Making Everything Easier!™

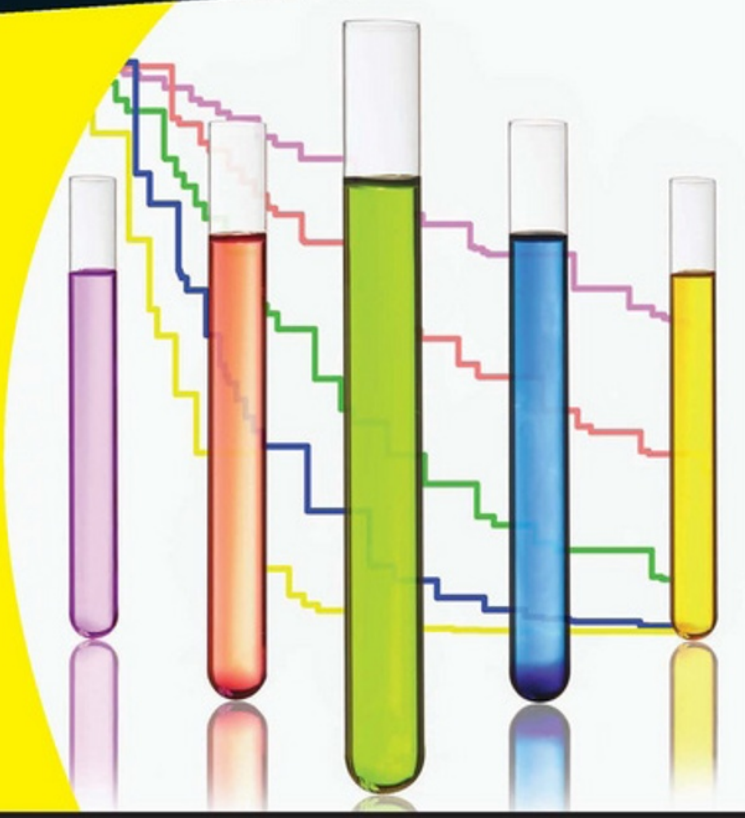
Biostatistics

FOR DUMMIES®

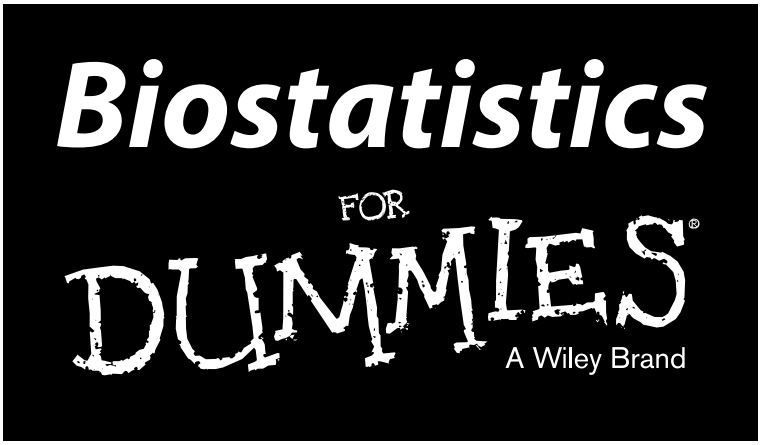
A Wiley Brand

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



by John C. Pezzullo, PhD

FOR DUMMIES
A Wiley Brand

Biostatistics For Dummies®
Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

Published simultaneously in Canada

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

Table of Contents

Introduction

Biostatistics 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`, 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 nonessential 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

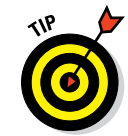
The final two chapters of this book provide "top-ten lists" of handy information and rules that you'll probably refer to often. Chapter 25 describes ten of the most common statistical distribution functions that you encounter in biostatistical research. Some of these distributions describe how your observed 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 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

Biostatistics 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 mathophobe, 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.