ON ϵ-COMPLETE CLASSES OF DECISION FUNCTIONS[1]

By J. Wolfowitz

Columbia University

An example of an "almost subminimax" solution[2] is given in a paper by Hodges and Lehmann ([1], Section 5, Problem 1). Another paper, written independently by Robbins [2], has put much stress on the idea and given a considerable discussion and many examples. Frank and Kiefer [3] have given a prescription for constructing almost subminimax solutions.

Let $\Omega = \{F\}$ be a set of distribution functions F. The statistician has to make one of a set of decisions in a space $D^t = \{d^t\}$. He takes observations in stages (finite subsets) on an infinite sequence of chance variables $X = X_1, X_2, \cdots$, distributed according to an unknown one of the F's. The statistician employs a decision function δ, a rule (which may involve randomization) which tells him when to stop taking observations and what decision to make when he has stopped taking observations. The risk $r(F, \delta)$ of a decision function (d.f.) δ when F is the distribution function of X is the sum of the expected values of the cost function and the loss function. All these ideas are described rigorously and in detail in the book [4] by Wald, whose notation we adopt. Some familiarity with this book and its ideas will be assumed. The brief résumé given in this paragraph was partly for the purpose of recalling some of the important notation.

An almost subminimax d.f. δ^* may be roughly described as follows. Let δ^{**} be, say, an admissible minimax d.f. For all F's in Ω we have $r(F, \delta^*) < r(F, \delta^{**}) + \epsilon$ with ϵ "small" and positive, while for "most" F's in Ω we have $r(F, \delta^{**}) - r(F, \delta^*)$ equal to a "large" positive number.

An important task of the mathematical statistician is to exhibit a complete class of d.f.'s for a problem under consideration; an essentially complete class is even more useful[3]. The choice of d.f. from among the members of an essentially complete class requires additional principles. A possible principle is to choose a minimax d.f. (There may be more than one minimax d.f., and even more than one admissible minimax d.f.) This might be the course of a very conservative statistician whose ignorance of F is complete. The appeal of an almost subminimax d.f. in preference to a minimax d.f. occurs when the statistician considers ϵ small, or the F's for which $r(F, \delta^*) > r(F, \delta^{**})$ rather "unlikely," or both. There will usually be little difficulty in deciding when ϵ is small, but perhaps considerable difficulty in deciding that some "few" F's are "unlikely" and just what is to be done about them.

Let $\epsilon > 0$ be a fixed number. A d.f. δ_1 will be called ϵ-equivalent to the d.f. δ_2 if $| r(F, \delta_1) - r(F, \delta_2) | \leqq \epsilon$ for all F in Ω. (The relation of ϵ-equivalence is obviously not transitive.) A d.f. δ_1 will be said to be ϵ-better than the d.f. δ_2 if δ_1 and δ_2 are not ϵ-equivalent and if $r(F, \delta_1) \leqq r(F, \delta_2) + \epsilon$ for every F in Ω. A

[1] Research under a contract with the Office of Naval Research.

[2] This represents a slight change in nomenclature from that of Robbins' paper [2].

[3] Provided, of course, that it is not complete.

class C of d.f.'s will be called essentially ϵ-complete if, for any d.f. δ_1 not in C, there exists a d.f. δ_2 which is a member of C and is such that either δ_2 is ϵ-equivalent to δ_1 or δ_2 is ϵ-better than δ_1.

If an almost subminimax d.f. δ^* exists, then δ^* is ϵ-better than the minimax decision function δ^{**}. If the statistician regards differences of at most ϵ in the risk function as negligible, and differences larger than ϵ as meaningful, he will prefer δ^* to δ^{**} and this preference will not depend upon which F's are "few" or "unlikely." We shall show that under certain conditions there exists an essentially ϵ-complete class C_0 which contains only a finite number of elements, and in the next paragraph we will formulate the conditions precisely.

We shall assume that Assumptions 3.3 to 3.7 of Wald [4] hold, and in addition that the available d.f.'s are all such that

$$r(F, \delta) < \infty \quad \text{for every} \quad F \text{ in } \Omega; \tag{1}$$

$$\lim_{k \to \infty} \sum_{j=k}^{\infty} P_j(F, \delta) \operatorname*{Sup}_{x, s^j} c(x; s^j) = 0 \tag{2}$$

uniformly in δ and the F in Ω. Here $P_j(F, \delta)$ is the probability according to F that, when δ is employed, a decision will be reached in exactly j stages of observations, and s^j is any j-stage set of observational indices such that the cost function $c(x; s^j)$ is not $+\infty$ identically in x. It follows from Assumptions 3.5 (II) and 3.6 (III) of [4] that $\operatorname{Sup}_{x, s^j} c(x; s^j)$ is always finite. Equations (1) and (2) are essentially Condition 7 on page 298 of [5]. Note that the statistician may use d.f.'s which require no observations at all.

Under these conditions we shall prove that there exists an essentially ϵ-complete class C_0 whose elements are finite in number.

Define

$$\rho_{1r}(F_1, F_2) = \operatorname*{Sup}_{R} | P\{R \mid F_1\} - P\{R \mid F_2\} |, \tag{3}$$

where R is any Borel set of the r-dimensional Euclidean space, and

$$\rho_2(F_1, F_2) = \operatorname*{Sup}_{d^t \epsilon D^t} | W(F_1, d^t) - W(F_2, d^t) |. \tag{4}$$

Finally, let

$$\rho(F_1, F_2) = \rho_2(F_1, F_2) + \sum_{r=1}^{\infty} 2^{-r} \rho_{1r}(F_1, F_2). \tag{5}$$

It follows easily from Assumption 3.7 that Ω is compact in the sense of the metric $\rho(F_1, F_2)$.

Let $\epsilon > 0$ be given arbitrarily. There exists an $m = m(\epsilon)$ such that

$$\operatorname*{Sup}_{F, \delta} | r(F, \delta^m) - r(F, \delta) | < \frac{\epsilon}{64}, \tag{6}$$

where δ^m is the d.f. δ truncated (in any manner whatever) after m stages of observations. In m stages of observations one can have observations only on the chance variables $X_1, \cdots, X_{m^*}$, where m^* is a finite-valued function of m. Just as in the proof of Theorem 3.3 of [4], one proves that $r(F, \delta^m)$ is continuous with respect to the metric $\rho(F_1, F_2)$ *uniformly* with respect to all available δ^m. We conclude from (6) that there exist a finite number q of F's, say $F_1^*, \cdots, F_q^*$, with the property that for every F in Ω there exists some element F_i^* such that

$$\text{Sup} \mid r(F, \delta) - r(F_i^*, \delta) \mid < \frac{3\epsilon}{64} \tag{7}$$

for all available δ.

We now define

$$\eta(d_1^t, d_2^t) = \operatorname*{Sup}_{F \in \Omega} \mid W(F, d_1^t) - W(F, d_2^t) \mid . \tag{8}$$

There exist a finite number of elements d^t, say $g_1, \cdots, g_l$, such that for every d^t in D^t there exists an element g_i such that $\eta(d^t, g_i) < \epsilon/64$. We shall now temporarily limit the statistician to decisions in $D^* = \{g_1, \cdots, g_l\}$. This means that a d.f. δ is to be replaced by δ^*, where δ^* is derived from δ in the following manner: δ and δ^* are the same except that, whenever δ tells the statistician to make a decision d^t in D^t, δ^* tells the statistician to make a decision in D^* nearest to d^t in the sense of (8). Obviously,

$$\mid r(F, \delta) - r(F, \delta^*) \mid < \frac{\epsilon}{64} \tag{9}$$

for every available δ and every F in Ω.

Consider now the following statistical problem: $\Omega^* = \{F_1^*, \cdots, F_q^*\}$ and $D^* = \{g_1, \cdots, g_l\}$ are the spaces of distributions and decisions, respectively. The statistician is allowed at most m^* observations in various possible steps. The cost function and other assumptions are as before. It follows that the set S of points in Euclidean q-dimensional space $\{r(F_1, \delta^{*m}), \cdots, r(F_q, \delta^{*m})\}$ for every available δ^{*m}, is bounded and convex—bounded because the loss and cost functions are bounded and convex because the totality of available d.f.'s is convex. Obviously a finite number of decision functions corresponding to a finite number of points of S which lie close to the periphery of S constitute an essentially $\epsilon/4$-complete class for the Ω^*, D^* problem with respect to all available d.f.'s which permit at most m stages of observations, and with respect to all d.f.'s which can be obtained from an available d.f. by truncation at the mth stage. Call the resulting class C_0.

It remains to prove that C_0 is essentially ϵ-complete for the problem Ω, D^t, and with respect to all available decision functions. Let δ_0 be any d.f. not in C_0, and consider the d.f. δ_0^{*m}. There are now two possibilities: (a) δ_0^{*m} is $\epsilon/4$-equivalent to some member h of C_0 with respect to the problem Ω^*, D^*; (b) δ_0^{*m} is not $\epsilon/4$-equivalent to any member of C_0 with respect to the problem Ω^*, D^*.

Suppose (a) holds. Let F be any element of Ω. By use of (7) we obtain that

$$| r(F, \delta_0^{*m}) - r(F, h) | < \frac{11\epsilon}{32}. \tag{10}$$

From (9) we have

$$| r(F, \delta_0^{m}) - r(F, h) | < \frac{23\epsilon}{64}. \tag{11}$$

Finally from (6) we have that

$$| r(F, \delta_0) - r(F, h) | < \frac{3\epsilon}{8}, \tag{12}$$

so that δ_0 and h are surely ϵ-equivalent.

Suppose now that (b) holds. Let t be $\epsilon/4$-better than δ_0^{*m} with respect to the problem Ω^*, D^*. For any F such that there exists a nearest F_i^* for which

$$| r(F_i^*, \delta_0^{*m}) - r(F_i^*, t) | < \frac{\epsilon}{4} \tag{13}$$

we have, just as in the previous case,

$$| r(F, \delta_0) - r(F, t) | < \frac{3\epsilon}{8}. \tag{14}$$

If for some F (13) does not hold we must have, for a nearest F_i^*, since t is $\epsilon/4$-better than δ_0^{*m} with respect to the problem Ω^*, D^*,

$$r(F_i^*, \delta_0^{*m}) > r(F_i^*, t) + \frac{\epsilon}{4}. \tag{15}$$

Now

$$| r(F, \delta) - r(F_i^*, \delta_0^{*m}) | < \frac{5\epsilon}{64},$$

$$| r(F, t) - r(F_i^*, t) | < \frac{3\epsilon}{64}.$$

Hence

$$r(F, \delta_0) > r(F, t) + \frac{\epsilon}{8}. \tag{16}$$

Since either (14) or (16) must hold we conclude that t is either ϵ-equivalent to δ_0 or t is ϵ-better than δ_0. This proves that C_0 is an essentially ϵ-complete class for the original problem.

REFERENCES

[1] J. L. Hodges and E. L. Lehmann, "Some problems in minimax point estimation," *Annals of Math. Stat.*, Vol. 21 (1950), pp. 182–197.

[2] HERBERT ROBBINS, "Asymptotically subminimax solutions of compound statistical decision problems," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1951.
[3] P. FRANK AND J. KIEFER, "Almost subminimax and biased minimax procedures," *Annals of Math. Stat.*, Vol. 22 (1951), pp. 465–468.
[4] A. WALD, *Statistical Decision Functions*, John Wiley and Sons, 1950.
[5] A. WALD, "Foundations of a general theory of sequential decision functions," *Econometrica*, Vol. 15 (1947), pp. 279–313.

ALMOST SUBMINIMAX AND BIASED MINIMAX PROCEDURES[1]

BY P. FRANK AND J. KIEFER

Columbia University

Robbins [1] emphasized the notion of an "almost subminimax" procedure[2] and gave an example of such a procedure. The examples in this paper have been constructed with a view to simplicity and to the indication of the underlying mechanism which makes subminimax solutions exist in certain decision problems. At the same time we point out another potentially undesirable property of a minimax procedure—biasedness.

All our examples fall within the following framework. A sample of one is taken from a population whose distribution is one of n given distributions: $F_1(x), F_2(x), \cdots, F_n(x)$. There are n decisions: $d_1, \cdots, d_n$. The weight function is $W(F_i, d_j) = 0$ if $i = j$ and $= 1$ otherwise. Instead of a finite number of F's, we may have a sequence of F's with a corresponding sequence of decisions. In all our examples each of the F's will be a uniform distribution over a finite interval of the x-axis, and our decision procedures will be randomized. These restrictions are made only for arithmetical simplicity.

With this setup, the risk when F_i is the true distribution is equal to the probability of not making decision d_i, which we will denote $r(F_i)$. We will not give an exact definition of an almost subminimax procedure, but just say that a procedure is almost subminimax if its maximum risk is "a little greater" than that of the minimax procedure (which risk is the same for all minimax procedures in our examples) and on the other hand its risk is "a lot less" than that of the minimax for "most of" the F's. Our examples will conform with this "definition" for almost any reasonable interpretation of the phrases in the quotes.

The first example will give an indication of the mechanism which makes a subminimax example possible. Let $F_1(x)$ be the uniform distribution on the interval $1 - a$ to 1, where $a > 0$ and small. Let $F_2(x)$ be the uniform distribution on the interval 0 to 1. An admissible minimax procedure to decide between d_1

[1] Research done under a contract with the Office of Naval Research.

[2] The examples of this paper fall into the framework of the definition in [1] of an "asymptotically subminimax solution" if each example is replaced by a sequence of examples whose a's approach zero. The present nomenclature was suggested as more suitable here.