MATHEMATICAL STATISTICS

A DECISION THEORETIC APPROACH

Thomas S. Ferguson

MATHEMATICAL STATISTICS

DECISION THEORETIC APPROACH

Thomas S. Ferguson

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
LOS ANGELES, CALIFORNIA

1967

ACADEMIC PRESS New York and London

A Subsidiary of Harcourt Brace Jovanovich, Publishers

ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 66-30080

PRINTED IN THE UNITED STATES OF AMERICA

Preface

The theory of games is a part of the rich mathematical legacy left by John von Neumann, one of the outstanding mathematicians of our era. Although others—notably Emil Borel—preceded him in formulating a theory of games, it was von Neumann who with the publication in 1927 of a proof of the minimax theorem for finite games laid the foundation for the theory of games as it is known today. Von Neumann's work culminated in a book written in collaboration with Oskar Morgenstern entitled *Theory of Games and Economic Behavior* published in 1944.

At about the same time, statistical theory was being given an increasingly rigorous mathematical foundation in a series of papers by J. Neyman and Egon Pearson. Statistical theory until that time, as developed by Karl Pearson, R. A. Fisher, and others had lacked the precise mathematical formulation, supplied by Neyman and Pearson, that allows the delicate foundational questions involved to be treated rigorously.

Apparently it was Abraham Wald who first appreciated the connections between the theory of games and the statistical theory of Neyman and Pearson, and who recognized the advantages of basing statistical theory on the theory of games. Wald's theory of statistical decisions, as it is called, generalizes and simplifies the Neyman-Pearson theory by unifying, that is, by treating problems considered as distinct in the Neyman-Pearson theory as special cases of the decision theory problem.

In the 1940's, Wald produced a prodigious amount of research that resulted in the publication of his book *Statistical Decision Functions* in 1950, the year of his tragic death in an airplane accident.

It is our objective to present the elements of Wald's decision theory and an investigation of the extent to which problems of mathematical statistics may be treated successfully by this approach. The main viewpoint is developed in the first two chapters and culminates in a rather general complete class theorem (Theorem 2.10.3). The remaining five chapters deal with statistical topics. No separate chapter on estimation is included since estimation is discussed as examples for general decision problems. It was originally intended that only those parts of statistical theory that could be justified from a decision-theoretic viewpoint would be included. Mainly, this entails the omission of those topics whose mathematical justification is given by large sample theory, such as maximum likelihood estimates, minimum χ^2 methods, and likelihood ratio tests. However, one exception is made. Although the theory of confidence sets as treated does not allow a decision-theoretic justification, it was felt that this topic "belongs" in any discourse on statistics wherein tests of hypotheses are treated. For purposes of comparison, the decision-theoretic notion of a set estimate is included in the exercises.

This book is intended for first-year graduate students in mathematics. It has been used in mimeographed form at UCLA in a two-semester or three-quarter course attended mainly by mathematicians, bio-statisticians, and engineers. I have generally finished the first four chapters in the first semester, deleting perhaps Sections 1.4 and 3.7, but I have never succeeded in completing the last three chapters in the second semester.

There are four suggested prerequisites.

(1) The main prerequisite is a *good* undergraduate course in probability. Ideally, this course should pay a little more attention to conditional expectation than the usual course. In particular, the formula $E(E(X \mid Y)) = E(X)$ should be stressed. Although the abstract approach to probability theory through measure theory is not used (except in Section 3.7, which may be omitted), it is assumed that the reader is acquainted with the notions of a σ-field of sets (as the natural domain of definition of a probability) and of a set of probability zero.

(2) An undergraduate course in analysis on Euclidean spaces is strongly recommended. It is assumed that the reader knows the con-

cepts of continuity, uniform continuity, open and closed sets, the Riemann integral, and so forth.

(3) An introductory undergraduate course in statistics is highly desirable as background material. Although the usual notions of test, power function, and so on, are defined as they arise, the discussion and illustration are rather abstract.

(4) A course in the algebra of matrices would be helpful to the student.

Rudimentary notes leading to this book have been in existence for about six years. Each succeeding generation of students has improved the quality of the text and removed errors overlooked by their predecessors. Without the criticism and interest of these students, too numerous to mention individually, this book would not have been written. Early versions of the notes benefitted from comments by Jack Kiefer and Herbert Robbins. The notes were used by Milton Sobel for a course at the University of Minnesota; his criticisms and those of his students were very useful. Further improvements followed when Paul Hoel used the notes in a course at UCLA. Finally, Gus Haggstrom gave the galleys a critical reading and caught several errors that eluded all previous readers. To all these, I express my deep appreciation.

THOMAS S. FERGUSON

Berkeley, California
February, 1967

Contents

CHAPTER 6. Multiple Decision Problems

CHAPTER 7. Sequential Decision Problems

CHAPTER 1

Game Theory and Decision Theory

1.1 Basic Elements

The elements of decision theory are similar to those of the theory of games. In particular, decision theory may be considered as the theory of a two-person game, in which nature takes the role of one of the players. The so-called normal form of a zero-sum two-person game, henceforth to be referred to as a *game*, consists of three basic elements:

1. A nonempty set, Θ, of possible states of nature, sometimes referred to as the parameter space.
2. A nonempty set, $\mathfrak{A}$, of actions available to the statistician.
3. A loss function, $L(\theta, a)$, a real-valued function defined on $\Theta \times \mathfrak{A}$.

A game in the mathematical sense is just such a triplet $(\Theta, \mathfrak{A}, L)$, and any such triplet defines a game, which is interpreted as follows. Nature chooses a point θ in Θ, and the statistician, without being informed of the choice nature has made, chooses an action a in $\mathfrak{A}$. As a consequence of these two choices, the statistician loses an amount $L(\theta, a)$. [The function $L(\theta, a)$ may take negative values. A negative loss may be interpreted as a gain, but throughout this book $L(\theta, a)$ represents the loss to the statistician if he takes action a when θ is the "true state of nature".] Simple though this definition may be, its scope is quite broad, as the following examples illustrate.

Example 1. Odd or Even. Two contestants simultaneously put up either one or two fingers. One of the players, call him player I, wins if the sum of the digits showing is odd, and the other player, player II, wins if the sum of the digits showing is even. The winner in all cases receives in dollars the sum of the digits showing, this being paid to him by the loser.

To create a triplet $(\Theta, \mathcal{A}, L)$ out of this game we give player I the label "nature" and player II the label "statistician". Each of these players has two possible choices, so that $\Theta = \{1, 2\} = \mathcal{A}$, in which "1" and "2" stand for the decisions to put up one and two fingers, respectively. The loss function is given by Table 1.1. Thus $L(1, 1) = -2$,

Table 1.1

Θ \ $\mathcal{A}$	1	2
1	−2	3
2	3	−4

$L(\theta, a)$

$L(1, 2) = 3$, $L(2, 1) = 3$, and $L(2, 2) = -4$. It is quite clear that this is a game in the sense described in the first paragraph. This example is discussed later in Section 1.7, in which it is shown that one of the players has a distinct advantage over the other. Can you tell which one it is? Which player would you rather be?

Example 2. Tic-Tac-Toe, Chess. In the game $(\Theta, \mathcal{A}, L)$ an element of the space Θ or $\mathcal{A}$ is sometimes referred to as a *strategy*. In some games strategies are built on a more elementary concept, that of a "move". Many parlor games illustrate this feature; for example, the games tic-tac-toe, chess, checkers, Battleship, Nim, Go, and so forth. A *move* is an action made by a specified player at a specified time during the game. The rules determine at each move the player whose turn it is to move and the choices of move available to that player at that time. For such a game a strategy is a rule that specifies for a given player the exact move he is to make each time it is his turn to move, for all possible histories of the game. The game of tic-tac-toe has at most nine moves, one player making five of them, the other making four. A player's strategy must tell him exactly what move to make in each possible position that may occur in the game. Because the number of possible games of tic-tac-toe is rather small (less than 9!), it is possible to write down an optimal strategy for each player. In this case each player has a strategy that guarantees its user at least a tie, no matter what his opponent does. Such strategies are called optimal strategies. Naturally, in the game of chess it is physically impossible to describe "all possible histories", for there are too many possible games of chess and many more strategies, in fact, than there are atoms in our solar system. We can write down strategies for the game of chess, but none so far constructed has much of a chance of beating the average amateur. When the two players have written down their strategies, they may be given to a referee who may play through the game and determine the winner. In the triplet $(\Theta, \mathcal{A}, L)$, which describes either tic-tac-toe or chess, the spaces Θ and $\mathcal{A}$ are the sets of all strategies for the two players, and the loss function $L(\theta, a)$ may be $+1$ if the strategy θ beats the strategy a, 0 for a draw, and -1 if a beats θ.

EXAMPLE 3. A GAME WITH BLUFFING. Another feature of many games, and one that is characteristic of card games, is the notion of a *chance move*. The dealing or drawing of cards, the rolling of dice, the spinning of a roulette wheel, and so on, are examples of chance moves. In the theory of games it is assumed that both players are aware of the probabilities of the various outcomes resulting from a chance move. Sometimes, as in card games, one player may be informed of the actual outcome of a chance move, whereas the other player is not. This leads to the possibility of "bluffing". The following example is a loose description of a situation which sometimes occurs in the game of stud poker.

Two players each put an "ante" of a units into a pot $(a > 0)$. Player I then draws a card from a deck, which gives him a winning or a losing card. Both players are aware of the probability P that the card drawn is a winning card $(0 < P < 1)$. Player I then may bet b units $(b > 0)$ by putting b units into the pot or he may check. If player I checks, he wins the pot if he has a winning card and loses the pot if he has a losing card. If player I bets, player II may call and put b units in the pot or he may fold. If player II folds, player I wins the pot whatever card he has drawn. If player II calls, player I wins the pot if he has a winning card and loses it otherwise.

If I receives a winning card, it is clear that he should bet: if he checks, he automatically receives total winnings of a units, whereas if he bets, he will receive at least a units and possibly more. For the purposes of our discussion we assume that the rules of the game enforce this condition: that if I receives a winning card, he must bet. This will eliminate some obviously poor strategies from player I's strategy set. With this restriction, player I has two possible strategies: (a) *the bluff strategy*—bet with a winning card or a losing card; and (b) *the honest strategy*—bet with a winning card, check with a losing card. The two strategies for player II are (a) *the call strategy*—if player I bets, call; and (b) *the fold strategy*—if player I bets, fold. Given a strategy for each player in a game with chance moves, a referee can play the game through as before, playing each chance move with the probability distribution specified, and determining who has won and by how much. The actual payoff in such games is thus a random quantity determined by the chance moves. In writing down a loss function, we replace these random quantities by their expected values in order to obtain a game as defined. (Further discussion of this may be found in Sections 1.3 and 1.4.) Table 1.2

Table 1.2

I \ II	Call	Fold
Bluff	$(2P - 1)(a + b)$	a
Honest	$(2P - 1)a + Pb$	$(2P - 1)a$

shows player I's expected winnings and player II's expected losses. For example, if I uses the honest strategy and II uses the call strategy, player II's loss will be $(a + b)$ with probability P(I receives a winning card) and $-a$ with probability $(1 - P)$(I receives a losing card). The expected loss is

$$(a + b)P - a(1 - P) = (2P - 1)a + Pb,$$

as found in the table. If player I is given the label "nature" and player II the label "statistician," the triplet $(\Theta, \mathcal{A}, L)$, in which $\Theta =$ (bluff, honest), $\mathcal{A} =$ (call, fold), and L is given by Table 1.2, defines a game that contains the main aspects of the bluffing game already described. This game is considered in Exercises 1.7.4 and 5.2.8.

1.2 A Comparison of Game Theory and Decision Theory

There are certain differences between game theory and decision theory that arise from the philosophical interpretation of the elements Θ, $\mathcal{A}$, and L. The main differences are these.

1. In a two-person game the two players are trying simultaneously to maximize their winnings (or to minimize their losses), whereas in decision theory nature chooses a state without this view in mind. This difference plays a role mainly in the interpretation of what is considered to be a good decision for the statistician and results in presenting him with a broader dilemma and a correspondingly wider class of what might be called "reasonable" decision rules. This is natural, for one can depend on an intelligent opponent to behave "rationally", that is to say, in a way profitable to him. However, a criterion of "rational" behavior for nature may not exist or, if it does, the statistician may not have knowledge of it. We do not assume that nature wins the amount $L(\theta, a)$ when θ and a are the points chosen by the players. An example will make this clear. Consider the game $(\Theta, \mathcal{A}, L)$ in which $\Theta = \{\theta_1, \theta_2\}$ and $\mathcal{A} = \{a_1, a_2\}$ and in which the loss function L is given by Table 1.3. In game

Table 1.3

	a_1	a_2
θ_1	4	1
θ_2	-3	0

$L(\theta, a)$

theory, in which the player choosing a point from Θ is assumed to be intelligent and his winnings in the game are given by the function L, the only "rational" choice for him is θ_1. No matter what his opponent does, he will gain more if he chooses θ_1 than if he chooses θ_2. Thus it is clear that the statistician should choose action a_2, instead of a_1, for he will lose only one instead of four. Again, this is the only reasonable thing for him to do. Now, suppose that the function L does not reflect the winnings of nature or that nature chooses a state without any clear objective in mind. Then we can no longer state categorically that the statistician should choose action a_2. If nature happens to choose θ_2, the statistician will prefer to take action a_1. This basic conceptual difference between game theory and decision theory is reflected in the difference between the theorems we have called fundamental for game theory and fundamental for decision theory (Sec. 2.2).

2. It is assumed that nature chooses the "true state" once and for all and that the statistician has at his disposal the possibility of gathering information on this choice by sampling or by performing an experiment. This difference between game theory and decision theory is more apparent than real, for one can easily imagine a game between two intelligent adversaries in which one of the players has an advantage given to him by the rules of the game by which he can get some information on the choice his opponent has made before he himself has to make a decision. It turns out (Sec. 1.3) that the over-all problem which allows the statistician to gain information by sampling may simply be viewed as a more complex game. However, all statistical games have this characteristic feature, and it is the exploitation of the structure which such gathering of information gives to a game that distinguishes decision theory from game theory proper.

For an entertaining introduction to finite games the delightful book *The Compleat Strategyst* by the late J. D. Williams (1954) is highly recommended. The more serious student should also consult the lucid accounts of game theory found in McKinsey (1952), Karlin (1959), and Luce and Raiffa (1957). An elementary text by Chernoff and Moses (1959) provides a good introduction to the main concepts of decision theory. The important book by Blackwell and Girshick (1954), which is a more advanced text, is recommended as collateral reading for this study.

1.3 Decision Function; Risk Function

To give a mathematical structure to this process of information gathering, we suppose that the statistician before making a decision is allowed to look at the observed value of a random variable or vector, X, whose distribution depends on the true state of nature, θ. Throughout most of this book *the sample space*, denoted by $\mathfrak{X}$, is taken to be (a Borel subset of) a finite dimensional Euclidean space, and the probability distributions of X are supposed to be defined on the Borel subsets, $\mathcal{B}$ of $\mathfrak{X}$. Thus for each $\theta \in \Theta$ there is a probability measure P_θ defined on $\mathcal{B}$, and a corresponding cumulative distribution function $F_X(x \mid \theta)$, which represents the distribution of X when θ is the true value of the parameter. [If X is an n-dimensional vector, it is best to consider X as a notation for $(X_1, \cdots, X_n)$ and $F_X(x \mid \theta)$ as a notation for the multivariate cumulative distribution function $F_{X_1, \cdots, X_n}(x_1, \cdots, x_n \mid \theta)$.]

A *statistical decision problem* or a *statistical game* is a game $(\Theta, \mathcal{A}, L)$ coupled with an experiment involving a random observable X whose distribution P_θ depends on the state $\theta \in \Theta$ chosen by nature.

On the basis of the outcome of the experiment $X = x$ (x is the observed value of X), the statistician chooses an action $d(x) \in \mathcal{A}$. Such a function d, which maps the sample space $\mathfrak{X}$ into $\mathcal{A}$, is an elementary strategy for the statistician in this situation. The loss is now the random quantity $L(\theta, d(X))$. The expected value of $L(\theta, d(X))$ when θ is the true state of nature is called *the risk function*

$$R(\theta, d) = E_\theta L(\theta, d(X)) \tag{1.1}$$

and represents the average loss to the statistician when the true state of nature is θ and the statistician uses the function d. Note that for some choices of the function d and some values of the parameter θ the expected value in (1.1) may be $\pm\infty$ or, worse, it may not even exist. As the following definition indicates, we do not bother ourselves about such functions.

Definition 1. Any function $d(x)$ that maps the sample space $\mathfrak{X}$ into $\mathcal{A}$ is called a nonrandomized *decision rule* or a nonrandomized *decision function*, provided the risk function $R(\theta, d)$ exists and is finite for all $\theta \in \Theta$. The class of all nonrandomized decision rules is denoted by D.

Unfortunately, the class D is not well defined unless we specify the sense in which the expectation in (1.1) is to be understood. The reader may take this expectation to be the Lebesgue integral,

$$R(\theta, d) = E_\theta L(\theta, d(X)) = \int L(\theta, d(x))\, dP_\theta(x).$$

With such an understanding, D consists of those functions d for which $L(\theta, d(x))$ is for each $\theta \in \Theta$ a Lebesgue integrable function of x. In particular, D contains all simple functions. (A function d from $\mathfrak{X}$ to $\mathfrak{A}$ is called simple if there is a finite partition of $\mathfrak{X}$ into measurable subsets $B_1, \cdots, B_m \in \mathfrak{B}$, and a finite subset $\{a_1, \cdots, a_m\}$ of $\mathfrak{A}$ such that for $x \in B_i$, $d(x) = a_i$ for $i = 1, \cdots, m$.) On the other hand, the expectation in (1.1) may be taken as the Riemann or the Riemann-Stieltjes integral,

$$R(\theta, d) = E_\theta L(\theta, d(X)) = \int L(\theta, d(x))\, dF_X(x \mid \theta).$$

In that case D would contain only functions d for which $L(\theta, d(x))$ is for each $\theta \in \Theta$ continuous on a set of probability one under $F_X(x \mid \theta)$. For the purposes of understanding what follows, it is not too important which of the various definitions is given to the expectation in (1.1). In most of the proofs of the theorems given later we use only certain linearity $[E(aX + Y) = aEX + EY]$ and ordering $(X > 0 \Rightarrow EX > 0)$ properties of the expectation; such proofs are equally valid for Lebesgue and Riemann integrals. Therefore we let the definition of the expectation be arbitrary (unless otherwise stated) and assume that the class D of decision rules is well defined.

EXAMPLE 1. The game of "odd or even" mentioned in Sec. 1.1 may be extended to a statistical decision problem. Suppose that before the game is played the player called "the statistician" is allowed to ask the player called "nature" how many fingers he intends to put up and that nature must answer truthfully with probability 3/4 (hence untruthfully with probability 1/4). The statistician therefore observes a random variable X (the answer nature gives) taking the values 1 or 2. If $\theta = 1$ is the true state of nature, the probability that $X = 1$ is 3/4; that is, $P_1\{X = 1\} = 3/4$. Similarly, $P_2\{X = 1\} = 1/4$. There are exactly four possible functions from $\mathfrak{X} = \{1, 2\}$ into $\mathfrak{A} = \{1, 2\}$. These are the four decision rules:

$$d_1(1) = 1, \quad d_1(2) = 1;$$
$$d_2(1) = 1, \quad d_2(2) = 2;$$
$$d_3(1) = 2, \quad d_3(2) = 1;$$
$$d_4(1) = 2, \quad d_4(2) = 2.$$

Rules d_1 and d_4 ignore the value of X. Rule d_2 reflects the belief of the