CONDITIONAL EXPECTATION AND UNBIASED SEQUENTIAL ESTIMATION[1]

By David Blackwell

Howard University

1. Summary. It is shown that $E[f(x)\ E(y \mid x)] = E(fy)$ whenever $E(fy)$ is finite, and that $\sigma^2 E(y \mid x) \leq \sigma^2 y$, where $E(y \mid x)$ denotes the conditional expectation of y with respect to x. These results imply that whenever there is a sufficient statistic u and an unbiased estimate t, not a function of u only, for a parameter θ, the function $E(t \mid u)$, which is a function of u only, is an unbiased estimate for θ with a variance smaller than that of t. A sequential unbiased estimate for a parameter is obtained, such that when the sequential test terminates after i observations, the estimate is a function of a sufficient statistic for the parameter with respect to these observations. A special case of this estimate is that obtained by Girshick, Mosteller, and Savage [4] for the parameter of a binomial distribution.

2. Conditional expectation. Denote by x any (not necessarily numerical) chance variable and by y any numerical chance variable for which $E(y)$ is finite. There exists a function of x, the conditional expectation of y with respect to x [3, pp. 95–101, 5, pp. 41–44] which we denote, as usual, by $E(y \mid x)$ and which is uniquely defined except for events of zero probability, such that whenever $f(x)$ is the characteristic function of an event F depending only on x (i.e. $f = 1$ when F occurs and $f = 0$ when F does not occur), the equation

$$E[f(x)E(y \mid x)] = E[f(x)y] \tag{1}$$

holds. Now if $f(x)$ is a simple function, i.e. a finite linear combination of characteristic functions, it is clear from the linearity of expectation that (1) continues to hold. Quite generally, we shall prove

Theorem 1: *The equation* (1) *holds for every function* $f(x)$ *for which* $E[f(x)y]$ *is finite.*

To simplify notation, we write $E(z \mid x) = E_x z$ for any chance variable z. The following corollary to Theorem 1 asserts simply that the operations E_x and multiplication by $f(x)$ are commutative. This fact, which is trivially equivalent to Theorem 1, has been stated by Kolmogoroff [5, p. 50].

Corollary: *If* $E[f(x)y]$ *is finite, then* $E_x[f(x)y] = f(x)E_x y$.

Proof of Corollary: If $g(x)$ is a characteristic function, then $E(gfE_x y) = E(gfy)$ by Theorem 1. Since $E_x(fy)$ is unique, the Corollary follows.

Proof of Theorem 1: Since Theorem 1 holds when $f(x)$ is a simple function and the product of a simple function and a characteristic function is a simple function, the Corollary holds when $f(x)$ is a simple function.

[1] The author is indebted to M. A. Girshick for suggesting the problem which led to this paper and for many helpful discussions.

Now let $f(x)$ be any function for which $E(fy)$ is finite. There is a sequence of simple functions $f_n(x)$ such that $f_n(x) \to f(x)$ and $|f_n(x)| \leq |f(x)|$. For instance we may define $f_n(x) = m/n$ when $m/n \leq f(x) < (m+1)/n$, $0 \leq m \leq n^2$, $f_n(x) = m/n$ when $(m-1)/n \leq f(x) < m/n$, $0 \geq m \geq -n^2$, $f_n(x) = 0$ otherwise.

We recall the following proposition of Doob [2, p. 296]:

$$|E_x y| \leq E_x |y| \tag{2}$$

with probability one. Then, using the Corollary (for simple functions) and (2), we have $|f_n E_x y| = |E_x(f_n y)| \leq E_x |f_n y| \leq E_x |fy|$. Also

$$E(f_n E_x y) = E(f_n y). \tag{3}$$

Since the two sequences of functions $f_n E_x y$, $f_n y$ are bounded in absolute value by the summable functions $E_x |fy|$, $|fy|$, Lebesgue's theorem [8, p. 29] applied to (3) yields (1).

In section 3 we shall use the fact that if u is a sufficient statistic for a parameter θ and f is any unbiased estimate for θ, then $E(f \mid u)$ (which, since u is a sufficient statistic, is a function of u independent of θ) is an unbiased estimate for θ. This is obvious, since it follows from the definition of conditional expectation that the two chance variables f and $E(f \mid u)$ have the same expected value. The interesting fact is that the estimate $E(f \mid u)$ is always a better estimate for than f in the sense of having a smaller variance, unless f is already a function of u only, in which case the two estimates f and $E(f \mid u)$ clearly coincide. This is simply the fact that the variance of the regression function of f on u is not greater than the variance of f. In the case of Gaussian variables, where the regression is linear, this fact has been noted by Doob [1, p. 231].[2] Our statement is embodied in

THEOREM 2: *If $\sigma^2 y$ is finite, so is $\sigma^2 E_x y$, and $\sigma^2 E_x y \leq \sigma^2 y$, with equality holding only if $E_x y = y$ with probability one.*

PROOF: Denote by m the common expected value of y and $E_x y$. Suppose for the moment that $\sigma^2 E_x y$ is finite. By the Schwarz inequality $E[yE_x y]$ is then finite. Then $\sigma^2 y = E(y-m)^2 = E[(y - E_x y) + (E_x y - m)]^2 = E(y - E_x y)^2 + \sigma^2 E_x y$, since $E[E_x y(E_x y - m)] = E[y(E_x y - m)]$ by Theorem 1. Thus $\sigma^2 y$ exceeds $\sigma^2 E_x y$ by $E(y - E_x y)^2$, which is positive unless $y = E_x y$, i.e. y is a function of x. Thus we obtain the usual decomposition: the variance of y is the variance of the regression of y on x plus the variance of y about the regression of y on x.

To show that $\sigma^2 E_x y$ is finite, we require the following

LEMMA (SCHWARZ INEQUALITY): *If $E(f^2)$ and $E(g^2)$ are finite, then, with probability one,*

$$E_x^2(fg) \leq E_x(f^2) E_x(g^2).$$

A proof can be constructed on the usual lines by considering the function $Q(x, \lambda) = E_x(f + \lambda g)^2$. There are, however, certain measure-theoretic difficulties

[2] For functions of finite variance it is possible to interpret conditional expectation as a projection in Hilbert space, when the statement becomes simply the Bessel inequality.

in handling simultaneously the conditional expectations of the family of chance variables $(f + \lambda g)^2$; instead we shall give a simple direct proof based on the ordinary Schwarz inequality for integrals.

We may suppose $f \geq 0$, $g \geq 0$ with probability one, since, from (2),

$$E_x^2(fg) \leq E_x^2(|f||g|)$$

with probability one. Unless the Lemma holds there are three positive numbers a, b, c with $a > bc$ for which the event

$$\{E_xfg > a^{\frac{1}{2}}, \quad E_x(f^2) < b, \quad E_x(g^2) < c\} = H$$

has positive probability. Then denoting by h the characteristic function of H and using the Schwarz inequality for integrals, we have

$$aP^2(H) \leq E^2[hE_x(fg)] = E^2(hfg) \leq E(hf^2)E(hg^2)$$
$$= E[hE_x(f^2)]E[hE_x(g^2)] \leq bcP^2(H),$$

which is impossible. This completes the proof of the Lemma.

The Lemma, with $f = y$, $g = 1$, yields $E_x^2(y) \leq E_x(y^2)$ with probability one, which implies the finiteness of $\sigma^2 E_x y$ and hence completes the proof of Theorem 2.

3. Unbiased sequential estimation. Consider a chance variable z whose distribution depends on a parameter θ. If we have an unbiased estimate $t(z)$ and a sufficient statistic $u(z)$ (not necessarily a single numerical chance variable) for θ, then, as mentioned in section 2, $v(u) = E(t \mid u)$ is an unbiased estimate for θ depending only on u.[3] We have shown that the variance of v is never greater than that of t, and we shall see that it is sometimes much smaller (see example II at the end of this section). The estimate obtained in this section for the parameter of a sequential process is of the v type; its importance lies in the fact that in many cases there is an unbiased estimate t (generally poor) which is a function of the first observation, and which will consequently be an unbiased estimate no matter what sequential test procedure is used.

Let $x_1, x_2, \cdots$ be a sequence of chance variables whose joint distribution is determined by an unknown point θ in a parameter space. A sequential sample (test) [9] is determined by specifying a sequence of mutually exclusive events $S_1, S_2, \cdots$, where S_i depends only on $x_1, \cdots, x_i$ and

$$\sum_{i=1}^{\infty} P(S_i) = 1 \quad \text{for all } \theta. \tag{4}$$

The event S_i is that sampling stops after the ith observation, and (4) ensures that sampling stops eventually. Thus if we define the chance variable $n = i$ when S_i occurs, n is the size of the sample.

[3] It was pointed out by the referee that, strictly speaking, u does not have to be sufficient; it is necessary only that $v(u)$ be independent of θ. The author is indebted to the referee for many valuable suggestions.

Denote by $u_1, u_2, \cdots$ any sequence of chance variables such that $u_i = u_i(x_1, \cdots, x_i)$ is a sufficient statistic for estimating θ from $x_1, \cdots, x_i$. There will of course be many such sequences $\{u_i\}$, but it often happens that there is one which arises in a natural way from the sequential process; if we are sampling from a binomial population, for instance, u_i = number of defectives in the first i observations is a sufficient statistic. We shall suppose that the sequential test satisfies the following condition

$$S_i = W_i C(S_1 + \cdots + S_{i-1}),^4 \tag{6}$$

where W_i is an event depending on u_i only. This condition means that when the ith observation is taken, the decision to stop at this point depends only on the ith sufficient statistic u_i. For the binomial example mentioned above, this means that the decision to stop after i observations depends only on the number of defectives observed at that stage, and not on the order in which they were observed. The Neyman criterion for u_i to be a sufficient statistic [7, 10, p. 135] shows that (6) is no restriction whatever for the sequential probability ratio test [9] since the ratio in terms of which the test is defined will be a function of u_i only.

Let $t_1, t_2, \cdots$ be any sequence of chance variables such that t_i is a function of $x_1, \cdots, x_i$; define $t = t_i$ when S_i occurs. If $E(t) = \theta$, t is said to be an unbiased estimate for θ (relative to the particular sequential test $\{S_i\}$). The theory of sequential sampling has been formulated primarily for testing hypotheses; a problem which arises naturally and often is the following: After a sequential sample has been obtained, is there an unbiased estimate for θ? Since a sample of constant size is a special case of a sequentially selected sample, we cannot hope to find unbiased estimates for arbitrary sequential samples unless such estimates exist for samples of every constant size. This is equivalent to the existence of a function $t(x_1)$ for which $E(t) = \theta$ for all θ. Our problem is to discover an unbiased estimate for θ which, when $n = i$, is a function of u_i alone. Such an estimate has been found by Girshick, Mosteller, and Savage [4] for sequential samples from a binomial population. It turns out that whenever there is any unbiased estimate at all for a particular sequential test, there is also one of the type described. Thus, if there is an unbiased estimate t for samples of fixed size N, there will be an unbiased estimate of the type described for every sequential test requiring at least N observations, since t is itself an unbiased estimate for such sequential tests.

Denote by t any unbiased estimate for θ relative to a particular sequential test $\{S_i\}$. Denote by w_i, h_i the characteristic functions of the events W_i, $C(S_1 + \cdots + S_i)$ respectively, and define $u = u_i$, $v = E(h_{i-1}t_i \mid u_i)/E(h_{i-1} \mid u_i)$ when $n = i$. To justify the definition of v we remark that the event $\{n = i, E(h_{i-1} \mid u_i) = 0\}$ has probability zero, since $qh_{i-1} \leq h_{i-1}$ with probability one, where q is the characteristic function of the event $\{E(h_{i-1} \mid u_i) > 0\}$, while

[4] For any event A, $C(A)$ denotes the event that A does not occur.

$$E(qh_{i-1}) = E[qE(h_{i-1} \mid u_i)] = E[E(h_{i-1} \mid u_i)] = E(h_{i-1}).$$

Since u_i is a sufficient statistic for θ with respect to $x_1, \cdots, x_i$, v is a function of u and n only, independent of θ. The main result of this section is

THEOREM 3. *v is an unbiased estimate for θ.*

PROOF: We shall show that $v = E(t \mid u, n)$. This not only shows that v is an unbiased estimate for θ, but also interprets v in a very simple way and, as mentioned above, implies that the variance of v does not exceed that of t. It must be verified that for every event D depending only on n and u, $E(dv) = E(dt)$, where d is the characteristic function of D. Now $D = \sum_{i=1}^{\infty} DS_i$, and $DS_i = D_iS_i$ where D_i is an event depending only on u_i. It is sufficient, then, to show $E(d_iw_ih_{i-1}v) = E(d_iw_ih_{i-1}t)$, where d_i is the characteristic function of D_i. Now

$$E(d_iw_ih_{i-1}v) = E[d_iw_ih_{i-1}E(h_{i-1}t_i \mid u_i)/E(h_{i-1} \mid u_i)],$$

using the definition of v. The function in brackets is h_{i-1} multiplied by a function of u_i; by Theorem 1 its expectation is unaltered if h_{i-1} is replaced by $E(h_{i-1} \mid u_i)$. Thus the right member of the last equality equals

$$E[d_iw_iE(h_{i-1}t_i \mid u_i)] = E(d_iw_ih_{i-1}t_i) = E(d_iw_ih_{i-1}t).$$

We conclude with two examples:

I. BINOMIAL AND POISSON DISTRIBUTIONS. Suppose $x_1, x_2, \cdots$ are independent with identical distributions, either binomial or Poisson, with parameter θ. Then $t = x_1$ ($= t_i$ for all i) is an unbiased estimate for θ, and it is well known that $u_i = x_1 + \cdots + x_i$ is a sufficient statistic for estimating θ from $x_1, \cdots, x_i$. For any sequential test satisfying (6) our unbiased estimate for θ will be

$$v = \frac{E(h_{i-1}x_1 \mid u_i = u)}{E(h_{i-1} \mid u_i = u)} = \frac{E(h_{i-1}x_1f)}{E(h_{i-1}f)}$$

when $n = i$, $u_i = u$, where f is the characteristic function of the event $u_i = u$. Then

$$v = \frac{\sum_{j=1}^{\infty} jk_j(u, i)}{\sum_{j=0}^{\infty} k_j(u, i)} \quad \text{for Poisson}$$

$$v = \frac{k_1(u, i)}{\sum_{j=0}^{1} k_j(u, i)} \quad \text{for binomial}$$

where $k_j(u, i)$ denotes the number of possible sequences $x_1, \cdots, x_i$ for which $n \geq i, x_1 + \cdots + x_i = u$, and $x_1 = j$. For the binomial case, this is the estimate found in [4].

II. SAMPLES OF CONSTANT SIZE. We consider the special case where a

sample of constant size N is selected, $x_1, \cdots, x_N$ are independent with identical distributions, and the density function for x_i has the form

$$p(x, \theta) = r(\theta)s(\theta)^{w(x)}q(x) \tag{7}$$

considered by Koopman [6][5]. Suppose further that there is an unbiased estimate $t(x_1)$ for θ. These conditions will be satisfied, for instance, if θ is the mean of a binomial, Poisson, or normal distribution, with $w(x) = t(x) = x$. Then $u_N = w(x_1) + \cdots + w(x_N)$ is a sufficient statistic. Our estimate v becomes simply $v = E[t(x_1) \mid u_N]$. Now $E[t(x_1) \mid u_N] = \cdots = E[t(x_N) \mid u_N]$, since u_N is a symmetric function of $x_1, \cdots, x_N$, which are independent with identical distributions. Consequently

$$v = E\left[\sum_{j=1}^{N} t(x_j)/N \mid u_N\right],$$

so that

$$\sigma^2(v) \leq \sigma^2\left(\sum_{j=1}^{N} t(x_j)/N\right) = \sigma^2 t(x_1)/N.$$

In the special case $w(x) = t(x) = x$, we have $v = \sum_{j=1}^{N} x_j/N$, i.e. our estimate is simply the mean of the N observations $x_1, \cdots, x_N$.

REFERENCES

[1] J. L. Doob, "The elementary Gaussian processes," *Annals of Math. Stat.*, Vol. 15 (1944), pp. 229–282.

[2] J. L. Doob, "The law of large numbers for continuous stochastic processes," *Duke Math. Jour.*, Vol. 6 (1940), pp. 290–306.

[3] J. L. Doob, "Stochastic processes with an integral-valued parameter," *Trans. Amer. Math. Soc.*, Vol. 44 (1938), pp. 87–150.

[4] M. A. Girshick, Frederick Mosteller, and L. J. Savage, "Unbiased estimates for certain binomial sampling problems, with applications," *Annals of Math. Stat.*, Vol. 17 (1946), pp. 13–23.

[5] A. Kolmogoroff, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Ergebnisse der Mathematik, Vol. 2 (1933).

[6] B. O. Koopman, "On distributions admitting a sufficient statistic," *Trans. Amer. Math. Soc.*, Vol. 39 (1936), pp. 399–409.

[7] J. Neyman, *Giornale dell Istituto Italiano degli Attuari*, Vol. 6 (1934), pp. 320–334.

[8] S. Saks, *Theory of the Integral*, Stechert, 1937.

[9] A. Wald, "Sequential tests of statistical hypotheses," *Annals of Math. Stat.*, Vol. 16 (1945), pp. 117–186.

[10] S. S. Wilks, *Mathematical Statistics*, Princeton Univ. Press, 1943.

[5] It has been shown by Koopman [6] that if there is a sufficient statistic satisfying certain regularity conditions, the density function for x must be of the form (7).