Springer Series in Statistics

Mark J. Schervish

Theory of Statistics

Springer

Mark J. Schervish

Theory of Statistics

With 26 Illustrations

Springer

Mark J. Schervish
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
USA

Library of Congress Cataloging-in-Publication Data
Schervish, Mark J.
Theory of Statistics / Mark J. Schervish
p. cm. – (Springer series in statistics)
Includes bibliographical references (p. -) and index.
ISBN 0-387-94546-6 (alk. paper)
1. Mathematical statistics. I. Title. II. Series.
QA276.S346 1995
519.5—dc20 95-11235

Printed on acid-free paper.

Production managed by Laura Carlson; manufacturing supervised by Joe Quatela.
Photocomposed pages prepared from the author's LaTeX files.
Printed and bound by Edwards Brothers, Inc., Ann Arbor, MI.
Printed in the United States of America.

9 8 7 6 5 4 3 2 (Corrected second printing, 1997)

ISBN 0-387-94546-6 Springer-Verlag New York Berlin Heidelberg SPIN 10552423

To Nancy, Margaret, and Meredith

Preface

This text has grown out of notes used for lectures in a course entitled *Advanced Statistical Theory* at Carnegie Mellon University over several years. The course (when taught by the author) has attempted to cover, in one academic year, those topics in estimation, testing, and large sample theory that are commonly taught to second year graduate students in a mathematically rigorous fashion. Most texts at this level fall into one of two categories. They either ignore the Bayesian point of view altogether or they cover Bayesian topics almost exclusively. This book covers topics in both classical[1] and Bayesian inference in a great deal of generality. My own point of view is Bayesian, but I believe that students need to learn both types of theory in order to achieve a fuller appreciation of the subject matter. Although many comparisons are made between classical and Bayesian methods, it is not a goal of the text to present a formal comparison of the two approaches as was done by Barnett (1982). Rather, the goal has been to prepare Ph.D. students to be able to understand and contribute to the literature of theoretical statistics with a broader perspective than would be achieved from a purely Bayesian or a purely classical course.

After a brief review of elementary statistical theory, the coverage of the subject matter begins with a detailed treatment of parametric statistical models as motivated by DeFinetti's representation theorem for exchangeable random variables (Chapter 1). In addition, Dirichlet processes and other tailfree processes are presented as examples of infinite-dimensional parameters. Chapter 2 introduces sufficient statistics from both Bayesian and non-Bayesian viewpoints. Exponential families are discussed here because of the important role sufficiency plays in these models. Also, the concept of information is introduced together with its relationship to sufficiency. A representation theorem is given for general distributions based on sufficient statistics. Decision theory is the subject of Chapter 3, which includes discussions of admissibility and minimaxity. Section 3.3 presents an axiomatic derivation of Bayesian decision theory, including the use of conditional probability. Chapter 4 covers hypothesis testing, including unbiased tests, P-values, and Bayes factors. We highlight the contrasts between the traditional "uniformly most powerful" (UMP) approach to testing and decision theoretic approaches (both Bayesian and classical). In particular, we

[1]What I call classical inference is called frequentist inference by some other authors.

see how the asymmetric treatment of hypotheses and alternatives in the UMP approach accounts for much of the difference. Point and set estimation are the topics of Chapter 5. This includes unbiased and maximum likelihood estimation as well as confidence, prediction, and tolerance sets. We also introduce robust estimation and the bootstrap. Equivariant decision rules are covered in Chapter 6. In Section 6.2.2, we debunk the common misconception of equivariant rules as means for preserving decisions under changes of measurement scale. Large sample theory is the subject of Chapter 7. This includes asymptotic properties of sample quantiles, maximum likelihood estimators, robust estimators, and posterior distributions. The last two chapters cover situations in which the random variables are not modeled as being exchangeable. Hierarchical models (Chapter 8) are useful for data arrays. Here, the parameters of the model can be modeled as exchangeable while the observables are only partially exchangeable. We introduce the popular computational tool known as Markov chain Monte Carlo, Gibbs sampling, or successive substitution sampling, which is very useful for fitting hierarchical models. Some topics in sequential analysis are presented in Chapter 9. These include classical tests, Bayesian decisions, confidence sets, and the issue of sampling to a foregone conclusion.

The presentation of material is intended to be very general and very precise. One of the goals of this book was to be the place where the proofs could be found for many of those theorems whose proofs were "beyond the scope of the course" in elementary or intermediate courses. For this reason, it is useful to rely on measure theoretic probability. Since many students have not studied measure theory and probability recently or at all, I have included appendices on measure theory (Appendix A) and probability theory (Appendix B).[2] Even those who have measure theory in their background can benefit from seeing these topics discussed briefly and working through some problems. At the beginnings of these two appendices, I have given overviews of the important definitions and results. These should serve as reminders for those who already know the material and as groundbreaking for those who do not. There are, however, some topics covered in Appendix B that are not part of traditional probability courses. In particular, there is the material in Section B.3.3 on conditional densities with respect to nonproduct measures. Also, there is Section B.6, which attempts to use the ideas of gambling to motivate the mathematical definition of probability. Since conditional independence and the law of total probability are so central to Bayesian predictive inference, readers may want to study the material in Sections B.3.4 and B.3.5 also.

Appendix C lists purely mathematical theorems that are used in the text

[2]These two appendices contain sufficient detail to serve as the basis for a full-semester (or more) course in measure and probability. They are included in this book to make it more self-contained for students who do not have a background in measure theory.

without proof, and Appendix D gives a brief summary of the distributions that are used throughout the text. An index is provided for notation and abbreviations that are used at a considerable distance from where they are defined. Throughout the book, I have added footnotes to those results that are of interest mainly through their value in proving other results. These footnotes indicate where the results are used explicitly elsewhere in the book. This is intended as an aid to instructors who wish to select which results to prove in detail and which to mention only in passing. A single numbering system is used within each chapter and includes theorems, lemmas, definitions, corollaries, propositions, assumptions, examples, tables, figures, and equations in order to make them easier to locate when needed.

I was reluctant to mark sections to indicate which ones could be skipped without interrupting the flow of the text because I was afraid that readers would interpret such markings as signs that the material was not important. However, because there may be too much material to cover, especially if the measure theory and probability appendices are covered, I have decided to mark two different kinds of sections whose material is used at most sparingly in other parts of the text. Those sections marked with a plus sign (+) make use of the theory of martingales. A lot of the material in some of these sections is used in other such sections, but the remainder of the text is relatively free of martingales. Martingales are particularly useful in proving limit theorems for conditional probabilities. The remaining sections that can be skipped or covered out of order without seriously interrupting the flow of material are marked with an asterisk (*). No such system is foolproof, however. For example, even though essentially all of the material dealing with equivariance is isolated in Chapter 6, there is one example in Chapter 7 and one exercise that make reference to the material. Similarly, the material from other sections marked with the asterisk may occasionally appear in examples later in the text. But these occurrences should be inconsequential. Of course, any instructor who feels that equivariance is an important topic should not be put off by the asterisk. In that same vein, students really ought to be made aware of what the main theorems in Section 3.3 say (Theorems 3.108 and 3.110), even though the section could be skipped without interrupting the flow of the material.

I would like to thank many people who helped me to write this book or who read early drafts. Many people have provided corrections and guidance for clarifying some of the discussions (not to mention corrections to some proofs). In particular, thanks are due to Chris Andrews, Bogdan Doytchinov, Petros Hadjicostas, Tao Jiang, Rob Kass, Agostino Nobile, Shingo Oue, and Thomas Short. Morris DeGroot helped me to understand what is really going on with equivariance. Teddy Seidenfeld introduced me to the axiomatic foundations of decision theory. Mel Novick introduced me to the writings of DeFinetti. Persi Diaconis and Bill Strawderman made valuable suggestions after reading drafts of the book, and those suggestions are incorporated here. Special thanks go to Larry Wasserman, who taught

from two early drafts of the text and provided invaluable feedback on the (lack of) clarity in various sections.

As a student at the University of Illinois at Urbana–Champaign, I learned statistical theory from Stephen Portnoy, Robert Wijsman, and Robert Bohrer (although some of these people may deny that fact after reading this book). Many of the proofs and results in this text bear startling resemblance to my notes taken as a student. Many, in turn, undoubtedly resemble works recorded in other places. Whenever I have essentially lifted, or cosmetically modified, or even only been deeply inspired by a published source, I have cited that source in the text. If results copied from my notes as a student or produced independently also resemble published results, I can only apologize for not having taken enough time to seek out the earliest published reference for every result and proof in the text. Similarly, the problems at the ends of each chapter have come from many sources. One source used often was the file of old qualifying exams from the Department of Statistics at Carnegie Mellon University. These problems, in turn, came from various sources unknown to me (even the ones I wrote). If I have used a problem without giving proper credit, please take it as a compliment. Some of the more challenging problems have been identified with an asterisk (*) after the problem number. Many of the plots in the text were produced using The New S Language and S–Plus [see Becker, Chambers, and Wilks (1988) and StatSci (1992)]. The original text processing was done using LaTeX, which was written by Lamport (1986) and was based on TeX by Knuth (1984).

Pittsburgh, Pennsylvania — MARK J. SCHERVISH
May 31, 1995

Several corrections needed to be made between the first and second printings of this book. During that time, I created a world-wide web page

`http://www.stat.cmu.edu/~mark/advt/`

on which readers may find up-to-date lists of any corrections that have been required. The most significant individual corrections made between the first and second printings are listed here:

- The discussion of the famous M-estimator on page 314 has been corrected.
- Theorems 7.108 and 7.116 each needed an additional condition concerning uniform boundedness of the derivatives of the H_n and H_n^* functions on a compact set. Only small changes were made to the proofs.
- The proofs of Theorems B.83 and B.133 were corrected, and small changes were made to Example 2.81 and Definition B.137.

Contents

---

*Sections and chapters marked with an asterisk may be skipped or covered out of order without interrupting the flow of ideas.

+Sections marked with a plus sign include results which rely on the theory of martingales. They may be skipped without interrupting the flow of ideas.

Chapter 1
Probability Models

1.1 Background

The purpose of this book is to cover important topics in the theory of statistics in a very thorough and general fashion. In this section, we will briefly review some of the basic theory of statistics with which many students are familiar. All that we do here will be repeated in a more precise manner at the appropriate place in the text.

1.1.1 General Concepts

Most paradigms for statistical inference make at least some use of the following structure. We suppose that some random variables $X_1, \ldots, X_n$ all have the same distribution, but we may be unwilling to say what that distribution is. Instead, we create a collection of distributions called a *parametric family* and denoted $\mathcal{P}_0$. For example, $\mathcal{P}_0$ might consist of all normal distributions, or just those normal distributions with variance 1, or all binomial distributions, or all Poisson distributions, and so forth. Each of these cases has the property that the collection of distributions can be indexed by a finite-dimensional real quantity, which is commonly called a *parameter*. For example, if the parametric family is all normal distributions, then the parameter can be denoted $\Theta = (\mathrm{M}, \Sigma)$, where M stands for the mean and Σ stands for the standard deviation. The set of all possible values of the parameter is called the *parameter space* and is often denoted by Ω. When $\Theta = \theta$, the distribution of the observations is denoted by P_θ. Expected values are denoted as $\mathrm{E}_\theta(\cdot)$.

We will denote observed data X. It might be that X is a vector of observations that are mutually independent and identically distributed (IID), or X might be some general quantity. The set of possible values for X is the *sample space* and is often denoted as $\mathcal{X}$. The members P_θ of the parametric family will be distributions over this space $\mathcal{X}$. If X is continuous or discrete, then densities or probability mass functions[1] exist. We will denote the density or mass function for P_θ by $f_{X|\Theta}(\cdot|\theta)$. For example, if X is a single random variable with continuous distribution, then

$$P_\theta(a < X \leq b) = \int_a^b f_{X|\Theta}(x|\theta)dx.$$

If $X = (X_1, \ldots, X_n)$, where the X_i are IID each with density (or mass function) $f_{X_1|\Theta}(\cdot|\theta)$ when $\Theta = \theta$, then

$$f_{X|\Theta}(x|\theta) = \prod_{i=1}^{n} f_{X_1|\Theta}(x_i|\theta), \tag{1.1}$$

where $x = (x_1, \ldots, x_n)$. After observing the data $X_1 = x_1, \ldots, X_n = x_n$, the function in (1.1), as a function of θ for fixed x, is called the *likelihood function*, denoted by $L(\theta)$. Section 1.3 is devoted to a motivation of the above structure based on the concept of *exchangeability* and DeFinetti's representation theorem 1.49. Exchangeability is discussed in detail in Section 1.2, and DeFinetti's theorem is the subject of Section 1.4.

1.1.2 Classical Statistics

Classical inferential techniques include tests of hypotheses, unbiased estimates, maximum likelihood estimates, confidence intervals and many other things. These will be covered in great detail in the text, but we remind the reader of a few of them here. Suppose that we are interested in whether or not the parameter lies in one portion Ω_H of the parameter space. We could then set up a *hypothesis* $H : \Theta \in \Omega_H$ with the corresponding *alternative* $A : \Theta \notin \Omega_H$. The simplest sort of *test* of this hypothesis would be to choose a subset $R \subseteq \mathcal{X}$, and then reject H if $x \in R$ is observed. The set R would be called the *rejection region* for the test. If $x \notin R$, we would say that we do not reject H. Tests are compared based on their power functions. The *power function* of a test with rejection region R is $\beta(\theta) = P_\theta(X \in R)$. The *size* of a test is $\sup_{\theta \in \Omega_H} \beta(\theta)$. Chapter 4 covers hypothesis testing in depth.

Example 1.2. Suppose that $X = (X_1, \ldots, X_n)$ and the X_i are IID with $N(\theta, 1)$ distribution under P_θ. The usual size α test of $H : \Theta = \theta_0$ versus $A : \Theta \neq \theta_0$ is

[1]Using the theory of measures (see Appendix A) we will be able to dispense with the distinction between densities and probability mass functions. They will both be special cases of a more general type of "density."

to reject H if $\overline{X} \in R$, where $\overline{X}$ is the sample average,

$$R = \left(-\infty, \theta_0 + \frac{1}{\sqrt{n}}\Phi^{-1}\left(\frac{\alpha}{2}\right)\right] \bigcup \left[\theta_0 + \frac{1}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right), \infty\right),$$

and Φ is the standard normal cumulative distribution function (CDF).

The notation and terminology in Chapter 4 are different from the above because we consider a more general class of tests called *randomized tests.* These are special cases of randomized decision rules, which are introduced in Chapter 3. The following example illustrates the reason that randomized decisions are introduced.

Example 1.3. Let $X \sim Bin(5, \theta)$ given $\Theta = \theta$. Suppose that we wish to test $H : \Theta \leq 1/2$ versus $A : \Theta > 1/2$. It might seem that the best test would be to reject H if $X > c$, where c is chosen to make the test have the desired level. Unfortunately, only six different levels are available for tests of this form. For example, if $c \in [4, 5)$, the test has level 1/32. If $c \in [3, 4)$, the test has level 3/16, and so on. If you desire a level such as 0.05, you must use a more complicated test.

A function of the data which takes values in the parameter space is called a *(point) estimator* of Θ. Section 5.1 considers point estimation in depth.

Example 1.4. Suppose that $X = (X_1, \ldots, X_n)$ and the X_i are IID with $N(\theta, 1)$ distribution under P_θ, then $\phi(x) = \sum_{i=1}^n x_i/n = \overline{x}$ takes values in the parameter space and can be considered an estimator of Θ.

Sometimes we wish to estimate a function g of Θ. An estimator ϕ of $g(\Theta)$ is *unbiased* if $\mathrm{E}_\theta[\phi(X)] = g(\theta)$ for all $\theta \in \Omega$. An estimator ϕ of Θ is a *maximum likelihood estimator (MLE)* if

$$\sup_{\theta \in \Omega} L(\theta) = L(\phi(x)),$$

for all $x \in \mathcal{X}$. An estimator ψ of $g(\Theta)$ is an MLE if $\psi(X) = g(\phi(X))$, where ϕ is an MLE of Θ. The reader should verify that the estimator ϕ in Example 1.4 is both an unbiased estimator and an MLE of Θ.

If the parameter Θ is real-valued, it is common to provide interval estimates of Θ. If (A, B) is a pair of random variables with $A \leq B$, and if

$$P_\theta(A \leq \theta \leq B) \geq \gamma,$$

for all $\theta \in \Omega$, then $[A, B]$ is called a *coefficient γ confidence interval* for Θ. Section 5.2 covers the theory of set estimation, which includes confidence intervals, prediction intervals, and tolerance intervals as special cases.

Example 1.5 (Continuation of Example 1.4). Suppose that $X = (X_1, \ldots, X_n)$ and the X_i are IID with $N(\theta, 1)$ distribution under P_θ, and let

$$A = \overline{X} - \frac{c}{\sqrt{n}}, \quad B = \overline{X} + \frac{c}{\sqrt{n}},$$

where $c > 0$. Then $[A, B]$ is a coefficient $2\Phi(-c)$ confidence interval for Θ, where Φ is the standard normal CDF.