THE UNIVERSITY
OF WISCONSIN

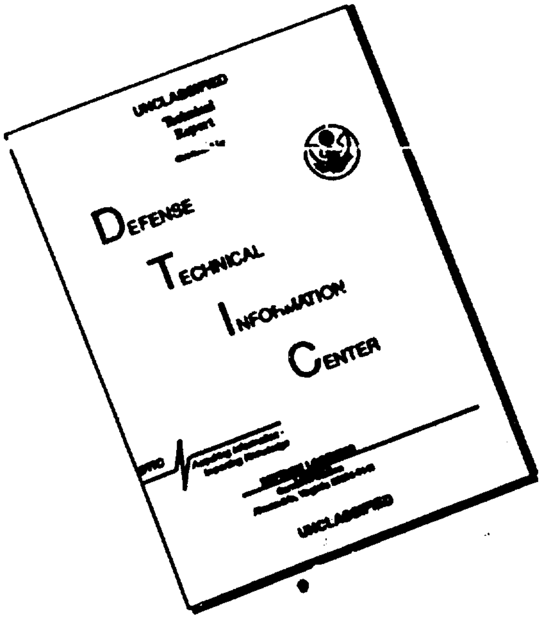
MATHEMATICS RESEARCH CENTER

Address:

Mathematics Research
Center
The University of Wisconsin
Madison, Wisconsin 53706
U.S.A.

30

DISCLAIMER NOTICE



UNCLASSIFIED

Technical Report

DEFENSE TECHNICAL INFORMATION CENTER

UNCLASSIFIED

THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

AR 70-31 Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Mathematics Research Center,
University of Wisconsin, Madison, Wis. 53706

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP
None

3. REPORT TITLE
MATHEMATICAL MODELS FOR STATISTICAL DECISION THEORY

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Summary Report: no specific reporting period.

5. AUTHOR(S) (First name, middle initial, last name)
Bernard Harris

6. REPORT DATE
December 1971

7a. TOTAL NO. OF PAGES
30

7b. NO. OF REFS
30

8a. CONTRACT OR GRANT NO.
Contract No. DA-31-124-ARO-D-462

b. PROJECT NO.
None

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
#1160

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
None

10. DISTRIBUTION STATEMENT
Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES
None

12. SPONSORING MILITARY ACTIVITY
Army Research Office-Durham, N.C.

13. ABSTRACT

Three methods of defining optimality of statistical decision rules are introduced. The first uses ideas of approximation theory by defining the optimal decision as that element of the risk set which best approximates an ideal rule. The second optimality principle defines optimality in terms of minimizing functionals. The third method is the axiomatization of optimality in statistical decision theory.

THE UNIVERSITY OF WISCONSIN
MATHEMATICS RESEARCH CENTER

Contract No.: DA-31-124-ARO-D-462

MATHEMATICAL MODELS FOR STATISTICAL DECISION THEORY

Bernard Harris

MRC Technical Summary Report #1160
December 1971

Received June 30, 1971

Madison, Wisconsin 53706

ABSTRACT

Three methods of defining optimality of statistical decision rules are introduced. The first uses ideas of approximation theory by defining the optimal decision as that element of the risk set which best approximates an ideal rule. The second optimality principle defines optimality in terms of minimizing functionals. The third method is the axiomatization of optimality in statistical decision theory.

MATHEMATICAL MODELS FOR STATISTICAL DECISION THEORY

Bernard Harris

1. Introduction. In the typical problem of statistical inference, the statistician is confronted with the problem of selecting one out of the vast number of possible decision rules. I will refer to this as the "fundamental problem of statistics". In this survey paper, I will try to give some mathematical characterizations of the problem of selecting an optimal decision procedure. This will not constitute a resolution of the "fundamental problem", inasmuch as this is intimately tied up with irresolvable philosophical difficulties. These arise, since in all but a few exceptional problems, there is no single procedure which can be regarded as uniformly dominating all other possible procedures. As a consequence, "reasonable people" have disagreed and will continue to disagree on the specific procedure that should be selected.

Despite all of the above difficulties, a great deal can still be done to formalize statistical decision theory. Thus, in what follows, I will give some characterizations of optimality in statistical inference. The discussion will of necessity be brief and is intended only to provide an introduction to these characterizations. More extensive treatments are in preparation and will appear subsequently.

The objective here will be to provide a mathematical structure in which the details of the original problem are replaced by abstract mathematical

Supported by the United States Army under Contract No.: DA-31-124-ARO-D-462 and by the Technological University, Eindhoven, Netherlands.

statements which retain those features common to large classes of problems. This enables us to isolate those aspects of decision theory which are relevant to the selection of a single decision rule.

For the most part, this paper does not really contain new mathematical results; instead it is largely concerned with the adaptation of known mathematical results to statistical problems.

Specifically, I will describe three ways of providing a mathematical model for statistical decision theory. The first is motivated by comparatively recent ideas of approximation theory. The second is obtained by representing the statistical problem as an optimization problem. The third approach is a discussion of some possible axiomatizations of statistical decision theory.

2. Preliminaries. Let Θ and $\mathcal{A}$ be topological spaces; θ and a will be used to denote generic elements of these spaces. Let $L(\theta, a)$ be a mapping from $\Theta \times \mathcal{A}$ into the reals. We will assume throughout that $L(\theta, a)$ is uniformly bounded from below; that is, there exists a real number $M \geq 0$, such that $L(\theta, a) \geq -M$ for all $\theta \epsilon \Theta$, $a \epsilon \mathcal{A}$. It is customary to refer to Θ as the parameter space, $\mathcal{A}$ as the space of actions, and L as the loss function. We also require a probability space $(\mathfrak{X}, \mathfrak{B}_{\mathfrak{X}}, P_\theta)$, where $\mathfrak{X}$ is a space of elements x and $\mathfrak{B}_{\mathfrak{X}}$ is a σ-algebra of subsets of $\mathfrak{X}$, and P_θ, $\theta \epsilon \Theta$ is one of a family of probability measures on $\mathfrak{B}_{\mathfrak{X}}$ indexed by Θ.

An experiment is conducted and the random variable X is observed. X is assumed to take values in $\mathfrak{X}$ and is distributed by P_θ, where θ is an element of Θ whose value is unknown to the statistician.

We can now outline the steps in a statistical problem. The statistician observes the event $X = x$, $x \in \mathfrak{X}$; then given the family of measures P_θ, but not the value of θ, he selects $a \in \mathcal{A}$ and is assessed the penalty $L(\theta, a)$. Thus, the objective is to choose a so that $L(\theta, a)$ is "kept small"; ideally, if there exists a_0 such that $L(\theta, a_0) = \min_{a \in \mathcal{A}} L(\theta, a)$, then one should choose a_0. However, it is obvious that in order to be able to choose a_0, in general, one would need to know which parameter $\theta \in \Theta$ prevailed. Since X is distributed by P_θ, the event $X = x$ contains information about θ, hence the choice of $a \in \mathcal{A}$ should generally depend on the outcome of the experiment. Thus, given the outcome of the experiment, a mapping $\delta: \mathfrak{X} \to \mathcal{A}$ is chosen. Since X is a random variable, $\delta(X)$ is a random variable, provided that δ is a measurable mapping. We denote the set of such measurable mappings by $\mathcal{D}$.

In repetitions of the experiment, X will change; hence, unless $\delta(X)$ is almost surely constant, $L(\theta, \delta(X))$ will vary. Thus, the average loss, $E_\theta\, L(\theta, \delta(X))$ is used as a criterion for choosing δ rather than the loss in any given experiment. We call $E_\theta\, L(\theta, \delta(X))$ the risk function (of δ) and denote it by $R_\delta(\theta) = R(\theta, \delta)$. Note that our assumptions (δ measurable and $L(\theta, a)$ uniformily bounded from below) insure that $R_\delta(\theta)$ "exists" for each θ; it may however, be $+\infty$.

It is desirable to augment $\mathcal{D}$ by introducing the "mixed" or "randomized" decision procedures Φ, whose elements will be denoted by φ. To do this, we introduce a σ-algebra $\mathcal{B}_{\mathcal{D}}$, and let Φ be the set of all probability distributions on $\mathfrak{A}_{\mathfrak{D}}$. Of necessity $\mathfrak{A}_{\mathfrak{D}}$ must include each $\{\delta\}$, $\delta \in \mathfrak{D}$, so that $\mathfrak{D} \subset \Phi$, by using distributions such that $P\{\delta\} = 1$. We will refer to the elements of Φ as decision rules or decision procedures; when it is necessary to specifically identify an element of Φ as being in $\mathfrak{D}$, we will refer to it as a pure decision rule.

Clearly, if $R(\theta, \delta)$ is $\mathfrak{A}_{\mathfrak{D}}$ measurable for each θ,

$$R_{\varphi}(\theta) = R(\theta, \varphi) = \int_{\mathfrak{D}} R(\theta, \delta) \, d\, \varphi(\delta)$$

is well-defined. We call $R_{\varphi}(\theta)$, as defined above, the risk function of φ . Then, the problem of selecting a decision procedure becomes the problem of selecting $\varphi \in \Phi$ so that $R(\theta, \varphi)$ is kept "small".

Let $S = \{R_{\varphi}(\theta), \ \varphi \in \Phi\}$ and let T be the mapping defined by $T\colon \Phi \rightarrow S$, that is $T(\varphi) = R_{\varphi}(\theta)$. We refer to S as the risk set.

If for φ_1, $\varphi_2 \in \Phi$, we have $R(\theta, \varphi_1) = R(\theta, \varphi_2)$ for all $\theta \in \Theta$, then φ_1 and φ_2 are said to be equivalent. Thus, the elements of S are equivalence classes of elements of Φ . Consequently, we can replace the problem of selecting $\varphi \in \Phi$ with the equivalent problem of selecting an element $s \in S$. Then we can choose any element of $T^{-1}(s)$ as the decision procedure to be employed.

In order to simplify notation, we will adopt the following conventions in the material that follows. For s_1, $s_2 \in S$, $s_1 < s_2$ means $T^{-1}(s_1) = R(\theta, \varphi_1) \leq R(\theta_2, \varphi_2) = T^{-1}(s_2)$ for all $\theta \in \Theta$ and for some $\theta_0 \in \Theta$, $R(\theta_0, \varphi_1) < R(\theta_0, \varphi_2)$; here φ_1 is any element of $T^{-1}(s_1)$ and φ_2 is any element of $T^{-1}(s_2)$. Similarly, $s_1 \leq s_2$ means $R(\theta, \varphi_1) \leq R(\theta, \varphi_2)$ for all $\theta \epsilon \Theta$.

In particular, we will say that $s_1 \epsilon S$ is inadmissible (and all $\varphi \epsilon T^{-1}(s_1)$ are inadmissible) if there is an $s_0 \epsilon S$ with $s_0 < s_1$. If there is no such s_0, then s_1, equivalently any $\varphi \epsilon T^{-1}(s_1)$, will be said to be admissible.

Clearly, if $s_1 < s_2$, then s_2 should not be employed by the statistician. Further, if there is an s_0 such that $s_0 \leq s$, for all $s \epsilon S$; then s_0 is to be chosen and there is no problem of selection. This is, unfortunately, an exceptional situation. In most problems, an ordering of this type is not available among the decision rules being considered; it is quite customary to find oneself with the problem of selecting one of a large set of mutually incomparable rules.

Note that S is of necessity a convex set. To see this, observe that Φ is a convex set. Then let $\varphi_0 = \lambda\varphi_1 + (1-\lambda)\varphi_2$, $\varphi_1, \varphi_2 \epsilon \Phi$, $\lambda \epsilon [0,1]$. Then

$$R_{\varphi_0}(\theta) = \int_{\mathcal{D}} R(\theta, \delta)\, d[(\lambda\varphi_1 + (1-\lambda)\varphi_2)(\delta)]$$

$$= \int_{\mathcal{D}} R(\theta, \delta)[\lambda d\varphi_1(\delta) + \lambda d\varphi_2(\delta)]$$

$$= \lambda R_{\varphi_1}(\theta) + (1-\lambda) R_{\varphi_2}(\theta) \ .$$

Since T is a linear mapping, we have that if

$$s_0 = T(\varphi_0) \ ,$$

then

$$s_0 = \lambda_{s_1} + (1-\lambda)\, s_2 \;,$$

where

$$s_1 = T(\varphi_1), \quad s_2 = T(\varphi_2) \;.$$

However, as S has been defined, it is not necessarily the convex hull of $T(\mathcal{D})$, denoted by $co(T(\mathcal{D}))$. To see this, let $\mathcal{D} = \{-\infty < \delta < \infty\}$, $\Theta = \{-\infty < \theta < \infty\}$ and

$$R_\delta(\theta) = \begin{cases} 0 & \theta < \delta \\ 1 & \theta \geq \delta \;. \end{cases}$$

That is, $T(\mathcal{D})$ is the set of all degenerate cumulative distribution functions. It is easily seen that $co(T(\mathcal{D}))$ is the set of all cumulative distribution functions with a finite number of jumps. However, S is the set of all cumulative distribution functions on $(-\infty, \infty)$.

In the discussion that follows, the "classical" decision criteria known as the minimax criterion, minimax regret criterion, and Laplace's criterion will be used repeatedly as illustrations. For the sake of completeness, they are defined below. Each of these corresponds to a different interpretation of what might be meant by "keeping $R_\varphi(\theta)$ small".

(1). <u>The minimax criterion</u>. Choose $\varphi_0 \in \Phi$ so that $\sup_\theta R_{\varphi_0}(\theta) \leq \sup_\theta R_\varphi(\theta)$ for all $\varphi \in \Phi$.

(2). <u>The minimax regret criterion</u>. Choose $\varphi_0 \in \Phi$ so that

$$\sup_{\theta}[R_{\varphi_0}(\theta) - \inf_{\gamma \epsilon \Phi} R_{\gamma}(\theta)] \leq \sup_{\theta}[R_{\varphi}(\theta) - \inf_{\gamma \epsilon \Phi} R_{\gamma}(\theta)]$$

for all $\varphi \epsilon \Phi$.

(3). Laplace's criterion. Choose $\varphi_0 \epsilon \Phi$ so that

$$\int_{\Theta} R_{\varphi_0}(\theta)\, d\mu(\theta) \leq \int_{\Theta} R_{\varphi}(\theta)\, d\mu(\theta)$$

for all $\varphi \epsilon \Phi$, where μ is the uniform measure on Θ . (If Θ is a compact subset of E_n, for example, $\int_{\Theta} R_{\varphi}(\theta)d\mu(\theta)$ is well-defined. In more general situations, some modifications to this definition may be required).

In each of the three cases above s_0, the "optimal" element of s , is defined by $s_0 = T(\varphi_0)$.

Each of these reflects a different interpretation of "keeping $R_{\varphi}(\theta)$ small", in ignorance of the value of θ . The minimax criterion guarantees that the largest value of $R_{\varphi}(\theta)$ is as small as possible. The minimax regret criterion identifies $\inf_{\varphi \epsilon \Phi} R_{\varphi}(\theta)$, the lower envelope function, as the smallest loss that you could incur if you knew θ; hence $R_{\varphi}(\theta) - \inf_{\gamma \epsilon \Phi} R_{\gamma}(\theta)$ is the additional loss that is incurred by one's ignorance of θ . Then you seek to make the maximum of this difference as small as possible. In Laplace's criterion, the philosophy is one of keeping the "average loss" small rather than the maximum — as is the case with minimax and minimax regret. Average is here identified with the uniform measure on Θ .

Historically, there have been two approaches employed in studying the question of what might be meant by a "best" $s \epsilon S$. They are

(1) Specify a criterion, then deduce its properties and see if they are satisfactory.

(2) List the properties that you would like a decision procedure to possess, then determine the *existence* and *construction* of decision rules meeting the required conditions.

In what follows, the third and fourth sections will follow the first approach and the fifth section will use the last approach.

For additional material on the above definitions and concepts, the reader is referred to standard treatises on decision theory, such as D. Blackwell and M. A. Girshick [2], T. S. Ferguson [11], and A. Wald [30].

3. Approximation theory and statistical decision theory

Let $\mathcal{L}$ be a normed linear space of real-valued functions of θ, $\theta \in \Theta$; for $x \in \mathcal{L}$, we denote the norm of x by $\|x\|_{\mathcal{L}}$. *When there is no danger* of confusion concerning the space under consideration, the subscript $\mathcal{L}$ will be deleted. Let $S_{\mathcal{L}}$ be a convex set in $\mathcal{L}$ and let v be a distinguished point in $\mathcal{L}$.

We say that $s_0 \in S_{\mathcal{L}}$ is a best approximation to v if $\|s_0 - v\| \leq \|s - v\|$ for all $s \in S_{\mathcal{L}}$. The existence and determination of s_0 is a well-known problem in approximation theory and there is a considerable literature about this topic. A few of the more significant of these results are summarized below. If we add the additional assumption that $v \leq s$ for all $s \in S_{\mathcal{L}}$, then, we will show that the notion of a best approximation to v is a possible interpretation of the concept of optimality in statistical decision theory.

Consequently, the theorems of this part of approximation theory frequently hav:e a natural reinterpretation in a statistical context.

For general discussions of properties of convex sets in normed linear spaces, the reader is referred to F. A. Valentine [29] and N. Dunford and J. T. Schwartz, Chapter V, [9]. The theory of best approximations by elements of convex sets is discussed in the book by I. Singer [26], (in particular see Appendix I) and in papers by F. R. Deutsch and P. H. Maserick [8], A. L. Garkavi [12,13], V. N. Burov [3, 4], and G. S. Rubinstein [23], to name a few.

We denote a hyperplane H in $\mathcal{L}$ as a set of the form

$$H = \{x \in \mathcal{L} : L(x) = c\}$$

where $L \in \mathcal{L}^*$, the adjoint space of $\mathcal{L}$, $L \neq 0$, and c is a real scalar.

Then, the best approximation in $S_{\mathcal{L}}$ to v can be characterized by the following theorems, which will be stated here without proofs.

Theorem 3.1 (I. Singer [26], F. R. Deutsch and P. H. Maserick[8]). Let $S_{\mathcal{L}}$ be a convex set in $\mathcal{L}$, a normed linear space, and let $v \in S_{\mathcal{L}}^c$, the complement of $S_{\mathcal{L}}$. Then, there exists an $s_0 \in S_{\mathcal{L}}$ which is a best approximation to v if and only if there exists a linear functional $L \in \mathcal{L}^*$ with

(1) $\|L\| = 1$,

(2) $L(s_0) = \inf_{s \in S_{\mathcal{L}}} L(s)$,

(3) $L(s_0 - v) = \|s_0 - v\|$.

Geometrically, this says that a point s_0 in $S_{\mathcal{L}}$ is a best approximation to $v \epsilon S^c_{\mathcal{L}}$ if and only if there is a hyperplane H separating v from $S_{\mathcal{L}}$, which supports $S_{\mathcal{L}}$ at s_0, and whose distance from v is the distance from v to s_0.

Secondly, we have the following characterization of a best approximation.

<u>Theorem 3.2</u> (A. L. Garkavi [13]). Let $S_{\mathcal{L}}$ be a convex set in the normed linear space $\mathcal{L}$. Then $s_0 \epsilon S_{\mathcal{L}}$ is the best approximation to $v \epsilon S^c_{\mathcal{L}}$ if and only if for each $s \epsilon S_{\mathcal{L}}$ there is a linear functional L_s in $\mathcal{L}^*$ such that

(1) L_s is an extreme point of the closed unit ball in $\mathcal{L}^*$,

(2) $L_s(s - s_0) \geq 0$,

(3) $L_s(s_0 - v) = \| s_0 - v \|$.

We now turn to the connection between the best approximation problem described above and optimality in statistical decision theory.

Let S be the risk set of a statistical decision problem and let $\mathcal{L}$ be a normed linear space. Let $S_{\mathcal{L}} = S \cap \mathcal{L} = \{s \epsilon S, \|s\|_{\mathcal{L}} < \infty \}$. We say that $s_0 \epsilon S$ is $(v, \mathcal{L})$ optimal if for $v \leq s$, for all $s \epsilon S_{\mathcal{L}}$, s_0 is the best approximation to v from $S_{\mathcal{L}}$. If $S_{\mathcal{L}}$ is empty, then we define every $s \epsilon S$ as $(v, \mathcal{L})$ optimal.

It remains to be shown that this is in fact a reasonable definition of optimality for statistical decision problems. We will try to justify this in two steps. First, we will show that minimax, minimax regret, and

Laplace's criterion are $(v, \mathcal{L})$ optimal procedures for particular choices of v and $\mathcal{L}$. Second, having established that this is in fact a generalization of these "classical" decision criteria, we will give an intuitive interpretation of the notion of $(v, \mathcal{L})$ optimality as a family of decision criteria. We will make the simplifying assumption that Θ is compact. Modifications to some definitions will be needed, when this is not the case, and can easily be made. However, these will be omitted here, since they do not serve the immediate purpose of this exposition.

Let $v = v(\theta) = -M$, $\theta \in \Theta$. Then, if we take $\mathcal{L}$ to be the space of bounded functions $f(\theta)$, $\theta \in \Theta$ with $\|f\|_{\mathcal{L}} = \sup_{\theta} |f(\theta)|$. Then, the best approximation to v from S, that is, the $(v, \mathcal{L})$ optimal decision rule for this case is the minimax decision procedure. Note that our hypotheses insure $v \leq s$ for all $s \in S$.

Now employ the same choice of v and let $\mathcal{L}$ be the space of μ-integrable functions of θ, where μ is the uniform measure on Θ. Here the $(v, \mathcal{L})$ optimal aecision rule is Laplace's criterion.

To obtain the identification of minimax regret as a $(v, \mathcal{L})$ optimal decision procedure, we define $v = v(\theta) = \inf_{\varphi \in \Phi} R_{\varphi}(\theta)$, the lower envelope function. Clearly $v \leq s$, for all $s \in S$. Then the same choice of $\mathcal{L}$ as in the representation of the minimax criterion provides the representation of the minimax regret criterion.

Hence, it is evident that this notion oi optimality is a generalization of some of the familiar notions of optimality.

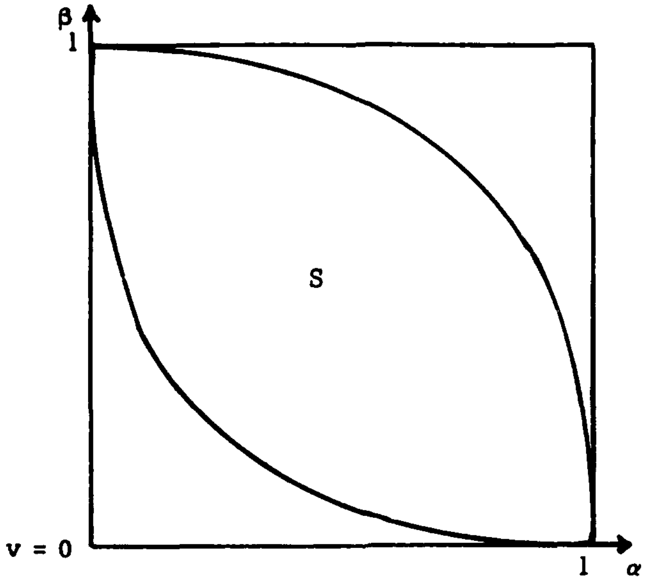
We now give an intuitive interpretation of $(v, \mathcal{L})$ optimality as a statistical optimality criterion. The statistician should interpret v as the "ideal" decision rule, that is, what he would like to be able to accomplish, such as in the case of "perfect information". The distance from v to S reflects his inability to accomplish this ideal as a consequence of uncertainty. Since v in general is not attainable, the suggestion is to choose that element of S which comes as close as possible to the ideal v, hence, "a best approximation to v".

To clarify these ideas, consider the following simple example. We consider a decision theoretic model for the problem of testing a simple hypothesis against a simple alternative. Thus $\Theta = (\theta_1, \theta_2)$ and $\mathcal{A} = (a_1, a_2)$.

Let

$$L(\theta, a) = \begin{array}{c} \\ a_1 \\ a_2 \end{array} \begin{array}{c} \begin{array}{cc} \theta_1 & \theta_2 \end{array} \\ \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array} .$$

Then the risk set is the set $(\alpha_\varphi, \beta_\varphi)$, corresponding to the probabilities of errors of the first and second kind using the decision rule $\varphi \in \Phi$. A typical risk set for such a problem is shown below.



Here it appears natural to set v = (0, 0), which corresponds to a "perfect" test, that is, one with size zero and power unity. The different choices of $\mathcal{L}$ correspond to different ways of defining the point in S which is closest to (0, 0) .

We now return to Theorems 3.1 and 3.2 to reexamine them in the light of statistical decision theory, rather than approximation theory. First note that the only significant distinction in the conversion to the statistical problem is the additional assumption $v \leq s$. This readily leads to the following, which we state as Theorem 3.3.

Theorem 3.3. Let $S_{\mathcal{L}}$ be a convex set in $\mathcal{L}$, a normed linear space and let $v \leq s$, for all $s \in S_{\mathcal{L}}$. Then, s_0 is a best approximation to v if and only if there exists a positive linear functional $L \in \mathcal{L}^*$ satisfying conditions

(1), (2) and (3) of Theorem 3.1. Similarly, s_0 is a best approximation to v if and only if for each $s \in S_{\mathcal{L}}$, there is a positive linear functional L_s in $\mathcal{L}^*$ satisfying conditions (1), (2) and (3) of Theorem 3.2.

Some of the more immediate consequences, for example, include the following well-known result in statistical decision theory. If all the risk functions in $S_{\mathcal{L}}$ are continuous and $S_{\mathcal{L}} \neq \emptyset$, then the space $\mathcal{L}$ determined by the sup norm coincides with $\mathcal{L}_\infty$ space, say if $\Theta \subset E_n$. The adjoint space is $\mathcal{L}_1$ space and the positive linear functionals of norm unity are then the probability distributions on Θ, using the usual Borel σ-algebra on Θ . Then the linear functional L of Theorem 3.1 is the least favorable distribution (for either minimax or minimax regret, depending on the choice of v from the two alternatives specified earlier). In a less restrictive context, the linear functional L of Theorem 3.1, gives a prior distribution against which s_0 is Bayes, provided one adds a few minor regularity conditions. Statistical interpretations of Theorem 3.2 are not as useful, although these can be inferred as well.

4. Opimization in statistical decision theory defined by minimizing functionals

In this section we introduce another method of defining optimization in statistical decision problems.

Let $\mathcal{H}$ be a class of extended real valued functionals h on a linear topological space $\mathfrak{M}$ of real valued functions of θ, $\theta \in \Theta$, with the following properties:

(1) $h(0) = 0$, $|h(-\overline{M})| < \infty$, where $-\overline{M}$ denotes the function $f(\theta) \equiv M$, $\theta \in \Theta$.

(2) If $x \leq y$, $h(x) \leq h(y)$.

Then we say that $s_0 \in S \subset \mathfrak{M}$ is h-optimal if $h(s_0) \leq h(s)$ for all $s \in S$. If $h(s) = +\infty$ for all $s \in S$, then every $s \in S$ is said to be h-optimal.

Since every norm is a functional, h-optimality reduces to $(v, \mathfrak{L})$ optimality in some particular instances. That is, if $h(s) = \|s - v\|_{\mathfrak{L}}$, when $\|s - v\|_{\mathfrak{L}} < \infty$ and $h(s) = \infty$, otherwise, then this is precisely $(v, \mathfrak{L})$ optimality. However, there are many instances of h-optimality which are not expressible in terms of norms, and hence, this is in fact a generalization of $(v, \mathfrak{L})$ optimality.

It is instructive to examine the principle of Bayesian inference in the light of the definition of h-optimality. Let μ be a probability measure (equivalently any measure μ with $\mu(\Theta) < \infty$) on the Borel sets of Θ . Then s_0 is the Bayes decision rule with respect to μ if

$$\int_\Theta R_{\varphi_0}(\theta)d\mu(\theta) \leq \int_\Theta R_\varphi(\theta)d\mu(\theta) \tag{4.1}$$

for all $\varphi \in \Phi$ and $s_0 = T(\varphi_0)$. Note that we need the additional assumption that $R_\varphi(\theta)$ is measurable with respect to the Borel σ-algebra on Θ for every $\varphi \in \Phi$.

Observe that (4.1) is a statement of minimization with respect to a linear functional. Since, as a consequence of the Riesz representation theorem, every positive linear functional has a representation of the form employed in (4.1) for compact Θ, Bayesian inference coincides in this case with h-optimality for linear h .

This observation provides an obvious explanation for the assertion that the solution to a Bayesian problem is obtained more easily than the solution using other criteria. Namely, in the sense employed here, Bayesian problems, are linear problems, that is, minimization with respect to linear functionals. Other optimization principles are generally non-linear in this sense.

We can relate this principle to a growing body of mathematical literature as well by noting that this is precisely the structure of mathematical programming problems. The function h becomes the objective function of the mathematical programmer and the convex set S is the set of feasible points. Hence Bayesian inference is linear programming from this point of view. However, the set of constraints required to generate S need not necessarily be finite. The general optimization problem is, in general, a problem of non-linear programming in an arbitrary linear space.

Some useful work in this area which can be exploited by statisticians are J. W. Daniel [6 , 7], K. Kirchgassner and K. Ritter [15], K. Ritter [22], and L. W. Neustadt [19]. In particular, it should be noted that some relationships between mathematical programming and the areas of statistics and probability are quite well-known. An extensive discussion of these and a substantial bibliography may be found in the survey paper by O. Krafft [16].

5. Axiomatizations of optimality in statistical decision theory. Another method of defining optimality in statistical decision problems is to list properties that you would like such a procedure to possess. I will briefly summarize the history of this topic and conclude with the statement of some recent results by E. E. Nordbrock and some of their consequences.

In H. Chernoff [5], a set of eight postulates is exhibited for a finite decision problem (Θ, $\mathcal{A}$ both finite). For these postulates, Laplace's criterion is the only rule which satisfies all eight. Chernoff notes that if an additional postulate were added to the list; a postulate of the "nature duplication" type to be discussed below, then a contradiction would result. Chernoff's results "justifying" Laplace's criterion were extended to more general decision problems by H. Uzawa [27,28].

In J. Milnor [18], a list of ten postulates for a finite decision problem are given. Subsets of these which characterize Laplace's criterion, minimax, and minimax regret are exhibited. It should be noted that minimax regret was proposed by L. J. Savage in [24,25], and is referred to as Savage's criterion by Milnor. Milnor also exhibited a set of eight postulates which are consistent and gave a construction of a rule which satisfies these postulates. His rule is a precursor of the rule used by Atkinson, Church, and Harris[1], which will be discussed in greater detail later. Good [14], proposed a restricted type of minimax rule. The thirteenth chapter of R. D. Luce and H. Raiffa [17] and the paper of R. Radner and J. Marschak [21] provide expository treatments of decision principles.

In Atkinson, Church, and Harris, the following set of postulates were proposed for the finite decision problem $\Theta = \{\theta_1, \ldots, \theta_n\}$ and $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$; here S is a convex polyhedron in E_n and is the convex hull of the row vectors of the matrix $A = \{a_{ij}\}$, where $a_{ij} = L(d_i, \theta_j)$.

1. The optimal class $Q(A)$ is non-empty.

2. Let π_1 and π_2 be permutations acting on Θ and $\mathcal{D}$ respectively. Then if $A' = \{a'_{ij}\} = \{a_{\pi_1^{-1}(i)\pi_2^{-1}(j)}\}$, $Q(A')$ is the set of points of S obtained by applying π_2 to the coordinates of points in $Q(A)$.

3. Every element of $Q(A)$ is admissible.

4. $Q(A)$ is a convex subset of S .

5. If
$$A_1 = \lambda A_0 + \begin{bmatrix} c_1 c_2 \cdots c_n \\ c_1 c_2 \cdots c_n \\ \vdots \\ c_1 c_2 \cdots c_n \end{bmatrix} , \quad \lambda > 0 ,$$
then $Q(A_1) = \{\lambda \tilde{x} + \tilde{c}, \ \tilde{x} \in Q(A_0)\}$, where $\tilde{c} = (c_1, c_2, \ldots, c_n)$.

6. Let $co(A_1^T) = co(A_2^T)$, where A^T is the transpose of A and $co(A)$ is the convex hull of the row vectors of A . If A_1 is obtained from A_2 by deleting j columns from A_2, then $Q(A_1)$ is obtained from $Q(A_2)$ by deleting the corresponding j coordinates from each element of $Q(A_2)$.

<u>Remark:</u> This type of postulate is usually called a "nature duplication" postulate.

7. If for two statistical decision problems, A_1 and A_2, with risk sets S_1 and S_2, the points of S_1 which are both extreme points and admissible

coincide with the points of S_2 which are both extreme points and admissible, then $Q(A_1) = Q(A_2)$.

8. If $\{A_n\}_{n=1}^{\infty}$ converges to A_0 and $x_n \in Q(A_n)$ for every n, then every limit point of $\{x_n\}_{n=1}^{\infty}$ is in $Q(A_0)$. (Here convergence of $\{A_n\}$ is element by element).

In [1], it was exhibited that these postulates are consistent for finite decision problems and a decision rule called "iterated minimax regret" (IMR) was shown to satisfy all eight postulates. IMR is closely related to the rule given by Milnor [18] and is described below.

The iterated minimax regret principle (IMR) selects any element $s_0 \in S$ as optimal which is obtained by the following process.

Let $v_1 = v_1(\theta) = \inf_{\varphi} R_{\varphi}(\theta)$ and let $\mathfrak{L}$ be the normed linear space with, for $s = T(\varphi)$, $\|s\| = \|R_{\varphi}(\theta)\| = \sup_{\theta \in \Theta} R_{\varphi}(\theta)$. Let $z_1 = \inf_{s \in S} \|s - v_1\|$ and let $\{\varepsilon_n\}_{n=1}^{\infty}$ be a sequence of positive numbers with $\lim_{n \to \infty} \varepsilon_n = 0$. Let $Q_1 = S$ and inductively, for $n \geq 1$, define $Q_{n+1} = \{s \in Q_n : \|s - v_n\| \leq z_n + \varepsilon_n z_1\}$ where $v_n = \inf_{\varphi \in T^{-1}(S_n)} R_{\varphi}(\theta)$ and $z_n = \inf_{s \in S_n} \|s - v_n\|$. If $z_1 = \infty$, then all $s \in S$ are said to be optimal. If $z_1 < \infty$, define $Q = \bigcap_{n=1}^{\infty} Q_n$ and choose as s_0 any element of Q .

B. Efron [10] extended these results in part to infinite decision problems. Here some modifications in the postulates must be made and these are listed below in the form given by E. E. Nordbrock [20].

2'. If $h: \Theta \to \Theta'$ is a homeomorphism and if $S' = \{R'_{\varphi}(\theta'); \theta' \in \Theta'; R' = R \circ h\}$, then $Q' = Q \circ h$.

5'. If $S' = \lambda S + c$, where $c = c(\theta)$ is a continuous function of θ, then $Q' = \lambda Q + c$.

6'. Let $S' = \{R_\varphi(\theta')\}$ and $S = \{R_\varphi(\theta) = R|\Theta\}$, where $R|\Theta$ means the restriction of the domain of R to Θ and $\Theta \subset \Theta'$. Then if for every $\theta_0 \in \Theta'$, there is a probability measure μ_{θ_0} on Θ such that for all $\varphi \in \Phi$, we have $R_\varphi(\theta_0) = \int_\Theta R_\varphi(\theta) d\mu_{\theta_0}(\theta)$, then $Q'|\Theta = Q$.

7' If S and S' have a common complete class, then $Q = Q'$.

8' Define $d(S, S') = \max\{\sup_{s \in S} \inf_{r \in S'} \|r - s\|, \sup_{r \in S'} \inf_{s \in S} \|r - s\|\}$. Then if $d(S_n, S) \to 0$ as $n \to \infty$ and if $s^{(n)} \in Q_n$ for all n, and if $d(s^{(n)}, s) \to 0$, then $s \in Q$.

Efron showed that these postulates (1, 2', 3, 4, 5', 6', 7', 8') are satisfied by IMR for S a closed bounded convex set in E_n. He also claimed that with the exception of postulate 1 and the weakening of the conclusion of postulate 8 to $s \in \bar{Q}$, this holds for closed bounded convex sets in L_∞. However, the following counter example shows that inadmissible decision procedures may result.

Example 5.1 Nordbrock [20]. Let $\Theta = \{1, 2, \dots\}$ and let $\mathcal{D} = \{0, 1, 2, \dots\}$. For $\delta = 1, 2, \dots,$ let

$$R_\delta(\theta) = \begin{cases} 0 & \delta = \theta \\ 1 & \delta \neq \theta \end{cases}$$

and let

$$R_0(\theta) = 1 \text{ for all } \theta \in \Theta .$$

Let S be the closed convex hull of $\{R_\delta(\theta)\}$. Then it is easily seen that $\bigcap_{n=1}^{\infty} Q_n = Q_1 = S$ and $s_0 = R_0(\theta)$ is inadmissible.

We now conclude with a statement of Nordbrock's results.

S is said to be weak intrinsically compact (Wald[30]), if for every sequence $\{s_n\} \epsilon S$, there is a subsequence $\{s_{n_k}\}$ and an $s' \epsilon S$ such that $\liminf R_{n_k}(\theta) \geq R'(\theta)$ for all θ, where $s_{n_k} = R_{n_k}(\theta)$ and $s' = R'(\theta)$.

Theorem 5.1. IMR satisfies properties 2', 4, 5', 6' generally. Property 1 holds if S is weak intrinsically compact. Property 8' holds if S is closed and properties 3 and 7 hold if S is compact.

6. Summary. In this exposition, I have attempted to give some illustrations of the possible directions in which the mathematical foundations of statistical decision theory might be developed. The limitations of this volume preclude the extensive development of these ideas which are necessary in order to determine its possible impact on the subject of theoretical statistics. However, it is hoped that this brief exposition will encourage research workers to further examine the implications of the ideas contained herein. At this stage it is not yet apparent whether the material of sections three and four will in fact produce new basic results in statistics as such. To date, it is possible to identify many familiar statistical results in the writings of functional analysts and further display the correspondence between these two areas. The notion of iterated minimax regret developed in the fifth section is handicapped by its apparent incomputability, except in rather artificial examples. The principle has not been successfully applied to any concrete statistical problem as yet.

The author is indebted to Professor Herman Chernoff for some discussions concerning his paper [5] and for calling the two papers by H. Uzawa [27, 28] to his attention.

REFERENCES

[1] F. V. Atkinson, J. D. Church, and B. Harris, Decision procedures for finite decision problems under complete ignorance, Ann. Math. Statist. 35 (1964), 1644-1655.

[2] D. Blackwell and M. A. Girshick, Theory of Games ana Statistical Decisions, John Wiley, New York, 1954.

[3] V. N. Burov, Approximation with constraints in linear normed spaces, I., Ukrain. Mat. Ž. 15 (1963), 3-12. (in Russian).

[4] V. N. Burov, Approximation with constraints in linear normed spaces. II., Ukrain. Mat. Ž. 15 (1963), 135-144. (in Russian).

[5] H. Chernoff, Rational selection of decision functions, Econometrica 22 (1954), 422-443.

[6] J. W. Daniel, Applications and methods for the minimization of functionals, in Non-linear Functional Analysis and Applications, 399-424, Editor L. B. Rall, Academic Press, New York, 1971.

[7] J. W. Daniel, The Approximate Minimization of Functionals, Prentice-Hall, Englewood Cliffs, N. J. (1971).

[8] F. R. Deutsch and P. H. Maserick, Applications of the Hahn-Banach theorem in approximation theory, SIAM Rev., 9 (1967), 516-530.

[9] N. Dunford and J. T. Schwartz, Linear Operators, I., Interscience Publishers, New York, 1957.

[10] B. Efron, Note on decision procedures for finite decision problems under complete ignorance, Ann. Math. Statist., 36 (1965), 691-697.

[11] T. S. Ferguson, Mathematical Statistics, A Decision Theoretic Approach, Academic Press, New York, 1967.

[12] A. L. Garkavi, On a criterion for an element of best approximation, Sibirsk Mat. Ž. 5 (1964), 472-476. (in Russian).

[13] A. L. Garkavi, Duality theorems for approximation by elements of convex sets, Uspehi Mat. Nauk, 16 (1961), 141-145. (in Russian).

[14] I. J. Good, Rational decisions, J. Royal Statist. Soc. Ser. B, 14 (1952), 107-114.

[15] K. Kirchgässner and K. Ritter, On stationary points of nonlinear maximum problems in Banach spaces, SIAM J. Control, 4 (1966), 732-739.

[16] O. Krafft, Programming methods in statistics and probability theory, in Nonlinear Programming, 425-446, Edited by J. B. Rosen, O. L. Mangasarian, and K. Ritter, Academic Press, New York, 1970.

[17] R. D. Luce and H. Raiffa, Games and Decisions, John Wiley, New York, 1957.

[18] J. Milnor, Games against nature, in Decision Processes, Edited by R. M. Thrall, C. H. Coombs, and R. L. Davis, John Wiley, New York, 1954.

[19] L. W. Neustadt, Sufficiency conditions and a duality theory for mathematical programming problems in arbitrary linear spaces, in Nonlinear Programming, 323-348, Edited by J. B. Rosen, O. L. Mangasarian, and K. Ritter, Academic Press, New York, 1970.

[20] E. E. Nordbrock, A rational approach to decision problems, MRC Technical Summary Report #1159, Mathematics Research Center, University of Wisconsin, 1971.

[21] R. Radner and J. Marschak, Note on some proposed decision criteria, in Decision Processes, 61-68, Edited by R. M. Thrall, C. H. Coombs, and R. L. Davis, John Wiley, New York, 1954.

[22] K. Ritter, Optimization theory in linear spaces, III. Mathematical programming in partially ordered Banach spaces, Math. Ann., 184 (1970), 133-154.

[23] G. S. Rubinstein, On a method of investigation of convex sets, Dokl. Akad. Nauk SSSR (N. S.), 102 (1955), 451-454. (in Russian).

[24] L. J. Savage, The theory of statistical decision, J. Amer. Stat. Assoc. 46 (1951), 55-67.

[25] L. J. Savage, The Foundations of Statistics, John Wiley, New York, 1954.

[26] I. Singer, Best approximation in normed linear spaces by elements of linear subspaces (translation of original Romanian book). Publishing House of the Academy of the Socialist Republic of Romania, Bucharest, 1970.

[27] H. Uzawa, A generalization of Laplace criterion for decision problems, Ann. Inst. Statist. Math., 7 (1956), 123-129.

[28] H. Uzawa, Note on the rational selection of decision functions, Econometrica, 25 (1957), 166-174.

[29] F. A. Valentine, Convex Sets, McGraw-Hill Book Company, New York, 1964.

[30] Abraham Wald, Statistical Decision Functions, John Wiley, New York, 1950.