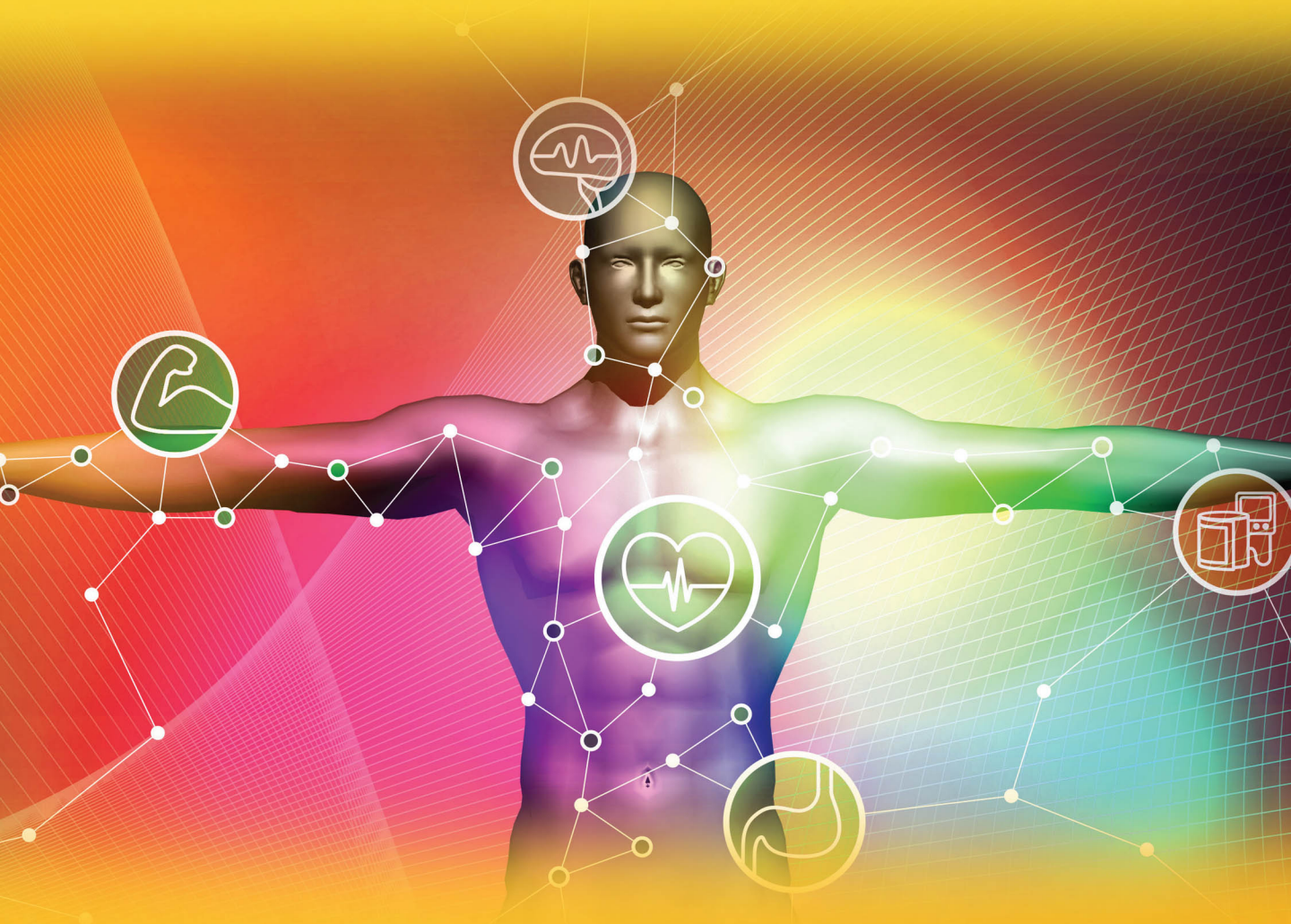
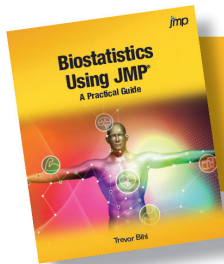


Biostatistics Using JMP®

A Practical Guide



Trevor Bihl



From *Biostatistics Using JMP: A Practical Guide*.
Full book available for purchase [here](#).

Contents

Dedication.....	iii
Acknowledgments	xi
About This Book.....	xiii
About the Author	xvii
Chapter 1: Introduction	1
1.1 Background and Overview.....	1
1.2 Getting Started with JMP	2
1.3 General Outline	4
1.4 How to Use This Book	5
1.5 Reference.....	5
Chapter 2: Data Wrangling: Data Collection	7
2.1 Introduction	7
2.2 Collecting Data from Files.....	8
2.2.1 JMP Native Files	8
2.2.2 SAS Format Files	9
2.2.3 Excel Spreadsheets.....	10
2.2.4 Text and CSV Format	11
2.3 Extracting Data from Internet Locations.....	14
2.3.1 Opening as Data	14
2.3.2 Opening as a Webpage.....	15
2.4 Data Modeling Types.....	17
2.4.1 Incorporating Expression and Contextual Data.....	18
2.5 References.....	19
Chapter 3: Data Wrangling: Data Cleaning	21
3.1 Introduction	21
3.2 Tables.....	21
3.2.1 Stacking Columns	24
3.2.2 Basic Table Organization.....	26
3.2.3 Column Properties	31
3.3 The Sorted Array	32

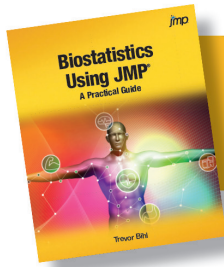
3.4 Restructuring Data.....	34
3.4.1 Combining Columns.....	35
3.4.2 Separating Out a Column (Text to Columns).....	36
3.4.3 Creating Indicator Columns	36
3.4.4 Grouping Inside Columns	38
3.5 References.....	41
Chapter 4: Initial Data Analysis with Descriptive Statistics	45
4.1 Introduction	45
4.2 Histograms and Distributions.....	45
4.2.1 Histograms.....	46
4.2.2 Box Plots	55
4.2.3 Stem-and-Leaf Plots	57
4.2.4 Pareto Charts.....	58
4.3 Descriptive Statistics.....	64
4.3.1 Sample Mean and Standard Deviation.....	66
4.3.2 Additional Statistical Measures	67
4.4 References.....	69
Chapter 5: Data Visualization Tools.....	71
5.1 Introduction	71
5.2 Scatter Plots.....	72
5.2.1 Coloring Points	75
5.2.2 Copying Better-Looking Figures.....	77
5.2.3 Multiple Scatter Plots.....	79
5.3 Charts.....	81
5.4 Multidimensional Plots	84
5.4.1 Parallel Plots	84
5.4.2 Cell Plots	87
5.5 Multivariate and Correlations Tool.....	89
5.5.1 Correlation Table	91
5.5.2 Correlation Heat Maps.....	92
5.5.3 Simple Statistics	93
5.5.4 Additional Multivariate Measures	93
5.6 Graph Builder and Custom Figures.....	94
5.6.1 Graph Builder Custom Colors	96
5.6.2 Incorporating Contextual Data.....	98
5.7 References.....	99

Chapter 6: Rates, Proportions, and Epidemiology	101
6.1 Introduction	101
6.2 Rates	101
6.2.1 Crude Rates	101
6.2.2 Adjusted Rates	105
6.3 Geographic Visualizations	108
6.3.1 National Visualizations.....	108
6.3.2 County and Lower Level Visualizations	116
6.4 References.....	120
Chapter 7: Statistical Tests and Confidence Intervals.....	123
7.1 Introduction	123
7.1.1 General Hypothesis Test Background	124
7.1.2 Selecting the Appropriate Method	125
7.2 Testing for Normality	126
7.2.1 Histogram Analysis	126
7.2.2 Normal Quantile/Probability Plot.....	128
7.2.3 Goodness-of-Fit Tests	131
7.2.4 Goodness-of-Fit for Other Distributions	132
7.3 General Hypothesis Tests	133
7.3.1 Z-Test Hypothesis Test of Mean.....	133
7.3.2 T-Test Hypothesis Test of Mean.....	135
7.3.3 Nonparametric Test of Mean (Wilcoxon Signed Rank)	136
7.3.4 Standard Deviation Hypothesis Test.....	140
7.3.5 Tests of Proportions.....	141
7.4 Confidence Intervals.....	144
7.4.1 Mean Confidence Intervals.....	144
7.4.2 Mean Confidence Intervals with Different Thresholds	144
7.4.3 Confidence Intervals for Proportions	145
7.5 Chi-Squared Analysis of Frequency and Contingency Tables.....	146
7.6 Two Sample Tests.....	150
7.6.1 Comparing Two Group Means	150
7.6.2 Paired Comparison, Matched Pairs.....	154
7.7 References.....	156
Chapter 8: Analysis of Variance (ANOVA) and Design of Experiments (DoE)	159
8.1 Introduction	159

8.2 One-Way ANOVA.....	161
8.2.1 One-Way ANOVA with Fit Y by X.....	161
8.2.2 Means Comparison, LSD Matrix, and Connecting Letters	165
8.2.3 Fit Y by X Changing Significance Levels.....	168
8.2.4 Multiple Comparisons, Multiple One-Way ANOVAs.....	169
8.2.5 One-Way ANOVA via Fit Model	171
8.2.6 One-Way ANOVA for Unequal Group Sizes (Unbalanced)	176
8.3 Blocking	179
8.3.1 One-Way ANOVA with Blocking via Fit Y by X.....	179
8.3.2 One-Way ANOVA with Blocking via Fit Model.....	182
8.3.3 Note on Blocking	183
8.4 Multiple Factors	183
8.4.1 Experimental Design Considerations	184
8.4.2 Multiple ANOVA	188
8.4.3 Feature Selection and Parsimonious Models.....	191
8.5 Multivariate ANOVA (MANOVA) and Repeated Measures	196
8.5.1 Repeated Measures MANOVA Background	196
8.5.2 MANOVA in Fit Model.....	197
8.6 References.....	201
Chapter 9: Regression and Curve Fitting.....	205
9.1 Introduction	205
9.2 Simple Linear Regression	206
9.2.1 Fit Y by X for Bivariate Fits (One X and One Y).....	206
9.2.2 Special Fitting Tools.....	208
9.3 Multiple Regression	211
9.3.1 Fit Model.....	211
9.3.2 Stepwise Feature Selection.....	214
9.3.3 Analysis of Covariance (ANCOVA).....	222
9.4 Nonlinear Curve Fitting and a Nonlinear Platform Example.....	226
9.5 References.....	232
Chapter 10: Diagnostic Methods for Regression, Curve Fitting, and ANOVA.....	233
10.1 Introduction	233
10.2 Computing Residuals with Fit Y by X and Fit Model	234
10.2.1 Fit Y by X.....	234
10.2.2 Fit Model.....	234
10.3 Checking for Normality	235

10.4 Checking for Nonconstant Error Variance (Heteroscedasticity)	236
10.5 Checking for Outliers.....	238
10.6 Checking for Nonindependence.....	242
10.7 Multiple Factor Diagnostics.....	243
10.8 Nonlinear Fit Residuals	245
10.9 Developing Appropriate Models.....	246
10.10 References.....	247
Chapter 11: Categorical Data Analysis	249
11.1 Introduction	249
11.2 Clustering.....	250
11.2.1 Hierarchical Clustering	250
11.2.2 K-means Clustering	260
11.3 Classification	263
11.3.1 JMP Data Preliminaries for Classification	265
11.3.2 Example Data Sets	267
11.4 Classification by Logistic Regression	268
11.4.1 Logistic Regression in Fit Y by X.....	268
11.4.2 Logistic Regression in Fit Model	270
11.5 Classification by Discriminant Analysis.....	273
11.5.1 Discriminant Analysis Loadings.....	275
11.5.2 Stepwise Discriminant Analysis.....	276
11.6 Classification with Tabulated Data	277
11.7 Classifier Performance Verification	280
11.8 References.....	284
Chapter 12: Advanced Modeling Methods	287
12.1 Introduction	287
12.2 Principal Components and Factor Analysis	288
12.2.1 Principal Components in JMP.....	288
12.2.2 Dimensionality Assessment	291
12.2.3 Factor Analysis in JMP	293
12.3 Partial Least Squares	296
12.4 Decision Trees.....	302
12.4.1 Classification Decision Trees in JMP.....	303
12.4.2 Predictive Decision Trees in JMP	308
12.5 Artificial Neural Networks	310
12.5.1 Neural Network Architecture	311
12.5.2 Classification Neural Networks in JMP.....	312
12.5.3 Predictive Neural Networks in JMP.....	315

12.6 Control Charts	317
12.7 References.....	321
Chapter 13: Survival Analysis.....	323
13.1 Introduction	323
13.2 Life Distributions	323
13.3 Kaplan-Meier Curves	327
13.3.1 Simple Survival Analysis	327
13.3.2 Multiple Groups	330
13.3.3 Censoring	331
13.3.4 Proportional Hazards	335
13.4 References.....	336
Chapter 14: Collaboration and Additional Functionality	339
14.1 Introduction	339
14.2 Saving Scripts and SAS Coding.....	339
14.2.1 Saving Scripts to Data Table	340
14.2.2 SAS Coding Functionality	341
14.3 Collaboration	342
14.3.1 Journals	342
14.3.2 Web Reports	344
14.4 Add-Ins.....	347
14.4.1 Finding Add-Ins.....	347
14.4.2 Developing Add-Ins	348
14.4.3 Example Add-In: Forest Plot / Meta-analysis	348
14.4.4 Add-In Version Control	351
14.5 References.....	352
Index	331



From *Biostatistics Using JMP: A Practical Guide*.
Full book available for purchase [here](#).

Chapter 1: Introduction

1.1 Background and Overview	1
1.2 Getting Started with JMP.....	2
1.3 General Outline.....	4
1.4 How to Use This Book.....	5
1.5 Reference	5

1.1 Background and Overview

This book evolved from personal experiences in both teaching and consulting in biostatistics. Although many biostatistics textbooks show computer outputs and results, they rarely show how to generate the results. Biostatistics instruction is also commonly theoretical and based on solving simple problems by hand. However, real-world data is usually more complicated than the simple examples, and I found that frequent collaborators—PhD-educated researchers who performed and managed experiments—needed more understanding of software to analyze their data. This is difficult because such researchers often spend a majority of their time performing experiments and use statistical software sparingly. They also often do not have access to a dedicated biostatistician in their office or are competing for the time of their office’s single biostatistician. Although such researchers might know the mechanics of a statistical method, they might not how to generate meaningful results using software.

Therefore, a practical, how-to guide to biostatistics was needed. There are many software applications available for statistics and biostatistics, so why JMP? As an educator, I found JMP an advantage to teaching. I could spend more time on theory and interpretation because JMP does not require scripts and syntax. As a collaborator and consultant, I found my colleagues would readily gravitate toward JMP and its results because of the graphical user interface (GUI) format and its ease of use. And finally, unless you want to code algorithms themselves, as a researcher, you will find JMP to be more user-friendly, correct, and developed when compared to many other competing packages. Incidentally, if you want to code, SAS programming abilities do exist in JMP. Thus, you can fully use JMP for analysis ranging from simple to complex and customized.

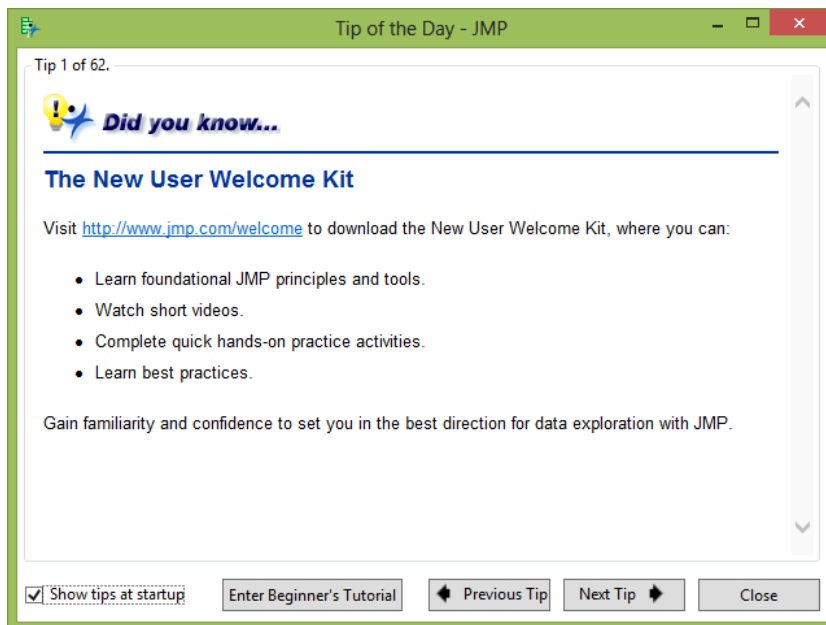
This book presents and solves problems germane to biostatistics with easy-to-reproduce examples. The book is also a general biostatistics reference that leverages the topics found in leading biostatistics books. This chapter introduces JMP, presents a general outline of the book contents, and provides a brief guide to using this book.

1.2 Getting Started with JMP

When you first run JMP, you will be greeted with a **Tip of the Day** (Figure 1.1). There are 62 tips of the day, and they show up whenever you start JMP. These tips can be useful to new JMP users in gaining familiarity with the software. However, if you don't want to see these tips further, you can do the following:

1. Clear **Show tips at start-up**.
2. Click **Close**.

Figure 1.1 Initial Tip of the Day



After you close the **Tip of the Day**, you are greeted with the primary JMP interface seen in Figure 1.2. Here, you can load data, create a new data table, or look for recently used files. If this is the first time you have opened JMP, there will be no recent files to consider. Thus, you must load or create a new data table.

To load a file:

- Click **File ► Open**.

or

- Click on the third icon on the taskbar.

To create a blank data table:

- Click **File ► New**.

or

- Click on the first icon on the taskbar.

Alternatively, if you want to load a built in JMP example data file, you can do so. A variety of files are available.

To load example data files:

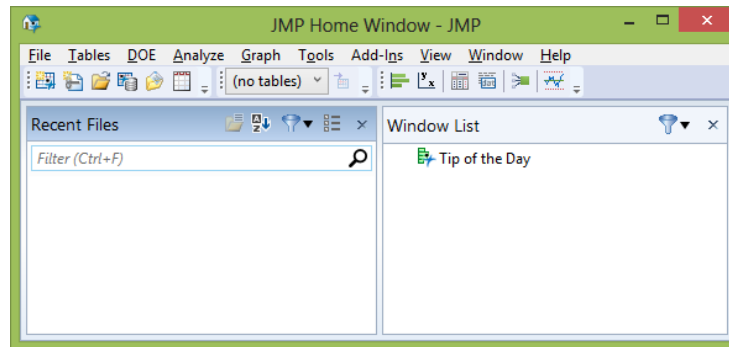
1. Click **Help ► Sample Data**.
2. Select a data file under the method of interest.

Also, you can select individual or multiple data tables in the **Window List** and then close all of these files. This is advantageous if you inadvertently opened many files, such as in a mistakenly setup Internet open.

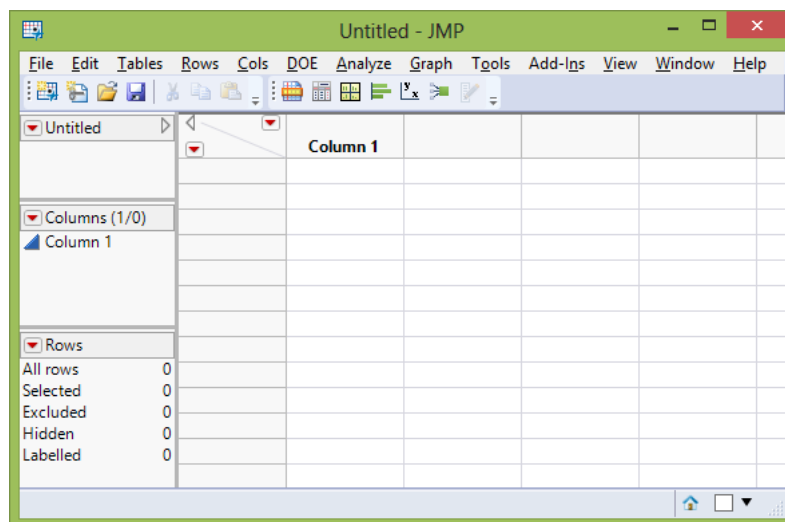
To close many open data tables:

1. Select the windows of interest.
2. Right-click and select **Close**.
3. You will then be prompted to save these files.

Figure 1.2 JMP Primary Interface



If you create a new data table, you will be presented with Figure 1.3. Here, you see that there is a spreadsheet-like table, with a *Column 1* ready for you to start considering. Also, when you have loaded and analyzed data, you can save these results to the JMP data table and instantly reload at a later date, as will be discussed in Section 14.2.

Figure 1.3 New Data Table

1.3 General Outline

With this basic usability knowledge from Section 1.2, you are now ready to consider bio-statistical data analysis. Biostatistics covers a wide variety of topics ranging from simple hypothesis tests to complex nonlinear algorithms. This book aims to cover the range of methods with varying levels of detail. To do so, this book is organized sequentially as outlined in Table 1.1.

Table 1.1 General Outline of *Biostatistics Using JMP: A Practical Guide*

Method	Chapter
Introduction	1
Data Wrangling: Data Collection	2
Data Wrangling: Data Cleaning	3
Initial Data Analysis with Descriptive Statistics	4
Data Visualization Tools	5
Rates, Proportions and Epidemiology	6
Statistical Tests and Confidence Intervals	7
Analysis of Variance (ANOVA) and Design of Experiments (DoE)	8
Regression and Curve Fitting	9
Diagnostic Methods for Regression, Curve Fitting and ANOVA	10
Categorical Data Analysis	11
Advanced Modeling Methods	12
Survival Analysis	13
Collaboration and Additional Functionality	14

1.4 How to Use This Book

In Chapters 2 and 3, this book moves to data-wrangling issues, such as data collection and cleaning. Since upward of 80% of your time can be spent in making messy data usable (Lohr, 2014), learning the tools JMP has for assembling and cleaning data is key and is covered in Chapters 2 and 3.

In Chapters 4 and 5, you can learn the basics of descriptive statistics and data visualizations in JMP. Following this, the primary focus is on modeling, which involves creating a mathematical representation (a model) of data or a system in order to make inferences about it.

After this, you have a few different paths available:

Chapter 6 discusses epidemiological and geographical interpretations.

Chapter 4 also discusses developing custom equations.

Chapter 7 discusses the various hypothesis test and confidence interval methods.

Chapters 8 to 10 discuss linear models such as analysis of variance (ANOVA), regression, and model validation.

Because of the interrelation of the underlying methods of regression (Chapter 9) and ANOVA (Chapter 8), diagnostic and remedial measures for these methods are discussed in Chapter 10.

Chapter 11 discusses classification methods, such as logistic regression, and clustering methods, such as k-means.

Chapters 7 to 10 largely deal with a continuous dependent (e.g., Y, variable (prediction)). Methods to analyze a discrete dependent variable are presented in Chapter 11.

Chapter 12 presents advanced modeling methods (e.g., factor analysis, neural networks, and control charts).

Chapter 13 introduces the vast array of survival analysis methods in JMP.

Chapter 14 presents methods that facilitate collaborating in addition to sources of additional functionality.

If you are using a previously created data set, then it is advantageous to start with data-wrangling methods and then look at the various analytical tools this book discusses. However, if you are starting a new experiment and will be collecting data, then you should start looking at Section 8.4.1, which discusses experimental design considerations and how to develop and select factor levels for an experiment.

1.5 Reference

Lohr, S. (2014, Aug. 18). For big-data scientists, ‘janitor work’ is key hurdle to insights. *New York Times*, p. B4.

About This Book

Rationale for This Book

This book focuses on the basics of statistical data analysis of biomedical/biological data using JMP. After both teaching and consulting in biostatistics, I saw a gap that existed between biostatistics books, which tend to be theoretical, and statistical software. To address this gap, I use statistical methods to analyze various biostatistics problems.

Importance of Statistical Analysis

Analytics, data mining, data science, and statistics are essentially synonyms, and describe finding meaning in data by developing mathematical models to find and describe relationships in the data. While many biostatistical applications are simple in nature, for example, a t-test to evaluate the mean differences in response due to a treatment, a wide variety of methods exists.

Biostatistics Focus

Biostatistics is the application of statistical methods to biological, or medical, data. While some methods see more frequent use in biostatistics, for example, survival analysis, these methods are not limited in use to just biostatistical problems. Essentially, all data is a matrix at the end of the day, and thus methods seen in biostatistical analysis can be applied to other domains.

The Power of JMP for Analytics

Familiarity with statistical methods enables one to analyze data via methods familiar in textbooks. However, many textbook examples are simple in nature, but real-world data rarely is. Thus, applying methods in a textbook can be frustrating if you have to wrestle both with the data and software.

This book was written with JMP due to the many advantages JMP has over other statistical software. JMP provides a GUI (graphical user interface) in which one can analyze data without coding algorithms. Additionally, the SAS underpinnings to JMP provide a wide and stable platform that can be trusted in its analysis. In total, JMP provides a tool that is easy to use and comes with a wide variety of built-in methods, the results of which can be trusted (something you can't say about all statistical software). And, for those who wish to code boutique algorithms, JMP also supports this as well.

Who Should Read This Book

This book is written for a variety of different persona groups. Although biostatistics is the focus, and is in the title, this book has broader appeal.

Biological/Medical Researchers and Laboratory Managers

Researchers in the sciences, for example, biology and medicine, spend a large majority of their time performing experiments and a small fraction of their time analyzing data. Remembering how to use software that is only accessed a few times a year can be challenging. Thus, this book is aimed particularly at this group and provides a practical guide to analyzing collected biological/medical data.

Statisticians and Data Scientists

This group might be interested in a broad look at how to use JMP to solve various problems and analyze data in JMP. While theory is light in this book, this group could easily learn the steps and nuances of JMP. Additionally, they would see practical data analysis and experimental data analysis using various JMP capabilities.

Students in Biostatistics or Statistics Classes

Many biostatistical courses use excellent textbooks that cover the theory and examples for a wide variety of problems. However, these textbooks rarely discuss how to solve the problems, leaving students with the need to either code equations or learn various statistical software programs on the fly. This book is written from a general standpoint and can thus be combined with any biostatistical textbook. Additionally, since the statistical methods themselves can be used in many domains, this book can be combined with multiple statistics courses and textbooks.

Biostatistics Methods and JMP Functionality Covered in This Book

This Book Covers the Following Biostatistics Methods

- Data Cleaning – Data Wrangling – Descriptive Statistics – Data Visualization
- Rates – Proportion – Geographical Visualization – Epidemiology
- Confidence Intervals – Hypothesis Tests
- Linear Regression – Curve Fitting – General Linear Models
- Analysis of Variance (ANOVA) – Analysis of Covariance (ANCOVA) – Remedial Measures for Regression and ANOVA
- Cluster Analysis – Hierarchical Clustering – K-means
- Classification Analysis – Logistic Regression – Discriminant Analysis
- Survival Analysis – Meta Analysis – Control Charts – Neural Networks – Decision Trees

Structure of This Book

Chapter 1 introduces this book and mirrors some content in this section. Additionally, Chapter 1 introduces how to start using JMP. Chapters 2 and 3 introduce data-wrangling issues, such as data collection and cleaning. These chapters are very helpful when analyzing real-world data using JMP. The basics of descriptive statistics and data visualization are presented in Chapters 4 and 5.

After Chapter 5, the focus of this book is on developing statistical models to describe data. Chapters 6 through 13 present various approaches, and your data and goals will drive which chapter you should read. Chapter 6 discusses epidemiology and geographical data analysis. Chapter 7 discusses hypothesis tests and confidence intervals. Chapters 8 to 10 present models such as analysis of variance, regression, curve fitting, and model validation. Chapter 11 discusses classification and clustering methods. Chapter 12 presents advanced modeling methods. Chapter 13 discusses survival analysis. Finally, Chapter 14 presents collaboration methods, incorporating custom JMP tools and meta-analysis, as an example.

Additional Resources

For downloads of sample data presented in this book, please visit my author page at:

<https://support.sas.com/bihl>

This site also includes downloadable color versions of selected figures that appear in this book. Since this book is printed in black and white, you might find that some color figures are easier to interpret and understand.

Please visit this site regularly, as I will provide updates on the content.

We Want to Hear from You

SAS Press books are written *by* SAS Users *for* SAS Users. We welcome your participation in their development and your feedback on SAS Press books that you are using. Please visit [sas.com/books](https://support.sas.com/books) to do the following:

- Sign up to review a book
- Recommend a topic
- Request information on how to become a SAS Press author
- Provide feedback on a book

Do you have questions about a SAS Press book that you are reading? Contact the author through saspress@sas.com or https://support.sas.com/author_feedback.

SAS has many resources to help you find answers and expand your knowledge. If you need additional help, see our list of resources at [sas.com/books](https://support.sas.com/books).

About the Author



Trevor Bihl is both a research scientist/engineer and an educator who teaches biostatistics, engineering statistics, and programming courses. He has been a SAS and JMP user since 2009 and provides various biostatistics and data mining consulting services. His background includes multivariate statistics, signal processing, data mining, and analytics. His educational background includes a BS and MS from Ohio University and a PhD from the Air Force Institute of Technology. He is the author of multiple journal and conference papers, book chapters, and technical reports.

Learn more about this author by visiting his author page at <http://support.sas.com/bihl>. There you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more.

Ready to take your SAS® and JMP® skills up a notch?



Be among the first to know about new books,
special events, and exclusive discounts.

support.sas.com/newbooks

Share your expertise. Write a book with SAS.

support.sas.com/publish

 sas.com/books
for additional books and resources.


THE POWER TO KNOW.

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. © 2017 SAS Institute Inc. All rights reserved. M1588358 US.0217