.

Dunn All Pairs for Joint Ranks and Dunn with Control for Joint Ranks

These comp are based on the rank of an observation in the entire data set. For the Dunn with Control for Joint Ranks tests, you must select a control level.

Level The first level of the X variable used in the pairwise comparison.

- Level he second level of the X variable used in the pairwise comparison.

Score Mean Difference The mean of the rank score of the observations in the first level (Level) minus the mean of the rank scores of the observations in the second level (-Level), where a continuity correction is applied. The ranks are obtained by ranking the observations within the entire sample. Tied ranks are averaged. The continuity correction is described in "Score Mean Difference" on page 166.

Std Error Dif The standard error of the Score Mean Difference.

Z The standardized test statistic, which has an asymptotic standard normal distribution under the null hypothesis of no difference in means.

p-Value The *p*-value for the asymptotic test based on Z.

Unequal Variances

When the variances across groups are not equal, the usual analysis of variance assumptions are not satisfied and the ANOVA *F* test is not valid. JMP provides four tests for equality of group variances and an ANOVA that is valid when the group sample variances are unequal. The concept behind the first three tests of equal variances is to perform an analysis of variance on a new response variable constructed to measure the spread in each group. The fourth test is Bartlett's test, which is similar to the likelihood ratio test under normal distributions.

Note: Another method to test for unequal variances is ANOMV. See "Analysis of Means Methods" on page 152.

The following Tests for Equal Variances are available:

O'Brien Constructs a dependent variable so that the group means of the new variable equal the group sample variances of the original response. An ANOVA on the O'Brien variable is actually an ANOVA on the group sample variances (O'Brien 1979, Olejnik, and Algina 1987).

Brown-Forsythe Shows the F test from an ANOVA where the response is the absolute value of the difference of each observation and the group median (Brown and Forsythe 1974).

Levene Shows the F test from an ANOVA where the response is the absolute value of the difference of each observation and the group mean (Levene 1960). The spread is measured as $z_{ij} = |y_{ij} - \bar{y}_i|$ (as opposed to the SAS default $z_{ij}^2 = (y_{ij} - \bar{y}_i)^2$).

Bartlett Compares the weighted arithmetic average of the sample variances to the weighted geometric average of the sample variances. The geometric average is always less than or equal to the arithmetic average with equality holding only when all sample variances are equal. The more variation there is among the group variances, the more these two averages differ. A function of these two averages is created, which approximates a χ^2-distribution (or, in fact, an F distribution under a certain formulation). Large values correspond to large values of the arithmetic or geometric ratio, and therefore to widely varying group variances. Dividing the Bartlett Chi-square test statistic by the degrees of freedom gives the F value shown in the table. Bartlett's test is not very robust to violations of the normality assumption (Bartlett and Kendall 1946).

If there are only two groups tested, then a standard F test for unequal variances is also performed. The F test is the ratio of the larger to the smaller variance estimate. The p-value from the F distribution is doubled to make it a two-sided test.

Note: If you have specified a **Block** column, then the variance tests are performed on data after it has been adjusted for the Block means.

Tests That the Variances Are Equal Report

The Tests That the Variances Are Equal report shows the differences between group means to the grand mean and to the median, and gives a summary of testing procedures.

If the equal variances test reveals that the group variances are significantly different, use Welch's test instead of the regular ANOVA test. The Welch statistic is based on the usual ANOVA F test. However, the means are weighted by the reciprocal of the group mean variances (Welch 1951; Brown and Forsythe 1974b; Asiribo, Osebekwin, and Gurland 1990). If there are only two levels, the Welch ANOVA is equivalent to an unequal variance t-test.

Description of the Variances Are Equal Report

Level Lists the factor levels occurring in the data.

Count Records the frequencies of each level.

Std Dev Records the standard deviations of the response for each factor level. The standard deviations are equal to the means of the O'Brien variable. If a level occurs only once in the data, no standard deviation is calculated.

MeanAbsDif to Mean Records the mean absolute difference of the response and group mean. The mean absolute differences are equal to the group means of the Levene variable.

MeanAbsDif to Median Records the absolute difference of the response and group median. The mean absolute differences are equal to the group means of the Brown-Forsythe variable.

Test Lists the names of the tests performed.

F Ratio Records a calculated *F* statistic for each test. See "Tests That the Variances Are Equal Report" on page 168.

DFNum Records the degrees of freedom in the numerator for each test. If a factor has *k* levels, the numerator has *k* - 1 degrees of freedom. Levels occurring only once in the data are not used in calculating test statistics for O'Brien, Brown-Forsythe, or Levene. The numerator degrees of freedom in this situation is the number of levels used in calculations minus one.

DFDen Records the degrees of freedom used in the denominator for each test. For O'Brien, Brown-Forsythe, and Levene, a degree of freedom is subtracted for each factor level used in calculating the test statistic. If a factor has k levels, the denominator degrees of freedom is *n* - *k*.

p-Value Probability of obtaining, by chance alone, an *F* value larger than the one calculated if in reality the variances are equal across all levels.

Description of the Welch's Test Report

F Ratio Shows the F test statistic for the equal variance test. See "Tests That the Variances Are Equal Report" on page 168.

DFNum Records the degrees of freedom in the numerator of the test. If a factor has *k* levels, the numerator has *k* - 1 degrees of freedom. Levels occurring only once in the data are not used in calculating the Welch ANOVA. The numerator degrees of freedom in this situation is the number of levels used in calculations minus one.

DFDen Records the degrees of freedom in the denominator of the test. See "Tests That the Variances Are Equal Report" on page 168.

Prob>F Probability of obtaining, by chance alone, an *F* value larger than the one calculated if in reality the means are equal across all levels. Observed significance probabilities of 0.05 or less are considered evidence of unequal means across the levels.

t Test Shows the relationship between the F ratio and the t Test. Calculated as the square root of the F ratio. Appears only if the *X* factor has two levels.

Related Information

- "Tests That the Variances Are Equal Report" on page 168

Equivalence Test

Equivalence tests assess whether there is a practical difference in means. You must pick a threshold difference for which smaller differences are considered practically equivalent. The most straightforward test to construct uses two one-sided *t*-tests from both sides of the difference interval. If both tests reject (or conclude that the difference in the means differs significantly from the threshold), then the groups are practically equivalent. The **Equivalence Test** option uses the Two One-Sided Tests (TOST) approach.

Related Information

- "Example of an Equivalence Test" on page 187

Robust

Outliers can lead to incorrect estimates and decisions. The **Robust** option provides two methods to reduce the influence of outliers in your data set: Robust Fit and Cauchy Fit.

Robust Fit

The Robust Fit option reduces the influence of outliers in the response variable. The Huber M-estimation method is used. Huber M-estimation finds parameter estimates that minimize the Huber loss function, which penalizes outliers. The Huber loss function increases as a quadratic for small errors and linearly for large errors. For more details about robust fitting, see Huber (1973) and Huber and Ronchetti (2009).

Related Information

- "Example of the Robust Fit Option" on page 188

Cauchy Fit

The Cauchy fit option assumes that the errors have a Cauchy distribution. A Cauchy distribution has fatter tails than the normal distribution, resulting in a reduced emphasis on outliers. This option can be useful if you have a large proportion of outliers in your data. However, if your data are close to normal with only a few outliers, this option can lead to incorrect inferences. The Cauchy option estimates parameters using maximum likelihood and a Cauchy link function.

Power

The **Power** option calculates statistical power and other details about a given hypothesis test.

- *LSV* (the Least Significant Value) is the value of some parameter or function of parameters that would produce a certain *p*-value alpha. Said another way, you want to know how small an effect would be declared significant at some *p*-value alpha. The LSV provides a measuring stick for significance on the scale of the parameter, rather than on a probability scale. It shows how sensitive the design and data are.
- *LSN* (the Least Significant Number) is the total number of observations that would produce a specified *p*-value alpha given that the data has the same form. The LSN is defined as the number of observations needed to reduce the variance of the estimates enough to achieve a significant result with the given values of alpha, sigma, and delta (the significance level, the standard deviation of the error, and the effect size). If you need more data to achieve significance, the LSN helps tell you how many more. The LSN is the total number of observations that yields approximately 50% power.
- *Power* is the probability of getting significance (*p*-value < alpha) when a real difference exists between groups. It is a function of the sample size, the effect size, the standard deviation of the error, and the significance level. The power tells you how likely your experiment is to detect a difference (effect size), at a given alpha level.

Note: When there are only two groups in a one-way layout, the LSV computed by the power facility is the same as the least significant difference (LSD) shown in the multiple-comparison tables.

Power Details Window and Reports

The Power Details window and reports are the same as those in the general fitting platform launched by the Fit Model platform. For more details about power calculation, see the Statistical Details appendix in the *Fitting Linear Models* book.

For each of four columns Alpha, Sigma, Delta, and Number, fill in a single value, two values, or the start, stop, and increment for a sequence of values. See Figure 6.31. Power calculations are performed on all possible combinations of the values that you specify.

Alpha (α) Significance level, between 0 and 1 (usually 0.05, 0.01, or 0.10). Initially, a value of 0.05 shows.

Sigma (σ) Standard error of the residual error in the model. Initially, RMSE, the estimate from the square root of the mean square error is supplied here.

Delta (δ) Raw effect size. For details about effect size computations, see the Standard Least Squares chapter in the *Fitting Linear Models* book. The first position is initially set to the square root of the sums of squares for the hypothesis divided by *n*; that is, $\delta = \sqrt{SS/n}$.

Number (n) Total sample size across all groups. Initially, the actual sample size is put in the first position.

Solve for Power Solves for the power (the probability of a significant result) as a function of all four values: α, σ, δ, and n.

Solve for Least Significant Number Solves for the number of observations needed to achieve approximately 50% power given α, σ, and δ.

Solve for Least Significant Value Solves for the value of the parameter or linear test that produces a p-value of α. This is a function of α, σ, n, and the standard error of the estimate. This feature is available only when the X factor has two levels and is usually used for individual parameters.

Adjusted Power and Confidence Interval When you look at power retrospectively, you use estimates of the standard error and the test parameters.

- Adjusted power is the power calculated from a more unbiased estimate of the non-centrality parameter.
- The confidence interval for the adjusted power is based on the confidence interval for the non-centrality estimate.

Adjusted power and confidence limits are computed only for the original Delta, because that is where the random variation is.

Related Information

- "Example of the Power Option" on page 190
- "Power" on page 203

Normal Quantile Plot

You can create two types of normal quantile plots:

- **Plot Actual by Quantile** creates a plot of the response values versus the normal quantile values. The quantiles are computed and plotted separately for each level of the X variable.
- **Plot Quantile by Actual** creates a plot of the normal quantile values versus the response values. The quantiles are computed and plotted separately for each level of the X variable.

The **Line of Fit** option shows or hides the lines of fit on the quantile plots.

Related Information

- "Example of a Normal Quantile Plot" on page 191

CDF Plot

A CDF plot shows the cumulative distribution function for all of the groups in the Oneway report. CDF plots are useful if you want to compare the distributions of the response across levels of the X factor.

Related Information

- "Example of a CDF Plot" on page 192

Densities

The **Densities** options provide several ways to compare the distribution and composition of the response across the levels of the X factor. There are three density options:

- **Compare Densities** shows a smooth curve estimating the density of each group. The smooth curve is the density estimate for each group.
- **Composition of Densities** shows the summed densities, weighted by each group's counts. At each X value, the Composition of Densities plot shows how each group contributes to the total.
- **Proportion of Densities** shows the contribution of the group as a proportion of the total at each X level.

Related Information

- "Example of the Densities Options" on page 193

Matching Column

Use the **Matching Column** option to specify a matching (ID) variable for a matching model analysis. The **Matching Column** option addresses the case when the data in a one-way analysis come from matched (paired) data, such as when observations in different groups come from the same subject.

Note: A special case of matching leads to the paired *t*-test. The **Matched Pairs** platform handles this type of data, but the data must be organized with the pairs in different columns, not in different rows.

The **Matching Column** option performs two primary actions:

- It fits an additive model (using an iterative proportional fitting algorithm) that includes both the grouping variable (the X variable in the Fit Y by X analysis) and the matching

variable that you select. The iterative proportional fitting algorithm makes a difference if there are hundreds of subjects, because the equivalent linear model would be very slow and would require huge memory resources.

- It draws lines between the points that match across the groups. If there are multiple observations with the same matching ID value, lines are drawn from the mean of the group of observations.

The **Matching Column** option automatically activates the **Matching Lines** option connecting the matching points. To turn the lines off, select **Display Options > Matching Lines**.

The Matching Fit report shows the effects with *F* tests. These are equivalent to the tests that you get with the Fit Model platform if you run two models, one with the interaction term and one without. If there are only two levels, then the *F* test is equivalent to the paired *t*-test.

Note: For details about the Fit Model platform, see the Model Specification chapter in the *Fitting Linear Models* book.

Related Information

- "Example of the Matching Column Option" on page 194

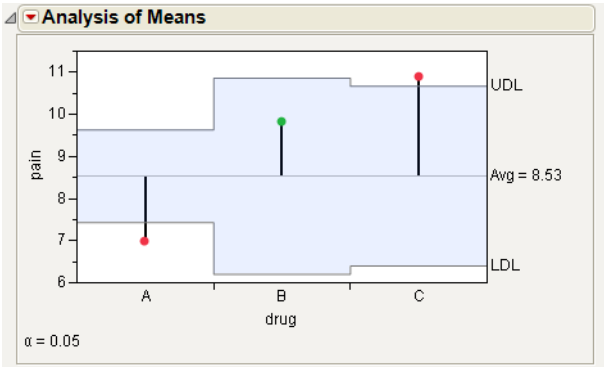
Additional Examples of the Oneway Platform

This section contains additional examples of selected options and reports in the Oneway platform.

Example of an Analysis of Means Chart

1. Select **Help > Sample Data Library** and open Analgesics.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select pain and click **Y, Response**.
4. Select drug and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **Analysis of Means Methods > ANOM**.

Figure 6.13 Example of Analysis of Means Chart



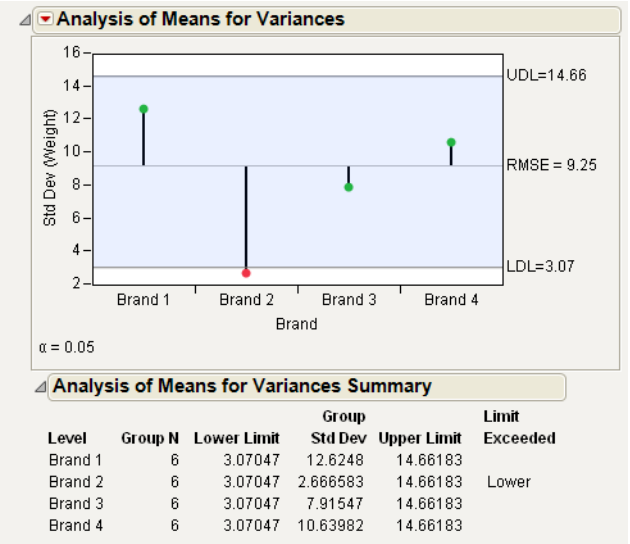
For the example in Figure 6.13, the means for drug A and C are statistically different from the overall mean. The drug A mean is lower and the drug C mean is higher. Note the decision limits for the drug types are not the same, due to different sample sizes.

Example of an Analysis of Means for Variances Chart

This example uses the Spring Data.jmp sample data table. Four different brands of springs were tested to see what weight is required to extend a spring 0.10 inches. Six springs of each brand were tested. The data was checked for normality, since the ANOMV test is not robust to non-normality. Examine the brands to determine whether the variability is significantly different between brands.

1. Select **Help > Sample Data Library** and open Spring Data.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Weight and click **Y, Response**.
4. Select Brand and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **Analysis of Means Methods > ANOM for Variances**.
7. From the red triangle menu next to Analysis of Means for Variances, select **Show Summary Report**.

Figure 6.14 Example of Analysis of Means for Variances Chart



Level	Group N	Lower Limit	Group Std Dev	Upper Limit	Limit Exceeded
Brand 1	6	3.07047	12.6248	14.66183	
Brand 2	6	3.07047	2.666583	14.66183	Lower
Brand 3	6	3.07047	7.91547	14.66183	
Brand 4	6	3.07047	10.63982	14.66183	

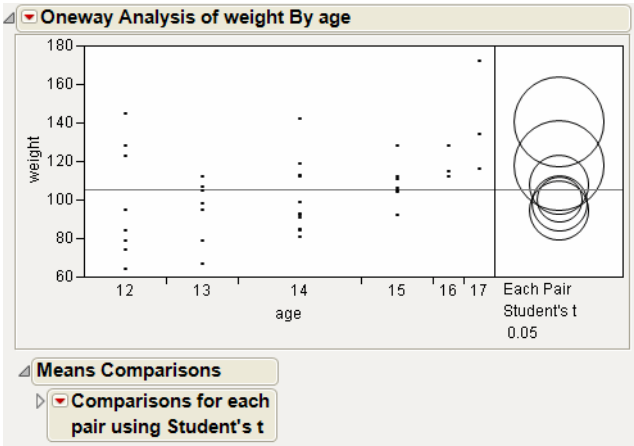
From Figure 6.14, notice that the standard deviation for Brand 2 exceeds the lower decision limit. Therefore, Brand 2 has significantly lower variance than the other brands.

Example of the Each Pair, Student's t Test

This example uses the Big Class.jmp sample data table. It shows a one-way layout of weight by age, and shows the group comparison using comparison circles that illustrate all possible *t*-tests.

1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response.**
4. Select age and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **Compare Means > Each Pair, Student's t**.

Figure 6.15 Example of Each Pair, Student's t Comparison Circles



The means comparison method can be thought of as seeing if the actual difference in the means is greater than the difference that would be significant. This difference is called the LSD (least significant difference). The LSD term is used for Student's *t* intervals and in context with intervals for other tests. In the comparison circles graph, the distance between the circles' centers represent the actual difference. The LSD is what the distance would be if the circles intersected at right angles.

Figure 6.16 Example of Means Comparisons Report for Each Pair, Student's *t*

Means Comparisons

Comparisons for each pair using Student's t

Confidence Quantile

t	Alpha
2.03224	0.05

LSD Threshold Matrix

Abs(Dif)-LSD

	17	16	15	14	12	13
17	-33.140	-10.807	4.373	13.634	14.189	17.944
16	-10.807	-33.140	-17.961	-8.699	-8.145	-4.389
15	4.373	-17.961	-21.695	-11.851	-11.720	-8.124
14	13.634	-8.699	-11.851	-16.570	-16.692	-13.184
12	14.189	-8.145	-11.720	-16.692	-20.294	-16.720
13	17.944	-4.389	-8.124	-13.184	-16.720	-21.695

Positive values show pairs of means that are significantly different.

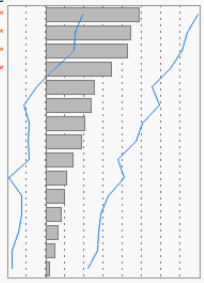
Connecting Letters Report

Level			Mean
17	A		140.66667
16	A	B	118.33333
15		B	108.28571
14		B	100.83333
12		B	99.00000
13		B	94.71429

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
17	13	45.95238	13.78194	17.9441	73.96066	0.0021*
17	12	41.66667	13.52106	14.1886	69.14477	0.0041*
17	14	39.83333	12.89183	13.6340	66.03268	0.0040*
17	15	32.38095	13.78194	4.3727	60.38923	0.0247*
16	13	23.61905	13.78194	-4.3892	51.62733	0.0957
17	16	22.33333	16.30702	-10.8065	55.47318	0.1798
16	12	19.33333	13.52106	-8.1448	46.81144	0.1619
16	14	17.50000	12.89183	-8.6993	43.69935	0.1836
15	13	13.57143	10.67545	-8.1237	35.26655	0.2123
16	15	10.04762	13.78194	-17.9607	38.05590	0.4710
15	12	9.28571	10.33646	-11.7205	30.29192	0.3753
15	14	7.45238	9.49855	-11.8510	26.75576	0.4381
14	13	6.11905	9.49855	-13.1843	25.42242	0.5238
12	13	4.28571	10.33646	-16.7205	25.29192	0.6810
14	12	1.83333	9.11590	-16.6924	20.35907	0.8418



In Figure 6.16, the LSD threshold table shows the difference between the absolute difference in the means and the LSD (least significant difference). If the values are positive, the difference in the two means is larger than the LSD, and the two groups are significantly different.

Example of the All Pairs, Tukey HSD Test

1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response.**
4. Select age and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **Compare Means > All Pairs, Tukey HSD**.

Figure 6.17 Example of All Pairs, Tukey HSD Comparison Circles

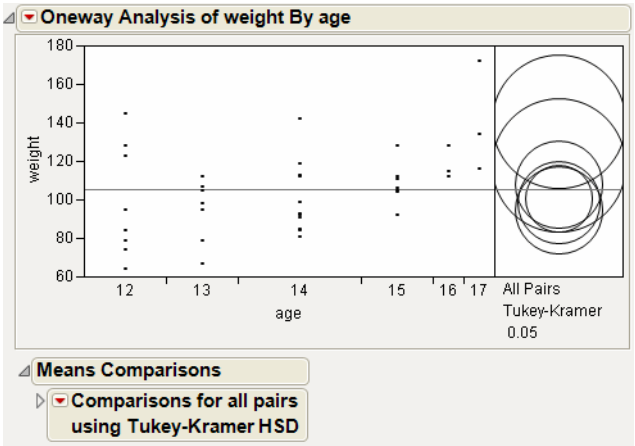


Figure 6.18 Example of Means Comparisons Report for All Pairs, Tukey HSD

Means Comparisons

Comparisons for all pairs using Tukey-Kramer HSD

Confidence Quantile

q*	Alpha
3.01825	0.05

LSD Threshold Matrix

Abs(Dif)-HSD

	17	16	15	14	12	13
17	-49.219	-26.885	-9.216	0.923	0.857	4.355
16	-26.885	-49.219	-31.550	-21.411	-21.477	-17.978
15	-9.216	-31.550	-32.221	-21.217	-21.912	-18.650
14	0.923	-21.411	-21.217	-24.609	-25.681	-22.550
12	0.857	-21.477	-21.912	-25.681	-30.140	-26.912
13	4.355	-17.978	-18.650	-22.550	-26.912	-32.221

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Level			Mean
17	A		140.66667
16	A	B	118.33333
15	A	B	108.28571
14		B	100.83333
12		B	99.00000
13		B	94.71429

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
17	13	45.95238	13.78194	4.3551	87.54968	0.0233*
17	12	41.66667	13.52106	0.8568	82.47657	0.0431*
17	14	39.83333	12.89183	0.9226	78.74404	0.0423*
17	15	32.38095	13.78194	-9.2163	73.97825	0.2029
16	13	23.61905	13.78194	-17.9783	65.21635	0.5325
17	16	22.33333	16.30702	-26.8853	71.55192	0.7443
16	12	19.33333	13.52106	-21.4766	60.14323	0.7091
16	14	17.50000	12.89183	-21.4107	56.41071	0.7512
15	13	13.57143	10.67545	-18.6497	45.79256	0.7982
16	15	10.04762	13.78194	-31.5497	51.64492	0.9769
15	12	9.28571	10.33646	-21.9123	40.48369	0.9442
15	14	7.45238	9.49855	-21.2166	36.12134	0.9683
14	13	6.11905	9.49855	-22.5499	34.78801	0.9867
12	13	4.28571	10.33646	-26.9123	35.48369	0.9983
14	12	1.83333	9.11590	-25.6807	29.34736	1.0000

In Figure 6.18, the Tukey-Kramer HSD Threshold matrix shows the actual absolute difference in the means minus the HSD, which is the difference that would be significant. Pairs with a positive value are significantly different. The q* (appearing above the HSD Threshold Matrix table) is the quantile that is used to scale the HSDs. It has a computational role comparable to a Student's *t*.

Example of the With Best, Hsu MCB Test

1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response.**
4. Select age and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **Compare Means > With Best, Hsu MCB**.

Figure 6.19 Examples of With Best, Hsu MCB Comparison Circles

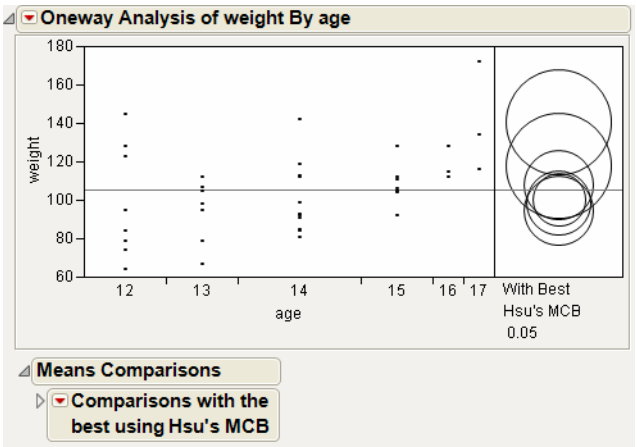


Figure 6.20 Example of Means Comparisons Report for With Best, Hsu MCB

Comparisons with the best using Hsu's MCB

Confidence Quantile

d	Alpha
2.23157	0.05
2.23157	
2.34056	
2.38159	
2.35278	
2.34056	

Comparisons with Max and Min

Level	with Max p-Value	with Min p-Value
17	0.9863	0.0039*
16	0.2219	0.1304
15	0.0491*	0.3172
14	0.0093*	0.6599
12	0.0091*	0.7216
13	0.0046*	0.9408

LSD Threshold Matrix

Mean[i]-Mean[j]-LSD

	17	16	15	14	12	13
17	-36.390	-14.057	0.124	9.130	9.855	13.695
16	-58.724	-36.390	-22.210	-13.203	-12.479	-8.638
15	-63.136	-40.803	-24.987	-15.169	-15.034	-11.415
14	-68.602	-46.269	-29.684	-19.418	-19.614	-16.113
12	-71.840	-49.507	-33.479	-23.544	-23.495	-19.907
13	-76.708	-54.374	-38.558	-28.741	-28.605	-24.987

If a column has any positive values, the mean is significantly less than the max.

Mean[i]-Mean[j]+LSD

	17	16	15	14	12	13
17	36.390	58.724	64.638	70.536	73.479	78.210
16	14.057	36.390	42.305	48.203	51.145	55.876
15	-1.626	20.708	24.987	30.074	33.605	38.558
14	-11.064	11.269	14.780	19.418	23.281	28.351
12	-11.493	10.840	14.907	19.877	23.495	28.479
13	-15.197	7.136	11.415	16.503	20.034	24.987

If a column has any negative values, the mean is significantly greater than the min.

The Comparison with Max and Min report compares the mean of each level to the maximum and the minimum of the means of the remaining levels. For example, the mean for age 15 differs significantly from the maximum of the means of the remaining levels. The mean for age 17 differs significantly from the minimum of the means of the remaining levels. The maximum mean could occur for age 16 or age 17, because neither mean differs significantly from the maximum mean. By the same reasoning, the minimum mean could correspond to any of the ages other than age 17.

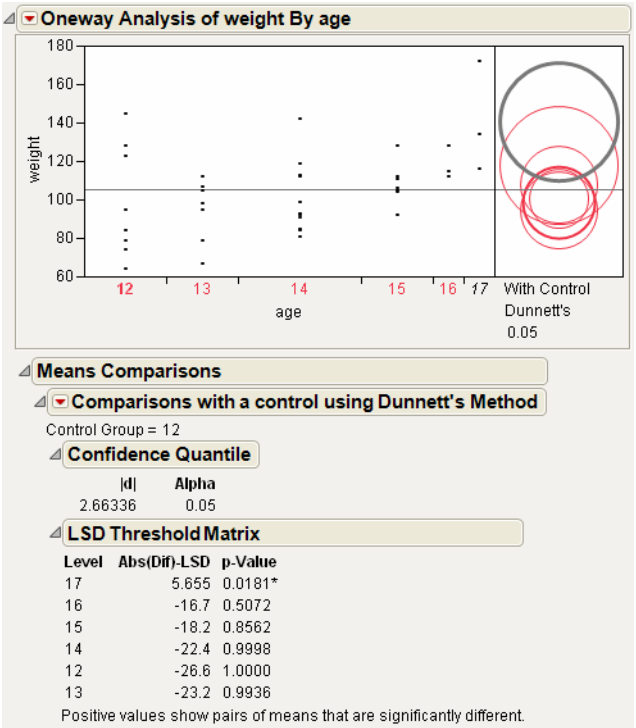
Example of the With Control, Dunnett's Test

1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response.**
4. Select age and click **X, Factor**.

5. Click **OK**.
6. From the red triangle menu, select **Compare Means > With Control, Dunnett's**.
7. Select the group to use as the control group. In this example, select age 12.

   Alternatively, click on a row to highlight it in the scatterplot before selecting the **Compare Means > With Control, Dunnett's** option. The test uses the selected row as the control group.
8. Click **OK**.

Figure 6.21 Example of With Control, Dunnett's Comparison Circles



Using the comparison circles in Figure 6.21, you can conclude that level 17 is the only level that is significantly different from the control level of 12.

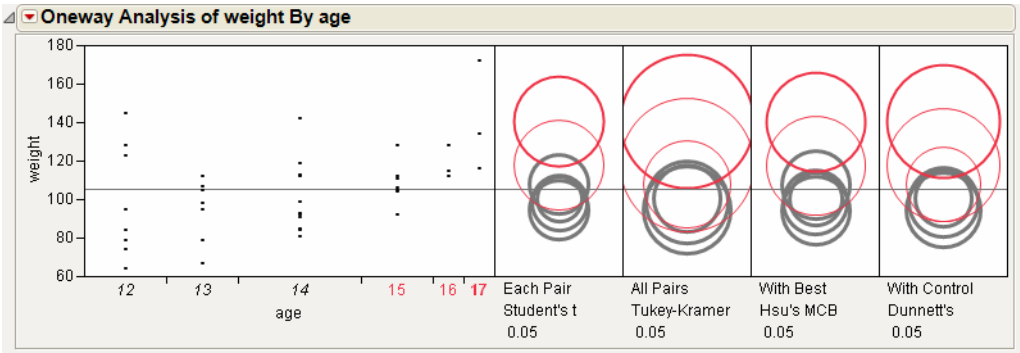
Example Contrasting All of the Compare Means Tests

1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select weight and click **Y, Response.**
4. Select age and click **X, Factor**.

5. Click **OK**.
6. From the red triangle menu, select each one of the **Compare Means** options.

Although the four methods all test differences between group means, different results can occur. Figure 6.22 shows the comparison circles for all four tests, with the age 17 group as the control group.

Figure 6.22 Comparison Circles for Four Multiple Comparison Tests



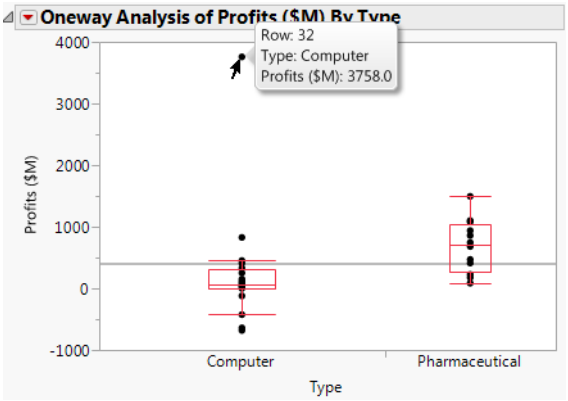
From Figure 6.22, notice that for the Student's t and Hsu methods, age group 15 (the third circle from the top) is significantly different from the control group and appears gray. But, for the Tukey and Dunnett method, age group 15 is not significantly different, and appears red.

Example of the Nonparametric Wilcoxon Test

Suppose you want to test whether the mean profit earned by companies differs by type of company. In Companies.jmp, the data consist of various metrics on two types of companies, Pharmaceutical (12 companies) and Computer (20 companies).

1. Select **Help > Sample Data Library** and open Companies.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Profits ($M) and click **Y, Response.**
4. Select Type and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **Display Options > Box Plots**.

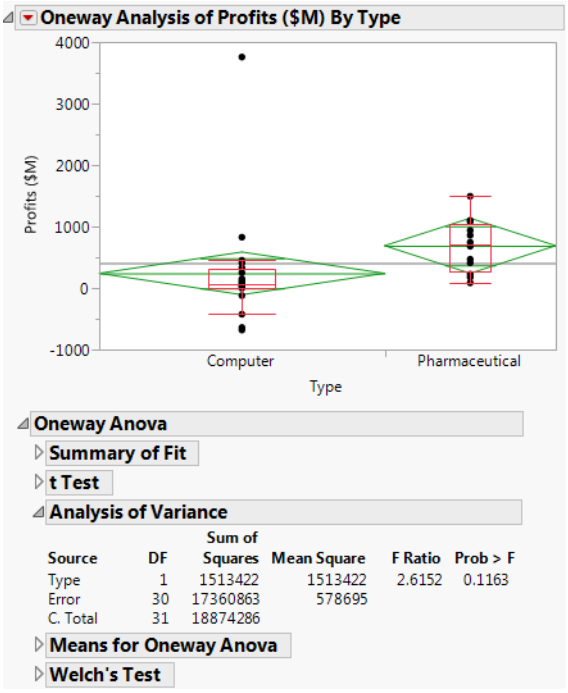
Figure 6.23 Computer Company Profit Distribution



The box plots suggest that the distributions are not normal or even symmetric. There is a very large value for the company in row 32 that might affect parametric tests.

7. From the red triangle menu, select **Means/ANOVA/Pooled t**.

Figure 6.24 Company Analysis of Variance



Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Type	1	1513422	1513422	2.6152	0.1163
Error	30	17360863	578695		
C. Total	31	18874286			

The F test shows no significance because the p-value is large ($p = 0.1163$). This might be due to the large value in row 32 and the possible violation of the normality assumption.

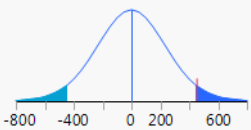
8. From the red triangle menu, select **t Test**.

Figure 6.25 t-Test Results

t Test

Pharmaceutical-Computer
Assuming unequal variances

Difference	449.21	t Ratio	1.902649
Std Err Dif	236.10	DF	28.98322
Upper CL Dif	932.10	Prob > |t|	0.0671
Lower CL Dif	-33.68	Prob > t	0.0335*
Confidence	0.95	Prob < t	0.9665



The Prob > |t| for a two-sided test is 0.0671. The t Test does not assume equal variances, but the unequal variances *t*-test is also a parametric test.

9. From the red triangle menu, select **Nonparametric > Wilcoxon Test**.

Figure 6.26 Wilcoxon Test Results

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
Computer	20	245.000	330.000	12.2500	-3.289
Pharmaceutical	12	283.000	198.000	23.5833	3.289

2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
283	3.28916	0.0010*

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
10.9470	1	0.0009*

The Wilcoxon test is a nonparametric test. It is based on ranks and so is resistant to outliers. Also, it does not require normality.

Both the normal and the chi-square approximations for the Wilcoxon test statistic indicate significance at a *p*-value of 0.0010. You conclude that there is a significant difference in the location of the distributions, and conclude that mean profit differs based on company type.

The normal and chi-square tests are based on the asymptotic distributions of the test statistics. If you have JMP Pro, you can conduct an exact test.

10. JMP PRO From the red triangle menu, select **Nonparametric > Exact Test > Wilcoxon Exact Test**.

Figure 6.27 Wilcoxon Exact Test Results

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
Computer	20	245.000	330.000	12.2500	-3.289
Pharmaceutical	12	283.000	198.000	23.5833	3.289

2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
283	3.28916	0.0010*

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
10.9470	1	0.0009*

2-Sample: Exact Test

S	Prob≥S	Prob≥|S-Mean|
283	0.0003*	0.0005*

The observed value of the test statistic is S = 283. This is the sum of the midranks for the level of Type with the smaller sample size (pharmaceuticals). The probability of observing an absolute difference from the mean midrank that exceeds the absolute value of S minus the mean of the midranks is 0.0005. This is a two-sided test for a difference in location and supports rejecting the hypothesis that profits do not differ by type of company.

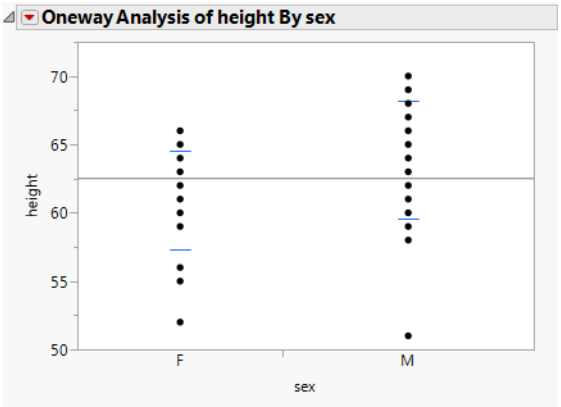
In this example, the nonparametric tests are more appropriate than the normality-based ANOVA test and the unequal variances *t*-test. The nonparametric tests are resistant to the large value in row 32 and do not require the underlying normality of Profits ($M) for each group.

Example of the Unequal Variances Option

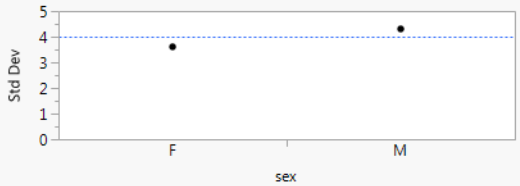
Suppose you want to test whether two variances (males and females) are equal, instead of two means.

1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select height and click **Y, Response.**
4. Select sex and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **Unequal Variances**.

Figure 6.28 Example of the Unequal Variances Report



Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
F	18	3.611890	2.592593	2.555556
M	22	4.308453	3.115702	3.090909

Test	F Ratio	DFNum	DFDen	p-Value
O'Brien[.5]	0.3018	1	38	0.5860
Brown-Forsythe	0.3615	1	38	0.5512
Levene	0.3716	1	38	0.5458
Bartlett	0.5593	1	.	0.4545
F Test 2-sided	1.4229	21	17	0.4644

Since the p-value from the 2-sided F-Test is large, you cannot conclude that the variances are unequal.

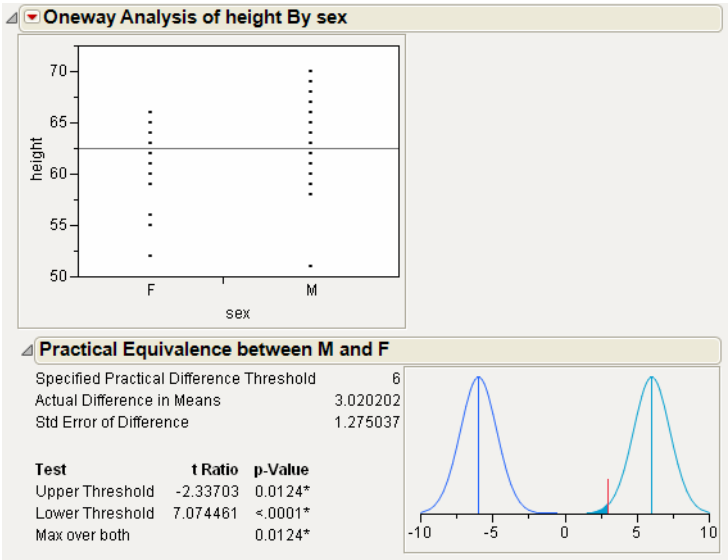
Example of an Equivalence Test

This example uses the Big Class.jmp sample data table. Examine if the difference in height between males and females is less than 6 inches.

1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select height and click **Y, Response.**
4. Select sex and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **Equivalence Test**.

7. Type 6 as the difference considered practically zero.
8. Click **OK**.

Figure 6.29 Example of an Equivalence Test



From Figure 6.29, notice the following:

- The test at the Upper Threshold is the lower one-tailed test for the null hypothesis that the mean difference is 6.
- The test at the Lower Threshold is the upper one-tailed test for the null hypothesis that the mean difference is -6.
- For both tests, the p-value is small. Therefore, you can conclude that the difference in population means is significantly located somewhere from 6 to -6. For your purposes, you can declare the means to be practically equivalent.

Example of the Robust Fit Option

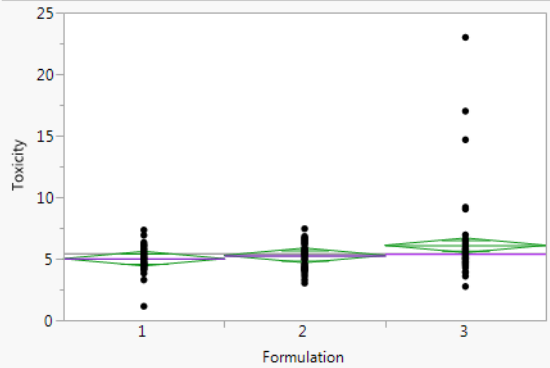
The data in the Drug Toxicity.jmp sample data table shows the toxicity levels for three different formulations of a drug.

1. Select **Help > Sample Data Library** and open Drug Toxicity.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Toxicity and click **Y, Response**.
4. Select Formulation and click **X, Factor**.
5. Click **OK**.

6. From the red triangle menu, select **Means/Anova**.
7. From the red triangle menu, select **Robust > Robust Fit**.

Figure 6.30 Example of Robust Fit

Oneway Analysis of Toxicity By Formulation



Oneway Anova

Summary of Fit

Rsquare	0.045786
Adj Rsquare	0.032803
Root Mean Square Error	2.128216
Mean of Response	5.460866
Observations (or Sum Wgts)	150

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Formulation	2	31.94736	15.9737	3.5267	0.0319*
Error	147	665.80744	4.5293		
C. Total	149	697.75480			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
1	50	5.01002	0.30098	4.4152	5.6048
2	50	5.27760	0.30098	4.6828	5.8724
3	50	6.09498	0.30098	5.5002	6.6898

Std Error uses a pooled estimate of error variance

Welch's Test

F Ratio	DFNum	DFDen	Prob > F
2.7695	2	89.359	0.0681

Welch Anova testing Means Equal, allowing Std Devs Not Equal

Robust Fit

Sigma	ChiSquare	PValue	LogWorth
1.01391	3.05069	0.2175	0.66245

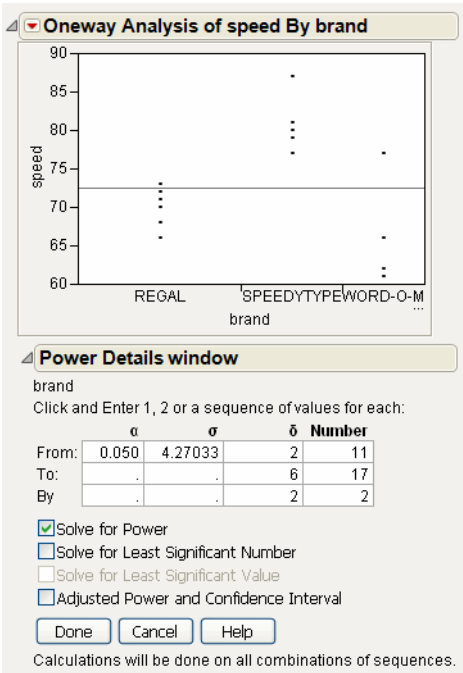
Level	Robust Mean	Std Error
1	5.04222	0.12241
2	5.27785	0.15014
3	5.39175	0.17766

If you look at the standard Analysis of Variance report, you might wrongly conclude that there is a difference between the three formulations, since the p-value is 0.0319. However, when you look at the Robust Fit report, you would not conclude that the three formulations are significantly different, because the p-value there is 0.21755. It appears that the toxicity for a few of the observations is unusually high, creating the undue influence on the data.

Example of the Power Option

1. Select **Help > Sample Data Library** and open Typing Data.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select speed and click **Y, Response.**
4. Select brand and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **Power**.
7. Within the From row, type 2 for Delta (the third box) and type 11 for Number.
8. Within the To row, type 6 for Delta, and type 17 in the Number box.
9. Within the By row, type 2 for both Delta and Number.
10. Select the **Solve for Power** check box.

Figure 6.31 Example of the Power Details Window



11. Click **Done**.

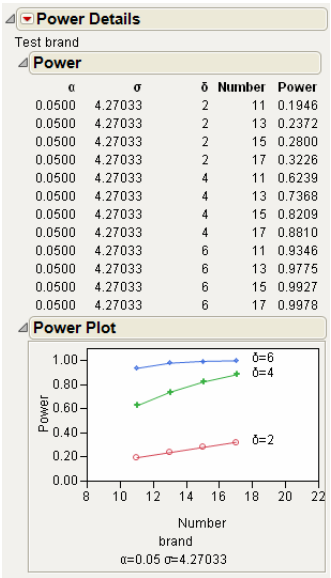
Note: The **Done** button remains dimmed until all of the necessary options are applied.

Power is computed for each combination of Delta and Number, and appears in the Power report.

To plot the Power values:

12. From the red triangle menu at the bottom of the report, select **Power Plot**.

Figure 6.32 Example of the Power Report



Power Details

Test brand

Power

α	σ	δ	Number	Power
0.0500	4.27033	2	11	0.1946
0.0500	4.27033	2	13	0.2372
0.0500	4.27033	2	15	0.2800
0.0500	4.27033	2	17	0.3226
0.0500	4.27033	4	11	0.6239
0.0500	4.27033	4	13	0.7368
0.0500	4.27033	4	15	0.8209
0.0500	4.27033	4	17	0.8810
0.0500	4.27033	6	11	0.9346
0.0500	4.27033	6	13	0.9775
0.0500	4.27033	6	15	0.9927
0.0500	4.27033	6	17	0.9978

Power Plot

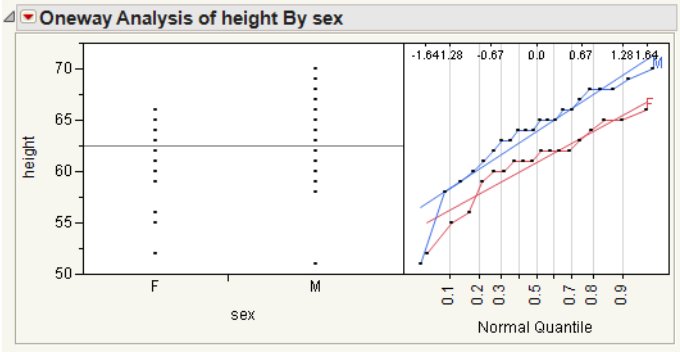
13. You might need to click and drag vertically on the Power axis to see all of the data in the plot.

Power is plotted for each combination of Delta and Number. As you might expect, the power rises for larger Number (sample sizes) values and for larger Delta values (difference in means).

Example of a Normal Quantile Plot

1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select height and click **Y, Response.**
4. Select sex and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **Normal Quantile Plot > Plot Actual by Quantile**.

Figure 6.33 Example of a Normal Quantile Plot



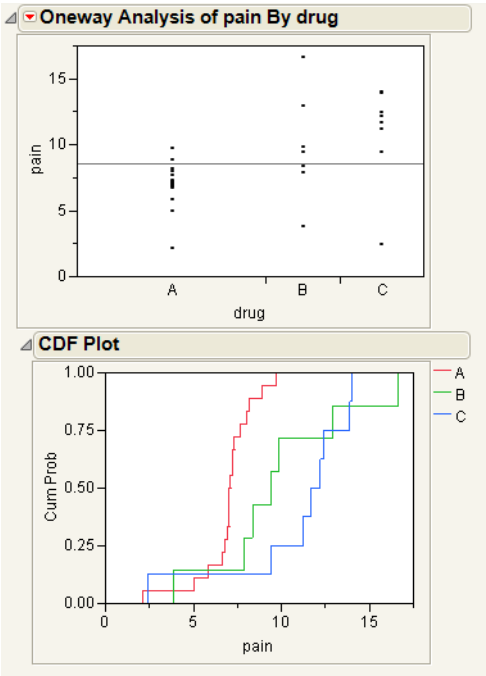
From Figure 6.33, notice the following:

- The Line of Fit appears by default.
- The data points track very closely to the line of fit, indicating a normal distribution.

Example of a CDF Plot

1. Select **Help > Sample Data Library** and open Analgesics.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select pain and click **Y, Response**.
4. Select drug and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **CDF Plot**.

Figure 6.34 Example of a CDF Plot

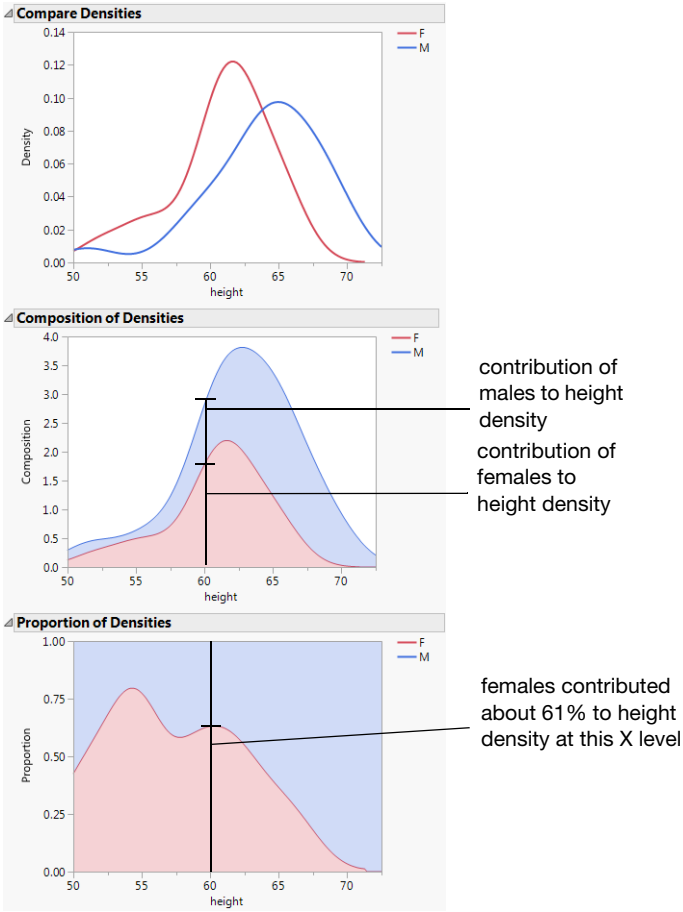


The levels of the X variables in the initial Oneway analysis appear in the CDF plot as different curves. The horizontal axis of the CDF plot uses the *y* value in the initial Oneway analysis.

Example of the Densities Options

1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select height and click **Y, Response.**
4. Select sex and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **Densities > Compare Densities**, **Densities > Composition of Densities**, and **Densities > Proportion of Densities**.

Figure 6.35 Example of the Densities Options



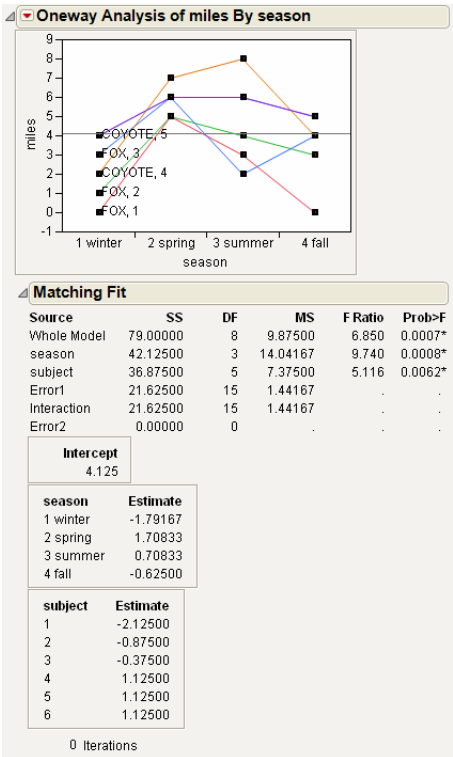
Example of the Matching Column Option

This example uses the Matching.jmp sample data table, which contains data on six animals and the miles that they travel during different seasons.

1. Select **Help > Sample Data Library** and open Matching.jmp.
2. Select **Analyze** > **Fit Y by X**.
3. Select miles and click **Y, Response**.
4. Select season and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **Matching Column**.
7. Select subject as the matching column.

8. Click **OK**.

Figure 6.36 Example of the Matching Column Report



Source	SS	DF	MS	F Ratio	Prob>F
Whole Model	79.00000	8	9.87500	6.850	0.0007*
season	42.12500	3	14.04167	9.740	0.0008*
subject	36.87500	5	7.37500	5.116	0.0062*
Error1	21.62500	15	1.44167	.	.
Interaction	21.62500	15	1.44167	.	.
Error2	0.00000	0	.	.	.

Intercept
4.125

season	Estimate
1 winter	-1.79167
2 spring	1.70833
3 summer	0.70833
4 fall	-0.62500

subject	Estimate
1	-2.12500
2	-0.87500
3	-0.37500
4	1.12500
5	1.12500
6	1.12500

The plot graphs the miles traveled by season, with subject as the matching variable. The labels next to the first measurement for each subject on the graph are determined by the species and subject variables.

The Matching Fit report shows the season and subject effects with *F* tests. These are equivalent to the tests that you get with the Fit Model platform if you run two models, one with the interaction term and one without. If there are only two levels, then the *F* test is equivalent to the paired *t*-test.

Note: For details about the Fit Model platform, see the Model Specification chapter in the *Fitting Linear Models* book.

Example of Stacking Data for a Oneway Analysis

When your data are in a format other than a JMP data table, sometimes they are arranged so that a row contains information for multiple observations. To analyze the data in JMP, you

must import the data and restructure it so that each row of the JMP data table contains information for a single observation. For example, suppose that your data are in a spreadsheet. The data for parts produced on three production lines are arranged in three sets of columns. In your JMP data table, you need to stack the data from the three production lines into a single set of columns so that each row represents the data for a single part.

Description and Goals

This example uses the file Fill Weights.xlsx, which contains the weights of cereal boxes randomly sampled from three different production lines. Figure 6.37 shows the format of the data.

- The ID columns contain an identifier for each cereal box that was measured.
- The Line columns contain the weights (in ounces) for boxes sampled from the corresponding production line.

Figure 6.37 Data Format

Weights					
ID	Line A	ID	Line B	ID	Line C
215	12.42	705	13.63	254	11.73
287	12.49	670	12.56	282	11.40
381	12.80	715	12.87	938	12.78
		683	13.09	597	12.19
		514	13.31	179	12.25
		517	12.64		
		946	12.75		

The target fill weight for the boxes is 12.5 ounces. Although you are interested in whether the three production lines are meeting the target, initially you want to see whether the three lines are achieving the same mean fill rate. You can use Oneway to test for differences among the mean fill weights.

To use the Oneway platform, you need to do the following:

1. Import the data into JMP. See "Import the Data" on page 196.
2. Reshape the data so that each row in the JMP data table reflects only a single observation. Reshaping the data requires that you stack the cereal box IDs, the line identifiers, and the weights into columns. See "Stack the Data" on page 198.

Import the Data

This example illustrates two ways to import data from Microsoft Excel into JMP. Select one method or explore both:

- Use the **File > Open** option to import data from a Microsoft Excel file using the Excel Import Wizard. See "Import the Data Using the Excel Import Wizard" on page 197. This method is convenient for any Excel file.
- Copy and paste data from Microsoft Excel into a new JMP data table. See "Copy and Paste the Data from Excel" on page 197. You can use this method with small data files.

For more information on how to import data from Microsoft Excel, see the Import Your Data chapter in the *Using JMP* book.

Import the Data Using the Excel Import Wizard

1. Select **Help > Sample Data Library** and open Fill Weights.xlsx located in the Samples/Import Data folder.

   The file opens in the Excel Import Wizard.

2. Type 3 next to **Column headers start on on row**.

   In the Excel file, row 1 contains information about the table and row 2 is blank. The column header information starts on row 3.

3. Type 2 for **Number of rows with column headers**.

   In the Excel file, rows 3 and 4 both contain column header information.

4. Click **Import**.

Figure 6.38 JMP Table Created Using Excel Import Wizard

	Weights-ID	Weights-Line A	Weights-ID 2	Weights-Line B	Weights-ID 3	Weights-Line C
1	215	12.42	705	13.63	254	11.73
2	287	12.49	670	12.56	282	11.4
3	381	12.8	715	12.87	938	12.78
4	•	•	683	13.09	597	12.19
5	•	•	514	13.31	179	12.25
6	•	•	517	12.64	•	•
7	•	•	946	12.75	•	•

   The data are placed in seven rows and multiple IDs appear in each row. For each of the three lines, there are an ID and Weight column, giving a total of six columns.

   Notice that the "Weights" part of the ID column name is unnecessary and misleading. You could rename the columns now, but it will be more efficient to rename the columns after you stack the data.

5. Proceed to "Stack the Data" on page 198.

Copy and Paste the Data from Excel

1. Open Fill Weights.xlsx in Microsoft Excel.
2. Select the data inside the table but exclude the unnecessary "Weights" heading.

3. Right-click and select **Copy**.
4. In JMP, select **File > New > Data Table**.
5. Select **Edit > Paste with Column Names**.

   The **Edit > Paste with Column Names** option is used when you have column names included in the selection on the clipboard.

Figure 6.39 JMP Table Created Using Paste with Column Names

	ID	Line A	ID 2	Line B	ID 3	Line C
1	215	12.42	705	13.63	254	11.73
2	287	12.49	670	12.56	282	11.4
3	381	12.8	715	12.87	938	12.78
4	•	•	683	13.09	597	12.19
5	•	•	514	13.31	179	12.25
6	•	•	517	12.64	•	•
7	•	•	946	12.75	•	•

6. Proceed to "Stack the Data" on page 198.

Stack the Data

Use the Stack option to place one observation in each row of a new data table. For more information on the Stack option, see the Reshape Data chapter in the *Using JMP* book.

1. In the JMP data table, select **Tables > Stack**.
2. Select all six columns and click **Stack Columns**.
3. Select **Multiple Series Stack**.

   You are stacking two series, ID and Line, so you do not change the Number of Series, which is set to 2 by default. The columns that contain the series are not contiguous. They alternate (ID, Line A, ID, Line B, ID, Line C). For this reason, you do not check Contiguous.
4. Deselect **Stack By Row**.
5. Select **Eliminate Missing Rows**.
6. Enter Stacked next to **Output table name**.
7. Click **OK**.

   In the new data table, Data and Data 2 are columns containing the ID and Weight data.
8. Right-click the Label column heading and select **Delete Columns**.

   The entries in the Label column were the column headings for the box IDs in the imported data table. These entries are not needed.
9. Rename each column by double-clicking on the column header. Change the column names as follows:
   - Data to ID

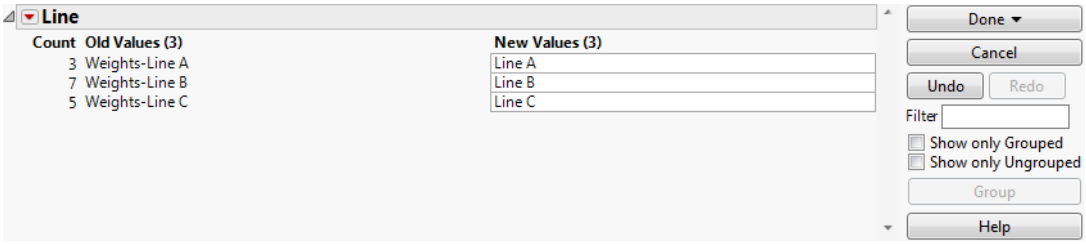
- Label 2 to Line
- Data 2 to Weight

10. In the Columns panel, click on the icon to the left of ID and select **Nominal**.

    Although ID is given as a number, it is an identifier and should be treated as nominal when modeling. This is not an issue in this example, but it is good practice to assign the appropriate modeling type to a column.

11. (Only applies if you imported the data from Excel using **File > Open.)** Do the following:
    1. Click the Line column header to select the column and select **Cols > Recode**.
    2. Change the values in the **New Values** column to match those in Figure 6.40 below.

Figure 6.40 Recode Column Values



    3. Click **Done > In place**.

Your new data table is now properly structured for JMP analysis. Each row contains data for a single cereal box. The first column gives the box ID, the second gives the production line, and the third gives the weight of the box (Figure 6.41).

Figure 6.41 Recoded Data Table

	ID	Line	Weight
1	215	Line A	12.42
2	287	Line A	12.49
3	381	Line A	12.8
4	705	Line B	13.63
5	670	Line B	12.56
6	715	Line B	12.87
7	683	Line B	13.09
8	514	Line B	13.31
9	517	Line B	12.64
10	946	Line B	12.75
11	254	Line C	11.73
12	282	Line C	11.4
13	938	Line C	12.78
14	597	Line C	12.19
15	179	Line C	12.25

Conduct the Oneway Analysis

For this part of the example, you will:

- Conduct a Oneway Analysis of Variance to test for differences in the mean fill weights among the three production lines.
- Obtain Comparison Circles to explore which lines might differ.
- Label points by ID in case you want to reweigh or further examine their boxes.

Before beginning, verify that you are using the Stacked data table.

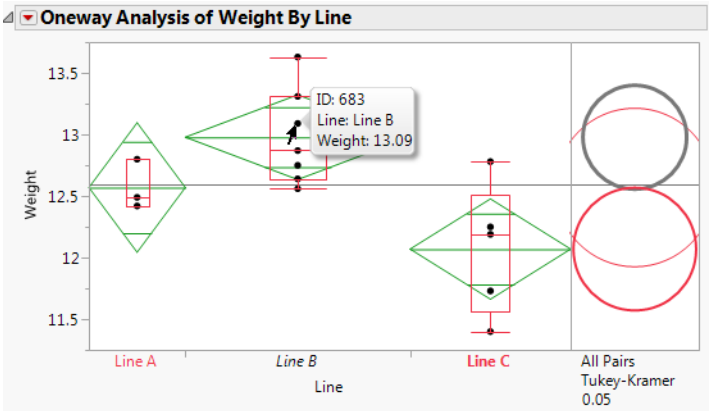
1. Select **Analyze > Fit Y by X**.
2. Select Weight and click **Y, Response**.
3. Select Line and click **X, Factor**.
4. Click **OK**.
5. From the red triangle menu, select **Means/Anova**.

   The mean diamonds in the plot show 95% confidence intervals for the production line means. The points that fall outside the mean diamonds might seem like outliers, however they are not. To see this, add box plots to the plot.
6. From the red triangle menu, select **Display Options > Box Plots**.

   All points fall within the box plots boundaries, therefore they are not outliers.
7. From the data table, in the Columns panel, right-click ID and select **Label/Unlabel**.
8. In the plot, hover over the points with your mouse to see their ID values, as well as their Line and Weight data. See Figure 6.42.
9. From the red triangle menu, select **Compare Means > All Pairs, Tukey HSD**.

   Comparison circles appear in a panel to the right of the plot.
10. Click the bottom comparison circle.

Figure 6.42 Oneway Analysis of Weight by Line



Oneway Anova

Summary of Fit

Rsquare	0.534629
Adj Rsquare	0.457068
Root Mean Square Error	0.418099
Mean of Response	12.594
Observations (or Sum Wgts)	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Line	2	2.4098743	1.20494	6.8930	0.0102*
Error	12	2.0976857	0.17481		
C. Total	14	4.5075600			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Line A	3	12.5700	0.24139	12.044	13.096
Line B	7	12.9786	0.15803	12.634	13.323
Line C	5	12.0700	0.18698	11.663	12.477

Std Error uses a pooled estimate of error variance

Welch's Test

Means Comparisons

In the Analysis of Variance report, the p-value of 0.0102 provides evidence that the means are not all equal. In the plot, the comparison circle for Line C is selected and appears red (in JMP default colors). Since the circle for Line B appears as thick gray, the mean for Line C differs from the mean for Line B at the 0.05 significance level. The means for Lines A and B do not show a statistically significant difference.

The mean diamonds shown in the plot span 95% confidence intervals for the means. The numeric bounds for the 95% confidence intervals are given in the Means for Oneway ANOVA report. Both of these indicate that the confidence intervals for Lines B and C do not contain the target fill weight of 12.5: Line B appears to overfill and Line C appears to underfill. For these two production lines, the underlying causes that result in off-target fill weights must be addressed.

Statistical Details for the Oneway Platform

The following sections provide statistical details for selected options and reports.

Comparison Circles

One approach to comparing two means is to determine whether their actual difference is greater than their *least significant difference* (LSD). This least significant difference is a Student's *t*-statistic multiplied by the standard error of the difference of the two means and is written as follows:

$$\text{LSD} = t_{\alpha/2}\text{std}(\hat{\mu}_1 - \hat{\mu}_2)$$

The standard error of the difference of two independent means is calculated from the following relationship:

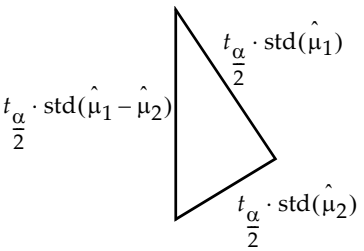
$$[\text{std}(\hat{\mu}_1 - \hat{\mu}_2)]^2 = [\text{std}(\hat{\mu}_1)]^2 + [\text{std}(\hat{\mu}_2)]^2$$

When the means are un correlated, these quantities have the following relationship:

$$\text{LSD}^2 = [t_{\alpha/2}\text{std}((\hat{\mu}_1 - \hat{\mu}_2))]^2 = [t_{\alpha/2}\text{std}(\hat{\mu}_1)]^2 + [t_{\alpha/2}\text{std}\hat{\mu}_2]^2$$

These squared values form a Pythagorean relationship, illustrated graphically by the right triangle shown in Figure 6.43.

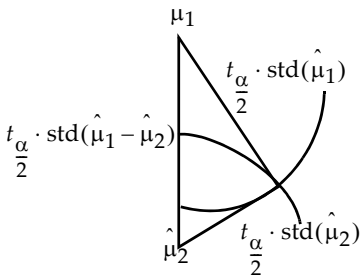
Figure 6.43 Relationship of the Difference between Two Means



The hypotenuse of this triangle is a measuring stick for comparing means. The means are significantly different if and only if the actual difference is greater than the hypotenuse (LSD).

Suppose that you have two means that are exactly on the borderline, where the actual difference is the same as the least significant difference. Draw the triangle with vertices at the means measured on a vertical scale. Also, draw circles around each mean so that the diameter of each is equal to the confidence interval for that mean.

Figure 6.44 Geometric Relationship of *t*-test Statistics



The radius of each circle is the length of the corresponding leg of the triangle, which is $t_{\alpha/2}\text{std}(\mu_i)$.

The circles must intersect at the same right angle as the triangle legs, giving the following relationship:

- If the means differ exactly by their least significant difference, then the confidence interval circles around each mean intersect at a right angle. That is, the angle of the tangents is a right angle.

Now, consider the way that these circles must intersect if the means are different by greater than or less than the least significant difference:

- If the circles intersect so that the outside angle is greater than a right angle, then the means *are not* significantly different. If the circles intersect so that the outside angle is less than a right angle, then the means *are* significantly different. An outside angle of less than 90 degrees indicates that the means are farther apart than the least significant difference.
- If the circles do not intersect, then they are significantly different. If they nest, they are not significantly different. See Figure 6.11.

The same graphical technique works for many multiple-comparison tests, substituting a different probability quantile value for the Student's *t*.

Power

To compute power, you make use of the noncentral *F* distribution. The formula (O'Brien and Lohr 1984) is given as follows:

Power = Prob($F > F_{crit}$, ν_1, ν_2, nc)

where:

- F is distributed as the noncentral $F(nc, v_1, v_2)$ and $F_{crit} = F_{(1-\alpha, v1, v2)}$ is the 1 - α quantile of the F distribution with ν_1 and ν_2 degrees of freedom.
- $\nu_1 = r - 1$ is the numerator df.

- $\nu_2 = r(n - 1)$ is the denominator df.
- n is the number per group.
- r is the number of groups.
- $nc = n(CSS)/\sigma^2$ is the non-centrality parameter.

  $CSS = \sum_{g=1}^{r} (\mu_g - \mu)^2$ is the corrected sum of squares.
- μ_g is the mean of the gth group.
- μ is the overall mean.
- σ^2 is estimated by the mean squared error (MSE).

Summary of Fit Report

Rsquare

Using quantities from the Analysis of Variance report for the model, the R^2 for any continuous response fit is always calculated as follows:

$$\frac{\text{Sum of Squares (Model)}}{\text{Sum of Squares (C Total)}}$$

Adj Rsquare

Adj Rsquare is a ratio of mean squares instead of sums of squares and is calculated as follows:

$$1 - \frac{\text{Mean Square (Error)}}{\text{Mean Square (C Total)}}$$

The mean square for Error is found in the Analysis of Variance report and the mean square for **C. Total** can be computed as the **C. Total** Sum of Squares divided by its respective degrees of freedom. See "The Analysis of Variance Report" on page 149.

Tests That the Variances Are Equal

F Ratio

O'Brien's test constructs a dependent variable so that the group means of the new variable equal the group sample variances of the original response. The O'Brien variable is computed as follows:

$$r_{ijk} = \frac{(n_{ij}-1.5)n_{ij}(y_{ijk}-\overline{y_{ij}})^2 - 0.5s_{ij}^2(n_{ij}-1)}{(n_{ij}-1)(n_{ij}-2)}$$

where n represents the number of y_{ijk} observations.

Brown-Forsythe is the model F statistic from an ANOVA on $z_{ij} = |y_{ij}-\tilde{y}_i|$ where $\tilde{y}_i$ is the median response for the ith level.

The Levene F is the model F statistic from an ANOVA on $z_{ij} = |y_{ij}-\overline{y_{i.}}|$ where $\overline{y_{i.}}$ is the mean response for the ith level.

Bartlett's test is calculated as follows:

$$T = \frac{v\log\left(\sum_i \frac{v_i}{v}s_i^2\right) - \sum_i v_i \log(s_i^2)}{1+\left(\frac{\sum_i \frac{1}{v_i} - \frac{1}{v}}{3(k-1)}\right)} \text{ where } v_i = n_i - 1 \text{ and } v = \sum_i v_i$$

and n_i is the count on the ith level and s_i^2 is the response sample variance on the ith level. The Bartlett statistic has a χ^2-distribution. Dividing the Chi-square test statistic by the degrees of freedom results in the reported F value.

Welch's Test F Ratio

The Welch's Test F Ratio is computed as follows:

$$F = \frac{\left[\frac{\sum_i w_i(\bar{y}_i - \tilde{y}_{..})^2}{k-1}\right]}{\left\{1+\frac{2(k-2)}{k^2-1}\left[\sum_i \frac{\left(1-\frac{w_i}{u}\right)^2}{n_i-1}\right]\right\}} \text{ where } w_i = \frac{n_i}{s_i^2},\ u = \sum_i w_i,\ \tilde{y}_{..} = \sum_i \frac{w_i \bar{y}_{i.}}{u},$$

and n_i is the count on the ith level, $\bar{y}_{i.}$ is the mean response for the ith level, and s_i^2 is the response sample variance for the ith level.

Welch's Test DF Den

The Welch approximation for the denominator degrees of freedom is as follows:

$$df = \frac{1}{\left(\frac{3}{k^2-1}\right)\left[\sum_i \frac{\left(1-\frac{w_i}{u}\right)^2}{n_i-1}\right]}$$

where w_i, n_i, and u are defined as in the F ratio formula.

Nonparametric Test Statistics

This section provides formulas for the test statistics used in the Wilcoxon, Median, and van der Waerden tests.

Notation

The tests are based on scores and use the following notation.

$j = 1,\ldots,n$ The observations in the entire sample.

$n_1, n_2, \ldots, n_k$ The numbers of observations in each of the k levels of X.

R_j The midrank of the jth observation. The *midrank* is the observation's rank if it is not tied and its average rank if it is tied.

α A function of the midranks used to define scores for the various tests.

$\alpha(R_j)$ The function α applied to the midrank of the jth observation.

The function α defines scores as follows:

Wilcoxon scores

$$\alpha(R_j) = R_j$$

Median scores

$$\alpha(R_j) = \begin{cases} 1 \text{ if } R_j > (n+1)/2 \\ 0 \text{ if } R_j < (n+1)/2 \\ t \text{ if } R_j = \text{median} \end{cases}$$

Let n_t denote the number of observations tied at the median. Then n_t is given by the following:

$$n_t = \begin{cases} (n_t/2) \text{ if } n_t \text{ is even} \\ (n_t+1)/(2n_t) \text{ if } n_t \text{ is odd} \end{cases}$$

van der Waerden scores

$$\alpha(R_j) = \text{Standard Normal Quantile}(R_j/(n+1))$$

Two-Sample Normal Approximations

Tests based on the normal approximation are given only when X has exactly two levels. The notation used in this section is defined in "Notation" on page 206. The statistics that appear in the Two-Sample Normal Approximation report are defined below.

S The statistic *S* is the sum of the values $\alpha(R_j)$ for the observations in the smaller group. If the two levels of X have the same numbers of observations, then the value of S corresponds to the last level of X in the value ordering.

Z The value of Z is given as follows:

$$Z = (S - E(S))/\sqrt{Var(S)}$$

Note: The Wilcoxon test adds a continuity correction. If (*S - E(S)*) is greater than zero, then 0.5 is subtracted from the numerator. If (*S - E(S)*) is less than zero, then 0.5 is added to the numerator.

E(S) The expected value of S under the null hypothesis. Denote the number of observations in the smaller level, or in the last level in the value ordering if the two groups have the same number of observations, by n_l:

$$E(S) = \frac{n_l}{n} \sum_{j=1}^{n} \alpha(R_j)$$

Var(S) Define *ave* to be the average score across all observations. Then the variance of S is given as follows:

$$Var(S) = \frac{n_1 n_2}{n(n-1)} \sum_{j=1}^{n} (\alpha(R_j) - ave)^2$$

One-Way ChiSquare Approximations

Note: The ChiSquare test based on the Wilcoxon scores is knows as the Kruskal-Wallis test.

The notation used in this section is defined in "Notation" on page 206. The following quantities are used in calculating the ChiSquare statistic:

T_i The total of the scores for the ith level of X.

$E(T_i)$ The expected value of the total score for level *i* under the null hypothesis of no difference in levels, given as follows:

$$E(T_i) = \frac{n_i}{n} \sum_{j=1}^{n} \alpha(R_j)$$

Var(T) Define *ave* to be the average score across all observations. Then the variance of T is given as follows:

$$Var(T) = \frac{1}{(n-1)} \sum_{j=1}^{n} (\alpha(R_j) - ave)^2$$

The value of the test statistic is given below. This statistic is asymptotically ChiSquare on *k - 1* degrees of freedom.

$$C = \left(\sum_{i=1}^{k} (T_i - E(T_i))^2 / n_i \right) / Var(T)$$

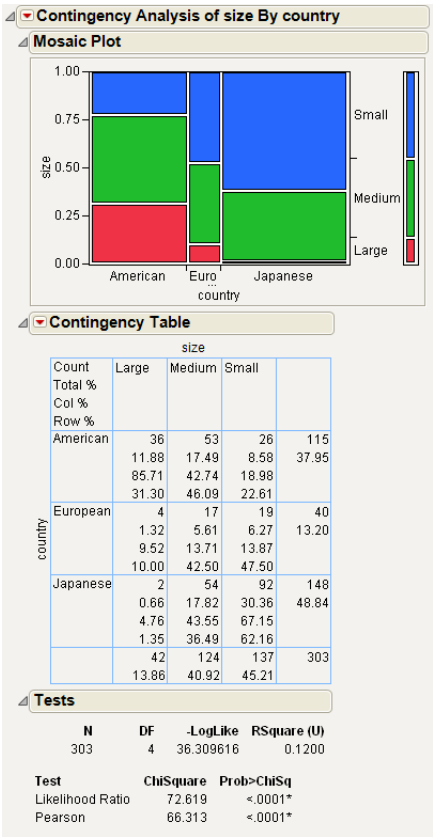
Chapter 7

Contingency Analysis

Examine Relationships between Two Categorical Variables

The Contingency or Fit Y by X platform lets you explore the distribution of a categorical (nominal or ordinal) variable Y across the levels of a second categorical variable X. The Contingency platform is the *categorical by categorical* personality of the Fit Y by X platform. The analysis results include a mosaic plot, frequency counts, and proportions. You can interactively perform additional analyses and tests on your data, such as an Analysis of Means for Proportions, a correspondence analysis plot, and so on.

Figure 7.1 Example of Contingency Analysis



Count Total % Col % Row %	Large	Medium	Small	
American	36 11.88 85.71 31.30	53 17.49 42.74 46.09	26 8.58 18.98 22.61	115 37.95
European	4 1.32 9.52 10.00	17 5.61 13.71 42.50	19 6.27 13.87 47.50	40 13.20
Japanese	2 0.66 4.76 1.35	54 17.82 43.55 36.49	92 30.36 67.15 62.16	148 48.84
	42 13.86	124 40.92	137 45.21	303

N	DF	-LogLike	RSquare (U)
303	4	36.309616	0.1200

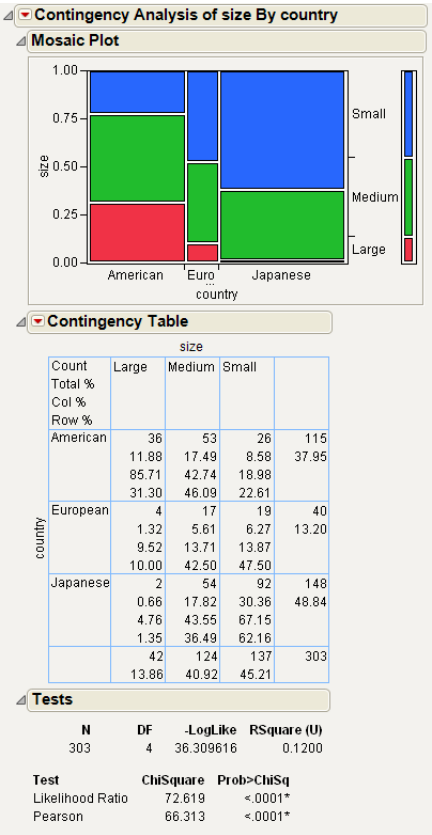
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	72.619	<.0001*
Pearson	66.313	<.0001*

Example of Contingency Analysis

This example uses the Car Poll.jmp sample data table, which contains data collected from car polls. The data include respondent attributes: sex, marital status, and age. The data also include attributes of the respondent's car: country of origin, the size, and the type. Examine the relationship between car sizes (small, medium, and large) and the cars' country of origin.

1. Select **Help > Sample Data Library** and open Car Poll.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select size and click **Y, Response**.
4. Select country and click **X, Factor**.
5. Click **OK**.

Figure 7.2 Example of Contingency Analysis



From the mosaic plot and legend in Figure 7.2, notice the following:

- Very few Japanese cars fall into the Large size category.
- The majority of the European cars fall into the Small and Medium size categories.
- The majority of the American cars fall into the Large and Medium size categories.

Launch the Contingency Platform

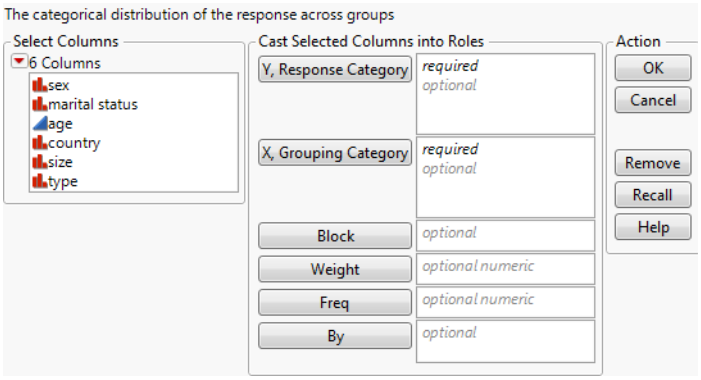
You can perform a contingency analysis using either the Fit Y by X platform or the Contingency platform. The two approaches are equivalent.

- To launch the Fit Y by X platform, select **Analyze > Fit Y by X**.

  or

- To launch the Contingency platform, from the JMP Starter window, click on the **Basic** category and click **Contingency**.

Figure 7.3 The Contingency Launch Window



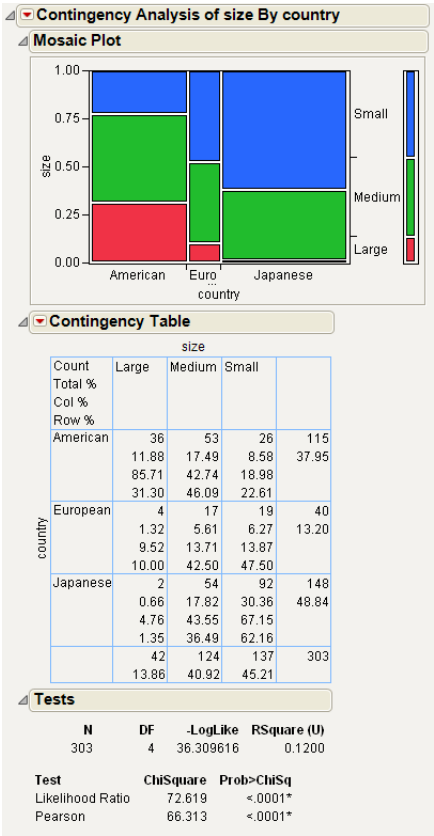
For more information about this launch window, see "Introduction to Fit Y by X" chapter on page 95.

After you click **OK**, the Contingency report window appears. See "The Contingency Report" on page 211.

The Contingency Report

To produce the plot shown in Figure 7.4, follow the instructions in "Example of Contingency Analysis" on page 210.

Figure 7.4 Example of a Contingency Report



Note: Any rows that are excluded in the data table are also hidden in the Mosaic Plot.

The Contingency report initially shows a Mosaic Plot, a Contingency Table, and a Tests report. You can add other analyses and tests using the options that are located within the red triangle menu. For details about all of these reports and options, see "Contingency Platform Options" on page 212.

Contingency Platform Options

Note: The Fit Group menu appears if you have specified multiple Y variables. Menu options allow you to arrange reports or order them by RSquare. See the Standard Least Squares chapter in the *Fitting Linear Models* book for more information.

Use the platform options within the red triangle menu next to Contingency Analysis to perform additional analyses and tests on your data.

Mosaic Plot A graphical representation of the data in the Contingency Table. See "Mosaic Plot" on page 214.

Contingency Table A two-way frequency table. There is a row for each factor level and a column for each response level. See "Contingency Table" on page 216.

Tests Analogous to the Analysis of Variance table for continuous data. The tests show that the response level rates are the same across X levels. See "Tests" on page 218.

Set α level Changes the alpha level used in confidence intervals. Select one of the common values (0.10, 0.05, 0.01) or select a specific value using the **Other** option.

Analysis of Means for Proportions Only appears if the response has exactly two levels. Compares response proportions for the X levels to the overall response proportion. See "Analysis of Means for Proportions" on page 219.

Correspondence Analysis Shows which rows or columns of a frequency table have similar patterns of counts. In the correspondence analysis plot, there is a point for each row and for each column of the contingency table. See "Correspondence Analysis" on page 219.

Cochran Mantel Haenszel Tests if there is a relationship between two categorical variables after blocking across a third classification. See "Cochran-Mantel-Haenszel Test" on page 221.

Agreement Statistic Only appears when both the X and Y variables have the same levels. Displays the Kappa statistic (Agresti 1990), its standard error, confidence interval, hypothesis test, and Bowker's test of symmetry, also know as McNemar's test. See "Agreement Statistic" on page 221.

Relative Risk Calculates risk ratios. Appears only when both the X and Y variables have only two levels. See "Relative Risk" on page 222.

Odds Ratio Appears only when there are exactly two levels for each variable. Produces a report of the odds ratio. For more information, see "Odds Ratio Option" on page 236.

The report also gives a confidence interval for this ratio. You can change the alpha level using the **Set α Level** option.

Two Sample Test for Proportions Performs a two-sample test for proportions. This test compares the proportions of the Y variable between the two levels of the X variable. Appears only when both the X and Y variables have only two levels. See "Two Sample Test for Proportions" on page 222.

Measures of Association Describes the association between the variables in the contingency table. See "Measures of Association" on page 223.

Cochran Armitage Trend Test Tests for trends in binomial proportions across levels of a single variable. This test is appropriate only when one variable has two levels and the other variable is ordinal. See "Cochran Armitage Trend Test" on page 224.

JMP PRO **Exact Test** Provides exact versions of the following tests:

- Fisher's Test
- Cochran Armitage Trend Test
- Agreement Test

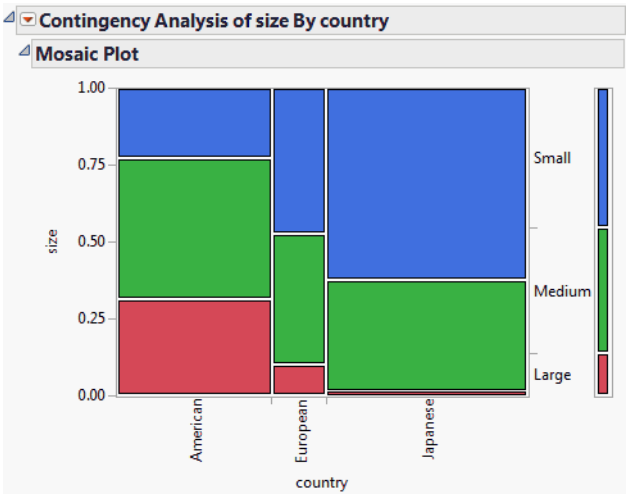
See "Exact Test" on page 225.

Mosaic Plot

The mosaic plot is a graphical representation of the two-way frequency table or Contingency Table. A mosaic plot is divided into rectangles, so that the vertical length of each rectangle is proportional to the proportions of the Y variable in each level of the X variable. The mosaic plot was introduced by Hartigan and Kleiner in 1981 and refined by Friendly (1994).

To produce the plot shown in Figure 7.5, follow the instructions in "Example of Contingency Analysis" on page 210.

Figure 7.5 Example of a Mosaic Plot



Note the following about the mosaic plot in Figure 7.5:

- The proportions on the *x*-axis represent the number of observations for each level of the X variable, which is country.
- The proportions on the *y*-axis at right represent the overall proportions of Small, Medium, and Large cars for the combined levels (American, European, and Japanese).

- The scale of the *y*-axis at left shows the response probability, with the whole axis being a probability of one (representing the total sample).

Clicking on a rectangle in the mosaic plot highlights the selection and highlights the corresponding data in the associated data table.

Replace variables in the mosaic plot by dragging and dropping a variable, in one of two ways: swap existing variables by dragging and dropping a variable from one axis to the other axis; or, click on a variable in the Columns panel of the associated data table and drag it onto an axis.

Context Menu

Right-click on the mosaic plot to change colors and label the cells.

Set Colors Shows the current assignment of colors to levels. See "Set Colors" on page 215.

Cell Labeling Specify a label to be drawn in the mosaic plot. Select one of the following options:

Unlabeled Shows no labels, and removes any of the other options.

Show Counts Shows the number of observations in each cell.

Show Percents Shows the percent of observations in each cell.

Show Labels Shows the levels of the Y variable corresponding to each cell.

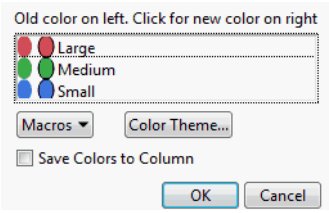
Show Row Labels Shows the row labels for all of the rows represented by the cell.

Note: For descriptions of the remainder of the right-click options, see the JMP Reports chapter in the *Using JMP* book.

Set Colors

When you select the **Set Colors** option, the Select Colors for Values window appears.

Figure 7.6 Select Colors for Values Window



The default mosaic colors depend on whether the response column is ordinal or nominal, and whether there is an existing Value Colors column property. To change the color for any level, click on the oval in the second column of colors and pick a new color.

Description of the Select Colors for Values Window

Macros Computes a color gradient between any two levels, as follows:

- If you select a range of levels (by dragging the mouse over the levels that you want to select, or pressing the Shift key and clicking the first and last level), the **Gradient Between Selected Points** option applies a color gradient to the levels that you have selected.
- The **Gradient Between Ends** option applies a gradient to all levels of the variable.
- Undo any of your changes by selecting **Revert to Old Colors**.

Color Theme Changes the colors for each value based on a color theme.

Save Colors to Column If you select this check box, a new column property (**Value Colors**) is added to the column in the associated data table. To edit this property from the data table, select **Cols > Column Info**.

Contingency Table

The Contingency Table is a two-way frequency table. There is a row for each factor level and a column for each response level.

To produce the plot shown in Figure 7.7, follow the instructions in "Example of Contingency Analysis" on page 210.

Figure 7.7 Example of a Contingency Table

Contingency Table

country	Count Total % Col % Row %	size: Large	Medium	Small	Total
American		36	53	26	115
		11.88	17.49	8.58	37.95
		85.71	42.74	18.98	
		31.30	46.09	22.61	
European		4	17	19	40
		1.32	5.61	6.27	13.20
		9.52	13.71	13.87	
		10.00	42.50	47.50	
Japanese		2	54	92	148
		0.66	17.82	30.36	48.84
		4.76	43.55	67.15	
		1.35	36.49	62.16	
Total		42	124	137	303
		13.86	40.92	45.21	

Note the following about Contingency tables:

- The Count, Total%, Col%, and Row% correspond to the data within each cell that has row and column headings (such as the cell under American and Large).
- The last column contains the total counts for each row and percentages for each row.

- The bottom row contains total counts for each column and percentages for each column.

For example, in Figure 7.7, focus on the cars that are large and come from America. The following table explains the conclusions that you can make about these cars using the Contingency Table.

Table 7.1 Conclusions Based on Example of a Contingency Table

Number	Description	Label in Table
36	Number of cars that are both large and come from America	Count
11.88%	Percentage of all cars that are both large and come from America (36/303)[a].	Total%
85.71%	Percentage of large cars that come from America (36/42)[b]	Col%
31.30%	Percentage of American cars that are large (36/115)[c].	Row%
37.95%	Percentage of all cars that come from America (115/303).	(none)
13.86%	Percentage of all cars that are large (42/303).	(none)

a. 303 is the total number of cars in the poll.
b. 42 is the total number of large cars in the poll.
c. 115 is the total number of American cars in the poll.

Tip: To show or hide data in the Contingency Table, from the red triangle menu next to Contingency Table, select the option that you want to show or hide.

Description of the Contingency Table

Count Cell frequency, margin total frequencies, and grand total (total sample size).

Total% Percent of cell counts and margin totals to the grand total.

Row% Percent of each cell count to its row total.

Col% Percent of each cell count to its column total.

Expected Expected frequency (E) of each cell under the assumption of independence. Computed as the product of the corresponding row total and column total divided by the grand total.

Deviation Observed cell frequency (O) minus the expected cell frequency (E).

Cell Chi Square Chi-square values computed for each cell as $(O - E)^2 / E$.

Col Cum Cumulative column total.

Col Cum% Cumulative column percentage.

Row Cum Cumulative row total.

Row Cum% Cumulative row percentage.

Tests

The Tests report shows the results for two tests to determine whether the response level rates are the same across X levels.

To produce the report shown in Figure 7.8, follow the instructions in "Example of Contingency Analysis" on page 210.

Figure 7.8 Example of a Tests Report

Tests

N	DF	-LogLike	RSquare (U)
303	4	36.309616	0.1200

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	72.619	<.0001*
Pearson	66.313	<.0001*

Note the following about the Chi-square statistics:

- When both categorical variables are responses (Y variables), the Chi-square statistics test that they are independent.
- You might have a Y variable with a fixed X variable. In this case, the Chi-square statistics test that the distribution of the Y variable is the same across each X level.

Description of the Tests Report

N Total number of observations.

DF Records the degrees of freedom associated with the test. The degrees of freedom are equal to $(c - 1)(r - 1)$, where c is the number of columns and r is the number of rows.

-LogLike Negative log-likelihood, which measures fit and uncertainty (much like sums of squares in continuous response situations).

Rsquare (U) Portion of the total uncertainty attributed to the model fit.

- An R^2 of 1 means that the factors completely predict the categorical response.
- An R^2 of 0 means that there is no gain from using the model instead of fixed background response rates.

For more information, see "Tests Report" on page 236.

Test Lists two Chi-square statistical tests of the hypothesis that the response rates are the same in each sample category. For more information, see "Tests Report" on page 236.

Prob>ChiSq Lists the probability of obtaining, by chance alone, a Chi-square value greater than the one computed if no relationship exists between the response and factor. If both variables have only two levels, Fisher's exact probabilities for the one-tailed tests and the two-tailed test also appear.

Fisher's Exact Test

This report gives the results of Fisher's exact test for a 2x2 table. The results appear automatically for 2x2 tables. For more details about Fisher's exact test, and for details about the test for *r* x *c* tables, see "Exact Test" on page 225.

Analysis of Means for Proportions

If the response has two levels, you can use this option to compare response proportions for the X levels to the overall response proportion. This method uses the normal approximation to the binomial. Therefore, if the sample sizes are too small, a warning appears in the results.

Note: For a description of Analysis of Means methods, see the document by Nelson, Wludyka, and Copeland (2005).

Set Alpha Level Selects the alpha level used in the analysis.

Show Summary Report Produces a report that shows the response proportions with decision limits for each level of the X variable. The report indicates whether a limit has been exceeded.

Switch Response Level for Proportion Changes the response category used in the analysis.

Display Options Shows or hides the decision limits, decision limit shading, center line, and point options.

Related Information

- "Example of Analysis of Means for Proportions" on page 225

Correspondence Analysis

Correspondence analysis is a graphical technique to show which rows or columns of a frequency table have similar patterns of counts. In the correspondence analysis plot, there is a

point for each row and for each column. Use Correspondence Analysis when you have many levels, making it difficult to derive useful information from the mosaic plot.

Understanding Correspondence Analysis Plots

The *row profile* can be defined as the set of rowwise rates, or in other words, the counts in a row divided by the total count for that row. If two rows have very similar row profiles, their points in the correspondence analysis plot are close together. Squared distances between row points are approximately proportional to Chi-square distances that test the homogeneity between the pair of rows.

Column and row profiles are alike because the problem is defined symmetrically. The distance between a row point and a column point has no meaning. However, the directions of columns and rows from the origin are meaningful, and the relationships help interpret the plot.

Correspondence Analysis Options

Use the options in the red triangle menu next to Correspondence Analysis to produce a 3-D scatterplot and add column properties to the data table.

3D Correspondence Analysis Produces a 3-D scatterplot.

Save Value Ordering Takes the order of the levels sorted by the first correspondence score coefficient and makes a column property for both the X and Y columns.

The Details Report

The Details report contains statistical information about the correspondence analysis and shows the values used in the plot.

Singular Value Provides the singular value decomposition of the contingency table. For the formula, see "Details Report in Correspondence Analysis" on page 237.

Inertia Lists the square of the singular values, reflecting the relative variation accounted for in the canonical dimensions.

Portion Portion of inertia with respect to the total inertia.

Cumulative Shows the cumulative portion of inertia. If the first two singular values capture the bulk of the inertia, then the 2-D correspondence analysis plot is sufficient to show the relationships in the table.

X variable c1, c2, c3 The values plotted on the Correspondence Analysis plot (Figure 7.11).

Y variable c1, c2, c3 The values plotted on the Correspondence Analysis plot (Figure 7.11).

Related Information

- "Example of Correspondence Analysis" on page 226

Cochran-Mantel-Haenszel Test

The Cochran-Mantel-Haenszel test discovers if there is a relationship between two categorical variables after blocking across a third classification.

Correlation of Scores Applicable when both Y or X are ordinal or interval. The alternative hypothesis is that there is a linear association between Y and X in at least one level of the blocking variable.

Row Score by Col Categories Applicable when Y is ordinal or interval. The alternative hypothesis is that, for at least one level of the blocking variable, the mean scores of the *r* rows are unequal.

Col Score by Row Categories Applicable when X is ordinal or interval. The alternative hypothesis is that, for at least one level of the blocking variable, the mean scores of the *c* columns are unequal.

General Assoc. of Categories Tests that for at least one level of the blocking variable, there is some type of association between X and Y.

Related Information

- "Example of a Cochran Mantel Haenszel Test" on page 229

Agreement Statistic

When the two variables have the same levels, the **Agreement Statistic** option is available. This option shows the Kappa statistic (Agresti 1990), its standard error, confidence interval, hypothesis test, and Bowker's test of symmetry.

The Kappa statistic and associated p-value given in this section are approximate. An exact version of the agreement test is available. See "Exact Test" on page 225.

Kappa Shows the Kappa statistic.

Std Err Shows the standard error of the Kappa statistic.

Lower 95% Shows the lower endpoint of the confidence interval for Kappa.

Upper 95% Shows the upper endpoint of the confidence interval for Kappa.

Prob>Z Shows the p-value for a one-sided test for Kappa. The null hypothesis tests if Kappa equals zero.

Prob>|Z| Shows the p-value for a two-sided test for Kappa.

ChiSquare Shows the test statistic for Bowker's test. For Bowker's test of symmetry, the null hypothesis is that the probabilities in the square table satisfy symmetry, or that $p_{ij}=p_{ji}$ for all pairs of table cells. When both X and Y have two levels, this test is equal to McNemar's test.

Prob>ChiSq Shows the p-value for the Bowker's test.

Related Information

- "Example of the Agreement Statistic Option" on page 230
- "Agreement Statistic Option" on page 235

Relative Risk

Calculate risk ratios for 2x2 contingency tables using the **Relative Risk** option. Confidence intervals also appear in the report. You can find more information about this method in Agresti (1990) section 3.4.2.

The Choose Relative Risk Categories window appears when you select the **Relative Risk** option. You can select a single response and factor combination, or you can calculate the risk ratios for all combinations of response and factor levels.

Related Information

- "Example of the Relative Risk Option" on page 231

Two Sample Test for Proportions

When both the X and Y variables have two levels, you can request a confidence interval for a difference between two proportions. It also outputs the test corresponding to the confidence interval.

Description Shows the test being performed.

Proportion Difference Shows the difference in the proportions between the levels of the X variable.

Lower 95% Shows the lower endpoint of the confidence interval for the difference. Based on the adjusted Wald confidence interval.

Upper 95% Shows the upper endpoint of the confidence interval for the difference. Based on the adjusted Wald confidence interval.

Adjusted Wald Test Shows two-tailed and one-tailed tests.

Prob Shows the p-values for the tests.

Response <variable> category of interest Select which response level to use in the test.

Related Information

- "Example of a Two Sample Test for Proportions" on page 233

Measures of Association

You can request several statistics that describe the association between the variables in the contingency table by selecting the **Measures of Association** option.

Gamma Based on the number of concordant and discordant pairs and ignores tied pairs. Takes values in the range -1 to 1.

Kendall's Tau-b Similar to Gamma and uses a correction for ties. Takes values in the range -1 to 1.

Stuart's Tau-c Similar to Gamma and uses an adjustment for table size and a correction for ties. Takes values in the range -1 to 1.

Somers' D An asymmetric modification of Tau-b.

- The C|R denotes that the row variable X is regarded as an independent variable and the column variable Y is regarded as dependent.
- Similarly, the R|C denotes that the column variable Y is regarded as an independent variable and the row variable X is dependent.

Somers' D differs from Tau-b in that it uses a correction for ties only when the pair is tied on the independent variable. It takes values in the range -1 to 1.

Lambda Asymmetric Differs for C|R and R|C.

- For C|R, is interpreted as the probable improvement in predicting the column variable Y given knowledge of the row variable X.
- For R|C, is interpreted as the probable improvement in predicting the row variable X given knowledge about the column variable Y.

Takes values in the range 0 to 1.

Lambda Symmetric Loosely interpreted as the average of the two Lambda Asymmetric measures. Takes values in the range 0 to 1.

Uncertainty Coef –For C|R, is the proportion of uncertainty in the column variable Y that is explained by the row variable X.

- For R|C, is interpreted as the proportion of uncertainty in the row variable X that is explained by the column variable Y.

Takes values in the range 0 to 1.

Uncertainty Coef Symmetric Symmetric version of the two Uncertainty Coef measures. Takes values in the range 0 to 1.

Notes:

Each statistic appears with its standard error and confidence interval.

- Gamma, Kendall's Tau-b, Stuart's Tau-c, and Somers' D are measures of ordinal association that consider whether the variable Y tends to increase as X increases. They classify pairs of observations as concordant or discordant. A pair is concordant if an observation with a larger value of X also has a larger value of Y. A pair is discordant if an observation with a larger value of X has a smaller value of Y. These measures are appropriate only when both variables are ordinal.
- The Lambda and Uncertainty measures are appropriate for ordinal and nominal variables.

For details about measures of association, see the following references:

- Brown and Benedetti (1977)
- Goodman and Kruskal (1979)
- Kendall and Stuart (1979)
- Snedecor and Cochran (1980)
- Somers (1962)

Related Information

- "Example of the Measures of Association Option" on page 234

Cochran Armitage Trend Test

This Cochran Armitage Trend tests for trends in binomial proportions across the levels of a single variable. This test is appropriate only when one variable has two levels and the other variable is ordinal. The two-level variable represents the response, and the other represents an explanatory variable with ordered levels. The null hypothesis is the hypothesis of no trend, which means that the binomial proportion is the same for all levels of the explanatory variable.

The test statistic and p-values given in this section are approximate. An exact version of the trend test is available. See "Exact Test" on page 225.

Related Information

- "Example of the Cochran Armitage Trend Test" on page 234

Exact Test

JMP PRO The following Exact tests are available in the Contingency platform:

Fisher's Exact Test Performs Fisher's Exact test for an *r* x *c* table. This is a test for association between two variables. Fisher's exact test assumes that the row and column totals are fixed, and uses the hypergeometric distribution to compute probabilities.

This test does not depend on any large-sample distribution assumptions. This means it is appropriate for situations where the Likelihood Ratio and Pearson tests become less reliable, like for small sample sizes or sparse tables.

The report includes the following information:

Table Probability (P) Gives the probability for the observed table. This is not the p-value for the test.

Two-sided Prob ≤ P Gives the p-value for the two-sided test.

For 2x2 tables, the Fisher's Exact test is automatically performed. See "Tests" on page 218.

Exact Cochran Armitage Trend Test Performs the exact version of the Cochran Armitage Trend Test. This test is available only when one of the variables has two levels. For more details about the trend test, see "Cochran Armitage Trend Test" on page 224.

Exact Agreement Test Performs an exact test for testing agreement between variables. This is an exact test for the Kappa statistic. This is available only when the two variables have the same levels. For more details about agreement testing, see "Agreement Statistic" on page 221.

Additional Examples of the Contingency Platform

This section contains additional examples using the options in the Contingency platform.

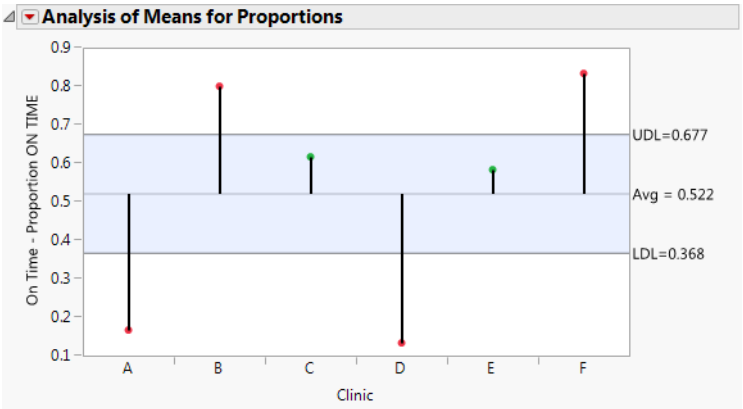
Example of Analysis of Means for Proportions

This example uses the Office Visits.jmp sample data table, which records late and on-time appointments for six clinics in a geographic region. 60 random appointments were selected from 1 week of records for each of the six clinics. To be considered on-time, the patient must be taken to an exam room within five minutes of their scheduled appointment time. Examine the proportion of patients that arrived on-time to their appointment.

1. Select **Help > Sample Data Library** and open Office Visits.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select On Time and click **Y, Response**.
4. Select Clinic and click **X, Factor**.

5. Select Frequency and click **Freq**.
6. Click **OK**.
7. From the red triangle menu next to Contingency Analysis, select **Analysis of Means for Proportions**.
8. From the red triangle menu next to Analysis of Means for Proportions, select **Show Summary Report** and **Switch Response Level for Proportion**.

Figure 7.9 Example of Analysis of Means for Proportions



Analysis of Means for Proportions Summary

Level	Group N	Lower Limit	Group Proportion	Upper Limit	Limit Exceeded
A	60	0.36791	0.166667	0.676534	Lower
B	60	0.36791	0.8	0.676534	Upper
C	60	0.36791	0.616667	0.676534	
D	60	0.36791	0.133333	0.676534	Lower
E	60	0.36791	0.583333	0.676534	
F	60	0.36791	0.833333	0.676534	Upper

Figure 7.9 shows the proportion of patients who were on-time from each clinic. From Figure 7.9, notice the following:

- The proportion of on-time arrivals is the highest for clinic F, followed by clinic B.
- Clinic D has the lowest proportion of on-time arrivals, followed by clinic A.
- Clinic E and clinic C are close to the average, and do not exceed the decision limits.

Example of Correspondence Analysis

This example uses the Cheese.jmp sample data table, which is taken from the Newell cheese tasting experiment, reported in McCullagh and Nelder (1989). The experiment records counts more than nine different response levels across four different cheese additives.

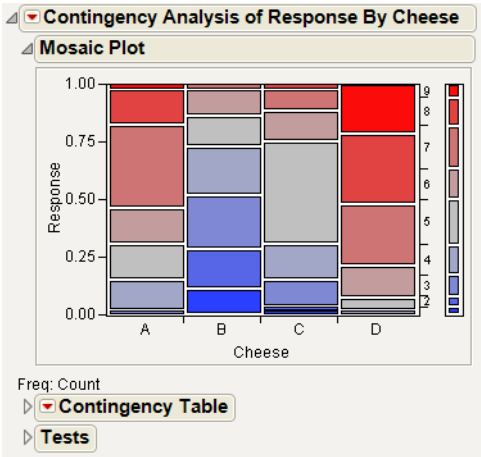
1. Select **Help > Sample Data Library** and open Cheese.jmp.

2. Select **Analyze > Fit Y by X**.
3. Select Response and click **Y, Response**.

   The Response values range from one to nine, where one is the least liked, and nine is the best liked.
4. Select Cheese and click **X, Factor**.

   A, B, C, and D represent four different cheese additives.
5. Select Count and click **Freq**.
6. Click **OK**.

Figure 7.10 Mosaic Plot for the Cheese Data



From the mosaic plot in Figure 7.10, you notice that the distributions do not appear alike. However, it is challenging to make sense of the mosaic plot across nine levels. A correspondence analysis can help define relationships in this type of situation.

7. To see the correspondence analysis plot, from the red triangle menu next to Contingency Analysis, select **Correspondence Analysis**.

Figure 7.11 Example of a Correspondence Analysis Plot

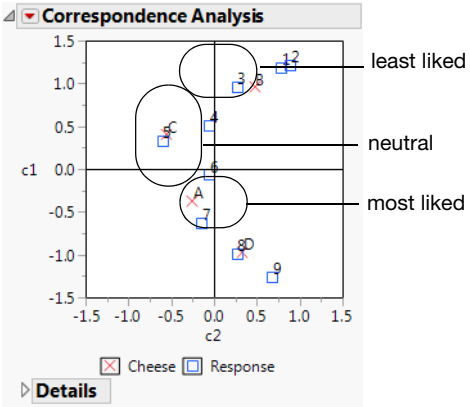
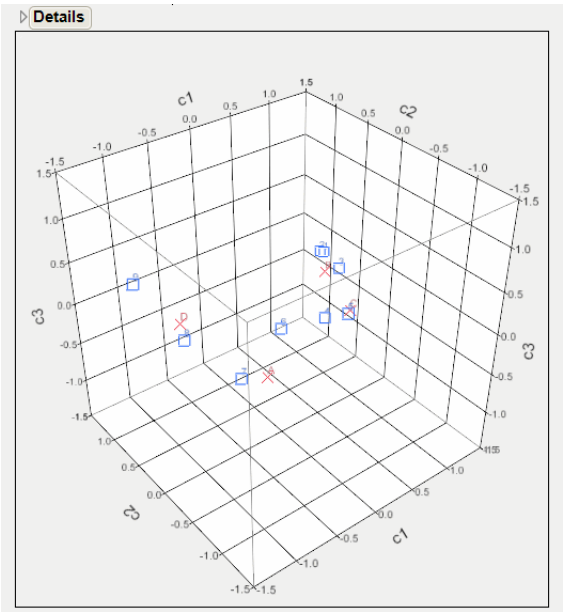


Figure 7.11 shows the correspondence analysis graphically, with the plot axes labeled c1 and c2. Notice the following:

- c1 seems to correspond to a general satisfaction level. The cheeses on the c1 axis go from least liked at the top to most liked at the bottom.
- Cheese D is the most liked cheese, with responses of 8 and 9.
- Cheese B is the least liked cheese, with responses of 1,2, and 3.
- Cheeses C and A are in the middle, with responses of 4,5,6, and 7.

8. From the red triangle menu next to Correspondence Analysis, select **3D Correspondence Analysis**.

Figure 7.12 Example of a 3-D Scatterplot



From Figure 7.12, notice the following:

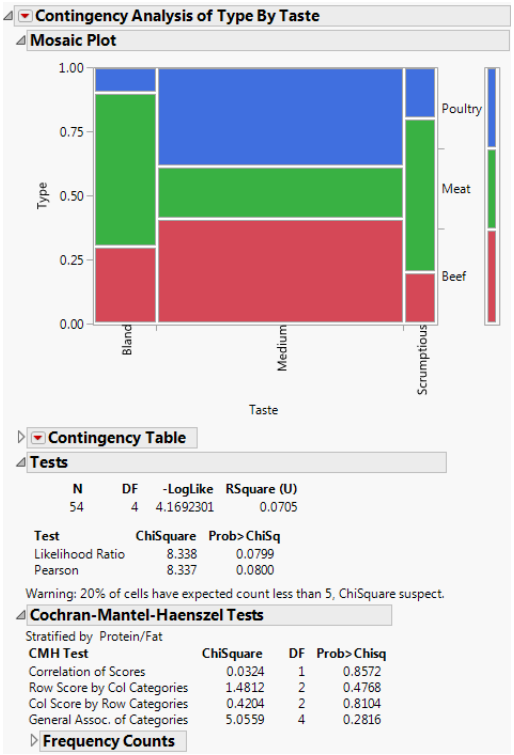
- Looking at the c1 axis, responses 1 through 5 appear to the right of 0 (positive). Responses 6 through 9 appear to the left of 0 (negative).
- Looking at the c2 axis, A and C appear to the right of 0 (positive). B and D appear to the left of 0 (negative).
- You can conclude that c1 corresponds to the general satisfaction (from least to most liked).

Example of a Cochran Mantel Haenszel Test

This example uses the Hot Dogs.jmp sample data table. Examine the relationship between hot dog type and taste.

1. Select **Help > Sample Data Library** and open Hot Dogs.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Type and click **Y, Response**.
4. Select Taste and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu next to Contingency Analysis, select **Cochran Mantel Haenszel**.
7. Select Protein/Fat as the grouping variable and click **OK**.

Figure 7.13 Example of a Cochran-Mantel-Haenszel Test



From Figure 7.13, you notice the following:

- The Tests report shows a marginally significant Chi-square probability of about 0.0799, indicating some significance in the relationship between hot dog taste and type.
- The Cochran Mantel Haenszel report shows that the p-value for the general association of categories is 0.2816, which is much larger than 5%.

Example of the Agreement Statistic Option

This example uses the Attribute Gauge.jmp sample data table. The data gives results from three people (raters) rating fifty parts three times each. Examine the relationship between raters A and B.

1. Select **Help > Sample Data Library** and open Attribute Gauge.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select A and click **Y, Response**.
4. Select B and click **X, Factor**.
5. Click **OK**.

6. From the red triangle menu next to Contingency Analysis, select **Agreement Statistic**.

Figure 7.14 Example of the Agreement Statistic Report

Contingency Analysis of A By B

Mosaic Plot

Contingency Table

B \ A	0	1	Total
Count Total % Col % Row %			
0	44 29.33 88.00 93.62	3 2.00 3.00 6.38	47 31.33
1	6 4.00 12.00 5.83	97 64.67 97.00 94.17	103 68.67
Total	50 33.33	100 66.67	150

Tests

Agreement Statistic

Kappa Coefficient

Degree of Agreement	Kappa	Std Err	Lower 95%	Upper 95%
	0.862944	0.044198	0.776317	0.949571

Asymptotic Test	Prob > z	Prob>|Z|
	<.0001*	<.0001*

Bowker's Test

Symmetry of Disagreement	ChiSquare	Prob>ChiSq
	1	0.3173

For 2-by-2 tables, Bowker's Test is equivalent to McNemar's Test.

From Figure 7.14, you notice that the agreement statistic of 0.86 is high (close to 1) and the p-value of <.0001 is small. This reinforces the high agreement seen by looking at the diagonal of the contingency table. Agreement between the raters occurs when both raters give a rating of 0 or both give a rating of 1.

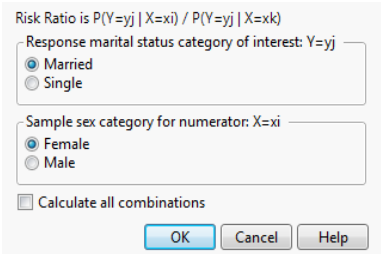
Example of the Relative Risk Option

This example uses the Car Poll.jmp sample data table. Examine the relative probabilities of being married and single for the participants in the poll.

1. Select **Help > Sample Data Library** and open Car Poll.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select marital status and click **Y, Response**.
4. Select sex and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu next to Contingency Analysis, select **Relative Risk**.

   The Choose Relative Risk Categories window appears.

Figure 7.15 The Choose Relative Risk Categories Window



Note the following about the Choose Relative Risk Categories window:

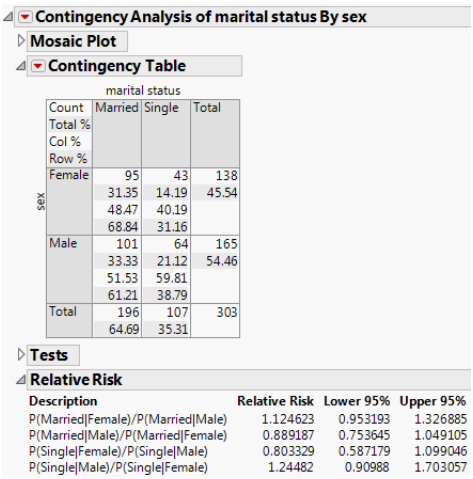
- If you are interested in only a single response and factor combination, you can select that here. For example, if you clicked **OK** in the window in Figure 7.15, the calculation would be as follows:

$$\frac{P(Y = \text{Married} | X = \text{Female})}{P(Y = \text{Married} | X = \text{Male})}$$

- If you would like to calculate the risk ratios for all ($2 \times 2 = 4$) combinations of response and factor levels, select the **Calculate All Combinations** check box. See Figure 7.16.

7. Ask for all combinations by selecting the **Calculate All Combinations** check box. Leave all other default selections as is.

Figure 7.16 Example of the Risk Ratio Report



Contingency Analysis of marital status By sex

Mosaic Plot

Contingency Table

marital status

sex	Count / Total % / Col % / Row %	Married	Single	Total
	Female	95	43	138
		31.35	14.19	45.54
		48.47	40.19	
		68.84	31.16	
	Male	101	64	165
		33.33	21.12	54.46
		51.53	59.81	
		61.21	38.79	
	Total	196	107	303
		64.69	35.31	

Tests

Relative Risk

Description	Relative Risk	Lower 95%	Upper 95%
P(Married|Female)/P(Married|Male)	1.124623	0.953193	1.326885
P(Married|Male)/P(Married|Female)	0.889187	0.753645	1.049105
P(Single|Female)/P(Single|Male)	0.803329	0.587179	1.099046
P(Single|Male)/P(Single|Female)	1.24482	0.90988	1.703057

To see how the relative risk is calculated, proceed as follows:

1. Examine the first entry in the Relative Risk report, which is P(Married | Female)/P(Married | Male).

2. You can find these probabilities in the Contingency Table. Since the probabilities are computed based on two levels of sex, which differs across the rows of the table, use the Row% to read the probabilities, as follows:

   P(Married | Female)=0.6884

   P(Married | Male) = 0.6121

   Therefore, the calculations are as follows:

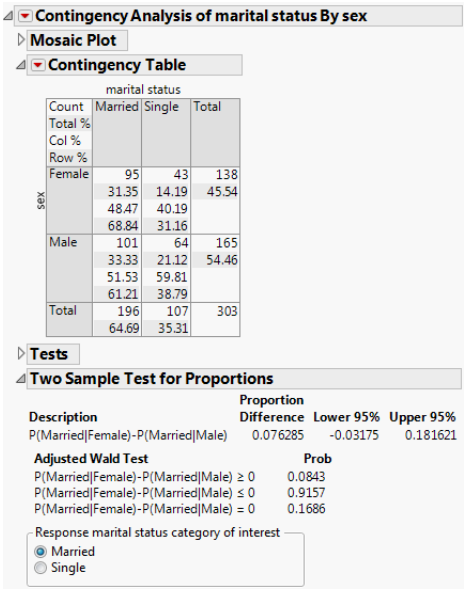
   $$P(\text{Married} \mid \text{Female})/P(\text{Married} \mid \text{Male}) = \frac{0.6884}{0.6121} = 1.1247$$

Example of a Two Sample Test for Proportions

This example uses the Car Poll.jmp sample data table. Examine the probability of being married for males and females.

1. Select **Help > Sample Data Library** and open Car Poll.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select marital status and click **Y, Response**.
4. Select sex and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu next to Contingency Analysis, select **Two Sample Test for Proportions**.

Figure 7.17 Example of the Two Sample Test for Proportions Report



Contingency Analysis of marital status By sex

Mosaic Plot

Contingency Table

sex	Count / Total % / Col % / Row %	Married	Single	Total
		marital status		
Female		95	43	138
		31.35	14.19	45.54
		48.47	40.19	
		68.84	31.16	
Male		101	64	165
		33.33	21.12	54.46
		51.53	59.81	
		61.21	38.79	
Total		196	107	303
		64.69	35.31	

Tests

Two Sample Test for Proportions

Description	Proportion Difference	Lower 95%	Upper 95%
P(Married|Female)-P(Married|Male)	0.076285	-0.03175	0.181621

Adjusted Wald Test	Prob
P(Married|Female)-P(Married|Male) ≥ 0	0.0843
P(Married|Female)-P(Married|Male) ≤ 0	0.9157
P(Married|Female)-P(Married|Male) = 0	0.1686

Response marital status category of interest
- (●) Married
- ( ) Single

In this example, you are comparing the probability of being married between females and males. See the Row% in the Contingency Table to obtain the following:

P(Married | Female)=0.6884

P(Married | Male) = 0.6121

The difference between these two numbers, 0.0763, is the Proportion Difference shown in the report. The two-sided confidence interval is [-0.03175, 0.181621]. The p-value by the adjusted Wald method corresponding to the confidence interval is 0.1686, which is close to the p-value (0.1665) by Pearson's Chi-square test. Generally, Pearson's Chi-square test is more popular than the modified Wald's test for testing the difference of two proportions.

Example of the Measures of Association Option

This example uses the Car Poll.jmp sample data table. Examine the probability of being married for males and females.

1. Select **Help > Sample Data Library** and open Car Poll.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select marital status and click **Y, Response**.
4. Select sex and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu next to Contingency Analysis, select **Measures of Association**.

Figure 7.18 Example of the Measures of Association Report

Contingency Analysis of marital status By sex

Mosaic Plot

Contingency Table

Tests

Measures of Association

Measure	Value	Std Error	Lower 95%	Upper 95%
Gamma	0.1667	0.1184	-0.0654	0.3987
Kendall's Tau-b	0.0795	0.0570	-0.0321	0.1911
Stuart's Tau-c	0.0757	0.0543	-0.0307	0.1821
Somers' D C|R	0.0763	0.0547	-0.0309	0.1835
Somers' D R|C	0.0828	0.0593	-0.0335	0.1991
Lambda Asymmetric C|R	0.0000	0.0000	0.0000	0.0000
Lambda Asymmetric R|C	0.0000	0.0000	0.0000	0.0000
Lambda Symmetric	0.0000	0.0000	0.0000	0.0000
Uncertainty Coef C|R	0.0049	0.0070	0.0000	0.0186
Uncertainty Coef R|C	0.0046	0.0066	0.0000	0.0176
Uncertainty Coef Symmetric	0.0047	0.0068	0.0000	0.0181

Since the variables that you want to examine (sex and marital status) are nominal, use the Lambda and Uncertainty measures. All of them are small, so it seems that there is a weak association between sex and marital status.

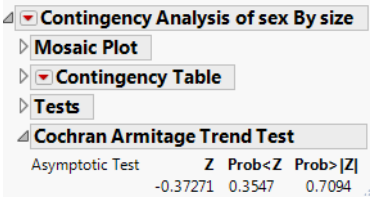
Example of the Cochran Armitage Trend Test

1. Select **Help > Sample Data Library** and open Car Poll.jmp.

For the purposes of this test, change size to an ordinal variable:

2. In the Columns panel, right-click on the icon next to size and select **Ordinal**.
3. Select **Analyze > Fit Y by X**.
4. Select sex and click **Y, Response**.
5. Select size and click **X, Factor**.
6. Click **OK**.
7. From the red triangle menu next to Contingency Analysis, select **Cochran Armitage Trend Test**.

Figure 7.19 Example of the Cochran Armitage Trend Test Report



The two-sided p-value (0.7094) is large. From this, you cannot conclude that there is a relationship in the proportion of male and females that purchase different sizes of cars.

Statistical Details for the Contingency Platform

This section contains statistical details for selected options and reports in the Contingency platform.

Agreement Statistic Option

Viewing the two response variables as two independent ratings of the n subjects, the Kappa coefficient equals +1 when there is complete agreement of the raters. When the observed agreement exceeds chance agreement, the Kappa coefficient is positive, with its magnitude reflecting the strength of agreement. Although unusual in practice, Kappa is negative when the observed agreement is less than chance agreement. The minimum value of Kappa is between -1 and 0, depending on the marginal proportions.

Quantities associated with the Kappa statistic are computed as follows:

$$\hat{\kappa} = \frac{P_0 - P_e}{1 - P_e} \text{ where } P_0 = \sum_i p_{ii} \text{ and } P_e = \sum_i p_{i.} p_{.i}$$

The asymptotic variance of the simple kappa coefficient is estimated by the following:

$$\text{var} = \frac{A+B-C}{(1-P_e)^2 n} \text{ where } A = \sum_i p_{ii}[1-(p_{i.}+p_{.i})(1-\hat{\kappa})]^2,\ B = (1-\hat{\kappa})^2 \sum_{i \neq j}\sum p_{ij}(p_{.i}+p_{j.})^2 \text{ and}$$

$$C = [\hat{\kappa} - P_e(1-\hat{\kappa})]^2$$

See Fleiss, Cohen, and Everitt (1969).

For Bowker's test of symmetry, the null hypothesis is that the probabilities in the two-by-two table satisfy symmetry ($p_{ij}=p_{ji}$).

Odds Ratio Option

The Odds Ratio is calculated as follows:

$$\frac{p_{11} \times p_{22}}{p_{12} \times p_{21}}$$

where p_{ij} is the count in the ith row and j^{th} column of the 2x2 table.

Tests Report

Rsquare (U)

Rsquare (U) is computed as follows:

$$\frac{-\text{log likelihood for Model}}{-\text{log likelihood for Corrected Total}}$$

The total negative log-likelihood is found by fitting fixed response rates across the total sample.

Test

The two Chi-square tests are as follows:

The **Likelihood Ratio** Chi-square test is computed as twice the negative log-likelihood for Model in the Tests table. Some books use the notation G^2 for this statistic. The difference of two negative log-likelihoods, one with *whole-population* response probabilities and one with *each-population* response rates, is written as follows:

$$G^2 = 2\left[\sum_{ij}(-n_{ij})\ln(p_j) - \sum_{ij} -n_{ij}\ln(p_{ij})\right] \text{ where } p_{ij} = \frac{n_{ij}}{N} \text{ and } p_j = \frac{N_j}{N}$$

This formula can be more compactly written as follows:

$$G^2 = 2\sum_i \sum_j n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right)$$

The **Pearson** Chi-square is calculated by summing the squares of the differences between the observed and expected cell counts. The Pearson Chi-square exploits the property that frequency counts tend to a normal distribution in very large samples. The familiar form of this Chi-square statistic is as follows:

$$\chi^2 = \sum \frac{(\mathrm{O} - \mathrm{E})^2}{\mathrm{E}}$$

where O is the observed cell counts and E is the expected cell counts. The summation is over all cells. There is no continuity correction done here, as is sometimes done in 2-by-2 tables.

Details Report in Correspondence Analysis

Lists the singular values of the following equation:

$$D_r^{-0.5}(\mathrm{P} - rc')\mathrm{D}_c^{-0.5}$$

where:

- P is the matrix of counts divided by the total frequency
- r and c are row and column sums of P
- the Ds are diagonal matrices of the values of r and c

Chapter 8

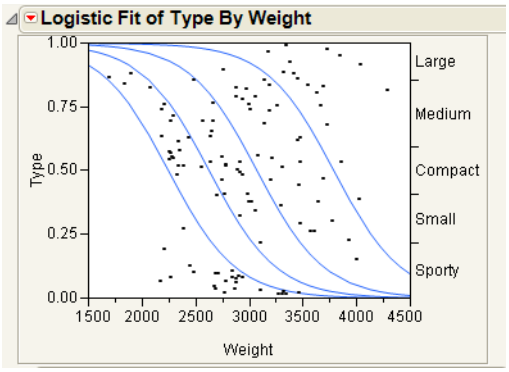
Logistic Analysis

Examine Relationships between a Categorical Y and a Continuous X Variable

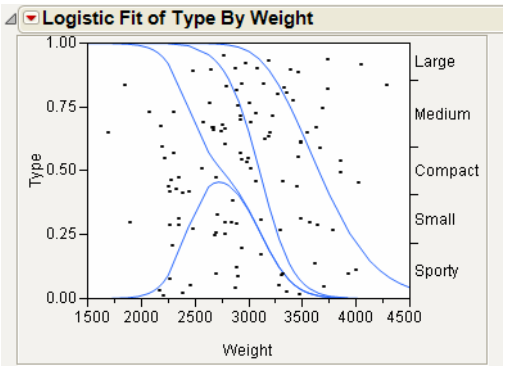
The Logistic platform fits the probabilities for response categories to a continuous *x* predictor. The fitted model estimates probabilities for each *x* value. The Logistic platform is the *nominal* or *ordinal* by *continuous* personality of the Fit Y by X platform. There is a distinction between nominal and ordinal responses on this platform:

- Nominal logistic regression estimates a set of curves to partition the probability among the responses.
- Ordinal logistic regression models the probability of being less than or equal to a given response. This has the effect of estimating a single logistic curve, which is shifted horizontally to produce probabilities for the ordered categories. This model is less complex and is recommended for ordered responses.

Figure 8.1 Examples of Logistic Regression



Ordinal Logistic Regression



Nominal Logistic Regression

Overview of Logistic Regression

Logistic regression has a long tradition with widely varying applications such as modeling dose-response data and purchase-choice data. Unfortunately, many introductory statistics courses do not cover this fairly simple method. Many texts in categorical statistics cover it (Agresti 1998), in addition to texts on logistic regression (Hosmer and Lemeshow 1989). Some analysts use the method with a different distribution function, the normal. In that case, it is called *probit analysis*. Some analysts use discriminant analysis instead of logistic regression because they prefer to think of the continuous variables as *Y*s and the categories as *X*s and work backwards. However, discriminant analysis assumes that the continuous data are normally distributed random responses, rather than fixed regressors.

Simple logistic regression is a more graphical and simplified version of the general facility for categorical responses in the Fit Model platform. For examples of more complex logistic regression models, see the Logistic Regression chapter in the *Fitting Linear Models* book.

Nominal Logistic Regression

Nominal logistic regression estimates the probability of choosing one of the response levels as a smooth function of the x factor. The fitted probabilities must be between 0 and 1, and must sum to 1 across the response levels for a given factor value.

In a logistic probability plot, the y-axis represents probability. For k response levels, k - 1 smooth curves partition the total probability (which equals 1) among the response levels. The fitting principle for a logistic regression minimizes the sum of the negative natural logarithms of the probabilities fitted to the response events that occur (that is, maximum likelihood).

Ordinal Logistic Regression

When Y is ordinal, a modified version of logistic regression is used for fitting. The cumulative probability of being at or below each response level is modeled by a curve. The curves are the same for each level except that they are shifted to the right or left.

The ordinal logistic model fits a different intercept, but the same slope, for each of r - 1 cumulative logistic comparisons, where r is the number of response levels. Each parameter estimate can be examined and tested individually, although this is seldom of much interest.

The ordinal model is preferred to the nominal model when it is appropriate because it has fewer parameters to estimate. In fact, it is practical to fit ordinal responses with hundreds of response levels.

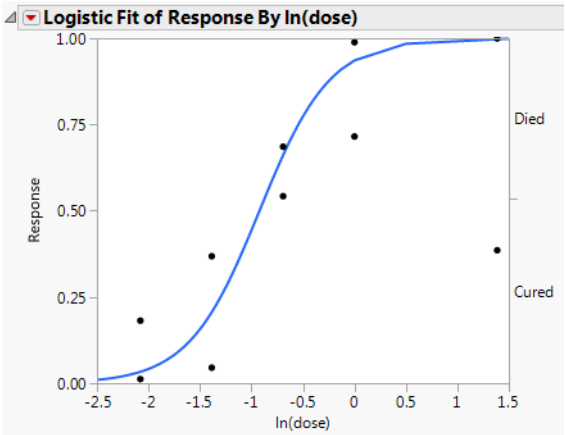
Example of Nominal Logistic Regression

This example uses the Penicillin.jmp sample data table. The data in this example comes from an experiment where 5 groups, each containing 12 rabbits, were injected with streptococcus bacteria. Once the rabbits were confirmed to have the bacteria in their system, they were given different doses of penicillin. You want to find out whether the natural log (ln(dose)) of dosage amounts has any effect on whether the rabbits are cured.

1. Select **Help > Sample Data Library** and open Penicillin.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Response and click **Y, Response**.
4. Select ln(Dose) and click **X, Factor**.

   Notice that JMP automatically fills in Count for **Freq**. Count was previously assigned the role of Freq.
5. Click **OK**.

Figure 8.2 Example of Nominal Logistic Report



Freq: Count

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	18.968295	1	37.93659	<.0001*
Full	18.313369			
Reduced	37.281664			

RSquare (U)	0.5088
AICc	40.862
BIC	44.6047
Observations (or Sum Wgts)	54

Fit Details

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	2.68517103	0.8253384	10.58	0.0011*
ln(dose)	2.9096072	0.7682096	14.35	0.0002*

For log odds of Cured/Died

Covariance of Estimates

The plot shows the fitted model, which is the predicted probability of being cured, as a function of ln(dose). The p-value is significant, indicating that the dosage amounts have a significant effect on whether the rabbits are cured.

Launch the Logistic Platform

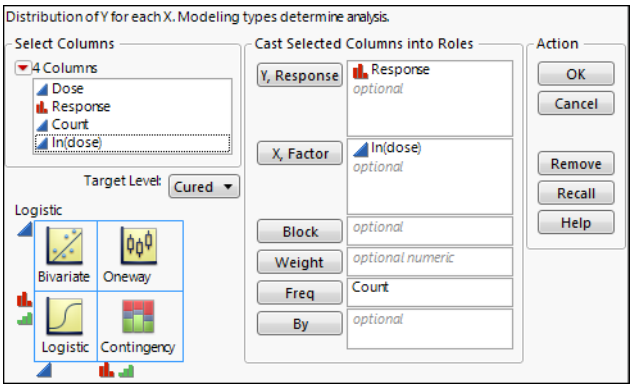
To perform a Logistic analysis, do the following:

1. Select **Analyze > Fit Y by X**.
2. Enter a nominal or ordinal column for Y, Response.
3. Enter a continuous column for X, factor.

The schematic indicates that you will be performing a logistic analysis.

Note: You can also launch a logistic analysis from the JMP Starter window. See "Launch Specific Analyses from the JMP Starter Window" on page 97.

Figure 8.3 Fit Y by X Logistic Launch Window



When the response is binary and has a nominal modeling type, a Target Level menu appears in the launch window. Use this menu to specify the level of the response whose probability you want to model.

For more information about the Fit Y by X launch window, see "Introduction to Fit Y by X" chapter on page 95.

After you click **OK**, the Logistic report appears. See "The Logistic Report" on page 243.

Data Structure

Your data can consist of unsummarized or summarized data:

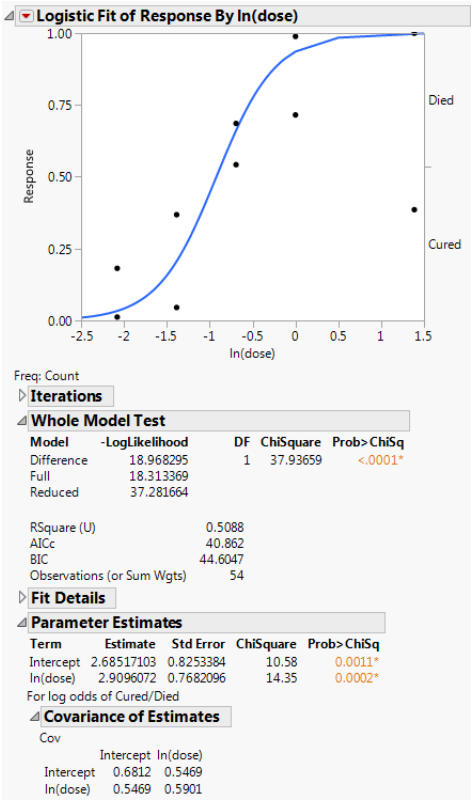
Unsummarized data There is one row for each observation containing its X value and its Y value.

Summarized data Each row represents a set of observations with common X and Y values. The data table must contain a column that gives the counts for each row. Enter this column as Freq in the launch window.

The Logistic Report

To produce the plot shown in Figure 8.4, follow the instructions in "Example of Nominal Logistic Regression" on page 241.

Figure 8.4 Example of a Logistic Report



The Logistic report window contains the Logistic plot, the Iterations report, the Whole Model Test report, and the Parameter Estimates report.

Note: The red triangle menu provides more options that can add to the initial report window. See "Logistic Platform Options" on page 247.

Logistic Plot

The logistic probability plot gives a complete picture of what the logistic model is fitting. At each x value, the probability scale in the y direction is divided up (partitioned) into probabilities for each response category. The probabilities are measured as the vertical distance between the curves, with the total across all Y category probabilities summing to 1.

Replace variables in the plot in one of two ways: swap existing variables by dragging and dropping a variable from one axis to the other axis; or, click on a variable in the Columns panel of the associated data table and drag it onto an axis.

Related Information

- "Additional Example of a Logistic Plot" on page 252

Iterations

The Iterations report shows each iteration and the evaluated criteria that determine whether the model has converged. Iterations appear only for nominal logistic regression.

Whole Model Test

The Whole Model Test report shows if the model fits better than constant response probabilities. This report is analogous to the Analysis of Variance report for a continuous response model. It is a specific likelihood ratio Chi-square test that evaluates how well the categorical model fits the data.

The negative sum of natural logs of the observed probabilities is called the negative log-likelihood (**–LogLikelihood**). The negative log-likelihood for categorical data plays the same role as sums of squares in continuous data: twice the difference in the negative log-likelihood from the model fitted by the data and the model with equal probabilities is a Chi-square statistic. This test statistic examines the hypothesis that the x variable has no effect on the responses.

Values of the **RSquare (U)** (sometimes denoted as R^2) range from 0 to 1. High R^2 values are indicative of a good model fit, and are rare in categorical models.

The Whole Model Test report contains the following columns:

Model Sometimes called Source.

- The **Reduced** model only contains an intercept.
- The **Full** model contains all of the effects as well as the intercept.
- The **Difference** is the difference of the log-likelihoods of the full and reduced models.

DF Records the degrees of freedom associated with the model.

–LogLikelihood Measures variation, sometimes called *uncertainty*, in the sample.

Full (the full model) is the negative log-likelihood (or uncertainty) calculated after fitting the model. The fitting process involves predicting response rates with a linear model and a logistic response function. This value is minimized by the fitting process.

Reduced (the reduced model) is the negative log-likelihood (or uncertainty) for the case when the probabilities are estimated by fixed background rates. This is the background uncertainty when the model has no effects.

The difference of these two negative log-likelihoods is the reduction due to fitting the model. Two times this value is the likelihood ratio Chi-square test statistic.

For more information, see "Likelihood, AICc, and BIC" the Statistical Details appendix in the *Fitting Linear Models* book.

Chi-Square The likelihood ratio Chi-square test of the hypothesis that the model fits no better than fixed response rates across the whole sample. It is twice the –LogLikelihood for the Difference Model. It is two times the difference of two negative log-likelihoods, one with whole-population response probabilities and one with each-population response rates. For more information, see "Whole Model Test Report" on page 258.

Prob>ChiSq The observed significance probability, often called the *p* value, for the Chi-square test. It is the probability of getting, by chance alone, a Chi-square value greater than the one computed. Models are often judged significant if this probability is below 0.05.

RSquare (U) The proportion of the total uncertainty that is attributed to the model fit, defined as the **Difference** negative log-likelihood value divided by the **Reduced** negative log-likelihood value. An RSquare (U) value of 1 indicates that the predicted probabilities for events that occur are equal to one: There is no uncertainty in predicted probabilities. Because certainty in the predicted probabilities is rare for logistic models, RSquare (U) tends to be small. For more information, see "Whole Model Test Report" on page 258.

Note: RSquare (U) is also know as *McFadden's pseudo R-square.*

AICc The corrected Akaike Information Criterion. See "Likelihood, AICc, and BIC" the Statistical Details appendix in the *Fitting Linear Models* book.

BIC The Bayesian Information Criterion. See "Likelihood, AICc, and BIC" the Statistical Details appendix in the *Fitting Linear Models* book.

Observations Sometimes called Sum Wgts. The total sample size used in computations. If you specified a **Weight** variable, this is the sum of the weights.

Measure The available measures of fit are as follows:

Entropy RSquare compares the log-likelihoods from the fitted model and the constant probability model.

Generalized RSquare is a measure that can be applied to general regression models. It is based on the likelihood function L and is scaled to have a maximum value of 1. The Generalized RSquare measure simplifies to the traditional RSquare for continuous normal responses in the standard least squares setting. Generalized RSquare is also known as the Nagelkerke or Craig and Uhler R^2, which is a normalized version of Cox and Snell's pseudo R^2. See Nagelkerke (1991).

Mean -Log p is the average of -log(p), where p is the fitted probability associated with the event that occurred.

RMSE is the root mean square error, where the differences are between the response and p (the fitted probability for the event that actually occurred).

Mean Abs Dev is the average of the absolute values of the differences between the response and p (the fitted probability for the event that actually occurred).

Misclassification Rate is the rate for which the response category with the highest fitted probability is not the observed category.

For Entropy RSquare and Generalized RSquare, values closer to 1 indicate a better fit. For Mean -Log p, RMSE, Mean Abs Dev, and Misclassification Rate, smaller values indicate a better fit.

Training The value of the measure of fit.

Definition The algebraic definition of the measure of fit.

Parameter Estimates

The nominal logistic model fits a parameter for the intercept and slope for each of $k-1$ logistic comparisons, where k is the number of response levels. The Parameter Estimates report lists these estimates. Each parameter estimate can be examined and tested individually, although this is seldom of much interest.

Term Lists each parameter in the logistic model. There is an intercept and a slope term for the factor at each level of the response variable, except the last level.

Estimate Lists the parameter estimates given by the logistic model.

Std Error Lists the standard error of each parameter estimate. They are used to compute the statistical tests that compare each term to zero.

Chi-Square Lists the Wald tests for the hypotheses that each of the parameters is zero. The Wald Chi-square is computed as $(\text{Estimate/Std Error})^2$.

Prob>ChiSq Lists the observed significance probabilities for the Chi-square tests.

Covariance of Estimates

Reports the estimated variances of the parameter estimates, and the estimated covariances between the parameter estimates. The square root of the variance estimates is the same as those given in the **Std Error** section.

Logistic Platform Options

Note: The Fit Group menu appears if you have specified multiple Y variables. Menu options allow you to arrange reports or order them by RSquare. See the Standard Least Squares chapter in the *Fitting Linear Models* book for more information.

Odds Ratios Adds odds ratios to the Parameter Estimates report. For more details, see the Logistic Regression chapter in the *Fitting Linear Models* book.

This option is available only for a response with two levels.

Inverse Prediction Prediction of *x* values from given *y* values. For more information, see "Inverse Prediction" on page 250.

This option is available only for a response with two levels.

Logistic Plot Shows or hides the logistic plot.

Plot Options The Plot Options menu includes the following options:

Show Points Toggles the points on or off.

Show Rate Curve Is useful only if you have several points for each *x*-value. In these cases, you get reasonable estimates of the rate at each value, and compare this rate with the fitted logistic curve. To prevent too many degenerate points, usually at zero or one, JMP only shows the rate value if there are at least three points at the *x*-value.

Line Color Enables you to pick the color of the plot curves.

ROC Curve A Receiver Operating Characteristic curve is a plot of sensitivity by (1 – specificity) for each value of x. See "ROC Curves" on page 249.

Lift Curve Produces a lift curve for the model. A lift curve shows the same information as a ROC curve, but in a way to dramatize the richness of the ordering at the beginning. The *Y*-axis shows the ratio of how rich that portion of the population is in the chosen response level compared to the rate of that response level as a whole. See the Logistic Regression chapter in the *Fitting Linear Models* book for details about lift curves.

Save Probability Formula Creates new data table columns that contain formulas. See "Save Probability Formula" on page 250.

See the JMP Reports chapter in the *Using JMP* book for more information about the following options:

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

ROC Curves

Suppose you have an x value that is a diagnostic measurement and you want to determine a threshold value of x that indicates the following:

- A condition exists if the x value is greater than the threshold.
- A condition does not exist if the x value is less than the threshold.

For example, you could measure a blood component level as a diagnostic test to predict a type of cancer. Now consider the diagnostic test as you vary the threshold and, thus, cause more or fewer false positives and false negatives. You then plot those rates. The ideal is to have a very narrow range of x criterion values that best divides true negatives and true positives. The Receiver Operating Characteristic (ROC) curve shows how rapidly this transition happens, with the goal being to have diagnostics that maximize the area under the curve.

Two standard definitions used in medicine are as follows:

- *Sensitivity*, the probability that a given x value (a test or measure) correctly predicts an existing condition. For a given x, the probability of incorrectly predicting the existence of a condition is 1 – sensitivity.
- *Specificity*, the probability that a test correctly predicts that a condition does not exist.

A ROC curve is a plot of sensitivity by (1 – specificity) for each value of x. The area under the ROC curve is a common index used to summarize the information contained in the curve.

When you do a simple logistic regression with a binary outcome, there is a platform option to request a ROC curve for that analysis. After selecting the **ROC Curve** option, a window asks you to specify which level to use as *positive*.

If a test predicted perfectly, it would have a value above which the entire abnormal population would fall and below which all normal values would fall. It would be perfectly sensitive and then pass through the point (0,1) on the grid. The closer the ROC curve comes to this ideal point, the better its discriminating ability. A test with no predictive ability produces a curve that follows the diagonal of the grid (DeLong, et al. 1988).

The ROC curve is a graphical representation of the relationship between false-positive and true-positive rates. A standard way to evaluate the relationship is with the area under the curve, shown below the plot in the report. In the plot, a yellow line is drawn at a 45 degree angle tangent to the ROC Curve. This marks a good cutoff point under the assumption that false negatives and false positives have similar costs.

Related Information

- "Example of ROC Curves" on page 254

Save Probability Formula

The **Save Probability Formula** option creates new data table columns. These data table columns save the following:

- formulas for linear combinations (typically called logits) of the x factor
- prediction formulas for the response level probabilities
- a prediction formula that gives the most likely response

Inverse Prediction

Inverse prediction is the opposite of prediction. It is the prediction of x values from given y values. But in logistic regression, instead of a y value, you have the probability attributed to one of the Y levels. This feature only works when there are two response categories (a binary response).

The Fit Model platform also has an option that gives an inverse prediction with confidence limits. The Standard Least Squares chapter in the *Fitting Linear Models* book gives more information about inverse prediction.

Related Information

- "Example of Inverse Prediction Using the Crosshair Tool" on page 255
- "Example of Inverse Prediction Using the Inverse Prediction Option" on page 256

Additional Examples of Logistic Regression

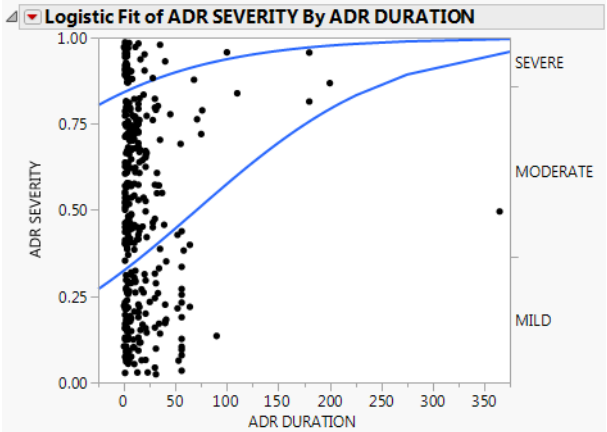
This section contains additional examples using logistic regression.

Example of Ordinal Logistic Regression

This example uses the AdverseR.jmp sample data table to illustrate an ordinal logistic regression. Suppose you want to model the severity of an adverse event as a function of treatment duration value.

1. Select **Help > Sample Data Library** and open AdverseR.jmp.
2. Right-click on the icon to the left of ADR SEVERITY and change the modeling type to ordinal.
3. Select **Analyze > Fit Y by X**.
4. Select ADR SEVERITY and click **Y, Response**.
5. Select ADR DURATION and click **X, Factor**.
6. Click **OK**.

Figure 8.5 Example of Ordinal Logistic Report



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	3.59292	1	7.185838	0.0073*
Full	350.19962			
Reduced	353.79254			

RSquare (U)	0.0102
AICc	706.467
BIC	718.032
Observations (or Sum Wgts)	357

Fit Details

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept[MILD]	-0.7315899	0.1309775	31.20	<.0001*
Intercept[MODERATE]	1.6713299	0.163161	104.93	<.0001*
ADR DURATION	0.0103692	0.0044661	5.39	0.0202*

You interpret this report the same way as the nominal report. See "The Logistic Report" on page 243.

In the plot, markers for the data are drawn at their x-coordinate. When several data points appear at the same y position, the points are jittered. That is, small spaces appear between the data points so you can see each point more clearly.

Where there are many points, the curves are pushed apart. Where there are few to no points, the curves are close together. The data pushes the curves in that way because the criterion that is maximized is the product of the probabilities fitted by the model. The fit tries to avoid points attributed to have a small probability, which are points crowded by the curves of fit. See the *Fitting Linear Models* book for more information about computational details.

For details about the Whole Model Test report and the Parameter Estimates report, see "The Logistic Report" on page 243. In the Parameter Estimates report, an intercept parameter is estimated for every response level except the last, but there is only one slope parameter. The intercept parameters show the spacing of the response levels. They always increase monotonically.

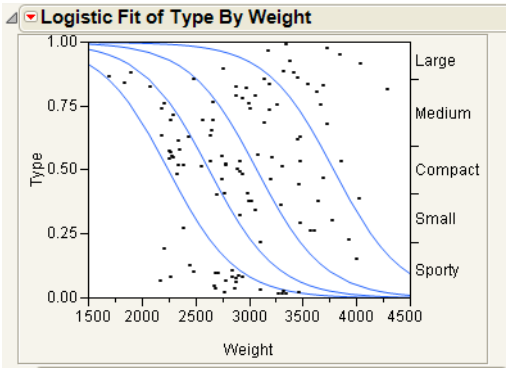
Additional Example of a Logistic Plot

This example uses the Car Physical Data.jmp sample data table to show an additional example of a logistic plot. Suppose you want to use weight to predict car size (Type) for 116 cars. Car size can be one of the following, from smallest to largest: Sporty, Small, Compact, Medium, or Large.

1. Select **Help > Sample Data Library** and open Car Physical Data.jmp.
2. In the Columns panel, right-click on the icon to the left of Type, and select **Ordinal**.
3. Right-click on Type, and select **Column Info**.
4. From the Column Properties menu, select **Value Ordering**.
5. Move the data in the following top-down order: Sporty, Small, Compact, Medium, Large.
6. Click **OK**.
7. Select **Analyze > Fit Y by X**.
8. Select Type and click **Y, Response**.
9. Select Weight and click **X, Factor**.
10. Click **OK**.

    The report window appears.

Figure 8.6 Example of Type by Weight Logistic Plot



In Figure 8.6, note the following observations:

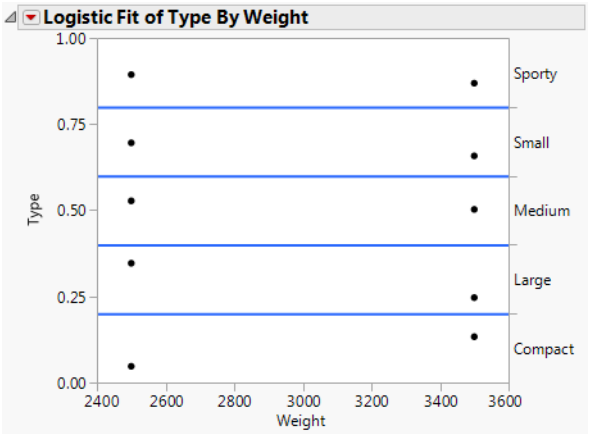
- The first (bottom) curve represents the probability that a car at a given weight is Sporty.
- The second curve represents the probability that a car is Small or Sporty. Looking only at the distance between the first and second curves corresponds to the probability of being Small.
- As you might expect, heavier cars are more likely to be Large.

- Markers for the data are drawn at their x-coordinate, with the y position jittered randomly within the range corresponding to the response category for that row.

If the x -variable has no effect on the response, then the fitted lines are horizontal and the probabilities are constant for each response across the continuous factor range. Figure 8.7 shows a logistic plot where Weight is not useful for predicting Type.

Figure 8.7 Examples of Sample Data Table and Logistic Plot Showing No y by x Relationship

	Type	Weight
1	Large	2500
2	Medium	2500
3	Compact	2500
4	Small	2500
5	Sporty	2500
6	Large	3500
7	Medium	3500
8	Compact	3500
9	Small	3500
10	Sporty	3500



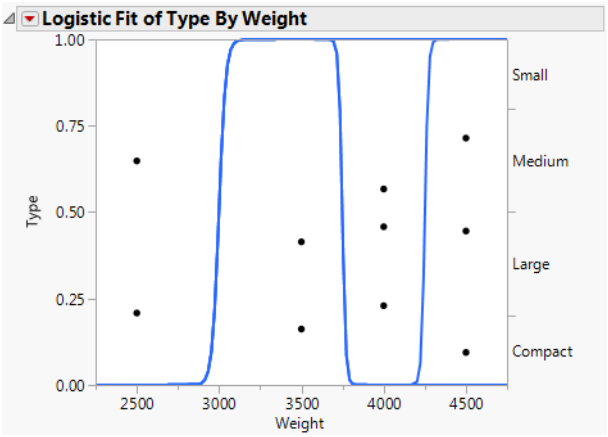
Note: To re-create the plots in Figure 8.7 and Figure 8.8, you must first create the data tables shown here, and then perform steps 7-10 at the beginning of this section.

If the response is completely predicted by the value of the factor, then the logistic curves are effectively vertical. The prediction of a response is near certain (the probability is almost 1) at each of the factor levels. Figure 8.8 shows a logistic plot where Weight almost perfectly predicts Type.

Note: In this case, the parameter estimates become very large and are marked *unstable* in the regression report.

Figure 8.8 Examples of Sample Data Table and Logistic Plot Showing an Almost Perfect *y* by *x* Relationship

	Type	Weight
1	Small	2500
2	Small	2500
3	Compact	3500
4	Compact	3500
5	Medium	4000
6	Medium	4000
7	Medium	4000
8	Large	4500
9	Large	4500
10	Large	4500



Example of ROC Curves

To demonstrate ROC curves, proceed as follows:

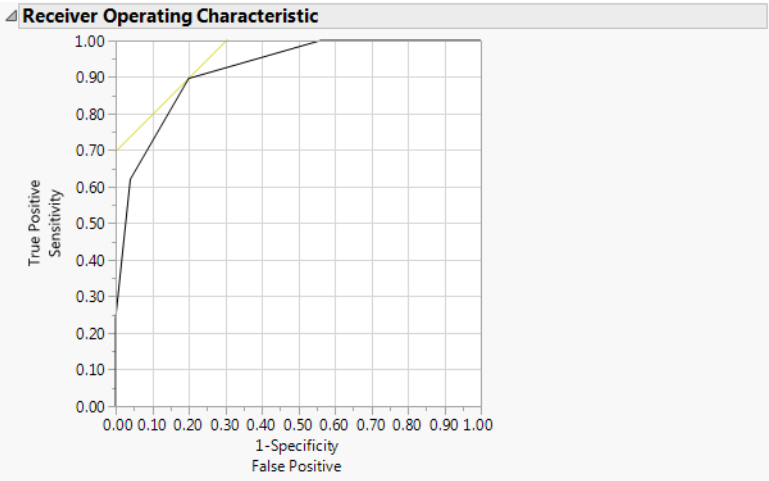
1. Select **Help > Sample Data Library** and open Penicillin.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Response and click **Y, Response**.
4. Select ln(Dose) and click **X, Factor**.

   Notice that JMP automatically fills in Count for **Freq**. Count was previously assigned the role of Freq.
5. Click **OK**.
6. From the red triangle menu, select **ROC Curve**.
7. Select Cured as the positive.
8. Click **OK**.

Note: This example shows a ROC Curve for a nominal response. For details about ordinal ROC curves, see the Partition chapter in the *Predictive and Specialized Modeling* book.

The results for the response by ln(Dose) example are shown here. The ROC curve plots the probabilities described above, for predicting response. Note that in the ROC Table, the row with the highest Sens-(1-Spec) is marked with an asterisk.

Figure 8.9 Examples of ROC Curve and Table



Using Response='Cured' to be the positive level

AUC
0.92000

ROC Table

X	Prob	1-Specificity	Sensitivity	Sens-(1-Spec)		True Pos	True Neg	False Pos	False Neg
.	.	0.0000	0.0000	0.0000		0	25	0	29
1.38629	0.9988	0.0000	0.2414	0.2414		7	25	0	22
0.00000	0.9361	0.0400	0.6207	0.5807		18	24	1	11
-0.69315	0.6611	0.2000	0.8966	0.6966	*	26	20	5	3
-1.38629	0.2061	0.5600	1.0000	0.4400		29	11	14	0
-2.07944	0.0334	1.0000	1.0000	0.0000		29	0	25	0
-2.07944	0.0334	1.0000	1.0000	0.0000		29	0	25	0

Since the ROC curve is well above a diagonal line, you conclude that the model has good predictive ability.

Example of Inverse Prediction Using the Crosshair Tool

In a study of rabbits who were given penicillin, you want to know what dose of penicillin results in a 0.5 probability of curing a rabbit. In this case, the inverse prediction for 0.5 is called the *ED50*, the effective dose corresponding to a 50% survival rate. Use the crosshair tool to visually approximate an inverse prediction.

To see which value of ln(dose) is equally likely either to cure or to be lethal, proceed as follows:

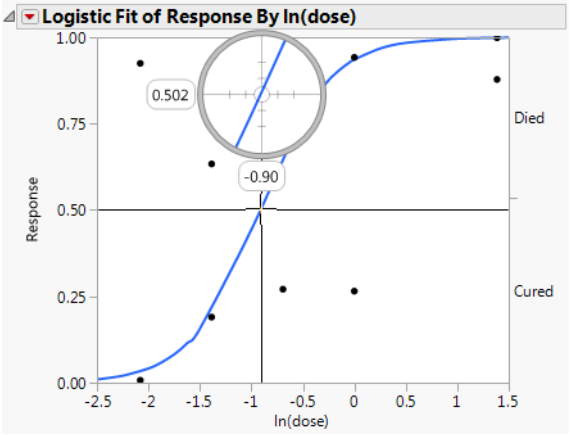
1. Select **Help > Sample Data Library** and open Penicillin.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Response and click **Y, Response**.
4. Select ln(Dose) and click **X, Factor**.

   Notice that JMP automatically fills in Count for **Freq**. Count was previously assigned the role of Freq.

5. Click **OK**.
6. Click on the crosshairs tool.
7. Place the horizontal crosshair line at about 0.5 on the vertical (Response) probability axis.
8. Move the cross-hair intersection to the prediction line, and read the ln(dose) value that shows on the horizontal axis.

In this example, a rabbit with a ln(dose) of approximately -0.9 is equally likely to be cured as it is to die.

Figure 8.10 Example of Crosshair Tool on Logistic Plot



Example of Inverse Prediction Using the Inverse Prediction Option

If your response has exactly two levels, the **Inverse Prediction** option enables you to request an exact inverse prediction. You are given the x value corresponding to a given probability of the lower response category, as well as a confidence interval for that x value.

To use the **Inverse Prediction** option, proceed as follows:

1. Select **Help > Sample Data Library** and open Penicillin.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Response and click **Y, Response**.
4. Select ln(Dose) and click **X, Factor**.

   Notice that JMP automatically fills in Count for **Freq**. Count was previously assigned the role of Freq.
5. Click **OK**.
6. From the red triangle menu, select **Inverse Prediction**. See Figure 8.11.
7. Type 0.95 for the **Confidence Level**.

8. Select **Two sided** for the confidence interval.
9. Request the response probability of interest. Type 0.5 and 0.9 for this example, which indicates you are requesting the values for ln(Dose) that correspond to a 0.5 and 0.9 probability of being cured.
10. Click **OK**.

    The Inverse Prediction plot appears.

Figure 8.11 Inverse Prediction Window

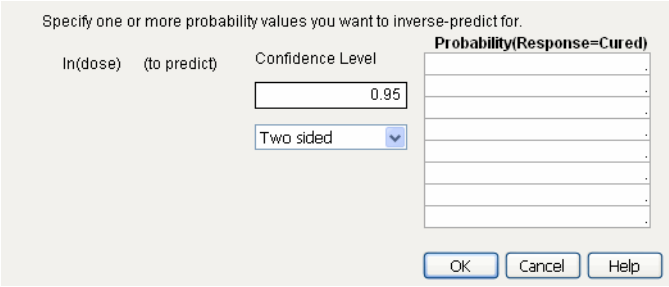
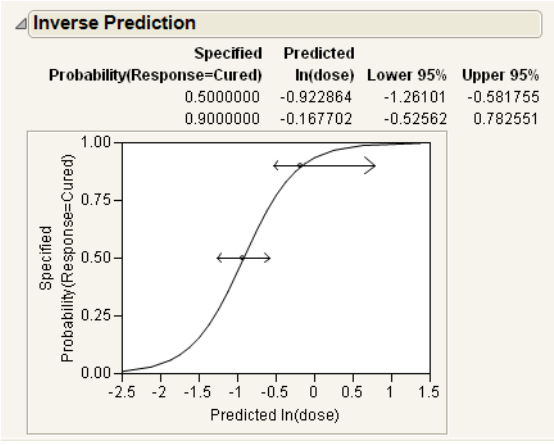


Figure 8.12 Example of Inverse Prediction Plot



Inverse Prediction

Specified Probability(Response=Cured)	Predicted ln(dose)	Lower 95%	Upper 95%
0.5000000	-0.922864	-1.26101	-0.581755
0.9000000	-0.167702	-0.52562	0.782551

The estimates of the x values and the confidence intervals are shown in the report as well as in the probability plot. For example, the value of ln(Dose) that results in a 90% probability of being cured is estimated to be between -0.526 and 0.783.

Statistical Details for the Logistic Platform

Whole Model Test Report

Chi-Square

The Chi-Square statistic is sometimes denoted G^2 and is written as follows:

$$G^2 = 2(\sum -\ln p(\text{background}) - \sum -\ln p(\text{model}))$$

where the summations are over all observations instead of all cells.

RSquare (U)

The ratio of this test statistic to the background log-likelihood is subtracted from 1 to calculate R^2. More simply, RSquare (U) is computed as follows:

$$\frac{\text{negative log-likelihood for Difference}}{\text{negative log-likelihood for Reduced}}$$

using quantities from the Whole Model Test report.

Note: RSquare (U) is also known as *McFadden's pseudo R-square*.

Chapter 9

Tabulate

Create Summary Tables Interactively

Use the Tabulate platform to interactively construct tables of descriptive statistics. The Tabulate platform is an easy and flexible way to present summary data in tabular form. Tables are built from grouping columns, analysis columns, and statistics keywords.

Figure 9.1 Tabulate Examples

	sex											
	F						M					
	age						age					
	12	13	14	15	16	17	12	13	14	15	16	17
	weight	weight	weight	weight	weight	weight	weight	weight	weight	weight	weight	weight
Min	64	67	81	92	112	116	79	79	92	104	128	134
Mean	100.2	95.3	96.6	102.0	113.5	116.0	97.0	94.3	103.9	110.8	128.0	153.0
Max	145	112	142	112	115	116	128	105	119	128	128	172

		sex							
		Female				Male			
		marital status				marital status			
		Married		Single		Married		Single	
		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
country	size	age	age	age	age	age	age	age	age
American	Large	33.6	8.107	41.0	.	34.7	3.931	32.0	6.265
	Medium	31.4	5.827	29.0	9.258	31.3	5.413	32.1	11.05
	Small	31.0	5.657	29.0	9.539	31.8	4.813	26.5	6.455
European	Large	34.0	7.071	28.0	.	.	.	26.0	.
	Medium	31.0	5.06	28.7	5.508	32.3	5.62	31.0	10.13
	Small	29.8	6.611	28.0	1.414	33.8	4.381	25.7	2.517
Japanese	Large	25.0	.	.	.	32.0	.	.	.
	Medium	30.5	4.993	28.0	3.071	32.3	3.878	27.4	5.016
	Small	29.6	4.251	31.1	9.562	29.8	5.357	28.7	4.739
country									
American		31.9	6.452	30.0	9.115	32.6	4.919	31.0	8.179
European		31.0	5.612	28.3	3.559	33.3	4.608	28.4	7.328
Japanese		29.8	4.54	30.1	8.113	30.9	4.822	28.3	4.781

		Mean		
Type	Size Co	Profits ($M)	Sales ($M)	profit/emp
Computer	big	1089.9	20597.48	4530.478
	medium	-85.75	3018.85	-3462.51
	small	44.94	1758.06	7998.815
	All	240.87	5652.02	6159.015
Pharmaceutical	big	894.42	7474.04	17140.70
	medium	698.98	4261.06	24035.11
	small	156.95	1083.75	38337.19
	All	690.08	5070.25	23546.12
All	All	409.32	5433.86	12679.18
Type				
Computer		240.87	5652.02	6159.015
Pharmaceutical		690.08	5070.25	23546.12
All		409.32	5433.86	12679.18

Example of the Tabulate Platform

You have data containing height measurements for male and female students. You want to create a table that shows the mean height for males and females and the aggregate mean for both sexes. You want the table to look like Figure 9.2.

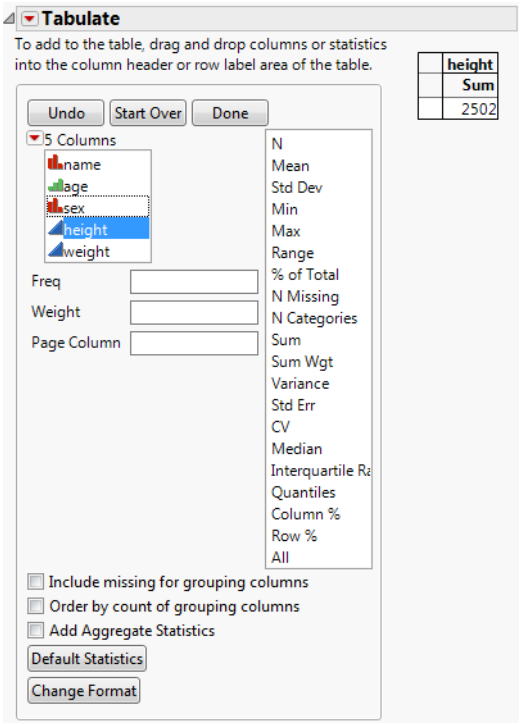
Figure 9.2 Table Showing Mean Height

	height
sex	Mean
F	60.9
M	63.9
All	62.6

1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Tabulate**.

   Since height is the variable you are examining, you want it to appear at the top of the table.
3. Click height and drag it into the Drop zone for columns.

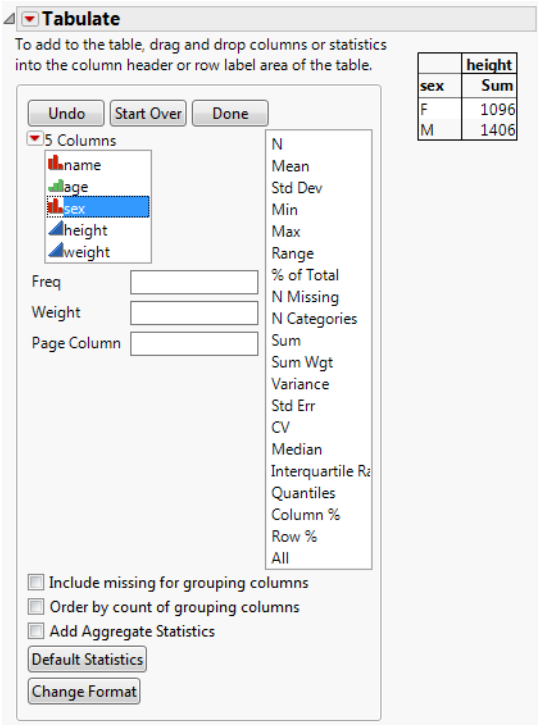
Figure 9.3 Height Variable Added



You want the statistics by sex, and you want sex to appear on the side.

4. Click sex and drag it into the blank cell next to the number 2502.

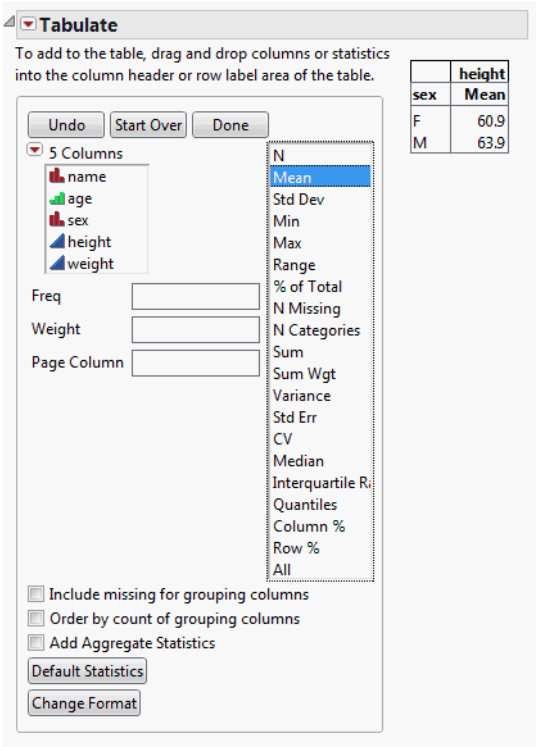
Figure 9.4 Sex Variable Added



Instead of the sum, you want it to show the mean.

5. Click Mean and drag it on top of Sum.

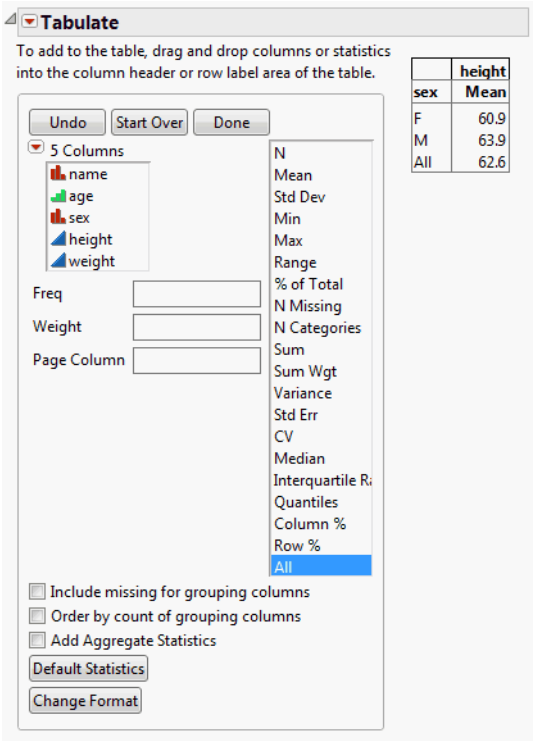
Figure 9.5 Mean Statistic Added



You also want to see the combined mean for males and females.

6. Click All and drag it on top of sex. Or, you can simply select the **Add Aggregate Statistics** check box.

Figure 9.6 All Statistic Added



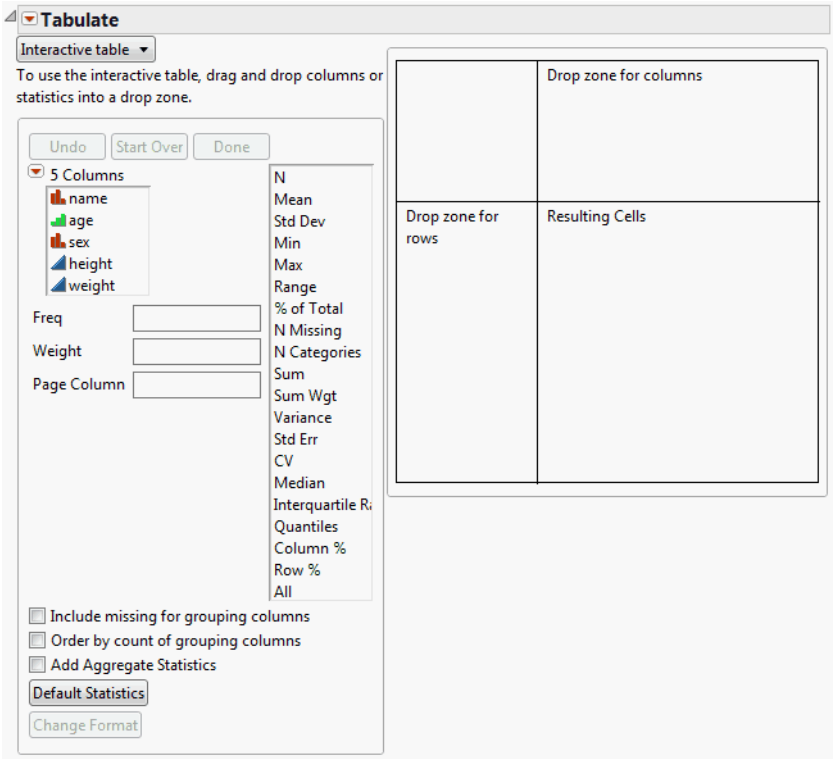
7. (Optional) Click **Done**.

The completed table shows the mean height for females, males, and the combined mean height for both.

Launch the Tabulate Platform

To launch the Tabulate platform, select **Analyze** > **Tabulate**.

Figure 9.7 The Tabulate Interactive Table



Note: For details about red triangle options, see "Tabulate Platform Options" on page 273.

The Tabulate window contains the following options:

Interactive table/dialog Switch between the two modes. Use the interactive table mode to drag and drop items, creating a custom table. Use the dialog mode to create a simple table using a fixed format. See "Use the Dialog" on page 265.

Statistics options Lists standard statistics. Drag any statistic from the list to the table to incorporate it. See "Add Statistics" on page 266.

Drop zone for columns Drag and drop columns or statistics here to create columns.

Note: If the data table contains columns with names equal to those in the Statistics options, be sure to drag and drop the column name from the column list; otherwise, JMP may substitute the statistic of the same name in the table.

Drop zone for rows Drag and drop columns or statistics here to create rows.

Resulting cells Shows the resulting cells based on the columns or statistics that you drag and drop.

Freq Identifies the data table column whose values assign a frequency to each row. This option is useful when a frequency is assigned to each row in summarized data.

Weight Identifies the data table column whose variables assign weight (such as importance or influence) to the data.

Page Column Generates separate tables for each category of a nominal or ordinal column. See "Example Using a Page Column" on page 281.

Include missing for grouping columns Creates a separate group for missing values in grouping columns. When unchecked, missing values are not included in the table. Note that any missing value codes that you have defined as column properties are taken into account.

Order by count of grouping columns Changes the order of the table to be in ascending order of the values of the grouping columns.

Add Aggregate Statistics Adds aggregate statistics for all rows and columns.

Default Statistics Enables you to change the default statistics that appear when you drag and drop analysis or non-analysis (for example, grouping) columns.

Change Format Enables you to change the numeric format for displaying specific statistics. See "Change Numeric Formats" on page 268.

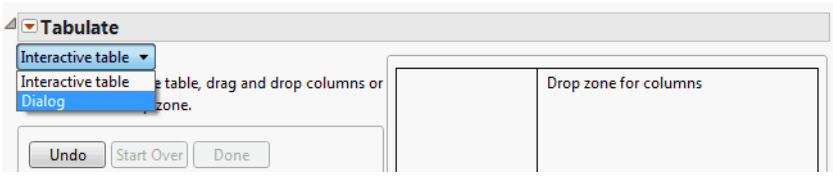
Change Plot Scale (Only appears if **Show Chart** is selected from the red triangle menu.) Enables you to specify a uniform custom scale.

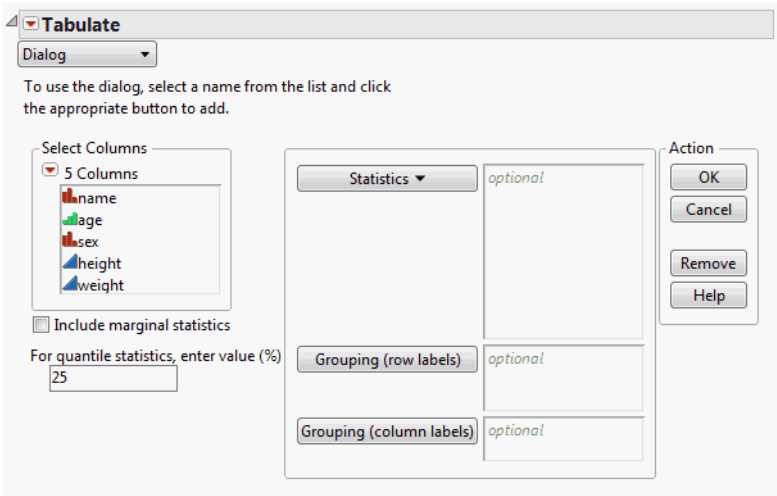
Uniform plot scale (Only appears if **Show Chart** is selected from the red triangle menu.) Deselect this box for each column of bars to use the scale determined separately from the data in each displayed column.

Use the Dialog

If you prefer not to drag and drop and build the table interactively, you can create a simple table using the Dialog interface. After selecting **Analyze > Tabulate**, select **Dialog** from the menu, as shown in Figure 9.8. You can make changes to the table by selecting **Show Control Panel** from the red triangle menu, and then drag and drop new items into the table.

Figure 9.8 Using the Dialog





The dialog contains the following options:

Include marginal statistics Aggregates summary information for categories of a grouping column.

For quantile statistics, enter value (%) Type the value at which the specific percentage of the argument is less than or equal to. For example, 75% of the data is less than the 75th quantile. This applies to all grouping columns.

Statistics Once you've selected a column, select a standard statistic to apply to that column. See "Add Statistics" on page 266

Grouping (row labels) Select the column to use as the row label.

Grouping (column labels) Select the column to use as the column label.

Add Statistics

Tip: You can select both a column and a statistic at the same time and drag them into the table.

Tabulate supports a list of standard statistics. The list is displayed in the control panel. You can drag any keyword from that list to the table, just like you do with the columns. Note the following:

- The statistics associated with each cell are calculated on values of the analysis columns from all observations in that category, as defined by the grouping columns.
- All of the requested statistics have to reside in the same dimension, either in the row table or in the column table.
- If you drag a continuous column into a data area, it is treated as an analysis column.

Tabulate uses the following keywords:

N Provides the number of nonmissing values in the column. This is the default statistic when there is no analysis column.

Mean Provides the arithmetic mean of a column's values. It is the sum of nonmissing values (and if defined, multiplied by the **weight** variable) divided by the **Sum Wgt**.

Std Dev Provides the sample standard deviation, computed for the nonmissing values. It is the square root of the sample variance.

Min Provides the smallest nonmissing value in a column.

Max Provides the largest nonmissing value in a column.

Range Provides the difference between **Max** and **Min**.

% of Total Computes the percentage of total of the whole population. The denominator used in the computation is the total of all the included observations, and the numerator is the total for the category. If there is no analysis column, the % of Total is the percentage of total of counts. If there is an analysis column, the % of Total is the percentage of the total of the sum of the analysis column. Thus, the denominator is the sum of the analysis column over all the included observations, and the numerator is the sum of the analysis column for that category. You can request different percentages by dragging the keyword into the table.

- Dropping one or more grouping columns from the table to the % of Total heading changes the denominator definition. For this, Tabulate uses the sum of these grouping columns for the denominator.
- To get the percentage of the column total, drag all the grouping columns on the row table and drop them onto the **% of Total** heading (same as Column %). Similarly, to get the percentage of the row total, drag all grouping columns on the column table and drop them onto the **% of Total** heading (same as Row %).

N Missing Provides the number of missing values.

N Categories Provides the number of distinct categories.

Sum Provides the sum of all values in the column. This is the default statistic for analysis columns when there are no other statistics for the table.

Sum Wgt Provides the sum of all weight values in a column. Or, if no column is assigned the weight role, **Sum Wgt** is the total number of nonmissing values.

Variance Provides the sample variance, computed for the nonmissing values. It is the sum of squared deviations from the mean, divided by the number of nonmissing values minus one.

Std Err Provides the standard error of the mean. It is the standard deviation divided by the square root of **N**. If a column is assigned the role of weight, then the denominator is the square root of the sum of the weights.

CV (Coefficient of Variation) Provides the measure of dispersion, which is the standard deviation divided by the mean multiplied by one hundred.

Median Provides the 50th percentile, which is the value where half the data are below and half are above or equal to the 50th quantile (median).

Geometric Mean Provides the *n*th root of the product of *n* numbers. Zero and negative numbers are treated like missing. For example, you might want to compare two companies based on varying metrics that come from different ranges. The statistic is also helpful when the data contains a large value in a skewed distribution.

Interquartile Range Provides the difference between the 3rd quartile and 1st quartile.

Quantiles Provides the value at which the specific percentage of the argument is less than or equal to. For example, 75% of the data is less than the 75th quantile. You can request different quantiles by clicking and dragging the **Quantiles** keyword into the table, and then entering the quantile into the box that appears.

Column % Provides the percent of each cell count to its column total if there is no analysis column. If there is an analysis column, the Column % is the percent of the column total of the sum of the analysis column. For tables with statistics on the top, you can add Column % to tables with multiple row tables (stacked vertically).

Row % Provides the percent of each cell count to its row total if there is no analysis column. If there is an analysis column, the Row % is the percent of the row total of the sum of the analysis column. For tables with statistics on the side, you can add Row % to tables with multiple column tables (side by side tables).

All Aggregates summary information for categories of a grouping column.

Change Numeric Formats

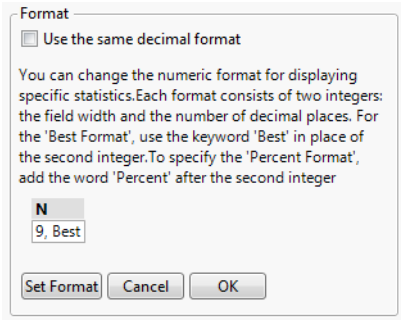
The formats of each cell depend on the analysis column and the statistics. For counts, the default format has no decimal digits. For each cell defined by some statistics, JMP tries to determine a reasonable format using the format of the analysis column and the statistics requested. To override the default format:

1. Click the **Change Format** button at the bottom of the Tabulate window.
2. In the panel that appears, enter the field width, a comma, and then the number of decimal places that you want displayed in the table. See Figure 9.9.

3. To exhibit the cell value in Percent format, add a comma after the number of decimal places and type the word **Percent**.
4. (Optional) If you would like JMP to determine the best format for you to use, type the word **Best** in the text box.

   JMP now considers the precision of each cell value and selects the best way to show it.
5. Click **OK** to implement the changes and close the Format section, or click **Set Format** to see the changes implemented without closing the Format section.

Figure 9.9 Changing Numeric Formats



The Tabulate Output

The Tabulate output consists of one or more column tables concatenated side by side, and one or more row tables concatenated top to bottom. The output might have only a column table or a row table.

Figure 9.10 Tabulate Output

	sex											
	F						M					
	age						age					
	12	13	14	15	16	17	12	13	14	15	16	17
	weight	weight	weight	weight	weight	weight	weight	weight	weight	weight	weight	weight
Min	64	67	81	92	112	116	79	79	92	104	128	134
Mean	100.2	95.3	96.6	102.0	113.5	116.0	97.0	94.3	103.9	110.8	128.0	153.0
Max	145	112	142	112	115	116	128	105	119	128	128	172

		sex							
		Female				Male			
		marital status				marital status			
		Married		Single		Married		Single	
		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
country	size	age	age	age	age	age	age	age	age
American	Large	33.6	8.107	41.0	.	34.7	3.931	32.0	6.265
	Medium	31.4	5.827	29.0	9.258	31.3	5.413	32.1	11.05
	Small	31.0	5.657	29.0	9.539	31.8	4.813	26.5	6.455
European	Large	34.0	7.071	28.0	.	.	.	26.0	.
	Medium	31.0	5.06	28.7	5.508	32.3	5.62	31.0	10.13
	Small	29.8	6.611	28.0	1.414	33.8	4.381	25.7	2.517
Japanese	Large	25.0	.	.	.	32.0	.	.	.
	Medium	30.5	4.993	28.0	3.071	32.3	3.878	27.4	5.016
	Small	29.6	4.251	31.1	9.562	29.8	5.357	28.7	4.739
country									
American		31.9	6.452	30.0	9.115	32.6	4.919	31.0	8.179
European		31.0	5.612	28.3	3.559	33.3	4.608	28.4	7.328
Japanese		29.8	4.54	30.1	8.113	30.9	4.822	28.3	4.781

		Mean		
Type	Size Co	Profits ($M)	Sales ($M)	profit/emp
Computer	big	1089.9	20597.48	4530.478
	medium	-85.75	3018.85	-3462.51
	small	44.94	1758.06	7998.815
	All	240.87	5652.02	6159.015
Pharmaceutical	big	894.42	7474.04	17140.70
	medium	698.98	4261.06	24035.11
	small	156.95	1083.75	38337.19
	All	690.08	5070.25	23546.12
All	All	409.32	5433.86	12679.18
Type				
Computer		240.87	5652.02	6159.015
Pharmaceutical		690.08	5070.25	23546.12
All		409.32	5433.86	12679.18

Creating a table interactively is an iterative process:

- Click the items (columns or statistics) from the appropriate list, and drag them into the drop zone (for rows or columns). See "Edit Tables" on page 273, and "Column and Row Tables" on page 272.
- Add to the table by repeating the drag and drop process. The table updates to reflect the latest addition. If there are already column headings or row labels, you can decide where the addition goes relative to the existing items.

Note the following about clicking and dragging:

- JMP uses the modeling type to determine a column's role. Continuous columns are assumed to be analysis columns. See "Analysis Columns" on page 271. Ordinal or nominal columns are assumed to be grouping columns. See "Grouping Columns" on page 271.
- When you drag and drop multiple columns into the initial table:
  - If the columns share a set of common values, they are combined into a single table. A crosstabulation of the column names and the categories gathered from these columns is generated. Each cell is defined by one of the columns and one of the categories.
  - If the columns do not share common values, they are put into separate tables.

  - You can always change the default action by right-clicking on a column and selecting **Combine Tables** or **Separate Tables**. For more details, see "Right-Click Menu for Columns" on page 275.
- To nest columns, create a table with the first column, and then drag the additional columns into the first column.
- In a properly created table, all grouping columns are together, all analysis columns are together, and all statistics are together. Therefore, JMP does not intersperse a statistics keyword within a list of analysis columns. JMP also does not insert an analysis column within a list of grouping columns.
- You can drag columns from the Table panel in the data table onto a Tabulate table instead of using the Tabulate Control Panel.

Note: The Tabulate table is updated when you add data to the open data table, delete rows, and recode the data.

Analysis Columns

Analysis columns are any numeric columns for which you want to compute statistics. They are continuous columns. Tabulate computes statistics on the analysis columns for each category formed from the grouping columns.

Note that all the analysis columns have to reside in the same dimension, either in the row table or in the column table.

Grouping Columns

Grouping columns are columns that you want to use to classify your data into categories of information. They can have character, integer, or even decimal values, but the number of unique values should be limited. Grouping columns are either nominal or ordinal.

Note the following:

- If grouping columns are nested, Tabulate constructs distinct categories from the hierarchical nesting of the values of the columns. For example, from the grouping columns Sex with values F and M, and the grouping column Marital Status with values Married and Single, Tabulate constructs four distinct categories: F and Married, F and Single, M and Married, M and Single.
- You can specify grouping columns for column tables as well as row tables. Together they generate the categories that define each table cell.
- Tabulate does not include observations with a missing value for one or more grouping columns by default. You can include them by checking the **Include missing for grouping columns** option.

- To specify codes or values that should be treated as missing, use the Missing Value Codes column property. You can include these by checking the **Include missing for grouping columns** option. For more details about Missing Value Codes, see The Column Info Window chapter in the *Using JMP* book.

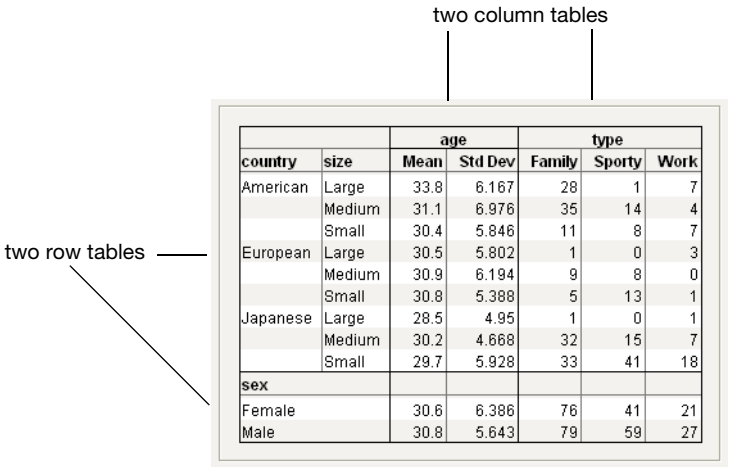
Column and Row Tables

In Tabulate, a table is defined by its column headings and row labels. These sub-tables are referred to as *row tables* and *column tables*. See Figure 9.11.

Example of Row and Column Tables

1. Select **Help > Sample Data Library** and open Car Poll.jmp.
2. Select **Analyze > Tabulate**.
3. Drag size into the Drop zone for rows.
4. Drag country to the left of the size heading.
5. Drag Mean over the N heading.
6. Drag Std Dev below the Mean heading.
7. Drag age above the Mean heading.
8. Drag type to the far right of the table.
9. Drag sex under the table.

Figure 9.11 Row and Column Tables



		age		type		
country	size	Mean	Std Dev	Family	Sporty	Work
American	Large	33.8	6.167	28	1	7
	Medium	31.1	6.976	35	14	4
	Small	30.4	5.846	11	8	7
European	Large	30.5	5.802	1	0	3
	Medium	30.9	6.194	9	8	0
	Small	30.8	5.388	5	13	1
Japanese	Large	28.5	4.95	1	0	1
	Medium	30.2	4.668	32	15	7
	Small	29.7	5.928	33	41	18
sex						
Female		30.6	6.386	76	41	21
Male		30.8	5.643	79	59	27

For multiple column tables, the labels on the side are shared across the column tables. In this instance, country and sex are shared across the tables. Similarly, for multiple row tables, the

headings on the top are shared among the row tables. In this instance, both age and type are shared among the tables.

Edit Tables

There are several ways to edit the items that you add to a table.

Delete Items

After you add items to the table, you can remove them in any one of the following ways:

- Drag the item away from the table.
- To remove the last item, click **Undo**.
- Right-click on an item and select **Delete**.

Remove Column Labels

Grouping columns display the column name on top of the categories associated with that column. For some columns, the column name might seem redundant. Remove the column name from the column table by right-clicking on the column name and selecting **Remove Column Label**. To re-insert the column label, right-click on one of its associated categories and select **Restore Column Label**.

Edit Statistical Key Words and Labels

You can edit a statistical key word or a statistical label. For example, instead of Mean, you might want to use the word Average. Right-click on the word that you want to edit and select **Change Item Label**. In the box that appears, type the new label. Alternatively, you can type directly into the edit box.

If you change one statistics keyword to another statistics keyword, JMP assumes that you actually want to change the statistics, not just the label. It would be as if you have deleted the statistics from the table and added the latter.

Tabulate Platform Options

The following options are available from the red triangle menu next to Tabulate:

Show Table Displays the summarized data in tabular form.

Show Chart Displays the summarized data in bar charts that mirrors the table of summary statistics. The simple bar chart enables visual comparison of the relative magnitude of the summary statistics. By default, all columns of bars share the same scale. You can have each column of bars use the scale determined separately from the data in each displayed

column, by clearing the **Uniform plot scale** check box. You can specify a uniform custom scale using the **Change Plot Scale** button. The charts are either 0-based or centered on 0. If the data are all nonnegative, or all non-positive, the charts baseline is at 0. Otherwise, the charts are centered on 0.

Show Control Panel Displays the control panel for further interaction.

Show Shading Displays gray shading boxes in the table when there are multiple rows.

Show Tooltip Displays tips that appear when you move the mouse over areas of the table.

Show Test Build Panel Displays the control area that lets you create a test build using a random sample from the original table. This is particularly useful when you have large amounts of data. See "Show Test Build Panel" on page 274.

Make Into Data Table Makes a data table from the report. There is one data table for each row table, because labels of different row tables might not be mapped to the same structure.

Full Path Column Name Uses the fully qualified column names of grouping columns for the column name in the created data table.

See the JMP Reports chapter in the *Using JMP* book for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

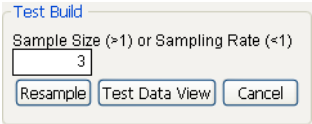
For a description of the options in the Select Columns red triangle menu, see the Get Started chapter in the *Using JMP* book.

Show Test Build Panel

If you have a very large data table, you might want to use a small subset of the data table to try out different table layouts to find one that best shows the summary information. In this case, JMP generates a random subset of the size as specified and uses that subset when it builds the table. To use the test build feature:

1. From the red triangle menu next to Tabulate, select **Show Test Build Panel**.
2. Enter the size of the sample that you want in the box under **Sample Size (>1) or Sampling Rate (<1)**, as shown in Figure 9.12. The size of the sample can be either the proportion of the active table that you enter or the number of rows from the active table.

Figure 9.12 The Test Build Panel



3. Click **Resample**.
4. To see the sampled data in a JMP data table, click the **Test Data View** button. When you dismiss the test build panel, Tabulate uses the full data table to regenerate the tables as designed.

Right-Click Menu for Columns

Right-click on a column in Tabulate to see the following options:

Delete Deletes the selected column.

Use as Grouping column Changes the analysis column to a grouping column.

Use as Analysis column Changes the grouping column to an analysis column.

Change Item Label (Appears only for separate or nested columns) Type a new label.

Combine Tables (Columns by Categories) (Appears only for separate or nested columns) Combines separate or nested columns. See "Example of Combining Columns into a Single Table" on page 279.

Separate Tables (Appears only for combined tables) Creates a separate table for each column.

Nest Grouping Columns Nests grouping columns vertically or horizontally.

Additional Examples of the Tabulate Platform

This example contains the following steps:

1. "Create a Table of Counts" on page 275
2. "Create a Table Showing Statistics" on page 276
3. "Rearrange the Table Contents" on page 278

Create a Table of Counts

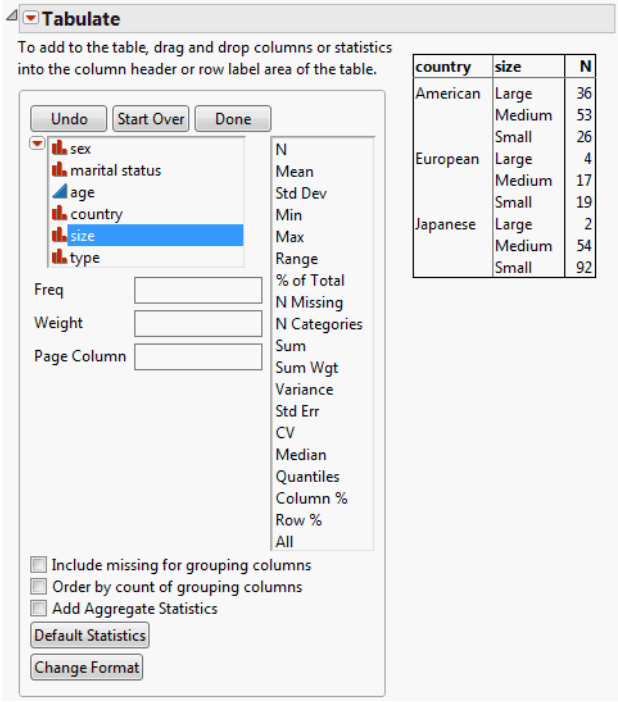
Suppose that you would like to view a table that contains counts for how many people in the survey own Japanese, European, and American cars, broken down by the size of the car. You want the table to look Figure 9.3.

Figure 9.13 Table Showing Counts of Car Ownership

country	size	N
American	Large	36
	Medium	53
	Small	26
European	Large	4
	Medium	17
	Small	19
Japanese	Large	2
	Medium	54
	Small	92

1. Select **Help > Sample Data Library** and open Car Poll.jmp.
2. Select **Analyze > Tabulate**.
3. Click country and drag it into the Drop zone for rows.
4. Click size and drag it to the right of the country heading.

Figure 9.14 Country and Size Added to the Table



Create a Table Showing Statistics

Suppose that you would like to see the mean (average) and the standard deviation of the age of people who own each size car. You want the table to look like Figure 9.15.

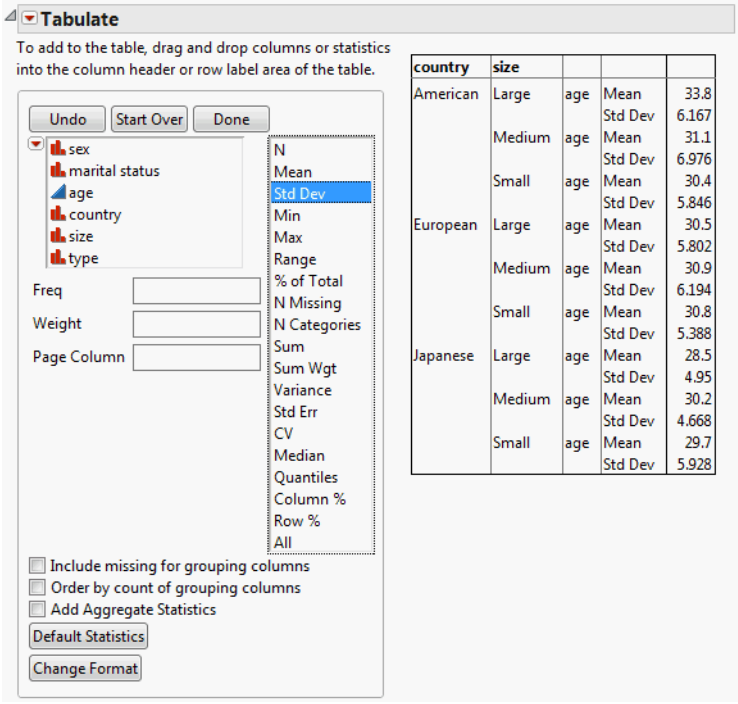
Figure 9.15 Table Showing Mean and Standard Deviation by Age

country	size			
American	Large	age	Mean	33.8
			Std Dev	6.2
	Medium	age	Mean	31.1
			Std Dev	7.0
	Small	age	Mean	30.4
			Std Dev	5.8
European	Large	age	Mean	30.5
			Std Dev	5.8
	Medium	age	Mean	30.9
			Std Dev	6.2
	Small	age	Mean	30.8
			Std Dev	5.4
Japanese	Large	age	Mean	28.5
			Std Dev	4.9
	Medium	age	Mean	30.2
			Std Dev	4.7
	Small	age	Mean	29.7
			Std Dev	5.9

1. Start from Figure 9.14. Click age and drag it to the right of the size heading.
2. Click Mean and drag it over Sum.
3. Click Std Dev and drag it below Mean.

   Std Dev is placed below Mean in the table. Dropping Std Dev above Mean places Std Dev above Mean in the table.

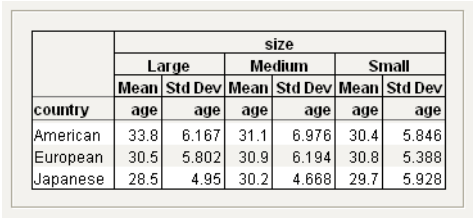
Figure 9.16 Age, Mean, and Std Dev Added to the Table



Rearrange the Table Contents

Suppose that you would prefer size to be on top, showing a crosstab layout. You want the table to look like Figure 9.17.

Figure 9.17 Size on Top



	size					
	Large		Medium		Small	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
country	age	age	age	age	age	age
American	33.8	6.167	31.1	6.976	30.4	5.846
European	30.5	5.802	30.9	6.194	30.8	5.388
Japanese	28.5	4.95	30.2	4.668	29.7	5.928

To rearrange the table contents, proceed as follows:

1. Start from Figure 9.16. Click on the size heading and drag it to the right of the table headings. See Figure 9.18.

Figure 9.18 Moving size

country	size			
American	Large	age	Mean	33.8
			Std Dev	6.167
	Medium	age	Mean	31.1
			Std Dev	6.976
	Small	age	Mean	30.4
			Std Dev	5.846
European	Large	age	Mean	30.5
			Std Dev	5.802
	Medium	age	Mean	30.9
			Std Dev	6.194
	Small	age	Mean	30.8
			Std Dev	5.388
Japanese	Large	age	Mean	28.5
			Std Dev	4.95
	Medium	age	Mean	30.2
			Std Dev	4.668
	Small	age	Mean	29.7
			Std Dev	5.928

			size		
country			Large	Medium	Small
American	age	Mean	33.8	31.1	30.4
		Std Dev	6.167	6.976	5.846
European	age	Mean	30.5	30.9	30.8
		Std Dev	5.802	6.194	5.388
Japanese	age	Mean	28.5	30.2	29.7
		Std Dev	4.95	4.668	5.928

2. Click on age and drag it under the Large Medium Small heading.
3. Select both Mean and Std Dev, and then drag them under the Large heading.

Now your table clearly presents the data. It is easier to see the mean and standard deviation of the car owner age broken down by car size and country.

Example of Combining Columns into a Single Table

You have data from students indicating the importance of self-reported factors in children's popularity (grades, sports, looks, money). Suppose that you want to see all of these factors in a single, combined table with additional statistics and factors. You want the table to look like Figure 9.19.

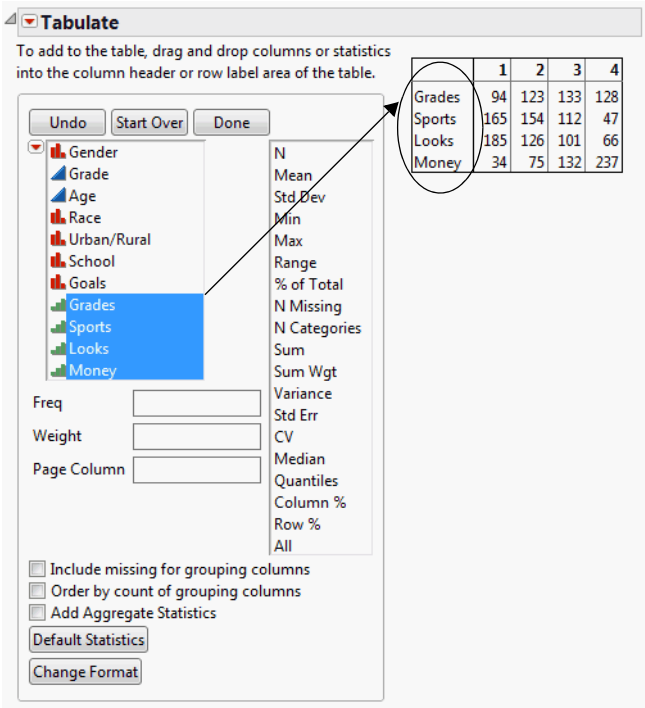
Figure 9.19 Adding Demographic Data

		% of Total														
		Urban/Rural														
		Rural					Suburban					Urban				
Gender		1	2	3	4	All	1	2	3	4	All	1	2	3	4	All
boy	Grades	2.30%	2.93%	3.56%	5.02%	13.81%	2.72%	5.86%	5.02%	5.02%	18.62%	3.14%	3.97%	5.44%	2.51%	15.06%
	Sports	6.49%	3.97%	2.72%	0.63%	13.81%	11.30%	4.60%	2.30%	0.42%	18.62%	8.79%	3.97%	1.46%	0.84%	15.06%
	Looks	3.14%	4.60%	2.72%	3.35%	13.81%	3.77%	5.86%	5.44%	3.56%	18.62%	2.30%	5.02%	4.18%	3.56%	15.06%
	Money	1.88%	2.30%	4.81%	4.81%	13.81%	0.84%	2.30%	5.86%	9.62%	18.62%	0.84%	2.09%	3.97%	8.16%	15.06%
girl	Grades	4.39%	4.39%	4.39%	4.18%	17.36%	2.51%	3.35%	2.93%	4.18%	12.97%	4.60%	5.23%	6.49%	5.86%	22.18%
	Sports	2.30%	6.69%	5.44%	2.93%	17.36%	1.67%	3.56%	5.65%	2.09%	12.97%	3.97%	9.41%	5.86%	2.93%	22.18%
	Looks	9.62%	3.35%	3.35%	1.05%	17.36%	7.95%	2.93%	1.26%	0.84%	12.97%	11.92%	4.60%	4.18%	1.46%	22.18%
	Money	1.05%	2.93%	4.18%	9.21%	17.36%	0.84%	3.14%	3.14%	5.86%	12.97%	1.67%	2.93%	5.65%	11.92%	22.18%

1. Select **Help > Sample Data Library** and open Children's Popularity.jmp.

2. Select **Analyze > Tabulate**.
3. Select Grades, Sports, Looks, and Money and drag them into the Drop zone for rows.

Figure 9.20 Columns by Categories



Notice that a single, combined table appears.

Tabulate the percentage of the one to four ratings of each category.

4. Drag Gender into the empty heading at left.
5. Drag % of Total above the numbered headings.
6. Drag All beside the number 4.

Figure 9.21 Gender, % of Total, and All Added to the Table

		% of Total				
Gender		**1**	**2**	**3**	**4**	**All**
boy	Grades	8.16%	12.76%	14.02%	12.55%	47.49%
	Sports	26.57%	12.55%	6.49%	1.88%	47.49%
	Looks	9.21%	15.48%	12.34%	10.46%	47.49%
	Money	3.56%	6.69%	14.64%	22.59%	47.49%
girl	Grades	11.51%	12.97%	13.81%	14.23%	52.51%
	Sports	7.95%	19.67%	16.95%	7.95%	52.51%
	Looks	29.50%	10.88%	8.79%	3.35%	52.51%
	Money	3.56%	9.00%	12.97%	26.99%	52.51%

Break down the tabulation further by adding demographic data.

7. Drag Urban/Rural below the % of Total heading.

Figure 9.22 Urban/Rural Added to the Table

		% of Total														
		Urban/Rural														
		Rural					Suburban					Urban				
Gender		1	2	3	4	All	1	2	3	4	All	1	2	3	4	All
boy	Grades	2.30%	2.93%	3.56%	5.02%	13.81%	2.72%	5.86%	5.02%	5.02%	18.62%	3.14%	3.97%	5.44%	2.51%	15.06%
	Sports	6.49%	3.97%	2.72%	0.63%	13.81%	11.30%	4.60%	2.30%	0.42%	18.62%	8.79%	3.97%	1.46%	0.84%	15.06%
	Looks	3.14%	4.60%	2.72%	3.35%	13.81%	3.77%	5.86%	5.44%	3.56%	18.62%	2.30%	5.02%	4.18%	3.56%	15.06%
	Money	1.88%	2.30%	4.81%	4.81%	13.81%	0.84%	2.30%	5.86%	9.62%	18.62%	0.84%	2.09%	3.97%	8.16%	15.06%
girl	Grades	4.39%	4.39%	4.39%	4.18%	17.36%	2.51%	3.35%	2.93%	4.18%	12.97%	4.60%	5.23%	6.49%	5.86%	22.18%
	Sports	2.30%	6.69%	5.44%	2.93%	17.36%	1.67%	3.56%	5.65%	2.09%	12.97%	3.97%	9.41%	5.86%	2.93%	22.18%
	Looks	9.62%	3.35%	3.35%	1.05%	17.36%	7.95%	2.93%	1.26%	0.84%	12.97%	11.92%	4.60%	4.18%	1.46%	22.18%
	Money	1.05%	2.93%	4.18%	9.21%	17.36%	0.84%	3.14%	3.14%	5.86%	12.97%	1.67%	2.93%	5.65%	11.92%	22.18%

You can see that for boys in rural, suburban, and urban areas, sports are the most important factor for popularity. For girls in rural, suburban, and urban areas, looks are the most important factor for popularity.

Example Using a Page Column

You have data containing height measurements for male and female students. You want to create a table that shows the mean height by the age of the students. Then you want to stratify your data by sex in different tables. To do so, add the stratification column as a page column, which will build the pages for each group. You want the table to look like Figure 9.23.

Figure 9.23 Mean Height of Students by Sex

sex = F

	height
age	Sum
12	293
13	177
14	313
15	126
16	125
17	62

Females

sex = M

	height
age	Sum
12	172
13	245
14	457
15	326
16	68
17	138

Males

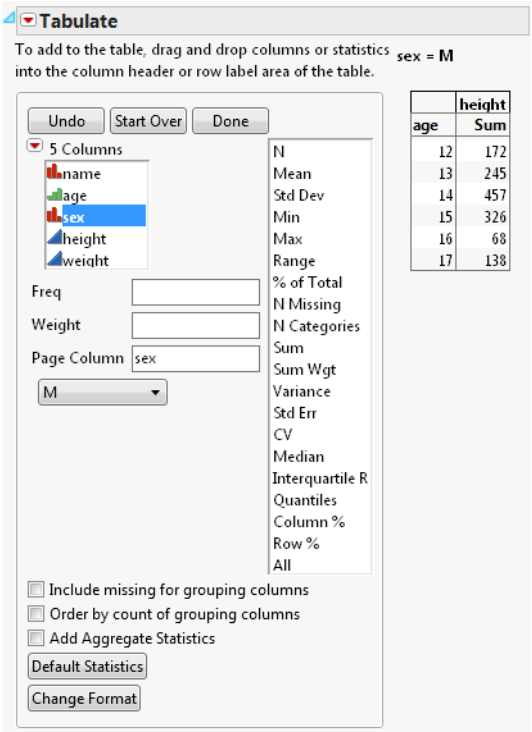
1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Tabulate**.

   Since height is the variable you are examining, you want it to appear at the top of the table.
3. Click height and drag it into the Drop zone for columns.

   You want the statistics by age, and you want age to appear on the side.

4. Click age and drag it into the blank cell next to the number 2502.
5. Click sex and drag it into **Page Column**.
6. Select F from the Page Column list to show the mean heights for only females.
7. Select M from the Page Column list to show the mean heights for only males. You can also select **None Selected** to show all values.

Figure 9.24 Using a Page Column



Chapter 10

JMP PRO Simulate

Answer Challenging Questions with Parametric Resampling

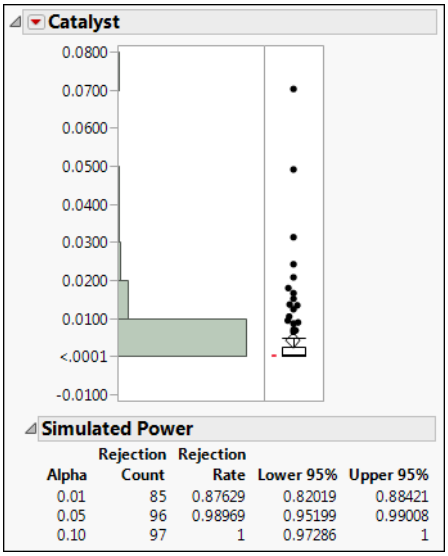
The Simulate platform is available only in JMP Pro.

The Simulate feature provides powerful parametric and nonparametric simulation capability. Use Simulate to do the following:

- Expand on the bootstrap to provide parametric bootstrapping.
- Obtain power calculations in nonstandard situations
- Approximate the distribution of statistics, such as predicted values, and confidence intervals, in nonstandard situations
- Conduct permutation tests
- Explore the effect of assumptions about predictors on models
- Explore various “what if” scenarios relative to your models
- Evaluate new or existing statistical methods

The Simulate option is available in many reports, including all of those that support Bootstrap. To access the Simulate option, right-click in a report.

Figure 10.1 Power Analysis Using Simulate



Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	85	0.87629	0.82019	0.88421
0.05	96	0.98969	0.95199	0.99008
0.10	97	1	0.97286	1

JMP PRO Overview of Simulate

Simulate provides simulated results for a column of statistics in a report. Right-click on a column of statistics in a report and select Simulate. In the Simulate window, specify a column in your data table that forms the basis for your simulation. This is the column that you *switch out*. This column can have any role in the analysis. In particular, it can be a response or a predictor in a model. You then specify a column in your data table that contains a formula that you want to use for the simulation. This is the column that you *switch in*. It functions as a surrogate for the column that you switched out.

Note: Your data table must contain a column that has a random component.

The method works as follows. A column of simulated values is generated based on the formula in the formula column that you switch in. The entire analysis that generated the report containing the statistics of interest is rerun using this new column of simulated values to replace the column that you switched out. This process is repeated *N* times, where *N* is the total number of samples that you specify.

The Simulate analysis produces an output data table showing a summary of the analysis.

- Each row of the data table represents the results of the analysis for one column of simulated values.
- There is a column for each row of the report table involved in the simulation.
- There are scripts to facilitate your analysis.

Tip: Simulate reruns the entire analysis that appears in the platform report from which Simulate is invoked. As a result, Simulate may run slowly for your selected column because of extraneous analyses in the report. If Simulate is taking a long time, remove extraneous options from the platform report prior to running Simulate.

JMP PRO Examples That Use Simulate

This section provides several example of the use of Simulate. Additional examples, also listed below, can be found in other books:

- "Construct an Accurate Confidence Interval for Variance Components" on page 285
- "Conduct a Permutation Test" on page 290
- "Explore Retaining a Factor in Generalized Regression" on page 293
- "Conduct Prospective Power Analysis for a Nonlinear Model" on page 298

- For an example that shows how to simulate a confidence interval for Ppk and the percent nonconforming for a non-normal variable, see the Process Capability chapter in the *Quality and Process Methods* book.

JMP PRO Construct an Accurate Confidence Interval for Variance Components

In this example, you are interested in the effects of temperature, time, and the amount of catalyst on a reaction. Temperature is a very-hard-to-change variable (whole plot factor), time is hard-to-change (subplot factor), and the amount of catalyst is easy-to-change. For information on whole plot and subplot factors, see the Custom Designs chapter in the *Design of Experiments Guide*.

Your goal is to obtain accurate confidence intervals for the whole-plot and sub-plot variance components. Previous studies have suggested that the whole-plot standard deviation is about twice the error standard deviation, while the sub-plot error is about 1.5 times the error standard deviation. The Wald intervals given in the REML report, which assume that the variance components are asymptotically normal, have poor coverage properties. You will obtain confidence intervals by simulating the distributions of the variance components.

In this example, you will do the following:

- Construct a custom design for your split-split-plot experiment. See "Construct the Design" on page 285.
- Fit a model using the REML method. See "Fit the Model" on page 288.
- Simulate variance component estimates in order to obtain percentile confidence intervals for the variance components. See"Explore Power" on page 288.

JMP PRO Construct the Design

If you prefer to skip the steps in this section, select **Help > Sample Data Library** and open Design Experiment/Catalyst Design.jmp. In the Catalyst Design.jmp data table, click the green triangle next to the **DOE Simulate** script. Then go to "Fit the Model" on page 288.

1. Select **DOE > Custom Design**.
2. In the Factors outline, type 3 next to **Add N Factors**.
3. Click **Add Factor > Continuous**.
4. Double-click to rename these factors Temperature, Time, and Catalyst.

   Keep the default Values of –1 and 1 for these factors.
5. For Temperature, click **Easy** and select **Very Hard**.

   This defines Temperature to be a whole plot factor.
6. For Time, click **Easy** and select **Hard** for Time.

This defines Time to be a subplot factor.

7. Click **Continue**.
8. In the Model outline, select **Interactions > 2nd**.

   This adds all two-way interactions to the model.
9. Click the Custom Design red triangle and select **Simulate Responses**.

   This opens the Simulate Responses window after you select Make Table to construct the design table.

Note: Setting the Random Seed in step 10 and Number of Starts in step 11 reproduces the exact results shown in this example. In constructing a design on your own, these steps are not necessary.

---

10. (Optional) Click the Custom Design red triangle and select **Set Random Seed**. Type 12345 and click **OK**.
11. (Optional) Click the Custom Design red triangle and select **Number of Starts**. Type 1000 and click **OK**.
12. Click **Make Design**.
13. Click **Make Table**.

Note: The entries in your Y and Y Simulated columns will differ from those that appear in Figure 10.2.

Figure 10.2 Design Table

Catalyst Design
Locked File C:\Program File
Design Custom Design
Criterion D Optimal
Model
Model for Y Simulated
Evaluate Design
DOE Simulate
DOE Dialog

Columns (7/0)
Whole Plots
Subplots
Temperature
Time
Catalyst
Y
Y Simulated

Rows
All rows 24
Selected 0
Excluded 0
Hidden 0
Labelled 0

	Whole Plots	Subplots	Temperature	Time	Catalyst	Y	Y Simulated
1	1	1	1	1	-1	-0.046768023	4.51599406
2	1	1	1	1	1	6.539690219	8.13483314
3	1	1	1	1	1	6.5397108257	6.83498659
4	1	2	1	-1	1	0.494801276	0.41850996
5	1	2	1	-1	-1	-2.65841644	-2.8766717
6	1	2	1	-1	-1	-0.200740508	-2.5418434
7	2	3	-1	1	1	3.1335699565	1.92256582
8	2	3	-1	1	1	2.2739679213	0.02864554
9	2	3	-1	1	-1	1.1704395547	-1.764573
10	2	4	-1	-1	-1	-0.693941035	1.00107343
11	2	4	-1	-1	-1	2.3525055616	0.95004034
12	2	4	-1	-1	1	-1.545149968	-1.7297909
13	3	5	-1	1	-1	-1.4838427	1.63438017
14	3	5	-1	1	-1	-1.375362439	1.61316385
15	3	5	-1	1	1	0.7434199911	4.38615961
16	3	6	-1	-1	1	-1.11989643	4.91184286
17	3	6	-1	-1	-1	3.4346315881	7.08612184
18	3	6	-1	-1	1	0.2987179478	4.54598547
19	4	7	1	-1	1	-1.245681444	3.06716433
20	4	7	1	-1	1	0.4951842993	2.73610368
21	4	7	1	-1	-1	-0.6602484	-0.0029228
22	4	8	1	1	-1	0.7779088461	-1.22983
23	4	8	1	1	1	7.9432230978	4.64542262
24	4	8	1	1	-1	1.0668195454	-0.9147606

Figure 10.3 Simulate Responses Window

Simulate Responses

Effects	Y
Intercept	1
Temperature	1
Time	1
Catalyst	1
Temperature*Time	1
Temperature*Catalyst	1
Time*Catalyst	1

Distribution

Normal
Error σ: 1
Whole Plots σ: 1
Subplots σ: 1

Binomial
Poisson

Apply

The design table and a Simulate Responses window appear. Notice that the design table contains a **DOE Simulate** script. At any time, you can run this script to specify different parameter values.

Continue to the next section, where you specify standard deviations for the whole plot and subplot errors, and fit a REML model to the first set of simulated values.

JMP PRO Fit the Model

Assume that the whole plot and subplot errors are normal. Based on your estimates of their standard deviations, if the error standard deviation is about 1 unit, the whole plot standard deviation is about 2 units and the subplot standard deviation is about 1.5 units. Since you are only interested in the whole- and sub-plot variation, you do not need to change the values assigned to Effects in the Simulate Responses outline.

1. In the Distribution panel (Figure 10.3), next to **Whole Plots σ**, type 2.

   Notice that the Normal distribution is selected by default, indicating that normal error will be added to the formula.

2. Next to **Subplots σ**, type 1.5.
3. Click **Apply**.

   In the data table, the formula for Y Simulated updates to reflect your specifications. To view the formula, click on the plus sign to the right of the column name in the Columns panel.

4. In the data table, click the green triangle next to the **Model** script.
5. Click the Y variable next to the **Y** button and click **Remove**.
6. Click Y Simulated and click the **Y** button.

   This action replaces Y with a column that contains a simulation formula.

7. Click **Run**.

   The model that is fit is based on a single set of simulated responses.

Note: Because the values in Y Simulated are randomly generated, the entries in your report will differ from those that appear in Figure 10.4.

Figure 10.4 REML Report Showing Wald Confidence Intervals

REML Variance Component Estimates

Random Effect	Var Ratio	Var Component	Std Error	95% Lower	95% Upper	Wald p-Value	Pct of Total
Whole Plots	3.8107856	1.556364	1.8007472	-1.973036	5.0857636	0.3874	68.445
Subplots	0.7568606	0.3091097	0.4661521	-0.604532	1.222751	0.5073	13.594
Residual		0.4084103	0.160228	0.2146172	1.060295		17.961
Total		2.2738839	1.8035082	0.7468087	28.115033		100.000

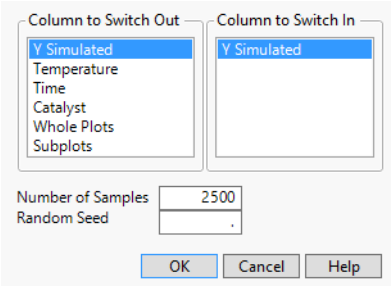
-2 LogLikelihood = 63.890021519
Note: Total is the sum of the positive variance components.
Total including negative estimates = 2.2738839

JMP PRO Explore Power

Next, simulate values for the variance components and use these to construct simulated percentile confidence intervals.

1. In the REML Variance Components Estimates outline, right-click in the **Var Component** column and select **Simulate**.

Figure 10.5 Simulate Window



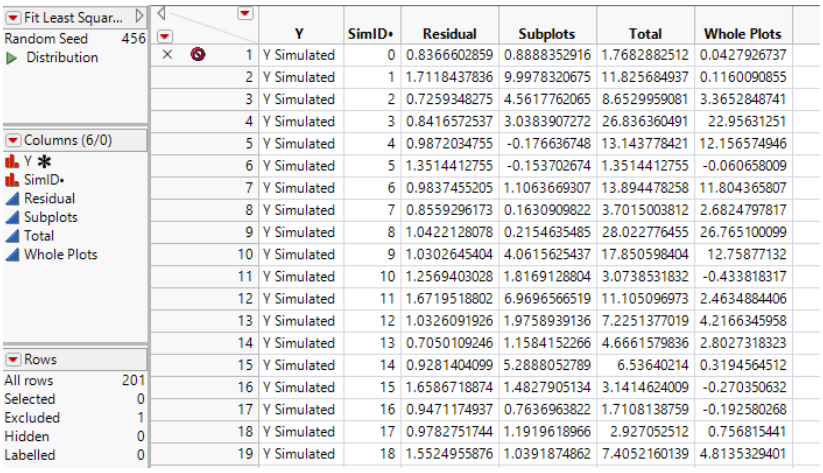
In your simulations, you replace the column Y Simulated, which you used to run your model, with a new instance of the column Y Simulated, which generates a new column of simulated values for each simulation. The column on which you right-clicked and that appears as selected, **Var Component**, will be simulated for each effect listed in the Parameter Estimates table.

2. Next to **Number of Samples**, enter **200**.
3. (Optional) Next to **Random Seed**, enter 456.

   This reproduces the values shown in Figure 10.6, except for the values in row 1.
4. Click **OK**.

   The entries in your row 1 will differ from those that appear in Figure 10.6.

Figure 10.6 Table of Simulated Results for Var Component

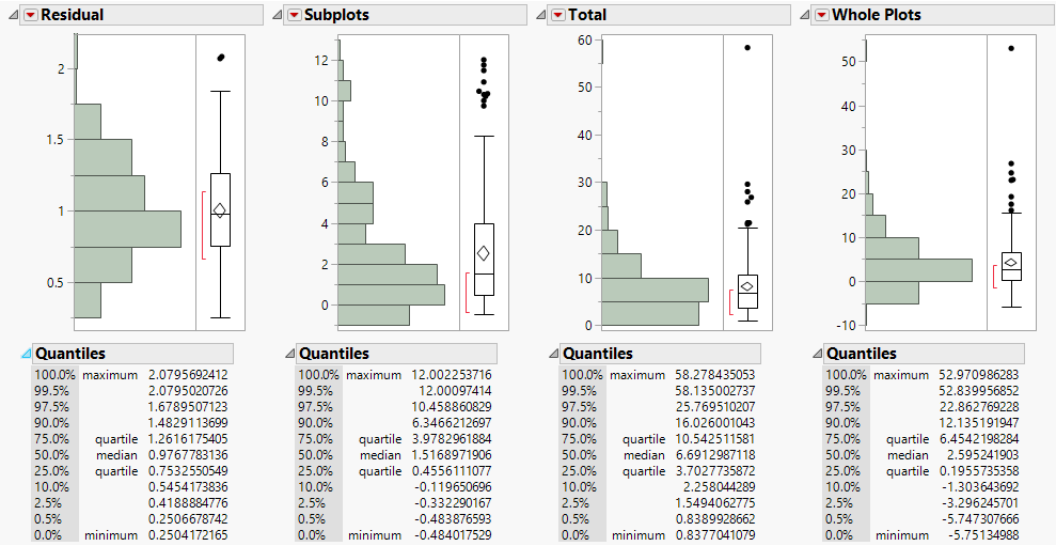


	Y	SimID•	Residual	Subplots	Total	Whole Plots
1	Y Simulated	0	0.8366602859	0.8888352916	1.7682882512	0.0427926737
2	Y Simulated	1	1.7118437836	9.9978320675	11.825684937	0.1160090855
3	Y Simulated	2	0.7259348275	4.5617762065	8.6529959081	3.3652848741
4	Y Simulated	3	0.8416572537	3.0383907272	26.836360491	22.95631251
5	Y Simulated	4	0.9872034755	-0.176636748	13.143778421	12.156574946
6	Y Simulated	5	1.3514412755	-0.153702674	1.3514412755	-0.060658009
7	Y Simulated	6	0.9837455205	1.1063669307	13.894478258	11.804365807
8	Y Simulated	7	0.8559296173	0.1630909822	3.7015003812	2.6824797817
9	Y Simulated	8	1.0422128078	0.2154635485	28.022776455	26.765100099
10	Y Simulated	9	1.0302645404	4.0615625437	17.850598404	12.75877132
11	Y Simulated	10	1.2569403028	1.8169128804	3.0738531832	-0.433818317
12	Y Simulated	11	1.6719518802	6.9696566519	11.105096973	2.4634884406
13	Y Simulated	12	1.0326091926	1.9758939136	7.2251377019	4.2166345958
14	Y Simulated	13	0.7050109246	1.1584152266	4.6661579836	2.8027318323
15	Y Simulated	14	0.9281404099	5.2888052789	6.53640214	0.3194564512
16	Y Simulated	15	1.6586718874	1.4827905134	3.1414624009	-0.270350632
17	Y Simulated	16	0.9471174937	0.7636963822	1.7108138759	-0.192580268
18	Y Simulated	17	0.9782751744	1.1919618966	2.927052512	0.756815441
19	Y Simulated	18	1.5524955876	1.0391874862	7.4052160139	4.8135329401

The first row of the Fit Least Squares Simulate Results (Var Component) data table contains the initial values of **Var Component** and is excluded. The remaining rows contain simulated values.

5. Run the **Distribution** script.

Figure 10.7 Distribution Plots for Variance Components



For each variance component, you can construct a 95% percentile confidence interval using the 2.5% and 97.5% quantiles. Compare these to the intervals given in the REML report (Figure 10.4):

– The simulated 95% percentile interval for the whole-plot variance component is -3.296 to 22.863. The Wald interval given in the REML report is -1.973 to 5.086.

– The simulated 95% percentile interval for the sub-plot variance component is -0.332 to 10.459. The Wald interval given in the REML report is -.605 to 1.22.

The intervals that you obtain using simulation are considerably wider than the REML interval calculated from your single set of values. For even more accurate intervals, consider running a larger number of simulations.

JMP PRO Conduct a Permutation Test

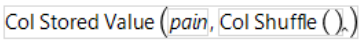
In this example, you are studying the effects of three drugs on pain. You are interested in whether they differ in their effects. Because you have a very small sample size and are somewhat concerned about violations of the usual ANOVA assumptions, you will use Simulate to conduct a permutation test.

First, you construct a formula that randomly shuffles the pain measurements among the three drugs. Under the null hypothesis of no effect, any of these allocations is as likely as any other. It follows that the *F* ratios obtained in this manner approximate the distribution of *F* ratios under the null hypothesis. Finally, you compare the observed value of the *F* ratio to the null distribution obtained by simulation.

JMP PRO Define the Simulation Formula

1. Select **Help > Sample Data Library** and open Analgesics.jmp.
2. Select **Cols > New Columns**.
3. Type Pain Shuffled for Column Name.
4. From the Column Properties list, select **Formula**.
5. In the function list, select **Row > Col Stored Value**.
6. Select pain from the Columns list.
7. Click the insert key (^) in the list of symbols above the editor panel.
8. From the list of functions, select **Random > Col Shuffle**.

Figure 10.8 Completed Formula



This formula randomly shuffles the entries in the pain column.

9. Click **OK** in the Formula Editor window.
10. Click **OK** in the Column Info window.

Perform the Permutation Test

1. Select **Analyze > Fit Y by X**.
2. Select pain and click **Y, Response**.
3. Select drug and click **X, Factor**.
4. Click **OK**.
5. Click the Oneway Analysis red triangle and select **Means/Anova**.

Figure 10.9 Analysis of Variance Report

Analysis of Variance

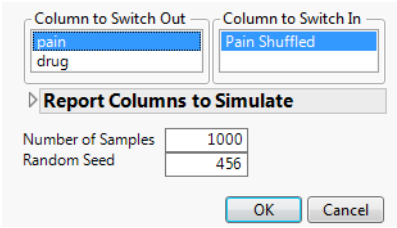
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
drug	2	99.89459	49.9473	6.2780	0.0053*
Error	30	238.67877	7.9560		
C. Total	32	338.57335			

Notice that the *F* ratio is 6.2780.

6. In the Analysis of Variance outline, right-click on the F Ratio column and select **Simulate**.
7. In the Column to Switch Out list, click pain.
8. Next to **Number of Samples**, enter 1000.
9. (Optional) Next to **Random Seed**, enter 456.

   This reproduces the values in this example.

Figure 10.10 Completed Simulate Window

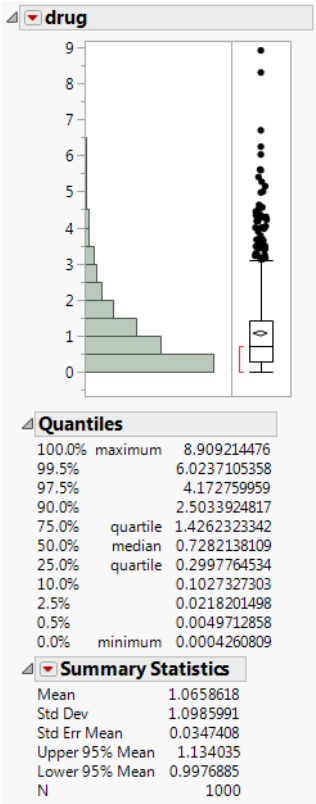


10. Click **OK**.

    In the table of simulated results, the columns for C. Total and Error are empty, since the *F* Ratio value in the Analysis of Variance table only applies to drug.

11. In the table of simulated values, run the **Distribution** script.

Figure 10.11 Simulated Distribution of *F* Ratios under the Null Distribution



The observed *F* ratio value of 6.2780 falls in the upper 0.5% of the simulated null distribution of *F* ratios. This presents strong evidence that the three drugs differ in their effects on pain.

JMP PRO Explore Retaining a Factor in Generalized Regression

In this example, a pharmaceutical manufacturer has historical information on the dissolution of a tablet inside the body and various factors that may affect the dissolution rate. A tablet with a dissolution rate below 70 is considered defective. You want to understand which factors affect dissolution rate.

In this example, you will do the following:

- Construct a generalized regression model.
- Fit a reduced model using the non-zeroed terms.
- Based on the reduced model, use simulation to explore the likelihood that one of the factors is included in the model.

JMP PRO Fit the Model

In this section, you fit a model using generalized regression. If you prefer not to work through the steps in this section, click the green triangle next to the **Generalized Regression** script in the Tablet Production.jmp data table to obtain the model.

1. Select **Help > Sample Data Library** and open Tablet Production.jmp.
2. Select **Analyze > Fit Model**.
3. Click Dissolution and click **Y**.
4. Select Mill Time through Atomizer Pressure and click **Add**.
5. From the Personality menu, select **Generalized Regression**.
6. Click **Run**.
7. In the Model Launch panel, click **Go**.

Figure 10.12 Model Based on Adaptive Lasso

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	110.35025	24.526587	20.242885	<.0001*	62.279021	158.42147
Mill Time	0.1291212	0.0283186	20.789802	<.0001*	0.0736177	0.1846247
Screen Size[3-5]	4.1694529	0.5433106	58.89275	<.0001*	3.1045837	5.234322
Screen Size[4-5]	2.3600819	0.5712373	17.069508	<.0001*	1.2404774	3.4796865
Mag. Stearate Supplier[Jones Inc-Smith Ind]	0	0	0	1.0000	0	0
Lactose Supplier[Bond Inc-James Ind]	0	0	0	1.0000	0	0
Sugar Supplier[Sour-Sweet]	0	0	0	1.0000	0	0
Talc Supplier[Rough-Smooth]	0	0	0	1.0000	0	0
Blend Time	0.7177884	0.1384385	26.883081	<.0001*	0.4464539	0.9891228
Blend Speed	0.1821224	0.2418817	0.5669188	0.4515	-0.291957	0.6562018
Compressor[Compress1-Compress2]	-0.495065	0.3993136	1.5370784	0.2151	-1.277705	0.2875754
Force	0	0	0	1.0000	0	0
Coating Supplier[Coat-Mac]	0	0	0	1.0000	0	0
Coating Supplier[Down-Mac]	0	0	0	1.0000	0	0
Coating Viscosity	0.181504	0.0503564	12.991611	0.0003*	0.0828073	0.2802007
Inlet Temp	0	0	0	1.0000	0	0
Exhaust Temp	0	0	0	1.0000	0	0
Spray Rate	-0.203293	0.0421928	23.214889	<.0001*	-0.285989	-0.120596
Atomizer Pressure	0	0	0	1.0000	0	0
Scale	2.0288847	0.1659965	149.38839	<.0001*	1.7035375	2.3542319

You are interested in the parameter estimates shown in the Adaptive Lasso with AICc Validation report. Based on the non-zero parameter estimates, the model suggests that Mill Time, Screen Size, Blend Time, Blend Speed, Compressor, Coating Viscosity, and Spray Rate are related to Dissolution.

JMP PRO Reduce the Model

Before reducing the model, ensure that no columns are selected in the Tablet Production.jmp data table. Selected columns are not deselected in the first step below. Ensuring that no columns are selected prevents the inadvertent inclusion of columns with zeroed terms.

If you prefer not to work through the steps in this section, click the green triangle next to the **Generalized Regression Reduced Model** script in the Tablet Production.jmp data table to obtain the reduced model.

1. Click the red triangle next to Adaptive Lasso with AICc Validation and select **Relaunch with Active Effects**.

   This opens a Fit Model window that places the terms with nonzero coefficient estimates in the Parameter Estimates reports into the Construct Model Effects list. The response is entered as Y. The Generalized Regression personality is selected.

2. Click **Run**.

3. In the Model Launch panel, click **Go**.

Figure 10.13 Reduced Model Using Adaptive Lasso

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	95.142391	23.430178	16.489099	<.0001*	49.220085	141.0647
Mill Time	0.1394103	0.0275609	25.585971	<.0001*	0.0853919	0.1934288
Screen Size[3-5]	4.3323833	0.534237	65.763638	<.0001*	3.285298	5.3794685
Screen Size[4-5]	2.6331283	0.5457852	23.275584	<.0001*	1.563409	3.7028476
Blend Time	0.7583048	0.1385246	29.966364	<.0001*	0.4868017	1.029808
Blend Speed	0.4575802	0.2289168	3.9955744	0.0456*	0.0089116	0.9062488
Compressor[Compress1-Compress2]	-0.877986	0.4127286	4.5252866	0.0334*	-1.686919	-0.069053
Coating Viscosity	0.198673	0.0486798	16.656386	<.0001*	0.1032624	0.2940836
Spray Rate	-0.212753	0.0410929	26.805238	<.0001*	-0.293294	-0.132213
Scale	1.9982636	0.1574951	160.97992	<.0001*	1.689579	2.3069483

Notice that the estimate for Blend Speed has a confidence interval (Lower 95%) that comes very close to including zero. Next, perform a simulation study to see how often Blend Speed would be included in the model had other data values from the dissolution distribution been observed.

JMP PRO Explore the Inclusion of Blend Speed in the Model

Use the report for the reduced model (Figure 10.13) in the steps below.

1. Click the red triangle next to Adaptive Lasso with AICc Validation and select **Save Columns > Save Simulation Formula**.

   This adds a new column called Dissolution Simulation Formula to the Tablet Production.jmp data table.

2. (Optional) In the data table Columns panel, click the plus sign to the right of Dissolution Simulation Formula.

Figure 10.14 Simulation Formula

```
Random Normal(
  95.142391047
  + 0.1394103417 • Mill Time
  + Match(Screen Size)(
      "3" ⇒ 4.3323832797
      "4" ⇒ 2.6331282831
      "5" ⇒ 0
      else ⇒ .)
  + 0.7583048404 • Blend Time
  + 0.4575801859 • Blend Speed
  + Match(Compressor)(
      "Compress1" ⇒ -0.877985955
      "Compress2" ⇒ 0
      else ⇒ .)
  + 0.1986730026 • Coating Viscosity
  + -0.212753206 • Spray Rate
  , 1.9982636301)
```

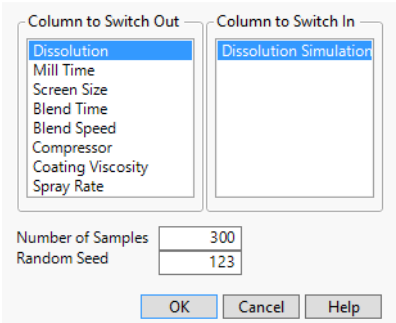
For each row, this formula simulates a value that could be obtained given the model and the distribution of Dissolution, which is estimated to be Normal with standard deviation about 1.998.

3. Click **Cancel**.
4. Go back to the reduced model report window. In the Parameter Estimates for Original Predictors report, right-click in the Estimate column and select **Simulate.**
5. Next to **Number of Samples**, enter 300.

   For the simulation, you ask JMP to replace the Dissolution column in each of 300 analyses with values simulated using the Dissolution Simulated column.
6. (Optional) Set the **Random Seed** to 123.

   This reproduces the values in this example.

Figure 10.15 Completed Simulation Window

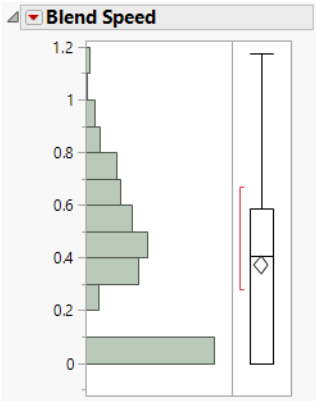


7. Click **OK**.

   The first row of the table contains the initial values of the Estimates and is excluded. The remaining rows contain simulated values.

8. Run the **Distribution** script.
9. Press the Ctrl key, click the Intercept red triangle menu, and select **Display Options > Customize Summary Statistics**.
10. Select **N Zero**.
11. Click **OK**.
12. Scroll to the Distribution report for Blend Speed.

Figure 10.16 Histogram of Simulated Blend Speed Coefficient Estimates



Quantiles

100.0%	maximum	1.1771951386
99.5%		1.1483727707
97.5%		0.9467869842
90.0%		0.7539992295
75.0%	quartile	0.5867988786
50.0%	median	0.4075510721
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.3717879
Std Dev	0.3016753
Std Err Mean	0.0174172
Upper 95% Mean	0.4060638
Lower 95% Mean	0.337512
N	300
N Zero	97

- The Summary Statistics report shows that for 97/300 = 32.3% of the simulations, the Blend Speed estimates are zero. This information suggests that it might make sense to retain Blend Speed in the model.

JMP PRO Conduct Prospective Power Analysis for a Nonlinear Model

In this example, you are interested in the main effects of six continuous factors on whether a part passes or fails inspection. The response is binomial and you can afford a total of 60 runs.

You will do the following:

1. Construct a custom design for your experiment. See "Construct the Design" on page 299.

   Note: Although a custom design is not optimal for a non-linear situation, in this example, for simplicity, you will use the Custom Design platform rather than the Nonlinear Design platform. For an example illustrating why a design constructed using the Nonlinear Design platform is better than an orthogonal design, see the Nonlinear Designs chapter in the *Design of Experiments Guide*.

2. Fit a logistic model using the Generalized Linear Model personality. See "Fit the Generalized Linear Model" on page 302.
3. Simulate likelihood ratio test *p*-values to explore the power of detecting a difference over a range of probability values that is determined by the linear predictor. See"Explore Power" on page 303.

Plan for the Example

You will model the probability of a success using a generalized linear model with the logit as a link function. The logit link function fits a logistic model:

$$\pi(\mathbf{X}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_6 X_6)}}$$

where $\pi(\mathbf{X})$ denotes the probability that a part passes at the given design settings $\mathbf{X} = (X_1, X_2, \ldots, X_6)$.

Denote the linear predictor by $L(\mathbf{X})$:

$$L(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_6 X_6$$

You will explore power for the following values of the coefficients of the linear predictor:

Coefficient	Value
β_0	0
β_1	1
β_2	.9
β_3	.8

Coefficient	Value
β_4	.7
β_5	.6
β_6	.5

Because the intercept in the linear predictor is 0, when all factors are set to 0, the probability of a passing part equals 50%. The probabilities associated with the levels of the ith factor, when all other factors are set to 0, are given below.

Factor	Percent Passing at $X_i = 1$	Percent Passing at $X_i = -1$	Difference
X_1	73.11%	26.89%	46.2%
X_2	71.09%	28.91%	42.2%
X_3	69.00%	31.00%	38.0%
X_4	66.82%	33.18%	33.6%
X_5	64.56%	35.43%	29.1%
X_6	62.25%	37.75%	24.5%

For example, when all factors other than X_1 are set to 0, the difference in pass rates that you want to detect is 46.2%. The smallest difference in pass rates that you want to detect occurs when all factors other than X_6 are set to zero and that difference is 24.5%.

JMP PRO Construct the Design

Note: If you prefer to skip the steps in this section, select **Help > Sample Data Library** and open Design Experiment/Binomial Experiment.jmp. Then go to "Define Simulated Responses" on page 301.

1. Select **DOE > Custom Design**.
2. In the Factors outline, type 6 next to **Add N Factors**.
3. Click **Add Factor > Continuous**.
4. Click **Continue**.

   You are constructing a main effects design, so make no changes to the Model outline.
5. Under Number of Runs, type 60 next to **User Specified**.
6. Click the Custom Design red triangle and select **Simulate Responses**.

This opens the Simulate Responses window after you select Make Table to construct the design table.

Note: Setting the Random Seed in step 7 and Number of Starts in step 8 reproduces the exact results shown in this example. In constructing a design on your own, these steps are not necessary.

7. (Optional) Click the Custom Design red triangle and select **Set Random Seed**. Type 12345 and click **OK**.
8. (Optional) Click the Custom Design red triangle and select **Number of Starts**. Type 1 and click **OK**.
9. Click **Make Design**.
10. Click **Make Table**.

Note: The entries in your Y and Y Simulated columns will differ from those that appear in Figure 10.17.

Figure 10.17 Partial View of Design Table

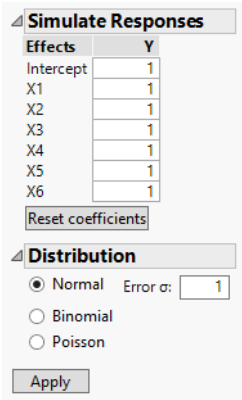
Custom Design
Design Custom Design
Criterion D Optimal
Model
Evaluate Design
DOE Simulate
DOE Dialog

Columns (8/0)
X1
X2
X3
X4
X5
X6
Y
Y Simulated

Rows
All rows 60
Selected 0
Excluded 0
Hidden 0
Labelled 0

	X1	X2	X3	X4	X5	X6	Y	Y Simulated
1	-1	-1	1	1	1	-1	0.5730432553	0.57304326
2	-1	1	-1	1	1	1	4.0344453836	4.03444538
3	-1	-1	-1	1	-1	1	-2.739733732	-2.7397337
4	1	-1	1	-1	1	-1	1.5218419259	1.52184193
5	-1	1	1	1	-1	-1	2.2954601779	2.29546018
6	-1	-1	-1	1	1	1	0.8583835188	0.85838352
7	-1	1	1	-1	-1	1	2.4624020338	2.46240203
8	1	-1	-1	1	1	1	4.7517212278	4.75172123
9	1	-1	-1	-1	-1	-1	-4.542667424	-4.5426674
10	1	-1	1	1	1	-1	2.5177895575	2.51778956
11	1	1	-1	1	1	-1	3.1954312663	3.19543127
12	-1	1	-1	-1	1	-1	-1.546124055	-1.5461241
13	1	1	1	1	1	1	6.0524139401	6.05241394
14	1	-1	1	-1	1	-1	1.7372052107	1.73720521
15	-1	1	-1	-1	1	-1	-1.427486276	-1.4274863
16	1	1	-1	1	1	-1	3.8036692724	3.80366927
17	1	1	-1	1	-1	-1	2.7041134812	2.70411348
18	1	1	-1	-1	-1	1	2.1959890994	2.1959891
19	-1	1	1	-1	1	1	3.2032518472	3.20325185
20	-1	-1	-1	-1	-1	-1	-3.088770371	-3.0887704
21	1	1	1	1	1	1	7.97396176	7.97396176
22	-1	-1	-1	-1	1	-1	-4.465207658	-4.4652077
23	-1	1	-1	1	1	1	4.3641989062	4.36419891

Figure 10.18 Simulate Responses Window



The design table and a Simulate Responses window appear. Two columns are added to the design table:

- Y contains a set of values simulated according to the specifications in the Simulate Responses window.
- Y Simulated contains a formula that calculates its values using the formula for the model that is specified in the Simulate Responses window. To view the formula, click on the plus sign to the right of the column name in the Columns panel.

Continue to the next section, where you simulate binomial responses and fit a generalized linear model to these simulated responses.

JMP PRO Define Simulated Responses

Your plan is to simulate binomial response data where the probability of success is given by a logistic model. For more information on Simulate Response, see the Custom Designs chapter in the *Design of Experiments Guide*.

Note: If you prefer to skip the steps in this section, click the green triangle next to the **Simulate Model Responses** script. Then go to "Fit the Generalized Linear Model" on page 302.

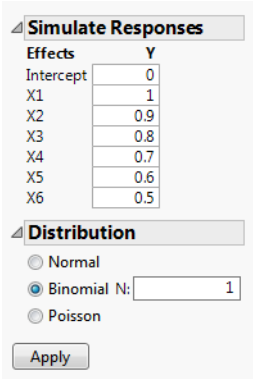
1. In the Simulate Responses window (Figure 10.19), type the following values under Y:
   - Next to Intercept, type 0.
   - Next to X1, 1 is entered by default. Keep that value.
   - Next to X2, type 0.9.
   - Next to X3, type 0.8.
   - Next to X4, type 0.7.
   - Next to X5, type 0.6.

– Next to X6, type 0.5.

2. In the Distribution outline, select **Binomial**.

   Leave the value for **N** set to 1, indicating that there is only one unit per trial.

Figure 10.19 Completed Simulate Responses Window

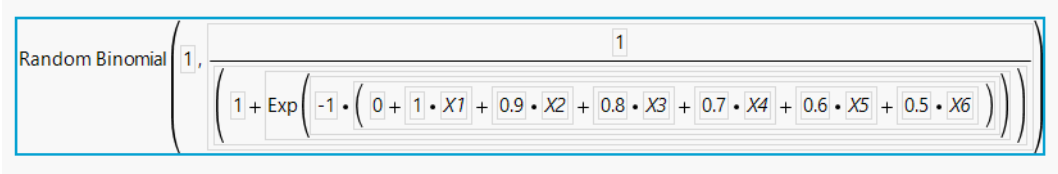


3. Click **Apply**.

   In the design data table, the Y Simulated column is replaced with a formula column that generates binomial values. A column called Y N Trials indicates the number of trials for each run.

4. (Optional) Click on the plus sign to the right of Y Simulated in the Columns panel.

Figure 10.20 Random Binomial Formula for Y Simulated



5. Click **Cancel**.

JMP PRO Fit the Generalized Linear Model

1. In the data table, click the green triangle next to the **Model** script.
2. Click the Y variable next to the **Y** button and click **Remove**.
3. Click Y Simulated and click the **Y** button.

   You are replacing Y with a column that contains randomly generated binomial values.

4. From the Personality menu, select **Generalized Linear Model**.

5. From the Distribution menu, select **Binomial**.

   Notice that the Logit function appears in the Link Function menu.

6. Click **Run**.

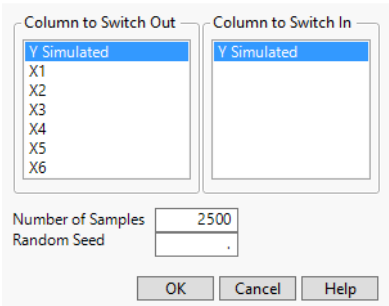
   The model that is fit is based on a single set of simulated binomial responses.

JMP PRO Explore Power

Next, explore the power of tests to detect a difference over the range of probability values determined by the linear predictor with the coefficient values given in "Plan for the Example" on page 298.

1. In the Effect Tests outline, right-click in the **Prob>ChiSq** column and select **Simulate**.

Figure 10.21 Simulate Window



   The column Y Simulated under the Column to Switch Out contains the values that were used to fit the model. When you select the column Y Simulated under Column to Switch In, for each simulation, you are telling JMP to replace the values in Y Simulated with a new column of values that are simulated using the formula in the column Y Simulated.

   The column you have selected in the report, **Prob>ChiSq**, is the *p*-value for a likelihood ratio test of whether the associated main effect is 0. The Prob>ChiSq value will be simulated for each effect listed in the Effect Tests table.

2. Next to **Number of Samples**, enter **500**.

3. Next to **Random Seed**, enter **123** and then click outside the text box.

   Setting the random seed ensures that all rows in your Simulate Results table, except the first, will match those in Figure 10.22.

4. Click **OK**.

   A Generalized Linear Model Simulate Results data table appears.

Note: Because response values are simulated, your simulated *p*-values will differ from those shown in Figure 10.22.

Figure 10.22 Table of Simulated Results, Partial View

Generalized Lin...
Random Seed 123
Distribution
Power Analysis

Columns (7/0)
SimID•
X1
X2
X3
X4
X5
X6

Rows
All rows 501
Selected 0
Excluded 1
Hidden 0
Labelled 0

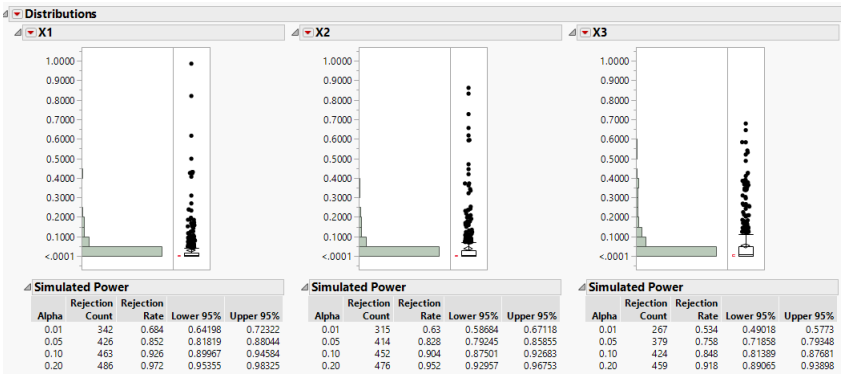
	SimID•	X1	X2	X3	X4	X5	X6
1	0	0.0034	0.0934	0.1127	0.1189	0.0160	0.1261
2	1	0.0001	0.0065	0.1445	0.9470	0.0263	0.1742
3	2	0.0152	0.0128	0.0129	0.2524	0.4177	0.0463
4	3	0.0017	0.0330	0.0011	0.0200	<.0001	0.1135
5	4	0.0002	0.0761	0.0003	0.1305	0.1993	0.6392
6	5	0.0179	0.1074	0.0310	0.0314	0.0765	0.1103
7	6	0.1678	0.0060	0.0004	0.0004	0.0015	0.1920
8	7	0.3097	0.0003	0.0010	0.1448	0.0003	0.0157
9	8	0.0183	0.0005	0.4225	0.0005	0.1902	0.3351
10	9	0.0213	<.0001	0.0076	0.0039	0.3077	0.2824
11	10	0.0873	0.0039	0.0107	0.0014	0.0006	0.4335
12	11	0.1144	0.1976	0.1020	0.0002	0.0459	0.5622
13	12	0.0371	0.0181	<.0001	0.2194	0.0032	0.2504
14	13	0.0005	0.0002	<.0001	0.0032	0.0028	0.9486
15	14	0.0022	0.0035	0.0529	0.0054	0.0016	0.1947
16	15	0.0184	0.0009	<.0001	0.0073	0.0747	0.9105
17	16	0.0012	<.0001	<.0001	0.0018	0.1585	0.0230
18	17	0.0044	<.0001	0.0018	0.0051	0.7172	0.0007
19	18	0.0049	0.0099	0.0009	0.2483	0.0822	0.4739

The first row of the table contains the initial values of **Prob>ChiSq** and is excluded. The remaining 500 rows contain simulated values.

5. Run the **Power Analysis** script.

Note: Because response values are simulated, your simulated power results will differ from those shown in Figure 10.23.

Figure 10.23 Distribution Plots for the First Three Effects



The histograms plot the 500 simulated Prob>ChiSq values for each main effect. The Simulated Power outline shows the simulated Rejection Rate in the 500 simulations.

For easier viewing, stack the reports and de-select the plots, as follows.

6. Click the Distributions red triangle and select **Stack**.

7. Press Ctrl and click the X1 red triangle, and de-select **Outlier Box Plot**.
8. Press Ctrl and click the X1 red triangle, then select **Histogram Options** and de-select **Histogram**.

Note: Because response values are simulated, your simulated power results will differ from those shown in Figure 10.24.

Figure 10.24 Power Results

Distributions

X1

Simulated Power

Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	342	0.684	0.64198	0.72322
0.05	426	0.852	0.81819	0.88044
0.10	463	0.926	0.89967	0.94584
0.20	486	0.972	0.95355	0.98325

X2

Simulated Power

Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	315	0.63	0.58684	0.67118
0.05	414	0.828	0.79245	0.85855
0.10	452	0.904	0.87501	0.92683
0.20	476	0.952	0.92957	0.96753

X3

Simulated Power

Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	267	0.534	0.49018	0.5773
0.05	379	0.758	0.71858	0.79348
0.10	424	0.848	0.81389	0.87681
0.20	459	0.918	0.89065	0.93898

X4

Simulated Power

Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	212	0.424	0.38142	0.46773
0.05	327	0.654	0.61127	0.69438
0.10	372	0.744	0.70399	0.78029
0.20	422	0.844	0.80959	0.87317

X5

Simulated Power

Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	135	0.27	0.23295	0.31056
0.05	244	0.488	0.44445	0.53174
0.10	304	0.608	0.56454	0.64981
0.20	361	0.722	0.68115	0.75946

X6

Simulated Power

Alpha	Rejection Count	Rejection Rate	Lower 95%	Upper 95%
0.01	86	0.172	0.14145	0.20755
0.05	186	0.372	0.33076	0.41519
0.10	240	0.48	0.43653	0.52378
0.20	306	0.612	0.56859	0.6537

In the Simulated Power outlines, the Rejection Rate for each row gives the proportion of *p*-values that are smaller that the corresponding Alpha. For example, for X3, which corresponds to a coefficient value of 0.8 and a probability difference of 38%, the simulated power for a 0.05 significance level is 379/500 = 0.758. Table 10.1 summarizes the estimated power at the 0.05 significance level for all effects. Notice how power decreases as the Difference to Detect decreases. Also notice that the power to detect an effect as large as 24.5% (X6) is only approximately 0.37.

Table 10.1 Simulated Power at Significance Level 0.05

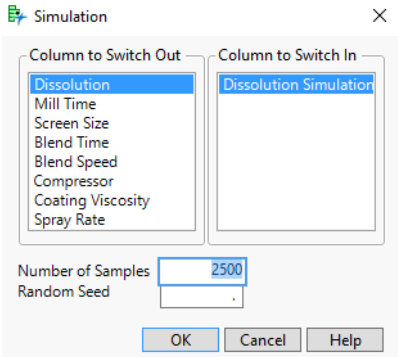
Factor	Percent Passing at $X_i = 1$	Percent Passing at $X_i = -1$	Difference to Detect	Simulated Power (Rejection Rate) at Alpha=0.05
X_1	73.11%	26.89%	46.2%	0.852
X_2	71.09%	28.91%	42.2%	0.828
X_3	69.00%	31.00%	38.0%	0.758
X_4	66.82%	33.18%	33.6%	0.654
X_5	64.56%	35.43%	29.1%	0.488
X_6	62.25%	37.75%	24.5%	0.372

JMP PRO Launch the Simulate Window

To launch the Simulate window, right-click on a column of calculated values in a report window and select Simulate. Simulate is available in many reports, including all reports that support bootstrapping. To use Simulate, the data table must contain a formula with a random component that simulates data.

The Simulate Window

Figure 10.25 Simulate Window for Tablet Production.jmp



The Simulate window contains these panels and options:

Column to Switch Out The column that is replaced by the Column to Switch In.

Column to Switch In The column that replaces replace the Column to Switch Out. The analysis will be repeated with values simulated according to the formula in the Column to Switch In. Only columns with formulas are listed in the Column to Switch In panel.

Number of Samples Number of times that the report is re-run for a set of simulated data. The default value is 2500.

Random Seed A value that controls the simulated results. The random seed makes the results reproducible.

When you click OK in the Simulation window, a window that shows a progress bar and a Stop Early button appears. The number of the sample being simulated is shown above the progress bar. If you click Stop Early, the simulated values that have been computed up to that point are presented in a Simulate Results table. The window also shows you which analyses are being run at any given time.

The Simulate Table

Simulate results appear in a table. Note the following:

- The first row of the table contains the values for the table items that appear in the report. For this reason, the first row is always excluded.
- The remaining rows give the simulation results. The number of remaining rows is equal to the Number of Samples you specified in the Simulate launch window.

- The rows in the report are identified by the first column in the report table that contains the selected column of calculated values. A column appears in the simulated results table for each item in this first column.
- The table contains the following scripts that facilitate analysis:
  - The **Distribution** script constructs a Distribution report consisting of plots, quantiles, and summary statistics for each column in the simulated results table. If the values in the simulated results table have a PValue format, a Simulated Power report is also provided. See "Simulated Power Report" on page 308.
  - The **Power Analysis** script appears only if you have simulated a column of *p*-values. It constructs a Distribution report showing histograms of *p*-values and provides a Simulated Power report. See "Simulated Power Report" on page 308.

JMP PRO Simulated Power Report

Alpha The significance level: 0.01, 0.05, 0.10, and 0.20.

Rejection Count The number of simulations where the test rejects at the corresponding significance level.

Rejection Rate The proportion of simulations where the test rejects at the corresponding significance level.

Lower 95% and Upper 95% Lower and upper limits for a 95% confidence interval for the simulated rejection rate. The interval is computed using the Wilson score method. See Wilson (1927).

Tip: Increase the Number of Samples for a narrower confidence interval.

Chapter 11

JMP PRO Bootstrapping

Approximate the Distribution of a Statistic through Resampling

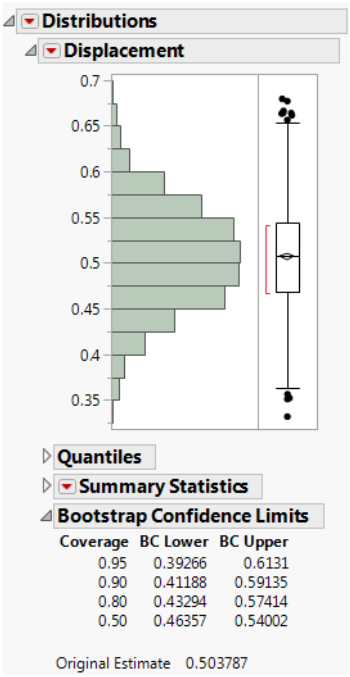
Bootstrapping is available only in JMP Pro.

Bootstrapping is a resampling method for approximating the sampling distribution of a statistic. You can use bootstrapping to estimate the distribution of a statistic and its properties, such as its mean, bias, standard error, and confidence intervals. Bootstrapping is especially useful in the following situations:

- The theoretical distribution of the statistic is complicated or unknown.
- Inference using parametric methods is not possible because of violations of assumptions.

Note: Bootstrap is available only from a right-click in a report. It is not a platform command.

Figure 11.1 Bootstrapping Results for a Slope Parameter



JMP PRO Overview of Bootstrapping

Bootstrapping repeatedly resamples the observations that are used in your report to construct an estimate of the distribution of a statistic or statistics. The observations are assumed to be independent.

In the simple bootstrap, the n observations are resampled with replacement to produce a bootstrap sample of size n. Note that some observations might not appear in the bootstrap sample, and others might appear multiple times. The number of times that an observation occurs in the bootstrap sample is called its *bootstrap weight*. For each bootstrap iteration, the entire analysis that produced the statistic of interest is rerun with these changes:

- the bootstrap sample of n observations is the data set
- the bootstrap weight is a frequency variable in the analysis platform

This process is repeated to produce a distribution of values for the statistic or statistics of interest.

However, the simple bootstrap can sometimes be inadequate. For example, suppose your data set is small or you have a logistic regression setting where you can encounter separation issues. In such cases, JMP enables you to conduct Bayesian bootstrapping using fractional weights. When fractional weights are used, a fractional weight is associated with each observation. The fractional weights sum to n. The statistic of interest is computed by treating the fractional weights as a frequency variable in the analysis platform. For information about fractional weights, see "Fractional Weights" on page 313 and "Calculation of Fractional Weights" on page 322.

To run a bootstrap analysis in a report, right-click in a table column that contains the statistic that you want to bootstrap and select Bootstrap.

Note: Bootstrap is available only from a right-click in a report. It is not a platform command.

JMP provides bootstrapping in most statistical platforms. The observations that comprise the sample are all observations that are used in the calculations for the statistics of interest. If the report uses a frequency column, the observations in that column are treated as if they were repeated the number of times indicated by the Freq variable. If the report uses a Weight variable, Bootstrap treats it as it was treated in the calculations for the report.

Tip: Bootstrap reruns the entire analysis that appears in the platform report from which Bootstrap is invoked. As a result, Bootstrap might run slowly for your selected column because of extraneous analyses in the report. If Bootstrap is running slowly, remove extraneous options from the platform report before running Bootstrap.

JMP PRO Example of Bootstrapping

This example uses the Car Physical Data.jmp sample data table. A tire manufacturer wants to predict an engine's horsepower from the engine's displacement (in^3). The company is most interested in estimating the slope of the relationship between the variables. The slope values help the company predict the corresponding change in horsepower when the displacement changes.

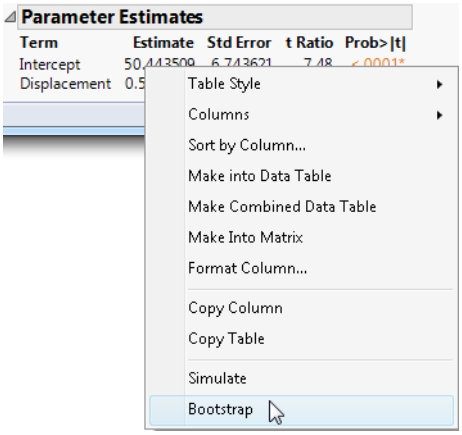
In this example, the regression assumption of homogeneity of variance is violated, so the confidence limits from the regression analysis for the slope might be misleading. For this reason, the company uses a bootstrap estimate of the confidence interval for the slope.

1. Select **Help > Sample Data Library** and open Car Physical Data.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Horsepower and click **Y, Response**.
4. Select Displacement and click **X, Factor**.
5. Click **OK**.
6. Select **Fit Line** from the Bivariate Fit red triangle menu.

   The slope estimate is 0.503787, approximately 0.504.
7. (Optional) Right-click in the **Parameter Estimates** report and select **Columns > Lower 95%**.
8. (Optional) Right-click in the **Parameter Estimates** report and select **Columns > Upper 95%**.

   The confidence limits from the regression analysis for the slope are 0.4249038 and 0.5826711.
9. Right-click the **Estimate** column in the **Parameter Estimates** report and select **Bootstrap** (Figure 11.2).

Figure 11.2 The Bootstrap Option



The column that you right-click is relevant when the **Split Selected Column** option is selected. For more information, see "Bootstrapping Window Options" on page 313.

10. Type 1000 for the **Number of Bootstrap Samples**.
11. (Optional) To match the results in Figure 11.3, type 12345 for the **Random Seed**.
12. Click **OK**.

    The bootstrap process runs and produces a Bootstrap Results data table with unstacked results for the slope and intercept.

    Next, analyze the bootstrapped slope.

13. In the Bootstrap Results table, run the Distribution script.

    The Distribution report includes the Bootstrap Confidence Limits report (Figure 11.3).

Figure 11.3 Bootstrap Report

Distributions

Displacement

Quantiles

100.0%	maximum	0.6800107545
99.5%		0.6483289879
97.5%		0.6212501769
90.0%		0.5781520893
75.0%	quartile	0.5423786474
50.0%	median	0.5051381996
25.0%	quartile	0.4692123772
10.0%		0.4374744231
2.5%		0.4035615438
0.5%		0.3686834278
0.0%	minimum	0.3494461047

Summary Statistics

Bootstrap Confidence Limits

Coverage	BC Lower	BC Upper
0.95	0.40028	0.61892
0.90	0.41646	0.59849
0.80	0.43405	0.57489
0.50	0.46765	0.54007

Original Estimate 0.503787

The estimate of the slope (step 6) is 0.504. Based on the bootstrap results for 95% coverage, the company can estimate the slope to be between 0.40028 and 0.61892. When the displacement is changed by one unit, with 95% confidence, the horsepower changes by some amount between 0.40028 and 0.61892. The bootstrap confidence interval for the slope (0.400 to 0.619) is slightly wider than the confidence interval (0.425 to 0.583) obtained using the usual regression assumptions in step 7 and step 8.

Note: The BC Lower and BC Upper columns in the Bootstrap Confidence Limits report refer to *bias-corrected intervals*. See "Bias-Corrected Percentile Intervals" on page 322.

JMP PRO Bootstrapping Window Options

To perform a bootstrap analysis, right-click on a numeric column of sample statistics in a table in a report window and select **Bootstrap**. The selected column is highlighted, and the Bootstrapping window appears. After you select options and click **OK** in the Bootstrapping window, bootstrap results for every statistic in the column appear in the default results table.

The Bootstrapping window contains the following options:

Number of Bootstrap Samples Sets the number of times that you want to resample the data and compute the statistics. A larger number results in more precise estimates of the statistics' properties. By default, the number of bootstrap samples is set to 2500.

Random Seed Sets a random seed that you can re-enter in subsequent runs of the bootstrap analysis to duplicate your current results. By default, no seed is set.

Fractional Weights Performs a Bayesian bootstrap analysis. In each bootstrap iteration, each observation is assigned a weight that is calculated as described in "Calculation of Fractional Weights" on page 322. The weighted observations are used in computing the statistics of interest. By default, the Fractional Weights option is not selected and a simple bootstrap analysis is conducted.

Tip: Use the Fractional Weights option if the number of observations that are used in your analysis is small or if you are concerned about separation in a logistic regression setting.

Suppose that Fractional Weights is selected. For each bootstrap iteration, each observation that is used in the report is assigned a nonzero weight. These weights sum to n, the number of observations used in the calculations of the statistics of interest. For more information about how the weights are calculated and used, see "Calculation of Fractional Weights" on page 322.

Split Selected Column Places bootstrap results for each statistic in the column that you selected for bootstrapping into a *separate* column in the Bootstrap Results table. Each row of the Bootstrap Results table (other than the first) corresponds to a single bootstrap sample.

If you deselect this option, a Stacked Bootstrap Results table appears. For each bootstrap iteration, this table contains results for the entire report table that contains the column that you selected for bootstrapping. Results for each row of the report table appear as rows in the Stacked Bootstrap Results table. Each column in the report table defines a column in the Stacked Bootstrap Results table. For an example, see "Stacked Results Table" on page 314.

Discard Stacked Table if Split Works (Applicable only if the **Split Selected Column** option is selected.) Determines the number of results tables produced by Bootstrap.

If the Discard Stacked Table if Split Works option *is not* selected, then two Bootstrap tables are shown:

– The Stacked Bootstrap Results table, which contains bootstrap results for each row of the table containing the column that you selected for bootstrapping. This table gives bootstrap results for every statistic in the report, where each column is defined by a statistic.

– The unstacked Bootstrap Results table, which is obtained by splitting the stacked table and providing results only for the column that is selected in the original report.

If the Discard Stacked Table if Split Works option *is* selected and if the **Split Selected Column** operation is successful, the Stacked Bootstrap Results table is not shown.

JMP PRO Stacked Results Table

The initial results of a bootstrap analysis appear in a stacked results table (Figure 11.4). This table might not appear if you have selected the **Discard Stacked Table if Split Works** option. Figure 11.4 shows a bootstrap table that is based on the Parameter Estimates report obtained by fitting a Bivariate model in Fit Y by X to Car Physical Data.jmp. See "Overview of Bootstrapping" on page 310.

Figure 11.4 Stacked Bootstrap Results Table

	X	Y	Term	~Bias	Estimate	Std Error	t Ratio	Prob>|t|	BootID•
1	Displacement	Horsepower	Intercept		50.443509471	6.7436209999	7.48	<.0001	0
2	Displacement	Horsepower	Displac...		0.5037874592	0.0398202611	12.65	<.0001	0
3	Displacement	Horsepower	Intercept		45.385604758	5.4724273046	8.29	<.0001	1
4	Displacement	Horsepower	Displac...		0.5451674092	0.0306151772	17.81	<.0001	1
5	Displacement	Horsepower	Intercept		40.843862813	7.0508187326	5.79	<.0001	2
6	Displacement	Horsepower	Displac...		0.5854988173	0.0457041828	12.81	<.0001	2
7	Displacement	Horsepower	Intercept		47.765642104	5.4677610087	8.74	<.0001	3
8	Displacement	Horsepower	Displac...		0.4943459579	0.0305765677	16.17	<.0001	3
9	Displacement	Horsepower	Intercept		59.758060908	7.0436720036	8.48	<.0001	4

Note the following about the stacked results table:

- For each bootstrap sample, there is a row for each value given in the first column of the report table. These values are shown in a column whose name is the name of the first column in the report table. In this example, for each bootstrap sample there is a row containing results for each Term: Intercept and Displacement, which appear in the Term column.
- The data table columns that are used in the analysis appear in the table. In this example, X is Displacement, and Y is Horsepower.

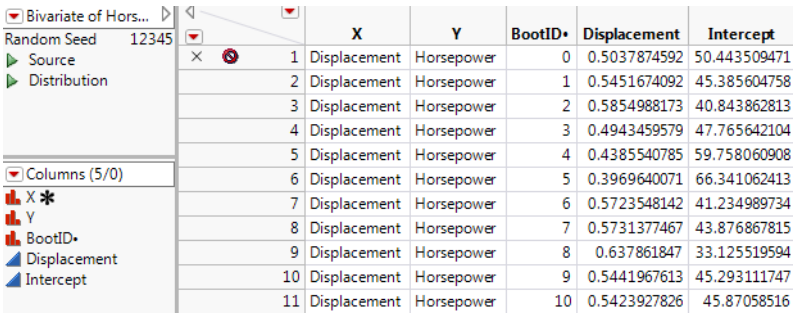
- There is a column for every column in the report table that you are bootstrapping. In this example, the columns are **~Bias**, Estimate, Std Error, t Ratio, and Prob>|t|. Note that **~Bias** is a column in the Fit Y by X report that is hidden unless one of the parameter estimates is biased.
- The BootID• column identifies the bootstrap sample. The rows where BootID• = 0 correspond to the original estimates. Those rows are marked with an X and have the excluded row state. In this example, each bootstrap sample is used to calculate results for two rows: the results for Intercept and the results for Horsepower.
- The data table name begins with "Stacked Bootstrap Results".

If you selected the **Split Selected Column** option, an unstacked results table might also appear. See "Unstacked Bootstrap Results Table" on page 315.

JMP PRO Unstacked Bootstrap Results Table

Select **Split Selected Column** to create a bootstrap table that contains separate columns for the report column that you selected. Each column corresponds to a Term in the report table. For example, in Figure 11.5, the Estimate column from Figure 11.4 is split into two columns (Displacement and Intercept), one for each level of Term.

Figure 11.5 Unstacked Bootstrap Results Table



	X	Y	BootID•	Displacement	Intercept
1	Displacement	Horsepower	0	0.5037874592	50.443509471
2	Displacement	Horsepower	1	0.5451674092	45.385604758
3	Displacement	Horsepower	2	0.5854988173	40.843862813
4	Displacement	Horsepower	3	0.4943459579	47.765642104
5	Displacement	Horsepower	4	0.4385540785	59.758060908
6	Displacement	Horsepower	5	0.3969640071	66.341062413
7	Displacement	Horsepower	6	0.5723548142	41.234989734
8	Displacement	Horsepower	7	0.5731377467	43.876867815
9	Displacement	Horsepower	8	0.637861847	33.125519594
10	Displacement	Horsepower	9	0.5441967613	45.293111747
11	Displacement	Horsepower	10	0.5423927826	45.87058516

Note the following about the unstacked results table:

- There is a single row for each bootstrap sample.
- The data table columns used in the analysis appear in the table. In this example, X is Displacement, and Y is Horsepower.
- There is a column for each row of the report that was bootstrapped.
- If you specified a Random Seed in the Bootstrapping window, the bootstrap results table contains a table variable called Random Seed that gives its value.

- The unstacked bootstrap results table contains a Source table script and a Distribution table script. The Distribution table script enables you to quickly obtain statistics based on the bootstrap samples, including bootstrap confidence intervals.
- The BootID• column identifies the bootstrap sample. The row where BootID• = 0 corresponds to the original estimates. That row is marked with an X and has the excluded row state. In the unstacked bootstrap table, each row is calculated from a single bootstrap sample.
- The data table name ends with "Bootstrap Results (<colname>)", where <colname> identifies the column in the report that was bootstrapped.

JMP PRO Analysis of Bootstrap Results

Analyze your bootstrap results using the Distribution platform:

- If your analysis produced an unstacked bootstrap results table, run the Distribution script in the table.
- If your analysis produced a stacked bootstrap results table, select **Analyze > Distribution** and assign the columns of interest to the appropriate roles. In most cases, it is appropriate to assign the column that corresponds to the first column in the report table to the By role.

The Distribution platform provides summary statistics for your bootstrap results. It also produces a Bootstrap Confidence Limits report for any table that contains a BootID• column (Figure 11.6).

You can use the Distribution report to obtain two types of bootstrap confidence intervals:

- The Quantiles report provides *percentile intervals*. For example, to construct a 95% confidence interval using the percentile method, use the 2.5% and 97.5% quantiles as the interval bounds.
- The Bootstrap Confidence Limits report provides *bias-corrected percentile intervals*. The report shows intervals with 95%, 90%, 80%, and 50% coverage levels. The BC Lower and BC Upper columns show the lower and upper endpoints, respectively. For more information about the computation of the bias-corrected percentile intervals, see "Bias-Corrected Percentile Intervals" on page 322.

Figure 11.6 Bootstrap Confidence Limits Report

Distributions

Displacement

Quantiles

100.0%	maximum	0.6800107545
99.5%		0.6483289879
97.5%		0.6212501769
90.0%		0.5781520893
75.0%	quartile	0.5423786474
50.0%	median	0.5051381996
25.0%	quartile	0.4692123772
10.0%		0.4374744231
2.5%		0.4035615438
0.5%		0.3686834278
0.0%	minimum	0.3494461047

Summary Statistics

Bootstrap Confidence Limits

Coverage	BC Lower	BC Upper
0.95	0.40028	0.61892
0.90	0.41646	0.59849
0.80	0.43405	0.57489
0.50	0.46765	0.54007

Original Estimate 0.503787

The **Original Estimate** at the bottom of the Bootstrap Confidence Limits report is the estimate of the statistic using the original data.

For more information about interpreting the Bootstrap Confidence Limits report, see "Overview of Bootstrapping" on page 310. Efron (1981) describes the methods for both the percentile interval and the bias-corrected percentile interval.

JMP PRO Additional Example of Bootstrapping

This example illustrates the benefits of the Fractional Weights (Bayesian Bootstrap) option for a small data table. The data consist of a response, Y, measured on three samples of each of seven different soil types. A scientist is interested in finding a confidence interval for the mean response for the wabash soil type.

Because each soil type has only three observations, the simple bootstrap has the potential to exclude all three of the observations for wabash from a bootstrap sample. The Fractional Weights option ensures that all observations for every soil type are represented in all bootstrap samples.

The scientist examines the distribution of wabash sample means from both bootstrap methods:

- "Simple Bootstrap Analysis" on page 318
- "Bayesian Bootstrap Analysis" on page 320

Simple Bootstrap Analysis

1. Select **Help > Sample Data Library** and open Snapdragon.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Y and click **Y, Response**.
4. Select Soil and click **X, Factor**.
5. Click **OK**.
6. Select **Means/Anova** from the Oneway Analysis red triangle menu.
7. In the **Means for Oneway Anova** report, right-click the **Mean** column and select **Bootstrap**.
8. Type 1000 for the **Number of Bootstrap Samples**.
9. (Optional) To match the results in Figure 11.7, type 12345 for the **Random Seed**.
10. Click **OK**.

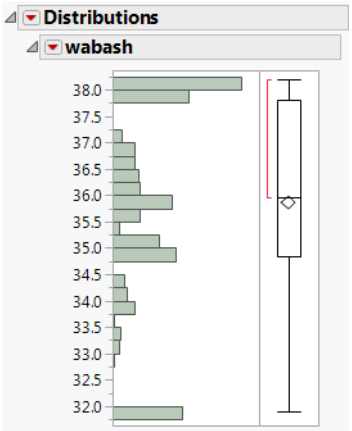
Figure 11.7 Bootstrap Results for a Simple Bootstrap

	X	Y	BootID•	clarion	clinton	compost	knox	o'neill	wabash	webster
1	Soil	Y	0	32.1667	30.3000	29.6667	34.9000	33.8000	35.9667	31.1000
2	Soil	Y	1	32.2333	30.9000	28.0000	35.7000	34.3500	31.9000	31.1000
3	Soil	Y	2	32.4333	30.3000	30.2000	33.9667	31.2000	38.0000	•
4	Soil	Y	3	32.5000	30.7500	29.2000	34.0333	32.8000	38.0000	31.1000
5	Soil	Y	4	31.5000	29.4000	29.9000	35.7500	35.0400	34.9500	•
6	Soil	Y	5	32.5000	30.6000	31.8000	35.7000	34.3000	35.0500	31.5667
7	Soil	Y	6	32.4000	30.1000	29.8500	34.4500	36.0000	38.2000	30.4000
8	Soil	Y	7	31.9000	29.3400	•	34.4000	33.6000	35.9667	31.4500
9	Soil	Y	8	32.1400	31.3500	29.6667	33.1000	35.1000	37.9333	30.6333
10	Soil	Y	9	32.7000	30.8000	30.2000	35.2600	31.2000	34.8500	32.5000
11	Soil	Y	10	32.1000	32.1000	28.0000	33.1000	34.1143	34.0000	31.8000
12	Soil	Y	11	31.5000	30.3000	•	33.1000	34.4000	37.2125	30.6333
13	Soil	Y	12	32.7000	30.4200	31.8000	34.0333	35.1000	35.0500	•
14	Soil	Y	13	32.1667	•	30.1000	34.9000	33.9000	38.2000	31.8000
15	Soil	Y	14	32.3000	•	29.2000	33.9667	34.4000	35.6000	31.9400
16	Soil	Y	15	32.2500	29.1000	29.6667	33.1000	35.1000	31.9000	•
17	Soil	Y	16	31.5000	30.4200	30.0000	•	35.2800	37.8000	31.1000
18	Soil	Y	17	32.4333	29.7000	29.2000	33.1000	34.4000	34.0000	31.1000
19	Soil	Y	18	32.3800	32.1000	29.9000	35.8000	34.6800	35.0500	31.8000
20	Soil	Y	19	32.0600	30.1000	30.2800	•	33.1500	31.9000	31.1000

The missing values in Figure 11.7 represent bootstrap iterations in which none of the observations for a given soil type were selected for the bootstrap sample.

11. Select **Analyze > Distribution**.
12. Select wabash and click **Y, Columns**.
13. Click **OK**.

Figure 11.8 Distribution of wabash Means from a Simple Bootstrap



Distributions

wabash

Quantiles

100.0%	maximum	38.2
99.5%		38.2
97.5%		38.2
90.0%		38.066666667
75.0%	quartile	37.8
50.0%	median	35.966666667
25.0%	quartile	34.85
10.0%		31.9
2.5%		31.9
0.5%		31.9
0.0%	minimum	31.9

Summary Statistics

Mean	35.869561
Std Dev	1.9461801
Std Err Mean	0.06278
Upper 95% Mean	35.992763
Lower 95% Mean	35.746359
N	961

Bootstrap Confidence Limits

Coverage	BC Lower	BC Upper
0.95	31.9	38.2
0.90	31.9	38.2
0.80	33.16	38.1
0.50	34.95	37.9333

Original Estimate 35.96667

Figure 11.8 shows the distribution of wabash means from the simple bootstrap analysis. Notice the following:

– The Summary Statistics report indicates that the number of rows containing bootstrap means for wabash is N = 961. Although you conducted 1,000 iterations, 39 bootstrap samples did not contain any of the three observations for wabash.

– The histogram of sample means is not smooth, with peaks at the two extremes. The three values for wabash are 38.2, 37.8, and 31.9. The peak at the low end of the distribution results from bootstrap samples that contain only the value 31.9. The peak at the high end results from bootstrap samples that contain one or both of the values 38.2 and 37.8.

Next, use the Fractional Weights (Bayesian Bootstrap) option to avoid obtaining missing values from the bootstrap samples and to smooth the distribution of bootstrapped means.

Bayesian Bootstrap Analysis

1. In the Oneway Analysis report, right-click the **Mean** column in the **Means for Oneway Anova** report and select **Bootstrap**.
2. Type 1000 for the **Number of Bootstrap Samples**.
3. (Optional) To match the results in Figure 11.9, type 12345 for the **Random Seed**.
4. Select the **Fractional Weights** option.
5. Click **OK**.

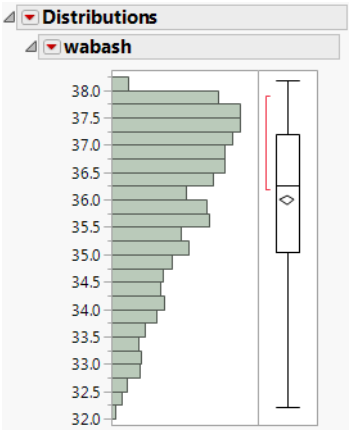
Figure 11.9 Bootstrap Results for a Bayesian Bootstrap

	X	Y	BootID	clarion	clinton	compost	knox	o'neill	wabash	webster
1	Soil	Y	0	32.1667	30.3000	29.6667	34.9000	33.8000	35.9667	31.1000
2	Soil	Y	1	31.9365	31.3493	29.3234	35.4497	34.7105	33.7270	31.4071
3	Soil	Y	2	32.4189	30.3474	29.5143	35.6281	32.7006	34.1027	32.1212
4	Soil	Y	3	32.2339	30.2001	31.4102	34.8758	33.6674	38.0389	31.5428
5	Soil	Y	4	32.4054	30.3242	30.6227	33.8495	32.9495	36.8344	31.9607
6	Soil	Y	5	32.2262	30.8672	29.7999	33.6759	32.2022	35.5792	31.9058
7	Soil	Y	6	32.3222	31.9732	28.8823	35.6307	34.8863	35.3014	30.1325
8	Soil	Y	7	31.9948	30.8828	29.2516	35.3424	33.2094	36.3367	30.8386
9	Soil	Y	8	31.6254	29.7677	28.6390	34.4697	33.0662	36.6183	31.4667
10	Soil	Y	9	32.3499	29.9416	29.5732	35.2564	31.9583	35.8246	29.9302
11	Soil	Y	10	32.5228	30.4506	28.4859	34.9088	35.8126	34.6317	31.6100
12	Soil	Y	11	31.9057	30.0711	29.0693	35.4018	34.4654	33.2086	31.0309
13	Soil	Y	12	31.7275	29.6189	29.1609	34.3984	33.6840	35.0815	31.5446
14	Soil	Y	13	32.5893	30.5210	28.6054	33.4594	34.0958	33.7692	31.7129
15	Soil	Y	14	32.2473	30.5199	31.6080	35.5617	34.1706	35.7146	31.8023
16	Soil	Y	15	32.2329	29.7275	30.1770	35.3286	32.7820	37.4866	30.9706
17	Soil	Y	16	31.5831	29.7508	28.7122	33.5304	34.5348	37.1092	31.2752
18	Soil	Y	17	32.3545	31.3237	29.1542	35.5890	32.2606	37.2005	30.8430
19	Soil	Y	18	32.3811	29.6241	30.6138	35.4308	33.2024	33.0787	31.2926
20	Soil	Y	19	31.7488	29.7763	28.7327	34.7007	33.8910	34.0573	29.9064

There are no missing values in the Bayesian Bootstrap results table. All 21 rows in the Snapdragon.jmp data table are included, with varying bootstrap weights, in each bootstrap sample.

6. Select **Analyze > Distribution**.
7. Select wabash and click **Y, Columns**.
8. Click **OK**.

Figure 11.10 Distribution of wabash Means from a Bayesian Bootstrap



Distributions

wabash

Quantiles

100.0%	maximum	38.184435516
99.5%		38.074322297
97.5%		37.8924836
90.0%		37.669948814
75.0%	quartile	37.185313748
50.0%	median	36.258063753
25.0%	quartile	35.034856614
10.0%		33.843030676
2.5%		32.846088771
0.5%		32.306055984
0.0%	minimum	32.204135264

Summary Statistics

Mean	35.992341
Std Dev	1.4269388
Std Err Mean	0.0451238
Upper 95% Mean	36.08089
Lower 95% Mean	35.903793
N	1000

Bootstrap Confidence Limits

Coverage	BC Lower	BC Upper
0.95	32.6396	37.8168
0.90	32.8837	37.7005
0.80	33.3884	37.4621
0.50	34.4179	36.8464

Original Estimate 35.96667

The Bayesian Bootstrap produces a much smoother distribution for the wabash sample means. All 1,000 bootstrap samples include the three observations for wabash. For each iteration, the wabash sample mean is calculated using different fractional weights.

The Bootstrap Confidence Limits report shows that a 95% confidence interval for the mean is 32.6396 to 37.8168.

JMP PRO Statistical Details for Bootstrapping

This section provides details about the Bootstrapping methods.

JMP PRO Calculation of Fractional Weights

The Fractional Weights option is based on the Bayesian bootstrap (Rubin, 1981). The number of times that an observation occurs in a given bootstrap sample is called its *bootstrap weight*. In the simple bootstrap, the bootstrap weights for each bootstrap sample are determined using simple random sampling with replacement.

In the Bayesian approach, sampling probabilities are treated as unknown parameters and their posterior distribution is obtained using a non-informative prior. Estimates of the probabilities are obtained by sampling from this posterior distribution. These estimates are used to construct the bootstrap weights, as follows:

- Randomly generate a vector of n values from a gamma distribution with shape parameter equal to $(n - 1)/n$ and scale parameter equal to 1.

  Note: Rubin (1981) uses 1 as the gamma shape parameter. The shape parameter that is used in JMP Pro ensures that the mean and variance of the fractional weights are equal to the mean and variance of the simple bootstrap weights.

- Compute S = sum of the n values.
- Compute the fractional weights by multiplying the vector of n values by n/S.
- If a Freq or Weight variable is specified for the analysis, multiply the fractional weights by the Freq or Weight values on a row-by-row basis. The sum of the values of the Freq or Weight variable must be greater than 1.

This procedure scales the fractional weights to have mean and variance equal to those of the simple bootstrap weights. The fractional bootstrap weights are positive, sum to n, and have a mean of 1.

JMP PRO Bias-Corrected Percentile Intervals

This section describes the calculation of the bias-corrected (BC) confidence intervals that appear in the Bootstrap Confidence Limits report when you run the Distribution script in the Bootstrap Results table. Bias-corrected percentile intervals improve on the ability of percentile intervals in accounting for asymmetry in the bootstrap distribution. See Efron (1981).

Notation

- p^* is the proportion of bootstrap samples with an estimate of the statistic of interest that is less than or equal to the original estimate.
- z_0 is the p^* quantile of a standard normal distribution.
- z_α is the α quantile of a standard normal distribution.

Bias-Corrected Confidence Interval Endpoints

The endpoints of a (1 - α) bias-corrected confidence intervals are given by quantiles of the bootstrap distribution:

- The lower endpoint is the $\Phi(2z_0 + z_\alpha)$ quantile.
- The upper endpoint is the $\Phi(2z_0 + z_{1-\alpha})$ quantile.

Chapter 12

Text Explorer

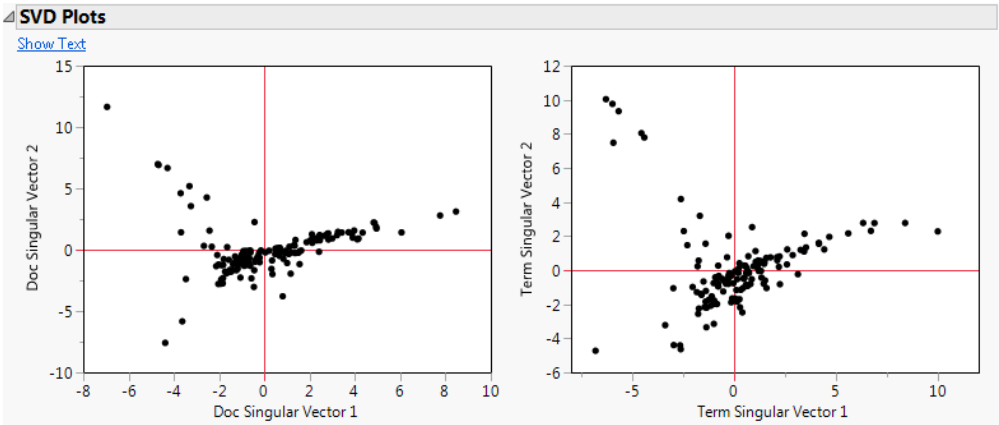
Explore Unstructured Text in Your Data

Many features in this platform are available only in JMP Pro and noted with this icon.

The Text Explorer platform enables you to analyze unstructured text, such as comment fields in surveys or incident reports. Interact with the text data by using tools to combine similar terms, recode misspecified terms, and understand the underlying patterns in your textual data.

The JMP Pro version of the platform also contains analysis tools that use singular value decomposition (SVD) to group similar documents into topics. You can cluster text documents or cluster terms that are in a collection of documents. You can also cluster documents using latent class analysis.

Figure 12.1 SVD Plots in Text Explorer



Text Explorer Platform Overview

Unstructured text data are common. For example, unstructured text data could result from a free response field in a survey, product review comments, or incident reports. The Text Explorer platform enables you to explore unstructured text in order to better understand its meaning. Text analysis is often an iterative process, so you might alternate between curating and analyzing the list of terms.

Curating the List of Terms

Text analysis uses some unique terminology. A *term* or *token* is the smallest piece of text, similar to a word in a sentence. However, you can define terms in many ways, including through the use of regular expressions; the process of breaking the text into terms is called *tokenization*.

- A *phrase* is a short collection of terms; the platform has options to manage phrases that are specified as terms in and of themselves.
- A *document* refers to a collection of words; in a JMP data table, the unstructured text in each row of the text column corresponds to a document.
- A *corpus* refers to a collection of documents.

It is often desirable to exclude some common words from the analysis. These excluded words are called *stop words*. The platform has a default list of stop words, but you can also add specific words as stop words. Although stop words are not eligible to be terms, they can be used in phrases.

You can also recode terms; this is useful for combining synonyms into one common term.

Stemming is the process of combining words with identical beginnings (*stems*) by removing the endings that differ. This results in “jump”, “jumped”, and “jumping” all being treated as the term “jump·”. The stemming procedure is similar to the procedure used in the Snowball string processing language. When a phrase is stemmed, each word in the phrase is stemmed as it would be stemmed as a stand-alone term.

Analyzing the List of Terms

Text analysis in the Text Explorer platform uses a *bag of words* approach. Other than in the formation of phrases, the order of terms is ignored. The analysis is based on the term counts.

After you curate the list of terms through the use of regular expressions, stop words, recoding, and stemming, you can perform analyses on the curated list of terms. The analysis options in the platform are based on the *document term matrix* (DTM). Each row in the DTM corresponds to a document (a cell in a text column of a JMP data table). Each column in the DTM corresponds to a term from the curated term list. This approach implements the bag of words approach since it ignores word ordering. In its simplest form, each cell of the DTM contains

the frequency (number of occurrences) of the column's term in the row's document. There are various other weighting schemes; these are described in "Save Options" on page 349.

JMP PRO The analysis options that are available in the platform first perform a singular value decomposition (SVD) on the document term matrix. This can greatly reduce the number of columns needed to represent the term information in the data. For more details about singular value decomposition, see the Statistical Details appendix in the *Multivariate Methods* book. Hierarchical clustering options are available for clustering the terms and for clustering the documents. These options enable you to group similar terms or documents together.

Platform Workflow

The expected steps for using the Text Explorer platform are as follows:

1. Specify the method for tokenizing (either built-in or customized regular expression).
2. Use the report to specify additional stop words, add phrases to the term list, perform recodes of terms, and specify exceptions to stemming rules.
3. Specify the preference for stemming.
4. Use word and phrase counts, SVD, and clustering approaches to identify important terms and phrases.

   Note: The SVD and clustering options are available only in JMP Pro.

5. Save results for use in further analysis: the term table, the DTM, the singular values, or other results.

   Note: The option to save the singular values is available only in JMP Pro.

6. Save Phrase, Recode, and Stop Words properties for use in future analyses of similar text data.

Text Processing Steps

The text is processed in three stages: tokenizing, phrasing, and terming.

Tokenizing Stage

The Tokenizing stage performs the following operations:

1. Convert text to lowercase.
2. Apply Tokenizing method (either Basic Words or Regex) to group characters into tokens.
3. Recode tokens based on specified recode definitions. Note that recoding occurs before stemming.

Phrasing Stage

The Phrasing stage collects phrases that occur in the corpus (collection of documents) and enables you to specify that individual phrases be treated as terms. Phrases cannot start or end with a stop word, but they can contain a stop word.

Terming Stage

The Terming stage creates the Term List from the tokens and phrases that result from the previous stages.

For each token, the Terming stage performs the following operations:

1. Check that the minimum and maximum length requirements specified in the launch window are met. Tokens that contain only numbers are excluded from this operation.
2. Check that the token is qualified to become a term; tokens parsed by the Basic Words tokenization method must contain at least one alphabetical or Unicode character. Tokens that contain only numbers are excluded from this operation. The Regex tokenization method uses regular expressions to determine what characters are part of a token.
3. Check that the token is not a stop word.
4. Apply stemming and stem exceptions.

For each phrase that you add, the Terming stage performs the following operations:

1. Add the phrase to the Term List. Phrases should apply stemming to each word in the phrase that is stemmed in the Term List. Phrases that have different raw tokens but the same stems are combined in the Term List.
2. Remove token term occurrences that appear in the phrase.

Example of the Text Explorer Platform

In this example, you want to explore the text responses from a survey about pets.

1. Select **Help > Sample Data Library** and open Pet Survey.jmp.
2. Select **Analyze > Text Explorer**.
3. Select Survey Response and click **Text Columns**.
4. Click **OK**.

Figure 12.2 Example of Initial Text Explorer Report

Text Explorer for Survey Response

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-empty Cases	Portion Non-empty per Case
372	194	1921	9.90206	150	0.7732

Term and Phrase Lists

Term	Count
cat	55
dogs	48
dog	46
cats	17
lap	14
barking	12
video	11
food	10
sit	10
walk	10
mice	9
just	8
like	8
take	8
bark	7
huskies	7
time	7
always	6
cattle	6
day	6
every	6
funny	6
house	6
hunting	6
one	6
sheep	6

Phrase	Count	N
video of the cat	5	4
sit in my lap	4	4
cat food	4	2
dog food	4	2
dog barks	3	2
duck hunting	3	2
funny video	3	2
great job	3	2
last week	3	2
stop barking	3	2
cat in my lap	2	4
dogs do a great	2	4
hunting with the dogs	2	4
sled in the winter	2	4
video of a cat	2	4
around the block	2	3
job of herding	2	3
lap and purr	2	3
take the huskies	2	3
us from intruders	2	3
wants to sit	2	3
went duck hunting	2	3
wish my cat	2	3
always trying	2	2
cat chasing	2	2
cat jumped	2	2

At a glance, you can see that there are 372 unique terms in 194 documents. In all, there are 1921 tokenized terms. The most common term is “cat”, and it occurs 55 times.

5. From the red triangle menu next to Text Explorer for Survey Response, select **Term Options** > **Stemming** > **Stem All Terms**.
6. In the Phrase List table, select **cat food** and **dog food**, right-click on the selection, and select **Add Phrase**.

   The terms cat food and dog food are included in the Term List.
7. Scroll down in the Term List and find the cat and dog food entries.

   You can see that there are four occurrences of each phrase.

Figure 12.3 Term List after Modifications and Scrolling

Term	Count
now·	4
cat· food·	4
dog· food·	4
anymor·	3
bath·	3
shake·	3
kid·	3
see·	3
box·	3
stay·	3
roll·	3
fluffi·	3
read·	3
hilari·	3
best·	3
leg·	3

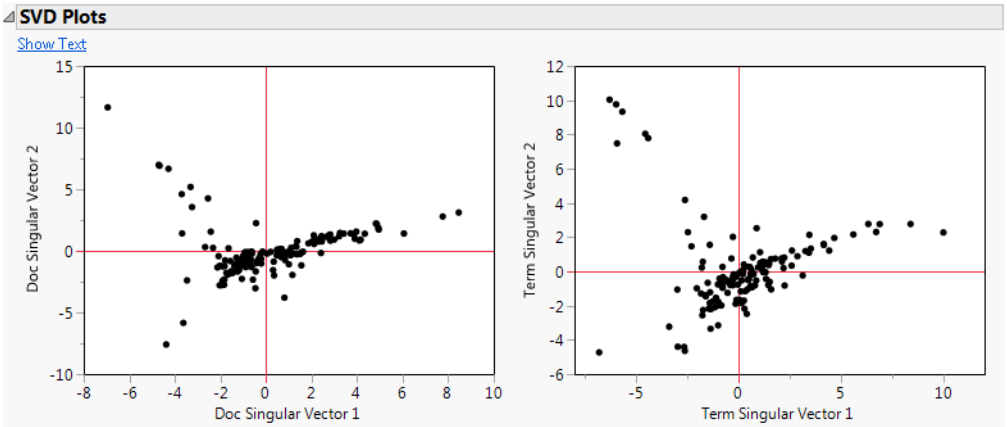
In the Phrase List, cat food and dog food are gray, since they are now locally being treated as terms in this Text Explorer report.

8. JMP PRO From the red triangle menu next to Text Explorer for Survey Response, select **Latent Semantic Analysis, SVD**.

9. JMP PRO Click **OK** to accept the default values.

   Two SVD Plots appear in the report, as shown in Figure 12.4. The one on the left shows the first two singular vectors in the document space. The one on the right shows the first two singular vectors in the term space.

Figure 12.4 SVD Plots



10. JMP PRO Select the three right-most points in the left SVD Plot.

    These three points represent survey responses that are clustered away from the rest of the points. To further investigate this cluster, you read the text of these responses.

11. JMP PRO Click the **Show Text** button that is below the SVD Plots outline title.

Figure 12.5 Text of Selected Documents

```
Context for Selected Rows:

There was this funny video of a cat trying to jump into someones lap, but fell into the pool instead.
[56]

The funny cat video where the cats jumped through the window right into the bathtub was hilarious.
[142]

We made this funny video of the cat trying to climb the wall to chase a laser pointer. [153]
```

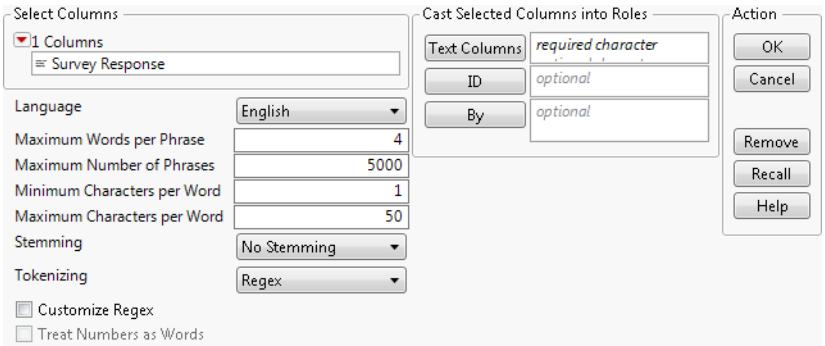
A window appears that contains the text of the three documents represented by the selected points. These survey responses are similar in that they all refer to some combination of “funny”, “cat”, and “video”. These documents have larger positive values for the first singular vector than the rest of the documents, which indicates that they are different from the rest of the documents in that dimension.

Further investigation of the singular vector dimensions could lead to interpretations of what the dimensions represent. For example, many of the documents on the far right of the plot are responses that are about cats. On the far left, many of the responses are about dogs. Therefore, the first singular vector is picking up differences based on whether the response was about a cat or a dog.

Launch the Text Explorer Platform

Launch the Text Explorer platform by selecting **Analyze > Text Explorer**.

Figure 12.6 The Text Explorer Launch Window



The Text Explorer launch window contains the following options:

Text Columns Assigns the columns that contain text data. If you specify multiple columns, a separate analysis is created for each column.

ID Assigns a column used to identify separate respondents in the Save Stacked DTM for Association output data table. This output data table is suitable for association analysis.

By Identifies a column that creates a report consisting of separate analyses for each level of the variable. If more than one By variable is assigned, a separate report is produced for each possible combination of the levels of the By variables.

Note: If you specify a By variable, the Customize Regex option and settings apply to all levels of the By variables.

Language Specifies the language used for text processing. This affects stemming and the built-in lists of stop words, recodes, and phrases. This option is independent of the language in which JMP is running. Unless the Language platform preference is set, the Language option is set according to the JMP Display Language preference. However, the Language option in Text Explorer does not support Chinese, Japanese, or Korean. If the JMP Display Language is Chinese, Japanese, or Korean, this option defaults to English.

Maximum Words per Phrase Specifies a maximum number of words that a phrase can contain to be included as a phrase in the analysis.

Maximum Number of Phrases Specifies the maximum number of phrases that appear in the Phrase List.

Minimum Characters per Word Specifies the number of characters that a word must contain to be included as a term in the analysis.

Maximum Characters per Word Specifies the largest number of characters that a word can contain to be included as a term in the analysis.

Stemming Specifies a method for combining terms with similar beginning characters but different endings. The following options are available:

No Stemming No terms are combined.

Stem for Combining Stems only the terms where two or more terms stem to the same term.

Stem All Terms Stems all terms.

Note: The use of the Stemming option also affects phrases that have been added to the Term List. Phrase identification occurs after terms within a phrase have been stemmed. For example, “dogs bark” and “dog barks” would both match the specified phrase “dog·bark·”.

Tokenizing Specifies a method for parsing the text into terms or tokens. The following tokenization options are available:

Regex Parses text using a default set of built-in regular expressions. If you want to add to, remove, or edit the set of regular expressions used to parse the text, select the **Customize Regex** option. See “Customize Regex: Regular Expression Editor” on page 333.

Basic Words Text is parsed into words based on a set of characters that typically separate words. These characters include spaces, tabs, new lines, and most punctuation marks. If you want numbers to be parsed into terms for the analysis, select the **Treat Numbers as Words** option. If you do not select this option, pieces of text between delimiters that contain only numbers are ignored in the tokenizing step.

Tip: You can view the default set of delimiters using the **Display Options > Show Delimiters** option in a Text Explorer report that uses the Basic Words Tokenizing method.

Customize Regex (Available only with the Regex Tokenizing method.) Enables you to use the Text Explorer Regular Expression Editor window to modify the Regex settings. Use this option to accommodate non-traditional words. Examples include phone numbers or words formed by a combination of characters and numbers. Using the Customize Regex option is not recommended unless the default Regex method is not giving you the results that you need. See "Customize Regex: Regular Expression Editor" on page 333.

Treat Numbers as Words (Available only with the Basic Words Tokenizing method.) Allows numbers to be tokenized as terms in the analysis. When this option is selected, the Minimum Characters per Word setting is ignored for terms that contain numeric digits.

After you click **OK** on the launch window, the Text Explorer Regular Expression Editor window appears if you selected **Customize Regex** in the launch window. Otherwise, the Text Explorer report appears.

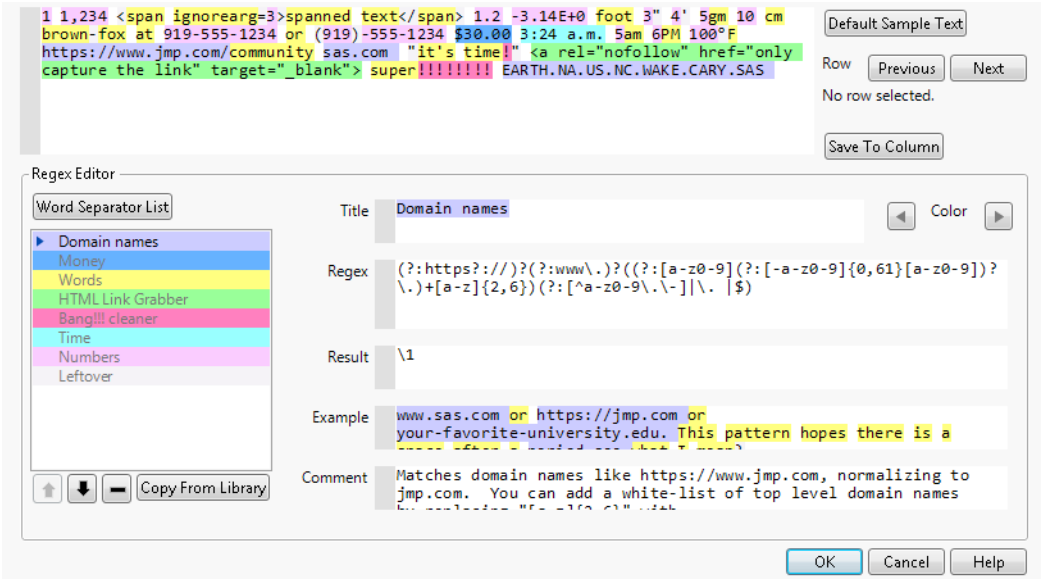
Note: The processing of text input is not case-sensitive. All text is converted to lowercase internally prior to tokenization. This conversion affects the processing of regular expressions and the aggregation of terms in the Text Explorer output.

Customize Regex: Regular Expression Editor

When you select the **Customize Regex** option, the Text Explorer Regular Expression Editor appears. Use this window to parse text documents using a wide variety of built-in regular expressions, such as phone numbers, times, or monetary values. You can also create your own regular expression definitions.

Note: Using the Customize Regex option is recommended only if you are not getting desired results from the default Regex method.

Figure 12.7 Text Explorer Regular Expression Editor



Parsing with the Script Editor Box

The script editor box at the top of the window shows you how the parsing would proceed for sample text. The results of parsing the regular expressions in the Regex Editor list are highlighted in colors that correspond to the colors in the Regex Editor list.

- Click the **Previous** and **Next** row buttons to populate the script editor box with text from your own data. This enables you to see how a given row of text data is parsed.
- Click the **Save to Column** button to save a new column to the data table that contains the result of the regular expression tokenization. For more information about specifying the result of the regular expression, see "Editing the Regular Expressions" on page 335.

Note: The **Save to Column** button uses only the regular expression to match text. The following settings are not used: stop words, recodes, stemming, phrases, or minimum and maximum characters per word to modify the output of the regular expression.

Adding Regular Expressions

To add a regular expression to be used in tokenization, click **Copy From Library**. The Regex Library Selections window appears. This window contains all the built-in regular expressions as well as any recently modified regular expressions that you created in previous instances of the Regular Expression Editor. Built-in regular expressions are labeled. Custom regular expressions that are saved in your library are labeled with the name that you specified. Only the most recent expression for a given name is stored in the Regex Library. Use the **Delete Selected Item** button to remove a custom regular expression from the Regex Library. The

Regex Library for each user is stored as a JSL file in a directory called TextExplorer. The location of this directory is based on your computer's operating system, as follows:

- Windows: "C:/Users/<username>/AppData/Roaming/SAS/JMP/TextExplorer/"
- Macintosh: "/Users/<username>/Library/Application Support/JMP/TextExplorer/"

These files can be shared with other users, but you should not edit the file directly. Use the Regular Expression Editor instead.

Editing the Regular Expressions

Terms are tokenized by processing the regular expressions in the order specified in the Regex Editor panel. To change the order of the regular expressions, select a regular expression in the list and click the up or down arrow buttons below the list. You can also drag and drop items in the regular expression list to change the order of execution. A blue triangle represents the currently selected regular expression. To remove a regular expression and exclude it from the tokenization, select it in the list and click the minus sign below the list. The "Leftover" regular expression cannot be removed and must appear last in the sequence of regular expressions.

When you select a regular expression in the list, the editable fields in the Regex Editor panel refer to the selected regular expression. Click and type in any of these fields to edit them.

Each regular expression has the following attributes:

Title Specifies a name used to identify the regular expression in the current window (as well as in the Regex Library later).

Regex Specifies the regular expression definition. The regular expression must have at least one set of parentheses to designate the regular expression capture.

Result Specifies what replaces the text matched by the regular expression. This value can be static text, blank, or the value of the regular expression capture. The regular expression capture is defined as the result of the Regex definition:

- To replace the matched text with static text, specify the static text in the Result field.
- To ignore the matched text, leave the Result field blank.
- To keep the text that results from the outer-most parentheses in the regular expression, use "\1" (without quotation marks) in the Result field.
- To keep the entire result of the regular expression, use "\0" (without quotation marks) in the Result field.

Example (Optional) Specifies an example text string with colors indicating the behavior of the regular expression.

Comment (Optional) Specifies a comment to explain the regular expression and its behavior.

Color Specifies the color used to identify matches of the regular expression in the text in the Script Editor box and in the Example field.

Note: If the regular expression definition in the Regex field is invalid, a red X appears next to the name of the regular expression in the list of regular expressions.

Creating a Custom Regular Expression

Follow these steps to create your own custom regular expression:

1. Click **Copy From Library**.
2. In the Regex Library Selections window, select a regular expression that seems close to what you need.
3. Click **OK**.
4. Edit the Regex definition in the Regex Editor panel.
5. Give your custom regular expression a different name in the Title field.

Tip: When editing the Regex definition field, it is helpful to have the Log window open and visible. Some error messages appear only in the Log window. To open the Log window, select **View > Log**. There are many Internet resources available for troubleshooting regular expressions, such as www.regexr.com.

The Word Separator List

The **Word Separator List** button enables you to specify a list of characters that occur between words in the tokenization process. The *between-word characters* cannot begin a word, but they can appear inside a word if one of the regular expressions allows it. You can add or remove characters from the list in the window that appears when you click the button. By default, the only character in the list is a whitespace character. In the Separator Characters window, click the **Reset** button to undo any modifications to the list of separator characters. Modifications to the list of separator characters are applied only to the current regular expression tokenization.

The processing of the specified regular expressions and the required “Leftover” regular expression proceeds as follows:

1. Compare the current character in the text stream to the list of separator characters.
   - If the character is in the list of separator characters, ignore the character, process any accumulated characters in the “Leftover” temporary string, move to the next character, and repeat step 1.
   - If the character is not in the list of separator characters, go to step 2.
2. Compare the string starting at the current character to each regular expression (one at a time, up to, but not including, the “Leftover” regular expression).
   - If the string starting at the current character matches one of the regular expressions, the following events occur. Any accumulated characters in the “Leftover” temporary string are processed. The value of the Result field is saved as a term. The current

character in the text stream becomes the character following the matched string. The processing returns to step 1.

  - If the string starting at the current character does not match any of the regular expressions up to the “Leftover” regular expression, go to step 3.

3. Collect characters into the “Leftover” temporary string by appending the current character and setting the current character to the next character in the text stream. Return to step 1.
   - The “Leftover” temporary string is accumulated one character at a time, until one of the other regular expressions produces a match.
   - The default Result of the “Leftover” regular expression is to discard the accumulated “Leftover” temporary string.

Tips:

- If you set the Result of the “Leftover” regular expression to \1, you might want to add more separator characters, such as punctuation marks. This ensures that your results do not include the specified punctuation marks.
- You might want to add more regular expressions from the Regex Library or create custom ones to capture terms of interest rather than changing the Result of the “Leftover” regular expression to \1.

The processing follows the above steps until reaching the end of the text string for each row in the data table.

Saving the Results to a Column in the Data Table

Click the **Save to Column** button to save to the data table a new column that contains the results of the regular expression tokenization. The new column is a character column with the same name as the text column specified in the Text Explorer launch window; a number is appended to the name so that the column names are unique.

Note: When you save the results of the custom regular expression tokenization to a column in the data table, the regular expression process is run on the original text in each row of the data table. It is not run on the version of the text string that was converted to lowercase.

Closing the Text Explorer Regular Expression Editor

After you click **OK** in the Text Explorer Regular Expression Editor window, the following events occur:

1. The custom regular expressions defined in the Text Explorer Regular Expression Editor window are saved to the Regex Library.

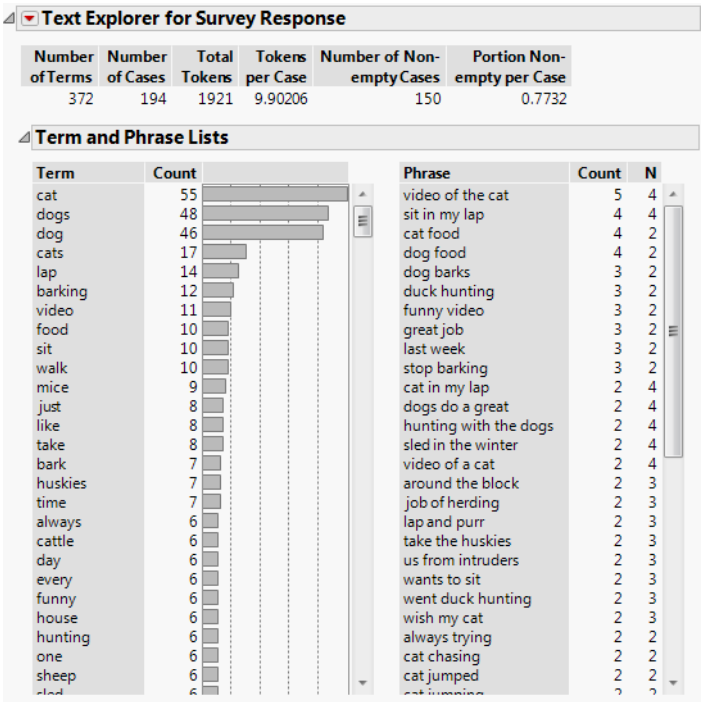
Caution: The custom Regex Library is saved only when you click **OK** and there are customized regular expressions. The most recently saved regular expressions will be available next time. Use unique names to keep additional regular expressions in the Regex Library. To ensure that a regular expression is available later, you can save a script from the Text Explorer report window.

2. The Text Explorer report appears. The report shows the result of using the specified regular expression settings to tokenize the text.

The Text Explorer Report

The Text Explorer report window contains the Summary Counts report and the Term and Phrase Lists report.

Figure 12.8 Example of a Text Explorer Report



Text Explorer for Survey Response

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-empty Cases	Portion Non-empty per Case
372	194	1921	9.90206	150	0.7732

Term and Phrase Lists

Term	Count
cat	55
dogs	48
dog	46
cats	17
lap	14
barking	12
video	11
food	10
sit	10
walk	10
mice	9
just	8
like	8
take	8
bark	7
huskies	7
time	7
always	6
cattle	6
day	6
every	6
funny	6
house	6
hunting	6
one	6
sheep	6

Phrase	Count	N
video of the cat	5	4
sit in my lap	4	4
cat food	4	2
dog food	4	2
dog barks	3	2
duck hunting	3	2
funny video	3	2
great job	3	2
last week	3	2
stop barking	3	2
cat in my lap	2	4
dogs do a great	2	4
hunting with the dogs	2	4
sled in the winter	2	4
video of a cat	2	4
around the block	2	3
job of herding	2	3
lap and purr	2	3
take the huskies	2	3
us from intruders	2	3
wants to sit	2	3
went duck hunting	2	3
wish my cat	2	3
always trying	2	2
cat chasing	2	2
cat jumped	2	2

Summary Counts Report

The first table in the Text Explorer report window contains the following summary statistics:

Number of Terms the number of terms in the Term List.

Number of Cases the number of documents in the corpus.

Total Tokens the total number of terms in the corpus.

Tokens per Case the number of tokens divided by the number of cases.

Number of Non-empty Cases the number of documents in the corpus that contain at least one term.

Portion Non-empty per Case the proportion of documents in the corpus that contain at least one term.

Term and Phrase Lists

The Term and Phrase Lists report contains tables of terms and phrases found in the text after tokenization has occurred. See Figure 12.8 for an example of the Term and Phrase Lists report. The Count column in the Term List indicates the number of occurrences of the term in the corpus. The Count column in the Phrase List indicates the number of occurrences of the phrase in the corpus; the N column indicates the number of words in the phrase.

By default, the Terms List is sorted in descending count order; terms that are tied in count are sorted alphabetically. The Phrases List is sorted in descending count order; phrases that are tied in count are then sorted in descending length (N) order. Further ties in the Phrases List are sorted alphabetically. The sort order of each list can be changed to alphabetical sorting using the options in each list.

The phrases that appear in the Phrase List are determined by the settings of the **Maximum Words per Phrase** and **Maximum Number of Phrases** options in the launch window. Phrases that occur only one time in the data table do not appear in the Phrase List.

Phrases can be specified as terms at various scopes. Phrases in the Phrase List that have been specified as terms are colored based on the scope of the phrase specification. See Table 12.1. For more information about specifying phrases in different scopes, see "Term Options Management Windows" on page 344.

Table 12.1 Colors for Specified Phrases

Scope	Color
Built-in	Red
User Library	Green
Column Property	Orange
Local	Gray

Actions for Terms and Phrases

You can access options in the Term List and Phrase List tables by selecting items and then right-clicking in the left-most column of each table. You can save each table as a data table by right-clicking in the Count column of each table and selecting Make into Data Table.

Term List Context Menu Options

When you right-click in the Term column of the Term List table, a context menu appears with the following options:

Select Rows Selects rows in the data table that contain the selected terms.

Show Text Shows the documents that contain the selected terms.

Note: By default, only the first 10,000 documents are shown. If there are more than 10,000 documents that contain a selected term, a window appears that enables you to increase this limit.

Alphabetical Order Toggles the sort order of the Term List between alphabetical order and descending Count order.

Find Enables you to search for a string in the Term List. When results of a Find operation are visible, you can return to the full list of terms by selecting Find again.

Copy Places the selected terms onto the clipboard.

Color Enables you to assign a color to the selected terms.

Label Places labels on the corresponding points in the Term SVD Plot for the selected terms.

Containing Phrases Selects the phrases in the Phrase List table that contain the selected terms.

Save Indicators Saves an indicator column to the data table for each term selected in the Term List. The value of the indicator column for each row is 1 if the document in that row contains the term and 0 otherwise.

Save Formula Saves a column formula to the data table for each term selected in the Term List. The column formula for each row evaluates to 1 if the document in that row contains the term and 0 otherwise. This is useful for new documents.

Recode Enables you to change the values for one or more terms. Select the terms in the list before selecting this option. After you select this option, the Recode window appears. See the Enter and Edit Data chapter in the *Using JMP* book.

Add Stop Word Adds the selected terms to the list of stop words and removes those terms from the Term List. This action also updates the Phrase List.

Add Stem Exception Adds the selected terms to the list of terms that are excluded from stemming.

Make into Data Table Creates a JMP data table from the report table.

Make Combined Data Table Searches the report for other tables like the one you selected and combines them into a single JMP data table.

Phrase List Context Menu Options

When you right-click in the Phrase column of the Phrase List table, a context menu appears with the following options:

Select Rows Selects rows in the data table that contain the selected phrases.

Show Text Shows the documents that contain the selected phrases.

Save Indicators Saves an indicator column to the data table for each phrase selected in the Phrase List. The value of the indicator column for each row is 1 if the document in that row contains the phrase and 0 otherwise.

Alphabetical Order Toggles the sort order of the Phrase List between alphabetical order and descending Count order.

Find Enables you to search for a string in the Phrase List. When results of a Find operation are visible, you can return to the full list of phrases by selecting Find again.

Copy Places the selected phrases onto the clipboard.

Select Contains Selects larger phrases in the Phrase List that contain the selected phrase.

Select Contained Selects smaller phrases in the Phrase List that are contained by the selected phrase.

Add Phrase Adds the selected phrases to the Term List and updates the Term Counts accordingly.

Add Stop Word Adds the selected phrases to the list of stop words and removes those phrases from the Phrase List. This action also updates the Term List.

Make into Data Table Creates a JMP data table from the report table.

Make Combined Data Table Searches the report for other tables like the one you selected and combines them into a single JMP data table.

Text Explorer Platform Options

This section describes the options available in the Text Explorer platform.

Text Preparation Options

The Text Explorer red triangle menu contains the following options for text preparation:

Display Options Shows a submenu of options to control the report display.

Show Word Cloud Shows or hides the Word Cloud report. The Word Cloud red triangle menu enables you to change the layout and font for the word cloud. See "Word Cloud Options" on page 343.

The word cloud can be interactively resized by changing the width. The height is then determined automatically. The rows in the Term List are linked to the terms in the Word Cloud.

Show Term List Shows or hides the Term List.

Show Phrase List Shows or hides the Phrase List.

Show Term and Phrase Options Shows buttons in the Term and Phrase Lists report corresponding to the options available in the context menus for each list. See "Term and Phrase Lists" on page 339.

Show Summary Counts Shows or hides the Summary Counts table. See "Summary Counts Report" on page 338.

Show Stop Words Shows or hides a list of the stop words used in the analysis. A built-in list of stop words is used initially. To add a stop word, right-click on it in the Term List and select **Add Stop Word** from the context menu. See "Term Options Management Windows" on page 344.

Show Recodes Shows or hides a list of the recoded terms. See "Term Options Management Windows" on page 344.

Show Specified Phrases Shows or hides a list of the phrases that have been specified by the user to be treated as terms. See "Term Options Management Windows" on page 344.

Show Stem Exceptions Shows or hides the terms that are excluded from stemming. See "Term Options Management Windows" on page 344.

Show Delimiters Shows or hides the delimiters used by the Basic Words Tokenizing method. To modify the set of delimiters used, you must use the `Add Delimiters()` or `Set Delimiters()` messages in JSL. This option is available only when the selected Tokenizing method is Basic Words.

Show Stem Report Shows or hides the Stemming report that contains two tables of stemming results. The table on the left maps each stem to the corresponding terms. The table on the right maps each term to its corresponding stem. This option is available only when the selected Stemming method is not No Stemming.

Term Options Shows a submenu of options that apply to the Term List.

Stemming See the description of stemming options in "Launch the Text Explorer Platform" on page 331.

Include Builtin Stop Words Specifies if the stop words used in the tokenizing process include built-in stop words or not.

Include Builtin Phrases Specifies if the phrases used in the tokenizing process include built-in phrases or not.

Manage Stop Words Shows a window that enables you to add or remove stop words. The changes made can be applied at the User, Column, and Local levels. You can also specify Local Exceptions that exclude stop words that are specified in any of the other levels. See "Term Options Management Windows" on page 344.

Manage Recodes Shows a window that enables you to add or remove recodes. The changes made can be applied at the User, Column, and Local levels. You can also specify Local Exceptions that exclude recodes that are specified in any of the other levels. See "Term Options Management Windows" on page 344.

Manage Phrases Shows a window that enables you to add or remove the phrases that are treated as terms. The changes made can be applied at the User, Column, and Local levels. You can also specify Local Exceptions that exclude phrases that are specified in any of the other levels. See "Term Options Management Windows" on page 344.

Manage Stem Exceptions Shows a window that enables you to add or remove exceptions to stemming. The changes made can be applied at the User, Column, and Local levels. You can also specify Local Exceptions that exclude stem exceptions that are specified in any of the other levels. See "Term Options Management Windows" on page 344.

Parsing Options Shows a submenu of options that apply to parsing and tokenization.

Tokenizing See the description of tokenizing options in "Launch the Text Explorer Platform" on page 331.

Customize Regex Shows the Customize Regex window. This option enables you to modify the Regex settings for the current Text Explorer report. This option is available only when the selected Tokenizing method is Regex.

Note: If you specified a By variable in the platform launch window, the Customize Regex option automatically broadcasts to all level of the By variables.

Treat Numbers as Words Allows numbers to be tokenized as terms in the analysis. Note that this option is affected by the setting for Minimum characters per word. This option is available only when the selected Tokenizing method is Basic Words.

Word Cloud Options

The Word Cloud red triangle menu contains the following options:

Layout (Set to Ordered by default.) Specifies the arrangement of the terms in the Word Cloud.

Ordered Presents the terms in horizontal lines ordered from most to least frequent.

Alphabetical Presents the terms in horizontal lines sorted in ascending alphabetical order.

Centered Presents the terms in a cloud and sized by frequency

Coloring (Set to None by default.) Specifies the coloring of the terms in the Word Cloud.

None Colors each term black. You can change this color in the Legend.

Arbitrary Grays Colors each term in varying shades of gray.

Arbitrary Colors Colors each term in various colors. You can adjust the colors in the Legend.

By column values Colors each term on a gradient color scale. The scale is based on the difference between the overall mean of another column and the value for a term in the Score Terms by Column command. You can adjust the colors and gradient in the Legend.

Font Specifies the font, style, and size of the terms in the Word Cloud.

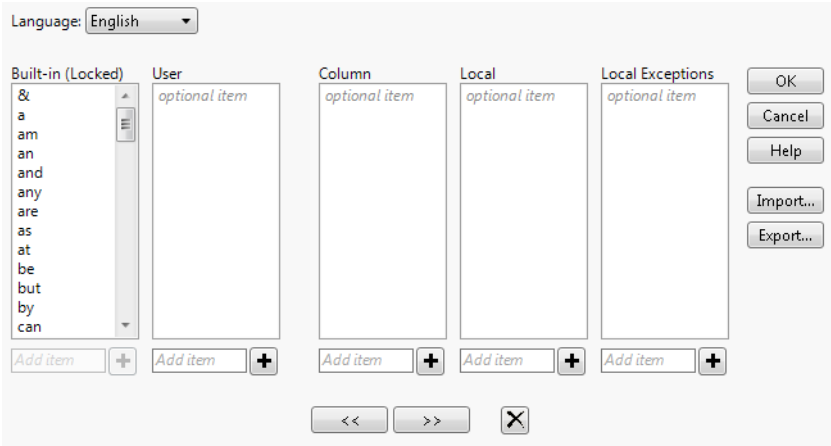
Show Legend (On by default.) Shows or hides the legend for the Word Cloud.

Term Options Management Windows

Phrase, stop word, recode, and stem exception information can be specified for many different scopes. They can be stored in the following locations: the Text Explorer user library (User scope), a column property for the analysis column (Column scope), or in a platform script (Local scope). You can save the local specifications and local exceptions for a specific instance of Text Explorer by saving the script for the Text Explorer report.

The Term Options management windows are four similar windows that enable you to manage the collections of stop words, recodes, phrases, and stem exceptions. Figure 12.9 shows the Manage Stop Words window. The Manage Phrases and Manage Stem Exceptions are identical to the Manage Stop Words window. The Manage Recodes window differs slightly. See "Manage Recodes" on page 346.

Figure 12.9 Manage Stop Words Window



Manage Stop Words

The Manage Stop Words window contains multiple lists of stop words that represent the different scopes (or locations) of specified stop words. Below each list is a text edit box and an add button. These controls enable you to add custom stop words to each scope. You can move stop words from one scope to another by dragging them. You can copy and paste items from one list to another list. Two buttons at the bottom of the window move the selected items from one scope to the next, either left or right. The X button removes the selected items from their current scope. You can edit existing items in the lists by double-clicking on an item and changing the text.

Language Specifies the list of Built-in stop words and to which language the user library selections are saved. If you select Apply Items for Language, the changes are saved to the master user library. The Language setting applies only to the Built-in and User scopes.

Built-in (Locked) Lists the built-in list of stop words for the specified language. You can exclude a built-in stop word by placing it in the Local Exceptions list.

User Lists the stop words in the user library for the specified language.

Column Lists the stop words in the “Stop Words” column property for the text column.

Local Lists the stop words in the local scope. They can be specified when Text Explorer is launched via JSL. These stop words are used only in the current Text Explorer platform report.

Local Exceptions Lists words that are not treated as stop words in the current Text Explorer platform. They can be specified when Text Explorer is launched via JSL. The words listed in Local Exceptions override words listed in all of the other scopes.

Import Enables you to import stop words from a text file. The stop words are copied to the clipboard. You can paste them into any of the lists other than Built-in.

Export Enables you to export stop words to the clipboard or to a text file. An Export window appears that enables you to select the scopes for which you would like to export stop words and the location of the export.

The user library files are located in a TextExplorer directory. The location of this directory is based on your computer's operating system:

- Windows: "C:/Users/<username>/AppData/Roaming/SAS/JMP/TextExplorer/<lang>/"
- Macintosh: "/Users/<username>/Library/Application Support/JMP/TextExplorer/<lang>/"

The master user library files are located in the TextExplorer directory itself. These files are not language-specific.

When you click **OK**, changes to the User and Column lists are saved to the user library and the column properties, respectively. Anything specified in the Local and Local Exceptions lists is saved only when you save the script of the Text Explorer report.

If saving Stop Words to the user library, the file is named stopwords.txt. If saving to a column property, the property is called "Stop Words".

Manage Recodes

The Manage Recodes window differs slightly from the Manage Stop Words window. Instead of one text edit box below each list, there are two text edit boxes. The old value (specified in the top box) is recoded to the new value (specified in the bottom box).

If saving Recodes to the user library, the file is named recodes.txt. If saving to a column property, the property is called "Recodes".

Manage Phrases

If saving Phrases to the user library, the file is named phrases.txt. If saving to a column property, the property is called "Phrases".

Manage Stem Exceptions

If saving Stem Exceptions to the user library, the file is named stemExceptions.txt. If saving to a column property, the property is called "Stem Exceptions".

Note: The Local Exceptions list in the Manage Stem Exceptions window lists stem exceptions that are excluded from the stem exception list. The words in this list are involved in the stemming operation.

JMP PRO Text Analysis Options

The Text Explorer red triangle menu contains the following analysis options:

Latent Class Analysis Performs a latent class analysis on the binary weighted document term matrix using sparse matrix routines.

When you select Latent Class Analysis from the Text Explorer red triangle menu, a Specifications window appears with the following options:

Maximum Number of Terms The maximum number of terms included in the latent class analysis.

Minimum Term Frequency The minimum number of occurrences a term must have to be included in the latent class analysis.

Number of Clusters The number of clusters in the latent class analysis.

See"Latent Class Analysis" on page 352.

Latent Semantic Analysis, SVD Performs a sparse singular value decomposition of the document term matrix. See "Latent Semantic Analysis (SVD)" on page 353.

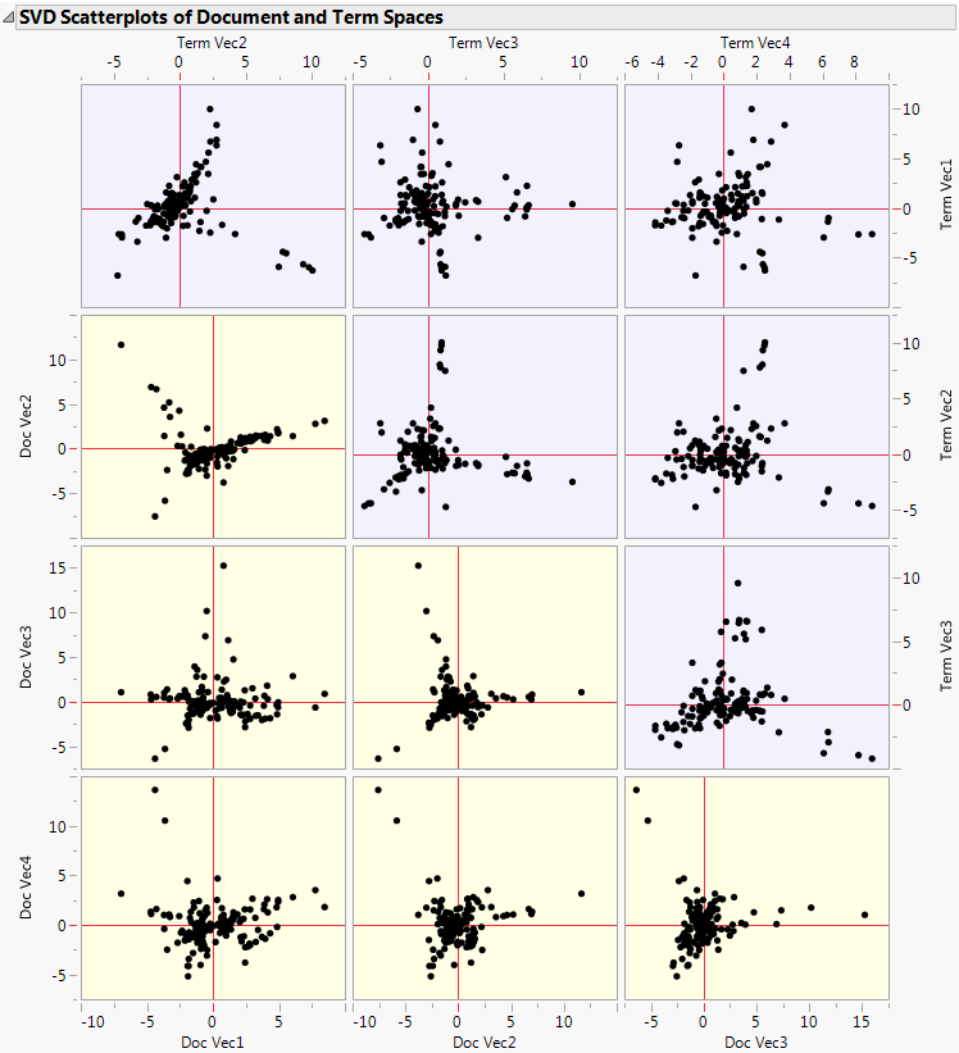
Topic Analysis, Rotated SVD Performs a varimax rotated singular value decomposition of the document term matrix to produce groups of terms called topics. See "Topic Analysis" on page 354.

Cluster Terms Shows or hides a hierarchical clustering analysis of the terms in the data. To the right of the dendrogram, there are options to set the number of clusters and save the clusters to a data table. For each term, this data table contains its frequency, the number of documents that contain it, and its assigned cluster.

Cluster Documents Shows or hides a hierarchical clustering analysis of the documents in the data. To the right of the dendrogram, there are options to set the number of clusters, save the clusters to a column in the data table, and show the documents in a selected branch of the dendrogram plot.

SVD Scatterplot Matrix (Available after selecting Latent Semantic Analysis, SVD.) Shows or hides a scatterplot matrix of the term and document singular value decomposition vectors. You are prompted to select the size of the scatterplot matrix when you select this option. This scatterplot matrix enables you to visualize more than the first two dimensions of the singular value decomposition.

Figure 12.10 SVD Scatterplots of Document and Term Spaces



Topic Scatterplot Matrix (Available after selecting Topic Analysis, Rotated SVD.) Shows or hides a scatterplot matrix of the rotated singular value decomposition vectors.

JMP PRO Singular Value Decomposition Specifications Windows

The analysis options in the Text Explorer platform are based on the Document Term Matrix (DTM). The DTM is formed by creating a column for each term in the Term List (up to a specified Maximum Number of Terms). Each text document (equivalent to a row in the data table) corresponds to a row of the DTM. The values in the cells of the DTM depend on the type of weighting specified by the user in the Specifications window.

Figure 12.11 shows the Singular Value Decomposition Specifications window. When you select options from the Text Explorer red triangle menu that perform a singular value decomposition on the document term matrix, the Specifications window appears with the following options:

Maximum Number of Terms The maximum number of terms included in the singular value decomposition.

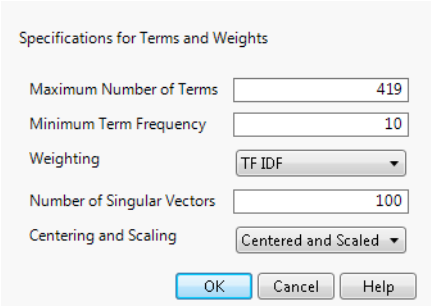
Minimum Term Frequency The minimum number of occurrences a term must have to be included in the singular value decomposition.

Weighting The weighting scheme that determines the values that go into the cells of the document term matrix. The weighting scheme options are described in "Document Term Matrix Specifications Window" on page 350.

Number of Singular Vectors The number of singular vectors in the singular value decomposition. The default value is the minimum of the number of documents, the number of terms, or 100.

Centering and Scaling Options for centering and scaling of the document term matrix. You can choose between **Centered**, **Centered and Scaled**, and **Uncentered**. By default, the document term matrix is both centered and scaled.

Figure 12.11 SVD Specification Window



Save Options

The Text Explorer red triangle menu contains the following options that are available to save information to data tables, table columns, and column properties:

Save Document Term Matrix Saves columns to the data table for each column of the document term matrix (up to a specified Maximum Number of Terms).

JMP PRO **Save Document Singular Vectors** Saves a user-specified number of singular vectors from the document singular value decomposition as columns to the data table.

JMP PRO **Save Document Topic Vectors** Saves a user-specified number of singular vectors from the rotated singular value decomposition as columns to the data table.

JMP PRO **Save Stacked DTM for Association** Saves a stacked version of the document-term matrix to a JMP data table. The stacked format is appropriate for analysis in the Association Analysis platform. For more information, see the Association Analysis chapter in the *Predictive and Specialized Modeling* book. If you specify an ID variable in the Text Explorer launch window, the ID variable is used to identify the rows that each term came from in the original text data table. The stacked table also contains a table script to launch Association Analysis.

Save DTM Formula Saves a vector-valued formula column to the data table. The length of the vector depends on user-specified options for the maximum number of terms, the minimum term frequency, and the weighting. The resulting column uses the `Text Score()` JSL function. For more information about this function, see Help > Scripting Index.

JMP PRO **Save Singular Vector Formula** (Available only after the Latent Semantic Analysis, SVD option has been selected.) Saves a vector-valued formula column containing the document singular value decomposition to the data table. The resulting column uses the `Text Score()` JSL function. For more information about this function, see Help > Scripting Index.

JMP PRO **Save Topic Vector Formula** (Available only after the Topic Analysis, Rotated SVD option has been selected.) Saves a vector-valued formula column containing the rotated singular value decomposition to the data table. The resulting column uses the `Text Score()` JSL function. For more information about this function, see Help > Scripting Index.

Save Term Table Creates a JMP data table that contains each term from the Term List, the number of occurrences, and the number of documents that contain each term. This data table also contains a column containing scores for each term if the Score Terms by Column option has been selected.

JMP PRO **Save Term Singular Vectors** Saves a user-specified number of singular vectors from the terms singular value decomposition as columns to a new data table where each row corresponds to a term. If a Term Table data table is already open, this option saves the columns to that data table.

JMP PRO **Save Term Topic Vectors** (Available only after the Topic Analysis, Rotated SVD option has been selected and the Term Table has been created.) Saves the topic vectors as columns to the data table created by the Save Term Table command.

Score Terms by Column Saves the Term List table with scores based on values in a specified column to a JMP data table. The scores for each term are the mean value of the specified column weighted by the number of occurrences of the term in each row.

Document Term Matrix Specifications Window

When you select the Save Document Term Matrix and Save DTM Formula options from the Text Explorer red triangle menu, the Document Term Matrix Specifications window appears with the following options:

Maximum Number of Terms The maximum number of terms included in the document term matrix.

Minimum Term Frequency The minimum number of occurrences a term must have to be included in the document term matrix.

Weighting The weighting scheme that determines the values that go into the cells of the document term matrix.

The following options are available for Weighting:

Binary Assigns 1 if a term occurs in each document and 0 otherwise.

Ternary Assigns 2 if a term occurs more than once in each document, 1 if it occurs only once and 0 otherwise.

Frequency Assigns the count of a term's occurrence in each document.

Log Freq Assigns *log10(1 + x)*, where *x* is the count of a term's occurrence in each document.

TF IDF Assigns *TF * log(nDoc / nDocTerm)*. Abbreviation for *term frequency - inverse document frequency*. This is the default weighting. The terms in the formula are defined as follows:

– *TF* = frequency of the term in the document

– *nDoc* = number of documents in the corpus

– *nDocTerm* = number of documents that contain the term

Report Options

See the JMP Reports chapter in the *Using JMP* book for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

JMP PRO Latent Class Analysis

Latent class analysis enables you to group the documents from the corpus into clusters of similar documents. The Latent Class Analysis report contains the Bayesian Information Criterion (BIC) value for the model and a Show Text button. If one or more clusters in the Cluster Mixture Probabilities table is selected, the Show Text button opens a window that contains the text of the documents that are deemed most likely to belong to the selected cluster.

The Latent Class Analysis red triangle menu contains the following options:

Display Options Specifies the contents of the Latent Class Analysis report. By default, all of the report options are shown.

Mixture Probabilities by Cluster Shows or hides a table of the probability of an observation belonging to each cluster.

Tip: You can select one or more rows in the Mixture Probabilities by Cluster table to select the observations assigned to the corresponding clusters.

Term Probabilities by Cluster Shows or hides a table of terms with an estimate for each cluster of the conditional probability that a document contains the term, given that the document belongs to a particular cluster. By default, the terms in this table are sorted by descending frequency in the corpus.

The Cluster Most Characteristic column shows the cluster that the term occurs in at the highest rate.

The Cluster Most Probable column shows the cluster in which a randomly chosen document that contains the term is most likely to be found.

Top Terms per Cluster Show or hides a table of the ten most frequent terms in each cluster.

MDS Plot Shows or hides a multidimensional scaling plot, which is a two-dimensional representation of the proximity of the clusters. For more information about MDS plots, see the Multidimensional Scaling chapter in the *Consumer Research* book.

Cluster Probabilities by Row Shows or hides the Mixture Probabilities table, which displays probabilities of cluster membership for each row. The Most Likely Cluster column indicates the cluster with the highest probability of membership for each row.

Save Probabilities Saves the values in the Mixture Probabilities table to the corresponding rows in the data table.

Save Probability Formulas Saves a formula column to the data table for each cluster as well as a formula column for the most likely cluster.

The score formula that is saved uses the `Text Score()` JSL function with the weighting argument set to "`LCA`".

Color by Cluster Colors each row in the data table according to its most likely cluster.

Remove Removes the Latent Class Analysis report from the Text Explorer report.

For more information about latent class analysis, see the Latent Class Analysis chapter in the *Multivariate Methods* book.

Note: The LCA algorithm that is used in the Text Explorer platform takes advantage of the specific structure of the document term matrix. For this reason, the LCA results in the Text Explorer platform do not exactly match the results in the Latent Class Analysis platform.

JMP PRO Latent Semantic Analysis (SVD)

Latent semantic analysis is a family of analysis techniques centered around computing a partial singular value decomposition (SVD) of the document term matrix (DTM). This decomposition reduces the text data into a manageable number of dimensions for analysis. Latent semantic analysis is similar to principal components analysis.

The singular value decomposition approximates the DTM using three matrices: **U**, **S**, and **V'**. The relationship between these matrices is defined as follows:

$$DTM \approx \mathbf{U} * \mathbf{S} * \mathbf{V'}$$

Define *nDoc* as the number of documents (rows) in the DTM, *nTerm* as the number of terms (columns) in the DTM, and *nVec* as the specified number of singular vectors. Note that *nVec* must be less than or equal to min(*nDoc*, *nTerm*). It follows that **U** is an *nDoc* by *nVec* matrix. **S** is a diagonal matrix of dimension *nVec*. The diagonal entries in **S** are the singular values in the SVD. **V'** is an *nVec* by *nDoc* matrix. The rows in **V'** are the singular vectors

The singular vectors capture connections among different words with similar meanings or topic areas. If three words tend to appear in the same documents, the SVD is likely to produce a singular vector in **V'** with large values for those three words. The **U** singular vectors represent the documents projected into this new term space.

Latent semantic analysis also captures indirect connections. If two words never appear together in the same document, but they generally appear in documents with another third word, the SVD is able to capture some of that connection. If two documents have no words in common but contain words that are connected in the dimension-reduced space, they map to similar vectors in the SVD output.

The SVD transforms text data into a fixed-dimensional vector space, making it amenable to all types of clustering, classification, and regression techniques. The Save options enable you to export this vector space to be analyzed in other JMP platforms.

The DTM, by default, is centered and scaled before the singular value decomposition. Centering and scaling can be turned off in the Specifications window. However, the SVD implementation takes full advantage of the sparsity of the DTM even when the DTM is centered.

JMP PRO SVD Plots Report

The Latent Semantic Analysis option produces two SVD Plots. The first plot shows the first two singular vectors for the documents, that is, the first two columns of the **U** matrix. The second plot shows the first two singular vectors for the terms, that is the first two rows of the **V'** matrix. The points in both of these plots can be selected. For the document SVD plot, the points correspond to rows in the data table. For the term SVD plot, the points correspond to rows in the Term List table.

Below the document and term SVD plots, a table of the singular values appears. These are the diagonal entries of the **S** matrix in the singular value decomposition of the document term matrix.

JMP PRO Topic Analysis

The Topic Analysis option performs a varimax rotation on the singular value decomposition (SVD) of the document term matrix (DTM). You must specify a number of rotated singular vectors, which corresponds to the number of topics. After you specify a number of topics, the Topic Words and Topic Scores reports appear. Topic analysis is similar to factor analysis.

The varimax rotation takes a set of singular vectors and rotates them to make them point more directly in the coordinate directions (toward the terms). This rotation makes the vectors help explain the text as each rotated vector orients toward a set of words. Negative values indicate a repulsion force. The terms with negative values occur in a topic less frequently compared to the terms with positive values.

JMP PRO Topic Words Report

The Topic Words report shows the terms that have the highest topic scores in each topic. There are additional reports that show the components of the rotated singular value decomposition

The report shows a table of terms in each topic that have the largest scores in absolute value. Each table is sorted in descending order by the absolute value of the score. These tables can be used to determine conceptual themes that correspond to each topic.

The Topic Words report also contains the following matrix reports:

Transform Contains the rotation matrix for the varimax rotation.

Rotated V Matrix Contains a matrix of term scores for each topic. The rotated **V** matrix results from a varimax rotation of the **V** matrix in the SVD analysis. See "Latent Semantic Analysis (SVD)" on page 353.

Rotated U Matrix Contains a matrix of document scores for each topic. Documents with higher scores in a topic are more likely to be associated with that topic.

Topic Portion Contains the topic portion values for each topic.

JMP PRO Topic Scores Report

The Topic Scores report contains a Show Text button and a topic scores plot. The Show Text button opens a window that contains the text of the selected documents.

The topic scores plot is a visual representation of the rotated **U** matrix. Each panel in the plot corresponds to one of the topics, or one of the rows of the rotated **U** matrix. Within each panel, each point corresponds to one of the document in the corpus, or one of the columns in the rotated **U** matrix.

JMP PRO Additional Example of the Text Explorer Platform

This example looks at aircraft incident reports from the National Transportation Safety Board for events occurring in 2001 in the United States. You want to explore the text that contains a description of the results of the investigation into the cause of each incident. You also want to find themes in the collection of incident reports.

1. Select **Help > Sample Data Library** and open Aircraft Incidents.jmp.
2. Select **Rows > Color or Mark by Column**.
3. Select Fatal from the columns list and click **OK**.

   The rows that contain accidents involving fatalities are colored red.
4. Select **Analyze > Text Explorer**.
5. Select Narrative Cause from the Select Columns list and click **Text Columns**.
6. From the Stemming list, select **Stem All Terms**.
7. From the Tokenizing list, select **Basic Words**.
8. Click **OK**.

Figure 12.12 Text Explorer Report for Narrative Cause

Text Explorer for Narrative Cause

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-empty Cases	Portion Non-empty per Case
1909	1906	50843	26.6752	1902	0.9979

Term and Phrase Lists

Term	Count
pilot·	1693
failur·	1045
factor·	998
land·	880
result·	818
maintain·	645
control·	610
loss·	488
engin·	464
terrain·	391
flight·	386
accid·	369
airplan·	363
power·	363
condit·	343
inadequ·	338
contribut·	333
fuel·	323
improp·	259
due·	244
aircraft·	243
direct·	239
takeoff·	222
reason·	218
subsequ·	216
undetermin·	214

Phrase	Count	N
pilot's failure	484	2
failure to maintain	458	3
pilot's failure to maintain	330	4
engine power	298	2
loss of engine	280	3
loss of engine power	273	4
directional control	221	2
forced landing	206	2
contributing factor	175	2
pilot's inadequate	174	2
maintain directional control	163	3
maintain directional	163	2
undetermined reasons	158	2
failure to maintain directional	129	4
pilot's improper	119	2
contributing factors	114	2
aircraft control	105	2
landing gear	102	2
inadvertent stall	88	2
fuel exhaustion	79	2
associated with the accident	76	4
control of the airplane	76	4
landing roll	73	2
loss of control	72	3
unsuitable terrain	72	2
power due	71	2

From the report, you see that there are almost 51,000 tokens and about 1,900 unique terms.

9. Select the top three terms in the Term List.

   Because there are approximately 51,000 tokens, the counts for these three terms represent almost 2% or more of all the terms.

10. Right-click on the selected terms and select **Add Stop Word**.

    Because these terms occur frequently compared to other terms, they do not provide much information to differentiate among documents.

11. JMP PRO From the red triangle menu next to Text Explorer for Narrative Cause, select **Latent Semantic Analysis, SVD**.

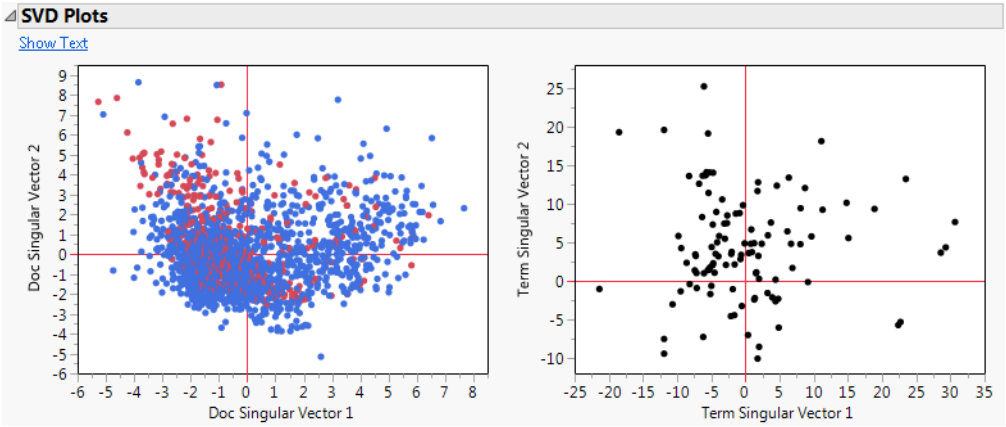
    This is the first analysis step toward topic analysis, which performs a rotation of the SVD.

12. JMP PRO In the Specifications window, type 50 for Minimum Term Frequency.

    Because there are approximately 51,000 tokens, this frequency is equivalent to a term that represents at least 0.1% of all the terms.

13. JMP PRO Click **OK**.

Figure 12.13 SVD Plots for Narrative Cause



There is not a lot of difference in the document SVD plot between fatal and non-fatal incidents.

14. JMP PRO From the Text Explorer red triangle menu, select **Topic Analysis, Rotated SVD**.

    You want to look for groups of terms that form topics.

15. JMP PRO Type 5 for Number of Topics.
16. JMP PRO Click **OK**.

Figure 12.14 Topic Words for Narrative Cause

Topic Words

Topic1		Topic2		Topic3		Topic4		Topic5	
Term	Score	Term	Score	Term	Score	Term	Score	Term	Score
power·	0.33882	factor·	0.2591	altitud·	0.2607	fuel·	0.2780	experi·	0.28338
loss·	0.33319	condit·	0.2368	low·	0.2427	personnel·	0.2558	control·	0.28050
engin·	0.30796	unsuit·	0.2306	dark·	0.2166	mainten·	0.2508	direct·	0.26325
forc·	0.30338	select·	0.2142	clearanc·	0.2106	result·	0.2289	lack·	0.25290
suitabl·	0.29425	associ·	0.2099	night·	0.2081	preflight·	0.2281	pilot·	0.23409
lack·	0.27698	failur·	-0.2049	pilot·	0.2034	exhaust·	0.2075	student·	0.21373
reason·	0.24825	area·	0.2009	instrument·	0.2034	reason·	-0.2065	maintain·	0.20806
undetermin·	0.24184	accid·	0.1954	maintain·	0.1969	undetermin·	-0.2046	factor·	0.18966
terrain·	0.18190	compens·	0.1899	airspe·	0.1738	inspect·	0.2042	aircraft·	0.18793
total·	0.16488	wind·	0.1755	stall·	0.1715	plan·	0.2041	failur·	0.18463
contribut·	0.14385	result·	-0.1686	maneuv·	0.1704	improp·	0.1938	instructor·	0.15164
		airspe·	-0.1656	continu·	0.1652	inadequ·	0.1916	crosswind·	0.14927
		inadequ·	0.1617	condit·	0.1600	subsequ·	0.1793	total·	0.14804
		stall·	-0.1613	land·	-0.1530	maintain·	-0.1683	airplan·	0.14728
		maintain·	-0.1525	weather·	0.1491				

The highest scoring words for each topic enable you to interpret whether the topic is capturing a theme in the incident reports.

For example, Topic 1 has high scores for power, loss, and engine, indicating a theme of losing power to the engine as a cause of the incident. This corresponds to the phrase “loss of engine power” occurring 273 times in the set of incident reports.

Based on the words with high scores in Topic 3, it can be described as being related to incidents that involved darkness or low altitude.

At this stage of the text analysis, you have many choices in how to proceed. Text analysis is an iterative process, so you might use topic information to further curate your term list by adding stop words or specifying phrases. You might save the weighted document-term matrix, the vectors from the SVD or rotated SVD as numeric columns in your data table and use them in other JMP analysis platforms. When you use these columns in other platforms, you can also include other columns from your data table in further analyses.

Appendix A

Statistical Details

Basic Analysis

This appendix contains statistical details that apply to multiple platforms in JMP Pro.

Platforms That Support Validation

The table below lists the types of crossvalidation available in each platform.

Parent Platform	Platform	Use Excluded Rows as Validation Holdback	Random Validation Holdback	K-Fold Cross-Validation	Validation Role Column
Fit Model	Fit Least Squares	No	No	No	Yes (only for model evaluation)
Fit Model	Forward Stepwise Regression	No	No	Yes (continuous response models only)	Yes
Fit Model	Logistic Regression	No	No	No	Yes (only for model evaluation)
Fit Model	Generalized Regression	Yes	Yes	Yes	Yes
Fit Model	Partial Least Squares	Yes	Yes	Yes	Yes
Partition	K Nearest Neighbors	Yes	Yes	Yes	Yes
Partition	Decision Tree	Yes	Yes	Yes	Yes
Partition	Bootstrap Forest	Yes	Yes	No	Yes
Partition	Boosted Tree	Yes	Yes	No	Yes
Neural	Neural	Yes	Yes	Yes	Yes

Appendix B

References

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.

Agresti, A., and Coull, B. (1998), "Approximate is Better Than 'Exact' for Interval Estimation of Binomial Proportions," *The American Statistician*, 52, 119–126

Akaike, H. (1974), "Factor Analysis and AIC," *Pschychometrika*, 52, 317–332.

Akaike, H. (1987), "A new Look at the Statistical Identification Model," *IEEE Transactions on Automatic Control*, 19, 716–723.

American Society for Quality Statistics Division (2004), *Glossary and Tables for Statistical Quality Control*, Fourth Edition, Milwaukee: Quality Press. "The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation," *JRSS* Suppl 8, 128–138.

Baglama, J. and Reichel, L. (2005), "Augmented Implicitly Restarted Lanczos Bidiagonalization Methods," *SIAM Journal on Scientific Computing* 27, 19-42.

Bartlett, M.S. (1966), *An Introduction to Stochastic Processes*, Second Edition, Cambridge: Cambridge University Press.

Bissell, A. F. (1990), "How Reliable is Your Capability Index?", *Applied Statistics*, 30, 331–340.

Brown, M.B. and Benedetti, J.K. (1977), "Sampling Behavior of Tests for Correlation in Two-Way Contingency Tables," *Journal of the American Statistical Association* 72, 305–315.

Brown, M.B. and Forsythe, A.B. (1974), "The Small Sample Behavior of Some Statistics Which Test the Equality of Several Means," *Technometrics* 16:1, 129–132.

Brown, M.B. and Forsythe, A.B. (1974), "Robust tests for the equality of variances," *Journal of the American Statistical Association*, 69, 364–367.

Chen, Song Xi and Hall, Peter. (1993), "Empirical Likelihood Confidence Intervals for Quantiles," *The Annals of Statistics* 21(3): 1166–1181.

Chou, Y.M., Owen, D.B., and Borrego, S.A. (1990), "Lower confidence limits on process capability indices," *Journal of Quality Technology*, 22(3): 223–229.

Cleveland, W.S. (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association* 74 (368): 829–836.

Cleveland, W.S. (1994). *Visualizing Data*, Summit, N.J: Hobart Press.

Cohen, J. (1960), "A coefficient of agreement for nominal scales," *Education Psychological Measurement*, 20: 37–46.

Cochran, W.G. and Cox, G.M. (1957). *Experimental Designs*, Second Edition, New York: John Wiley and Sons.

Conover, W. J. (1972). "A Kolmogorov Goodness-of-fit Test for Discontinuous Distributions". *Journal of the American Statistical Association* 67: 591–596.

Conover, W.J. (1980), *Practical Nonparametric Statistics*, New York: John Wiley and Sons, Inc.

Conover, W.J. (1999), *Practical Nonparametric Statistics*, 3rd Edition, New York: John Wiley and Sons, Inc.

Cureton, E.E. (1967), "The Normal Approximation to the Signed-Rank Sampling Distribution when Zero Differences are Present," *Journal of the American Statistical Association*, 62:319, 1068–1069.

DeLong, E., Delong, D, and Clarke-Pearson, D.L. (1988), "Comparing the Areas Under Two or more Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," *Biometrics* 44, 837–845.

Devore, J. L. (1995), *Probability and Statistics for Engineering and the Sciences*, Duxbury Press, CA.

Dunn, O.J. (1964), "Multiple Comparisons Using Rank Sums," *Technometrics* 6, 241–252.

Dunnett, C.W. (1955), "A multiple comparison procedure for comparing several treatments with a control" *Journal of the American Statistical Association*, 50, 1096–1121.

Dwass, M. (1955), "A Note on Simultaneous Confidence Intervals," *Annals of Mathematical Statistics* 26: 146–147.

Efron, B. (1981), "Nonparametric Standard Errors and Confidence Intervals," *The Canadian Journal of Statistics* 9:2, 139–158.

Eubank, R.L. (1999), *Nonparametric Regression and Spline Smoothing*, Second Edition, Boca Raton, Florida: CRC.

Fisher, L. and Van Ness, J.W. (1971), "Admissible Clustering Procedures," *Biometrika*, 58, 91–104.

Fleiss, J.L., Cohen J., and Everitt, B.S. (1969), "Large-Sample Standard Errors of Kappa and Weighted Kappa," *Psychological Bulletin*, 72: 323–327.

Friendly, M. (1991), "Mosaic Displays for Multiway Contingency Tables," *New York University Department of Psychology Reports*: 195.

Golub, G.H. and van der Vorst, H.A., (2000), "Eigenvalue Computation in the 20th Century," *Journal of Computational and Applied Mathematics* 123, 35-65.

Goodman, L.A. and Kruskal, W.H. (1979), *Measures of Association for Cross Classification*, New York: Springer-Verlag (reprint of JASA articles).

Gupta, S.S. (1965), On Some Multiple Decision (selection and ranking), Rules., *Technometrics* 7, 225–245.

Hahn, G. J. and Meeker, W. Q. (1991) *Statistical Intervals: A Guide for Practitioners*. New York: Wiley.

Hajek, J. (1969), *A Course in Nonparametric Statistics*, San Francisco: Holden–Day.

Hartigan, J.A. and Kleiner, B. (1981), "Mosaics for Contingency Tables," *Proceedings of the 13th Symposium on the Interface between Computer Science and Statistics*, Ed. Eddy, W. F., New York: Springer–Verlag, 268–273.

Hayter, A.J. (1984), "A proof of the conjecture that the Tukey–Kramer multiple comparisons procedure is conservative," *Annals of Mathematical Statistics*, 12 61–75.

Hosmer, D.W. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley and Sons.

"Hot Dogs," (1986), *Consumer Reports* (June), 364–367.

Hsu, J. (1981), "Simultaneous confidence intervals for all distances from the 'best'," *Annals of Statistics*, 9, 1026–1034.

Hsu, J. (1989), *Tutorial Notes on Multiple Comparisons*, American Statistical Association, Washington, DC.

Hsu, J.C. (1996), *Multiple Comparisons: Theory and Methods*, Chapman and Hall.

Huber, Peter J. (1973), "Robust Regression: Asymptotics, Conjecture, and Monte Carlo," *Annals of Statistics*, Volume 1, Number 5, 799–821.

Huber, P.J. and Ronchetti, E.M. (2009), *Robust Statistics,* Second Edition, Wiley.

Iman, R.L. (1974), "Use of a t-statistic as an Approximation to the Exact Distribution of Wilcoxon Signed Ranks Test Statistic," *Communications in Statistics—Simulation and Computation*, 795–806.

Kendall, M. and Stuart, A. (1979), *The Advanced Theory of Statistics, Volume 2*, New York: Macmillan Publishing Company, Inc.

Kramer, C.Y. (1956), "Extension of multiple range tests to group means with unequal numbers of replications," *Biometrics*, 12, 309–310.

Lehmann, E.L. (2006), Nonparametrics: Statistical Methods Based on Ranks, 2nd edition, Springer.

Levene, H. (1960), "Robust tests for the equality of variances" In I. Olkin (ed), *Contributions to probability and statistics*, Stanford Univ. Press.

Mason, R.L. and Young, J.C. (2002), *Multivariate Statistical Process Control with Industrial Applications,* Philadelphia: ASA-SIAM.

McCullagh, P. and Nelder, J.A. (1983), *Generalized Linear Models*, London: Chapman and Hall Ltd.

Meeker, W.Q. and Escobar, L.A. (1998), *Statistical Methods for Reliability Data*, pp. 60–62, New York: John Wiley and Sons.

Nagelkerke, N.J.D. (1991), "A Note on a General Definition of the Coefficient of Determination," *Biometrika*, 78:3 691–692.

Nelson, P.R., Wludyka, P.S., and Copeland, K.A.F. (2005), *The Analysis of Means: A Graphical Method for Comparing Means, Rates, and Proportions*, Philadelphia: Society for Industrial and Applied Mathematics.

Neter, J., Wasserman, W. and Kutner, M.H. (1990), *Applied Linear Statistical Models*, Third Edition, Boston: Irwin, Inc.

O'Brien, R.G. (1979), "A general ANOVA method for robust tests of additive models for variances," *Journal of the American Statistical Association*, 74, 877–880.

O'Brien, R., and Lohr, V. (1984), "Power Analysis For Linear Models: The Time Has Come," *Proceedings of the Ninth Annual SAS User's Group International Conference*, 840–846.

Odeh, R. E. and Owen, D. B. (1980) *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*. New York: Marcel Dekker, Inc.

Olejnik, S.F. and Algina, J. (1987), "Type I Error Rates and Power Estimates of Selected Parametric and Nonparametric Tests of Scale," *Journal of Educational Statistics* 12, 45–61.

Pratt, J.W. (1959), "Remarks on Zeros and Ties in the Wilcoxon Signed Rank Procedures", Journal of the American Statistical Association, 54:287, 655–667.

Reinsch, C.H. (1967), *Smoothing by Spline Functions*, Numerische Mathematik, 10, 177–183.

Rousseuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley and Sons.

Rubin, D. (1981), "The Bayesian Bootstrap", *The Annals of Statistics*, 9:1, 130–134.

SAS Institute (2008), *SAS/STAT 9.2 User's Guide Introduction to Nonparametric Analysis*, Chapter 16, SAS Institute Inc., Cary NC. Retrieved May 27, 2016 from https://support.sas.com/documentation/cdl/en/statugnonparam/61764/PDF/default/statugnonparam.pdf.

Schafer, J. and Strimmer, K. (2005), "A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics," *Statistical Applications in Genetics and Molecular Biology*, 4:1, 1–30.

Slifker, J. F. and Shapiro, S. S. (1980). *Technometrics*, 22, 239–246.

Snedecor, G.W. and Cochran, W.G. (1980), *Statistical Methods*, 7th edition, Ames, Iowa: Iowa State University Press.

Somers, R.H. (1962), "A New Asymmetric Measure of Association for Ordinal Variables," *American Sociological Review*, 27, 799–811.

Tan, Charles Y., and Iglewicz, Boris (1999), "Measurement-methods Comparisons and Linear Statistical Relationship," *Technometrics*, 41:3, 192–201.

Tamhane, A. C. and Dunlop, D. D. (2000) *Statistics and Data Analysis*, Prentice Hall.

Tukey, J. (1953), "A problem of multiple comparisons," Dittoed manuscript of 396 pages, Princeton University.

Welch, B.L. (1951), "On the comparison of several mean values: an alternative approach," *Biometrika* 38, 330–336.

Wheeler, D.J. (2003), *Range Based Analysis of Means*, SPC Press, Knoxville, TN.

Wilson, E. B. (1927), "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, 22, 209-212.

Wludyka, P.S. and Nelson, P.R. (1997), "An Analysis-of-Means-Type Test for Variances From Normal Populations", *Technometrics*, 39:3, 274–285.

Wludyka, P.S. and Nelson, P.R. (1999), "Two Non-Parametric, Analysis-of-Means-Type Tests for Homogeneity of Variances," *Journal of Applied Statistics*, 26:2, 243–256.

Index
Basic Analysis

Numerics

3D Correspondence Analysis option 220–229
5% Contours option 124

A

Agreement Statistic option 213, 221–222, 235
All fits 62–89
All Graphs option 144–145
All Pairs, Tukey HSD test 156, 158–180
Analysis of Means
  charts 154–175
  for Proportions option 213, 219
  options 155
Analysis of Means Methods 142, 152–155
Analysis of Variance
  See also Oneway
  report 110–111, 149–150
ANOM
  See also Analysis of Means
  for Variances 175–176
  Transformed Ranks 153
ANOMV display options 155

B

Bartlett test 168
Beta Binomial fit 91–92
Beta fit 86
Binomial fit 91
Bivariate Normal Ellipse menu 122–124
Bivariate platform 99
  example 100
  launching 100
  options 102–105, 122–125
  report window 101
Block button 97
Block Means report 150
bootstrapping 309–323
Box plots 144
Brown-Forsythe test 168

C

Capability Analysis option 59–61, 79–83
CDF plot 143–144, 173
Cell Labeling option 215
Change Format option 265
Change Item Label 273
Change Plot Scale option 265
Cluster Documents, Text Explorer 347
Cluster Terms, Text Explorer 347
Cochran Armitage Trend Test 214, 224
Cochran Mantel Haenszel test 213, 221
Coefficient of Variation 268
Color by Density Quantile option 125
Color Theme option 216
column tables 272
Compare Densities option 173
Compare Means options 143, 156–161
comparison circles 145, 157–158, 202–203
Composition of Densities option 173
Confid Curves Fit option 123
Confid Curves Indiv option 123
Confid Quantile option 160
Confid Shaded Fit option 124
Confid Shaded Indiv option 124
Confidence Interval options 47, 56
Confidence Limits option 65
Connect Means option 145–146
Connecting Letters Report 160
Contingency platform 209
  example 210–211
  launching 211
  options 213–214
  report window 211

Contingency Table 213, 216–218
Continuous Fit options 61–89
Contour Fill option 124
Contour Lines option 124
Correlation report 120
Correspondence Analysis option 213, 219
Count Axis option 46, 49
Covariance of Estimates report 247

D

Densities options 143, 173
Density Axis option 46, 49
Density Curve option 63
Density Ellipse option 119–120
density estimation fits 104
Detailed Comparisons Report 161
Details report 220
Diagnostic Plot option 63–65
Difference Matrix option 160
Discrete Fit options 62–92
Display Options 144–145, 155, 219
  for categorical variables 46
  for continuous variables 47–48
Distribution platform 33
  categorical variables in 34
  continuous variables in 34
  example 34–36
  launching 36
  options 45–58
  report window 37–39
Document Term Matrix, Text Explorer 326

E

Each Pair, Student's t test 156, 158–178
Equal Variances option 117
Equivalence Test option 55, 143, 170–188
Exact Test 214, 225
  See also Fisher's Exact Test
  options 162
Exponential fit 85
Extreme Value fit 85

F

Fisher's Exact Test 219
Fit Each Value
  command 116–117
  menu 122–123
  report 117
Fit Line command 103, 106–113, 133
Fit Mean
  command 103, 105–106
  menu 122–123
  report 106
Fit Orthogonal command 117–118, 133
Fit Polynomial command 104, 106–136
Fit Special command 104, 113–114
Fit Spline command 115–116, 133
Fit X to Y option 117
Fit Y by X platform 95
Fix Parameters option 63
Frequencies
  option 46
  report 42

G

Gamma fit 85
Gamma Poisson fit 90–91
GLog fit 88–89
Goodness of Fit
  option 63
  tests 65–93
Gradient Between Ends option 216
Gradient Between Selected Points option 216
Grand Mean option 145
Graph in Variance Scale option 155
Group By option 102, 121–122

H

Histogram Options 48
histograms 39–66
  borders option 102, 105
  color option 46, 48
  creating subsets 39
  highlighting data 39, 41
  option in Oneway 145
  options for continuous variables 48–50
  rescaling 39
  resizing 39–40
  selecting data 41

specifying data 39
Horizontal Layout option 46, 48

I

Include missing for grouping columns option 265
Inverse Prediction option 248
Iterations report 245

J

jitter 251
Johnson fits 87–88

K

Kernel Control option 124
Kolmogorov Smirnov Test 162
Kruskal-Wallis test, *see* Wilcoxon Test

L

Lack of Fit report 109–110
language processing, Text Explorer 332
Latent Class Analysis option, Text Explorer 347, 352
Latent Semantic Analysis, SVD in Text Explorer 347
Levene test 168
Lift Curve option 248
Line Color option 123
Line of Fit option 65, 123, 172
Line Style option 123
Line Width option 123
Linear Fit
menu 122–124
report 107–136
Logistic platform 239
See also logistic regression
Covariance of Estimates report 247
examples 241–242, 250–254
Iterations report 245
launching 242
logistic plot 244, 248
options 247–250
Parameter Estimates report 247
report window 243–247
Whole Model Test report 245–247
logistic regression 239–240
LogNormal fit 84
LSD Threshold Matrix option 160

M

Macros option 216
Make Into Data Table option 274
Mann-Whitney test, *see* Wilcoxon Test
Matching Column option 144, 173–174
Matching Dotted Lines option 145
Matching Lines option 145, 174
Max (summary statistics) 267
Mean CI Lines option 145
Mean Diamonds option 145, 150–151
Mean Error Bars option 145, 151
Mean Lines option 145, 151
Mean of Means option 145
Means and Std Dev option 142
Means for Oneway Anova report 150
Means/Anova option 142, 147
Means/Anova/Pooled t option 142, 147
Measures of Association option 213, 223–224
Median (summary statistics) 268
Median Reference Line option 65
Median Test 161
Mesh Plot option 125
Min (summary statistics) 267
Model Clustering option 125
Mosaic Plot
in Contingency 213–216
option in Distribution 47

N

N and N Missing (summary statistics) 267
N Categories (summary statistics) 267
Nonpar Density command 120–121
Nonparametric
options 143
tests 161
Nonparametric Bivariate Density report 121
Normal fit 83
Normal Mixtures fits 86
Normal Quantile Plot 49–50, 143, 172
Number of Cases, Text Explorer 339

Number of Non-empty Cases, Text Explorer 339

O

O'Brien test 167
Odds Ratios option 213, 248
Oneway platform 137–138
  example 138–140
  launching 140
  options 141–145
  plot 140
Order by count of grouping columns option 265
Order By option 47
Ordered Differences Report 160
Orthogonal Fit Ratio menu 122–123
Orthogonal Regression report 117
Outlier Box Plot 51–52

P

Parameter Estimates report 112–136, 247
parametric resampling, Simulate platform 283
Parsing Options, Text Explorer 343
Plot Actual by Quantile option 172
Plot Options menu in Logistic 248
Plot Quantile by Actual option 172
Plot Residuals option 124
Points Jittered option 145
Points option 144
Points Spread option 145
Poisson fit 90
Polynomial Fit Degree menu 122
Polynomial Fit report 107–136
Portion of Non-empty Cases, Text Explorer 339
Power option 143, 171–172, 203–204
Prediction Interval option 58–70, 78
Prob Axis option 46, 49
Proportion of Densities option 173

Q

Quantile Box Plot 52
Quantile Density Contours menu 122–125
Quantiles
  option 47, 63, 73, 142, 146–147
  report 42

R

ranges, summary statistics 267
regression fits for Bivariate 104
regular expression editor, Text Explorer 333
Relative Risk option 213, 222
Remove Column Label 273
Remove Fit option 64, 123
Report option 123
Restore Column Label 273
Revert to Old Colors option 216
ROC Curve option 248
Rotate option 65
row profile 220
row tables 272

S

sample size, in tabulate 274
Save Coefficients option 124
Save Colors to Column option 216
Save commands in Distribution 57–58
Save Density Formula command 63
Save Density Grid option 125
Save Density Quantile option 125
Save Fitted Quantiles command 63
Save for Adobe Flash platform (.SWF) option 45
Save options
  in Oneway 144
Save Predicteds option 123
Save Probability Formula option 248
Save Residuals option 123
Save Spec Limits command 64
Save Value Ordering option 220
Saved Transformed command 64
Select Colors for Values window 215
Select Points by Density option 124
Select Points Inside option 124
Select Points Outside option 124
Separate Bars option 46
Set Alpha Level option
  for Bivariate 124
  for Contingency 213, 219
  for Oneway 143, 155

Set Bin Width option 48
Set Colors option 215
Set Spec Limits for K Sigma option 63
Shaded Contour option 124
Shadowgram option 48
Show Center Line option 155
Show Chart option 273
Show Control Panel option 274
Show Counts option 46, 49
Show Decision Limit Shading option 155
Show Decision Limits option 155
Show Delimiters, Text Explorer 342
Show Percents option 46, 49
Show Points option
  for Bivariate 102
Show Shading option 274
Show Stem Report, Text Explorer 342
Show Summary Report option 155, 219
Show Table option 273
Show Test Build Panel option 274
Show tooltip option 273
Simulate platform data table results 307
Smooth Curve fit 87
Smoothing Spline Fit
  menu 122
  report 115
Spec Limits option 63, 65
Specified Variance Ratio option 117
Specify Transformation or Constraint
  window 113, 126
Stack option 45
Standard Deviation 267
Standard Error 268
Std Dev and Std Err (summary statistics) 267
Std Dev Lines option 145, 151
Std Error Bars option 46, 48
Steel With Control test 165
Steel-Dwass All Pairs test 165
Stem and Leaf plot 52
Stemming, Text Explorer 326, 332
stop words, Text Explorer 326, 345
Student's t test 156, 158–178
Subset option 39
Summary of Fit report 108, 147
Summary Statistics
  customize 48
  options 45
  report 42, 48
SVD Scatterplot Matrix, Text Explorer 347
Switch Response Level for Proportion
  option 219

T

t test
  option 142
  report 148–149
Tabulate 259–282
term (token), Text Explorer 326
Term Options, Text Explorer 342
Test Mean option 53–54
Test Probabilities option 47–68
Test Std Dev option 54–55
Tests
  option in Contingency 213
  report 218–219
Text Explorer
  Cluster Documents 347
  Cluster Terms 347
  editing regular expressions 333
  language processing 332
  Latent Class Analysis option 347, 352
  Latent Semantic Analysis, SVD 347
  Number of Cases 339
  Number of Non-empty Cases 339
  Parsing Options 343
  platform options 341
  Portion of Non-empty Cases 339
  Show Delimiters 342
  Show Stem Report 342
  Stemming 326, 332
  stop words 326, 345
  SVD Scatterplot Matrix 347
  term (token) 326
  Term Options 342
  Tokenizing 332
  Tokens per Case 339
  Topic Analysis, Rotated SVD 347
  Topic Scatterplot Matrix 348
  Total Tokens 339
Text Explorer, Word Separator List 336
Tokenizing, Text Explorer 332
Tokens per Case, Text Explorer 339

Tolerance Interval option 59–71, 78–79
Topic Analysis, Rotated SVD in Text
    Explorer 347
Topic Scatterplot Matrix, Text Explorer 348
Total Tokens, Text Explorer 339
Transformed Fit
    menu 122
    report 114
Treat Numbers as Words, Text Explorer 333
Two Sample Test for Proportions 213, 222–223

U

Unequal Variances 143, 167–168
Uniform plot scale option 265
Uniform Scaling option 45
Univariate Variances, Prin Comp option 117

V

van der Waerden Test 161
Variance (summary statistics) 268
Vertical option 46, 48

W-Z

Weibull fits 84
Welch's test 168
Whole Model Test report 245–247
Wilcoxon
    Each Pair test 165
    Test 161
With Best, Hsu MCB test 156, 159–181
With Control, Dunnett's test 156, 160
Word Separator List, Text Explorer 336
X Axis proportional option 145, 150–151
X, Continuous Regressor button 97
X, Grouping button 97
X, Grouping Category button 97
X, Regressor button 97
Y, Categorical Response button 97
Y, Response Category button 97