

Springer Texts in Statistics

Series Editors:

G. Casella

S. Fienberg

I. Olkin

Springer Texts in Statistics

- Athreya/Lahiri*: Measure Theory and Probability Theory
Bilodeaul/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/Lopuhaäl/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

Editorial Board

George Casella
Department of Statistics
University of Florida
Gainesville, FL 32611-8545
USA

Stephen Fienberg
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890
USA

Ingram Olkin
Department of Statistics
Stanford University
Stanford, CA 94305
USA

ISBN 978-0-387-98735-4

e-ISBN 978-0-387-73251-0

Library of Congress Control Number: 2007939169

© 2008 Springer Science+Business Media, LLC

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY, 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

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
September 2007

Nick Longford

Contents

Preface	VII
1 ANOVA and Ordinary Regression	1
1.1 Analysis of Variance	1
1.1.1 Synthetic Estimation	4
1.2 Ordinary Regression	7
1.2.1 Prediction	12
1.3 Model Diagnostics	16
1.3.1 Simulation-Based Diagnostics	19
1.4 Toward Causal Inference	23
1.5 Designing Regression Studies	24
1.6 Observational Studies and Experiments	26
1.6.1 Human Subjects	28
1.6.2 Observational Studies and Estimation of Effects	29
Suggested Reading	31
Problems and Exercises	31
2 Maximum Likelihood Estimation	37
2.1 Likelihood	37
2.1.1 Consistency	40
2.1.2 Asymptotic Efficiency and Normality	41
2.2 Sufficient Statistics	44
2.3 Synthetic Estimation	48
2.4 Model Selection	51
2.4.1 Hypothesis Testing	54
2.4.2 Inference Following Model Selection	55
2.5 Model Selection Criteria Related to Likelihood	57
Suggested Reading	62
Problems and Exercises	62

3	Sampling Methods	67
3.1	Preliminaries	67
3.2	Horvitz–Thompson Estimator	69
	3.2.1 Simple Random Sampling	72
	3.2.2 Systematic Sampling Designs	73
	3.2.3 Some Other Sampling Designs	75
3.3	Stratification	76
3.4	Clustering	77
	3.4.1 Two-Stage Clustered Sampling	78
3.5	Planned and Realised Sampling Designs	80
	3.5.1 Adjusting and Trimming the Weights	81
3.6	Sample Size Calculation	85
3.7	Using Auxiliary Information	87
	3.7.1 Fitting Models to Survey Data	88
	3.7.2 Regression Estimators of the Population Mean	90
3.8	Small-Area Estimation	92
	Suggested Reading	96
	Problems and Exercises	97
4	The Bayesian Paradigm	103
4.1	The Updating Mechanism	103
	4.1.1 Setting the Prior	104
	4.1.2 Conjugate Priors	109
4.2	Computational Issues	109
	4.2.1 Sampling from Multivariate Distributions	112
4.3	Coherence	114
4.4	Bayesian Study Design	118
4.5	Model Diagnostics and Prediction	121
	Suggested Reading	123
	Problems and Exercises	124
5	Incomplete Data	129
5.1	Terminology and Notation	129
5.2	Dealing with Incomplete Data	135
5.3	Nonresponse Mechanisms	138
	5.3.1 Models for Incomplete Data	140
5.4	EM Algorithm	141
5.5	Multiple Imputation	144
	5.5.1 Modelling Nonresponse	146
	5.5.2 Working with Plausible Values	146
	5.5.3 Monotone Response Patterns and Chained Equations	148
	5.5.4 Sensitivity Analysis with Respect to NMAR	149
5.6	MI for Categorical and Longitudinal Data	150
5.7	Other Applications of the Missing-Data Idea	152
	5.7.1 Finite Mixtures	153

5.7.2	Outliers and Contaminants: Data Editing	155
5.7.3	Balanced Complete Data	156
	Suggested Reading	156
	Problems and Exercises	157
6	Imperfect Measurement	163
6.1	The Measurement Process	163
6.1.1	Information About the Measurement Process	167
6.2	Attributes of a Good Manifest Variable	168
6.2.1	Impartiality	168
6.3	Linear Regression with Manifest Variables	170
6.3.1	Latent Outcome Variable	170
6.3.2	Latent Regression Variable	172
6.3.3	Estimating the Latent Values	174
6.4	Categorical Manifest Variables	175
6.5	Measurement Error as Incompleteness	177
6.6	Simulation–Extrapolation	183
6.7	Coarse Data	186
6.7.1	Inference with Coarse Variables	187
6.7.2	Bootstrap	190
	Suggested Reading	194
	Problems and Exercises	195
7	Experiments and Observational Studies	201
7.1	Comparing Treatments	201
7.1.1	Experimental Design	203
7.1.2	Assignment and Sampling Designs	207
7.1.3	SUTVA Assumptions and Cluster-Randomisation	207
7.1.4	The Scale for Comparison	208
7.2	Block-Randomisation	209
7.2.1	When and How to Block	211
7.2.2	From Sample to Population	212
7.3	Observational Studies	213
7.3.1	Matching	218
7.4	Imperfect Experiments	220
7.5	Comparing Institutions	222
7.5.1	Comparing Performances over Time	225
7.6	Regression Methods	226
	Suggested Reading	227
	Problems and Exercises	227

8	Clinical Trials	233
8.1	The Context	233
8.2	Models and Inference	234
8.2.1	Comparing Two Groups	237
8.3	Crossover Design	239
8.3.1	Estimation	241
8.3.2	Minimax Estimation	243
8.4	Treatment Heterogeneity	246
8.5	Bioequivalence	247
8.6	Incorporating Utilities	253
8.7	Example	255
	Suggested Reading	260
	Problems and Exercises	261
9	Random Coefficients	265
9.1	Introduction	265
9.2	Patterns of Variation	267
9.2.1	Invariance with Respect to Linear Transformations ...	268
9.3	Maximum Likelihood Estimation	270
9.3.1	Restricted Maximum Likelihood	276
9.3.2	Residuals	277
9.3.3	Borrowing Strength	278
9.3.4	EM Algorithm—A Connection to Missing Data	279
9.3.5	Technical Details	280
9.4	Model Validity	282
9.5	Inference About Variation	285
9.5.1	Confounding in Cluster-Level Variation	286
9.6	Multilevel Models and Other Extensions	287
9.7	Estimating Many Quantities	290
9.8	Some Applications	292
9.8.1	Small-Area Estimation	292
9.8.2	Performance Assessment of Institutions	293
9.8.3	Progression over Time	293
9.8.4	Studying Families	294
	Suggested Reading	295
	Problems and Exercises	295
10	Generalised Linear Models	301
10.1	Introduction	301
10.2	Examples of GLMs	306
10.3	Maximum Likelihood Estimation	307
10.3.1	Overdispersion	309
10.3.2	Model Selection	311
10.4	Residuals	315
10.4.1	Overall Assessment of Fit	317

10.5	Random Coefficients	318
10.5.1	Fitting GLMM	321
10.5.2	Exact Derivatives of the Log-Likelihood	323
10.5.3	Laplace Approximation	324
10.5.4	Some Generalisations and Alternatives	325
	Suggested Reading	329
	Problems and Exercises	330
11	Longitudinal and Time-Series Analysis	335
11.1	Introduction	335
11.2	Markov Property	340
11.3	Time Series	342
11.3.1	Moving Average	343
11.4	Targets, Designs, and Models	344
11.4.1	Panels	348
11.5	Irregular Time Points and Time-Specific Covariates	349
11.6	Analysis	352
11.6.1	Missing Values	354
11.6.2	Multivariate and Cluster-Longitudinal Data	355
11.6.3	Mixture Models	356
11.7	Example: House Prices in New Zealand	359
11.7.1	Adjustment for Rating Valuation	364
	Suggested Reading	366
	Problems and Exercises	367
12	Meta-Analysis and Estimating Many Quantities	371
12.1	Introduction	371
12.2	Studies with Identical Targets	373
12.3	Study-Specific Targets	375
12.4	Maximum Likelihood Estimation	378
12.5	Publication Bias	380
12.5.1	Funnel Plot	381
12.6	Inference for a Particular Context	384
12.6.1	Prediction for an Unrealised Context	387
12.7	Multivariate Meta-Analysis	387
12.8	Estimating Many Quantities	389
	Suggested Reading	392
	Problems and Exercises	392
Appendix. A	Refresher	395
A.1	Populations and Variables	395
A.2	Replications and Randomness	399
A.2.1	Efficiency	400
A.3	Notation	402
A.4	Distributions	404

A.4.1	Describing Distributions	406
A.4.2	Approximating the Distribution by a Histogram	410
A.5	Sampling Design	411
A.5.1	Complex Sampling Designs	413
A.5.2	Sampling Frame	413
A.5.3	The Planned and Realised Sampling Processes	414
A.6	Measurement Processes	416
A.7	Infinite Populations	419
A.7.1	Continuous Distributions	421
A.7.2	Superpopulations: Models	423
A.8	Distributions	424
A.8.1	Simulations	427
A.9	Classes of Distributions and Models	428
A.10	Normal Distributions	429
A.10.1	Log-Normal Distributions	431
A.11	Uniform Distributions	432
A.12	Beta and Gamma Distributions	433
A.13	Classes of Discrete Distributions	434
A.13.1	Discrete Uniform Distributions	435
A.14	Discrete Bivariate Distributions	436
A.14.1	Conditional Distributions	437
A.15	Bivariate Continuous Distributions	438
A.16	Operating with Bivariate Distributions	440
A.17	Random Samples	443
A.18	Regression	444
A.19	Multivariate Distributions	445
A.19.1	Multivariate Normal Distributions	446
A.19.2	Regression with Normally Distributed Variables	447
A.20	Formulating Inferences	448
A.20.1	Confidence Intervals	449
	Problems and Exercises	451
	References	459
	Index	469