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

COPYRIGHTED MATERIAL

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.

Concentrating on Clinical Research

REMEMBER

This book covers topics that are applicable to all areas of biostatistics, concentrating on methods that are especially relevant to *clinical research* — studies involving people. If you're going to do research on human subjects, you'll want to check out two chapters that deal with clinical trials (and specifically drug development trials). These studies are among the most rigorously designed, closely regulated, expensive, and consequential of all types of scientific research — a mistake here can have disastrous human and financial consequences. So even if you don't expect to ever take part in drug development research, clinical trials (and the statistical issues they entail) are worth a close look.

Two chapters look at clinical research — one from the *inside,* and one from the *outside.*

- Chapter 5 describes the statistical aspects of clinical trials:
  - **Designing the study:** This aspect includes formulating goals, objectives, and hypotheses; estimating the required sample size; and composing the protocol.
  - **Executing the study:** During this phase, you're dealing with regulatory and subject protection groups, randomization and blinding, and collecting data.
  - **Analyzing the data from the study:** At this point, you're validating data, dealing with missing data and multiplicity, and handling interim analyses.
- Chapter 6 describes the whole drug development process, from the initial exploration of promising compounds to the final regulatory approval and the subsequent long-term monitoring of the safety of marketed products. It describes the different kinds of clinical trials that are carried out, in a logical progression, at different phases of the development process.

Many researchers have run into problems while analyzing their data because of decisions they made (or failed to make) while designing and executing their study. Many of these early errors arise from not understanding, or appreciating, the different kinds of data that their study can generate. Chapter 7 shows you how to recognize the kinds of data you encounter in biological research (numerical, categorical, and date- and time-oriented data), and how to collect and validate your data. Then in Chapter 8 you see how to summarize each type of data and display it graphically; your choices include 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 small 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

Chapters 9 and 10 deal with *statistical estimation theory,* which addresses the question of how accurately and precisely you can estimate some population parameter (like the mean blood hemoglobin concentration in all adult males, or the true correlation coefficient between body weight and blood pressure in all adult females) from the values you observe in your sample.

- In Chapter 9, you discover the difference between accuracy and precision (they're not synonymous!), and find out how to calculate the *standard error* (a measure of how precise, or imprecise, your observed value is) for the things you measure or count from your sample.
- In Chapter 10, you find out how to construct a *confidence interval* (the range that is likely to include the true population parameter) for anything you can measure or count.

But often the thing you measure (or count) isn't what you're really interested in. You may measure height and weight, but really be interested in body mass index, which is calculated from height and weight by a simple formula. If every number you acquire directly has some degree of imprecision, then anything you calculate from those numbers will also be imprecise, to a greater or lesser extent. Chapter 11 explains how random errors propagate through mathematical expressions and shows you how to calculate the standard error (and confidence interval) for anything you calculate from your raw data.

Statistical decision theory

Much of the rest of this book deals with *statistical decision theory* — how to decide whether some effect you've observed in your data (such as the difference in the average value of a variable between two groups or the association between two variables) reflects a real difference or association in the population or is merely the result of random fluctuations in your data or sampling.

Decision theory, as covered in this book, can also be divided into two broad sub-categories — comparing means and proportions between groups (in Part III), and understanding the relationship between two or more variables (in Part IV).

Comparing groups

In Part III, you meet (or get reacquainted with) some of the famous-name tests.

- In Chapter 12, you see how to compare *average values* between two or more groups by using t tests and ANOVAs, and their counterparts (Wilcoxon, Mann-Whitney, and Kruskal-Wallis tests) that can be used with skewed or other non-normally distributed data.
- Chapter 13 shows how to compare *proportions* (like cure rates) between two or more groups, using the chi-square and Fisher Exact tests on cross-tabulated data.
- Chapter 14 focuses on one specific kind of cross-tab — the *fourfold table* (having two rows and two columns). It turns out that you can get a lot of very useful information from a fourfold table, so it's worth a chapter of its own.
- In Chapter 15, you see how *event rates* (also called *person-time* data) can be estimated and compared between groups.
- Chapter 16 wraps up Part III with a description of a special kind of analysis that occurs often in biological research — *equivalence* and *non-inferiority* testing, where you try to show that two treatments or products aren't really different from each other or that one isn't any worse than the other.

Looking for relationships between variables

Science is, at its heart, the search for relationships, and regression analysis is the part of statistics that deals with the nature of relationships between different variables:

- You may want to know whether there's a *significant association* between two variables: Do smokers have a greater risk of developing liver cancer than nonsmokers, or is age associated with diastolic blood pressure?
- You may want to develop a formula for predicting the value of a variable from the observed values of one or more other variables: Can you predict the duration of a woman's labor if you know how far along the pregnancy is (the gestational age), how many other children she has had in the past (her *parity*), and how much the baby-to-be weighs (from ultrasound measurements)?

- You may be fitting a theoretical formula to some data in order to estimate one of the parameters appearing in that formula — like determining how fast the kidneys can remove a drug from the body (a terminal elimination rate constant), from measurements of drug concentration in the blood at various times after taking a dose of the drug.

Regression analysis can handle all these tasks, and many more besides. Regression is so important in biological research that this book devotes Part IV to it. But most Stats 101 courses either omit regression analysis entirely or cover only the very simplest type — fitting a *straight line* to a set of points. Even second semester statistics courses may go only as far as *multivariate linear regression,* where you can have more than one predictor variable.

If you know nothing of correlation and regression analysis, read Chapter 17, which provides an introduction to these topics. I cover simple straight-line regression in Chapter 18; I extend that coverage to more than one predictor variable in Chapter 19. These three chapters deal with ordinary linear regression, where you're trying to predict the value of a numerical outcome variable (like blood pressure or serum glucose) from one or more other variables (such as age, weight, and gender) by using 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 clinical research, you encounter more complicated relationships. Chapter 20 describes *logistic regression,* where the outcome is the occurrence or nonoccurrence of some kind of event, and you want to predict the probability that the event will occur. And you find out about several other kinds of regression in Chapter 21:

- *Poisson regression,* where the outcome can be the number of events that occur in an interval of time
- *Nonlinear least-squares regression,* where the relationship can be more complicated than a simple summation of terms in a linear model
- *LOWESS curve-fitting,* where you may have no explicit formula at all that describes the data

A Matter of Life and Death: Working with Survival Data

Sooner or later, all living things die. And in biological research, it becomes very important to characterize that sooner-or-later part as accurately as possible. But this characterization can get tricky. It's not enough to say that people live an average of 5.3 years after acquiring a certain disease. Does everyone tend to last five or six years, or do half the people die within the first few months, and the other half survive ten years or more? And how do you analyze your data when some subjects may far outlive your clinical study (that is, they're still alive when you have to finish your study and write up the results)? And how do you analyze people who skip town after a few months, so you don't know whether they lived or died after that?

The existence of problems like these led to the development of a special set of techniques specifically designed to deal with survival data. More generally, they also apply to the time of the first occurrence of other (non-death) events as well, like remission or recurrence of cancers, heart attacks, strokes, and first bowel movement after abdominal surgery. These techniques, which span the whole data analysis process, are all collected in Part V.

To discover how to acquire survival data properly (it's not as obvious as you may think), read Chapter 22, where I also show how to summarize and graph survival data, and how to estimate such things as mean and median survival time and percent survival to specified time points. A special statistical test for comparing survival among groups of subjects is covered in Chapter 23. And in Chapter 24, I describe Cox proportional-hazards regression — a special kind of regression analysis for survival data.

Figuring Out How Many Subjects You Need

Of all the statistical challenges a researcher may encounter, none seems to instill as much apprehension and insecurity as calculating the number of subjects needed to provide a sufficiently powered study — one that provides a high probability of yielding a statistically significant result if the hoped-for 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 clinical research. As I describe each statistical test in Parts III, IV, and V, I explain how to estimate the number of subjects needed to provide sufficient power for that test. In addition, Chapter 26 describes ten simple rules for getting a "quick and dirty" estimate of the required sample size.

Getting to Know Statistical Distributions

What statistics book would be complete without a set of tables? Back in the not-so-good old days, when people had to do statistical calculations by hand, they needed tables of the common statistical distributions (Normal, Student t, chi-square, Fisher F, and so on) in order to complete the calculation of the significance test. But now the computer does all this for you, including calculating the exact p value, so these tables aren't nearly as necessary as they once were.

But you should still be familiar with the common statistical distributions that describe how your observations may fluctuate or that may come up in the course of performing a statistical calculation. So Chapter 25 contains a list of the most well-known distribution functions, with explanations of where you can expect to encounter those distributions, what they look like, what some of their more interesting properties are, and how they're related to other distributions. Some of them are accompanied by a small table of critical values, corresponding to significance at the 5 percent level (that is, $p = 0.05$).