CERTAIN INEQUALITIES IN INFORMATION THEORY AND THE CRAMÉR-RAO INEQUALITY

By S. Kullback

The George Washington University

1. Summary and Introduction. The Cramér-Rao inequality provides, under certain regularity conditions, a lower bound for the variance of an estimator [7], [15]. Various generalizations, extensions and improvements in the bound have been made, by Barankin [1], [2], Bhattacharyya [3], Chapman and Robbins [5], Fraser and Guttman [11], Kiefer [12], and Wolfowitz [16], among others.

Further considerations of certain inequality properties of a measure of information, discussed by Kullback and Leibler [14], yields a greater lower bound for the information measure (formula (4.11)), and leads to a result which may be considered a generalization of the Cramér-Rao inequality, the latter following as a special case. The results are used to define discrimination efficiency and estimation efficiency at a point in parameter space.

2. The first inequality. We use the notation and terminology of [14]. Consider the measurable transformations T_N of the probability spaces $(\mathfrak{X}, \mathfrak{B}, \mu_i)$ onto the probability spaces $(\mathcal{Y}, \mathfrak{Z}, \nu_i^{(N)})$, and suppose for $G \in \mathfrak{Z}$ that $\nu_i^{(N)}(G) = \mu_i(T_N^{-1}G)$ for $i = 1$ or 2.

THEOREM 2.1. *Let the T_N be such that*

$$\lim_{N\to\infty} \nu_i^{(N)}(G) = \nu_i(G), \qquad i = 1, 2; \qquad G \in \mathfrak{Z}, \tag{2.1}$$

Then

$$I(1:2; x) \geqq \liminf_{N\to\infty} I'_N(1:2; y) \geqq I'(1:2; y); \tag{2.2}$$

$$J(1, 2; x) \geqq \liminf_{N\to\infty} J'_N(1, 2; y) \geqq J'(1, 2; y). \tag{2.2'}$$

PROOF. We first derive a result which is similar to a lemma used by Doob [8]. Using Lemma 3.2 of [14], we have

$$I'_N(1:2; y) \geqq \sum \nu_1^{(N)}(G_j) \log \frac{\nu_1^{(N)}(G_j)}{\nu_2^{(N)}(G_j)}, \tag{2.3}$$

where the sum is taken over any set of pairwise disjunct G_j such that $\bigcup_j G_j = \mathcal{Y}$. Accordingly,

$$\liminf_{N\to\infty} I'_N(1:2; y) \geqq \sum \nu_1(G_j) \log \frac{\nu_1(G_j)}{\nu_2(G_j)}, \tag{2.4}$$

Received 10/12/53, revised 4/20/54.

and therefore

$$\liminf_{N\to\infty} I'_N(1:2;y) \geqq I'(1:2;y), \tag{2.5}$$

since the right member of (2.5) is the l.u.b. of the right member of (2.4). In conjunction with Theorem 4.1 and paragraph 5 of [14] and (2.5), the inequalities (2.2) and (2.2′) follow. These are used herein only in Section 3.

3. An example. Consider N independent observations from the binomial distributions $B(p_i, q_i)$, for $i = 1$, or 2, which as $N \to \infty$ approach as limits the Poisson exponential distributions with means $m_i = Np_i$, for $i = 1$ or 2. It may be verified readily that

$$I'_N(1:2;y) = \sum \frac{N!}{y!(N-y)!} p_1^y q_1^{N-y} \log \frac{p_1^y q_1^{N-y}}{p_2^y q_2^{N-y}} = N\left(p_1 \log \frac{p_1}{p_2} + q_1 \log \frac{q_1}{q_2}\right), \tag{3.1}$$

$$I'(1:2;y) = \sum \frac{m_1^y e^{-m_1}}{y!} \log \frac{m_1^y e^{-m_1}}{m_2^y e^{-m_2}} = (m_2 - m_1) + m_1 \log \frac{m_1}{m_2}. \tag{3.2}$$

Using the well known inequality $x_1 \log (x_1/x_2) \geqq x_1 - x_2$, and $m_i = Np_i$ for $i = 1$ or 2, it is found that

$$Np_1 \log \frac{p_1}{p_2} + Nq_1 \log \frac{q_1}{q_2} = m_1 \log \frac{m_1}{m_2} + N\left(1 - \frac{m_1}{N}\right) \log \frac{1 - m_1/N}{1 - m_2/N} \geqq m_1 \log \frac{m_1}{m_2} + N\left(\frac{m_2}{N} - \frac{m_1}{N}\right) = m_1 \log \frac{m_1}{m_2} + (m_2 - m_1), \tag{3.3}$$

or

$$\liminf_{N\to\infty} I'_N(1:2;y) \geqq I'(1:2;y). \tag{3.4}$$

As a matter of fact, for this particular case, as may be readily seen from the first two members of (3.3),

$$\lim_{N\to\infty} I'_N(1:2;y) = I'(1:2;y). \tag{3.5}$$

4. The second inequality. Suppose $g_1(y)$, $g_2(y)$, and $g^*(y)$ are densities satisfying the conditions of paragraph 4 of [14]. Then using Lemma 3.1 of [14],

$$\int g_1(y) \log \frac{g_1(y)}{g_2(y)}\, d\gamma(y) + \int g_1(y) \log \frac{g_2(y)}{g^*(y)}\, d\gamma(y) = \int g_1(y) \log \frac{g_1(y)}{g^*(y)}\, d\gamma(y) \geqq 0, \tag{4.1}$$

or

$$\int g_1(y) \log \frac{g_1(y)}{g_2(y)}\, d\gamma(y) \geqq \int g_1(y) \log \frac{g^*(y)}{g_2(y)}\, d\gamma(y) \tag{4.2}$$

In particular, let us take, for real t,

$$g^*(y) = \frac{e^{ty}\, g_2(y)}{M_2(t)}, \qquad M_2(t) = \int e^{ty} g_2(y)\, d\gamma(y), \tag{4.3}$$

so that (4.2) becomes

$$I'(1:2; y) \geqq at - \log M_2(t), \qquad a = E_1(y), \tag{4.4}$$

with equality if and only if

$$g_1(y) = g^*(y) = \frac{e^{ty} g_2(y)}{M_2(t)} \quad [\gamma(y)]. \tag{4.5}$$

To investigate further the right member of (4.4) we will use the notation, and, in particular, the results of paragraphs 4 and 6 of Chernoff [6]. Clearly

$$I'(1:2; y) \geqq \sup_t (at - \log M_2(t)) = -\log m_2(a), \tag{4.6}$$

where $m_2(a) = \inf_t e^{-at} M_2(t)$. Note that for the value of t satisfying $a = N_2(t(a)) / M_2(t(a))$, we have

$$-\log m_2(a) = \int g^*(y) \log \frac{g^*(y)}{g_2(y)}\, d\gamma(y) \geqq 0.$$

From this, or the results of Lemma 7 of Chernoff [6], it follows that $-\log m_2(a)$ is a convex function of a. Limiting ourselves to statistics y for which $E_2(y)$ and $\mathrm{Var}_2(y)$ are finite, the results of Chernoff [6] may also be derived for the case $a \geqq E_2(y)$.

We can write

$$\log m_2(a) = \log m_2(E_2(y)) + (a - E_2(y)) \frac{d}{da} \log m_2(a) \Big|_{a=E_2(y)} + \frac{(a - E_2(y))^2}{2!} \left(-\frac{dt(b)}{db}\right), \tag{4.7}$$

where b is between a and $E_2(y)$. But as Chernoff [6] has shown,

$$\log m_2(E_2(y)) = 0, \qquad \frac{d}{da} \log m_2(a) \Big|_{a=E_2(y)} = 0, \qquad \frac{dt(a)}{da} \Big|_{a=E_2(y)} = \frac{1}{\mathrm{Var}_2(y)}, \qquad \frac{dt(b)}{db} = \frac{1}{\mathrm{Var}_*(y)}, \tag{4.8}$$

where $\mathrm{Var}_*(y)$ is the variance of y for the distribution defined by

$$g_*(y) = e^{t(b)y} g_2(y) / M_2(t(b)). \tag{4.9}$$

From (4.6), (4.7), and (4.8) it follows that

$$I'(1:2; y) \geqq (E_1(y) - E_2(y))^2 / 2\, \mathrm{Var}_*(y), \tag{4.10}$$

where [13] the right side is the value of $I(1:2)$ for two normal distributions with common variance $\mathrm{Var}_*(y)$ and means $E_1(y)$ and $E_2(y)$.

We take y as the linear function $y = c_1y_1 + c_2y_2 + \cdots + c_ky_k$, where the random variables $y_1, y_2, \cdots, y_k$ are such that the requirements already imposed on y are satisfied and $\text{Var}_*(y) = \sum_{i,j=1}^{k} c_ic_j \text{ cov}_*(y_i, y_j)$. Then, as is known [13], the l.u.b. of the right member of (4.10) for possible values of the c's is given by the quadratic form $\frac{1}{2}\delta'\sigma_*^{-1}\delta$, where δ is the one column matrix of the differences $\delta_j = E_1(y_j) - E_2(y_j)$ for $j = 1, 2, \cdots; k$, and δ' is the transpose of δ while σ_* is the matrix of variances and covariance of the y_j for $j = 1, 2, \cdots, k$ in the distribution defined by (4.9).

We thus have the second inequality

$$I(1:2; x) \geqq I'(1:2; y) \geqq \tfrac{1}{2}\delta'\sigma_*^{-1}\delta. \tag{4.11}$$

For the binomial distribution, this yields

$$p_1 \log \frac{p_1}{p_2} + q_1 \log \frac{q_1}{q_2} \geqq \frac{(p_1 - p_2)^2}{2p_* q_*}, \quad p_* = \frac{p_2 e^t}{p_2 e^t + q_2}, \quad q_* = \frac{q_2}{p_2 e^t + q_2}, \tag{4.12}$$

for some value of t between 0 (when $p_* = p_2$) and $\log p_1q_2/q_1p_2$ (when $p_* = p_1$). Note that $p_* = b$, and that from our derivation b is between p_1 and p_2.

5. The Cramér-Rao inequality. For the parametric case, where the populations are neighboring points θ and $\theta + \Delta\theta$ in the k-dimensional parameter space and the y_j for $j = 1, 2, \cdots, k$ are unbiased estimators of the parameters, (4.11) yields, under suitable regularity conditions [14],

$$(\Delta\theta)'G(\Delta\theta) \geqq (\Delta\theta)'H(\Delta\theta) \geqq (\Delta\theta)'\sigma^{-1}(\Delta\theta), \tag{5.1}$$

where $\Delta\theta$ is the one column matrix of the $\Delta\theta_j$ for $j = 1, 2, \cdots, k$ and $(\Delta\theta)'$ is its transpose, while G and H are respectively the matrices $(g_{\alpha\beta})$ and $(h_{\alpha\beta})$, for $\alpha, \beta = 1, 2, \cdots, k$, where

$$g_{\alpha\beta} = \int f(x) \left(\frac{\partial}{\partial\theta_\alpha} \log f(x)\right)\left(\frac{\partial}{\partial\theta_\beta} \log f(x)\right) d\lambda(x),$$

$$h_{\alpha\beta} = \int g(y) \left(\frac{\partial}{\partial\theta_\alpha} \log g(y)\right)\left(\frac{\partial}{\partial\theta_\beta} \log g(y)\right) d\gamma(y),$$

and σ is the matrix of variances and covariances of the estimators.

It should be observed that the discussion in Sections 4 and 5 holds whether we are dealing with a fixed sample size or sequential procedure. For the latter case, ([16] p. 216) let $\mathfrak{X}$ of the probability spaces $(\mathfrak{X}, \mathfrak{S}, \mu_i)$, be the space of all possible infinite sequences (x) of observations $x_1, x_2, \cdots$. Let there be given an infinite sequence of Borel measurable functions $\phi_1(x_1), \phi_2(x_1, x_2), \cdots, \phi_j(x_1, x_2, \cdots, x_j)$ $\cdots$, defined for all observable sequences in $\mathfrak{X}$ such that each takes only the values zero and one. We further assume that at least one of the functions $\phi_1(x_1)$, $\phi_2(x_1, x_2), \cdots$ takes the value one $[\lambda(x)]$, and let n be the smallest integer for which this occurs. Thus $n(x)$ is a chance variable.

The sequential process is then defined as follows. Take an observation and find $\phi_1(x_1)$. If it is unity, the sampling process stops; otherwise sampling con-

tinues. If a second observation is taken and the value of $\phi_2(x_1, x_2)$ is unity, the process stops; otherwise it continues, and so on. In general, after taking j observations,

$\phi_i(x_1, x_2, \cdots, x_i) = 0$ for $i = 1, 2, \cdots, j - 1$. If

$\phi_j(x_1, x_2, \cdots, x_j) = 1$, sampling stops; otherwise it is continued.

If R_j denotes the set of all points $(x_1, x_2, \cdots)$ for which the process stops with the jth observation, then $\mathfrak{X} = \bigcup_j R_j$. The variable y is taken as a function of the observations $x_1, x_2, \cdots, x_n$ (those obtained prior to the termination of the process of drawing observations).

Thus the results in (4.11) and (5.1) hold for fixed sample size or sequential procedures.

6. Quadratic forms. Certain useful results with respect to quadratic forms, which are essentially corollaries of known theorems, are needed for the subsequent discussion.

LEMMA 6.1. *If both $X'AX$ and $X'CX$ are positive definite quadratic forms (matric notation) such that $X'AX \geqq X'CX$, then*

(a) *the roots of $|A - \lambda C| = 0$ are real and $\geqq 1$;*

(b) $|A| \geqq |C|$;

(c) *any principal minor of A is not less than the corresponding principal minor of C, (determinant or quadratic form);*

(d) $Y'C^{-1}Y \geqq Y'A^{-1}Y$;

(e) *any principal minor of C^{-1} is not less than the corresponding principal minor of A^{-1} (determinant or quadratic form).*

PROOF. Results (a), (b), and (c) are immediate corollaries of theorems 44 and 48 in Ferrar [10]. Since $A^{-1} = C^{-1}CA^{-1}$ and $C^{-1} = C^{-1}AA^{-1}$, there exists a non-singular matrix B such that (Bôcher [4], p. 301) $C^{-1} = B'AB$ and $A^{-1} = B'CB$. Thus applying the transformation $X = BY$ gives

$$X'AX = Y'B'ABY = Y'C^{-1}Y, \qquad X'CX = Y'B'CBY = Y'A^{-1}Y,$$

and (d) and (e) then follow.

7. Efficiency. With respect to the estimators y_j of Section 5, the *discrimination efficiency* at a point P in the k-dimensional parameter space (P.S.) is defined by

$$\lambda = \frac{(d\theta)'H(d\theta)}{(d\theta)'G(d\theta)}. \tag{7.1}$$

We take $(d\theta)'G(d\theta)$ as the basis of the metric of (P.S.). The $g_{\alpha\beta}$ for $\alpha, \beta = 1, 2, \cdots, k$, are the components of a covariant tensor of the second order which is called the *fundamental tensor* of the metric (Eisenhart [9]). Since $(d\theta)'H(d\theta) \leqq (d\theta)'G(d\theta)$ and both forms are positive definite, the roots of

$$|H - \lambda G| = 0, \tag{7.2}$$

are real, positive, and all $\leqq 1$. Accordingly there exists a real transformation of the θ's such that at a point P in (P.S.) the forms in (7.1) may be written as

$$\lambda = \frac{\lambda_1 \, d\psi_1^2 + \cdots + \lambda_k \, d\psi_k^2}{d\psi_1^2 + \cdots + d\psi_k^2} \tag{7.3}$$

and $\lambda_1, \lambda_2, \cdots, \lambda_k$, are the roots of (7.2) (Eisenhart [9] p. 108). If we write

$$\cos^2 \alpha_i = \frac{d\psi_i^2}{d\psi_1^2 + \cdots + d\psi_k^2}, \qquad i = 1, 2, \cdots, k, \tag{7.4}$$

then (7.3) may be written as

$$\lambda = \lambda_1 \cos^2 \alpha_1 + \lambda_2 \cos^2 \alpha_2 + \cdots + \lambda_k \cos^2 \alpha_k . \tag{7.5}$$

The directions at the point P determined by $\cos \alpha_1 = 1$, $\cos \alpha_2 = 1, \cdots$, are known as the *principal directions* determined by the tensor $h_{\alpha\beta}$ (Eisenhart [9], p. 110). Furthermore, at the point P the finite maxima and minima of λ defined by (7.1) are given by the principal directions at the point and are indeed the roots of (7.2). Since $(d\theta)'G(d\theta)$ is positive definite, λ is finite for all directions (Eisenhart [9], par. 33).

As the *estimation efficiency* of the estimators $y_1, y_2, \cdots, y_k$, we take the product of the discrimination efficiencies for the principal directions at the point P, that is,

$$\text{Eff} = \lambda_1 \lambda_2 \cdots \lambda_k = |H| \, / \, |G| \leqq 1, \tag{7.6}$$

which is invariant for all nonsingular transformations of the parameters, with equality holding if and only if the estimators are sufficient [14].

8. Asymptotic efficiency. Suppose we have n independent observations from an l-variate population with k parameters. It is also of interest to consider, instead of (7.1), the *asymptotic discrimination efficiency* at a point P in (P.S.) defined by

$$\lambda = \frac{(d\theta)'\sigma^{-1}(d\theta)}{n(d\theta)' \, G(d\theta)}, \qquad n \text{ large}, \tag{8.1}$$

where the elements of the matrix G are computed for a single observation from the l-variate population. Since $(d\theta)'\sigma^{-1}(d\theta) \leqq n(d\theta)'G(d\theta)$ and both forms are positive definite, the roots of

$$|\sigma^{-1} - \lambda n G| = 0 \tag{8.2}$$

are real, positive and $\leqq 1$. As in Section 7, the roots of (8.2) are the finite maxima and minima of (8.1) at a point P in (P.S.) and are given by the principal directions determined by the tensor $\sigma^{\alpha\beta}$ at the point.

As the *asymptotic estimation efficiency* of the unbiased estimators $y_1, y_2, \cdots, y_k$

(cf. Cramér [7], pp. 489, 494) we take the product of the asymptotic discrimination efficiencies for the principal directions at the point P, that is,

$$\text{Asymp Eff} = \lambda_1\lambda_2 \cdots \lambda_k = |\sigma^{-1}| \,/\, |nG| \leqq 1, \qquad n \text{ large},$$

the equality holding for all n if the estimators are sufficient and (4.5) is satisfied. If $|\sigma|\ |G| \to n^{-k}$, then the asymptotic efficiency approaches unity and $\lambda_i \to 1$ for $i = 1, 2, \cdots, k$.

REFERENCES

[1] E. W. Barankin, "Locally best unbiased estimates," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 477–501.

[2] E. W. Barankin, "Concerning some inequalities in the theory of statistical estimation," *Skand. Aktuarietids*, Vol. 34 (1951), pp. 35–40.

[3] A. Bhattacharyya, "On some analogues of the amount of information and their use in statistical estimation," *Sankhyā*, Vol. 8 (1946), pp. 1–14, (1947), pp. 201–218, (1948) pp. 315–328.

[4] M. Bôcher, *Higher Algebra*, MacMillan, 1924.

[5] D. G. Chapman and H. Robbins, "Minimum variance estimation without regularity assumptions," *Ann. Math. Stat.*, Vol. 22 (1951), pp. 581–586.

[6] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 493–507.

[7] H. Cramér, *Mathematical Methods of Statistics*, Princeton Univ. Press, 1946.

[8] J. L. Doob, "Statistical estimation," *Trans. Amer. Math. Soc.*, Vol. 39 (1936), pp. 410–421.

[9] L. P. Eisenhart, *Riemannian Geometry*, Princeton Univ. Press, 1926.

[10] W. L. Ferrar, *Algebra*, Oxford Univ. Press, 1941.

[11] D. A. S. Fraser and I. Guttman, "Bhattacharyya bounds without regularity assumptions," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 629–632.

[12] J. Kiefer, "On minimum variance estimators," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 627–629.

[13] S. Kullback, "An application of information theory to multivariate analysis," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 88–102.

[14] S. Kullback and R. A. Leibler, "Information and sufficiency," *Ann. Math. Stat.*, Vol. 22 (1951), pp. 79–86.

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

[16] J. Wolfowitz, "The efficiency of sequential estimates and Wald's equation for sequential processes," *Ann. Math. Stat.*, Vol. 18 (1947), pp. 215–230.