Springer Texts in Statistics

Jay L. Devore
Kenneth N. Berk

Modern Mathematical Statistics with Applications

Second Edition

Springer

Springer Texts in Statistics

Series Editors:

G. Casella
S. Fienberg
I. Olkin

For further volumes:
http://www.springer.com/series/417

Modern Mathematical Statistics with Applications

Second Edition

Jay L. Devore
California Polytechnic State University

Kenneth N. Berk
Illinois State University

Springer

Jay L. Devore
California Polytechnic State University
Statistics Department
San Luis Obispo California
USA
jdevore@calpoly.edu

Kenneth N. Berk
Illinois State University
Department of Mathematics
Normal Illinois
USA
kberk@ilstu.edu

Additional material to this book can be downloaded from http://extras.springer.com

ISBN 978-1-4614-0390-6 e-ISBN 978-1-4614-0391-3
DOI 10.1007/978-1-4614-0391-3
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011936004

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my wife Carol

whose continuing support of my writing efforts
over the years has made all the difference.

To my wife Laura

who, as a successful author, is my mentor and role model.

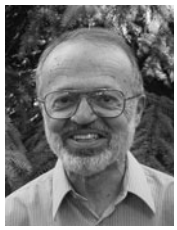
About the Authors

Jay L. Devore

Jay Devore received a B.S. in Engineering Science from the University of California, Berkeley, and a Ph.D. in Statistics from Stanford University. He previously taught at the University of Florida and Oberlin College, and has had visiting positions at Stanford, Harvard, the University of Washington, New York University, and Columbia. He has been at California Polytechnic State University, San Luis Obispo, since 1977, where he was chair of the Department of Statistics for 7 years and recently achieved the exalted status of Professor Emeritus.

Jay has previously authored or coauthored five other books, including *Probability and Statistics for Engineering and the Sciences*, which won a McGuffey Longevity Award from the Text and Academic Authors Association for demonstrated excellence over time. He is a Fellow of the American Statistical Association, has been an associate editor for both the *Journal of the American Statistical Association* and *The American Statistician*, and received the Distinguished Teaching Award from Cal Poly in 1991. His recreational interests include reading, playing tennis, traveling, and cooking and eating good food.

Kenneth N. Berk



Ken Berk has a B.S. in Physics from Carnegie Tech (now Carnegie Mellon) and a Ph.D. in Mathematics from the University of Minnesota. He is Professor Emeritus of Mathematics at Illinois State University and a Fellow of the American Statistical Association. He founded the Software Reviews section of *The American Statistician* and edited it for 6 years. He served as secretary/treasurer, program chair, and chair of the Statistical Computing Section of the American Statistical Association, and he twice co-chaired the Interface Symposium, the main annual meeting in statistical computing. His published work includes papers on time series, statistical computing, regression analysis, and statistical graphics, as well as the book *Data Analysis with Microsoft Excel* (with Patrick Carey).

Contents

Preface

Purpose

Our objective is to provide a postcalculus introduction to the discipline of statistics that

- Has mathematical integrity and contains some underlying theory.
- Shows students a broad range of applications involving real data.
- Is very current in its selection of topics.
- Illustrates the importance of statistical software.
- Is accessible to a wide audience, including mathematics and statistics majors (yes, there are a few of the latter), prospective engineers and scientists, and those business and social science majors interested in the quantitative aspects of their disciplines.

A number of currently available mathematical statistics texts are heavily oriented toward a rigorous mathematical development of probability and statistics, with much emphasis on theorems, proofs, and derivations. The focus is more on mathematics than on statistical practice. Even when applied material is included, the scenarios are often contrived (many examples and exercises involving dice, coins, cards, widgets, or a comparison of treatment A to treatment B).

So in our exposition we have tried to achieve a balance between mathematical foundations and statistical practice. Some may feel discomfort on grounds that because a mathematical statistics course has traditionally been a feeder into graduate programs in statistics, students coming out of such a course must be well prepared for that path. But that view presumes that the mathematics will provide the hook to get students interested in our discipline. This may happen for a few mathematics majors. However, our experience is that the application of statistics to real-world problems is far more persuasive in getting quantitatively oriented students to pursue a career or take further coursework in statistics. Let's first draw them in with intriguing problem scenarios and applications. Opportunities for exposing them to mathematical foundations will follow in due course. We believe it is more important for students coming out of this course to be able to carry out and interpret the results of a two-sample t test or simple regression analysis than to manipulate joint moment generating functions or discourse on various modes of convergence.

Content

The book certainly does include core material in probability (Chapter 2), random variables and their distributions (Chapters 3–5), and sampling theory (Chapter 6). But our desire to balance theory with application/data analysis is reflected in the way the book starts out, with a chapter on descriptive and exploratory statistical

techniques rather than an immediate foray into the axioms of probability and their consequences. After the distributional infrastructure is in place, the remaining statistical chapters cover the basics of inference. In addition to introducing core ideas from estimation and hypothesis testing (Chapters 7–10), there is emphasis on checking assumptions and examining the data prior to formal analysis. Modern topics such as bootstrapping, permutation tests, residual analysis, and logistic regression are included. Our treatment of regression, analysis of variance, and categorical data analysis (Chapters 11–13) is definitely more oriented to dealing with real data than with theoretical properties of models. We also show many examples of output from commonly used statistical software packages, something noticeably absent in most other books pitched at this audience and level.

Mathematical Level

The challenge for students at this level should lie with mastery of statistical concepts as well as with mathematical wizardry. Consequently, the mathematical prerequisites and demands are reasonably modest. Mathematical sophistication and quantitative reasoning ability are, of course, crucial to the enterprise. Students with a solid grounding in univariate calculus and some exposure to multivariate calculus should feel comfortable with what we are asking of them. The several sections where matrix algebra appears (transformations in Chapter 5 and the matrix approach to regression in the last section of Chapter 12) can easily be deemphasized or skipped entirely.

Our goal is to redress the balance between mathematics and statistics by putting more emphasis on the latter. The concepts, arguments, and notation contained herein will certainly stretch the intellects of many students. And a solid mastery of the material will be required in order for them to solve many of the roughly 1,300 exercises included in the book. Proofs and derivations are included where appropriate, but we think it likely that obtaining a conceptual understanding of the statistical enterprise will be the major challenge for readers.

Recommended Coverage

There should be more than enough material in our book for a year-long course. Those wanting to emphasize some of the more theoretical aspects of the subject (e.g., moment generating functions, conditional expectation, transformations, order statistics, sufficiency) should plan to spend correspondingly less time on inferential methodology in the latter part of the book. We have opted not to mark certain sections as optional, preferring instead to rely on the experience and tastes of individual instructors in deciding what should be presented. We would also like to think that students could be asked to read an occasional subsection or even section on their own and then work exercises to demonstrate understanding, so that not everything would need to be presented in class. Remember that there is never enough time in a course of any duration to teach students all that we'd like them to know!

Acknowledgments

We gratefully acknowledge the plentiful feedback provided by reviewers and colleagues. A special salute goes to Bruce Trumbo for going way beyond his mandate in providing us an incredibly thoughtful review of 40+ pages containing

many wonderful ideas and pertinent criticisms. Our emphasis on real data would not have come to fruition without help from the many individuals who provided us with data in published sources or in personal communications. We very much appreciate the editorial and production services provided by the folks at Springer, in particular Marc Strauss, Kathryn Schell, and Felix Portnoy.

A Final Thought

It is our hope that students completing a course taught from this book will feel as passionately about the subject of statistics as we still do after so many years in the profession. Only teachers can really appreciate how gratifying it is to hear from a student after he or she has completed a course that the experience had a positive impact and maybe even affected a career choice.

Jay L. Devore
Kenneth N. Berk

CHAPTER ONE

Overview and Descriptive Statistics

Introduction

Statistical concepts and methods are not only useful but indeed often indispensable in understanding the world around us. They provide ways of gaining new insights into the behavior of many phenomena that you will encounter in your chosen field of specialization.

The discipline of statistics teaches us how to make intelligent judgments and informed decisions in the presence of uncertainty and variation. Without uncertainty or variation, there would be little need for statistical methods or statisticians. If the yield of a crop were the same in every field, if all individuals reacted the same way to a drug, if everyone gave the same response to an opinion survey, and so on, then a single observation would reveal all desired information.

An interesting example of variation arises in the course of performing emissions testing on motor vehicles. The expense and time requirements of the Federal Test Procedure (FTP) preclude its widespread use in vehicle inspection programs. As a result, many agencies have developed less costly and quicker tests, which it is hoped replicate FTP results. According to the journal article "Motor Vehicle Emissions Variability" (*J. Air Waste Manage. Assoc.*, 1996: 667–675), the acceptance of the FTP as a gold standard has led to the widespread belief that repeated measurements on the same vehicle would yield identical (or nearly identical) results. The authors of the article applied the FTP to seven vehicles characterized as "high emitters." Here are the results of four hydrocarbon and carbon dioxide tests on one such vehicle:

HC *(g/mile)*	13.8	18.3	32.2	32.5
CO *(g/mile)*	118	149	232	236

J.L. Devore and K.N. Berk, *Modern Mathematical Statistics with Applications*, Springer Texts in Statistics,
DOI 10.1007/978-1-4614-0391-3_1,

The substantial variation in both the HC and CO measurements casts considerable doubt on conventional wisdom and makes it much more difficult to make precise assessments about emissions levels.

How can statistical techniques be used to gather information and draw conclusions? Suppose, for example, that a biochemist has developed a medication for relieving headaches. If this medication is given to different individuals, variation in conditions and in the people themselves will result in more substantial relief for some individuals than for others. Methods of statistical analysis could be used on data from such an experiment to determine on the average how much relief to expect.

Alternatively, suppose the biochemist has developed a headache medication in the belief that it will be superior to the currently best medication. A comparative experiment could be carried out to investigate this issue by giving the current medication to some headache sufferers and the new medication to others. This must be done with care lest the wrong conclusion emerge. For example, perhaps really the two medications are equally effective. However, the new medication may be applied to people who have less severe headaches and have less stressful lives. The investigator would then likely observe a difference between the two medications attributable not to the medications themselves, but to a poor choice of test groups. Statistics offers not only methods for analyzing the results of experiments once they have been carried out but also suggestions for how experiments can be performed in an efficient manner to lessen the effects of variation and have a better chance of producing correct conclusions.

1.1 Populations and Samples

We are constantly exposed to collections of facts, or **data**, both in our professional capacities and in everyday activities. The discipline of statistics provides methods for organizing and summarizing data and for drawing conclusions based on information contained in the data.

An investigation will typically focus on a well-defined collection of objects constituting a **population** of interest. In one study, the population might consist of all gelatin capsules of a particular type produced during a specified period. Another investigation might involve the population consisting of all individuals who received a B.S. in mathematics during the most recent academic year. When desired information is available for all objects in the population, we have what is called a **census**. Constraints on time, money, and other scarce resources usually make a census impractical or infeasible. Instead, a subset of the population—a **sample**—is selected in some prescribed manner. Thus we might obtain a sample of pills from a particular production run as a basis for investigating whether pills are conforming to manufacturing specifications, or we might select a sample of last year's graduates to obtain feedback about the quality of the curriculum.