

Preface

This book will discuss basic statistical analysis methods through a series of biological examples using R and R-Commander as computational tools. The book is intended for a wide range of readers, from people with relatively strong analytical background who want to learn about statistics and its application in biology, to nonstatistician scientists who use statistical methods in their research.

While the theoretical aspects of statistics are intriguing and interesting on their own, we believe that what separates statistics from other branches of mathematics is its intimate relationship with other fields, such as biology, economics, and social sciences, and its widespread application in these areas. In statistics, a theoretical work is usually inspired by applied problems, and new theories usually find immediate applications in real-world problems. This interweaving of theory and application has put statistics in a special place in the scientific world.

In this book, most topics are motivated by real examples first. We believe that learning a new topic becomes easier if it is motivated by interesting and engaging applied problems. We also hope that this approach helps students to improve their critical thinking and problem-solving skills for situations where they are presented with new problems. To this end, we motivate each new topic with a relevant problem from biology. We then try to reach the solution intuitively before discussing the related statistical methods. For example, when discussing Bayes' theorem, we first present a biological problem (finding the probability of lung cancer for smokers) and find the answer to that problem intuitively based on what we already know. Then, we introduce Bayes' theorem as a general form of our solution for this type of problem.

While discussing statistical methods and their applications, our goal is to keep a balance between mathematical rigor and readability. To accomplish this, we have moved concepts that tend to be more complex with limited applications in everyday analysis to the end of each chapter in "Advanced" sections. For the most part, these sections could be skipped in the first reading of this book.

Throughout the book, we use R-Commander, a free and publicly available computer program, to show how statistical methods can be used in practice. We believe that using these methods while learning them could help with the learning process.

Most of the examples discussed in this book are based on scientific studies whose data are publicly available. For each example, we provide the step-by-step application of R-Commander. Readers are encouraged to follow these steps while reading the book so that they can learn statistics through hands-on experience.

For some examples, the data are available through R and R-Commander. For these examples, we provide the steps required to obtain the data. For some other examples, the data are available online and can be downloaded from <http://extras.springer.com>. Appendix A shows the steps for installing and using R-Commander. Before reading the chapters, readers should follow these steps to install R-Commander on their computer.

The chapters are arranged according to what a typical statistical analysis involves. We usually start with some specific scientific questions in mind. Then, we design a scientific study to answer those questions. In Chap. 1, we very briefly discuss different types of studies and their objectives. We also present an overview of typical steps we take from raising a scientific question to answering it through statistical methods. These steps always involve identifying a *target population*, which is the group of individuals we want to study (e.g., population of humans, orange trees, cells).

Because the target populations are usually very large, we conduct our studies on a relatively small number of individuals randomly *sampled* (i.e., selected) from the population. From these individuals we collect information in the form of measurements of some specific characteristics such as age, size, and counts. We refer to the information obtained from these individuals as *data* collectively. In Chaps. 2 and 3, we discuss several *data exploration* techniques, which involve summarizing and visualizing data to obtain a high-level understanding of the data and the target population.

We want to generalize what we learn from the individuals participating in our study (i.e., the randomly selected individuals) to the whole population. This generalization should always be accompanied by our acknowledgement that we are not completely certain about our findings since our knowledge of the population is based on a relatively small sample of individuals from that population. Specifically, we always present our findings along with some measurements that reflect the extent of our *uncertainty*. To this end, we use *probability* as a powerful mathematical tool to measure uncertainty. We discuss probability in Chaps. 4 and 5.

The process of analyzing data to learn about the whole population is referred to as *statistical inference*. This usually involves guessing some unknown values, drawing conclusions, and making decisions. Chapters 6, 7, and 8 discuss some basic methods of inferential statistics. Chapters 9, 10, and 11 provide slightly more advanced statistical inference methods, which for the most part could be considered as the generalization of topics covered in Chap. 8.

Finally, Chap. 12 discusses *clustering* methods, and Chap. 13 discusses *Bayesian analysis* very briefly. These topics are not traditionally included in introductory books on statistics. We decided to include these topics due to their immense importance in scientific studies. While this book does not do justice to these two topics, we hope that it serves as an introduction for interested readers.

As mentioned above, we use R-Commander to show how statistical methods can be used for real problems. Using R-Commander does not require any computer programming. For readers who are comfortable with learning a programming language, we discuss the equivalent R programs at the end of each chapter in Advanced sections. These readers should start from Appendix B, where we provide a brief introduction to R programming.

The methods discussed in this book have been developed by many researchers over many years. To avoid overburdening the reader, we provide only a small number of references, mainly for related books that go beyond what we have covered here, and also for real problems that we have used as examples.

Writing this book has helped me to improve my teaching, and the feedback I have received from my students has helped me to improve the book during the past several years. I would like to thank all my students who challenged me with their questions; they have been my toughest critics.

I would like to thank John Fox for developing R-Commander. This is an extremely useful tool for teaching basic statistics to students without programming background.

I would also like to thank Jessica Utts, Michael Phelan, Sam Behseta, and Wesley Johnson for reviewing the book and providing thoughtful comments and constructive criticisms to improve the quality of this book. I am very grateful to have such good friends and supportive colleagues.

A very special thanks goes to Laura Balzer, who is currently a graduate student at UC Berkeley. She has been extremely helpful in the process of preparing the initial draft and editing the book.

Finally, I would like to thank my family for being patient and supportive throughout the process of writing this book; it would not have been possible without their love and support.

Irvine, CA
November 11, 2011

Babak Shahbaba

Biostatistics with R

An Introduction to Statistics Through Biological Data

Shahbaba, B.

2012, XVI, 352 p. 162 illus., 73 illus. in color., Softcover

ISBN: 978-1-4614-1301-1