UNBIASED ESTIMATES WITH MINIMUM VARIANCE

By Charles Stein

University of California, Berkeley

Summary. Subject to certain restrictions, a characterization of unbiased estimates with minimum variance is obtained. For two fairly broad classes of problems, solutions are given which are more readily applicable. These are used to obtain such estimates in some particular cases. The applicability of the results to problems of sequential estimation is pointed out. The problem of unbiased estimation is not at present of much practical importance, but is of some theoretical interest and has been treated by many statisticians. Also, the method used in this paper may be applicable to other problems in statistics.

1. Introduction.

Let R be a space of points x, B an additive class of subsets C of R and μ a measure over B such that R can be represented as the union of a countable collection of elements of B each of which has finite μ-measure. Let Ω be a set called the parameter space and let X be a random variable distributed in accordance with the probability density function $p(x \mid \theta)$ for some $\theta \epsilon \Omega$, so that for any $C \epsilon B$

$$P\{X \epsilon C \mid \theta\} = \int_C p(x \mid \theta)\, d\mu(x).$$

A measurable real-valued function $f(x)$ on R is called an unbiased estimate of the real-valued function $g(\theta)$ on Ω if, for every $\theta \epsilon \Omega$

$$E(f(X) \mid \theta) = \int f(x) p(x \mid \theta)\, d\mu(x) = g(\theta). \tag{1}$$

The problem considered in this paper is that of finding an unbiased estimate f^* of g which minimizes the variance at θ_0. Since this variance is

$$\begin{aligned} &E([f(X) - g(\theta_0)]^2 \mid \theta_0) \\ &\qquad = \int [f(x) - g(\theta_0)]^2 p(x \mid \theta)\, d\mu(x) \\ &\qquad = \int [f(x)]^2 p(x \mid \theta)\, d\mu(x) - \left[\int f(x) p(x \mid \theta_0)\, d\mu(x)\right]^2, \end{aligned} \tag{2}$$

this problem is equivalent to minimizing

$$\int [f(x)]^2 p(x \mid \theta_0)\, d\mu(x) \tag{3}$$

subject to (1). It will be convenient to introduce the measure

$$\nu(C) = \int_C p(x \mid \theta_0)\, d\mu(x) \tag{4}$$

and the probability ratios

$$\pi(x \mid \theta) = \frac{p(x \mid \theta)}{p(x \mid \theta_0)}. \tag{5}$$

We suppose $\pi(x \mid \theta)$ finite for almost all x, and all θ. When we say "for almost all x," we mean "except for a set of μ-measure 0."

In most practical problems, the set R is a subset of some finite-dimensional Euclidean space and μ is either ordinary Lebesgue measure or, in the case where R is countable, counting measure which makes the measure of a set the number of points it contains. An exception is the application to sequential analysis considered in section 3 below, in which R is a countable union of sets, each of which is a subset of a finite dimensional Euclidean space. For the basic notation and concepts of the theory of integration see Saks [2], Ch. I.

We shall define

$$A(\theta_1, \theta_2) = \int \pi(x \mid \theta_1)\pi(x \mid \theta_2)\, d\nu(x), \tag{6}$$

and suppose

$$A(\theta, \theta) < \infty \text{ for all } \theta. \tag{7}$$

By Schwartz's inequality this implies that $A(\theta_1, \theta_2) < \infty$ for all θ_1, θ_2. If (7) is not true then it may happen that there exists no unbiased estimate with minimum variance even though there exist unbiased estimates. Consider, for example, the case where Ω consists of two point, 0 and 1, and $g(\theta) = \theta$, and

$$p(x \mid 0) = \begin{cases} 1 \text{ for } 0 < x < 1 \\ 0 \text{ otherwise} \end{cases}$$

$$p(x \mid 1) = \begin{cases} \frac{1}{2}x^{-\frac{1}{2}} \text{ for } 0 < x < 1 \\ 0 \text{ otherwise} \end{cases}$$

and μ is ordinary Lebesgue measure. It is clear that there exist unbiased estimates of θ with arbitrarily small positive variance at $\theta = 0$ but there exists none with 0 variance.

2. The principal theorem. In accordance with the usual terminology we denote by L_2 the class of all measurable functions ϕ such that

$$\int [\phi(x)]^2\, d\nu(x) < \infty. \tag{8}$$

Finally, G is the class of all functions ψ expressible in the form

$$\psi(\theta) = \int \phi(x)\pi(x \mid \theta)\, d\nu(x) \quad \text{with} \quad \phi \in L_2. \tag{9}$$

THEOREM 1. *If $\pi(x \mid \theta)$ is finite for all θ and almost all x, and (7) is satisfied, and there exists an unbiased estimate of g, then there exists an unbiased estimate*

f^* of g which minimizes (3). *If f^* has finite variance then any other unbiased estimate of g with minimum variance at θ_0 is essentially equal to f^*, that is, differs from f^* only on a set of μ-measure* 0. *A function f is an unbiased estimate of g with minimum variance at θ_0 if and only if there exists a real-valued functional T on G for which*

$$TA(\theta, \theta_1) = g(\theta_1) \text{ for all } \theta_1 \in \Omega, \tag{10}$$

$$T \int \phi(x)\pi(x \mid \theta)\, d\nu(x) = \int \phi(x)f(x)\, d\nu(x) \text{ for all } \phi \in L_2. \tag{11}$$

(*The preceding sentence does not assume the existence of an unbiased estimate of g.*) *The minimum variance is $Tg(\theta) - [g(\theta_0)]^2$.*

Proof. Let $\{f_i\}$ be a sequence of unbiased estimates of g such that

$$\lim_{i\to\infty} \int [f_i(x)]^2\, d\nu(x) = \underset{f}{\text{g.l.b.}} \int [f(x)]^2\, d\nu(x)$$

where f ranges over all unbiased estimates of g. Then by the weak compactness of every sphere in L_2 (see [1], p. 10) there exists $f^* \in L_2$ and an increasing sequence $\{n_i\}$ of integers for which

$$\int \phi f^*\, d\nu = \lim_{i\to\infty} \int \phi f_{n_i}\, d\nu \text{ for all } \phi \in L_2.$$

Since $\pi(x \mid \theta) \in L_2$ by (7), this implies that f^* is an unbiased estimate of g. Also

$$\int [f^*]^2\, d\nu \leq \lim_{i\to\infty} \int f_{n_i}^2\, d\nu = \underset{f}{\text{g.l.b.}} \int f^2\, d\nu. \tag{12}$$

Thus f^* is an unbiased estimate of g with minimum variance.

Let $\phi_1 \in L_2$ be such that

$$\int \phi_1(x)\pi(x \mid \theta)\, d\nu(x) = 0 \text{ for all } \theta \in \Omega. \tag{13}$$

Then, using the f^* defined in the last paragraph, we obtain for any real ε

$$0 \leq \int (f^* + \varepsilon\phi_1)^2\, d\nu - \int [f^*]^2\, d\nu = 2\varepsilon \int \phi_1 f^*\, d\nu + \varepsilon^2 \int \phi_1^2\, d\nu \tag{14}$$

since $f^* + \varepsilon\phi_1$ is an unbiased estimate of g. Dividing (14) by ε and letting $\varepsilon \to 0$ we obtain

$$\int \varphi_1 f^*\, d\nu = 0. \tag{15}$$

If a function in G can be represented in two ways,

$$\int \phi(x)\pi(x \mid \theta)\, d\nu(x) = \int \varphi'(x)\pi(x \mid \theta)\, d\nu(x),$$

and consequently $\phi_1 = \phi' - \phi$ satisfies (13) and (15). Thus (11) defines a functional on G in a consistent way. Also, this functional satisfies (10) since

$$TA(\theta, \theta_1) = T \int \pi(x \mid \theta_1)\pi(x \mid \theta)\, d\nu(x)$$

$$= \int \pi(x \mid \theta_1) f^*(x)\, d\nu(x) = g(\theta_1).$$

By (2) and (11) the minimum variance is

$$\int [f^*(x)]^2\, d\nu(x) - [g(\theta_0)]^2 = T \int f^*(x)\pi(x \mid \theta)\, d\nu(x) - [g(\theta_0)]^2$$

$$= Tg(\theta) - [g(\theta_0)]^2.$$

To prove the converse, let f^* be any function in L_2 for which there exists a functional T satisfying (10) and (11). By (11) with $\phi(x) = \pi(x \mid \theta_1)$,

$$\int f^*(x)\pi(x \mid \theta_1)\, d\nu(x) = T \int \pi(x \mid \theta)\pi(x \mid \theta_1)\, d\nu(x)$$

$$= TA(\theta, \theta_1) = g(\theta_1)$$

by (10), so that f^* is an unbiased estimate of g. Any other unbiased estimate f of g with finite variance at θ_0 is an element of L_2. Thus from (1) and (11) we obtain

$$Tg(\theta) = \int ff^*\, d\nu$$

$$= \int [f^*]^2\, d\nu.$$

Applying Schwartz's inequality to the middle expression we obtain

$$\int [f^*]^2\, d\nu \leq \int [f]^2\, d\nu$$

with strict inequality unless f is essentially equal to f^*.

COROLLARY 1. *Suppose $\pi(x \mid \theta)$ is finite for all θ and almost all x and (7) holds. Let $H_1(x, d)$ be the set of all $\theta \in \Omega$ such that $\pi(x \mid \theta) > d$, and let H be the smallest additive class containing all $H_1(x, d)$. Suppose there exists an additive set function λ over H such that there exists a finite collection of parameter points θ_k and positive number c_k such that*

$$\int \pi(x \mid \theta) \mid d\lambda(\theta) \mid \leq \sum c_k \pi(x \mid \theta_k) \tag{16}$$

for almost all x, and

$$\int A(\theta, \theta_1)\, d\lambda(\theta) = g(\theta_1). \tag{17}$$

Then the unbiased estimate of $g(\theta)$ with minimum variance at θ_0 is

$$f^*(x) = \int \pi(x \mid \theta)\, d\lambda(\theta) \tag{18}$$

and the minimum variance is

$$\int g(\theta)\, d\lambda(\theta) - [g(\theta_0)]^2. \tag{19}$$

Proof: We need only show that (10) and (11) are satisfied by

$$T\psi(\theta) = \int \psi(\theta)\, d\lambda(\theta)$$

and (18). But

$$TA(\theta, \theta_1) = \int A(\theta, \theta_1)\, d\lambda(\theta) = g(\theta_1) \tag{20}$$

by (17) and

$$\begin{aligned} T \int \phi(x)\pi(x \mid \theta)\, d\nu(x) &= \int d\lambda(\theta) \int \phi(x)\pi(x \mid \theta)\, d\nu(x) \\ &= \int \phi(x)\, d\nu(x) \int \pi(x \mid \theta)\, d\lambda(\theta) = \int \phi(x) f^*(x)\, d\nu(x). \end{aligned} \tag{21}$$

Since each of the functions $\phi(x)$, $\pi(x \mid \theta)$ considered as a function of x and θ is measurable (BH), their product is also. The interchange of order of integration in (21) is justified by Fubini's Theorem (Saks [2], p. 87) and (16) which by (9) implies that $\int \mid d\lambda(\theta) \mid \int \phi(x)\pi(x \mid \theta)\, d\nu(x) < \infty$. The equations (20) and (21) are equivalent to (10) and (11) respectively.

Corollary 2. *Suppose $\pi(x \mid \theta)$ is finite for all θ and almost all x and* (7) *holds. Suppose also that Ω is a set of real numbers and:*

(i) *for some m, either a positive integer or $+\infty$, $\pi(x \mid \theta)$ is, for almost all x, differentiable m times with respect to θ at $\theta = \theta_0$,*

(ii) *for each $n < m$ there exists a finite collection of parameter values $\theta_{n,k}$ and positive constants $c_{n,k}$ such that*

$$\left| \frac{\pi^{(n)}(x \mid \theta_0 + \delta) - \pi^{(n)}(x \mid \theta_0)}{\delta} \right| \leq \sum_k c_{n,k}\pi(x \mid \theta_{n,k}) \tag{22}$$

for all δ whose absolute value is sufficiently small and almost all x,

(iii) *there exist constants a_n such that for all θ_1,*

$$g(\theta_1) = \sum_{n=0}^{m} a_n \left[\frac{\partial^n}{\partial \theta^n} A(\theta, \theta_1) \right]_{\theta=\theta_0}, \tag{23}$$

(iv) *there exists a finite collection of parameter values θ_k and positive constants*

c_k such that

$$\sum_{n=0}^{m} \left| a_n \left[\frac{\partial^n}{\partial \theta^n} \pi(x \mid \theta) \right]_{\theta=\theta_0} \right| \leq \sum_k c_k \pi(x \mid \theta_k). \tag{24}$$

Then the unbiased estimate of $g(\theta)$ with minimum variance at θ_0 is

$$f^*(x) = \sum_{n=0}^{m} a_n \left[\frac{\partial^n}{\partial \theta^n} \pi(x \mid \theta) \right]_{\theta=\theta_0}. \tag{25}$$

The minimum variance is

$$\sum_{n=0}^{m} a_n \left[\frac{\partial^n}{\partial \theta^n} g(\theta) \right]_{\theta=\theta_0}.$$

PROOF. We need only show that the functional T defined by

$$T \int \phi(x)\pi(x \mid \theta)\, d\nu(x) = \sum_{n=0}^{m} a_n \frac{\partial^n}{\partial \theta^n} \int \phi(x)\pi(x \mid \theta)\, d\nu(x) \Bigg]_{\theta=\theta_0} \tag{26}$$

satisfies (10) and (11) with f^* given by (25). Equation (23) yields (11) immediately. Also

$$T \int \phi(x)\pi(x \mid \theta)\, d\nu(x) = \sum_{n=0}^{m} a_n \frac{\partial^n}{\partial \theta^n} \int \phi(x)\pi(x \mid \theta)\, d\nu(x) \Bigg]_{\theta=\theta_0}$$

$$= \Sigma\, a_n \int \phi(x) \frac{\partial^n}{\partial \theta^n} \pi(x \mid \theta) \Bigg]_{\theta=\theta_0} d\nu(x)$$

by (9), (i), (22) and Lebesgue's Theorem on term by term integration (Saks [2] p. 29.). Using (24) and Lebesgue's Theorem, we find that this is equal to

$$\int \phi(x) \Sigma\, a_n \frac{\partial^n}{\partial \theta^n} \pi(x \mid \theta) \Bigg]_{\theta=\theta_0} d\nu(x) = \int \phi(x) f^*(x)\, d\nu(x).$$

which completes the proof.

There is an obvious combination of Corollaries 1 and 2 which will not be stated explicitly. Also Corollary 2 can be extended to involve differentiation with respect to several parameters. It would be of considerable interest to obtain a characterization of all possible functionals T in terms of the usual operations such as integration and differentiation. Also, the methods used here should be applicable, with some modifications, to other problems of minimization subject to an infinite set of side conditions.

COROLLARY 3. *Suppose that subject to the condition of Theorem* 1, *for $i = 1, 2, f_i^*$ are unbiased estimates of g_i with minimum variance at θ_0. Then $f_1^* + f_2^*$ is an unbiased estimate of $g_1 + g_2$ with minimum variance at θ_0.*

This follows immediately from (11) and (12) in Theorem 1. Actually, the restriction to problems satisfying the conditions of Theorem 1 is unnecessary, but we shall not prove this here.

3. Some special cases. We first consider a problem which is of little practical interest but serves well as an illustration of Corollary 1. Let X be a single observation from a uniform distribution on the interval $(\theta, \theta + 1)$, i.e.

$$p(x \mid \theta) = \begin{cases} 1 \text{ if } \theta < x < \theta + 1 \\ 0 \text{ otherwise.} \end{cases}$$

We suppose θ lies in the interval $(-N, N - 1)$ where N is a given positive integer, and take as the distribution for which the variance is to be minimized

$$p(x \mid \theta_0) = \begin{cases} \dfrac{1}{2N} \text{ if } -N < x < N \\ 0 \text{ otherwise.} \end{cases}$$

This is the same as using the original p.d.f. $p(x \mid \theta)$ with θ a random variable taking on the values $-N, -N + 1, \cdots, N - 1$ with equal probability. The measure μ is of course ordinary Lebesgue measure. Then

$$\pi(x \mid \theta) = \begin{cases} 2N \text{ if } \theta < x < \theta + 1 \\ 0 \text{ otherwise} \end{cases} \tag{27}$$

and

$$\frac{1}{2N}A(\theta_1, \theta_2) = \begin{cases} 0 & \text{if } \theta_1 < \theta_2 - 1 \\ \theta_1 - \theta_2 + 1 & \text{if } \theta_2 - 1 < \theta_1 < \theta_2 \\ \theta_2 - \theta_1 + 1 & \text{if } \theta_2 < \theta_1 < \theta_2 + 1 \\ 0 & \text{if } \theta_2 + 1 < \theta_1 . \end{cases} \tag{28}$$

For $-N < \theta_1 < N - 1$, equation (17) becomes

$$\int_{\max(-N,\, \theta_1 - 1)}^{\theta} (\theta - \theta_1 + 1)\, d\lambda(\theta) + \int_{\theta_1}^{\min(N-1,\, \theta_1+1)} (\theta_1 - \theta + 1)\, d\lambda(\theta) = g(\theta_1)/2N \tag{29}$$

and (18) becomes

$$f^*(x)/2N = \lambda(\min[N - 1, x]) - \lambda(\max[-N, x - 1]). \tag{30}$$

The reader will not be confused by the use of λ as a point function here, and as a set function in Corollary 1. Using (30) and integration by parts (Saks [2], p. 102) we can rewrite (29) as

$$\int_{\theta_1}^{\theta_1+1} f^*(x)\, dx = g(\theta_1), \tag{31}$$

which is merely the condition that f^* be an unbiased estimate of g. It is clear from (31) that g admits an unbiased estimate if and only if it is absolutely continuous. Differentiating (31) we obtain

$$f^*(\theta + 1) - f^*(\theta) = g(\theta). \tag{32}$$

Consequently the general solution of (31) is

$$f^*(\theta) = \sum_{i=1}^{[\theta]+N} g'(\theta - i) + \gamma(\theta), \tag{33}$$

where γ is a function of period 1 such that

$$\int_0^1 \gamma(\theta)\, d\theta = 0. \tag{34}$$

Here, contrary to the usual convention, $[\theta]$ denotes the largest integer less than θ. The one of (33) which minimizes the variance at θ_0 is determined by the condition that there exist λ satisfying (30). Let y be any number on the half-closed interval $(-N, -N + 1)$, and sum (30) for $x = y, y + 1 \cdots y + 2N - 1$. This yields

$$\frac{1}{2N} \sum_{j=0}^{2N-1} f^*(y + j) = \lambda(N - 1) - \lambda(-N). \tag{35}$$

Carrying out the same computation on (33) we obtain

$$\frac{1}{2N} \sum_{j=0}^{2N-1} \sum_{i=1}^{j+N} g'(y + j - i) + \gamma(y) = \lambda(N - 1) - \lambda(-N). \tag{36}$$

Combining (34) and (35) we find that the proper choice of γ is that which gives

$$\begin{aligned} f^*(x) = {} & \sum_{i=0}^{[x]+N+1} \frac{1 + i}{2N} g'(x - [x] - N + i) \\ & + \sum_{i=x+N}^{2N-2} \left(\frac{1 + i}{2N} - 1\right) g'(x - [x] - N + i) \\ & + \frac{1}{2N} \sum_{j=1}^{2N-1} \{g(j - N) - g(-N)\}. \end{aligned} \tag{37}$$

If the limit of (37) as $N \to \infty$ exists, it agrees with Nörlund's simplest definitions of the principal solution of (32) (see Milne-Thompson [3] formula (2) p. 201) whenever the latter is applicable. The author has not checked the agreement with Nörlund's more general definitions.

Next we consider the problem of obtaining an unbiased estimate of $g(\theta)$ with minimum variance at θ_0 when X consists of n independent observations, each uniformly distributed over the interval $(0, \theta)$. Here θ is an unknown positive number. The result is independent of the choice of θ_0. Clearly a necessary and sufficient condition for the existence of an unbiased estimate of g is that g be absolutely continuous. Corollary 1 can be applied to obtain as the best unbiased estimate $g(Y) + \frac{Y}{n} g'(Y)$ where $Y = \max(X_1 \cdots X_n)$. However, this result can be obtained much more simply by observing that, given any sufficient statistic Z, there exists an unbiased estimate with minimum variance which is a function only of Z. A proof of this is given by Blackwell [4]. But Y is a sufficient statistic, and the condition that $f^*(Y)$ be an unbiased estimate of g is that

$$\frac{n}{\theta^n} \int_0^\theta f^*(y) y^{n-1}\, dy = g(\theta).$$

This has as its unique solution that given above.

A similar situation holds when the X_i, $i = 1 \cdots n$, are independently normally distributed with unknown common mean θ and unit variance. Here Corollary 1 is not applicable, but Corollary 2 is. The result can again be obtained more simply as the unique solution of the integral equation

$$\frac{1}{\sqrt{2\pi}} \int f_0^*(y)^{-\frac{1}{2}(y-\theta\sqrt{n})^2}\, dy = g(\theta)$$

with

$$f^*(x_1, \cdots, x_n) = f_0^*(y), \qquad y = \frac{1}{\sqrt{n}} \sum_1^n x_i .$$

It should be observed that the methods of section 2 are applicable also to problems of sequential estimation. Let $X_1, X_2, \cdots$ be a sequence of real-valued random variables such that $(X_1, \cdots, X_n)$ have the joint p.d.f. $p_n(x_1, \cdots, x_n \mid \theta)$ for some unknown $\theta \epsilon \Omega$. Suppose it has been decided to terminate the procedure on the m^{th} observation if $(X_1, \cdots, X_m) \epsilon \bar{R}_m$ for some given sets $\bar{R}_m$ in m space, and suppose these sets are so chosen that the probability of termination is 1 for all θ. Then we can define the space $R = \bigcup_m \bar{R}_m$, the union of the $\bar{R}_m$, the measure

$$\mu(A) = \sum_m \mu_m(AR_m)$$

for any set $A \subset R$ for which the intersections $A \cap R_m$ are Borel sets, where μ_m is ordinary m-dimensional Lebesgue measure, and the probability density functions

$$p(x \mid \theta) = p_m(x_1 \cdots x_m \mid \theta) \quad \text{if} \quad x = (x_1 \cdots x_m) \epsilon \bar{R}_m .$$

The previous results are then applicable. Most of the familiar results in the theory of statistical inference can be extended to sequential problems in the same way. Of course the interesting and difficult problems of sequential analysis are usually concerned chiefly with the appropriate choice of the regions $\bar{R}_m$.

4. Connections with the work of other authors. Many lower bounds for the variance of an unbiased estimate were obtained by Bhattacharyya [5], and some results were obtained earlier by others whose results are referred to by Bhattacharyya. His work has been extended to sequential problems as indicated in section 3 above by G. R. Seth in a doctoral dissertation at Columbia University. This leads to results analogous to, but in some respects more general than those of Wolfowitz [6]. Among other papers on sequential estimation, there are the one by Blackwell [4] already referred to, and the one by Girshick, Mosteller, and Savage [7]. These deal mainly with problems in which there is a unique unbiased estimate based on a sufficient statistic.

The author is indebted to A. Wald, J. L. Hodges, E. Barankin, and H. Rubin for some helpful suggestions and comments.

REFERENCES

[1] B. v. Sz. Nagy, "Spektraldarstellung linearer Transformationen des Hilbertschen Raumes," *Ergebnisse der Mathematik*, Vol. 5, No. 5 (1942).

[2] S. Saks, *Theory of the Integral*, Monografie Matematyczne, Tom VII, Warsaw, 1937.

[3] L. M. Milne-Thompson, *The Calculus of Finite Differences*. Macmillan, London, 1933.

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

[5] A. Bhattacharyya, "On some analogues of the amount of information and their use in statistical estimation," *Sankhya*, Vol. 8 (1946), p. 1 and Vol. 8 (1947), p. 201.

[6] J. Wolfowitz, "The efficiency of sequential estimates and Wald's equation for sequential processes," *Annals of Math. Stat.*, Vol. 18 (1947), p. 215.

[7] M. Girshick, F. Mosteller, and L. Savage, "Unbiased estimates for certain binomial sampling problems," *Annals of Math. Stat.*, Vol. 17 (1946), p. 13.