Springer Texts in Statistics

Series Editors:
G. Casella
S. Fienberg
I. Olkin

Springer Texts in Statistics

Athreya/Lahiri: Measure Theory and Probability Theory
Bilodeau/Brenner: Theory of Multivariate Statistics
Brockwell/Davis: An Introduction to Time Series and Forecasting
Carmona: Statistical Analysis of Financial Data in S-PLUS
Chow/Teicher: Probability Theory: Independence, Interchangeability, Martingales, Third Edition
Christensen: Advanced Linear Modeling: Multivariate, Time Series, and Spatial Data; Nonparametric Regression and Response Surface Maximization, Second Edition
Christensen: Log-Linear Models and Logistic Regression, Second Edition
Christensen: Plane Answers to Complex Questions: The Theory of Linear Models, Second Edition
Davis: Statistical Methods for the Analysis of Repeated Measurements
Dean/Voss: Design and Analysis of Experiments
Dekking/Kraaikamp/Lopuhaä/Meester: A Modern Introduction to Probability and Statistics
Durrett: Essentials of Stochastic Processes
Edwards: Introduction to Graphical Modeling, Second Edition
Everitt: An R and S-PLUS Companion to Multivariate Analysis
Gentle: Matrix Algebra: Theory, Computations, and Applications in Statistics
Ghosh/Delampady/Samanta: An Introduction to Bayesian Analysis
Gut: Probability: A Graduate Course
Heiberger/Holland: Statistical Analysis and Data Display; An Intermediate Course with Examples in S-PLUS, R, and SAS
Jobson: Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design
Jobson: Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods
Karr: Probability
Kulkarni: Modeling, Analysis, Design, and Control of Stochastic Systems
Lange: Applied Probability
Lange: Optimization
Lehmann: Elements of Large Sample Theory
Lehmann/Romano: Testing Statistical Hypotheses, Third Edition
Lehmann/Casella: Theory of Point Estimation, Second Edition
Longford: Studying Human Populations: An Advanced Course in Statistics
Marin/Robert: Bayesian Core: A Practical Approach to Computational Bayesian Statistics
Nolan/Speed: Stat Labs: Mathematical Statistics Through Applications
Pitman: Probability
Rawlings/Pantula/Dickey: Applied Regression Analysis

(continued after index)

Springer Texts in Statistics

(continued from p. ii)

Robert: The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, Second Edition
Robert/Casella: Monte Carlo Statistical Methods, Second Edition
Rose/Smith: Mathematical Statistics with *Mathematica*
Ruppert: Statistics and Finance: An Introduction
Sen/Srivastava: Regression Analysis: Theory, Methods, and Applications
Shao: Mathematical Statistics, Second Edition
Shorack: Probability for Statisticians
Shumway/Stoffer: Time Series Analysis and Its Applications, Second Edition
Simonoff: Analyzing Categorical Data
Terrell: Mathematical Statistics: A Unified Introduction
Timm: Applied Multivariate Analysis
Toutenberg: Statistical Analysis of Designed Experiments, Second Edition
Wasserman: All of Nonparametric Statistics
Wasserman: All of Statistics: A Concise Course in Statistical Inference
Weiss: Modeling Longitudinal Data
Whittle: Probability via Expectation, Fourth Edition

Nicholas T. Longford

Studying Human Populations

An Advanced Course in Statistics

Springer

N.T. Longford
SNTL Statistics Research and Consulting, Reading
England

ISBN 978-0-387-98735-4 e-ISBN 978-0-387-73251-0

Library of Congress Control Number: 2007939169

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

To Shenki Xhadni and his sponsors

Preface

This monograph is for postgraduate students of statistics, statistical analysts, and other professionals who are interested in the design and analysis of studies in which responses are elicited from human subjects. Emphasis is placed on dealing with data that arise in imperfectly conducted studies. The reasons for imperfection include a sampling plan that cannot be implemented, measurement or elicitation of information by imperfect instruments, poor motivation of the subjects and their unwillingness to cooperate, and a multitude of other unavoidable shortcomings in relation to textbook-like settings that would be easy to analyse.

The subject of statistics is defined as making decisions in the presence of uncertainty. The context of a population and one or several variables defined for each member of this population is presented, and complete information is at first defined as having established the values of these variables for every member of the population. Making decisions with such complete information is regarded as a task outside the remit of statistics and is assumed to be a resolved problem or a problem for another profession. The raison d'être for statistics is that the available resources (time, manpower, expertise, funding, respondents' goodwill, and the like) are not sufficient for collecting the complete information.

With insufficient resources, we may establish the values of the variables for only some of the members of the population, and we may establish them imprecisely using imperfect instruments. Estimation is defined as forming a summary of the collected incomplete information (the data) with the purpose of getting as close as possible to the complete-information quantity of interest (the target). The quality of such a process (efficiency of the estimator) is described in frequentist terms by the mean squared error (MSE), defined by replications of the data-generating and estimation processes. Study design is defined generally as doing the best that can be done with the available resources. 'Doing the best' entails designing a study, implementing it (collecting the data), and estimating the target with the smallest possible MSE.

This scheme can be adapted for other forms of inference (such as confidence intervals and hypothesis tests) and measures of quality different from MSE.

The text assumes that the reader is familiar with the basics of statistics: the frequentist perspective; the definition of discrete and continuous distributions, including conditional and multivariate distributions; the concepts of independence, density, and distribution function; the common classes of distributions (normal and distributions derived from it, uniform, beta, gamma, binomial, and Poisson); sampling design and measurement process; the elementary statistical calculus (evaluating expectations and variances and fitting ordinary regression); and hypothesis testing and confidence intervals for some simple settings. This material is condensely presented in the Appendix, intended both for revision before reading Chapter 1 and for reference throughout the study. The exercises at the end of the Appendix are a suitable material for an entrance or revision exam.

Chapter 1 follows the standard curriculum of the analysis of variance and ordinary regression but parts company with the established solutions by adhering to the goals of efficient estimation and unbiased assessment of the efficiency. Chapter 2 introduces maximum likelihood as a general method of estimation, presents the basic results (without proofs), and discusses model selection and model uncertainty, issues broached in the previous chapter.

With limited resources, we can record the values of the relevant variables for only some members of the population and may have to do so imprecisely. These two forms of incompleteness lead to two general topics: survey sampling (Chapter 3) and measurement processes (Chapter 6). Between them, Chapter 4 introduces the Bayesian perspective as an alternative to the frequentist one, although it can be argued that there are three perspectives—model-based, design-based, and Bayesian, introduced in the respective Chapters 2, 3, and 4.

Chapter 5 returns to the frequentist perspective to discuss data incompleteness as a ubiquitous problem in implementing a design for studying a human population and introduces methods for dealing with missing data, data that we intended to collect but failed to. Complete information is defined here as the result of a perfectly implemented study design, a dataset that would be relatively easy to analyse. EM algorithm and multiple imputation are presented as two generic methods for dealing with incompleteness. Some other applications of these methods are outlined. In Chapter 6, imperfect measurement is presented as one of them.

Chapter 7 discusses experiments and observational studies and highlights the importance of the treatment-assignment process. Chapter 8 deals with clinical trials and presents them as a model example of experiments, emphasising the key role of their design, in the context of high ethical costs. Here, as well as in some earlier chapters, hypothesis testing is discussed, with the criticism that it fails to integrate information about the consequences (severity) of the two kinds of error that may be committed. Model selection criteria are subjected to similar criticism.

Chapters 9 and 10 discuss methods for multilevel and generalised linear models, respectively, as two indispensable elements of a statistician's analytical (computational) armoury. Chapter 11 deals with longitudinal and time-series analysis, treating them as applications of the methods presented in the previous two chapters.

Chapter 12 concludes with meta-analysis, a method for summarising the results of studies with a common or similar inferential agenda. The multivariate version of meta-analysis is discussed and connected to the problem of estimating one or several of a large number of interrelated quantities.

The chapters are designed so that they can be read or studied in order, with logical stopping points after Chapters 6, 8, and 10, which are followed by increasingly demanding material. They are intended as both a textbook for a semester, with some of the last few chapters optional, and a reference, with chapters as self-contained units. Chapters 1–8 can be covered in an academic quarter.

Several themes straddle the chapters. First among them is the view of nonstandard problems as involving missing data. That is, the problem at hand would be (more) tractable if some additional information were available. With the EM algorithm and multiple imputation, this is a natural approach to expanding the horizon of problems that we can deal with. Second is the pursuit of efficiency (small MSE) and of honesty (unbiased estimation of MSE) in estimation. Combining estimators (synthesis) is presented as an alternative to model selection, and their properties are compared in several settings, starting with the analysis of variance (ANOVA) in Chapter 1. Third is that we should be concerned with analysis of information, not merely analysis of one dataset at a time, and that study design is much more important than analysis. There is no reprieve for the deficiencies in the study design, whereas a reanalysis is a relatively inexpensive affair. The value of computing, for simulations in particular, and graphics, for effective data exploration and to summarise the results, is emphasised as a companion and, in some instances, an alternative, to (mathematical) analytical effort.

Background in elementary calculus and linear algebra is assumed, and experience in some statistical software, such as `R` [151] or `S-plus` [191], at an introductory level at least, is essential. In the spirit of object orientation, I tried to avoid subscripting whenever possible by defining suitable vectors and matrices. At a slower pace, the text could be combined with a course in `R` or other software for statistical analysis and graphics. Although all the computing and graphics was prepared in `R`, the text has very few references to `R`, and all the examples in the text, including simulations, can be reproduced with other software. The code used for the analyses and illustrations, mostly in the form of `R` functions, and the datasets for the exercises can be downloaded from `www.sntl.co.uk/BookA`.

Each chapter has a few references for further reading and more detailed study (for example, the monographs [168] for Chapter 3, [110] for Chapter 5, [113] for Chapter 9, [132] for Chapter 10, [37] for Chapter 11, and [72]

for Chapter 12) and 16–26 exercises, some directly connected to the text of the chapter and to its examples in particular. They range in difficulty and complexity from those for solving within a few minutes to open-ended problems suitable for projects for individual or small groups of students.

I have thought hard about the notation, whether to design rules that could be used consistently throughout the book or to adhere to the conventions that are consistent within narrow subject areas represented by the chapters but not across them. For example, capital letters are used for population quantities and lowercase for sample quantities in survey sampling, whereas in linear models capital letters are used for matrices and lowercase for vectors. I have settled for the prevailing conventions, with a few exceptions. As is common, I use the same notation for a random variable (estimator, dataset) and its realisation (estimate, realised dataset), but preface the latter by the term 'value of' whenever the two might be confused. In a few instances I simply ran out of suitable symbols or wanted to stick to established conventions and had to reuse some symbols. For example, β is used for both regression parameters and the power of a selection (or a test) in Chapter 2.

I could not avoid a few forward references in the text. None of them requires a detailed study of the section referred to, and when the section is reached later, the introduction made earlier is useful because the topic is not completely new. To smooth the text, I have set aside some mathematical niceties in favour of terms that are commonly used, but strictly speaking are not correct. Thus, by continuous distribution I mean throughout absolutely continuous distribution, and every one-to-one continuous function is assumed to be monotone.

I want to thank University Pompeu Fabra (UPF), Barcelona, Spain, and other institutions for opportunities to use draft chapters of this book in my lectures. I wrote and revised most of the manuscript in 2006 at UPF. I have benefited from attending the annual Applied Statistics Weeks organised by UPF and from eye-opening lectures by Don Rubin in particular. Support for this work by grants from the Spanish Ministry of Education and Science is acknowledged. Comments and encouragement from Anna Cuxart, Albert Satorra, and Frederic Udina, my colleagues at UPF, are acknowledged.

I had a fair number of false starts and postponed deadlines, and I want to commend Springer-Verlag for its near-asymptotic patience.

Reading, England *Nick Longford*
September 2007

Contents