

LEARNING MADE EASY



2nd Edition

# Biostatistics

for  
**dummies**<sup>®</sup>  
A Wiley Brand



Learn how to  
analyze data properly

Know the important steps in  
conducting clinical research

Use software to analyze  
large data sets

**Monika Wahi, MPH**  
**John C. Pezzullo, PhD**





# Biostatistics

2nd Edition

**by Monika Wahj, MPH and  
John C. Pezzullo, PhD**

for  
**dummies**<sup>®</sup>  
A Wiley Brand

## Biostatistics For Dummies®, 2nd Edition

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

Copyright © 2024 by John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial technologies or similar technologies.

Media and software compilation copyright © 2024 by John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial technologies or similar technologies.

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, For Dummies, the Dummies Man logo, [Dummies.com](http://Dummies.com), Making Everything Easier, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. 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 <https://hub.wiley.com/community/support/dummies>.

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: 202493911

ISBN 978-1-394-25146-9 (pbk); ISBN 978-1-394-25147-6 (ebk); ISBN 978-1-394-25148-3 (ebk)

# Contents at a Glance

<b>Introduction</b> .....	1
<b>Part 1: Getting Started with Biostatistics</b> .....	5
CHAPTER 1: Biostatistics 101 .....	7
CHAPTER 2: Overcoming Mathophobia: Reading and Understanding Mathematical Expressions .....	15
CHAPTER 3: Getting Statistical: A Short Review of Basic Statistics .....	29
<b>Part 2: Examining Tools and Processes</b> .....	51
CHAPTER 4: Counting on Statistical Software .....	53
CHAPTER 5: Conducting Clinical Research .....	61
CHAPTER 6: Taking All Kinds of Samples .....	77
CHAPTER 7: Having Designs on Study Design .....	87
<b>Part 3: Getting Down and Dirty with Data</b> .....	99
CHAPTER 8: Getting Your Data into the Computer .....	101
CHAPTER 9: Summarizing and Graphing Your Data .....	111
CHAPTER 10: Having Confidence in Your Results .....	129
<b>Part 4: Comparing Groups</b> .....	139
CHAPTER 11: Comparing Average Values between Groups .....	141
CHAPTER 12: Comparing Proportions and Analyzing Cross-Tabulations .....	159
CHAPTER 13: Taking a Closer Look at Fourfold Tables .....	173
CHAPTER 14: Analyzing Incidence and Prevalence Rates in Epidemiologic Data. . .	191
<b>Part 5: Looking for Relationships with Correlation and Regression</b> .....	199
CHAPTER 15: Introducing Correlation and Regression .....	201
CHAPTER 16: Getting Straight Talk on Straight-Line Regression .....	213
CHAPTER 17: More of a Good Thing: Multiple Regression .....	233
CHAPTER 18: A Yes-or-No Proposition: Logistic Regression .....	249
CHAPTER 19: Other Useful Kinds of Regression .....	271
CHAPTER 20: Getting the Hint from Epidemiologic Inference .....	291

<b>Part 6: Analyzing Survival Data</b> .....	299
CHAPTER 21: Summarizing and Graphing Survival Data .....	301
CHAPTER 22: Comparing Survival Times .....	317
CHAPTER 23: Survival Regression .....	327
<b>Part 7: The Part of Tens</b> .....	349
CHAPTER 24: Ten Distributions Worth Knowing .....	351
CHAPTER 25: Ten Easy Ways to Estimate How Many Participants You Need .....	361
<b>Index</b> .....	369

# Table of Contents

<b>INTRODUCTION</b> .....	1
About This Book .....	1
Foolish Assumptions .....	2
Icons Used in This Book .....	3
Beyond the Book .....	3
Where to Go from Here .....	3
<b>PART 1: GETTING STARTED WITH BIOSTATISTICS</b> .....	5
<b>CHAPTER 1: Biostatistics 101</b> .....	7
Brushing Up on Math and Stats Basics .....	7
Doing Calculations with the Greatest of Ease .....	8
Concentrating on Epidemiologic Research .....	9
Drawing Conclusions from Your Data .....	10
Statistical estimation theory .....	10
Statistical decision theory .....	10
A Matter of Life and Death: Working with Survival Data .....	13
Getting to Know Statistical Distributions .....	13
Figuring Out How Many Participants You Need .....	14
<b>CHAPTER 2: Overcoming Mathophobia: Reading and Understanding Mathematical Expressions</b> .....	15
Breaking Down the Basics of Mathematical Formulas .....	16
Displaying formulas in different ways .....	16
Checking out the building blocks of formulas .....	17
Focusing on Operations Found in Formulas .....	18
Basic mathematical operations .....	18
Powers, roots, and logarithms .....	20
Factorials and absolute values .....	22
Functions .....	23
Simple and complicated formulas .....	24
Equations .....	24
Counting on Collections of Numbers .....	25
One-dimensional arrays .....	25
Higher-dimensional arrays .....	26
Arrays in formulas .....	26
Sums and products of the elements of an array .....	27

<b>CHAPTER 3: Getting Statistical: A Short Review of Basic Statistics</b> .....	29
Taking a Chance on Probability .....	30
Thinking of probability as a number .....	30
Following a few basic rules of probabilities .....	30
Comparing odds versus probability .....	32
Some Random Thoughts about Randomness .....	33
Selecting Samples from Populations .....	33
Recognizing that sampling isn't perfect .....	34
Digging into probability distributions .....	35
Introducing Statistical Inference .....	37
Statistical estimation theory .....	37
Statistical decision theory .....	39
Honing In on Hypothesis Testing .....	40
Getting the language down .....	40
Testing for significance .....	41
Understanding the meaning of "p value" as the result of a test ..	42
Examining Type I and Type II errors .....	42
Grasping the power of a test .....	44
Going Outside the Norm with Nonparametric Statistics .....	47
 <b>PART 2: EXAMINING TOOLS AND PROCESSES</b> .....	 51
 <b>CHAPTER 4: Counting on Statistical Software</b> .....	 53
Considering the Evolution of Statistical Software .....	54
Comparing Commercial to Open-Source Software .....	54
Checking Out Commercial Software .....	55
SAS .....	55
SPSS .....	56
Microsoft Excel .....	57
Online analytics platforms .....	57
Focusing on Open-Source and Free Software .....	58
Open-source software .....	58
Other free statistical software .....	59
Choosing Between Code-based and Non-Code-Based Methods ...	60
Storing Data in the Cloud .....	60
 <b>CHAPTER 5: Conducting Clinical Research</b> .....	 61
Designing a Clinical Trial .....	61
Identifying aims, objectives, hypotheses, and variables .....	62
Deciding who is eligible for the study .....	64
Choosing the structure of a clinical trial .....	65
Using randomization .....	65
Selecting the analyses to use .....	67



	Determining how many participants to enroll in a clinical trial . . .	68
	Assembling the study protocol . . . . .	68
	Carrying Out a Clinical Trial . . . . .	71
	Protecting clinical trial participants. . . . .	71
	Collecting and validating data . . . . .	74
	Analyzing Your Data . . . . .	74
	Dealing with missing data . . . . .	74
	Handling multiplicity . . . . .	75
<b>CHAPTER 6:</b>	<b>Taking All Kinds of Samples. . . . .</b>	<b>77</b>
	Making Forgivable (and Non-Forgivable) Errors . . . . .	78
	Framing Your Sample . . . . .	78
	Sampling for Success . . . . .	79
	Taking a simple random sample. . . . .	80
	Taking a stratified sample . . . . .	81
	Engaging in systematic sampling . . . . .	82
	Sampling clusters . . . . .	83
	Sampling at your convenience . . . . .	84
	Sampling in multiple stages. . . . .	85
<b>CHAPTER 7:</b>	<b>Having Designs on Study Design . . . . .</b>	<b>87</b>
	Presenting the Study Design Hierarchy . . . . .	88
	Describing what we see . . . . .	89
	Getting analytical. . . . .	90
	Going from observational to experimental . . . . .	90
	Climbing the Evidence Pyramid. . . . .	90
	Starting at the base: Expert opinion . . . . .	91
	Making the case with case studies . . . . .	91
	Making statements about the population . . . . .	92
	Going from case series to case-control . . . . .	93
	Following a cohort over time. . . . .	95
	Advancing to the clinical trial stage. . . . .	97
	Reaching the top: Systematic reviews and meta-analyses . . . . .	97
	<b>PART 3: GETTING DOWN AND DIRTY WITH DATA . . . . .</b>	<b>99</b>
<b>CHAPTER 8:</b>	<b>Getting Your Data into the Computer . . . . .</b>	<b>101</b>
	Looking at Levels of Measurement. . . . .	102
	Classifying and Recording Different Kinds of Data. . . . .	103
	Dealing with free-text data. . . . .	103
	Assigning participant study identification (ID) numbers . . . . .	104
	Organizing name and address data in the study ID crosswalk. . . . .	104
	Collecting categorical data in your research database . . . . .	105

	Recording numerical data . . . . .	107
	Entering date and time data . . . . .	108
	Checking Your Entered Data for Errors . . . . .	109
	Creating a File that Describes Your Data File . . . . .	110
<b>CHAPTER 9:</b>	<b>Summarizing and Graphing Your Data . . . . .</b>	<b>111</b>
	Summarizing and Graphing Categorical Data . . . . .	112
	Summarizing Numerical Data . . . . .	114
	Locating the center of your data . . . . .	115
	Describing the spread of your data . . . . .	119
	Numerically expressing the symmetry and shape of the distribution . . . . .	120
	Structuring Numerical Summaries into Descriptive Tables . . . . .	123
	Graphing Numerical Data . . . . .	124
	Showing the distribution with histograms . . . . .	124
	Summarizing grouped data with bars, boxes, and whiskers . . . . .	126
	Depicting the relationships between numerical variables with other graphs . . . . .	128
<b>CHAPTER 10:</b>	<b>Having Confidence in Your Results . . . . .</b>	<b>129</b>
	Feeling Confident about Confidence Interval Basics . . . . .	130
	Defining confidence intervals . . . . .	130
	Understanding and interpreting confidence levels . . . . .	131
	Taking sides with confidence intervals . . . . .	132
	Calculating Confidence Intervals . . . . .	133
	Before you begin: Formulas for confidence limits in large samples . . . . .	133
	The confidence interval around a mean . . . . .	134
	The confidence interval around a proportion . . . . .	135
	The confidence interval around an event count or rate . . . . .	136
	Relating Confidence Intervals and Significance Testing . . . . .	137
	<b>PART 4: COMPARING GROUPS . . . . .</b>	<b>139</b>
<b>CHAPTER 11:</b>	<b>Comparing Average Values between Groups . . . . .</b>	<b>141</b>
	Grasping Why Different Situations Need Different Tests . . . . .	142
	Comparing the mean of a group of numbers to a hypothesized value . . . . .	142
	Comparing the mean of two groups of numbers . . . . .	143
	Comparing the means of three or more groups of numbers . . . . .	143
	Comparing means in data grouped on several different variables . . . . .	144
	Adjusting for a confounding variable when comparing means . . . . .	145
	Comparing means from sets of matched numbers . . . . .	145
	Comparing means of matched pairs . . . . .	146

	Using Statistical Tests for Comparing Averages . . . . .	146
	Surveying Student t tests . . . . .	147
	Assessing the ANOVA . . . . .	152
	Running nonparametric tests . . . . .	157
	Estimating the Sample Size You Need for Comparing Averages . . . . .	158
	Using formulas for manual calculation . . . . .	158
	Software and web pages . . . . .	158
<b>CHAPTER 12:</b>	<b>Comparing Proportions and Analyzing</b>	
	<b>Cross-Tabulations . . . . .</b>	<b>159</b>
	Examining Two Variables with the Pearson Chi-Square Test . . . . .	161
	Understanding how the chi-square test works . . . . .	161
	Pointing out the pros and cons of the chi-square test . . . . .	167
	Modifying the chi-square test: The Yates continuity correction . . . . .	168
	Focusing on the Fisher Exact Test . . . . .	169
	Understanding how the Fisher Exact test works . . . . .	169
	Noting the pros and cons of the Fisher Exact test . . . . .	170
	Calculating Power and Sample Size for Chi-Square and Fisher Exact Tests . . . . .	171
<b>CHAPTER 13:</b>	<b>Taking a Closer Look at Fourfold Tables . . . . .</b>	<b>173</b>
	Focusing on the Fundamentals of Fourfold Tables . . . . .	174
	Choosing the Correct Sampling Strategy . . . . .	175
	Producing Fourfold Tables in a Variety of Situations . . . . .	176
	Describing the association between two binary variables . . . . .	177
	Quantifying associations . . . . .	178
	Evaluating diagnostic procedures . . . . .	183
	Investigating treatments . . . . .	187
	Looking at inter- and intra-rater reliability . . . . .	188
<b>CHAPTER 14:</b>	<b>Analyzing Incidence and Prevalence Rates in</b>	
	<b>Epidemiologic Data . . . . .</b>	<b>191</b>
	Understanding Incidence and Prevalence . . . . .	192
	Prevalence: The fraction of a population with a particular condition . . . . .	192
	Incidence: Counting new cases . . . . .	192
	Understanding how incidence and prevalence are related . . . . .	194
	Analyzing Incidence Rates . . . . .	194
	Expressing the precision of an incidence rate . . . . .	194
	Comparing incidences with the rate ratio . . . . .	195
	Calculating confidence intervals for a rate ratio . . . . .	196
	Comparing two event rates . . . . .	196
	Comparing two event counts with identical exposure . . . . .	197
	Estimating the Required Sample Size . . . . .	198

<b>PART 5: LOOKING FOR RELATIONSHIPS WITH CORRELATION AND REGRESSION</b> .....	199
<b>CHAPTER 15: Introducing Correlation and Regression</b> .....	201
Correlation: Estimating How Strongly Two Variables Are Associated .....	202
Lining up the Pearson correlation coefficient.....	202
Analyzing correlation coefficients.....	203
Regression: Discovering the Equation that Connects the Variables .....	207
Understanding the purpose of regression analysis.....	207
Talking about terminology and mathematical notation .....	208
Classifying different kinds of regression .....	209
<b>CHAPTER 16: Getting Straight Talk on Straight-Line Regression</b> .....	213
Knowing When to Use Straight-Line Regression.....	213
Understanding the Basics of Straight-Line Regression .....	215
Running a Straight-Line Regression .....	216
Taking a few basic steps.....	217
Walking through an example.....	217
Interpreting the Output of Straight-Line Regression .....	220
Seeing what you told the program to do.....	220
Evaluating residuals .....	221
Making your way through the regression table .....	224
Wrapping up with measures of goodness-of-fit .....	227
Scientific fortune-telling with the prediction formula .....	228
Recognizing What Can Go Wrong with Straight-Line Regression.....	229
Calculating the Sample Size You Need.....	230
<b>CHAPTER 17: More of a Good Thing: Multiple Regression</b> .....	233
Understanding the Basics of Multiple Regression .....	234
Defining a few important terms .....	234
Being aware of how the calculations work .....	235
Executing a Multiple Regression Analysis in Software.....	236
Preparing categorical variables .....	236
Recoding categorical variables as numerical.....	237
Creating scatter charts before you jump into multiple regression analysis .....	238
Taking a few steps with your software.....	241
Interpreting the Output of a Multiple Regression Analysis.....	241
Examining typical multiple regression output.....	241
Checking out optional output to request.....	243
Deciding whether your data are suitable for regression analysis .....	243
Determining how well the model fits the data .....	244

Watching Out for Special Situations that Arise in Multiple Regression . . . . .	245
Synergy and anti-synergy . . . . .	246
Collinearity and the mystery of the disappearing significance. . .	246
Calculating How Many Participants You Need . . . . .	247
<b>CHAPTER 18: A Yes-or-No Proposition: Logistic Regression. . . . .</b>	<b>249</b>
Using Logistic Regression. . . . .	250
Understanding the Basics of Logistic Regression. . . . .	251
Fitting a function with an S shape to your data . . . . .	252
Handling multiple predictors in your logistic model . . . . .	255
Running a Logistic Regression Model with Software . . . . .	256
Interpreting the Output of Logistic Regression. . . . .	257
Seeing summary information about the variables. . . . .	258
Assessing the adequacy of the model . . . . .	258
Checking out the table of regression coefficients. . . . .	259
Predicting probabilities with the fitted logistic formula. . . . .	260
Making yes or no predictions. . . . .	262
Heads Up: Knowing What Can Go Wrong with Logistic Regression . . . . .	266
Don't misinterpret odds ratios for numerical predictors . . . . .	267
Beware of the complete separation problem . . . . .	267
Figuring Out the Sample Size You Need for Logistic Regression . . .	268
<b>CHAPTER 19: Other Useful Kinds of Regression. . . . .</b>	<b>271</b>
Analyzing Counts and Rates with Poisson Regression. . . . .	271
Introducing the generalized linear model . . . . .	272
Running a Poisson regression . . . . .	273
Interpreting the Poisson regression output . . . . .	275
Discovering other uses for Poisson regression. . . . .	276
Anything Goes with Nonlinear Regression . . . . .	279
Distinguishing nonlinear regression from other kinds . . . . .	279
Checking out an example from drug research . . . . .	280
Running a nonlinear regression . . . . .	282
Interpreting the output. . . . .	283
Using equivalent functions to fit the parameters you really want. . . . .	285
Smoothing Nonparametric Data with LOWESS. . . . .	286
Running LOWESS. . . . .	287
Adjusting the amount of smoothing. . . . .	289
<b>CHAPTER 20: Getting the Hint from Epidemiologic Inference. . . . .</b>	<b>291</b>
Staying Clearheaded about Confounding . . . . .	292
Avoiding overloading . . . . .	293
Adjusting for confounders . . . . .	294
Understanding Interaction (Effect Modification) . . . . .	296

Getting Casual about Cause . . . . .	297
Rothman's causal pie . . . . .	297
Bradford Hill's criteria of causality . . . . .	298
<b>PART 6: ANALYZING SURVIVAL DATA . . . . .</b>	<b>299</b>
<b>CHAPTER 21: Summarizing and Graphing Survival Data . . . . .</b>	<b>301</b>
Understanding the Basics of Survival Data . . . . .	302
Examining how survival times are intervals . . . . .	302
Recognizing that survival times aren't normally distributed. . . . .	303
Considering censoring . . . . .	303
Looking at the Life-Table Method . . . . .	307
Making a life table . . . . .	307
Interpreting a life table . . . . .	311
Graphing hazard rates and survival probabilities from a life table . . . . .	312
Digging Deeper with the Kaplan-Meier Method . . . . .	313
Heeding a Few Guidelines for Life-Tables and the Kaplan-Meier Method . . . . .	315
Recording survival times correctly . . . . .	315
Miscoding censoring information . . . . .	316
<b>CHAPTER 22: Comparing Survival Times . . . . .</b>	<b>317</b>
Comparing Survival between Two Groups with the Log-Rank Test. . . . .	319
Understanding what the log-rank test is doing. . . . .	319
Running the log-rank test on software. . . . .	320
Looking at the calculations. . . . .	320
Assessing the assumptions . . . . .	323
Considering More Complicated Comparisons . . . . .	324
Estimating the Sample Size Needed for Survival Comparisons . . . . .	324
<b>CHAPTER 23: Survival Regression . . . . .</b>	<b>327</b>
Knowing When to Use Survival Regression . . . . .	328
Grasping the Concepts behind Survival Regression. . . . .	329
The steps to perform a PH regression . . . . .	330
Hazard ratios . . . . .	334
Executing a Survival Regression . . . . .	335
Interpreting the Output of a Survival Regression. . . . .	337
Testing the validity of the assumptions . . . . .	339
Checking out the table of regression coefficients. . . . .	340
Homing in on hazard ratios and their confidence intervals . . . . .	341
Assessing goodness-of-fit and predictive ability of the model . . . . .	342
Focusing on baseline survival and hazard functions . . . . .	342

How Long Have I Got, Doc? Constructing Prognosis Curves. . . . .	343
Obtaining the necessary output . . . . .	343
Finding h . . . . .	345
Estimating the Required Sample Size for a Survival Regression . . .	346
<b>PART 7: THE PART OF TENS</b> . . . . .	349
<b>CHAPTER 24: Ten Distributions Worth Knowing</b> . . . . .	351
The Uniform Distribution . . . . .	352
The Normal Distribution. . . . .	353
The Log-Normal Distribution . . . . .	353
The Binomial Distribution . . . . .	354
The Poisson Distribution . . . . .	355
The Exponential Distribution . . . . .	356
The Weibull Distribution. . . . .	356
The Student t Distribution . . . . .	357
The Chi-Square Distribution . . . . .	358
The Fisher F Distribution . . . . .	359
<b>CHAPTER 25: Ten Easy Ways to Estimate How Many Participants You Need</b> . . . . .	361
Comparing Means between Two Groups . . . . .	362
Comparing Means among Three, Four, or Five Groups. . . . .	362
Comparing Paired Values. . . . .	363
Comparing Proportions between Two Groups. . . . .	363
Testing for a Significant Correlation . . . . .	363
Comparing Survival between Two Groups . . . . .	364
Scaling from 80 Percent to Some Other Power . . . . .	365
Scaling from 0.05 to Some Other Alpha Level. . . . .	365
Adjusting for Unequal Group Sizes. . . . .	366
Allowing for Attrition. . . . .	366
<b>INDEX</b> . . . . .	369





# Introduction

---

**B**iostatistics is the practical application of statistical concepts and techniques to topics in the biology and life sciences fields. Because these are broad fields, biostatistics covers a very wide area. It is used when studying many types of experimental units, from viruses to trees to fleas to mice to people. Biostatistics involves designing research studies, safely conducting human research, collecting and verifying research data, summarizing and displaying the data, and analyzing the data to answer research hypotheses and draw meaningful conclusions.

It is not possible to cover all the subspecialties of biostatistics in one book, because such a book would have to include chapters on molecular biology, genetics, agricultural studies, animal research (both inside and outside the lab), clinical trials, and epidemiological research. So instead, we focus on the most widely applicable topics of biostatistics and on the topics that are most relevant to human research based on a survey of graduate-level biostatistics curricula from major universities.

## About This Book

---

We wrote this book to be used as a reference. Our intention was for you to pull out this book when you want information about a particular topic. This means you don't have to read it from beginning to end to find it useful. In fact, you can jump directly to any part that interests you. We hope you'll be inclined to look through the book from time to time, open it to a page at random, read a page or two, and get a useful reminder or pick up a new fact.

Only in a few places does this book provide detailed steps about how to perform a particular statistical calculation by hand. Instruction like that may have been necessary in the mid-1900s. Back then, statistics students spent hours in a *computing lab*, which is a room that had an adding machine. Thankfully, we now have statistical software to do this for us (see Chapter 4 for advice on choosing statistical software). When describing statistical tests, our focus is always on the concepts behind the method, how to prepare your data for analysis, and how to interpret the results. We keep mathematical formulas and derivations to a minimum. We

only include them when we think 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 study design is crucial for the success of any research, this book gives special attention to the design of both epidemiologic studies and clinical trials. We also pay special attention to providing advice on how to calculate the number of participants you need for your study. You will find easy-to-apply examples of sample-size calculations in the chapters describing significance tests in Parts 4, 5, and 6, and in Chapter 25.

## Foolish Assumptions

We wrote this book to help several kinds of people. We 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
- » Professionals who have had no formal biostatistical training, and possibly no statistical training at all, who now must analyze biological or research data as part of their work
- » Doctors, nurses, and other healthcare professionals who want to carry out human research

If you're interested in biostatistics, then you're no dummy! But perhaps you sometimes *feel* like a dummy when it comes to biostatistics, or statistics in general, or even mathematics. Don't feel bad. We both have felt that way many times over the years. In fact, we still feel like that whenever we are propelled into an area of biostatistics with which we are unfamiliar, because it is new to us. (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.)

What is important to keep in mind when learning biostatistics is that you don't have to be a math genius to be a good biostatistician. You also don't need any special math skills to be an excellent research scientist who can intelligently design research studies, execute them well, collect and analyze data properly, and draw valid conclusions. You just have to have a solid grasp of the basic concepts and know how to utilize statistical software properly to obtain the output you need and interpret it.

# Icons Used in This Book

Icons are the little graphics in the margins of this book, and are used to draw your attention to certain kinds of material. Here's what they mean:



REMEMBER

This icon signals information especially worth keeping in mind. Your main take-aways from this book should be the material marked with this icon.



TECHNICAL  
STUFF

We use this icon to flag explanations of technical topics, such as derivations and computational formulas that you don't have to know to do biostatistics. They are included to give you deeper insight into the material.



TIP

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 a topic or concept.



WARNING

This icon alerts you to discussion of a controversial topic, a concept that is often misunderstood, or a pitfall or common mistake to guard against in biostatistics.

# Beyond the Book

In addition to the abundance of information and guidance related to using biostatistics for analysis of research data that we provide in this book, you get access to even more help and information online at [Dummies.com](http://Dummies.com). Check out this book's online Cheat Sheet. Just go to [www.dummies.com](http://www.dummies.com) and search for "Biostatistics For Dummies Cheat Sheet."

# 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. For a more detailed list of topics, 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 headings for the sections and subsections of that chapter.

If you want to get the big picture of what biostatistics encompasses and the areas of biostatistics covered in this book, then read Chapter 1. This is a top-level overview of the book's topics. Here are a few other special parts of this book you may want to jump into first, depending on your interest:

- » If you're uncomfortable with mathematical notation, then Chapter 2 is the place to start.
- » If you want a quick refresher on basic statistics like what you would learn in a typical introductory course, then read Chapter 3.
- » You can get an introduction to human research and clinical trials in Chapters 5, 7, and 20.
- » If you want to learn about collecting, summarizing, and graphing data, jump to Part 3.
- » If you need to know about working with survival data, you can go right to Part 6.
- » If you're puzzled about a particular statistical distribution function, then look at Chapter 24.
- » And if you need to calculate some quick sample-size estimates, turn to Chapter 25.

# 1 Getting Started with Biostatistics

### **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 you may have learned previously, such as probability, randomness, populations, samples, statistical inference, and more.

#### IN THIS CHAPTER

- » Getting up to speed on the prerequisites for biostatistics
- » Understanding the human research environment
- » Surveying the specific procedures used to analyze biological data
- » Estimating how many participants you need
- » Working with distributions

## Chapter **1**

# Biostatistics 101

**B**iostatistics deals with the design and execution of scientific studies involving biology, the acquisition and analysis of data from those studies, 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 teach how to analyze data from epidemiologic studies and clinical trials, this book focuses on that as well. In this first chapter, we 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 give you supplementary information or context that you may find useful while 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 statistics books, we include 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 we include Chapter 2, "Overcoming Mathophobia" to show you how to read and understand the basic mathematical notation we use in this book. We 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 feel unsure 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, 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

For instructional purposes, some chapters in this book include step-by-step instructions for performing statistical tests and analyses by hand. We include such instruction only to illustrate the concepts that are involved in the procedure or to demonstrate calculations that are simple to do manually.

However, we demonstrate many of the statistical functions we talk about in this book using R, which is a free, open-source software package. If you are in a class and assigned a particular software package to use, you will have to use that software for the course, which may be commercial software associated with a fee. However, if you are learning on your own, you may choose to use open-source software, which is free. Chapter 4 provides guidance on both commercial and free software.



# Concentrating on Epidemiologic Research



REMEMBER

This book covers topics that are applicable to all areas of biostatistics, concentrating on methods that are especially relevant to *epidemiologic research* — studies involving people. This includes *clinical trials*, which are experiments done to develop therapeutic interventions such as drugs. Because policy in healthcare is often based on the results from clinical trials, if you make mistake analyzing clinical trial data, it can have disastrous and wide-ranging human and financial consequences. Even if you don't expect to ever work in a domain that relies heavily on clinical trials (such as drug development research), ensuring that you have a working knowledge of how to manage the statistical issues seen in clinical trials is critical.

Three chapters discuss clinical trials:

- » Chapter 5 describes the statistical aspects of clinical trials as three phases. First, it covers the design phase, where a study protocol is written. Next, it describes the execution phase, where data are collected, and efforts are made to prevent invalid or missing data. In the final phase, data from the study are analyzed and interpreted to answer the hypotheses.
- » Chapter 7 presents epidemiologic study designs and explains the importance of the clinical trial as a study design.
- » Chapter 20 explains the role well-designed clinical trials play in accruing evidence of causal inference in biostatistics.

Much of the work in biostatistics is using data from samples to make inferences about the background population from which the sample was drawn. Now that we have large databases, it is possible to easily take samples of data. Chapter 6 provides guidance on different ways to take samples of larger populations so you can make valid population-based estimates from these samples. Sampling is especially important when doing observational studies. While clinical trials covered are experiments, where participants are assigned interventions, in observational studies, participants are merely observed, with data collected and statistics performed to make inferences. Chapter 7 describes these observational study designs, and the statistical issues that need to be considered when analyzing data arising from such studies.

Data used in biostatistics are often collected in online databases, but some data are still collected on paper. Regardless of the source of the data, they must be put into electronic format and arranged in a certain way to be able to be analyzed using statistical software. Chapter 8 is devoted to describing how to get your data into the computer and arrange it properly so it can be analyzed correctly. It also

describes how to collect and validate your data. Then in Chapter 9, we show you how to summarize each type of data and display it graphically. We explain how to make bar charts, box-and-whiskers charts, and more.

## Drawing Conclusions from Your Data

Most statistical analysis involves *inferring*, or drawing conclusions about the population at large based on your observations of a sample drawn from that population. The theory of *statistical inference* is often divided into two broad sub-theories: *estimation* theory and *decision* theory.

### Statistical estimation theory

Chapter 10 deals with *statistical estimation theory*, which addresses the question of how accurately and precisely you can estimate a population parameter from the values you observe in your sample. For example, you may want to estimate the mean blood hemoglobin concentration in adults with Type II diabetes, or the true correlation coefficient between body weight and height in certain pediatric populations. Chapter 10 describes how to estimate these parameters by constructing a *confidence interval* around your estimate. The confidence interval is the range that is likely to include the true population parameter, which provides an idea of the precision of your estimate.

### Statistical decision theory

Much of the rest of this book deals with *statistical decision theory*, which is how to decide whether some effect you've observed in your data reflects a real difference or association in the background population or is merely the result of random fluctuations in your data or sampling. If you measure the mean blood hemoglobin concentration in two different samples of adults with Type II diabetes, you will likely get a different number. But does this difference reflect a real difference between the groups in terms of blood hemoglobin concentration? Or is this difference a result of random fluctuations? Statistical decision theory helps you decide.

In Part 4, we cover statistical decision theory in terms of comparing means and proportions between groups, as well as understanding the relationship between two or more variables.

## Comparing groups

In Part 4, we show you different ways to compare groups statistically.

- » In Chapter 11, you see how to compare *average values* between two or more groups by using t tests and ANOVAs. We also describe their nonparametric counterparts that can be used with skewed or other non-normally distributed data.
- » Chapter 12 shows how to compare *proportions* between two or more groups, such as the proportions of patients responding to two different drugs, using the chi-square and Fisher Exact tests on cross-tabulated (cross-tab) data.
- » Chapter 13 focuses on one specific kind of cross-tab called the *fourfold table*, which has exactly two rows and two columns. Because the fourfold table provides the opportunity for some particularly insightful calculations, it's worth a chapter of its own.
- » In Chapter 14, you discover how the terminology used in epidemiologic studies is applied to specifically formatted fourfold tables to calculate incidence and prevalence rates.

## Looking for relationships between variables

Epidemiology and biostatistics are interested in *causal inference*, which means trying to figure out what causes particular outcomes in biological research. While it is possible to look at the relationship between two variables in a *bivariate analysis*, regression analysis is the part of statistics that enables you to explore the relationship between multiple variables and one outcome in the same model so you can evaluate their relative cause of the outcome. Here are some use-cases for regression:

- » You may want to know whether there's a *statistically significant association* between one or more variables and an outcome, even if there are other variables in the model. You may ask: Does being overweight increase the likelihood of getting liver cancer? Or: Is exercising fewer hours per week associated with higher blood pressure measurements? In answering both of those questions, you may want to control other variables known to influence the outcome.
- » You may want to develop a formula for predicting the value of a variable from the observed values of one or more other variables. For example, you may want to predict how long a newly diagnosed cancer patient may survive based on their age, obesity status, and medical history.

» You may be fitting a theoretical formula to some data to estimate one of the parameters appearing in that formula. An example of such a problem is determining how fast the kidneys can remove a drug from the body, which is called a terminal elimination rate constant. This can be estimated from measurements of drug concentration in the blood taken at various times after taking a dose of the drug.

Regression analysis can manage all these tasks and many more. Regression is so important in biological research that all the chapters in Part 5 are focused on some aspect of regression.



TIP

If you have never learned correlation and regression analysis, read Chapter 15, which introduces these topics. We cover simple straight-line regression in Chapter 16, which includes one predictor variable. We extend that to cover multiple regression with more than one predictor variable in Chapter 17. These three chapters deal with ordinary linear regression, where you're trying to predict the value of a numerical outcome variable from one or more other variables. An example would be trying to predict mean blood hemoglobin concentration using variables like age, blood pressure level, and Type II diabetes status. Ordinary linear regression uses a formula that's a simple summation of terms, each of which consists of a predictor variable multiplied by a regression coefficient.

But in real-world biological and epidemiologic research, you encounter more complicated relationships. Chapter 18 describes *logistic regression*, where the outcome is the occurrence or non-occurrence of an event (such as being diagnosed with Type II diabetes), and you want to predict the probability that the event will occur. You also find out about several other kinds of regression in Chapter 19:

- » *Poisson regression*, where the outcome is the number of events that occur in an interval of time
- » *Nonlinear least-squares regression*, where the relationship between the predictors and numerical outcome can be more complicated than a simple summation of terms in a linear model
- » *LOWESS curve-fitting*, where you fit a custom function to describe your data

Finally, Part 5 ends with Chapter 20, which provides guidance on the mechanics of regression modeling, including how to develop a modeling plan, and how to choose variables to include in models.

# A Matter of Life and Death: Working with Survival Data

Sooner or later, everyone dies, and in biological research, it becomes especially important to characterize that sooner-or-later part as accurately as possible using survival analysis techniques. But characterizing survival can get tricky. It's possible to say that patients may live an average of 5.3 years after they are diagnosed with a particular disease. But what is the exact survival experience? Imagine you do a study with patients who have this disease. You may ask: Do all patients tend to live around five or six years, or do half the patients die within the first few months, and the other half survive ten years or more? And what if some patients live longer than the observational period of your study? How do you include them in your analysis? And what about participants who stopped returning calls from your study staff? You do not know if these dropouts went on to live or die. How do you include their data in your analysis?



REMEMBER

The need to study survival with data like these led to the development of survival analysis techniques. But survival analysis is not only intended to study the outcome of death. You can use survival analysis to study the time to the first occurrence of non-death events as well, like remission or recurrence of cancer, the diagnosis of a particular condition, or the resolution of a particular condition. Survival analysis techniques are presented in Part 6.

## Getting to Know Statistical Distributions

Statistics books always contain tables, so why should this one be any different? Back in the not-so-good old days, when analysts had to do statistical calculations by hand, they needed to use tables of the common statistical distributions to complete the calculation of the significance test. They needed tables for the normal distribution, Student t, chi-square, Fisher F, and others. Now, software does all this for you, including calculating exact p values, so these printed tables aren't necessary anymore.

But you should still be familiar with the common statistical distributions that may describe the fluctuations in your data, or that may be referenced in the course of performing a statistical calculation. Chapter 24 contains a list of commonly used distribution functions, with explanations of where you can expect to encounter those distributions and what they look like. We also include a description of some of their properties and how they're related to other distributions. Some of them are accompanied by a small table of critical values, corresponding to statistical significance at  $\alpha = 0.05$ .

# Figuring Out How Many Participants You Need

---

Of all the statistical challenges a researcher may encounter, none seems to instill as much apprehension and insecurity as having to estimate the number of participants needed for a study. While smaller sample sizes mean less data collection work, you want to make sure your target sample size is large enough so that in the end, your study has sufficient power. You want to conduct a study with a high probability of yielding a statistically significant result if the hypothesized effect is truly present in the population.



TIP

Because sample-size estimation is such an important part of the design of any research project, this book shows you how to make those estimates for the situations you're likely to encounter when doing biological research. As we describe each statistical test in Parts 4, 5, 6, and 7, we explain how to estimate the number of participants needed to provide sufficient power for that test. In addition, Chapter 25 describes ten simple rules for getting a “quick and dirty” estimate of the required sample size.