LEARNING MADE EASY

2nd Edition

Biostatistics

for dummies®

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 Wahi, MPH and John C. Pezzullo, PhD

for dummies®
A Wiley Brand

Biostatistics For Dummies®, 2nd Edition

Published by: **John Wiley & Sons, Inc.,** 111 River Street, Hoboken, NJ 07030-5774, www.wiley.com

Published simultaneously in Canada

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.

Library of Congress Control Number: 2024939111

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

Table of Contents

Introduction

Biostatistics 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:

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

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.

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.

This icon alerts you to discussion of a controversial topic, a concept that is often misunderstand, 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`. Check out this book's online Cheat Sheet. Just go to `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

Biostatistics 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 mathophobe, 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.