AN OPTIMUM PROPERTY OF REGULAR MAXIMUM LIKELIHOOD ESTIMATION

By V. P. Godambe

Science College, Nagpur, India

Let x be a point of some abstract space $\mathfrak{X}$, on which is defined a measure μ. Further, let $p(x, \theta)$ denote a probability density function with respect to μ, which, but for an unknown parameter, θ, is completely specified for all $x \varepsilon \mathfrak{X}$. It is taken for granted that $\theta \varepsilon \Omega$, a given index set. We make the following assumptions:

(a) Ω is an open interval of the real line.

(b) For almost all $x(\mu)$, $(\partial \log p(x, \theta)/\partial\theta)$ and $(\partial^2 \log p(x, \theta)/\partial\theta^2)$ exist for all $\theta \varepsilon \Omega$.

(c) $\int p(x, \theta)\, d\mu$ and $\int (\partial \log p(x, \theta)/\partial\theta) p(x, \theta)\, d\mu$ are differentiable under the integral sign.

(d) $E[(\partial \log p(x, \theta)/\partial\theta)^2 \mid \theta] > 0$ for all $\theta \varepsilon \Omega$.

To estimate θ, we construct a function, $g(x, \theta)$, on $\mathfrak{X} \times \Omega$ such that

(i) $E[g(x, \theta) \mid \theta] = 0$ for all $\theta \varepsilon \Omega$,

(ii) for almost all $x(\mu)$, $\partial g/\partial\theta$ exists for all $\theta \varepsilon \Omega$,

(iii) $\int g(x, \theta) p(x, \theta)\, d\mu$ is differentiable under the integral sign,

(iv) $[E(\partial g/\partial\theta \mid \theta)]^2 > 0$ for all $\theta \varepsilon \Omega$.

The condition (i) above is not much of a restriction, for, from any function, $f(x, \theta)$, such that, $E[f(x, \theta) \mid \theta] < \infty$, we can construct a $g(x, \theta)$ satisfying (i).

Now, given such a function g, the procedure of estimating θ, is as follows: If the observed value of $x = X$, an estimate of θ is given by $\theta_g(X)$, where the equation

$$g(X, \theta) = 0, \tag{1}$$

is satisfied by $\theta = \theta_g(X)$.

For this reason, any function g satisfying conditions (i)–(iv) above, will be called a *regular estimating function*. Let $\mathfrak{S}$ denote the class of all the regular estimating functions g. Now from $\mathfrak{S}$ we propose to select some g^*, which, in a certain sense is an *optimum estimating function.*

Definition: A g^* belonging to $\mathfrak{S}$ is said to be an optimum estimating function if

$$\frac{E(g^{*2} \mid \theta)}{\left[E\left(\frac{\partial g^*}{\partial\theta} \mid \theta\right)\right]^2} \leq \frac{E(g^2 \mid \theta)}{\left[E\left(\frac{\partial g}{\partial\theta} \mid \theta\right)\right]^2} \quad \text{for all } g \,\varepsilon\, \mathfrak{S} \text{ and } \theta \,\varepsilon\, \Omega. \tag{2}$$

Of course there is certain amount of arbitrariness in the above definition of "optimum". It is not however altogether unreasonable. We want the values of

Received July 28, 1959; revised May 17, 1960.

g to cluster around 0, as much as possible (i.e. $E(g^2 \mid \theta)$ should be as small as possible), and at the same time it is desirable that $E[g(x, \theta + \delta\theta) \mid \theta]$ should be as far away from 0 as possible. (This is conveniently translated as

$$[E(\partial g/\partial\theta \mid \theta)]^2$$

should be as large as possible). This justifies the definition (2). The following Theorem establishes that $g^* = (\partial \log p)/\partial\theta$[1] satisfies (2) above. Hence follows the optimality of the maximum likelihood estimation. *It is important to note that the validity of the Theorem below is independent of the discussion of "optimality" in this paragraph.*

THEOREM: *For all $g \in \mathfrak{S}$,*

$$\text{(A)} \qquad \frac{E(g^2 \mid \theta)}{\left[E\,\frac{\partial g}{\partial\theta} \mid \theta\right]^2} \geqq \frac{1}{E\left[\left(\frac{\partial \log p}{\partial\theta}\right)^2 \mid \theta\right]},$$

the equality being attained for

$$\text{(B)} \qquad g^* = \frac{\partial \log p}{\partial\theta}.$$

Then, obviously,

$$\text{(C)} \qquad \frac{E(g^{*2} \mid \theta)}{\left[E\left(\frac{\partial g^*}{\partial\theta} \mid \theta\right)\right]^2} \leqq \frac{E(g^2 \mid \theta)}{\left[E\left(\frac{\partial g}{\partial g} \mid \theta\right)\right]^2}$$

for all $g \in \mathfrak{S}$.

PROOF: We have from (i), for $\theta \in \Omega$,

$$\text{(3)} \qquad \int g(x, \theta)\, p(x, \theta)\, d\mu = 0.$$

Differentiating (3) under the integral sign (this is permissible because of (a), (b), (ii) and (iii)), we have

$$\text{(4)} \qquad \int \frac{\partial g}{\partial\theta}\, p\, d\mu + \int g\, \frac{\partial \log p}{\partial\theta}\, p\, d\mu = 0.$$

Further, because of (c),

$$\text{(5)} \qquad E\left(\frac{\partial \log p}{\partial\theta} \mid \theta\right) = 0.$$

Thus second integral in (4) is the covariance of g and $\partial \log p/\partial\theta$. It then follows from (d), (iv) and the Cauchy-Schwarz inequality, that, for $\theta \in \Omega$,

$$\text{(6)} \qquad \frac{E(g^2 \mid \theta)}{\left[E\left(\frac{\partial g}{\partial\theta} \mid \theta\right)\right]^2} \geqq \frac{1}{E\left[\left(\frac{\partial \log p}{\partial\theta}\right)^2 \mid \theta\right]}.$$

[1] It follows from (a)-(d) that $\partial \log p/\partial\theta \in \mathfrak{S}$.

This proves part (A) of the Theorem. Of course, instead of (iv), when,

$$E(\partial g/\partial\theta \mid \theta) = 0,$$

(A) is true trivially. Further, if

$$g^* = \frac{\partial \log p}{\partial\theta}, \tag{7}$$

we have from (a), (b) and (c) that

$$-E\left[\frac{\partial^2 \log p}{\partial\theta^2} \mid \theta\right] = E\left[\left(\frac{\partial \log p}{\partial\theta}\right)^2 \mid \theta\right], \tag{8}$$

and hence

$$\frac{E(g^{*2} \mid \theta)}{\left[E\left(\frac{\partial g^*}{\partial\theta} \mid \theta\right)\right]^2} = \frac{1}{E\left[\left(\frac{\partial \log p}{\partial\theta}\right)^2 \mid \theta\right]}. \tag{9}$$

This proves part (B) of the Theorem.

It is interesting to note that the above theorem generalizes the Cramér-Rao inequality [1], [2]. For let

$$g_1(x, \theta) = f(x) - e_f(\theta), \tag{10}$$

where f is a function on $\mathfrak{X}$, and

$$e_f(\theta) = E(f \mid \theta). \tag{11}$$

We assume $g_1 \varepsilon \mathfrak{S}$. Now if the above Theorem (A) is applied to g_1, we get,

$$\frac{\text{Var } (f \mid \theta)}{[\partial e_f(\theta)/\partial\theta]^2} \geqq \frac{1}{E[(\partial \log p/\partial\theta)^2 \mid \theta]}. \tag{12}$$

This is the Cramér-Rao inequality. Part (B) of the above Theorem is valid under general regularity conditions, while the sign of equality in (12) holds, when and only when, in addition to regularity conditions,

$$\frac{\partial \log p(x, \theta)}{\partial\theta} = \lambda(\theta)(f(x) - e_f(\theta)), \tag{13}$$

where $\lambda(\theta)$ is some function of θ alone (Cramér [1], p. 480; Fend [3]). Obviously, when the sign of equality holds in (12), g_1 in (10) is given by,

$$\lambda(\theta)g_1(x, 6) = \frac{\partial \log p(x, \theta)}{\partial\theta}. \tag{14}$$

The author acknowledges with pleasure that G. A. Barnard communicated to the Royal Statistical Society, London, a result similar to the preceding Theorem, independently, and at nearly the same time when this paper was written. Barnard's result is explained in [4], p. 145.

REFERENCES

[1] HARALD CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, N. J., 1954.

[2] C. R. RAO, "Information and accuracy attainable in estimation of statistical parameters," *Bull. Cal. Math. Soc.*, Vol. 37 (1945), pp. 81–91.

[3] A. V. FEND, "On the attainment of Cramér-Rao and Bhattacharyya bounds for the variance of an estimate," *Ann. Math. Stat.*, Vol. 30 (1959), pp. 381–388.

[4] J. DURBIN, "Estimation of parameters in time series regression models," *J. Roy. Stat. Soc., Ser. B*, Vol. 22 (1960), pp. 139–153.