HZ BOOKS
华章教育

经典原版书库

数理统计与数据分析

（英文版·第2版）

Mathematical Statistics and Data Analysis

Second Edition

John A. Rice

DUXBURY PRESS

（美）John A. Rice 加利福尼亚大学伯克利分校 著

机械工业出版社
China Machine Press

THOMSON

经典原版书库

数理统计与数据分析

（英文版·第2版）

Mathematical Statistics and Data Analysis

(Second Edition)

(美) John A. Rice 著
加利福尼亚大学伯克利分校

机械工业出版社
China Machine Press

John A. Rice: Mathematical Statistics and Data Analysis, Second Edition (ISBN: 0-534-20934-3).

本书版权登记号：图字：01-2002-1057

图书在版编目（CIP）数据

数理统计与数据分析（英文版·第2版）/（美）赖斯（Rice, J. A.）著. –北京：机械工业出版社，2003.7

（经典原版书库）

书名原文：Mathematical Statistics and Data Analysis, Second Edition

ISBN: 7-111-12413-8

Ⅰ.数… Ⅱ.赖… Ⅲ.①数理统计–高等学校–教材–英文 ②统计数据–统计分析（数学）–高等学校–教材–英文 Ⅳ.O212

中国版本图书馆CIP数据核字（2003）第050106号

机械工业出版社（北京市西城区百万庄大街22号 邮政编码 100037）

责任编辑：杨海玲

北京忠信诚印刷厂印刷·新华书店北京发行所发行

2003年7月第1版第1次印刷

787mm × 1092mm 1/16 · 42 印张

印数：0 001 - 2 000册

定价：78.00 元

凡购本书，如有倒页、脱页、缺页，由本社发行部调换

We must be careful not to confuse data with the abstractions we use to analyze them.

WILLIAM JAMES (1842–1910)

Preface

Intended Audience

This text is intended for juniors, seniors, or beginning graduate students in statistics, mathematics, natural sciences, and engineering as well as for adequately prepared students in the social sciences and economics. A year of calculus, including Taylor Series and multivariable calculus, and an introductory course in linear algebra are prerequisites.

This Book's Objectives

This book reflects my view of what a first, and for many students last, course in statistics should be. Such a course should include some traditional topics in mathematical statistics (such as methods based on likelihood), topics in descriptive statistics and data analysis with special attention to graphical displays, aspects of experimental design, and realistic applications of some complexity. It should also reflect the quickly growing use of computers in statistics. These themes, properly interwoven, can give students a view of the nature of modern statistics. The alternative of teaching two separate courses, one on theory and one on data analysis, seems to me artificial. Furthermore, many students take only one course in statistics and do not have time for two or more.

Analysis of Data and the Practice of Statistics

In order to draw the above themes together, I have endeavored to write a book closely tied to the practice of statistics. It is in the analysis of real data that one sees the roles played by both formal theory and informal data analytic methods. I have organized this book around various kinds of problems that entail the use of statistical methods and have included many real examples to motivate and introduce the theory. Among the advantages of such an approach are that theoretical constructs are presented in meaningful contexts, that they are gradually supplemented and reinforced, and that they are integrated with more informal methods. This is, I think, a fitting approach to statistics, the historical development of which has been spurred on primarily by practical needs rather than abstract or aesthetic considerations. At the same time, I have not shied away from using the mathematics that the students are supposed to know.

This Revision

The basic intent and structure of the book remain the same. In composing the second edition, I have focused my efforts in two areas: improving the existing material pedagogically and incorporating new material. Thus, in the first area, I have expanded and revised discussions where I thought the existing discussion too terse, and I have included new examples in the text where I thought they would be helpful to the reader. For example, I have revised the introduction of confidence intervals in Chapter 7 and their reintroduction in Chapter 8. The introduction of the Mann-Whitney test in Chapter 11 has been rewritten in order to make the ideas clearer. More than 150 new problems have been added. In particular, to help students check their comprehension, these include a large number of routine exercises, such as true-false questions. Some more advanced problems have been added as well.

One of the most influential developments in statistics in the last decade has been the introduction and rapid dissemination of bootstrap methods. This development is of such fundamental importance that, to my mind, its inclusion in an introductory course at the level of this text is mandatory. I introduce the bootstrap in Chapter 8, where the parametric bootstrap arises quite naturally. As well as being of great practical importance, introduction of the bootstrap at this point reinforces the concept of a sampling distribution. The nonparametric bootstrap is introduced in Chapter 10 in the context of estimating the standard error of a location estimate. It arises again in Chapter 11 as a method for assessing the variability of a shift estimate, in Chapter 13 for assessing the variability of the estimate of an odds ratio (a new section of Chapter 13 is devoted to the odds ratio), and finally in Chapter 14 in a discussion of the "random X model" for regression. New problems throughout these chapters ask students how to use the bootstrap to estimate standard errors and confidence intervals for various functionals.

Brief Outline

A complete outline can be found, of course, in the table of contents. Here I will just highlight some points and indicate various curricular options for the instructor.

The first six chapters contain an introduction to probability theory, particularly those aspects most relevant to statistics. Chapter 1 introduces the basic ingredients of probability theory and elementary combinatorial methods from a non-measure theoretic point of view. In this and the other probability chapters, I have tried to use real-world examples rather than balls and urns whenever possible.

The concept of a random variable is introduced in Chapter 2. I chose to discuss discrete and continuous random variables together, instead of putting off the continuous case until later. Several common distributions are introduced. An advantage of this approach is that it provides something to work with and develop in later chapters.

Chapter 3 continues the treatment of random variables by going into joint distributions. The instructor may wish to skip lightly over Jacobians; this can be done with little loss of continuity, since they are utilized rarely in the rest of the book. The material in Section 3.7 on extrema and order statistics can be omitted if the instructor is willing to do a little backtracking later.

Expectation, variance, covariance, conditional expectation, and moment-generating functions are taken up in Chapter 4. The instructor may wish to pass lightly over conditional expectation and prediction, especially if he or she does not plan to cover sufficiency later. The last section of this chapter introduces the δ method, or the method of propagation of error. This method is used several times in the statistics chapters.

The law of large numbers and the central limit theorem are proved in Chapter 5 under fairly strong assumptions.

Chapter 6 is a compendium of the common distributions related to the normal and sampling distributions of statistics computed from the usual normal random sample. I don't spend a lot of time on this material here but do develop the necessary facts as they are needed in the statistics chapters. It is useful for students to have these distributions collected in one place.

Chapter 7 is on survey sampling, an unconventional, but in some ways natural, beginning to the study of statistics. Survey sampling is an area of statistics with which most students have some vague familiarity, and a set of fairly specific, concrete statistical problems can be naturally posed. It is a context in which, historically, many important statistical concepts have developed, and it can be used as a vehicle for introducing concepts and techniques that are developed further in later chapters, for example:

- The idea of an estimate as a random variable with an associated sampling distribution
- The concepts of bias, standard error, and mean squared error
- Confidence intervals and the application of the central limit theorem
- An exposure to notions of experimental design via the study of stratified estimates and the concept of relative efficiency
- Calculation of expectations, variances, and covariances

One of the unattractive aspects of survey sampling is that the calculations are rather grubby. However, there is a certain virtue in this grubbiness, and students are given practice in such calculations. The instructor has quite a lot of flexibility as to how deeply to cover this chapter. The sections on ratio estimation and stratification are optional and can be skipped entirely or returned to at a later time without loss of continuity.

Chapter 8 is concerned with parameter estimation, a subject that is motivated and illustrated by the problem of fitting probability laws to data. The method of moments and the traditional method of maximum likelihood are developed. The concept of efficiency is introduced, and the Cramer-Rao Inequality is proved. Section 8.7 introduces the concept of sufficiency and some of its ramifications. The material on the Cramer-Rao lower bound and on sufficiency can be skipped; to my mind, the importance of sufficiency is usually overstated. Section 8.6.1 (the negative binomial distribution) can also be skipped.

Chapter 9 is an introduction to hypothesis testing with particular application to testing for goodness of fit, which ties in with Chapter 8. (This subject is further developed in Chapter 11.) Informal, graphical methods are presented here as well. Several of the last sections of this chapter can be skipped if the instructor is pressed for time. These include Section 9.7 (the Poisson dispersion test), Section 9.8 (hanging rootograms), and Section 9.10 (tests for normality).

A variety of descriptive methods are introduced in Chapter 10. Many of these techniques are used in later chapters. The importance of graphical procedures is stressed, and notions of robustness are introduced. The placement of a chapter on descriptive methods this late in a book may seem strange. I have chosen to do so because descriptive procedures usually have a stochastic side and, having been through the three chapters preceding this one, students are by now better equipped to study the statistical behavior of various summary statistics (for example, a confidence interval for the median). If the instructor wishes, the material on survival and hazard functions can be skipped.

Classical and nonparametric methods for two-sample problems are introduced in Chapter 11. The concepts of hypothesis testing, first introduced in Chapter 9, are further developed. The chapter concludes with some discussion of experimental design and the interpretation of observational studies.

The first eleven chapters are the heart of an introductory course; the theoretical constructs of estimation and hypothesis testing have been developed, graphical and descriptive methods have been introduced, and aspects of experimental design have been discussed.

The instructor has much more freedom in selecting material from Chapters 12 through 15. In particular, it is not necessary to proceed through these chapters in the order in which they are presented.

Chapter 12 treats the one-way and two-way layouts via analysis of variance and nonparametric techniques. The problem of multiple comparisons, first introduced at the end of Chapter 11, is discussed.

Chapter 13 is a rather brief treatment of the analysis of categorical data. Likelihood ratio tests are developed for homogeneity and independence.

Chapter 14 concerns linear least squares. Simple linear regression is developed first and is followed by a more general treatment using linear algebra. I have chosen to employ matrix algebra but have kept the level of the discussion as simple and concrete as possible, not going beyond concepts typically taught in an introductory one-quarter course. In particular, I have *not developed* a geometric analysis of the general linear model or made any attempt to unify regression and analysis of variance. Throughout this chapter, theoretical results are balanced by more qualitative data analytic procedures based on analysis of residuals.

Chapter 15 is an introduction to decision theory and the Bayesian approach. I believe that students are most likely to appreciate this material after having been exposed to the classical material developed in earlier chapters.

Computer Use and Problem Solving

A VAX and a SUN were used in working the examples in the text and are used by my students on most problems involving real data. My students and I have used both S and Minitab; other packages could be used as well. I have not discussed any particular packages in the text and leave the choice of what, if any, package to use up to the instructor. However, a floppy disk containing the data sets in the text will be available to instructors who have adopted it. Contact the publisher for further information.

This book includes a fairly large number of problems, some of which will be quite difficult for students. I think that problem solving, especially of nonroutine problems, is very important. The course as I teach it includes three hours of lecturing per week and a one-hour section for problem solving and instruction on the use of the computer.

Acknowledgments

I am indebted to a large number of people who contributed directly and indirectly to the first edition. Earlier versions were used in courses taught by Richard Olshen, Yosi Rinnot, Donald Ylvisaker, Len Haff, and David Lane, who made many helpful comments. Students in their classes and in my own had many constructive comments. Teaching assistants, especially Joan Staniswalis, Roger Johnson, Terri Bittner, and Peter Kim, worked through many of the problems and found numerous errors. Many reviewers provided useful suggestions: Rollin Brant, University of Toronto; George Casella, Cornell University; Howard B. Christensen, Brigham Young University; David Fairley, Ohio State University; Peter Guttorp, University of Washington; Hari Iyer, Colorado State University; Douglas G. Kelly, University of North Carolina; Thomas Leonard, University of Wisconsin; Albert S. Paulson, Rensselaer Polytechnic Institute; Charles Peters, University of Houston, University Park; Andrew Rukhin, University of Massachusetts, Amherst; Robert Schaefer, Miami University; and Ruth Williams, University of California, San Diego. Richard Royall and W. G. Cumberland kindly provided the data sets used in chapter 7 on survey sampling. Several other data sets were brought to my attention by statisticians at the National Bureau of Standards, where I was fortunate to spend a year while on sabbatical. I deeply appreciate the patience, persistence, and faith of my editor, John Kimmel, in bringing this project to fruition.

The candid comments of many students and faculty who used the first edition of the book were influential in its revision. In particular I would like to thank Ian Abramson, Edward Bedrick, Jon Frank, Richard Gill, Roger Johnson, Torgny Lindvall, Michael Martin, Deb Nolan, Roger Pinkham, Yosi Rinott, Philip Stark, and Bin Yu; I apologize to any individuals who have inadvertently been left off this list. Finally, I would like to thank Alex Kugushev for his encouragement and support in carrying out the revision and the work done by Terri Bittner in carefully reading the manuscript for accuracy and in the solutions of the new problems. Any remaining errors are, of course, my responsibility.

Contents