

*Making Everything Easier!™*

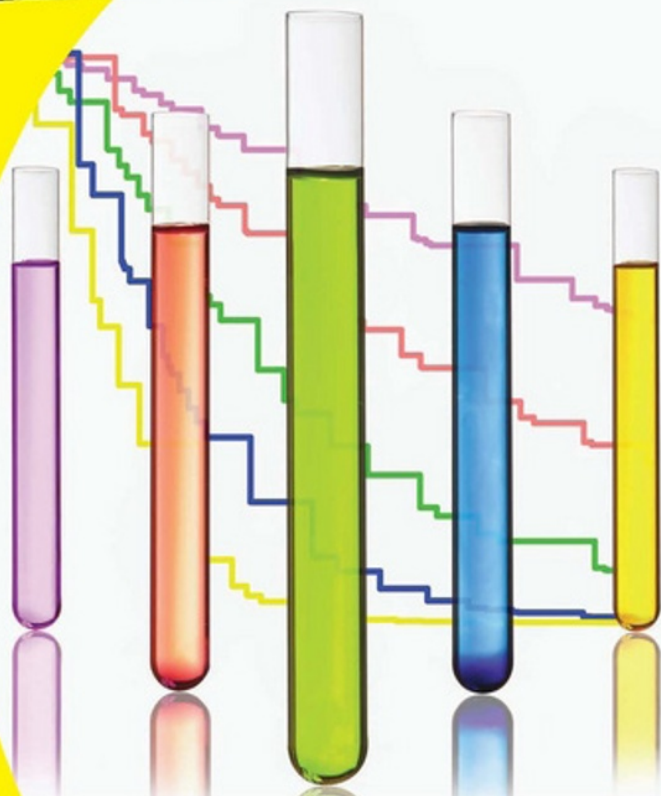
# ***Biostatistics***

FOR  
**DUMMIES®**  
A Wiley Brand

## ***Learn to:***

- Make sense of complex material
- Understand key concepts in statistics as they relate to biological concepts
- Score your highest in your biostatistics course

**John Pezzullo, PhD**





***Biostatistics***

FOR  
**DUMMIES<sup>®</sup>**  
A Wiley Brand

**by John C. Pezzullo, PhD**

FOR  
**DUMMIES<sup>®</sup>**  
A Wiley Brand

## **Biostatistics For Dummies®**

Published by  
**John Wiley & Sons, Inc.**  
111 River St.  
Hoboken, NJ 07030-5774  
[www.wiley.com](http://www.wiley.com)

Copyright © 2013 by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

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, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the Publisher. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

**Trademarks:** Wiley, the Wiley logo, For Dummies, the Dummies Man logo, A Reference for the Rest of Us!, The Dummies Way, Dummies Daily, The Fun and Easy Way, Dummies.com, Making Everything Easier, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc., and/or its affiliates in the United States and other countries, and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc., is not associated with any product or vendor mentioned in this book.

**LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.**

For general information on our other products and services, please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002.

For technical support, please visit [www.wiley.com/techsupport](http://www.wiley.com/techsupport).

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit [www.wiley.com](http://www.wiley.com).

Library of Congress Control Number: 2013936422

ISBN 978-1-118-55398-5 (pbk); ISBN 978-1-118-55395-4 (ebk); ISBN 978-1-118-55396-1 (ebk); ISBN 978-1-118-55399-2 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2

## *About the Author*

**John C. Pezzullo, PhD**, is an adjunct associate professor at Georgetown University. He has had a half-century of experience supporting researchers in the physical, biological, and social sciences. For more than 25 years, he led a dual life at Rhode Island Hospital as an information technology programmer/analyst (and later director) while also providing statistical and other technical support to biological and clinical researchers at the hospital. He then joined the faculty at Georgetown University as informatics director of the National Institute of Child Health and Human Development's Perinatology Research Branch. He has held faculty appointments in the departments of obstetrics and gynecology, biomathematics and biostatistics, pharmacology, nursing, and internal medicine. He is now semi-retired and living in Florida, but he still teaches biostatistics and clinical trial design to Georgetown students over the Internet. He created the `StatPages.info` website, which provides online statistical calculating capability and other statistics-related resources.



## *Dedication*

To my wife, Betty: Without your steadfast support and encouragement, I would never have been able to complete this book. To Mom and Dad, who made it all possible. And to our kids, our grandkids, and our great-grandkids!

## *Author's Acknowledgments*

My heartfelt thanks to Matt Wagner of Fresh Books, Inc., and to Lindsay Lefevere for the opportunity to write this book; to Tonya Cupp, my special editor, who tutored me in the “Wiley ways” during the first quarter of the chapter-writing phase of the project; to Georgette Beatty, my project editor, who kept me on the path and on target (and mostly on time) throughout the process; to Christy Pingleton, the copy editor, for making sure what I said was intelligible; and to William Miller and Donatello Telesca, the technical reviewers, for making sure that what I said was correct.

Special thanks to Darrell Abernethy for his invaluable suggestions in Chapter 6.

And a special word of appreciation to all my family and friends, who provided so much support and encouragement throughout the whole project.

## **Publisher's Acknowledgments**

We're proud of this book; please send us your comments at <http://dummies.custhelp.com>. For other comments, please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002.

Some of the people who helped bring this book to market include the following:

### ***Acquisitions, Editorial, and Vertical Websites***

**Senior Project Editor:** Georgette Beatty

**Executive Editor:** Lindsay Sandman Lefevere

**Copy Editor:** Christine Pingleton

**Assistant Editor:** David Lutton

**Editorial Program Coordinator:** Joe Niesen

**Technical Editors:** Dr. William G. Miller,  
Donatello Telesca

**Editorial Manager:** Michelle Hacker

**Editorial Assistant:** Alexa Koschier

**Cover Photo:** Test tubes © Mike Kemp/  
jupiterimages; Graph courtesy of  
John Pezzullo

### ***Composition Services***

**Project Coordinator:** Sheree Montgomery

**Layout and Graphics:** Carrie A. Cesavice

**Proofreaders:** Debbye Butler, John Greenough

**Indexer:** Ty Koontz

### ***Special Help***

Tonya Cupp, Sarah Faulkner, Todd Lothery,  
Danielle Voirol

---

## **Publishing and Editorial for Consumer Dummies**

**Kathleen Nebenhaus**, Vice President and Executive Publisher

**David Palmer**, Associate Publisher

**Kristin Ferguson-Wagstaffe**, Product Development Director

## **Publishing for Technology Dummies**

**Andy Cummings**, Vice President and Publisher

## **Composition Services**

**Debbie Stailey**, Director of Composition Services



# Contents at a Glance

<b>Introduction .....</b>	<b>1</b>
<b>Part I: Beginning with Biostatistics Basics .....</b>	<b>7</b>
Chapter 1: Biostatistics 101 .....	9
Chapter 2: Overcoming Mathophobia: Reading and Understanding Mathematical Expressions .....	17
Chapter 3: Getting Statistical: A Short Review of Basic Statistics .....	31
Chapter 4: Counting on Statistical Software .....	51
Chapter 5: Conducting Clinical Research .....	61
Chapter 6: Looking at Clinical Trials and Drug Development .....	77
<b>Part II: Getting Down and Dirty with Data .....</b>	<b>91</b>
Chapter 7: Getting Your Data into the Computer .....	93
Chapter 8: Summarizing and Graphing Your Data .....	103
Chapter 9: Aiming for Accuracy and Precision .....	121
Chapter 10: Having Confidence in Your Results .....	133
Chapter 11: Fuzzy In Equals Fuzzy Out: Pushing Imprecision through a Formula .....	143
<b>Part III: Comparing Groups .....</b>	<b>153</b>
Chapter 12: Comparing Average Values between Groups .....	155
Chapter 13: Comparing Proportions and Analyzing Cross-Tabulations .....	173
Chapter 14: Taking a Closer Look at Fourfold Tables .....	189
Chapter 15: Analyzing Incidence and Prevalence Rates in Epidemiologic Data .....	203
Chapter 16: Feeling Noninferior (Or Equivalent) .....	211
<b>Part IV: Looking for Relationships with Correlation and Regression .....</b>	<b>219</b>
Chapter 17: Introducing Correlation and Regression .....	221
Chapter 18: Getting Straight Talk on Straight-Line Regression .....	233
Chapter 19: More of a Good Thing: Multiple Regression .....	251
Chapter 20: A Yes-or-No Proposition: Logistic Regression .....	267
Chapter 21: Other Useful Kinds of Regression .....	291

<b><i>Part V: Analyzing Survival Data</i></b> .....	<b>311</b>
Chapter 22: Summarizing and Graphing Survival Data.....	313
Chapter 23: Comparing Survival Times .....	331
Chapter 24: Survival Regression.....	339
<b><i>Part VI: The Part of Tens</i></b> .....	<b>357</b>
Chapter 25: Ten Distributions Worth Knowing .....	359
Chapter 26: Ten Easy Ways to Estimate How Many Subjects You Need .....	369
<b><i>Index</i></b> .....	<b>375</b>

# Table of Contents

---

## ***Introduction* ..... 1**

About This Book .....	1
Conventions Used in This Book .....	2
What You're Not to Read .....	2
Foolish Assumptions .....	3
How This Book Is Organized .....	3
Part I: Beginning with Biostatistics Basics .....	3
Part II: Getting Down and Dirty with Data .....	4
Part III: Comparing Groups .....	4
Part IV: Looking for Relationships with Correlation and Regression .....	4
Part V: Analyzing Survival Data .....	5
Part VI: The Part of Tens .....	5
Icons Used in This Book .....	5
Where to Go from Here .....	6

## ***Part I: Beginning with Biostatistics Basics*..... 7**

### **Chapter 1: Biostatistics 101 ..... 9**

Brushing Up on Math and Stats Basics .....	9
Doing Calculations with the Greatest of Ease .....	10
Concentrating on Clinical Research .....	11
Drawing Conclusions from Your Data .....	12
Statistical estimation theory .....	12
Statistical decision theory .....	12
A Matter of Life and Death: Working with Survival Data .....	14
Figuring Out How Many Subjects You Need .....	15
Getting to Know Statistical Distributions .....	16

### **Chapter 2: Overcoming Mathophobia: Reading and Understanding Mathematical Expressions..... 17**

Breaking Down the Basics of Mathematical Formulas .....	18
Displaying formulas in different ways .....	18
Checking out the building blocks of formulas .....	18
Focusing on Operations Found in Formulas .....	20
Basic mathematical operations .....	20
Powers, roots, and logarithms .....	22
Factorials and absolute values .....	24



Functions .....	25
Simple and complicated formulas .....	25
Equations .....	26
Counting on Collections of Numbers .....	26
One-dimensional arrays .....	26
Higher-dimensional arrays .....	27
Arrays in formulas .....	28
Sums and products of the elements of an array .....	28

### **Chapter 3: Getting Statistical: A Short Review of Basic Statistics. . . 31**

Taking a Chance on Probability.....	31
Thinking of probability as a number .....	32
Following a few basic rules.....	32
Comparing odds versus probability.....	33
Some Random Thoughts about Randomness .....	34
Picking Samples from Populations .....	35
Recognizing that sampling isn't perfect.....	36
Digging into probability distributions.....	37
Introducing Statistical Inference .....	38
Statistical estimation theory .....	38
Statistical decision theory .....	40
Homing In on Hypothesis Testing .....	41
Getting the language down .....	41
Testing for significance.....	42
Understanding the meaning of "p value" as the result of a test ....	43
Examining Type I and Type II errors .....	43
Grasping the power of a test .....	45
Going Outside the Norm with Nonparametric Statistics.....	48

### **Chapter 4: Counting on Statistical Software. . . . . 51**

Desk Job: Personal Computer Software.....	51
Checking out commercial software .....	52
Focusing on free software.....	55
On the Go: Calculators and Mobile Devices.....	57
Scientific and programmable calculators .....	57
Mobile devices .....	58
Gone Surfin': Web-Based Software .....	58
On Paper: Printed Calculators .....	59

### **Chapter 5: Conducting Clinical Research . . . . . 61**

Designing a Clinical Study .....	61
Identifying aims, objectives, hypotheses, and variables .....	61
Deciding who will be in the study.....	63
Choosing the structure of the study .....	64
Using randomization .....	64
Selecting the analyses to use .....	66

Defining analytical populations.....	67
Determining how many subjects to enroll.....	67
Putting together the protocol .....	68
Carrying Out a Clinical Study.....	70
Protecting your subjects.....	70
Collecting and validating data.....	72
Analyzing Your Data.....	73
Dealing with missing data .....	74
Handling multiplicity .....	74
Incorporating interim analyses .....	76

## **Chapter 6: Looking at Clinical Trials and Drug Development . . . . . 77**

Not Ready for Human Consumption: Doing Preclinical Studies.....	78
Testing on People during Clinical Trials to Check a Drug's	
Safety and Efficacy .....	80
Phase I: Determining the maximum tolerated dose .....	80
Phase II: Finding out about the drug's performance .....	82
Phase III: Proving that the drug works .....	84
Phase IV: Keeping an eye on the marketed drug .....	85
Holding Other Kinds of Clinical Trials .....	86
Pharmacokinetics and pharmacodynamics (PK/PD studies) .....	86
Bioequivalence studies .....	88
Thorough QT studies .....	88

## ***Part II: Getting Down and Dirty with Data..... 91***

### **Chapter 7: Getting Your Data into the Computer . . . . . 93**

Looking at Levels of Measurement.....	94
Classifying and Recording Different Kinds of Data .....	95
Dealing with free-text data.....	95
Assigning subject identification (ID) numbers.....	95
Organizing name and address data .....	96
Collecting categorical data .....	96
Recording numerical data.....	98
Entering date and time data .....	99
Checking Your Entered Data for Errors.....	101
Creating a File that Describes Your Data File .....	102

### **Chapter 8: Summarizing and Graphing Your Data . . . . . 103**

Summarizing and Graphing Categorical Data .....	104
Summarizing Numerical Data.....	106
Locating the center of your data .....	107
Describing the spread of your data.....	110
Showing the symmetry and shape of the distribution.....	113
Structuring Numerical Summaries into Descriptive Tables.....	114

Graphing Numerical Data .....	115
Showing the distribution with histograms .....	116
Summarizing grouped data with bars, boxes, and whiskers.....	118
Depicting the relationships between numerical variables with other graphs.....	120

## **Chapter 9: Aiming for Accuracy and Precision .....121**

Beginning with the Basics of Accuracy and Precision.....	121
Getting to know sample statistics and population parameters...	121
Understanding accuracy and precision in terms of the sampling distribution .....	122
Thinking of measurement as a kind of sampling .....	123
Expressing errors in terms of accuracy and precision.....	124
Improving Accuracy and Precision .....	126
Enhancing sampling accuracy.....	126
Getting more accurate measurements .....	126
Improving sampling precision.....	127
Increasing the precision of your measurements .....	128
Calculating Standard Errors for Different Sample Statistics .....	129
A mean.....	129
A proportion .....	130
Event counts and rates .....	130
A regression coefficient .....	131

## **Chapter 10: Having Confidence in Your Results .....133**

Feeling Confident about Confidence Interval Basics.....	133
Defining confidence intervals .....	134
Looking at confidence levels .....	134
Taking sides with confidence intervals.....	135
Calculating Confidence Intervals .....	136
Before you begin: Formulas for confidence limits in large samples.....	136
The confidence interval around a mean .....	137
The confidence interval around a proportion .....	139
The confidence interval around an event count or rate.....	140
The confidence interval around a regression coefficient .....	141
Relating Confidence Intervals and Significance Testing.....	141

## **Chapter 11: Fuzzy In Equals Fuzzy Out: Pushing Imprecision through a Formula .....143**

Understanding the Concept of Error Propagation .....	144
Using Simple Error Propagation Formulas for Simple Expressions.....	146
Adding or subtracting a constant doesn't change the SE.....	146
Multiplying (or dividing) by a constant multiplies (or divides) the SE by the same amount.....	147
For sums and differences: Add the squares of SEs together.....	147

For averages: The square root law takes over .....	148
For products and ratios: Squares of relative SEs are added together .....	148
For powers and roots: Multiply the relative SE by the power .....	149
Handling More Complicated Expressions .....	149
Using the simple rules consecutively .....	150
Checking out an online calculator .....	150
Simulating error propagation — easy, accurate, and versatile ...	152

## ***Part III: Comparing Groups* ..... 153**

### **Chapter 12: Comparing Average Values between Groups . . . . . 155**

Knowing That Different Situations Need Different Tests .....	155
Comparing the mean of a group of numbers to a hypothesized value .....	156
Comparing two groups of numbers .....	156
Comparing three or more groups of numbers .....	157
Analyzing data grouped on several different variables .....	158
Adjusting for a “nuisance variable” when comparing numbers .....	158
Comparing sets of matched numbers .....	159
Comparing within-group changes between groups .....	160
Trying the Tests Used for Comparing Averages .....	161
Surveying Student t tests .....	161
Assessing the ANOVA .....	164
Running Student t tests and ANOVAs from summary data .....	168
Running nonparametric tests .....	169
Estimating the Sample Size You Need for Comparing Averages .....	169
Simple formulas .....	169
Software and web pages .....	170
A sample-size nomogram .....	170

### **Chapter 13: Comparing Proportions and Analyzing Cross-Tabulations . . . . . 173**

Examining Two Variables with the Pearson Chi-Square Test .....	174
Understanding how the chi-square test works .....	175
Pointing out the pros and cons of the chi-square test .....	180
Modifying the chi-square test: The Yates continuity correction ...	181
Focusing on the Fisher Exact Test .....	181
Understanding how the Fisher Exact test works .....	181
Noting the pros and cons of the Fisher Exact test .....	182
Calculating Power and Sample Size for Chi-Square and Fisher Exact Tests .....	183
Analyzing Ordinal Categorical Data with the Kendall Test .....	185
Studying Stratified Data with the Mantel-Haenszel Chi-Square Test ....	187

**Chapter 14: Taking a Closer Look at Fourfold Tables . . . . .189**

Focusing on the Fundamentals of Fourfold Tables .....	190
Choosing the Right Sampling Strategy .....	191
Producing Fourfold Tables in a Variety of Situations .....	192
Describing the association between two binary variables .....	193
Assessing risk factors .....	194
Evaluating diagnostic procedures .....	197
Investigating treatments .....	199
Looking at inter- and intra-rater reliability .....	201

**Chapter 15: Analyzing Incidence and Prevalence Rates  
in Epidemiologic Data . . . . .203**

Understanding Incidence and Prevalence .....	203
Prevalence: The fraction of a population with a particular condition .....	204
Incidence: Counting new cases .....	204
Understanding how incidence and prevalence are related .....	205
Analyzing Incidence Rates .....	205
Expressing the precision of an incidence rate .....	205
Comparing incidences with the rate ratio .....	206
Calculating confidence intervals for a rate ratio .....	207
Comparing two event rates .....	207
Comparing two event counts with identical exposure .....	209
Estimating the Required Sample Size .....	209

**Chapter 16: Feeling Noninferior (Or Equivalent). . . . .211**

Understanding the Absence of an Effect .....	212
Defining the effect size: How different are the groups? .....	212
Defining an important effect size: How close is close enough? ...	213
Recognizing effects: Can you spot a difference if there really is one? .....	213
Proving Equivalence and Noninferiority .....	214
Using significance tests .....	214
Using confidence intervals .....	215
Some precautions about noninferiority testing .....	217

***Part IV: Looking for Relationships  
with Correlation and Regression . . . . . 219*****Chapter 17: Introducing Correlation and Regression . . . . .221**

Correlation: How Strongly Are Two Variables Associated? .....	222
Lining up the Pearson correlation coefficient .....	222
Analyzing correlation coefficients .....	223



Regression: What Equation Connects the Variables? .....	227
Understanding the purpose of regression analysis .....	227
Talking about terminology and mathematical notation .....	228
Classifying different kinds of regression .....	229

## **Chapter 18: Getting Straight Talk on Straight-Line Regression . . . 233**

Knowing When to Use Straight-Line Regression .....	234
Understanding the Basics of Straight-Line Regression .....	235
Running a Straight-Line Regression .....	236
Taking a few basic steps .....	237
Walking through an example .....	237
Interpreting the Output of Straight-Line Regression .....	239
Seeing what you told the program to do .....	240
Looking at residuals .....	241
Making your way through the regression table .....	243
Wrapping up with measures of goodness-of-fit .....	247
Scientific fortune-telling with the prediction formula .....	248
Recognizing What Can Go Wrong with Straight-Line Regression .....	249
Figuring Out the Sample Size You Need .....	249

## **Chapter 19: More of a Good Thing: Multiple Regression . . . . . 251**

Understanding the Basics of Multiple Regression .....	251
Defining a few important terms .....	252
Knowing when to use multiple regression .....	253
Being aware of how the calculations work .....	253
Running Multiple Regression Software .....	254
Preparing categorical variables .....	254
Recoding categorical variables as numerical .....	255
Creating scatter plots before you jump into your multiple regression .....	256
Taking a few steps with your software .....	258
Interpreting the Output of a Multiple Regression .....	258
Examining typical output from most programs .....	259
Checking out optional output available from some programs ....	260
Deciding whether your data is suitable for regression analysis ...	261
Determining how well the model fits the data .....	262
Watching Out for Special Situations that Arise in Multiple Regression .....	263
Synergy and anti-synergy .....	263
Collinearity and the mystery of the disappearing significance ...	263
Figuring How Many Subjects You Need .....	265

**Chapter 20: A Yes-or-No Proposition: Logistic Regression . . . . . 267**

Using Logistic Regression.....	267
Understanding the Basics of Logistic Regression .....	268
Gathering and graphing your data.....	268
Fitting a function with an S shape to your data .....	270
Handling multiple predictors in your logistic model .....	274
Running a Logistic Regression with Software.....	274
Interpreting the Output of Logistic Regression .....	275
Seeing summary information about the variables.....	276
Assessing the adequacy of the model.....	276
Checking out the table of regression coefficients .....	278
Predicting probabilities with the fitted logistic formula .....	278
Making yes or no predictions.....	280
Heads Up: Knowing What Can Go Wrong with Logistic Regression .....	285
Don't fit a logistic function to nonlogistic data.....	285
Watch out for collinearity and disappearing significance.....	285
Check for inadvertent reverse-coding of the outcome variable ...	286
Don't misinterpret odds ratios for numerical predictors.....	286
Don't misinterpret odds ratios for categorical predictors.....	286
Beware the complete separation problem .....	287
Figuring Out the Sample Size You Need for Logistic Regression .....	288

**Chapter 21: Other Useful Kinds of Regression . . . . . 291**

Analyzing Counts and Rates with Poisson Regression .....	291
Introducing the generalized linear model.....	292
Running a Poisson regression .....	293
Interpreting the Poisson regression output .....	295
Discovering other things that Poisson regression can do .....	296
Anything Goes with Nonlinear Regression.....	298
Distinguishing nonlinear regression from other kinds .....	299
Checking out an example from drug research .....	300
Running a nonlinear regression .....	302
Interpreting the output .....	304
Using equivalent functions to fit the parameters you really want .....	305
Smoothing Nonparametric Data with LOWESS.....	306
Running LOWESS .....	307
Adjusting the amount of smoothing.....	309

***Part V: Analyzing Survival Data ..... 311*****Chapter 22: Summarizing and Graphing Survival Data .....313**

Understanding the Basics of Survival Data .....	314
Knowing that survival times are intervals.....	314
Recognizing that survival times aren't normally distributed .....	314
Considering censoring .....	315
Looking at the Life-Table Method.....	318
Making a life table .....	319
Interpreting a life table.....	323
Graphing hazard rates and survival probabilities from a life table.....	324
Digging Deeper with the Kaplan-Meier Method.....	324
Heeding a Few Guidelines for Life Tables and the Kaplan-Meier Method .....	326
Recording survival times the right way .....	327
Recording censoring information correctly .....	327
Interpreting those strange-looking survival curves .....	328
Doing Even More with Survival Data.....	329

**Chapter 23: Comparing Survival Times ..... 331**

Comparing Survival between Two Groups with the Log-Rank Test.....	332
Understanding what the log-rank test is doing .....	333
Running the log-rank test on software .....	333
Looking at the calculations.....	334
Assessing the assumptions .....	336
Considering More Complicated Comparisons.....	337
Coming Up with the Sample Size Needed for Survival Comparisons....	337

**Chapter 24: Survival Regression .....339**

Knowing When to Use Survival Regression.....	339
Explaining the Concepts behind Survival Regression .....	340
The steps of Cox PH regression .....	341
Hazard ratios .....	345
Running a Survival Regression .....	346
Interpreting the Output of a Survival Regression.....	347
Testing the validity of the assumptions.....	349
Checking out the table of regression coefficients .....	350
Homing in on hazard ratios and their confidence intervals.....	350
Assessing goodness-of-fit and predictive ability of the model ....	351
Focusing on baseline survival and hazard functions .....	352

How Long Have I Got, Doc? Constructing Prognosis Curves ..... 353

    Running the proportional-hazards regression..... 353

    Finding h ..... 354

Estimating the Required Sample Size for a Survival Regression ..... 356

**Part VI: The Part of Tens..... 357**

**Chapter 25: Ten Distributions Worth Knowing . . . . .359**

The Uniform Distribution ..... 360

The Normal Distribution..... 360

The Log-Normal Distribution ..... 361

The Binomial Distribution ..... 362

The Poisson Distribution..... 362

The Exponential Distribution..... 363

The Weibull Distribution ..... 364

The Student t Distribution..... 364

The Chi-Square Distribution..... 366

The Fisher F Distribution..... 367

**Chapter 26: Ten Easy Ways to Estimate How Many  
Subjects You Need. . . . .369**

Comparing Means between Two Groups ..... 370

Comparing Means among Three, Four, or Five Groups..... 370

Comparing Paired Values ..... 370

Comparing Proportions between Two Groups..... 371

Testing for a Significant Correlation ..... 371

Comparing Survival between Two Groups..... 371

Scaling from 80 Percent to Some Other Power..... 372

Scaling from 0.05 to Some Other Alpha Level ..... 373

Making Adjustments for Unequal Group Sizes ..... 373

Allowing for Attrition ..... 374

**Index ..... 375**

# Introduction

---

**B**iostatistics is the practical application of statistical concepts and techniques to topics in biology. Because biology is such a broad field — studying all forms of life from viruses to trees to fleas to mice to people — biostatistics covers a very wide area, including designing biological experiments, safely conducting research on human beings, collecting and verifying data from those studies, summarizing and displaying that data, and analyzing the data to draw meaningful conclusions from it.

No book of reasonable size can hope to span all the subspecialties of biostatistics, including molecular biology, genetics, agricultural studies, animal research (in the lab and in the wild), clinical trials on humans, and epidemiological research. So I've concentrated on the most widely applicable topics and on the topics that are most relevant to research on humans (that is, *clinical* research). I chose these topics on the basis of a survey of graduate-level biostatistics curricula from major universities. I hope it covers most of the topics you're most interested in; but if it doesn't, please tell me what you wish I had included. You can e-mail me at [jcp12345@gmail.com](mailto:jcp12345@gmail.com), and I'll try to respond to your message.

## About This Book

I wrote this book as a reference — something you go to when you want information about a particular topic. So you don't have to read it from beginning to end; you can jump directly to the part you're interested in. In fact, I hope you'll be inclined to pick it up from time to time, open it to a page at random, read a page or two, and get a little something useful from it.

This book generally doesn't show you the detailed steps to perform every statistical calculation by hand. That may have been necessary in the mid-1900s, when statistics students spent hours in a “computing lab” (that is, a room that had an adding machine in it) calculating a correlation coefficient, but nowadays computers do all the computing. (See Chapter 4 for advice on choosing statistical software.) When describing statistical tests, my focus is always on the concepts behind the method, how to prepare your data for analysis, and how to interpret the results. I keep mathematical formulas and derivations to a minimum in this book; I include them only when they help explain what's going on. If you really want to see them, you can find them in many biostatistics textbooks, and they're readily available online.

Because good experimental design is crucial for the success of any research, this book gives special attention to the design of clinical trials and, specifically, to calculating the number of subjects you need to study. You find easy-to-apply examples of sample-size calculations in the chapters describing significance tests in Parts III, IV, and V and in Chapter 26.

## Conventions Used in This Book

Here are some typographic conventions I use throughout this book:

- ✓ When I introduce a new term, I put the term in *italics* and define it. I also use italics occasionally to emphasize important information.
- ✓ In bulleted lists, I often place the most important word or phrase of each bulleted item in **boldface** text. The action parts of numbered steps are also boldface.
- ✓ I show web links (URLs) as monotype text.
- ✓ When this book was printed, some web addresses may have needed to break across two lines of text. If that happened, rest assured that I haven't put in any extra characters (like hyphens) to indicate the break. So, when using one of these web addresses, just type in exactly what you see in this book, pretending as though the line break doesn't exist.
- ✓ Whenever you see the abbreviation *sd* or *SD*, it always refers to the *standard deviation*.
- ✓ Anytime you see the word *significant* in reference to a p value, it means  $p \leq 0.05$ .
- ✓ When you see the lowercase italicized letter *e* in a formula, it refers to the mathematical constant 2.718..., which I describe in Chapter 2. (On the very rare occasions that it stands for something else, I say so.)
- ✓ I alternate between using male and female pronouns (instead of saying "he or she," "him or her," and so on) throughout the book. No gender preference is intended.

## What You're Not to Read

Although I try to keep technical (that is, mathematical) details to a minimum, I do include them occasionally. The more complicated ones are marked by a Technical Stuff icon. You can skip over these paragraphs, and it won't prevent you from understanding the rest of the material. You can also skip over anything that's in a sidebar (text that resides in a box). Sidebars contain non-essential but interesting stuff, like historical trivia and other "asides."

## *Foolish Assumptions*

I wrote this book to help several kinds of people, and I assume you fall into one of the following categories:

- ✓ Students at the undergraduate or graduate level who are taking a course in biostatistics and want help with the topics they're studying in class
- ✓ People who have had no formal biostatistical training (perhaps no statistical training at all) but find themselves having to deal with data from biological or clinical studies as part of their job
- ✓ Doctors, nurses, and other healthcare professionals who want to carry out clinical research

If you're interested in biostatistics, then you're no dummy. But I bet you sometimes *feel* like a dummy when it comes to biostatistics, or statistics in general, or mathematics. Don't feel bad — I've felt that way many times over the years, and still feel like that whenever I'm propelled into an area of biostatistics I haven't encountered before. (If you haven't taken a basic statistics course yet, you may want to get *Statistics For Dummies* by Deborah J. Rumsey, PhD — published by Wiley — and read parts of that book first.)

The important thing to keep in mind is that you don't have to be a math genius to be a good biological or clinical scientist — one who can intelligently design experiments, execute them well, collect and analyze data properly, and draw valid conclusions. You just have to have a good grasp of the basic concepts and know how to utilize the sophisticated statistical software that has become so widely available.

## *How This Book Is Organized*

I've divided this book into six parts, and each part contains several chapters. The following sections describe what you find in each part.

### *Part I: Beginning with Biostatistics Basics*

This part can be thought of as providing preparation and context for the remainder of this book. Here, I bring you up to speed on math and statistics concepts so that you're comfortable with them throughout this book. Then I provide advice on selecting statistical software. And finally I describe one major setting in which biostatistics is utilized — clinical research.

## ***Part II: Getting Down and Dirty with Data***

This part focuses on the raw material that biostatistical analysis works with — *data*. You probably already know the two main types of data: numerical (or quantitative) data, such as ages and heights, and non-numerical data, such as names and genders. Part II gets into the more subtle (but very important) distinctions between different data types.

You discover how to collect data properly, how to summarize it concisely and display it as tables and graphs, and how to describe the quality of the data (its precision and the uncertainties associated with your measured values). And you find out how the precision of your raw data affects the precision of other things you calculate from that data.

## ***Part III: Comparing Groups***

This part describes some of the most common statistical analyses you carry out on your data — comparing variables between groups. You discover how to answer questions like these: Does an arthritis medication reduce joint pain more than a placebo? Does a history of diabetes in a parent predict the likelihood of diabetes in the child? And if so, by how much?

You also find out how to show that there's *no meaningful difference* between two groups. Is a generic drug really equivalent to the name brand? Does a new drug *not* interfere with normal heart rhythm? This endeavor entails more than just not proving that there *is* a difference — absence of proof is not proof of absence, and there are special ways to prove that there's no important difference in your data.

Throughout this part, I discuss common statistical techniques for comparing groups such as t tests, ANOVAs, chi-square tests, and the Fisher Exact test.

## ***Part IV: Looking for Relationships with Correlation and Regression***

This part takes you through the very broad field of regression analysis — studying the relationships that can exist between variables. You find out how to test for a significant association between two or more variables and how to express that relationship in terms of a formula or equation that predicts the likely value of one variable from the observed values of one or more other variables. You see how useful such an equation can be, both for understanding the underlying science and for doing all kinds of practical things based on that relationship.



After reviewing the simple straight-line and multiple linear regression techniques you probably encountered in a basic stats course, you discover how to handle the more advanced problems that occur in the real world of biological research — *logistic regression* for analyzing yes-or-no kinds of outcomes, like “had a miscarriage”; *Poisson regression* for analyzing the frequency of recurring events, such as the number of hospitalizations for emphysema patients; and *nonlinear regression* when the relationship between the variables can take on a complicated mathematical form.

## Part V: Analyzing Survival Data

This part is devoted to the analysis of one very special and important kind of data in biological research — *survival time* (or, more generally, the time to the first occurrence of some particular kind of event). You see what makes this type of data so special and why special methods are needed to deal with it correctly. You see how to calculate survival curves, test for a significant difference in survival between two or more groups of subjects, and apply the powerful and general methods of regression analysis to survival data.

## Part VI: The Part of Tens

The final two chapters of this book provide “top-ten lists” of handy information and rules that you’ll probably refer to often. Chapter 25 describes ten of the most common statistical distribution functions that you encounter in biostatistical research. Some of these distributions describe how your observed data values are likely to fluctuate, and some are used primarily in conjunction with the common significance tests (t-tests, chi-square tests, and ANOVAs). Chapter 26 contains a set of handy rules of thumb you can use to get quick estimates of the number of subjects you need to study in order to have a good chance of obtaining significant results.

## Icons Used in This Book



Icons (the little drawings in the margins of this book) are used to draw your attention to certain kinds of material. Here’s what they mean:

This icon signals something that’s really worth keeping in mind. If you take away anything from this book, it should be the material marked with this icon.



I use this icon to flag things like derivations and computational formulas that you don’t have to know or understand but that may give you a deeper insight into other material. Feel free to skip over any information with this icon.



This icon refers to helpful hints, ideas, shortcuts, and rules of thumb that you can use to save time or make a task easier. It also highlights different ways of thinking about some topic or concept.



This icon alerts you to a topic that can be tricky or a concept that people often misunderstand.

## *Where to Go from Here*

You're already off to a good start — you've read this introduction, so you have a good idea of what this book is all about (at least what the major parts of the book are all about). For an even better idea of what's in it, take a look at the Contents at a Glance — this drills down into each part, and shows you what each chapter is all about. Finally, skim through the full-blown table of contents, which drills further down into each chapter, showing you the sections and subsections of that chapter.

If you want to get the big picture of what biostatistics encompasses (at least those parts of biostatistics covered in this book), then read Chapter 1. This is a top-level overview of the basic concepts that make up this entire book. Here are a few other special places you may want to jump into:

- ✓ If you're uncomfortable with mathematical notation, then Chapter 2 is the place to start.
- ✓ If you want a quick refresher on basic statistics (the kind of stuff that would be taught in a Stats 101 course), then read Chapter 3.
- ✓ You can get an introduction to clinical research in Chapters 5 and 6.
- ✓ If you want to know about collecting, summarizing, and graphing data, jump to Part II.
- ✓ If you need to know about working with survival data, you can go right to Part V.
- ✓ If you're puzzled about some particular statistical distribution function, then look at Chapter 25.
- ✓ And if you need to do some quick sample-size estimates, turn to Chapter 26.

Part I

# Beginning with Biostatistics Basics



Visit [www.dummies.com](http://www.dummies.com) for great (and free!) Dummies content online.

## *In this part . . .*

- ✓ Get comfortable with mathematical notation that uses numbers, special constants, variables, and mathematical symbols — a must for all you mathophobes.
- ✓ Review basic statistical concepts — such as probability, randomness, populations, samples, statistical inference, and more — to get ready for the study of biostatistics.
- ✓ Choose and acquire statistical software (both commercial and free), and discover other ways to do statistical calculations, such as calculators, mobile devices, and web-based programs.
- ✓ Understand clinical research — how biostatistics influences the design and execution of clinical trials and how treatments are developed and approved.

# Chapter 1

## Biostatistics 101

---

### *In This Chapter*

- ▶ Getting up to speed on the prerequisites for biostatistics
  - ▶ Understanding the clinical research environment
  - ▶ Surveying the special procedures used to analyze biological data
  - ▶ Estimating how many subjects you need
  - ▶ Working with distributions
- 

**B**iostatistics deals with the design and execution of scientific experiments on living creatures, the acquisition and analysis of data from those experiments, and the interpretation and presentation of the results of those analyses.

This book is meant to be a useful and easy-to-understand companion to the more formal textbooks used in graduate-level biostatistics courses. Because most of these courses concentrate on the more clinical areas of biostatistics, this book focuses on that area as well. In this chapter, I introduce you to the fundamentals of biostatistics.

### *Brushing Up on Math and Stats Basics*

Chapters 2 and 3 are designed to bring you up to speed on the basic math and statistical background that's needed to understand biostatistics and to give you some supplementary information (or "context") that you may find generally useful while you're reading the rest of this book.

- ✓ Many people feel unsure of themselves when it comes to understanding mathematical formulas and equations. Although this book contains fewer formulas than many other statistics books do, I do use them when they help illustrate a concept or describe a calculation that's simple enough to do by hand. But if you're a real mathphobe, you probably dread looking at *any* chapter that has a math expression anywhere in it. That's why

I include Chapter 2 — to show you how to read and understand the basic mathematical notation that I use in this book. I cover everything from basic mathematical operations to functions and beyond.

- ✓ If you're in a graduate-level biostatistics course, you've probably already taken one or two introductory statistics courses. But that may have been a while ago, and you may not feel too sure of your knowledge of the basic statistical concepts. Or you may have little or no formal statistical training, but now find yourself in a work situation where you interact with clinical researchers, participate in the design of research projects, or work with the results from biological research. If so, then you definitely want to read Chapter 3, which provides an overview of the fundamental concepts and terminology of statistics. There, you get the scoop on topics such as probability, randomness, populations, samples, statistical inference, accuracy, precision, hypothesis testing, nonparametric statistics, and simulation techniques.

## *Doing Calculations with the Greatest of Ease*

This book generally doesn't have step-by-step instructions for performing statistical tests and analyses by hand. That's because in the 21st century you shouldn't be doing those calculations by hand; there are lots of ways to get a computer to do them for you. So this book describes calculations only to illustrate the concepts that are involved in the procedure, or when the calculations are simple enough that it's feasible to do them by hand (or even in your head!).

Unlike some statistics books that assume that you're using a specific software package (like SPSS, SAS, Minitab, and so on), this book makes no such assumption. You may be a student at a school that provides a commercial package at an attractive price or requires that you use a specific product (regardless of the price). Or you may be on your own, with limited financial resources, and the big programs may be out of your reach. Fortunately, you have several options. You can download some excellent free programs from the Internet. And you can also find a lot of web pages that perform specific statistical tests and procedures; collectively they can be thought of as the equivalence of a free online statistical software package. Chapter 4 describes some of these options — commercial products, free programs, web-based calculators, and others.