The Theory of Statistical Inference

SHELEMYAHU ZACKS

The Theory of Statistical Inference

SHELEMYAHU ZACKS

Professor of Mathematics and Statistics
Case Western Reserve University
Cleveland, Ohio

JOHN WILEY & SONS, INC.

New York · London · Sydney · Toronto

Library of Congress Catalog Card Number: 77-132227

ISBN 0-471-98103-6

Printed in the United States of America

10 9 8 7 6 5 4 3 2

Preface

The theory of statistical inference has been developed during the last three decades to such an extent that even a great number of volumes will not suffice to cover every theorem and result of some importance. The library shelves are stacked every year with journals in which hundreds of papers present studies in the general field of statistical inference. Concurrently, many textbooks are published every year on the methods of statistical inference and on the basic theory. Generally these textbooks are written for the population of undergraduate and beginning graduate students. As a consequence of this general trend, there are only a few advanced books on the theory of statistical inference. In this book I attempt to provide the advanced graduate student and the researcher in mathematical statistics with some of the material that is in the literature but to a large extent has not yet been compiled in a book. To attain this objective I have chosen to write this book on an advanced level. The reader therefore has to be proficient in the theory of probability and the theory of statistical distribution functions and have some background in advanced stochastic processes for Chapters 9 and 10. In addition, general knowledge of advanced calculus, real variables, and functional analysis is helpful.

T' book has ten chapters. The first chapter gives a general view of the material in the book. The following nine chapters deal with the topics under consideration. These chapters are to a large degree self-contained. One can read each chapter almost independently of the others. Chapter 2, however, should be considered as a basic one, and is desirable to read the material in this chapter first. Each chapter contains a number of problems for solution. The problems are neither uniformly difficult nor easy, and not all of them are of high statistical interest. All the problems, however, reflect the material in the various sections. For the convenience of the reader, a list of the references cited is provided at the end of each chapter.

I began writing this book during the spring semester of 1967. At that time I was working in the Department of Statistics at Kansas State University. I would like to acknowledge Professor H. C. Fryer, the chairman of that department, for providing me with excellent working conditions and some secretarial help. My sincere gratitude is extended also to Professors J. R. Blum and L. H. Koopmans who, as chairman of the Mathematics and

Statistics Department at the University of New Mexico, provided me with the necessary help and support.

I am also grateful to Professor Geoffry S. Watson who introduced my book to John Wiley and Sons for publication; to Professor Debabrata Basu for some interesting discussions of various topics of statistical inference; to Miss Beatrice Shube, editor in the Wiley-Interscience Division; and to Mrs. Arlene Conkle for an excellent typing of the manuscript.

SHELEMYAHU ZACKS

Cleveland, Ohio
November, 1970

Commonly Used Abbreviations

a.s.	:	almost surely
i.i.d.	:	independent identically distributed
r.v.	:	random variable
d.f.	:	distribution function
s.t.	:	such that
m.g.f.	:	moment generating function
U.M.V.U.	:	uniformly minimum variance unbiased estimator
L.M.V.U.	:	locally minimum variance unbiased
L.U.E.	:	linear unbiased estimator
B.L.U.E.	:	best linear unbiased estimator
L.S.E.	:	least squares estimator
M.L.E.	:	maximum likelihood estimator
M.S.E.	:	mean square error
B.A.N.	:	best asymptotically normal
C.A.N.	:	consistent asymptotically normal
U.C.A.N.	:	uniformly consistent asymptotically normal
A.O.V.	:	analysis of variance
U.M.P.	:	uniformly most powerful (test)
U.M.P.U.	:	uniformly most powerful unbiased (test)
U.M.A.	:	uniformly most accurate (confidence intervals)
U.M.A.U.	:	uniformly most accurate unbiased (confidence intervals)
S.P.R.T.	:	sequential probability ratio test
M.L.R.	:	monotone likelihood ratio
I.F.R.A.	:	increasing failure rate average
I.F.R.	:	increasing failure rate
D.F.R.	:	decreasing failure rate
D.F.R.A.	:	decreasing failure rate average
L.C.	:	log-convex

Contents

CHAPTER 1

Synopsis

1.1. GENERAL INTRODUCTION

This book represents the author's attempt to summarize in an integrated form the major results in certain fields of the theory of statistical inference. There are virtually thousands of papers in the statistical literature on various topics of statistical inference, many of which have not yet been accounted for in textbooks. Results of many of these interesting papers are reported and presented here. We do not however, report all the studies which have been published throughout the years on the topics which we discuss. Only a relatively small number of papers are cited and studied in each chapter. The choice of which papers to discuss and what results should be presented is in itself a difficult decision. The author has probably committed the sin of overlooking some important results, or not reporting them. In the present book we concentrate on nine basic branches of the theory of statistical inference. These include the following: sufficient statistics; unbiased estimation; the efficiency of estimators under quadratic loss; maximum likelihood estimation; Bayes and minimax estimation; equivariant estimation; admissibility of estimators; testing statistical hypotheses; confidence and tolerance estimation. A chapter is devoted to each one of these nine subjects. In order to introduce the reader to the subject matter of each area, we provide introductory sections which summarize the material discussed in the corresponding chapters. These sections are designed to present a general view of what can be found in this book.

Important subject areas of statistical inference which are not discussed in this book are the following: nonparametric procedures of testing and estimation (only a few sections in the book deal with distribution free methods), and in particular we do not discuss the important area of robust procedures. The subject of sampling finite populations is discussed only briefly. There is no account of many of the recent important contributions to the theory of simultaneous testing of hypotheses and simultaneous confidence intervals estimation. (An introduction to this subject is found in the textbook by Miller [1].) Special multivariate techniques are not discussed, and in particular we do not treat the subjects of optimal design of experiments, optimal statistical control, or adaptive procedures, although in several chapters we do mention certain results from these areas. A reasonably complete account of all these areas would span several volumes. The subjects we treat in this book are, on the other hand, fundamental to all areas of statistical inference. Eight out of nine chapters of subject matter deal with problems of estimation, but only one chapter is devoted to testing statistical hypotheses. This fact should not be considered a reflection of the author's value judgment concerning the relative importance of the various areas of statistical inference. It merely reflects the fact that an excellent book is available on testing statistical hypotheses (Lehmann [2]), and we therefore present in one chapter only certain subjects that are either not discussed or only briefly mentioned in Lehmann's book. On the other hand, there is no book on the theory of statistical estimation that discusses the subjects mentioned above on the theoretical level of this book.

The optimality and admissibility of estimators and of test procedures are defined and discussed relative to a loss function, which expresses the "regret" for erroneous decisions in a quantitative manner. The types of loss functions used are the ones on which results are available in the literature. We do not discuss the question of the proper choice of a loss function. The same remark applies to the types of prior distributions applied in the various examples in this book. The choice of a prior distribution is not guided by any normative principle. We proceed now to a synopsis of the various chapters.

1.2. SUFFICIENT STATISTICS

The concept of sufficient statistics was introduced by Fisher [2] in 1922 and has been ever since a subject of many investigations. Chapter 2 is devoted to this important and basic notion of the theory of statistical inference. We start, in Section 2.1, with several examples of statistical models and their corresponding sufficient statistics. As originally indicated by Fisher, a statistic (a function of the observed sample values) is sufficient for the objectives of statistical inference if it contains, in a certain sense, all the "information" on the parent distribution (the distribution function according to which the sample values have been generated). In what sense do we use the word "information" here? We assume in the statistical model that the observed random variables have a certain joint distribution function which belongs to a specified family $\mathcal{F}$ of distribution functions. The actual (true) parent distribution is, however, unknown. Suppose that $T(X)$ is a statistic (a properly measurable function of the observed random variables X), such that the conditional expectation of *any* other statistic $Z(X)$, given $T(X)$, is independent of $\mathcal{F}$. $T(X)$ is called then a sufficient statistic for $\mathcal{F}$. If the sample value of $T(X)$ is known, the values of any other statistic $Z(X)$ do not add any further relevant statistical information on $\mathcal{F}$. Furthermore, we can say that a knowledge of the sample value of a sufficient statistic $T(X)$ provides the statistician with all the empirical information required for regenerating an equivalent sample. This can be done by the common methods of Monte Carlo simulation. Indeed, since the conditional joint distribution of the observed random variables $X_1, \ldots, X_n$, given $T(X)$, is independent of $\mathcal{F}$, we can generate for each value of $T(X)$ a sample $Y_1, \ldots, Y_n$ whose conditional distribution, given $T(X)$, is like that of $X_1, \ldots, X_n$. Thus one method of verifying whether $T(X)$ is a sufficient statistic is to determine the conditional distribution of $X_1, \ldots, X_n$, given $T(X)$. This method could very often be difficult and laborious.

Fisher [2] and later Neyman [1] provided a simple criterion with which we can generally determine whether a family $\mathcal{F}$ of distribution functions admits a nontrivial sufficient statistic, and what is the form of this statistic. This criterion is provided by the celebrated Neyman-Fisher factorization theorem. A rigorous measure-theoretic proof of the Neyman-Fisher factorization theorem was given only in 1949 by Halmos and Savage [1], and in some generalization by Bahadur [2]. We provide in Section 2.2 some preliminary measure-theoretic framework for the proof of the factorization theorem, that is presented in Section 2.3. Conditions are given for the existence of sufficient statistics. The usefulness of the theorem is illustrated with examples.

Section 2.4 is devoted to the subject of minimal sufficient statistics. The notion is explained first by a simple example, Example 2.6, in which we show the relevance of the correspondence between random variables, or statistics, and the sample space partitioning. The idea of minimal sufficient statistics is then explained in terms of the associated contours in the sample space, which contain the contours of all other sufficient statistics. A minimal sufficient statistic partitions the sample space to the smallest number of contours. The theory of Lehmann and Scheffé [1], published in 1950, is then discussed. This theory establishes a method of constructing the contours of minimal sufficient statistics (if exist). It also provides criteria for deciding whether a nontrivial sufficient statistic exists. [A trivial sufficient statistic is either the whole sample $(X_1, \ldots, X_n)$ or the corresponding order statistic $(X_{(1)}, \ldots, X_{(n)})$, $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$.] As we prove in Theorems 2.4.1 and 2.4.4, the Lehmann-Scheffé method of constructing minimal sufficient statistics is valid if the (dominated) family of probability measures is countable; or if the corresponding family of density functions is separable, in an $\mathcal{L}_1$-metric, $\delta(f, g) = \int |f(x) - g(x)|\, \mu(dx)$. These conditions are quite mild. [Whenever we have a parametric family of distribution functions for random variables (vectors) with a parameter space, which is an open set in a Euclidean k-space absolutely continuous with respect to the linear Lebesgue measure, the separability condition holds.] Bahadur [2] extended the above results and proved that whenever the family $\mathcal{F}$ of probability measures is dominated the log-likelihood function is a minimal sufficient statistic [see (2.4.26)]. This result has many interesting applications. In particular see Godambe [1] for a formulation of this result in the theory of sampling finite populations.

Minimal sufficient statistics are closely related to the exponential family of distribution functions. This relationship is studied in Section 2.5 where we present a theory developed by Dynkin [1] and published in 1951. This theory prevails only in cases of families of probability measures which are regular in Dynkin's special sense. In Dynkin's theory the log-likelihood function plays an important role. Each probability measure P in a regular family $\mathcal{F}$ is associated with a log-likelihood function $g(x, P)$, where the domain of the variable x is an interval Δ on the real line. We consider the linear space $\mathcal{L}(\mathcal{G}, \Delta)$ generated by the class $\mathcal{G} = \{g(x, P); P \in \mathcal{F}\}$. The dimension of this linear space gives the dimension of the minimal sufficient statistic. Whenever $\mathcal{L}(\mathcal{G}, \Delta)$ is an infinite dimensional space, the only sufficient statistic is a trivial one. On the other hand, if the dimension of $\mathcal{L}(\mathcal{G}, \Delta)$ is $r = k + 1$ and the sample size n is greater than k, the family of distribution functions is a k-parameter exponential (the Koopmans-Darmois) type. Several examples exhibit the main result. An interesting result of this theory is that whenever a family $\mathcal{F}$ of probability measures consists of mixtures of probability measures the only sufficient statistic is a trivial one (see Example 2.11). The implication of this is that sufficient statistics are very sensitive to the assumptions on which the statistical models are based. For example, if $\mathcal{F}$ consists of all probability measures corresponding to the normal $\mathcal{N}(\theta, 1)$ distributions, $-\infty < \theta < \infty$, a minimal sufficient statistic is $T_n = \sum_{i=1}^{n} X_i$. On the other hand, if $\mathcal{F}$ consists of probability measures with corresponding distribution functions $F_{\alpha,\theta}(x) = \alpha\Phi(x - \theta) + (1 - \alpha)\Phi(x)$, where α, $0 < \alpha < 1$, is known or unknown and $\Phi(x)$ is the $\mathcal{N}(0, 1)$ distribution function, then the only sufficient statistic is a trivial one. This shows that even if α is very close to 1, but still smaller than 1, we cannot reduce the data without loss of information. From a pragmatic point of view, however, if α is very close to 1 so that probability measure of Borel sets B contributed by $\bar{\alpha}\,\Phi(x)$ is negligible, we could approximate the mixtures $F_{\alpha,\theta}(x)$ by $F(x - \theta)$ and use the sufficient statistic T_n. Another implication from the main result of Section 2.5 is that in many statistical models that are widely used in applications (e.g., those with the Wiebull distributions, various types of the Pearson system, Laplace distributions with unknown location parameters) the only sufficient statistics are the trivial ones. This is a source of many complications in estimation and in testing of hypotheses, which will be further discussed later on.

Section 2.6 deals with completeness and sufficiency. The properties of completeness and sufficiency are of fundamental importance in the theory of uniformly minimum risk unbiased estimation, which is discussed in Chapter 3. Several theorems are proven in this section concerning the conditions under which families of the induced distribution functions of sufficient statistics are complete. Several examples are provided.

In Section 2.7 we present the theory of sufficiency and invariance. The results of this theory play a basic role in the theory of minimax estimation (Chapter 6), equivariant estimation (Chapter 7), invariant testing of hypotheses (Chapter 9), and so forth. The main problem studied in this section is the following: Under what conditions does the maximal invariant reduction of a sufficient statistic, with respect to a group of transformations $\mathcal{G}$, yield an equivalent statistic to the one obtained by a sufficiency reduction of a maximal invariant statistic? The conditions are stated in a theorem of Stein (Theorem 2.7.1), proven recently by Hall, Wijsman, and Ghosh [1]. The theory of sufficiency and invariance can be used in many cases to prove that certain statistics are independent. Among the problems of Section 2.11 are several that can be easily solved by using the results of this section.

Section 2.8 concerns the property of transitivity of a sequence of sufficient statistics. The results of this section have important applications in the theory of sequential estimation and sequential testing of hypotheses.

Section 2.9 is devoted to the notion of sufficient experiments, introduced by Blackwell [1] in 1951. This notion has various applications in the theory of optimum design of experiments. Such applications have been shown by DeGroot [3].

Section 2.10 discusses the role of sufficient statistics in Bayesian analysis. The results of this section are well known and appear in various books and papers. In particular we mention the book of Raiffa and Schlaifer [1] and the interesting application of the notion of Bayes sufficiency in the theory of sampling finite populations by Godambe [3].

1.3. UNBIASED ESTIMATION

Chapter 3 is devoted to the problem of unbiased estimation of a functional $g: \mathcal{P} \rightarrow E^{(k)}$, where $E^{(k)}$, $k \geqslant 1$, is the Euclidean k-space. In other words, a function g, whose domain is a specified family $\mathcal{P}$ of probability measures, is considered. The family $\mathcal{P}$ could be a parametric or a nonparametric one, and the function g ascribes each P of $\mathcal{P}$ a real or a vector valued parameter $g(P)$ (e.g., the first k moments of P, if exist). The objective is to estimate the point $g(P)$ on the basis of the observed random variables $X_1, X_2, \ldots$. The class of all estimators $\hat{g}(X)$, which are statistics on $\mathfrak{X} = \mathfrak{X}_1 \times \mathfrak{X}_2 \times \cdots$, mapping $\mathfrak{X}$ to the range of $g(P)$, say Θ, and satisfying

$$E_P\{\hat{g}(X)\} = g(P), \qquad \text{all } P \in \mathfrak{I},$$

is called the *class of unbiased estimators of g, relative to* $\mathfrak{I}$.

Lehmann [2], p. 11, defines an unbiased point estimator in a more general manner. He considers a loss function $L(g(P), \hat{g}(X))$ and says that $\hat{g}$ is an unbiased estimator if,

$$E_P\{L(g(P), \hat{g}(X))\} \leq E_P\{L(g(P), \tilde{g}(X))\}$$

for all $P \in \mathfrak{I}$, where $\tilde{g}(X)$ is any other estimator of $g(P)$. It is easy to verify that the two definitions coincide when the loss function $L(g, \hat{g})$ is quadratic, that is, in the real case $L(g, \hat{g}) = \lambda(g)(\hat{g} - g)^2$, $0 < \lambda(g) < \infty$. In this book we adopted a more common approach, in which an unbiased estimator is any estimator whose expectation, under P, is equal to $g(P)$ *for all* $P \in \mathfrak{I}$. This definition is independent of any particular choice of a loss function $L(g, \hat{g})$. Moreover, if we adopt Lehmann's more general definition we may find, in most cases of interest, that an unbiased estimator does not exist. In this book we call an estimator $\hat{g}$ that minimizes the risk function $E_P\{L(g(P), \hat{g}(X))\}$, at all $P \in \mathfrak{I}$, a *uniformly minimum risk estimator.*

As will be shown in examples (see in particular Chapter 7) uniformly minimum risk estimators (if exist) are not necessarily unbiased in the classical sense. An estimator $\hat{g}(X)$ is called a *uniformly minimum risk unbiased* if it minimizes the risk $E_P\{L(g(P), \hat{g}(X))\}$ uniformly in $P \in \mathfrak{I}$, with respect to the class of *unbiased* estimators only. As mentioned above, uniformly minimum risk unbiased estimators do not coincide, necessarily, with uniformly minimum risk estimators.

In Section 3.1 we formulate the general theory of uniformly minimum risk unbiased estimators for quadratic or convex loss functions when the family $\mathfrak{I}$ of probability measures is a parametric one, that is, when $\mathfrak{I} = \{P_\theta; \theta \in \Theta\}$ where Θ is some open interval (rectangle) in a k-dimensional Euclidean space $E^{(k)}$. We prove the celebrated Blackwell-Rao-Lehmann-Scheffé theorem in a general framework for quadratic loss functions of the form

$$L(g, \hat{g}) = (\hat{g} - g)'\mathbf{A}(\hat{g} - g),$$

where $\mathbf{A}$ is a $k \times k$ positive definite matrix. The main result is then generalized to cases with any loss function $L(g, \hat{g})$, which is convex in $\hat{g}$ for each g. As shown in this section, the existence of a uniformly minimum risk unbiased estimator depends to a large extent on whether the family $\mathfrak{I}$ admits a complete sufficient statistic.

Section 3.2 presents several examples of the method of deriving uniformly minimum risk unbiased estimators in various models. In these examples we restrict attention to real parameter cases with squared-error loss functions, that is, $L(g, \hat{g}) = (\hat{g} - g)^2$. Since $\hat{g}$ is unbiased, the risk function $E_P\{(\hat{g}(X) - g(P))^2\}$ is the variance of $\hat{g}(X)$. Thus the best unbiased estimators for this loss function are those that are uniformly minimum variance unbiased (U.M.V.U.).

In Section 3.3 we discuss locally minimum variance unbiased (L.M.V.U.) estimators for cases in which U.M.V.U. estimators do not exist. An unbiased estimator $\hat{g}(X, \theta_0)$ is a minimum variance unbiased estimator at $\theta = \theta_0$ if $\text{Var}_{\theta_0}\{\hat{g}(X; \theta_0)\} \leq \text{Var}_{\theta_0}\{\tilde{g}(X)\}$, where $\tilde{g}(X)$ is any unbiased estimator of $g(\theta)$. We restrict attention to parametric cases. A more general definition is given for cases of vector valued parameters $g(\theta)$. We start with a theorem that establishes a necessary and sufficient condition for an unbiased estimator $\hat{g}$ to be L.M.V.U. at $\theta = \theta_0$. Many examples are available in the literature of statistical models for which no U.M.V.U. estimators exist but one can construct L.M.V.U. estimators that attain a zero variance at one or more values of θ. We show such an example (Example 3.9) from Sethuraman [1], in which an L.M.V.U. estimator attains a zero variance at an infinite sequence of θ values. The question raised is whether we can provide a general method of constructing L.M.V.U. estimators that attain a zero variance at one or more points of θ. We present such a method, apparently attributable to Takeuchi and published by Morimoto and Sibuya [1]. This method is limited to models of absolutely continuous distribution functions whose support is an open interval $(\theta, b(\theta))$, where $b(\theta)$ is a differentiable function. By the method presented we can construct the required estimator in a recursive manner on intervals of Θ. Example 3.8 illustrates this method. In Section 3.3 we also provide a theorem of Stein [3] that establishes necessary and sufficient conditions, in terms of the likelihood ratio function, for an estimator $\hat{g}$ to be L.M.V.U. at $\theta = \theta_0$. The theorem is restricted to dominated families of probability measures, with (generalized) density functions satisfying certain regularity conditions in terms of some linear operators on an L^2-space. In these cases we show that Stein's theorem is a generalization of the previous theorem of this section. Whenever the theorem is applicable it also yields the variance of the L.M.V.U. estimator, in terms of the linear operators used.

Kitagawa [1] provided a general theory of linear translatable operators on function spaces that can yield in certain statistical models U.M.V.U. estimators. Washio, Morimoto, and Ikeda [1] applied the theory of Kitagawa to the case of 1-parameter exponential families of distribution functions. We know that in this case a U.M.V.U. estimator exists, since a minimal sufficient statistic is complete. The main theorem (Theorem 3.4.2) from