SOME PROBLEMS IN MINIMAX POINT ESTIMATION

By J. L. Hodges, Jr.,[1] and E. L. Lehmann

University of California, Berkeley

1. Summary. In the present paper the problem of point estimation is considered in terms of risk functions, without the customary restriction to unbiased estimates. It is shown that, whenever the loss is a convex function of the estimate, it suffices from the risk viewpoint to consider only nonrandomized estimates. For a number of specific problems the minimax estimates are found explicitly, using the squared error as loss. Certain minimax prediction problems are also solved.

2. Introduction. The principles most commonly applied in the selection of a point estimate are the principles of maximum likelihood (R. A. Fisher) and of minimum variance unbiased estimation (Markoff).[2] Both of these principles are intuitively appealing, but neither of them can be justified very well in a systematic development of statistics. This holds also for some modifications of these principles proposed by G. W. Brown [1], as the author himself points out.

In an important early paper [2], Wald indicated a more systematic approach to the problem, which he later developed into his general theory of statistical decision problems [3, 4, 5]. Consider a random variable X distributed over a space $\mathfrak{X}$ according to a distribution P_θ^X with $\theta \in \Omega$. It is desired to estimate some $g(\theta)$. If the value x of X is observed one makes an estimate, say $f(x)$, and thereby incurs a loss of $W[g(\theta), f(x)]$ when θ is the true value of the parameter. We shall assume that the loss function is nonnegative. It then follows that the expectation of the loss will always exist (although it may be infinite). The risk associated with the estimate f is defined to be the expected loss, as given by

$$R_f(\theta) = E_\theta W[g(\theta), f(x)] = \int_{\mathfrak{X}} W[g(\theta), f(x)]\, dP_\theta^X(x). \tag{2.1}$$

The choice of estimate should then be made according to the risk function. As a particular possibility Wald suggests the use of minimax estimates, i.e. estimates which minimize $\sup_\theta R_f(\theta)$.

The main purpose of the present paper is to obtain minimax estimates for a number of specific problems. Only few such problems have been worked out so far, the emphasis in Wald's work having been on the general theory. In [2] Wald obtained the minimax estimate of an unknown location parameter. Stein and Wald [6] treated the sequential problem of estimating the mean of a normal dis-

[1] This work was supported in part by the Office of Naval Research.

[2] Actually, the principle of minimum variance unbiased estimation goes back to Gauss. For discussions of the history of these ideas, see E. Czuber, *Theorie der Beobachtungsfehler*, Leipzig, 1891, and R. L. Plackett, "A historical note on the method of least squares", *Biometrika*, Vol. 36 (1950), p. 458.

tribution with known variance, and in his forthcoming book Wald considers as an example the sequential problem of estimating the mean of a random variable distributed uniformly over an interval of length 1.

It seems worthwhile to consider further special problems both because one may obtain estimates that in some cases are preferable to the conventional ones, and because these examples throw some light on the general desirability of the minimax principle. As we shall see below, it does not seem possible to reach any definite conclusions on this latter point, and to obtain a generally valid comparison between the minimax estimate and, for example, the unbiased estimate with uniformly smallest variance (when such an estimate exists).

Consider, for example, the problem of estimating the probability of success from a number of independent trials each of which may be a success or a failure, when the loss-function is the squared error. If the number of trials is one, the minimax estimate (as is shown below) is given by $f(X) = \frac{1}{2}X + \frac{1}{4}$, where X is 1 or 0 as the trial is a success or failure. As is easily seen, this estimate has smaller risk than the usual estimate $f^*(X) = X$ whenever $0.07 \leqq p \leqq 0.93$. On the other hand, when the number of trials is large the standard estimate $\bar{X}$ has smaller risk than the minimax estimate nearly everywhere. The minimax estimate is only slightly better in a small interval centered at $p = \frac{1}{2}$, whose length tends to zero as the number of trials tends to infinity, and is worse everywhere else.

For our purpose it is convenient to formulate the problem of point estimation as follows (see in this connection [7]). A random variable X is distributed over a space $\mathfrak{X}$ according to a distribution P belonging to a family $\mathfrak{F}$. We wish to estimate $g(P)$ where g is a function whose domain is $\mathfrak{F}$ and whose range is contained in some space $\mathcal{Y}$ (in any example $\mathcal{Y}$ is usually a Euclidean space, mostly even a one dimensional Euclidean space). An estimate is a statistic $f(X)$ taking on values in $\mathcal{Y}$. We denote by $W[g(P), f(x)]$ the loss which results from making the estimate $f(x)$ when P is the true distribution, and we define the risk function of the estimate f by

$$R_f(P) = E_P W[g(P), f(X)]. \tag{2.2}$$

The problem is to determine f so as to minimize $\sup_{P \epsilon \mathfrak{F}} R_f(P)$.

Our principal tool will be the following theorem, which is essentially contained in Wald's work but which is not stated there explicitly. The theorem is a slight modification of one used for the theory of testing in [8].

THEOREM 2.1. *Let* $\{P_\theta\}$, $\theta \epsilon \omega$ *(where* ω *is a subset of a Euclidean space), be a parametric subfamily of* $\mathfrak{F}$, *and let* λ *be a probability measure over* ω. *Suppose that* f *minimizes*

$$\int_\omega E_\theta W[g(P_\theta), f(X)]\, d\lambda(\theta) \tag{2.3}$$

and that

(i) $E_\theta W[g(P_\theta), f(X)]$ *is constant (say* c*) for all* $\theta \epsilon \omega$,

(ii) $E_P W[g(P), f(X)] \leq c$ *for all* P *in* $\mathfrak{F}$.

Then f *is a minimax estimate for estimating* g.

PROOF. Let f^* be any other estimate of g. Then

$$\begin{aligned}\sup_{P\in\mathfrak{F}} E_P W[g(P), f(X)] &= \int_\omega E_\theta W[g(P_\theta), f(X)]\, d\lambda(\theta) \\ &\leqq \int_\omega E_\theta W[g(P_\theta), f^*(X)]\, d\lambda(\theta) \qquad (2.4) \\ &\leqq \sup_{P\in\mathfrak{F}} E_P W[g(P), f^*(X)].\end{aligned}$$

We note that if f is the unique function minimizing (2.3), then the first inequality in (2.4) becomes strict, and hence f is the unique minimax estimate of **g**.

Following Wald we shall call the function f that minimizes (2.3) the Bayes estimate of g associated with the a priori distribution λ. As a corollary to theorem 2.1, we note that a Bayes estimate whose risk function is constant, is a minimax estimate.

3. Randomization. In the formulation of the problem of point estimation given above, the estimate $f(x)$ is assumed to be completely determined by the observed value x of the random variable X. In the present section a broader formulation of the problem will be considered, in which the estimate corresponding to x may itself be a random variable, say T_x. This extension is a special case of the notion of randomized decision function introduced by Wald in his general decision theory. We associate with each x in $\mathfrak{X}$ a probability distribution F_x, with the convention that when X is observed to have the value x, we estimate $g(P)$ by means of a random variable T_x which is distributed according to F_x. Estimates of this latter kind we shall call *randomized*, and the fixed estimates $f(x)$ *nonrandomized.*

The motivation behind the admission of randomized estimates (or more generally of randomized statistical decision funtions) is that in some problems of statistical inference the performance of the decision function is considerably improved by randomization. It is clear however that the randomized functions are more complicated, and hence that it is useful to know when their consideration is not necessary. Before investigating this question we give the following definition, which makes precise a sense in which certain estimates may be omitted from consideration. (See Wald [9]).

DEFINITION. For a given estimation problem a class C of estimates will be said to be essentially complete with respect to a class D of estimates, if for every estimate g in D there exists an estimate f in C such that $R_f(P) \leqq R_g(P)$ for all P in $\mathfrak{F}$. If D is the class of all randomized estimates we simply say that C is essentially complete for the given problem.

It is clear that if one adopts the risk function point of view, one loses nothing by restricting consideration to an essentially complete class of estimates. In the present section we find conditions under which the totality of nonrandomized estimates forms an essentially complete class.

For this purpose we need the notion of convexity. A set S in a k-dimensional Euclidean space is said to be convex if, whenever P and Q are in S, then all points on the line segment from P to Q are also in S. A real valued function ψ defined over a k-dimensional Euclidean space is said to be convex, if for any points $(x_1, \cdots, x_k)$ and $(y_1, \cdots, y_k)$ of the space, and any number $0 < \alpha < 1$ we have

$$(3.1)\quad \begin{aligned} &\alpha\psi(x_1, \cdots, x_k) + (1 - \alpha)\psi(y_1, \cdots, y_k) \geqq \\ &\qquad\qquad \psi(\alpha x_1 + (1 - \alpha)y_1, \cdots, \alpha x_k + (1 - \alpha)y_k). \end{aligned}$$

We use the following notation for conditional expectation. If U and V are two random variables which have a joint distribution, then $E(U|v)$ denotes the conditional expectation of U given that $V = v$; $E(U|S)$ denotes the conditional expectation of U given that V is in S. Let $\Phi(v) = E(U|v)$; then for $\Phi(V)$ we write $E(U \mid V)$.

LEMMA 3.1. *Let U, V be two random variables with a joint distribution, such that U is distributed in a k-dimensional space and $E(U)$ is finite. Let ψ be a real-valued convex function defined over this space and bounded from below. Then*

$$E\{\psi\{E(U \mid V)\}\} \leqq E\{\psi(U)\}.$$

PROOF. The proof is immediate in the special case that, for almost all v, there exists a determination of the conditional probability distribution of U given v which is a measure. We then know, from the convexity of ψ, that for almost all values v of V, $\psi\{E(U \mid v)\} \leqq E\{\psi(U) \mid v\}$. Replacing v by V and taking expectations of both sides, we obtain the desired result.

If we do not assume the existence of conditional measures, the proof is more complicated. Since $E(U)$ is finite, there exists a function $E(U \mid v)$ such that for any set S, $E(U \mid S) = E\{E(U \mid V) \mid S\}$; see [10], p. 47. Since ψ is convex it is measurable, and since ψ is bounded from below $E\{\psi(U)\}$ exists. Excluding the trivial case $E\{\psi(U)\} = \infty$, we know there exists a function $E\{\psi(U) \mid v\}$ such that for any set S, $E\{\psi(U) \mid S\} = E\{E\{\psi(U) \mid V\} \mid S\}$.

If the lemma were false, we should have $E\{E\{\psi(U) \mid V\}\} < E\{\psi\{E(U \mid V)\}\}$, and could find an $\epsilon > 0$ and a set A of positive V measure such that for every $v \epsilon A$, $E\{\psi(U) \mid v\} + 2\epsilon < \psi\{E(U \mid v)\}$. This implies the existence of a number d and a set B of positive V measure such that for every $v \epsilon B$, $E\{\psi(U) \mid v\} \leqq d$ and $d + \epsilon \leqq \psi\{E(U \mid v)\}$. Since ψ is convex, the domain D of points P for which $\psi(P) < d + \epsilon$ is convex, and we may find a subset C of B, of positive V measure, for which the set of points $E(U \mid v)$, $v \epsilon C$, lies in a convex domain E disjoint of D. It follows that $E(U \mid C)$ lies in E, and hence that $\psi\{E(U \mid C)\} \geqq d + \epsilon$. Clearly $d \geqq E\{\psi(U) \mid C\}$. Thus we have the contradiction $E\{\psi(U) \mid C\} \geq \psi\{E(U \mid C)\}$.

DEFINITION. A loss function W will be called convex if for every $u \epsilon \mathcal{Y}$, $W(u, v)$ is a convex function of the estimate v.

An example of a convex loss function is provided by the Markoff principle of estimation. The variance of an unbiased estimate may be considered as a risk function if we take the loss function to be the squared error, i.e. the square of the difference between the true value $g(P)$ and the estimated value $f(x)$ or T_x; and this loss function is clearly convex.

THEOREM 3.2. *If the loss function W is convex, if $\mathcal{Y}$ is in a Euclidean space, and if we consider only estimates having finite expectation, then the class of nonrandomized estimates is essentially complete.*

PROOF. Let T_X be any randomized estimate such that $E(T_X)$ exists and is finite. Applying lemma 3.1 we see that $E(T_X \mid X)$, which as a function of X only is a nonrandomized estimate, has a risk never greater than that of T_X.

The restriction in theorem 3.2 to estimates having finite expectation may be replaced by the requirement that for each $u \in \mathcal{Y}$ there exist a number M_u such that if $| v - u | = M_u$ then $W(u, v) > W(u, u)$. With this requirement and the convexity assumption, it follows that the risk associated with T_X is infinite whenever $E(T_X)$ is infinite.

Theorem 3.2 is related to a generalization of a theorem of Blackwell. If Y is a sufficient statistic for $g(P)$, and if for almost all y the conditional distribution of X given y exists in the sense of measure, we may regard estimation of $g(P)$ based on X as randomized estimation of $g(P)$ based on Y; and if the assumptions of theorem 3.1 are satisfied, we may apply this theorem to conclude the essential completeness of the class of nonrandomized estimates based on Y. In the general case we may resort again to lemma 3.1 to prove the following theorem; the proof is the same as that of theorem 3.2 if X is replaced by Y throughout.

THEOREM 3.3. *If the loss function W is convex, if $\mathcal{Y}$ is in a Euclidean space, if we consider only estimates having a finite expectation, and if Y is a sufficient statistic for $\mathcal{F}$, then the class of nonrandomized estimates which are functions of Y only is essentially complete.*

Blackwell [11] proved that if U is a sufficient statistic for a real-valued parameter θ, and if T is an unbiased estimate for θ, then $E(T \mid U)$, which is a function of U only and also an unbiased estimate for θ, has a variance which never exceeds that of T. Observing that the theorems above hold true when we restrict attention to unbiased estimates, Blackwell's result may be obtained from theorem 3.3 by letting $\mathcal{Y}$ be one-dimensional, letting W be the squared error, and restricting ourselves to unbiased estimates. In a similar manner we can get from theorem 3.3 an extension of Blackwell's theorem given by Barankin [12], who treated the case in which $W(\theta, t) = | \theta - t |^s$, $s \geq 1$. It is clear that these loss functions are convex.

If the convexity assumption is removed, theorems 3.2 and 3.3 cease to be true. For example, if $\mathcal{X}$ has only n points, if $\mathcal{Y}$ is a finite line segment of length greater than $2n\alpha$, and if the loss is 0 whenever $| g(P) - f(x) | \leq \alpha$, and 1 otherwise, then the minimax risk among nonrandomized estimates is 1. By admitting randomization, however, the maximum risk can be brought below 1 without using X at all; if our estimate T is uniformly distributed over $\mathcal{Y}$, then the maximum risk will be $1 - \alpha/(\text{length of } \mathcal{Y})$.

The example just given may seem inappropriate, in that with the specified loss function the problem would customarily be considered one of interval estimation rather than point estimation. This objection does not apply however to the loss functions considered in the following theorem.

THEOREM 3.4. *Let* $\mathfrak{X} = \{0, 1, \cdots, n\}$, $n \geq 1$. *Let* $\mathcal{F}$ *be the set of binomial distributions* P_p *defined by* $P_p(X = x) = \binom{n}{x} p^x(1 - p)^{n-x}$, $0 \leq p \leq 1$, *Let* $\mathcal{Y}$ *be the closed interval* $[0, 1]$ *and* $g(P_p) = p$. *Let* $W(p, t) = | p - t |^s$, $0 < s < 1$. *Then no minimax estimate can be nonrandomized, and the class of nonrandomized estimates is not essentially complete.*

PROOF. For any nonrandomized estimate f, $R_f(p)$, being a sum of products of continuous functions of p, is itself a continuous function of p. The nonrandomized minimax risk is less than 1, as may be shown by considering any estimate of the following kind: $f(0) = 0$, $f(n) = 1$, and $0 \leq f(x) \leq 1$ for all x. Here $R_f(0) = R_f(1) = 0$, while if $0 < p < 1$, $R_f(p) \leq \max_x | p - f(x) |^s < 1$. By continuity $\sup_{0 \leq p \leq 1} R_f(p) < 1$.

It is easy to see that there exists among the nonrandomized estimates a minimax estimate, say h. Let the corresponding minimax risk be denoted by M. We know that $M = \sup_{0 \leq p \leq 1} R_h(p) < 1$; it is obvious that $M > 0$. Observe that $h(0) < 1$, since $h(0) \geq 1$ leads to the contradiction $R_h(0) = | h(0) |^s \geq 1$. We can write

$$R_h(p) = \sum_{h(x)=h(0)} P_p(X = x) \cdot | p - h(x) |^s + \sum_{h(x)\neq h(0)} P_p(X = x) \cdot | p - h(x)|^s.$$

The second sum has a finite derivative with respect to p at $p = h(0)$, while the first sum increases with infinite speed as p is moved away from $h(0)$. Therefore $R_h\{h(0)\} < M$; and by an exactly symmetrical argument, $0 < h(n)$ and $R_h\{h(n)\} < M$. Using the continuity of R_h, we can find a positive number ω so small that $R_h(p) < M$ whenever $| p - h(0) | < \omega$ or $| p - h(n) | < \omega$.

Consider now the randomized estimate T_X defined by $T_x = h(x)$ if $0 < x < n$, and by $T_x = h(x) + \alpha Y$ otherwise, where Y is a random variable independent of X and taking on the values 1 and -1 each with probability $\frac{1}{2}$, and where $0 < \alpha < \omega$. Observe

$$R_{T_X}(p) - R_h(p) = (1 - p)^n[\tfrac{1}{2}\{ | p - h(0) + \alpha |^s + | p - h(0) - \alpha |^s\} - | p - h(0) |^s] + p^n[\tfrac{1}{2}\{| p - h(n) + \alpha |^s + | p - h(n) - \alpha |^s\} - | p - h(n) |^s].$$

By the concavity of the functions involved, the first square bracketted term is negative whenever $| p - h(0) | \geq \alpha$, and the second is negative whenever $| p - h(n) | \geq \alpha$. We can choose α so small that whenever either $| p - h(0) |$ or $| p - h(n) |$ is less than α, $R_{T_X}(p) - R_h(p) < \omega$. A continuity argument now shows that $\sup_{0 \leq p \leq 1} R_{T_X}(p) < M$. But this proves that no minimax estimate, with randomization permitted, can be nonrandomized. It is also now obvious that the class of nonrandomized estimates is not essentially complete: every nonrandomized estimate must have a risk function which somewhere exceeds $\sup_{0 \leq p \leq 1} R_{T_X}(p)$.

4. General properties of minimax estimation. Whether a principle such as the minimax principle is a desirable one has to be decided mainly on two criteria:

(i) its general properties, and

(ii) its performance in many particular instances.

It has already been remarked that in the second respect the minimax principle does not seem entirely satisfactory. With regard to the former, one great advantage of this principle is that when there is a unique minimax estimate, it is admissible. Here an estimate f is said to be admissible (see [3]) if there exists no other estimate f^* such that $R_{f^*}(P) \leq R_f(P)$ for all P in $\mathcal{F}$ with strict inequality holding for some P. It is interesting that, as we shall show below, this admissibility property is not shared by either the principle of unbiasedness or the maximum likelihood principle.

In this connection we begin by proving another theorem concerning essentially complete classes.

THEOREM 4.1. *Suppose that the space $\mathcal{Y}$ is a finite interval $[a, b]$ on the real line, and that for each $u \in \mathcal{Y}$, $W(u, v)$ is a non-decreasing function of v when $v > u$ and a non-increasing function of v when $v < u$. Then the class of estimates whose range is contained in $\mathcal{Y}$ is essentially complete with respect to the class of all real valued estimates.*

PROOF. If T is any real-valued estimate, define T^* by

$$T^* = \begin{cases} T & \text{if} \quad T \in \mathcal{Y}, \\ a & \text{if} \quad T < a, \\ b & \text{if} \quad t > b. \end{cases} \tag{4.1}$$

It is clear that $R_{T^*}(P) \leq R_T(P)$ for every $P \in \mathcal{F}$.

Halmos [7] has provided an example in which the minimum variance unbiased estimate takes on, with positive probability, values outside the range of the parameter. It can be shown from the proof of theorem 4.1 that in this case any unbiased estimate is inadmissible, provided the loss function is of the kind described in theorem 4.1.

That the maximum likelihood principle may also lead to inadmissible estimates is easy to show, since this is the case in many familiar situations. The following example may be of interest in that here the maximum likelihood estimate is uniformly worst among all estimates which one would consider using.

Example. Let X be a random variable with only 0 and 1 as possible values, and let $P(X = 1) = p$. Assume it to be known that $\frac{1}{3} \leq p \leq \frac{2}{3}$. Then the maximum likelihood estimate for p is easily seen to be $\frac{1}{3}(X + 1)$, and, if the loss function is the squared error, the associated risk function is $\frac{1}{3}(p - \frac{1}{2})^2 + \frac{1}{36}$. This risk function is, for every possible value of p, greater than that of any estimate $f(x)$ satisfying: $\frac{1}{3} \leq f(0) \leq f(1) = 1 - f(0) \leqq \frac{2}{3}$.

The selection of loss function in any problem should in theory be governed by metastatistical considerations, but in fact the circumstances of statistical problems do not usually offer compelling reasons for using one loss function rather than another. Considerations of mathematical facility are often determining. Thus, various classical unbiased estimates become minimax estimates when the loss function is judiciously chosen. For, if we take as loss function the ratio of squared error to the variance of the unbiased estimate, the risk becomes constant, and we can easily obtain the classical estimates as minimax estimates in the familiar binomial, Poisson, and rectangular problems, and in some of the nonparametric problems considered in section 6.

However, this approach seems to be somewhat artificial, and hereafter we shall restrict ourselves to a single loss function, namely the squared error. There are two reasons for this choice. With squared error for the loss, the mathematical problems are rather simple. And as was remarked above, squared error (if one restricts oneself to unbiased estimates) is the traditional loss function. Fortunately, the squared error loss function is convex, and hence theorem 3.2 permits us to avoid considering randomized estimates.

When the loss function is squared error, we have the following obvious linearity property, which for later reference we state as

THEOREM 4.2. *If $f(X)$ is the minimax estimate for $g(P)$, then $af(X) + b$ is the minimax estimate for $a \cdot g(P) + b$.*

However, as we shall show by an example in the next section, it need not be true that if $X_1, \cdots, X_n$ are independent and $f_i(X_i)$ is the minimax estimate for $g_i(P_i)$, $i = 1, \cdots, n$, then $\sum_{i=1}^n a_i f_i(X_i)$ is the minimax estimate for $\sum_{i=1}^n a_i g_i(P_i)$. This is a definite disadvantage of the minimax principle as compared with the Markoff principle which does possess the linearity property mentioned.

We conclude this section with an explicit solution of the Bayes problem in the squared error case. If the distribution P is itself a random variable distributed over $\mathcal{F}$ according to some distribution λ, we may compare estimates f by means of their expected loss $Q(f) = E[g(P) - f(X)]^2$. Since $Q(f) = E\{E[g(P) - f(X)]^2 \mid X\}$, it is well known that $Q(f)$ is minimized by using the estimate $f(x) = E[g(P) \mid x]$, provided the conditional measures exist. In fact, this result holds even without this assumption.

THEOREM 4.3. *$E[g(P) - f(X)]^2$ is minimized by $f(x) = E[g(P) \mid x]$.*

PROOF.
$$E[g(P) - f(X)]^2 - E\{g(P) - E[g(P) \mid X]\}^2 = E\{E[g(P) \mid X] - f(X)\}^2$$
$$+ 2E[E\{g(P) - E[g(P) \mid X]\}\{E[g(P) \mid X] - f(X)\} \mid X] \geq 0.$$

In applications it is convenient to write $E[g(P) \mid X]$ more explicitly. Suppose that with respect to some measure μ over $\mathfrak{X}$, each distribution $P \in \mathcal{F}$ has a generalized probability density p_P, so that for any A, the probability that $X \in A$ computed for P, is given by

$$\int_A p_P(x)\, d\mu(x) .$$

Minimizing a quadratic expression shows that

$$\frac{\int_{\mathfrak{F}} g(P)p_P(x)\,d\lambda(P)}{\int_{\mathfrak{F}} p_P(x)\,d\lambda(P)} \tag{4.2}$$

is a Bayes solution.

5. Binomial and hypergeometric distributions. In the present section we shall consider three discrete minimax problems.

PROBLEM 1. (Binomial.) Let X be a binomial random variable witd parameter p, $0 \leq p \leq 1$, so that $P(X = k) = \binom{n}{k}p^k(1 - p)^{n-k}$. We shall show that the minimax estimate for p is

$$\frac{X}{n} \cdot \frac{\sqrt{n}}{(\sqrt{n} + 1)} + \frac{1}{2(\sqrt{n} + 1)}. \tag{5.1}$$

Consider any linear estimate $\alpha X + \beta$. The risk $E_p(\alpha X + \beta - p)^2$ is a quadratic function of p which is constantly equal to β^2 when $\alpha = \frac{1}{\sqrt{n}\,(1 + \sqrt{n})}$ and $\beta = \frac{1}{2(1 + \sqrt{n})}$. Hence (5.1) is a constant risk estimate of p. Since it is easily seen that

$$\frac{\int_0^1 p \cdot p^k q^{n-k} \cdot p^{a-1} q^{b-1}\,dp}{\int_0^1 p^k q^{n-k} \cdot p^{a-1} q^{b-1}\,dp} = \frac{a + k}{a + b + n}, \qquad (q = 1 - p),$$

it follows that (5.1) is the Bayes estimate when p is distributed with probability density $C(pq)^{(\sqrt{n}/2)-1}$, and hence by Theorem 2.1 we conclude that (5.1) is the minimax estimate of p.

After obtaining this result we were informed that it had been obtained earlier by H. Rubin, to whom, therefore, the priority belongs.

It is interesting to compare the risk of the above estimate with that of the standard unbiased estimate X/n. We have

$$E\left(\frac{X}{n} - p\right)^2 = \frac{pq}{n},$$

$$E\left[\frac{1}{1 + \sqrt{n}}\left(\frac{X}{\sqrt{n}} + \frac{1}{2}\right) - p\right]^2 = \frac{1}{4(1 + \sqrt{n})^2}.$$

As is easily seen, $\frac{pq}{n} \leq \frac{1}{4(1 + \sqrt{n})^2}$ if and only if

$$\left|p - \frac{1}{2}\right| \geq \frac{\sqrt{1 + 2\sqrt{n}}}{2(1 + \sqrt{n})}.$$

Thus the standard estimate is better than the minimax estimate outside an interval around $p = \frac{1}{2}$ whose length decreases with increasing n, tending to 0 as n tends to infinity. However, for very small values of n the minimax estimate has the smaller risk over nearly the whole range.

PROBLEM 2. (Difference of binomials.) Let X and Y be independent binomial random variables, where $P(X = k) = \binom{n}{k} p_1^k (1 - p_1)^{n-k}$ and $P(Y = l) = \binom{n}{l} p_2^l(1 - p_2)^{n-l}$. By use of theorem 2.1 we shall show that the minimax estimate for $p_1 - p_2$ is $\frac{\sqrt{2n}}{1 + \sqrt{2n}}\left(\frac{X}{n} - \frac{Y}{n}\right)$. For the set ω of theorem 2.1 we take $p_1 = p$, $p_2 = 1 - p$, $0 \leq p \leq 1$, and we let $Z = X + n - Y$. Applying the result of Problem 1 to Z, we find the minimax estimate of p to be $\alpha_{2n} \cdot Z + \beta_{2n}$, and by Theorem 4.2 the minimax estimate based on Z for $p_1 - p_2 = 2p - 1$, is $\frac{\sqrt{2n}}{1 + \sqrt{2n}}\left(\frac{X}{n} - \frac{Y}{n}\right)$, and the risk of this estimate is constant over ω.

To prove that this is also the minimax estimate of $p_1 - p_2$ for the original problem, we consider the risk as a function of p_1 and p_2. It is easy to show that $(1 + \sqrt{2n})^2 R(p_1, p_2) = 2\cdot[p_1(1 - p_1) + p_2(1 - p_2)] + (p_1 - p_2)^2$. Finally it can be shown that $p_1(1 - p_1) + p_2(1 - p_2)$ is maximized, subject to the condition that $p_1 - p_2$ be constant, when $p_1 + p_2 = 1$.

PROBLEM 3. (Hypergeometric.) We finally consider the problem of estimating the number of defectives in a lot from a sample drawn from this lot at random. We denote by N and n the number of elements in lot and sample respectively, and by D and X the corresponding number of defectives. For later reference we note

$$P(X = k) = \frac{\binom{D}{k}\binom{n - D}{n - k}}{\binom{N}{n}},$$

$$E(X) = n\frac{D}{N},$$

$$\sigma_x^2 = \frac{nD(N - n)(N - D)}{N^2(N - 1)}.$$

As in Problem 1 we easily find a linear function of X whose risk is constant. In fact

$$E_D(\alpha X + \beta - D)^2 \equiv \beta^2$$

when

$$\alpha = \frac{N}{n + \sqrt{\frac{n(N - n)}{N - 1}}}, \beta = \frac{N}{2}\left(1 - \frac{\alpha n}{N}\right).$$

To prove that $\alpha X + \beta$ is the minimax estimate of D we shall show that it is the Bayes estimate corresponding to

$$P(D = d) = \int_0^1 \binom{N}{d} p^d q^{N-d} \cdot C\, p^{a-1} q^{b-1}\, dp, \tag{5.2}$$

where $a, b > 0$, and

$$C = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}.$$

In this connection it is useful to notice that since (5.2) is a distribution

$$\sum_{d=0}^{N} \binom{N}{d} \frac{\Gamma(a+d)\,\Gamma(N+b-d)}{\Gamma(N+a+b)} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} = 1. \tag{5.3}$$

Using theorem 4.3, we find the Bayes estimate associated with (5.2) to be

$$f(k) = \frac{\sum_{d=k}^{N-n+k} d \binom{d}{k}\binom{N-d}{n-k}\binom{N}{d} \Gamma(a+d)\,\Gamma(N+b-d)}{\sum_{d=k}^{N-n+k} \binom{d}{k}\binom{N-d}{n-k}\binom{N}{d} \Gamma(a+d)\,\Gamma(N+b-d)}.$$

Replacing d by $(d-a)+a$, and using the relation

$$\binom{d}{k}\binom{N-d}{n-k}\binom{N}{d} = \binom{N-n}{d-k} \cdot \text{(terms not involving } d\text{)},$$

we find:

$$f(k) = \frac{\sum_{i=0}^{N-n} \binom{N-n}{i} \Gamma(d+a+1)\,\Gamma(N+b-d)}{\sum_{i=0}^{N-n} \binom{N-n}{i} \Gamma(d+a)\,\Gamma(N+b-d)} - a.$$

Now apply (5.3) to numerator and denominator separately; then

$$f(k) = k\,\frac{a+b+N}{a+b+n} + \frac{a(N-n)}{a+b+n}.$$

Putting $\dfrac{a+b+N}{a+b+n} = \alpha$, $\dfrac{a(N-n)}{a+b+n} = \beta$ one obtains easily

$$a = \frac{\beta}{\alpha - 1}, \quad b = \frac{N - \alpha n - \beta}{\alpha - 1}.$$

Substituting the values of α and β one finds that $\beta > 0$, $N > \alpha n + \beta$ and that $\alpha > 1$ provided $N > n + 1$. In the special case $N = n$ the result is immediate, while if $N = n + 1$, the result is obtained by giving to D a binomial distribution with $p = \frac{1}{2}$.

In a similar manner it is easily shown that the least squares estimate for a linear function of one or more of the θ's, is the minimax estimate.

Theorem 6.5 gives a justification of the least squares estimate different from that of the Markoff theorem. In the Markoff theorem, it is shown that the least squares estimate has *uniformly* smallest risk among all linear unbiased estimates; here it is shown that the least squares estimate minimizes the maximum risk among all estimates. (The assumptions concerning variances also differ.)

7. Prediction problems. Frequently one is interested in estimating the value of a random variable rather than that of a parameter. A customary method for this is to estimate the expectation of the random variable (a parameter) and then to "identify" the variable and its expectation; i.e., to use the estimate of the expectation as a prediction for the variable. As we shall see below one is led to this procedure if one adopts the point of view of unbiased estimation, so that from this point of view prediction poses no new problem. This however is no longer true when one employs the minimax principle.

Consider a pair X, Y of random variables having a joint distribution P belonging to a family $\mathfrak{F}$ of distributions. It is desired to use the observed X to predict, say, $g(Y)$. We are interested in minimax predictions; i.e., functions $f(X)$ which minimize $\sup_{P\epsilon\mathfrak{F}} E_P W[g(Y), f(X)]$. To obtain minimax predictions we need the following analogue of Theorem 2.1.

THEOREM 7.1. *Let $\{P_\theta\}$, $\theta \epsilon \omega$ be a parametric subfamily of $\mathfrak{F}$, and let λ be a probability measure over ω. Suppose that f is such that $\int E_\theta W[g(Y), f(X)]\, d\lambda(\theta)$ is minimum, and that*

(i) *$E_\theta W[g(Y), f(X)]$ is constant, say $= c$, for all $\theta \epsilon \omega$,*

(ii) *$E_P W[g(Y), f(X)] \leqq c$ for all $P \epsilon \mathfrak{F}$.*

Then f is a minimax prediction for $g(Y)$.

The proof is the exact analogue of that of theorem 2.1.

COROLLARY 7.2. *A constant risk Bayes prediction is a minimax prediction.*

Suppose now that X and Y are independent and that $W[g(y), f(x)] = [g(y) - f(x)]^2$. Consider the problem first from the point of view of unbiasedness. A prediction could reasonably be called unbiased if $E_P f(X) = E_P g(Y)$. Subject to unbiasedness, the risk is given by $E_P[g(Y) - f(X)]^2 = \sigma_P^2 f(X) + \sigma_P^2\, g(Y)$. But $\sigma_P^2 g(Y)$ is a known function of P, and hence the problem of minimizing (for a particular P) the expected squared error reduces to that of finding an unbiased estimate of $E_P g(Y)$ with minimum variance at P. In a similar way one sees, without any restriction to unbiased predictions, that the Bayes prediction for $g(Y)$ is the same as the Bayes estimate for $E_P g(Y)$, and hence that formula (4.2), with $g(P)$ replaced by $E_P g(Y)$, may be used if the assumptions there made are valid.

One might expect that as in the unbiased theory the prediction will coincide with the estimate. This however is not the case since the λ's that give constant risk in the two cases will usually be distinct. In fact the two problems are rather different in that the "least favorable" λ for the prediction problem must not only take into account the difficulty of finding the correct value of θ for various a priori distributions but also the difficulty of predicting $g(Y)$ when θ is known.

As a first example consider the prediction analogue of problem 1 of section 5. Let X, Y be independent binomial variables such that $P(X = k) = \binom{m}{k} p^k(1-p)^{m-k}$ and $P(Y = l) = \binom{n}{l} p^l(1-p)^{m-l}$. We shall obtain the minimax prediction of Y in a manner quite analogous to the one in which we determined the minimax estimate of p. Actually, the present problem is a generalization of the earlier one, to which it can be reduced by letting $n \to \infty$. First it is easily seen that

$$E\left(\alpha \frac{X}{m} + \beta - \frac{Y}{n}\right)^2$$

is a quadratic function of p, which when $m > 1$ is constant for

$$\alpha = \frac{m}{m-1}\left[1 - \sqrt{\frac{1}{m} + \frac{1}{n} - \frac{1}{mn}}\right],$$
$$\beta = \frac{1-\alpha}{2}.$$

But we have already seen that $\alpha \frac{X}{m} + \beta$ is the Bayes solution corresponding to $Cp^{a-1}q^{b-1}$ where $\alpha = \frac{m}{m+a+b}$, $\beta = \frac{a}{m+a+b}$. Clearly $\beta = \frac{1-\alpha}{2}$ when $a = b$, and $a > 0$ provided $0 < \alpha < 1$, which is easily verified when $m, n > 1$. We note that as $n \to \infty$, the values of α, β tend to those of the minimax estimate of P.

When $m = 1$, $E\left(\alpha \frac{X}{m} + \beta - \frac{Y}{n}\right)^2$ is constant for $\alpha = \frac{n-1}{2n}$, $\beta = \frac{1-\alpha}{2}$, and again $\alpha \frac{X}{m} + \beta$ is the Bayes estimate of a beta distribution when $n > 1$, and hence minimax.

Finally in the case $n = 1$, the situation degenerates. Since $E(\frac{1}{2} - Y)^2 = \frac{1}{4}$, the prediction $f(X) = \frac{1}{2}$ has constant risk. In addition it is the Bayes prediction corresponding to the distribution which assigns probability 1 to $p = \frac{1}{2}$. Hence in this case, regardless of the value of X one would predict for Y the value $\frac{1}{2}$.

It is interesting that the above prediction problem can be interpreted also as an estimation problem in the following manner. Suppose a lot of size $N = m + n$ is such that the number of defectives follow a binomial distribution; this is the case when the items making up the lot are produced by a manufacturing process that is in statistical control. It is desired to estimate from a sample of size m taken from this lot, the proportion of defectives in the remainder. That this is equivalent to the prediction problem treated above follows from a remark of Mood [13] that in such a lot the number of defectives in the sample and in the remainder are independently distributed according binomial distributions with common p.

We can again use the binomial results to obtain the solutions of certain nonparametric problems. For example, let $X_1, \cdots, X_m$ be independently and identically distributed on [0, 1] and let $Y_1, \cdots, Y_n$ be another sample from the same distribution. Then the minimax prediction for $\bar{Y}$ is given by $\alpha\bar{X} + \beta$ with $\alpha = \frac{m}{m-1}\left[1 - \sqrt{\frac{1}{m} + \frac{1}{n} - \frac{1}{mn}}\right]$, $\beta = \frac{1-\alpha}{2}$. This follows from the fact that

$$
\begin{aligned}
E(\alpha\bar{X} + \beta - \bar{Y})^2 &= E[\alpha(\bar{X} - \bar{\mu}) - (\bar{Y} - \mu) + (\beta + (\alpha - 1)\mu)]^2 \\
&= \alpha^2\left(\frac{1}{m} + \frac{1}{n}\right)\sigma^2 + [\beta + (\alpha - 1)\mu]^2 \\
&\leq \alpha^2\left(\frac{1}{m} + \frac{1}{n}\right)\mu(1 - \mu) + [\beta + (\alpha - 1)\mu]^2.
\end{aligned}
$$

An analogous modification clearly is possible for theorem 6.4.

For the situation considered in 6.5, the prediction problem gives the same result as the estimation problem. For consider first two samples $X_1, \cdots, X_m$; $Y_1, \cdots, Y_n$ from a normal distribution with known variance σ^2. Here

$$E_\theta[f(X_1, \cdots, X_n) - \bar{Y}]^2 = E_\theta[f(X_1, \cdots, X_m) - \theta]^2 + \frac{\sigma^2}{n},$$

and hence the risk differs from that of the estimation problem only by a constant. Thus $\bar{X}$ is the minimax prediction of $\bar{Y}$, and it is then seen immediately that it is also the minimax prediction for $\bar{Y}$ when of the underlying common distribution of the X's and Y's it is assumed only that the variance is bounded.

REFERENCES

[1] G. W. BROWN, "On small sample estimation," *Annals of Math. Stat.*, Vol. 18 (1949), p. 514.

[2] A. WALD, "Contributions to the theory of statistical estimation and testing hypotheses," *Annals of Math. Stat.*, Vol. 10 (1939), p. 299.

[3] A. WALD, *On the Principles of Statistical Inference*, Notre Dame Math. Lectures, No. 1 (1942).

[4] A. WALD, "Statistical decision functions which minimize the maximum risk," *Annals of Math.*, Vol. 46 (1945), p. 265.

[5] A. WALD, "Statistical decision functions," *Annals of Math. Stat.*, Vol. 20 (1949), p. 165.

[6] C. STEIN AND A. WALD, "Sequential confidence intervals for the mean of a normal distribution with known variance," *Annals of Math. Stat.*, Vol. 18 (1947), p. 427.

[7] P. R. HALMOS, "The theory of unbiased estimation," *Annals of Math. Stat.*, Vol. 17 (1946), p. 34.

[8] E. L. LEHMANN AND C. STEIN, "Most powerful tests of composite hypotheses. I. Normal distributions," *Annals of Math. Stat.*, Vol. 19 (1948), p. 495.

[9] A. WALD, "An essentially complete class of admissible decision functions," *Annals of Math. Stat.*, Vol. 18 (1947), p. 549.

[10] A. KOLMOGOROFF, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Berlin, 1933.

[11] D. BLACKWELL, "Conditional expectation and unbiased sequential estimation," *Annals of Math. Stat.*, Vol. 18 (1947), p. 105.

[12] E. W. BARANKIN, "Extension of a theorem of Blackwell," *Annals of Math. Stat.*, Vol. 21 (1950), p. 280.

[13] A. M. MOOD, "On the dependence of sampling inspection plans upon population distributions," *Annals of Math. Stat.*, Vol. 14 (1943), p. 145.