ASYMPTOTICALLY SUBMINIMAX SOLUTIONS OF COMPOUND STATISTICAL DECISION PROBLEMS

HERBERT ROBBINS
UNIVERSITY OF NORTH CAROLINA

1. Summary

When statistical decision problems of the same type are considered in large groups the minimax solution may not be the "best," since there may exist solutions which are "asymptotically subminimax." This is shown in detail for a classical problem in the theory of testing hypotheses.

2. Introduction

Consider the following *simple statistical decision problem.* The random variable x is normally distributed with variance 1 and mean θ, where θ is known to have one of the two values ± 1. It is required to decide, on the basis of a single observation on x, whether the true value of θ is 1 or -1, in such a way as to minimize the probability of error.

For any decision rule R the probability of error will depend on the true value of θ. Let

$$(1) \qquad \eta(R) = P[\text{error} \mid R, \theta = -1], \quad \delta(R) = P[\text{error} \mid R, \theta = 1].$$

By a suitable choice of R we can give to $\eta(R)$ any desired value between 0 and 1; unfortunately, if R is chosen so that $\eta(R)$ is near 0 then $\delta(R)$ will be near 1, and in this circumstance lies the problem.

For any constant c let R_c be the decision rule which asserts "$\theta = \operatorname{sgn}(x - c)$"; thus in using R_c we assert "$\theta = 1$" if $x > c$ and "$\theta = -1$" if $x < c$. Then

$$(2) \qquad \eta(R_c) = \int_c^\infty f(x+1)\,dx = F(-1-c),$$

$$\delta(R_c) = \int_{-\infty}^c f(x-1)\,dx = F(-1+c),$$

where we have set

$$(3) \qquad f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad F(x) = \int_{-\infty}^x f(y)\,dy = 1 - F(-x).$$

It is clear from (2) that

(4) for any number η between 0 and 1 there exists a number $c = c(\eta)$ such that $\eta(R_c) = \eta$.

Moreover, using the fundamental lemma of Neyman and Pearson it can be shown that

(5) for any c and any decision rule R such that $\eta(R) \leqq \eta(R_c)$, $\delta(R) \geqq \delta(R_c)$.

It follows from (4) and (5) that we need only admit into competition decision rules of the form R_c, but it remains to choose the proper value for c.

An examination of (2) shows that the value $c = 0$ is of particular interest. Let us denote by $\tilde{R}$ the rule R_c with $c = 0$; thus in using $\tilde{R}$ we assert "$\theta = \text{sgn}(x)$." Now for any c,

$$(6)\quad \max[\eta(R_c), \delta(R_c)] = \max[F(-1-c), F(-1+c)] = F(-1+|c|),$$

and this attains its minimum value $F(-1) = .1587$ for $c = 0$. It follows from (4)–(6) that for any decision rule R (not necessarily of the form R_c),

$$(7)\quad \max[\eta(R), \delta(R)] \geqq \max[\eta(\tilde{R}), \delta(\tilde{R})] = \eta(\tilde{R}) = \delta(\tilde{R}) = F(-1),$$

and it can be shown that this inequality is strict unless $R = \tilde{R}$, where we regard two decision rules as equal if and only if they arrive at the same decision with probability 1 for all values of θ. Thus $\tilde{R}$ is the unique decision rule which minimizes the maximum possible probability of error, or, in Wald's terminology, $\tilde{R}$ is the unique *minimax* decision rule. As is often the case with minimax solutions, $\tilde{R}$ has the agreeable property that the probability of error is independent of the value of θ.

The unique minimax property of $\tilde{R}$ is a strong argument in favor of $\tilde{R}$ but is not in itself a compelling reason for regarding $\tilde{R}$ as the "best" solution of the decision problem. Suppose, for the sake of argument, that there existed another rule R for which

$$(8)\quad \eta(R) = F(-1) + \epsilon_1, \quad \delta(R) = \epsilon_2,$$

where both ϵ_1 and ϵ_2 are small positive numbers, say less than .001. This would not contradict the minimax property (7) of $\tilde{R}$. Still, there would be little doubt that R is preferable to $\tilde{R}$, for in using R we would achieve a *much smaller* probability of error when $\theta = 1$ at the cost of only a *slightly greater* probability of error when $\theta = -1$.

Of course, there is no such rule R. In fact, it follows easily from the previous discussion that for any rule R such that (8) holds,

$$(9)\quad \epsilon_1 + \epsilon_2 \geqq F(-1),$$

equality holding only for $\epsilon_1 = 0$, $\epsilon_2 = F(-1)$, $R = \tilde{R}$. Hence ϵ_1 and ϵ_2 cannot *both* be made small, and the gain, $F(-1) - \epsilon_2$, of any rule R over $\tilde{R}$ when $\theta = 1$ is more than balanced by the loss, ϵ_1, when $\theta = -1$. The fact that any improvement over $\tilde{R}$ when $\theta = 1$ must be accompanied by an even greater deterioration when $\theta = -1$, goes beyond the minimax property of $\tilde{R}$ and greatly strengthens the view that $\tilde{R}$ is in fact the "best" decision rule.

Statistical decision problems often occur, or can be considered, in large groups. Thus let $x_1, \ldots, x_n$ be independent random variables, each normally distributed with variance 1, and with respective means $\theta_1, \ldots, \theta_n$, where $\theta_i = \pm 1$, $i = 1,$

. . . , n. *No relation whatever is assumed to hold among the unknown parameters* θ_i. To emphasize this point, x_1 could be an observation on a butterfly in Ecuador, x_2 on an oyster in Maryland, x_3 the temperature of a star, and so on, all observations being taken at different times. Let it be required to decide, on the basis of the observed values $x_1, \ldots, x_n$, for every $i = 1, \ldots, n$ whether $\theta_i = 1$ or -1, in such a way as to minimize the expected total number of errors. The parameter space Ω of this *compound statistical decision problem* consists of the 2^n points $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, $\theta_i = \pm 1$. It is natural to suppose that the "best" solution of the compound problem consists in applying to each of the x_i the "best" solution of the original simple problem and therefore in asserting "$\theta_i = \operatorname{sgn}(x_i)$, $i = 1, \ldots, n$." Let us again call this (compound) decision rule $\tilde{R}$. It is indeed true that $\tilde{R}$ remains for every n the unique minimax solution, in that for any rule $R \neq \tilde{R}$ which may be applied in the compound problem,

$$\max_{\boldsymbol{\theta} \in \Omega} [\text{exp. no. of errors} \,|\, R, \boldsymbol{\theta}] > \max_{\boldsymbol{\theta} \in \Omega} [\text{exp. no. of errors} \,|\, \tilde{R}, \boldsymbol{\theta}] .$$

We shall see, however, that *for large* n, $\tilde{R}$ *can no longer be regarded as the "best" decision rule in the compound problem.* Nor is this due to any special property of the simple decision problem with which we began; it lies rather in the fundamental operation of "compounding" and will occur in a large class of compound decision problems.

3. Statement of the compound decision problem. The rule $\tilde{R}$

Let

$$x_1, \ldots, x_n \tag{10}$$

be independent random variables, each normally distributed with variance 1, and with respective means

$$\theta_1, \ldots, \theta_n, \qquad \theta_i = \pm 1 . \tag{11}$$

On the basis of the observed sample (10) we are to decide for every $i = 1, \ldots, n$ whether the true value of θ_i is 1 or -1, in such a way as to minimize the expected total number of errors.

Denote by Ω the set of all 2^n possible parameter points $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, $\theta_i = \pm 1$. For any $\boldsymbol{\theta}$ in Ω the density function of the sample vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ is, by hypothesis,

$$\phi(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}} e^{-\sum_{i=1}^{n} (x_i - \theta_i)^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-(\boldsymbol{x}^2 + n)/2} \cdot e^{\boldsymbol{\theta}\boldsymbol{x}} . \tag{12}$$

If $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are any two points of Ω take

$$w(\boldsymbol{\theta}', \boldsymbol{\theta}) = \frac{1}{n} (\text{no. of } i \text{ for which } \theta_i' \neq \theta_i) = \frac{1}{2n} \sum_{i=1}^{n} \left| \theta_i' - \theta_i \right| \tag{13}$$

as a measure of the loss involved when the true parameter point is $\boldsymbol{\theta}$ and the decision "$\boldsymbol{\theta} = \boldsymbol{\theta}'$" is taken. [The factor $1/n$ in (13) is used in order to stabilize certain later formulas as n varies.] Order the points of Ω arbitrarily as $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(2^n)}$.

The most general (randomized) decision rule R amounts to specifying as a function of x a probability distribution $p_j(x), j = 1, \ldots, 2^n$, on Ω:

$$(14)\quad R:\ p_j(x), \qquad j = 1, \ldots, 2^n, \qquad p_j(x) \geqq 0, \qquad \sum_{j=1}^{2^n} p_j(x) \equiv 1.$$

For given x the rule R asserts "$\theta = \theta^{(j)}$" with probability $p_j(x)$. When θ is the true parameter point the expected loss in using R is given by the *risk function*

$$(15)\quad L(R, \theta) = \int \Big[\sum_{j=1}^{2^n} p_j(x) w(\theta^{(j)}, \theta) \Big] \phi(x, \theta)\, dx$$

$$= \frac{1}{2n} \sum_{i=1}^{n} \int \Big[\sum_{j=1}^{2^n} p_j(x) \left| \theta_i^{(j)} - \theta_i \right| \Big] \phi(x, \theta)\, dx.$$

This can be put into a more convenient form as follows. Let

$$(16)\quad u_i(x) = \frac{1}{2} \sum_{j=1}^{2^n} p_j(x)(1 + \theta_i^{(j)}), \qquad i = 1, \ldots, n,$$

$$= \text{conditional probability, given } x, \text{ of deciding that } \theta_i = 1,\ 0 \leqq u_i(x) \leqq 1.$$

Then

$$(17)\quad \sum_{j=1}^{2^n} p_j(x) \left| \theta_i^{(j)} - \theta_i \right| = \begin{cases} 2u_i(x) & \text{if } \theta_i = -1, \\ 2[1 - u_i(x)] & \text{if } \theta_i = 1, \end{cases}$$

$$= 1 + \theta_i - 2\,\mathrm{sgn}(\theta_i)\, u_i(x) \quad \text{for } \theta_i = \pm 1.$$

For any θ in Ω let

$$(18)\quad p(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (1 + \theta_i) = \frac{1}{n} (\text{no. of } i \text{ for which } \theta_i = 1),\ 0 \leqq p(\theta) \leqq 1.$$

Then from (15)–(18) we have

$$(19)\quad L(R, \theta) = p(\theta) - \frac{1}{n} \sum_{i=1}^{n} \mathrm{sgn}(\theta_i) \int \phi(x, \theta)\, u_i(x)\, dx.$$

This shows that $L(R, \theta)$ (although in general not R itself) depends only on the n functions (16).

The maximum likelihood estimate of the true parameter point θ is, by (12),

$$\hat{\theta} = [\mathrm{sgn}(x_1), \ldots, \mathrm{sgn}(x_n)].$$

The corresponding decision rule will be denoted by $\tilde{R}$:

$$(20)\quad \tilde{R}: \text{"}\theta_i = \mathrm{sgn}(x_i), \qquad i = 1, \ldots, n.\text{"}$$

For the rule $\tilde{R}$, $u_i(x) = 1$ or 0 according as $x_i > 0$ or $x_i < 0$, so that from (19),

$$(21)\quad L(\tilde{R}, \theta) = \frac{1}{n} \sum_{i=1}^{n} \Big[\frac{1 + \theta_i}{2} - \mathrm{sgn}(\theta_i) \int_{x_i > 0} \phi(x, \theta)\, dx \Big].$$

Using the notation of (3), (21) becomes

$$(22)\qquad L(\tilde{R}, \theta) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1+\theta_i}{2} - \operatorname{sgn}(\theta_i)\int_0^\infty f(x_i - \theta_i)\,dx_i\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1+\theta_i}{2} - \operatorname{sgn}(\theta_i)\,F(\theta_i)\right] = \frac{1}{n}\sum_{i=1}^{n}F(-1) \equiv F(-1) = .1587$$

for every θ in Ω. Thus $\tilde{R}$ has the constant risk $F(-1)$ no matter what the true parameter point θ.

Returning to the general case where R is any decision rule with associated functions (16) we shall consider certain weighted sums of $L(R, \theta)$ taken over all θ in Ω. For any $k = 0, 1, \ldots, n$ let Ω_k denote the set of all θ in Ω for which $p(\theta) = k/n$; thus $\theta \in \Omega_k$ if exactly k of its components are 1. Let a function $h(\theta) \geqq 0$, $\not\equiv 0$ be defined on Ω such that $h(\theta) = \text{constant} = b_k$ for $\theta \in \Omega_k$, $k = 0, 1, \ldots, n$. Then from (19) we have

$$(23)\qquad \sum_{\Omega} h(\theta) L(R, \theta) = \sum_{\Omega} h(\theta)\, p(\theta)$$

$$-\frac{1}{n}\sum_{i=1}^{n}\int\left[\sum_{\Omega} h(\theta)\operatorname{sgn}(\theta_i)\,\phi(x, \theta)\right]u_i(x)\,dx\,.$$

This will be a minimum with respect to R for given $h(\theta)$ [in Wald's terminology, R will be a "Bayes solution" corresponding to $h(\theta)$] if and only if for a.e. x,

$$(24)\qquad u_i(x) = \begin{cases} 1 \text{ if } \displaystyle\sum_{\Omega} h(\theta)\operatorname{sgn}(\theta_i)\,\phi(x, \theta) > 0\,, \\ 0 \text{ otherwise}\,. \end{cases}$$

Let

$$(25)\qquad \Omega_{k,i}^{+} = \text{all } \theta \text{ in } \Omega_k \text{ for which } \theta_i = 1\,,$$

$$\Omega_{k,i}^{-} = \text{all } \theta \text{ in } \Omega_k \text{ for which } \theta_i = -1\,,$$

so that

$$\Omega_k = \Omega_{k,i}^{+} + \Omega_{k,i}^{-}, \quad k = 0, 1, \ldots, n\,; \quad i = 1, \ldots, n\,.$$

Then (24) asserts that $u_i(x) = 1$ when

$$(26)\qquad \sum_{k=0}^{n} b_k\left[\sum_{\Omega_{k,i}^{+}}\phi(x, \theta) - \sum_{\Omega_{k,i}^{-}}\phi(x, \theta)\right] > 0\,.$$

Multiplying by the positive factor $(2\pi)^{n/2}e^{(x^2+n)/2}\cdot e^{1x}$, where $1 = (1, \ldots, 1)$, (26) is seen to be equivalent to

$$(27)\qquad \sum_{k=0}^{n} b_k\left[\sum_{\Omega_{k,i}^{+}} e^{(1+\theta)x} - \sum_{\Omega_{k,i}^{-}} e^{(1+\theta)x}\right] > 0\,.$$

Let

$$(28)\qquad S_k^{(i)} = \sum e^{2(x_{j_1}+\cdots+x_{j_k})}\,, \qquad k = 1, \ldots, n-1\,,$$

$$S_{-1}^{(i)} = S_n^{(i)} = 0\,, \qquad S_0^{(i)} = 1$$

where the summation is over all $\binom{n-1}{k}$ combinations of the integers $1, \ldots, i-1, i+1, \ldots, n$ taken k at a time. Then (27) may be written as

$$\sum_{k=0}^{n} b_k \left[e^{2x_i} S_{k-1}^{(i)} - S_k^{(i)} \right] > 0 ,$$

or, finally, as

$$x_i > \tfrac{1}{2} \ln \frac{\sum_{k=0}^{n-1} b_k S_k^{(i)}}{\sum_{k=0}^{n-1} b_{k+1} S_k^{(i)}} . \tag{29}$$

It follows that $\sum_{\Omega} h(\theta) L(R, \theta) =$ *min. for the (nonrandomized) rule*

$$R: \text{ ``}\theta_i = \operatorname{sgn}\left(x_i - \tfrac{1}{2} \ln \frac{\sum_{k=0}^{n-1} b_k S_k^{(i)}}{\sum_{k=0}^{n-1} b_{k+1} S_k^{(i)}} \right), \qquad i = 1, \ldots, n \text{ .''} \tag{30}$$

If we regard two rules as equal if and only if they give the same decision with probability 1 for all θ in Ω, then the minimizing rule (30) is unique.

Example 1. $b_k = 1, k = 0, 1, \ldots, n$. In this case (30) shows that $\sum_{\Omega} L(R, \theta) =$ min. for $R = \tilde{R}$ defined by (20). Since

$$\sum_{\Omega} L(R, \theta) > \sum_{\Omega} L(\tilde{R}, \theta) \qquad \text{for } R \neq \tilde{R} , \tag{31}$$

$$L(\tilde{R}, \theta) \equiv F(-1) ,$$

it follows that $\tilde{R}$ is the *unique minimax decision rule:*

$$\max_{\theta \in \Omega} L(R, \theta) > \max_{\theta \in \Omega} L(\tilde{R}, \theta) = F(-1) \qquad \text{for every } R \neq \tilde{R} . \tag{32}$$

The result (32) for $n = 1$ is well known. Our purpose in proving it for arbitrary n is to show that $\tilde{R}$ remains the unique minimax solution even when we admit rules of the general type (14) which make the decision on each θ_i depend on the whole sample (10). For $\tilde{R}$, the decision on θ_i depends only on x_i. This agrees with the fact that, since the components (10) are independent, and since no relation is assumed to hold among the components (11) (that is, since the true parameter point θ can be any point in Ω), it is x_i alone which contains "information" about the value of θ_i.

From the minimax point of view our decision problem is completely solved in favor of $\tilde{R}$ by (32). Nevertheless, we shall consider two other examples of decision rules obtained by minimizing weighted sums.

We shall call a decision rule R *symmetric* if $L(R, \theta) =$ constant $= c_k$ for $\theta \in \Omega_k$, $k = 0, 1, \ldots, n$. [Any rule of the form (30) is easily seen to be symmetric.] In

the class of symmetric rules it is of interest to minimize the sum $\sum_{k=0}^{n} c_k$. This will be done in the following example.

Example 2. $b_k = \binom{n}{k}^{-1}$, $k = 0, 1, \ldots, n$. (30) shows that

$$\sum_{k=0}^{n} \binom{n}{k}^{-1} \sum_{\Omega_k} L(R, \theta) = \min.$$

in the class of all decision rules, symmetric or not, for the symmetric rule

$$\bar{R}\text{: ``}\theta_i = \operatorname{sgn}\left(x_i - \tfrac{1}{2} \ln \frac{\sum_{k=0}^{n-1} \binom{n}{k}^{-1} S_k^{(i)}}{\sum_{k=0}^{n-1} \binom{n}{k+1}^{-1} S_k^{(i)}} \right), \qquad i = 1, \ldots, n \text{ .''} \tag{33}$$

Let $L(\bar{R}, \theta) = \bar{c}_k$ for θ in Ω_k, $k = 0, 1, \ldots, n$. If R is *any* symmetric rule with $L(R, \theta) = c_k$ for θ in Ω_k then

$$\sum_{k=0}^{n} c_k = \sum_{k=0}^{n} \binom{n}{k}^{-1} \sum_{\Omega_k} L(R, \theta),$$

since there are $\binom{n}{k}$ points in Ω_k. *Hence* $\bar{R}$ *defined by* (33) *minimizes* $\sum_{k=0}^{n} c_k$ *in the class of all symmetric decision rules.* Since $\bar{R} \neq \tilde{R}$ it follows incidentally that

$$\frac{1}{n+1} \sum_{k=0}^{n} \bar{c}_k < F(-1). \tag{34}$$

As a final example we shall consider the decision problem when the parameter space is some fixed Ω_k. This corresponds to the case when the *number* k (but not the positions) of the values 1 (and hence also of the values -1) in the sequence (11) is known.

Example 3. $b_k = 1$ for some fixed k, $0 \leqq k \leqq n$, and $b_j = 0$ for $j \neq k$. Then $\sum_{\Omega_k} L(R, \theta) = \min.$ uniquely for the symmetric rule

$$\tilde{R}_k\text{: ``}\theta_i = \tilde{R}_k \operatorname{sgn}\left(x_i - \tfrac{1}{2} \ln \frac{S_k^{(i)}}{S_{k-1}^{(i)}} \right), \qquad i = 1, \ldots, n \text{ .''} \tag{35}$$

We shall not attempt to determine numerically the constant $L(\tilde{R}_k, \theta)$, $\theta \in \Omega_k$ [it is, of course, $< F(-1)$]. As in example 1 it follows that $\tilde{R}_k$ is the unique minimax decision rule when θ is restricted to Ω_k:

$$\max_{\theta \in \Omega_k} L(R, \theta) > \max_{\theta \in \Omega_k} L(\tilde{R}_k, \theta) \qquad \text{for every } R \neq \tilde{R}_k. \tag{36}$$

This result is somewhat surprising, as the following considerations show. For

definiteness take $k = 1$, so that $\tilde{R}_k$ becomes

$$\tilde{R}_1\text{: ``}\theta_i = \operatorname{sgn}\left(e^{2x_i} - \sum_{j \neq i} e^{2x_j}\right), \qquad i = 1, \ldots, n\text{ .''}$$

For $n > 2$ the probability of the decision "$\theta = (-1, -1, \ldots, -1)$" is positive (since this decision will be taken when all the x_i are nearly equal) even though it is known to involve exactly one error!

A more plausible rule than $\tilde{R}_1$ when θ is known to lie in Ω_1 would be to assign the value $\theta_i = 1$ to that i for which $x_i = \max(x_1, \ldots, x_n)$ and $\theta_j = -1$ for $j \neq i$. This rule, call it R, always assigns to θ a value in Ω_1 and has a constant risk in Ω_1, as does $\tilde{R}_1$, but from (36) it follows that $L(R, \theta) > L(\tilde{R}_1, \theta)$ for $\theta \in \Omega_1$, so that $\tilde{R}_1$ is uniformly better than R in Ω_1. Of course, if one is restricted to decision rules which assign to θ a value in Ω_1 then R is presumably minimax. Corresponding remarks hold for $k = 2, \ldots, n - 1$.

4. R^*, a competitor of $\tilde{R}$

We have proved (32) that for the unrestricted compound decision problem where θ is known only to lie in Ω, the rule $\tilde{R}$ defined by (20) is the unique minimax solution. We now make the, perhaps surprising, statement that *for large values of n there are strong reasons for regarding $\tilde{R}$ as a relatively poor decision rule.* In support of this assertion we propose the following rule R^* as a competitor of R. Let

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{37}$$

$$x^* = \begin{cases} \infty & \text{if } \bar{x} \leqq -1, \\ \frac{1}{2}\ln\dfrac{1-\bar{x}}{1+\bar{x}} & \text{if } -1 < x < 1, \\ -\infty & \text{if } \bar{x} \geqq 1, \end{cases}$$

$$R^*\text{: ``}\theta_i = \operatorname{sgn}(x_i - x^*), \qquad i = 1, \ldots, n\text{ .''}$$

Observe that R^* makes the decision on each θ_i depend on all the components (10) and not solely on x_i. Now it may be that the n populations from which the x_i are drawn are entirely different and completely unrelated, as in the last paragraph of section 2. The use of the "hybrid" mean $\bar{x}$ might then seem to be meaningless physically and pointless statistically. Furthermore, the rule R^* is not "admissible" in Wald's sense; that is, there exists a rule R such that $L(R, \theta) \leqq L(R^*, \theta)$ for every θ in Ω, the strict inequality holding for at least one, and possibly all, θ. This is known to follow from the fact that R^* is not of the form (30) of the "Bayes solutions" of the present problem. Nevertheless, on the principle that the proof of the pudding lies in the eating, let us compare R^* with R by computing the risk function $L(R^*, \theta)$.

For any $0 \leqq p \leqq 1$ and any $n = 1, 2, \ldots$, let [see (3) for notation]

$$(38)\quad h(p,n) = pF(-2p\sqrt{n}) + (1-p)F[-2(1-p)\sqrt{n}] + \int_{-2(1-p)\sqrt{n}}^{2p\sqrt{n}} \left\{ pF\left[\sqrt{\frac{n}{n-1}}\left(-1-\frac{x}{\sqrt{n}}+\tfrac{1}{2}\ln\frac{1-p+\frac{x}{2\sqrt{n}}}{p-\frac{x}{2\sqrt{n}}}\right)\right] + (1-p)F\left[\sqrt{\frac{n}{n-1}}\left(-1-\frac{x}{\sqrt{n}}-\tfrac{1}{2}\ln\frac{1-p+\frac{x}{2\sqrt{n}}}{p-\frac{x}{2\sqrt{n}}}\right)\right]\right\} f(x)\,dx .$$

It is plausible from inspection of (38), and it can be proved rigorously, that

$$(39)\quad \lim_{n\to\infty} h(p,n) = h(p) = pF\left(-1+\tfrac{1}{2}\ln\frac{1-p}{p}\right) + (1-p)F\left(-1-\tfrac{1}{2}\ln\frac{1-p}{p}\right)$$

uniformly for all $0 \leqq p \leqq 1$. We note also that

$$(40)\quad h(p,n) = h(1-p,n),\quad h(p) = h(1-p),\quad h(0) = h(1) = 0,$$
$$h(.5,n) > F(-1),\quad h(p) < F(-1), \text{ for } p \neq .5,\quad h(.5) = F(-1).$$

By elementary calculation which we omit here it can be shown that

$$(41)\qquad L(R^*,\theta) = h[p(\theta),n],$$

from which it follows that

$$(42)\qquad \lim_{n\to\infty} \{L(R^*,\theta) - h[p(\theta)]\} = 0$$

uniformly for all θ in Ω.

A few values of $h(p)$ and $h(p, 100)$ are given in table I, computed by Mr. J. F. Hannan. [The entries for $h(p, 100)$ are averages of strict upper and lower bounds and are not guaranteed beyond two significant figures.] From the table we see that

TABLE I

p	$F(-1)$	$h(p)$	$h(p, 100)$	$h(p, 1000)$
0.0 or 1.0	.1587	0	.0041	
.1 or .9	.1587	.0691	.0763	
.2 or .8	.1587	.1121	.1174	
.3 or .7	.1587	.1387	.1439	
.4 or .6	.1587	.1538	.1591	
.5	.1587	.1587	.1628	.1591

for $n = 100$, R^* has a *slightly higher* risk than $\tilde{R}$ for p near .5 and a *much lower* risk for p near 0 or 1. As $n \to \infty$, this phenomenon becomes more pronounced. Since (39) and the last two relations in (40) hold, we call R^* an *asymptotically subminimax* decision rule as $n \to \infty$.

A statistical decision problem is sometimes regarded as a game between the statistician S and Nature [1]. In the present problem if S should use the decision rule R^* then Nature could counter by seeing to it that $p(\theta) \simeq .5$. Since $L(R^*, \theta) > L(\tilde{R}, \theta) = F(-1)$ for $p(\theta) \simeq .5$, S would do better, as far as expectations are

concerned, to use $\tilde{R}$. But if Nature is not an opponent but a neutral observer of the game then $p(\theta)$ may not be $\simeq .5$, and in using R^* rather than $\tilde{R}$, S would be balancing the possibility of a slightly higher risk in return for that of a much lower one. As $n \to \infty$, the set of values of $p(\theta)$ for which $L(R^*, \theta) > L(\tilde{R}, \theta)$ converges to the single point .5 and the excess of $L(R^*, \theta)$ over $L(\tilde{R}, \theta)$ in the neighborhood of this point tends to 0, while the excess of $L(\tilde{R}, \theta)$ over $L(R^*, \theta)$ near $p(\theta) = 0$ or 1 tends to $F(-1)$. Even for large n this is not, of course, a compelling reason for preferring R^* to $\tilde{R}$, especially if there is reason to believe that $p(\theta)$ *is* near .5, but we shall not labor this point here.

The reader will have observed that R^* can only be used in applications in which all the values (10) are at hand before any of the individual decisions concerning the θ_i are to be made. This will often be the case in practice. Even when it is not, R^* can be used, after all the values (10) are known, to supersede preliminary decisions based, say, on $\tilde{R}$, or perhaps on some rule which uses the values $x_1, \ldots, x_i$ to decide the value of θ_i.

We emphasize that R^* is by no means advanced as in any sense a "best" rule. Its chief virtue as an asymptotically subminimax rule is its comparative simplicity, both in application and in the computation of its risk function (38). A possible candidate for a rule superior to R^* in every respect save simplicity is the rule $\bar{R}$ defined by (33); unfortunately, the risk function $L(\bar{R}, \theta)$ seems difficult to compute. It is possible that $\bar{R}$ is uniformly better than R^*. On the other hand, it may be that the limiting value of $L(\bar{R}, \theta)$ as $n \to \infty$ and $p(\theta) \to p$ is $h(p)$, in which case $\bar{R}$ and R^* would be asymptotically equivalent in performance. Finally, it is possible that no rule has a limiting risk function uniformly below $h(p)$, in which case R^* would be "asymptotically admissible."

In the preceding discussion the rule R^* was introduced without motivation. In what follows we shall show how R^* came to be considered, in a way which indicates that the existence of asymptotically subminimax decision functions is to be expected in a wide class of problems.

5. Heuristic motivation for R^*

A decision rule R with corresponding functions (16) will be called *simple* if for some function $u(x)$,

$$u_i(\boldsymbol{x}) = u(x_i), \qquad i = 1, \ldots, n. \tag{43}$$

(For $n = 1$ any rule is simple. For $n > 1$, of the specific rules $\tilde{R}$, $\bar{R}$, $\tilde{R}_k$, R^* considered thus far, only $\tilde{R}$ is simple.) For any simple rule R (19) becomes

$$\begin{aligned} L(R, \theta) &= p(\theta) - \frac{1}{n}\sum_{i=1}^{n} \operatorname{sgn}(\theta_i) \int \phi(\boldsymbol{x}, \theta)\, u(x_i)\, d\boldsymbol{x} \\ &= p(\theta) - \int \{ p(\theta) f(x-1) - [1 - p(\theta)] f(x+1) \} u(x)\, dx. \end{aligned} \tag{44}$$

This shows incidentally that every simple rule is "symmetric" (definition in section 3, preceding example 2) and that for fixed R, $L(R, \theta)$ is a linear function of $p(\theta)$.

Now let λ be any constant, $0 \leqq \lambda \leqq 1$, and choose the function $u(x)$, $0 \leqq u(x) \leqq 1$, so as to maximize the integral

$$\int [\lambda f(x+1) - (1-\lambda) f(x-1)]\, u(x)\, dx . \tag{45}$$

This occurs if and only if for a.e. x,

$$u(x) = u_\lambda(x) = \begin{cases} 1 & \text{if } \lambda f(x+1) - (1-\lambda) f(x-1) > 0 , \\ 0 & \text{otherwise} , \end{cases} \tag{46}$$

which determines the simple rule

$$R_\lambda\text{: } \text{“}\theta_i = \operatorname{sgn}\left(x_i - \tfrac{1}{2}\ln\frac{1-\lambda}{\lambda}\right), \qquad i = 1, \ldots, n \text{.”} \tag{47}$$

It follows that *when* $p(\boldsymbol{\theta}) = \lambda$, R_λ *minimizes* $L(R, \boldsymbol{\theta})$ *in the class of all simple rules.* The risk function of R_λ is, by (44), for any λ and any $\boldsymbol{\theta}$,

$$\begin{aligned} L(R_\lambda, \boldsymbol{\theta}) = {} & p(\boldsymbol{\theta}) F\left(-1 + \tfrac{1}{2}\ln\frac{1-\lambda}{\lambda}\right) \\ & + [1 - p(\boldsymbol{\theta})] F\left(-1 - \tfrac{1}{2}\ln\frac{1-\lambda}{\lambda}\right). \end{aligned} \tag{48}$$

The family of simple rules R_λ, $0 \leqq \lambda \leqq 1$, is "complete" in Wald's sense: if R is any simple rule then there exists a λ such that

$$L(R_\lambda, \boldsymbol{\theta}) \leqq L(R, \boldsymbol{\theta}) \quad \text{for every } \boldsymbol{\theta} \text{ in } \Omega . \tag{49}$$

To show this directly in the present case, take any simple rule R with associated function $u(x)$ and choose that λ for which

$$\int f(x+1)\, u_\lambda(x)\, dx = F\left(-1 + \tfrac{1}{2}\ln\frac{1-\lambda}{\lambda}\right) = \int f(x+1)\, u(x)\, dx ; \tag{50}$$

then necessarily

$$\int f(x-1)\, u_\lambda(x)\, dx = F\left(-1 - \tfrac{1}{2}\ln\frac{1-\lambda}{\lambda}\right) \leqq \int f(x-1)\, u(x)\, dx , \tag{51}$$

since otherwise (45) would not be maximized by $u_\lambda(x)$. Now (49) follows directly from (44), (50), (51). We note that $R_\lambda = \tilde{R}$ for $\lambda = .5$, and that $L(R_{.5}, \boldsymbol{\theta}) \equiv F(-1)$.

It follows from (49) that in the class of simple rules we may confine ourselves to the family R_λ. It remains to choose λ. From (48) we see that for fixed λ, $L(R_\lambda, \boldsymbol{\theta})$ is a linear function of $p(\boldsymbol{\theta})$ with extreme values $F\left(-1 - \tfrac{1}{2}\ln\frac{1-\lambda}{\lambda}\right)$, $F\left(-1 + \tfrac{1}{2}\ln\frac{1-\lambda}{\lambda}\right)$ assumed respectively when $p(\boldsymbol{\theta}) = 0, 1$. It follows that for $\lambda \neq .5$,

$$\begin{aligned} \max_{\boldsymbol{\theta} \in \Omega} L(R_\lambda, \boldsymbol{\theta}) &= \max\left[F\left(-1 - \tfrac{1}{2}\ln\frac{1-\lambda}{\lambda}\right), F\left(-1 + \tfrac{1}{2}\ln\frac{1-\lambda}{\lambda}\right)\right] \\ &> \tfrac{1}{2}\left[F\left(-1 - \tfrac{1}{2}\ln\frac{1-\lambda}{\lambda}\right) + F\left(-1 + \tfrac{1}{2}\ln\frac{1-\lambda}{\lambda}\right)\right] > F(-1) , \end{aligned} \tag{52}$$

which is a stronger inequality than the minimax character of $R_{.5} = \tilde{R}$, previously shown to hold in the class of all decision rules. Since by (52) the linear risk function $L(R_\lambda, \theta)$ for $\lambda \neq .5$ lies above $F(-1)$ for more than half the interval $0 \leqq p(\theta) \leqq 1$, and rises above it at one end by more than it falls below it at the other, it seems reasonable (when nothing is known about θ) to regard $\tilde{R} = R_{.5}$ as the "best" of the rules R_λ and hence of all simple rules. Thus when we are restricted to *simple* rules the minimax rule $\tilde{R}$ *is* the "best," and asymptotically subminimax rules do not exist. On the other hand, if $p(\theta)$ is known then by the previous discussion culminating in (48),

$$(53) \qquad L(R_{p(\theta)}, \theta) = h[p(\theta)] \quad [\text{see } (39)] < F(-1) \text{ for } p(\theta) \neq .5 .$$

Thus if $p(\theta)$ were known we could, by using the simple rule $R_{p(\theta)}$, achieve the risk function $h[p(\theta)]$ which lies below the risk function $F(-1)$ of $\tilde{R}$. {Of course, by using the nonsimple rule $\tilde{R}_k$ with $k = np(\theta)$ [see (35)] we could still further reduce the risk function.} In fact, the curve $y = h(p)$ is the envelope of the one parameter family of straight lines

$$y = y(p, \lambda) = pF\left(-1 + \tfrac{1}{2}\ln\frac{1-\lambda}{\lambda}\right) + (1-p)F\left(-1 - \tfrac{1}{2}\ln\frac{1-\lambda}{\lambda}\right)$$

and lies below each of them, including the line $y = y(p, .5) = F(-1)$.

In practice, of course, $p(\theta)$ will rarely be known. However, and this is the key to the matter, *we can estimate $p(\theta)$ from the sample* (10) *and then use the rule R_λ with λ replaced by our estimate of $p(\theta)$*. Let us see how this attempt to lift ourselves by our own bootstraps works out.

We must first choose some estimator of $p(\theta)$. One's first thought is to use the method of maximum likelihood. As was pointed out in section 3, the maximum likelihood estimate of θ is

$$\hat{\theta} = \hat{\theta}(x) = [\text{sgn}(x_1), \ldots, \text{sgn}(x_n)],$$

so that (presumably)

$$p(\hat{\theta}) = \frac{1}{n}(\text{no. of } i \text{ for which } x_i > 0)$$

is the maximum likelihood estimate of $p(\theta)$. It is easily seen that

$$E[p(\hat{\theta}) \mid \theta] = [1 - 2F(-1)]\, p(\theta) + F(-1),$$

so that for $p(\theta) \neq .5$, $p(\hat{\theta})$ is a biased estimator whose bias does not tend to 0 as $n \to \infty$ unless $p(\theta) \to .5$. We can correct for bias by using the unbiased estimator

$$z = \frac{p(\hat{\theta}) - F(-1)}{1 - 2F(-1)}$$

which has variance

$$\text{var}[z \mid \theta] = \frac{1}{n}\frac{F(-1)[1 - F(-1)]}{[1 - 2F(-1)]^2} \simeq \frac{.29}{n}$$

and is nearly normal for large n.

Consider instead, however, the unbiased estimator

$$v = \tfrac{1}{2}(1 + \bar{x}), \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

which is normal with variance

$$\text{var}\,[\,v\,|\,\theta\,] \doteq \frac{.25}{n}$$

less than that of z. Since $0 \leqq p(\theta) \leqq 1$ it seems natural to truncate v at 0 and 1 and to use the modified estimator

$$v' = \begin{cases} 0 & \text{if } v \leqq 0\,, \\ v & \text{if } 0 < v < 1\,, \\ 1 & \text{if } v \geqq 1\,, \end{cases}$$

which, though biased, seems "better" than v. We shall adopt, arbitrarily, v' as our estimator for $p(\theta)$. v' is clearly consistent as $n \to \infty$; in fact

$$\lim_{n\to\infty} E\{\,[\,v' - p\,(\theta)\,]^2\,|\,\theta\,\} = 0$$

uniformly for all θ in Ω. Now the rule R_λ with λ replaced by the estimate v' of $p(\theta)$ is simply R^* defined by (37), and this provides the motivation for considering R^*.

Since v' is a consistent estimator of $p(\theta)$ it is clear from (53) that (42) must hold. The performance of R^* for finite n must, of course, be worked out by computation on the basis of (41) and (38), as was done in the table of section 4.

There are, of course, other ways in which we might obtain decision rules to compete with $\tilde{R}$. For example, we might use an integer-valued estimate of $k = np(\theta)$ in the rule $\tilde{R}_k$. Again, we might use an iterative process: first estimating $p(\theta)$ then using R_λ with λ replaced by its estimate to obtain the decision "$\theta = \theta^{(k)}$" for some k, and finally using R_λ again with λ replaced by $p(\theta^{(k)})$. There is also the rule $\bar{R}$ derived in section 3 as the solution of a minimum problem. Compared with any of these, R^* has at least the advantage of simplicity.

Because of the demonstrated properties of R^* it seems safe to say that for "large" n there exist, among the nonsimple rules, worthy competitors of the minimax rule $\tilde{R}$. If this be admitted one then has the problem of finding the "best" decision rule. The definition of "best" is an open question at the moment, but at least it appears that "best" does not equal "minimax."

The existence of asymptotically subminimax decision functions is not confined to problems of the "compound" type. For example, let x have a binomial distribution (n, θ) and let it be required to estimate θ by some function $t = t(x)$ so as to minimize the quantity

$$L\,(t,\ \theta)\ = nE\,[\,(t-\theta)^2\,|\,\theta\,] = n\sum_{x=0}^{n}\binom{n}{x}\theta^x\,(1-\theta)^{n-x}\,[\,t\,(x)\ -\theta\,]^2\,.$$

For the conventional estimator $t_1 = x/n$ we have

$$L\,(t_1,\ \theta)\ = \theta\,(1-\theta)\,, \tag{54}$$

while the minimax estimator is

$$t_2 = \frac{x + \dfrac{\sqrt{n}}{2}}{n + \sqrt{n}}\,,$$

for which

$$L(t_2, \theta) = \frac{1}{4\left(1+\frac{1}{\sqrt{n}}\right)^2}. \tag{55}$$

As $n \to \infty$, (55) $\to \frac{1}{4}$ which is greater than (54) except for $\theta = .5$. Thus t_1 is asymptotically subminimax, although in this case it is the minimax risk function (55) which varies with n. The question of whether t_1 is "better" than t_2 has been raised by Hodges and Lehmann [2].

6. General remarks on compound decision problems

A wide class of statistical decision problems can be brought under a general scheme, due to Wald, in which there is given (1) a sample space X of points x, (2) a parameter space ω of points θ such that for every θ in ω there corresponds a probability distribution P_θ on X, (3) a class $\mathfrak{D}$ of decisions D, (4) a loss function $w(D, \theta) \geqq 0$ representing the cost of taking the decision D when the true value of the parameter is θ. Any function $u = u(x)$ with values in $\mathfrak{D}$ is called a decision function, and the function

$$L(u, \theta) = \int w[u(x), \theta]\, dP_\theta(x) \tag{56}$$

is called the risk function. The statistical decision problem ρ is to find the decision function u which in some sense minimizes the risk function (56) over ω. For example, we may seek the u which minimizes the quantity

$$\int_\omega L(u, \theta)\, dG(\theta), \tag{57}$$

where $G(\theta)$ is a given distribution on ω, and which is called the *Bayes solution* of ρ corresponding to $G(\theta)$. Or we may require that u be a *minimax* solution for which the quantity

$$\max_{\theta \in \omega} L(u, \theta)$$

is a minimum.

It often happens that one deals with a set of n independent and, in general, unrelated, decision problems of the same mathematical form. Thus, let $x_1, \ldots, x_n$ be independent random variables such that each x_i presents the same problem ρ. Each x_i will be distributed in X with a distribution P_{θ_i}, $\theta_i \in \omega$, but no relation is assumed to hold among the n parameter values $\theta_1, \ldots, \theta_n$. For each x_i a decision $D_i \in \mathfrak{D}$ must be taken. We shall take the quantity

$$\frac{1}{n}\sum_{i=1}^{n} w(D_i, \theta_i)$$

as a measure of the loss incurred by any set of decisions $D_1, \ldots, D_n$ when the true parameter values are respectively $\theta_1, \ldots, \theta_n$. If a decision function u_i depending on x_i alone is used for the i-th decision then, setting $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and $\boldsymbol{u} = (u_1, \ldots, u_n)$, the risk function will be

$$\begin{aligned} L(\boldsymbol{u}, \boldsymbol{\theta}) &= \int \cdots \int \frac{1}{n}\sum_{i=1}^{n} w[u_i(x_i), \theta_i]\, dP_{\theta_1}(x_1) \cdots dP_{\theta_n}(x_n) \\ &= \frac{1}{n}\sum_{i=1}^{n} \int w[u_i(x), \theta_i]\, dP_{\theta_i}(x). \end{aligned} \tag{58}$$

The problem of minimizing (58) by proper choice of the $u_i(x)$ is essentially the same as that of minimizing (56). However, if the whole set of values $x_1, \ldots, x_n$ is known before the individual decisions are to be made, then we can permit u_i to depend on *all* the values $x_1, \ldots, x_n$, so that the risk function will be

$$(59)\quad L(\boldsymbol{u}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \int \cdots \int w[u_i(x_1, \ldots, x_n), \theta_i]\, dP_{\theta_1}(x_1) \cdots dP_{\theta_n}(x_n).$$

The problem of minimizing (59) over the n-fold Cartesian product Ω of ω with itself, consisting of all ordered n-tuples $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, $\theta_i \in \omega$, is quite different from the original problem $\boldsymbol{\rho}$ involving (56). We shall denote the problem of minimizing (59) by $\boldsymbol{\rho}^{(n)}$ and call it the *compound decision problem* corresponding to the simple problem $\boldsymbol{\rho}$.

At first sight it may seem that the use of decision functions of the general form $u_i(x_1, \ldots, x_n)$ is pointless, since the values x_j for $j \neq i$ can contribute no information concerning θ_i; this because the distribution P_{θ_i} of x_j depends only on θ_j which was not assumed to be in any way related to θ_i. From this point of view we should stick to *simple* decision functions of the form $u_i(x_1, \ldots, x_n) = u(x_i)$ where $u(x)$ is the "best" solution of $\boldsymbol{\rho}$. The example of section 4, however, shows that there may be great advantages in using nonsimple decision functions of the general form $u_i(x_1, \ldots, x_n)$. In that example the minimax solution of $\boldsymbol{\rho}^{(n)}$ is afforded by the simple decision functions $\tilde{u}_i(x_1, \ldots, x_n) = \tilde{u}(x_i)$, where $\tilde{u}(x)$ is the minimax solution of $\boldsymbol{\rho}$, but although $\tilde{u}$ was seen to be the "best" solution of $\boldsymbol{\rho}$, the existence as $n \to \infty$ of an asymptotically subminimax solution of $\boldsymbol{\rho}^{(n)}$ showed that the minimax solution of $\boldsymbol{\rho}^{(n)}$ was not the "best" for large n. This phenomenon is to be expected in many cases, as we shall see.

The most interesting Bayes solutions of $\boldsymbol{\rho}^{(n)}$ are those obtained by minimizing the integral of (59) over Ω with respect to some distribution $G(\boldsymbol{\theta})$ which is invariant under all permutations of the components $\theta_1, \ldots, \theta_n$ of $\boldsymbol{\theta}$; the corresponding Bayes solutions $\boldsymbol{u}$ will then be *symmetric* in the sense that the risk function (59) will be invariant under all permutations of $\theta_1, \ldots, \theta_n$. In general the Bayes solutions may be expected to be complicated in structure and difficult to evaluate in performance.

We shall now give a heuristic rule for constructing certain nonsimple solutions of $\boldsymbol{\rho}^{(n)}$. For any $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ in Ω let $G_{\boldsymbol{\theta}}(\theta)$ be the cumulative distribution function of the probability distribution of a random variable θ for which $P[\theta = \theta_i] = 1/n$, $i = 1, \ldots, n$; that is, if ω is the real line,

$$(60)\quad G_{\boldsymbol{\theta}}(\theta) = \frac{1}{n}(\text{no. of } i \text{ for which } \theta_i \leqq \theta), \quad -\infty < \theta < \infty.$$

Suppose that the distribution P_θ on X has a density function $f(x, \theta)$; then for any simple rule $\boldsymbol{u}$ such that $u_i(x_1, \ldots, x_n) = u(x_i)$, (58) becomes

$$(61)\quad L(\boldsymbol{u}, \boldsymbol{\theta}) = \int \left[\frac{1}{n}\sum_{i=1}^{n} w[u(x), \theta_i]\, f(x, \theta_i)\right] dx$$

$$= \int \left[\int w[u(x), \theta]\, f(x, \theta)\, dG_{\boldsymbol{\theta}}(\theta)\right] dx.$$

Now if $G_{\boldsymbol{\theta}}(\theta)$ is known (that is, if the components θ_i of $\boldsymbol{\theta}$ are known apart from their order) then (61) will be a minimum for that function $u(x) = u[x; G_{\boldsymbol{\theta}}(\theta)]$ such that for every fixed x, $u(x) = t$, where t is that number for which

$$\int w(t, \theta) f(x, \theta) dG_{\boldsymbol{\theta}}(\theta) = \min. \tag{62}$$

Denote the simple decision rule for which $u_i(x_1, \ldots, x_n) = u[x_i; G_{\boldsymbol{\theta}}(\theta)]$ by $u[G_{\boldsymbol{\theta}}(\theta)]$; this will clearly be better than any simple rule which depends on $x_1, \ldots, x_n$ alone. Of course, in order to use the rule $u[G_{\boldsymbol{\theta}}(\theta)]$ we must know $G_{\boldsymbol{\theta}}(\theta)$, which depends on the unknown $\boldsymbol{\theta}$. *Thus we must devise a method for estimating* $G_{\boldsymbol{\theta}}(\theta)$ *from the observed values* $x_1, \ldots, x_n$. [Actually, we need only be able to estimate the left hand side of (62) for every x and t.] This involves finding a solution to the following problem:

(I). Let $x_1, \ldots, x_n$ be independent random variables such that the density function of x_i is $f(x, \theta_i)$, where $\theta_1, \ldots, \theta_n$ are n arbitrary unknown elements of a parameter set ω. The joint density function of the x_i is therefore $\prod_{i=1}^{n} f(x_i, \theta_i)$. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. From the observed values $x_1, \ldots, x_n$ we are to form an estimator $G(\theta; x_1, \ldots, x_n)$ of the cumulative distribution function (60) which for large n shall be "close" to (60) with probability near 1 for all possible values of $\boldsymbol{\theta}$.

Assuming problem (I) to be solved we can apply in the compound decision problem $\boldsymbol{\rho}^{(n)}$ the nonsimple decision rule

$$u^*(x_1, \ldots, x_n) = \{u[x_1, G(\theta; x_1, \ldots, x_n)], \ldots, u[x_n, G(\theta; x_1, \ldots, x_n)]\}; \tag{63}$$

that is, $u[G_{\boldsymbol{\theta}}(\theta)]$ with $G_{\boldsymbol{\theta}}(\theta)$ replaced by its estimate $G(\theta; x_1, \ldots, x_n)$. If our solution of problem (I) is a good one then for large n (63) will be better than any simple rule. In particular, if the minimax solution (assumed unique) of $\boldsymbol{\rho}$ is denoted by $u(x)$, then (63) will be better than the simple rule

$$u(x_1, \ldots, x_n) = [u(x_1), \ldots, u(x_n)]. \tag{64}$$

If (64) is the minimax solution of $\boldsymbol{\rho}^{(n)}$ in the class of all decision rules, nonsimple included, then (63) will be *asymptotically subminimax.*

We have seen in section 5 that problem (I) can be solved in the very simple case in which ω consists of only two elements, ± 1, and $f(x, \theta)$ is the normal density function with mean θ and variance 1. The function $G_{\boldsymbol{\theta}}(\theta)$ is then completely determined by the number $p(\boldsymbol{\theta}) =$ (no. of i for which $\theta_i = 1)/n$, of which a consistent estimator is $(1 + \bar{x})/2$, where $\bar{x} = \sum_{i=1}^{n} x_i/n$.

Before proceeding further with problem (I) let us consider a different but analogous problem:

(II). Let $x_1, \ldots, x_n$ be independent random variables each with a common density function

$$h_G(x) = \int f(x, \theta) dG(\theta)$$

where $f(x, \theta)$ is the same as in problem (I) and $G(\theta)$ is an unknown distribution on ω. The joint density function of the x_i is therefore $\prod_{i=1}^{n} h_G(x_i)$. From the observed values $x_1, \ldots, x_n$ we are to form an estimator $G(\theta; x_1, \ldots, x_n)$ of the unknown $G(\theta)$ which for large n will be "close" to $G(\theta)$ with probability near 1 for all $G(\theta)$ in some class $\mathcal{G}$.

Problem (II) is a generalization of a classical problem in the theory of estimation. Let $\mathcal{G}$ be the class of distributions concentrated at some *single point* of ω; then the joint density function of the x_i is simply $\prod_{i=1}^{n} f(x_i, \theta)$, with θ unknown, and we require a consistent estimator of θ. Under certain conditions on $f(x, \theta)$ and ω, the *method of maximum likelihood* provides a solution:

$$(65) \qquad \hat{\theta}(x_1, \ldots, x_n) = \text{that } \theta \text{ in } \omega \text{ for which } \prod_{i=1}^{n} f(x_i, \theta) = \max .$$

More generally, it has been announced in an abstract [3] that under certain conditions the "generalized method of maximum likelihood" provides a solution of problem (II):

$$(66) \quad \hat{G}(\theta; x_1, \ldots, x_n) = \text{that } G(\theta) \text{ in } \mathcal{G} \text{ for which } \prod_{i=1}^{n} h_G(x_i) = \max .$$

Problem (II) is itself of interest in statistical decision problems in which there is a *prior distribution of parameters*. Returning to the problem ρ stated at the beginning of this section, if θ is itself a random variable with known distribution $G(\theta)$ on ω then the best solution of ρ is that u which minimizes the integral (57). However, if $G(\theta)$ exists but the statistician knows only that it belongs to some class $\mathcal{G}$, then in the problem $\rho^{(n)}$ he can estimate $G(\theta)$ by solving problem (II) and then determine $u(x)$ by minimizing (57) with $G(\theta)$ replaced by the estimate $G(\theta; x_1, \ldots, x_n)$ [3]. However, even the assumption of an existing but unknown prior distribution $G(\theta)$ will be questionable in most applications of statistics, and we merely mention the matter here.

We have stated that under certain conditions problem (II) can be solved by the generalized method of maximum likelihood. Problem (I) is more difficult, and it is easily seen that a solution of problem (I) would in general provide a solution of problem (II). Conversely, however, as a heuristic principle *we can in some cases solve problem* (I) *by acting "as though" it were problem* (II); in fact, any solution $G(\theta; x_1, \ldots, x_n)$ of problem (II) which is a symmetric function of $x_1, \ldots, x_n$ [as, for example, (66)] will at the same time provide a possible solution of problem (I). In justification of this principle we point out that if $\theta_1, \ldots, \theta_n$ form a random sample from a distribution $G(\theta)$, then for large n the empirical cumulative distribution function of $\theta_1, \ldots, \theta_n$ will tend uniformly to $G(\theta)$ with probability 1 as $n \to \infty$. Hence the random variables $x_1, \ldots, x_n$ of problem (II) will act much like those of problem (I), insofar as symmetric functions of the x_i are concerned. Questions of uniformity, of course, have to be considered before any precise theorem can be stated, and the whole subject of problems (I) and (II) requires and

will repay a careful treatment. In particular, the generalized method of maximum likelihood, even if in theory it provides a solution to these problems, will in practice be extremely difficult to apply.

REFERENCES

[1] A. WALD, "Statistical decision functions," *Annals of Math. Stat.*, Vol. 20 (1949), pp. 165–205.
[2] J. L. HODGES, JR. and E. L. LEHMANN, "Some problems in minimax point estimation," *Annals of Math. Stat.*, Vol. 21 (1950), pp. 182–197.
[3] H. ROBBINS, "A generalization of the method of maximum likelihood: estimating a mixing distribution," abstract, *Annals of Math. Stat.*, Vol. 21 (1950), pp. 314–315.