James O. Berger

Statistical Decision Theory

Foundations, Concepts, and Methods

James O. Berger

Statistical Decision Theory

Foundations, Concepts, and Methods

Springer-Verlag
New York Heidelberg Berlin

James O. Berger
Department of Statistics
Purdue University
West Lafayette, Indiana 47907
USA

AMS Classification: 60CXX

With 20 Figures.

Library of Congress Cataloging in Publication Data

Berger, James.
Statistical decision theory, foundations, concepts, and methods.

(Springer series in statistics)
Bibliography: p.
Includes index.
1. Statistical decision. 2. Bayesian
statistical decision theory. I. Title. II. Series.
QA279.4.B47 1980 519.5′42 79-27052

Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-90471-9 Springer-Verlag New York
ISBN 3-540-90471-9 Springer-Verlag Berlin Heidelberg

Preface

Decision theory is generally taught in one of two very different ways. When taught by theoretical statisticians, it tends to be presented as a set of optimality principles, together with a collection of mathematical techniques useful in establishing the optimality of various statistical procedures. When taught by applied decision theorists, it is usually a course in Bayesian analysis, showing how this one decision principle can be applied in various practical situations. The original goal I had in writing this book was to find some middle ground. I wanted a book which discussed the more theoretical ideas and techniques of decision theory, but in a manner that was constantly oriented towards solving statistical problems. In particular, it seemed crucial to include a discussion of when and why the various decision principles should be used, and indeed why decision theory is needed at all.

This original goal seemed indicated by my philosophical position at the time, which can best be described as basically neutral. I felt that no one approach to decision theory (or statistics) was clearly superior to the others, and so planned a rather low key and impartial presentation of the competing ideas. In the course of writing the book, however, I turned into a rabid Bayesian. There was no single cause for this conversion; just a gradual realization that things seemed to ultimately make sense only when looked at from the Bayesian viewpoint. Specific considerations that I found particularly compelling were: (i) The Bayesian measures of accuracy of a conclusion seem more realistic than the classical measures (the final precision versus initial precision argument). (ii) In most circumstances any reasonable statistical procedure corresponds to a Bayes procedure with respect to some prior distribution, and if this prior distribution is quite unrealistic, then use of the statistical procedure seems suspect. (iii) Principles of rational behavior seem to imply that one must act as if he had a prior distribution.

My original reservations about Bayesian analysis included the concerns that the analysis can give very bad results if an incorrect prior distribution is chosen, and that Bayesian conclusions lack objectivity. The first of these concerns I still have, but the second has been replaced by the belief that only through the Bayesian viewpoint can objectivity, if desired, be achieved. The justification for this belief is again the second point mentioned earlier, that virtually any reasonable procedure will be a Bayes procedure with respect to some prior distribution (or at least some positive measure). When this is the case, it seems unreasonable to claim that a proposed procedure is objective unless it corresponds to some "objective prior distribution."

These remarks are by no means meant to be conclusive (or even understandable) arguments in themselves. Instead they are meant to convey something of the basic philosophy of the book. The first four chapters of the book describe this basic philosophy, and present the Bayesian approach to statistics. Included are often neglected subjects of importance in the implementation of the Bayesian approach, such as the construction of loss functions and the quantification of prior information. This material is a blend of arguments justifying the Bayesian viewpoint (including criticisms of classical statistical methods), presentations of actual Bayesian techniques (including the oft-maligned but usually quite reasonable use of noninformative prior distributions), and discussions of the dangers in using the Bayesian approach. In particular, a "robust" Bayesian approach is advocated, in which one attempts to protect against misspecification of the prior distribution.

The remaining four chapters of the book turn to other aspects of decision theory. Chapters 5 and 6 present the minimax and invariance approaches to decision theory. These approaches are not only interesting theoretically, but can also be of considerable aid in the analysis of decision problems, particularly when there is no, or only very vague, prior information. It is shown, however, that use of these methods with no regard for Bayesian considerations can be very bad. Chapter 7 discusses (from a decision-theoretic viewpoint) the important topic of sequential analysis. Chapter 8 introduces the more theoretical side of decision theory, with a discussion of complete class theorems.

The book is meant to be an applied book on decision theory. Theoretical results are, for the most part, presented only to the extent that they can be usefully applied to real statistical problems, and pains are taken to show how to apply any such theoretical techniques. For this reason, the mathematical and statistical level of the book is kept as low as possible. It is assumed that the reader knows calculus and some probability (at least, say, what expectations and conditional probability distributions are), and is familiar with certain basic statistical concepts such as estimation, testing, and (to a degree) sufficiency. From time to time we will pay lip service to things like measurability, and (especially in later chapters) may employ some higher mathematical facts. Knowledge of such things is not needed, however, to follow and understand the basic material.

To avoid mathematical complications, the proofs of some theorems are not given, and indeed some theorems are not even stated completely rigorously. (Such theorems are usually called "Results.") It is perhaps advisable for the reader with no or little previous exposure to decision theory to omit some of the more delicate philosophical issues and examples on a first reading, and return to them after obtaining a fairly firm grasp of the essentials.

In terms of teaching, the book is designed to be usable as a text at two different graduate levels. First, it can be used for applied statistical students (say applied masters students) as a basic introduction into the use and methodology of decision theory. As such, its purpose is also to describe the inadequacies of certain classical statistical procedures, and to instill a concern for correct formulation of a statistical problem.

At the second level, the book is intended to provide a combined applied and theoretical introduction to decision theory for doctoral students. Such students may need not only the basic methodology, but also the ability to investigate the decision-theoretic optimality properties of statistical procedures.

The book is organized so as to be easy to use at either of the above levels. Listed below are various suggested programs at each level. The indicated numbers are chapter or section numbers.

Level 1 (Applied)

1 Quarter: 1 (except 1.4, 1.7, 1.8), 2, 3 (except 3.2.3, 3.2.4, 3.2.5), 4 (except 4.4.5, 4.5, 4.6.2, 4.6.5), selected parts of 7.

1 Semester: 1 (except 1.4, 1.7, 1.8), 2, 3 (except 3.2.3), 4, 7 (except 7.4.9, 7.4.10, 7.5.4, 7.5.5, 7.6).

1 Semester (including minimax): 1 (except 1.7, 1.8), 2, 3 (except 3.2.3, 3.2.4, 3.2.5), 4 (except 4.4.5, 4.5, 4.6.2, 4.6.5), 5 (except 5.2.3, 5.2.4, 5.2.5, 5.2.6, 5.3.3), 7 (except 7.2, 7.4.7, 7.4.8, 7.4.9, 7.4.10, 7.5.4, 7.5.5, 7.6).

2 Quarters: 1 (except 1.7, 1.8), 2, 3, 4, 5 (except 5.2.5, 5.2.6), 7 (except 7.4.9, 7.4.10).

Level 2 (More Theoretical)

1 Quarter: 1, 2 (except 2.3), 3 (except 3.2.3, 3.2.4, 3.2.5), 4 (except 4.4.5, 4.6.2, 4.6.4, 4.6.5), 5 (except 5.2.3, 5.4.5), selected parts of 7.

1 Semester: 1, 2, 3 (except 3.2.3, 3.2.4, 3.2.5), 4 (except 4.4.5, 4.6.4), 5 (except 5.2.3, 5.4.5), 7 (except 7.4.9, 7.4.10), selected parts of 6 and 8.

2 Quarters: 1, 2, 3 (except 3.2.3, 3.2.4, 3.2.5), 4 (except 4.6.4), 5 (except 5.2.3), 6 (except 6.6), 7 (except 7.4.9, 7.4.10), selected parts of 8.

3 Quarters or 2 Semesters: Entire book.

When writing a book containing many intuitive ideas about statistics, there is always a problem in acknowledging the originators of the ideas. Many ideas become part of the folklore or have been independently rediscovered by a number of people. An attempt was made to acknowledge the originator of an idea or technique whenever the originator was known to me. I apologize for the undoubtedly numerous omissions. Particularly galling will no doubt be those ideas I seem to implicitly claim to have originated, which in reality were developed years ago by a score of people. Besides my apologies, all I can say is you should have written a book.

I am very grateful to a number of people who contributed in one way or another to the existence of this book. Lawrence Brown, Roger Farrell, Leon Gleser, Jack Kiefer, Herman Rubin, and William Studden were particularly helpful. Larry Brown, with penetrating insight, pointed out some logical inconsistencies in an earlier version of the book. Roger Farrell and Leon Gleser shared their thoughts in many interesting discussions. Jack Kiefer was kind enough to read and comment upon certain portions of the manuscript. Bill Studden provided very helpful critical comments for the first five chapters, and discovered a major blunder. Finally, Herman Rubin served as a constant source of aid and inspiration throughout the writing of the book. He helped me over many rough spots and tried his best to keep me from saying foolish things. Of course, none of the above people necessarily agree with what is presented in the book. (Indeed only Herman Rubin claims to be a Bayesian.) I feel that our common ground is substantial, however.

I would also like to thank Lou Anne Scott, Norma Lucas, Kathy Woods and Carolyn Knutsen for their excellent typing of the manuscript. Several graduate students, particularly T. C. Kao and Don Wallace, were also very helpful, pointing out many errors and checking a large portion of the exercises. Finally, I would like to express my appreciation to the John Simon Guggenheim Memorial Foundation and the Alfred P. Sloan Foundation. Much of the book was written during the tenure of a Guggenheim Fellowship, and parts during the tenure of a Sloan Fellowship.

West Lafayette, Indiana
March 1980

JAMES BERGER

Contents

CHAPTER 1

Basic Concepts

1.1 Introduction

Decision theory, as the name implies, is concerned with the problem of making decisions. Statistical decision theory is concerned with the making of decisions in the presence of statistical knowledge which sheds light on some of the uncertainties involved in the decision problem. We will, for the most part, assume that these uncertainties can be considered to be unknown numerical quantities, and will represent them by θ (possibly a vector or matrix).

As an example, consider the situation of a drug company deciding whether or not to market a new pain reliever. Two of the many factors affecting its decision are the proportion of people for which the drug will prove effective (θ_1), and the proportion of the market the drug will capture (θ_2). Both θ_1 and θ_2 will be generally unknown, though typically experiments can be conducted to obtain statistical information about them. This problem is one of decision theory in that the ultimate purpose is to decide whether or not to market the drug, how much to market, what price to charge, etc.

Classical statistics is directed towards the use of sample information (the data arising from the statistical investigation) in making inferences about θ. These classical inferences are, for the most part, made without regard to the use to which they are to be put. In decision theory, on the other hand, an attempt is made to combine the sample information with other relevant aspects of the problem in order to make the best decision.

In addition to the sample information, two other types of information are typically relevant. The first is a knowledge of the possible consequences of the decisions. Often this knowledge can be quantified by determining the loss that would be incurred for each possible decision and for the various possible values of θ. (Statisticians seem to be pessimistic creatures who think in terms of losses. Decision theorists in economics and business talk instead in terms of gains (utility). As our orientation will be mainly statistical, we will use the loss function terminology. Note that a gain is just a negative loss, so there is no real difference between the two approaches.)

The incorporation of a loss function into statistical analysis was first studied extensively by Abraham Wald. Indeed he can be considered to be the founder of statistical decision theory.

In the drug example, the losses involved in deciding whether or not to market the drug will be complicated functions of θ_1, θ_2, and many other factors. A somewhat simpler situation to consider is that of estimating θ_1, for use, say, in an advertising campaign. The loss in underestimating θ_1 arises from making the product appear worse than it really is (adversely affecting sales), while the loss in overestimating θ_1 would be based on the risks of possible penalties for misleading advertising.

The second source of nonsample information that is useful to consider is called prior information. This is information about θ arising from sources other than the statistical investigation. Generally, prior information comes from past experience about similar situations involving similar θ. In the drug example, for instance, there is probably a great deal of information available about θ_1 and θ_2 from different but similar pain relievers.

A compelling example of the possible importance of prior information was given by L. J. Savage (1961). He considered the following three statistical experiments.

1. A lady, who adds milk to her tea, claims to be able to tell whether the tea or the milk was poured into the cup first. In all of ten trials conducted to test this, she correctly determines which was poured first.
2. A music expert claims to be able to distinguish a page of Haydn score from a page of Mozart score. In ten trials conducted to test this, he makes a correct determination each time.
3. A drunken friend says he can predict the outcome of a flip of a fair coin. In ten trials conducted to test this, he is correct each time.

In all three situations, the unknown quantity θ is the probability of the person answering correctly. A classical significance test of the various claims would consider the null hypothesis (H_0) that $\theta = 0.5$ (i.e., the person is guessing). In all three situations this hypothesis would be rejected with a (one-tailed) significance level of 2^{-10}. Thus the above experiments give strong evidence that the various claims are valid.

In situation 2 we would have no reason to doubt this conclusion. (The outcome is quite plausible with respect to our prior beliefs.) In situation 3, however, our prior opinion that this prediction is impossible (barring a belief in extrasensory perception) would tend to cause us to ignore the experimental evidence as being a lucky streak. In situation 1 it is not quite clear what to think, and different people will draw different conclusions according to their prior beliefs of the plausibility of the claim. In these three identical statistical situations, prior information clearly cannot be ignored.

1.2 Basic Elements

The unknown quantity θ which affects the decision process is commonly called the *state of nature*. In making decisions it is clearly important to consider what the possible states of nature are. The symbol Θ will be used to denote the set of all possible states of nature. Typically, when experiments are performed to obtain information about θ, the experiments are designed so that the observations are distributed according to some probability distribution which has θ as an unknown parameter. In such situations θ will be called the *parameter* and Θ the *parameter space*.

Decisions are more commonly called *actions* in the literature. Particular actions will be denoted by a, while the set of all possible actions under consideration will be denoted $\mathscr{A}$.

As mentioned in the introduction, a key element of decision theory is the loss function. If a particular action a_1 is taken and θ_1 turns out to be the true state of nature, then a loss $L(\theta_1, a_1)$ will be incurred. Thus we will assume a *loss function* $L(\theta, a)$ is defined for all $(\theta, a) \in \Theta \times \mathscr{A}$. For technical convenience, only loss functions satisfying $L(\theta, a) \geq -K > -\infty$ will be considered. This condition is satisfied by all loss functions of interest. Chapter 2 will be concerned with showing why a loss function will typically exist in a decision problem, and with indicating how a loss function can be determined.

When a statistical investigation is performed to obtain information about θ, the outcome (a random variable) will be denoted X. Often X will be a vector, as when $X = (X_1, X_2, \ldots, X_n)$, the X_i being independent observations from a common distribution. (From now on vectors will appear in boldface type; thus $\mathbf{X}$.) A particular realization of X will be denoted x. The set of possible outcomes is the *sample space*, and will be denoted $\mathscr{X}$. (Usually $\mathscr{X}$ will be a subset of R^n, n-dimensional Euclidean space.)

The probability distribution of X will, of course, depend upon the unknown state of nature θ. Let $P_\theta(A)$ or $P_\theta(X \in A)$ denote the probability of the event A ($A \subset \mathscr{X}$), when θ is the true state of nature. For simplicity, X will be assumed to be either a continuous or a discrete random variable, with density $f(x|\theta)$. Thus if X is continuous (i.e., has a density with respect to Lebesgue measure), then

$$P_\theta(A) = \int_A f(x|\theta)dx,$$

while if X is discrete, then

$$P_\theta(A) = \sum_{x \in A} f(x|\theta).$$

Certain common probability densities and their relevant properties are given in Appendix 1.

It will frequently be necessary to consider expectations over random variables. The expectation (over X) of a function $h(x)$, for a given value of θ, is defined to be

$$E_\theta[h(X)] = \begin{cases} \int_{\mathscr{X}} h(x) f(x \mid \theta) dx & \text{(continuous case),} \\ \sum_{x \in \mathscr{X}} h(x) f(x \mid \theta) & \text{(discrete case).} \end{cases}$$

It would be cumbersome to have to deal separately with these two different expressions for $E_\theta[h(X)]$. Therefore, as a convenience, we will define

$$E_\theta[h(X)] = \int_{\mathscr{X}} h(x) dF^X(x \mid \theta),$$

where the right-hand side is to be interpreted as in the earlier expression for $E_\theta[h(X)]$. (This integral can, of course, be considered a Riemann–Stieltjes integral, where $F^X(x \mid \theta)$ is the cumulative distribution function of X. Readers not familiar with such terms can just treat the integral as a notational device.) Note that, in the same way, we can write

$$P_\theta(A) = \int_A dF^X(x \mid \theta).$$

Frequently, it will be necessary to clarify the random variables over which an expectation or probability is being taken. Superscripts on E or P will serve this role. (A superscript could be the random variable, its density, its distribution function, or its probability measure, whichever is more convenient.) Subscripts on E will denote parameter values at which the expectation is to be taken. When obvious, subscripts or superscripts will be omitted.

The third type of information discussed in the introduction was prior information concerning θ. A useful way of talking about prior information is in terms of a probability distribution on Θ. (Prior information about θ is seldom very precise. Therefore, it is rather natural to state prior beliefs in terms of probabilities of various possible values of θ being true.) The symbol $\pi(\theta)$ will be used to represent a prior density of θ (again for either the continuous or discrete case). Thus if $A \subset \Theta$,

$$P(\theta \in A) = \int_A dF^\pi(\theta) = \begin{cases} \int_A \pi(\theta) d\theta & \text{(continuous case),} \\ \sum_{\theta \in A} \pi(\theta) & \text{(discrete case).} \end{cases}$$

Chapter 3 discusses the construction of prior probability distributions, and also indicates what is meant by probabilities concerning θ. (After all, in most situations there is nothing "random" about θ. A typical example is when θ is an unknown but fixed physical constant (say the speed of light) which is to be determined. The basic idea is that probability statements concerning θ are then to be interpreted as "personal probabilities" reflecting the degree of personal belief in the likelihood of the given statement.)

Three examples of use of the above terminology follow.

EXAMPLE 1. In the drug example of the introduction, assume it is desired to estimate θ_2. Since θ_2 is a proportion, it is clear that $\Theta = \{\theta_2 : 0 \le \theta_2 \le 1\} = [0, 1]$. Since the goal is to estimate θ_2, the action taken will simply be the choice of a number as an estimate for θ_2. Hence $\mathscr{A} = [0, 1]$. (Usually $\mathscr{A} = \Theta$ for estimation problems.) The company might determine the loss function to be

$$L(\theta_2, a) = \begin{cases} \theta_2 - a & \text{if } \theta_2 - a \ge 0, \\ 2(a - \theta_2) & \text{if } \theta_2 - a \le 0. \end{cases}$$

(The loss is in units of "utility," a concept that will be discussed in Chapter 2.) Note that an overestimate of demand (and hence overproduction of the drug) is considered twice as costly as an underestimate of demand, and that otherwise the loss is linear in the error.

A reasonable experiment which could be performed to obtain sample information about θ_2 would be to conduct a sample survey. For example, assume n people are interviewed, and the number X who would buy the drug is observed. It might be reasonable to assume that X is $\mathscr{B}(n, \theta_2)$ (see Appendix 1), in which case the sample density is

$$f(x|\theta_2) = \binom{n}{x}\theta_2^x(1 - \theta_2)^{n-x}.$$

There could well be considerable prior information about θ_2, arising from previous introductions of new similar drugs into the market. Let's say that, in the past, new drugs tended to capture between $\frac{1}{10}$ and $\frac{1}{5}$ of the market, with all values between $\frac{1}{10}$ and $\frac{1}{5}$ being equally likely. This prior information could be modeled by giving θ_2 a $\mathscr{U}(0.1, 0.2)$ prior density, i.e., letting

$$\pi(\theta_2) = 10 I_{(0.1, 0.2)}(\theta_2).$$

The above development of L, f, and π is quite crude, and usually much more detailed constructions are required to obtain satisfactory results. The techniques for doing this will be developed as we proceed.

EXAMPLE 2. A shipment of transistors is received by a radio company. It is too expensive to check the performance of each transistor separately, so a sampling plan is used to check the shipment as a whole. A random sample of n transistors is chosen from the shipment and tested. Based upon X, the number of defective transistors in the sample, the shipment will be accepted or rejected. Thus there are two possible actions: a_1—accept the shipment, and