*Sankhyā : The Indian Journal of Statistics*
1985, Volume 47, Series A, Pt. 3, pp. 310-325.

# EFFICIENCY OF ESTIMATES - PART I

*By* J. K. GHOSH

*Indian Statistical Institute*

***SUMMARY.*** The paper provides a partly expository, partly historical review of various results and techniques in the area of asymptotic efficiency. The following topics are discussed. Fisher's and LeCam's inequality, Cramer-Rao inequality, Haick-Inagaki convolution theorem, local asymptotic minimaxity, asymptotic normality of posteriors, approximate Bayes character of maximum likelihood estimates, and limiting experiments.

## 1. INTRODUCTION

The modern study of efficiency in estimation theory or what would perhaps be called to-day first order asymptotic efficiency, began with the seminal paper of LeCam (1953) and reached a more or less final form in Hájek (1972), the only significant work after that being Millar's (1983) elucidation and extension of some basic results of LeCam (1972). At about the same time, various groups of workers—Akahira and Takeuchi (1976), Ghosh and Subramanyam (1974), Efron (1975), Pfanzagl (1973, 1975)—began a systematic study of higher order efficiency to discriminate between efficient estimates. It was but natural that much attention would focus on the pioneering work of Rao (1961, 1962, 1963). In the last ten years or so, this theory too has reached a fairly definitive form, the most general results being those of Bickel, Chibishov and Van Zwet (1981) and Bickel, Goetze and Van Zwet (1984).

In at least two ways, then, this seems to be the right time to survey and make clear to the general statistical world what has been achieved in these areas. I shall attempt such a survey in a two part article, addressed to the general reader of our subject rather than the specialist. In the present paper, which is Part I of the proposed survey, I shall be mainly concerned with the classical work on (first order) efficiency, trying to present the main results with a sketch of proof in most cases. Though no thorough review of the subject has appeared so far, the excellent introduction in Hájek (1972), Chapter 5 of Roussas (1972), parts of the first thirty pages or so of Pfanzagl (1980), Ibragimov and Khasminski (1981) and Lehmann (1983, Chapter 6), contain partial reviews.

*AMS* (1980) *subject classification :* 62F105.

*Key words and phrases :* Cramer-Rao inequality, asymptotic efficiency, local asymptotic minimaxity, limiting experiment.

Much of the work on efficiency was prompted by an attempt to understand how well the maximum likelihood estimate performs as $n$ goes to infinity. So we begin the next section with a brief account of the maximum likelihood estimate.

## 2. Assumptions, notations and the maximum likelihood estimate

One considers a sequence of i.i.d. random variables $\{X_i\}$ of which $n$ are observed. The common density is $f_\theta$, for convenience we assume the parameter space $\Theta$ is the whole real line $R$. It is assumed, unless otherwise stated, that the usual regularity conditions[1] for existence of a consistent solution of the likelihood equation, vide Rao (1965), or Serfling (1980), hold. The usual proofs of existence have a small gap but it is well-known that the gap can be removed, vide Ghosh (1983). In most of the literature on efficiency it is this consistent solution, denoted by $\hat{\theta}_n$, which is called the maximum likelihood estimate; this will be the convention here also. As everyone knows $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal, $AN(0, I^{-1}(\theta))$ where $I(\theta)$ is the Fisher information $E_\theta\left\{\frac{f'_\theta}{f_\theta}\right\}^2$, assumed to be positive and finite. This property is often referred to as the efficiency of the maximum likelihood estimates. We shall examine in the next few sections how far this terminology is justified.

The symbol $N(\mu, \sigma^2)$ will indicate both a normally distributed random variable with mean $\mu$ and variance $\sigma^2$ as well as its distribution function.

## 3. Inequalities of Fisher, LeCam and Cramer-Rao

Let $T_n = T_n(X_1, \ldots, X_n)$ be an estimate of $\theta$. A natural question to ask is whether

$$\sqrt{n}(T_n - \theta) \text{ is } A.N.(0, v(\theta)) \ \forall\ \theta \qquad \ldots (1)$$

entails

$$v(\theta) \geqslant I^{-1}(\theta) \ \forall\ \theta. \qquad \ldots (2)$$

If the answer were yes, one would be justified in calling $\hat{\theta}_n$ efficient in the class of all estimates satisfying (1). Unfortunately, as everyone knows today the following famous example due to Hodges, vide LeCam (1953) shows the answer is no under the conditions assumed in Section 2: Choose $\theta_0$ and let

$$T_n = \begin{cases} \hat{\theta}_n & \text{if } |\hat{\theta}_n - \theta_0| > g(n) \\ \theta_0 & \text{if } |\hat{\theta}_n - \theta_0| \leqslant g(n) \end{cases}$$

where $g(n) \to 0$ but $n^{1/2}g(n) \to \infty$ (e.g. $g(n) = n^{-1/4}$). Then (1) holds with $v(\theta) = I^{-1}(\theta)$ if $\theta \neq \theta_0$ and $v(\theta_0) = 0$. So not only is (2) violated but it is

[1] These imply among other things that $\theta_0 + h/\sqrt{n}$ is contiguous to $\theta_0$, vide Hájek (1972).

clear that subject only to (1) the lower bound to $v(\theta)$ at any $\theta_0$ must be zero. On the other hand LeCam (1953) proved the remarkable result that (1) does entail $v(\theta) \geqslant I^{-1}(\theta)$ almost everywhere (with respect to the Lebesgue measure). I shall refer to this as LeCam's inequality. If the stronger result (2) holds, $T_n$ will be said to satisfy Fisher's inequality, in honour of Fisher whose pioneering arguments were the basis of expecting (2).

The fifties and sixties were full of different attempts to get round the Hodges example. This was done either by putting more conditions on the estimates—Fisher consistency, Kallianpur and Rao (1955) or asymptotic normality uniformly on compact sets, Rao (1963) etc.—or by permitting all estimates (not necessarily with an asymptotic distribution) but evaluating them in a different way. The first approach will be reviewed in Section 4 and the second in Section 5. It was intuitively clear that the Hodges example was pathological and both the sufficient conditions for (2) given in Section 4 and local minimax criterion of Section 5 make clear mathematically what are the desirable things that the Hodges estimate lacks.

Most of us have been brought up on the tradition that Fisher's inequality (2) is a plausible even though not rigorous consequence of the Cramer-Rao (which should perhaps be called Cramer-Frechet-Rao since Frechet was an independent discoverer). Historically however it wasn't so, the Cramer-Rao being a relative late comer as a small sample analogue of (2). Indeed Fisher had a proof, which is satisfactory even by modern standards, for the estimates which are now called $M$-estimates. In fact both the inequality (2) and the essence of its proof for $M$-estimates can be traced even further back to Edgeworth; Pratt (1976) provides a well-documented study of the pioneering efforts of these giants.

Even retrospectively the Cramer-Rao doesn't seem well adapted to prove (2). Let us see why. To derive (2) rigorously from

$$\text{var}(\sqrt{m}(T_m - \theta)) \geqslant (1 + b'_m(\theta))^2 / I(\theta) \qquad \ldots (3)$$

one has to assume two things : (A) the limit of variance of $\sqrt{n}(T_n - \theta)$ is the same as the variance $v(\theta)$ of the asymptotic distribution and (B) $b'_n \to 0$. Both (A) and (B), particularly (B), are unpleasant assumptions with nothing but mathematical convenience in their favour and both would be rather hard to check for the common estimates. Walker (1963) provides a sufficient condition for (A) which involves, roughly speaking, the finiteness of the $(2+\delta)$-th moment of $\sqrt{n}(T_n - \theta)$ for all $n$; the proof seems to require more—that the integrals be bounded by a fixed constant for all $n$. This stronger

condition is of course well-known to be sufficient for (A). Incidentally, as pointed out by Walker, for $F_\theta = N(\theta, 1)$, the Hodges example satisfies (A) but violates (B), confirming one's suspicion that (B) is the more serious restriction. For all these reasons the sufficient conditions to be presented in the next section turn out to be of an altogether different kind and their justification will not invoke (3).

In the other direction, one can, following Pratt (1976), base a proof of the Cramer-Rao on Fisher's inequality. Without loss of generality we consider the unbiased case, i.e., $b_m = 0$. Suppose, then, $f_\theta$ satisfies regularity conditions as explained in Section 2 and $T_m$ is unbiased, $\text{var}_\theta(T_m) < \infty \; \forall \; \theta$ and $E_\theta |T|^3$ is bounded on compacts. Under these condtions let me prove (3). Let $n = km$ and $T_n = k^{-1}\{T_m(X_1, \ldots, X_m) + \ldots + T_m(X_{n-m+1}, \ldots, X_n)\}$. Then by the Borry-Esseen theorem, $\sqrt{n}(T_n - \theta)$ is $A.N.(0, m \text{ var}_\theta(T_m))$ uniformly on compacts. By Rao's theorem, vide Section 4, Fisher's inequality holds for $v(\theta) = m \text{ var}_\theta(T_m)$ which is just (3).

There is an interesting heuristic proof of (2) which is, by tradition attributed to Fisher (but Pratt (1976) apparently doesn't think so). It goes as follows. Let $I_{T_n}(\theta)$ be the Fisher information contained in the marginal distribution of $T_n$. Fisher had proved as everyone knows, $I_{T_n}(\theta) \leqslant nI(\theta)$, $nI$ being the total information in the sample. Observe that if $Y$ is $N(\theta, \sigma^2)$ then $I_Y(\theta) = \dfrac{1}{\sigma^2}$ and so (1) suitably strengthened should imply $I_{T_n}$ is approximately $\dfrac{n}{v(\theta)}$. Hence (2) follows from $I_{T_n} \leqslant nI$. This argument is implicit in Rao (1961) and forms the basis of his approach to efficiency and second order efficiency in that paper.

## 4. Sufficient conditions for Fisher's inequality

The simpiest sufficient condition for (2)—multinomial probability and smooth Fisher consistent estimates—could have been due to Fisher but is actually due to Kallianpur and Rao (1955). The proof is so beautifully simple that it can be included in undergraduate courses. I can't resist sketching the argument for those who have not seen it before. Let $X_1$ take $k$ possible values with probability $\pi_1(\theta), \ldots, \pi_k(\theta)$. A sufficient statistic based on $X_1, \ldots, X_n$ is $(p_1^n, \ldots, p_k^n)$ where $p_j^n$ is the proportion of the $j$-th value among $X_1, X_2, \ldots, X_n$. The log likelihood $\log L_n$ is $\sum_1^k np_j^n \log \pi_j(\theta)$. Consider only estimates of form $T_n = g(p^n)$. Such a $T_n$ will be called Fisher consistent if

$$g(\pi(\theta)) = \theta \qquad \ldots \quad (4)$$

and $g$ is continuous in $p^m$. If in addition $g$ is smooth in the sense of being continuously differentiable, then an application of the delta method shows $\sqrt{n}(T_n-\theta)$ and $n^{-1/2}d \log L_n/d\theta$ have in the limit a bivariate normal distribution with covariance

$$\Sigma \frac{\partial g(\pi(\theta))}{\partial \pi_i} \cdot \pi_i'(\theta)$$

which on differentiating (4) is one; (2) now follows from the Cauchy-Schwarz exactly as in the proof of the Cramer-Rao.

This sort of approach was extended in Kallianpur and Rao (1955) to Frechet differentiable statistics and in Kallianpur (1963) to statistics differentiable in the sense of von Mises. However in parametric estimation theory these sufficient conditions do not seem to be satisfactory, they are certainly rather restrictive. From the point of view of a statistician, the most appealing sufficient condition is the one due to Rao (1963) who strengthens (1) by requiring that it be uniform on compacts. The motivation for this is that if one is to use the normal approximation to the distribution function of $T_n$, one must be able to choose an $n_0$ such that for $n > n_0$, $\sqrt{n}(T_n-\theta)$ is normal to the degree of approximation required. If $n_0$ depends on $\theta$ which is unknown, the approximation would be useless. With uniformity on compacts we can get an $n_0$ which works since in practice we can often find a bounded set that contains the unknown $\theta$. Ideally one would like to have uniformity over the whole parameter space but this is rarely attainable except for location parameter problems.

Assuming $\sqrt{n}(T_n-\theta)$ is $A.N.(0, v(\theta))$ uniformly on compacts, Rao proves (2). As noted by Rao (1963), uniformity and the fact that $f_\theta$ is continuous in $\theta$ ensure that $v(\theta)$ is continuous. Since he also assumes, among other things, continuity of $I(\theta)$, his result follows from LeCam's inequality. Rao's own proof is different and much more innovative. The proof that he gives can be shown to hold under the weaker regularity conditions assumed in Section 2, see for example the calculations in Bahadur (1964). In particular one doesn't have to assume $I(\theta)$ is continuous; then continuity of $f_\theta$ and $f_\theta'$ gives, via Fatou's lemma, only lower semicontinuity of $I(\theta)$. In order to see why uniformity prevents superefficiency, it would be interesting to get a fairly direct proof of Rao's result from LeCam's inequality under the regularity conditions assumed in Section 2.

Rao's own proof involves a clever application of the Neyman-Pearson lemma. Fix $\theta_0$ and $1 > \alpha > 0$ and compute the limiting power of the most powerful test of size $\alpha$ for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_0 + \delta/\sqrt{n}$. Compute also

the limiting power of the test which rejects $H_0$ if $T_n > c_n$ with suitable randomisation at the boundary to achieve size $\alpha$. The fact that the limiting power of the most powerful test must be the bigger of the two quantities is just the inequality (2).

The same idea of evaluating an estimate by comparing the limiting power of an associated test with the limiting power of the optimum test has been used independently by Bahadur (1964) to prove LeCam's inequality. Bahadur first proves by the method outlined above that $v(\theta_0) \geqslant I^{-1}(\theta_0)$ at any $\theta_0$ where $T_n$ satisfies

$$\underline{lim}\ P_{\theta_0+n^{-\frac{1}{2}}}\{T_n < \theta_0+n^{-\frac{1}{2}}\} \leqslant \tfrac{1}{2}. \qquad \ldots \quad (5)$$

The associated test accepts $\theta = \theta_0$ iff $T_n < \theta_0+n^{-\frac{1}{2}}$ and (5) bounds its error of second kind. Bahadur then shows that (5) holds for almost all $\theta_0$, thus proving LeCam's inequality. Bahadur's proof that (5) holds almost everywhere also leads to the following fact (which should be compared with the Hájek-Inagaki invariance condition (7)): Suppose that under $\theta$, $\sqrt{n}(T_n-\theta)$ converges in law to $G_\theta$ for all $\theta$; then for fixed real $\delta$ there exists a subsequence $\{m\} \subset \{n\}$ such that under $\theta+\delta/\sqrt{m}$,

$$\text{the distribution of } \sqrt{m}(T_m-\theta-\delta/\sqrt{m})$$

$$\text{converges weakly to } G_\theta \text{ for almost all } \theta. \qquad \ldots \quad (6)$$

Rao's result has been extended in two directions. The limiting law for $\sqrt{n}(T_n-\theta)$ is allowed to be non-normal but with zero as the unique median and instead of comparing the asymptotic variance $v(\theta)$ with $I^{-1}(\theta)$, one compares the limiting concentration probability $\lim P_\theta\{-r < \sqrt{n}(T_n-\theta) < r\}$ with the corresponding probability for $N(0, I^{-1}(\theta))$. Predictably, there is also an analogue of LeCam's inequality. The basic technique in all cases is an application of the Neyman-Pearson lemma. The importnt papers are Wolfowitz (1965), Schmetterer (1966) (who introduces the interesting and useful notion of continuous convergence) and Pfanzagl (1970).

There is a deep result of a rather different sort—the Hájek-Inagaki convolution theorem—from which most, if not all, of the above results can be obtained as a corollary via Anderson's inequality. Assume that $\forall\ \delta$, under $\theta+\delta/\sqrt{n}$,

$$\text{the distribution of } \sqrt{n}(T_n-\theta-\delta/\sqrt{n}) \xrightarrow{w} G_\theta \qquad \ldots \quad (7)$$

where $G_\theta$ does not depend on $\delta$; then $G_\theta$ is the convolution of $N(0, I^{-1}(\theta))$ and some other probability distribution. Clearly this means $G_\theta$ must be

more spread out than $N(0, I^{-1}(\theta))$. The result was obtained at about the same time by Hájek (1970) and Inagaki (1970), Inagaki's conditions being much stronger than needed. An elegant proof based on the idea of limiting experiments and Kakutani's fixed point theorem is sketched in LeCam (1972); a more detailed exposition of the same proof is available in Millar (1983). An elegant elementary proof due to Bickel is presented in Roussas (1972). Jegannathan (1980) attributes to LeCam the remark that the existence of a limiting distribution of $\sqrt{n}(T_n-\theta)\ \forall\ \theta$ implies the Hájek-Inagaki invariance condition (7) holds for almost all $\theta$. In the proof of the convolution theorem, e.g., Bickel's as reproduced in the next page, one needs (7) only for a countably dense set of $\delta$'s and for some subsequence $\{m\} \subset \{n\}$. This weaker condition follows easily from (6) by the standard diagonal selection procedure.

The convolution theorem may be thought of as a modern version of another of Fisher's old results which says that difference between an inefficient estimate $T_n$ and the efficient estimate $\hat{\theta}_n$ is (asymptotically) independent of $\hat{\theta}_n$. A rigorously treated special case and heuristic argument for the general may clarify what is involved in proving the convolution theorem. Fix $\theta_0$ and consider all $\theta$'s which are of the form $\theta = \theta_0+\delta/\sqrt{n}$, $\delta$ any real number. For such $\theta$'s the scaled score function

$$V_n = \frac{1}{\sqrt{n}} \left.\frac{d \log L_n}{d\theta}\right|_{\theta_0} /I(\theta_0)$$

is asymptotically sufficient (in the LeCam-Wald sense, explained, say, in Roussas, (1972, Chapter 3).

We first consider the important (but technically rather trivial) special case where in addition to (7) $\sqrt{n}(T_n-\theta_0)$ and $V_n$ are (under $\theta_0$) asymptotically bivariate normal. Let $T$, $V$ denote the random variables corresponding to the limiting distribution. By a well-known result on contiguity, vide Roussas (1972, Ch. 1), $\sqrt{n}(T_n-\theta_0)$ and $V_n$ are asymptotically bivariate normal under $\theta = \theta_0+\delta/\sqrt{n}$ with same dispersion matrix as under $\theta_0$ and mean of $V_n$ equal to $\delta$, mean of $\sqrt{n}(T_n-\theta_0)$ equal to $I(\theta_0)\sigma_{12}\cdot\delta$, where $\sigma_{12}$ is the covariance of the bivarate normal. By (7), $\sigma_{12}$ must be $1/I(\theta_0)$. This immediately implies $(T-V)$ is orthogonal to $V$ and hence independent of $V$; this, of course, is the convolution theorem. It is worth noting that here among unbiased estimates of $\delta$, based on $T$ and $V$, $V$ is the best and so must be orthogonal to $T-V$ which is an unbiased estimate of zero.

We now turn to some heuristics in the general case. Suppose $T_n$ satisfies (7). Passing to a subsequence we can assumet hat under $\theta_0$, the joint distribution function of $\sqrt{n}(T_n-\theta_0)$ and $V_n$ converges weakly to, say,

$G_\delta(t, v)$. As before, this implies the existence of a limiting distribution $G_\delta$ under $\theta_0+\delta/\sqrt{n}$. Let $T, V$ be random variables having distribution $G_\delta$ under $\delta$. Then one expects as before that $V$ is a complete sufficient statistic, $N(\delta, I^{-1}(\theta_0))$, and by (7) the marginal distribution of $(T-V)$, under $\delta$, is free of $\delta$. From this one would have to conclude that $V$ and $(T-V)$ are independent. To do this suppose the conditional distribution of $T$ given $V$ has a density. Then, using sufficiency of $V$,

$$dG_\delta(t, v) = g(t \mid v)dtN(\delta, I^{-1}(\delta_0))(d\,v). \qquad \ldots \quad (8)$$

Clearly it is enough to prove that $g(t \mid v)$ is of the form $h(t-v)$ for then the joint distribution factors in the way needed for independence of $(T-V)$ and $V$. Observe that for any bounded measurable $\psi$, (7) and an elementary change of variables show

$$\iint \psi(t-\delta)dG_\delta - \iint \psi(t)dG_0 =$$

$$\int\int \psi(t)\{g(t+\delta \mid v+\delta) - g(t \mid v)\}dt\, N(0, I^{-1}(\theta_0))(dv) = 0. \qquad \ldots \quad (9)$$

This implies, via Stein's lemma, Lehmann (1959, p. 225), $g(t+\delta \mid v+\delta)$ is invariant under translations and hence must be a function of $(t-v)$ only. The proof suggested by LeCam (1972) is a rigorous version of this.

Bickel's proof depends on a different kind of tricky calculation, which often comes in handy in problems involving contiguous distributions. Passing to a subsequence assume as before that, under $\theta_0$, $\sqrt{n}(T_n-\theta_0)$ and $\sqrt{n}V_n$ have a limiting joint distribution $G_0$. Let $\phi(h_1, h_2)$ denote the m.g.f. of $G_0$ defined for all complex $h_1$, $h_2$ for which the integral in question exists. Clearly the convolution theorem would follow if we can show

$$\phi(ih, 0) = \psi(h)e^{-\frac{h^2}{2I(\theta_0)}} \qquad \ldots \quad (10)$$

where $\psi(h) = \phi(ih, 0)\, e^{h^2/2I(\theta_0)}$ must be a characteristic function of some random variable. To prove this use (7), and observe

$$\begin{aligned}\phi(ih_1, 0) &= \lim E_{\theta_0+h_2/\sqrt{n}}[\exp\{ih_1\sqrt{n}(T_n-\theta_0-h_2/\sqrt{n})\}] \\ &= \lim E_{\theta_0}[\exp\{ih_1\sqrt{n}(T_n-\theta_0)-ih_1h_2\}]\left[\frac{p_{\theta_0+h_2/\sqrt{n}}}{p_{\theta_0}}\right] \\ &= \phi(ih_1, h_2I)\exp\{-ih_1h_2-(h_2^2/2)I\} \qquad \ldots \quad (10a)\end{aligned}$$

where in the last step one uses the stochastic expansion

$$\log\{p_{\theta_0+h_2/\sqrt{n}}/p_{\theta_0}\} = I\,h_2\sqrt{n}V_n - \tfrac{1}{2}h_2^2I + o_p(1)$$

and

$$I = I(\theta_0).$$

By analyticity (10$a$) holds for all complex $h_2$. Putting $h_1 = h$, $h_2 = -ih/I$, one sees that $\psi(h)$ as defined in (10) equals $\phi(ih, -ih)$ which is the characteristic function of $T-V$.

## 5. Asymptotic local minimax theorem

I now turn finally to the other method of dealing with the Hodges examples in which one considers all estimates but evaluates them by a limiting local minimax criterion. I begin with some notations.

Let $l(y)$ be a symmetric bowl-shaped loss function, i.e., $l(y) = l(-y)$, $l(|y|)$ is non-decreasing in $|y|$ and $l(0) = 0$. I also assume for convenience that $l$ is bounded. What is actually needed is that it is integrable with respect to all $N(0, \sigma^2)$, the more general case following from the bounded case by an easy truncation argument. For any estimate $T_n$, let the risk be $R(\theta, T_n) = E_\theta\{l(\sqrt{n}(T_n - \theta))\}$. The performance at $\theta$ is however evaluated not by $R(\theta, T_n)$ but by the supremum of $R(\theta', T_n)$ over the set $|\theta' - \theta| < \delta$, where $\delta$ is a small positive number that will eventually tend to zero. This is rather like the smoothing out of a bad function. The asymptotic performance at $\theta$ is measured by

$$\lim_{\delta \downarrow 0} \lim_{n \to \infty} \sup_{|\theta' - \theta| < \delta} R(\theta', T_n) = \rho(\theta, \{T_n\}), \text{ say.}$$

This seems a fairly satisfactory thing to do except for the arbitrariness of the loss function $l$ and the fact that the loss in estimating $\theta$ by $T_n$ is taken to be $l(\sqrt{n}(T_n - \theta))$. It turns out (in view of (11) below) that arbitrariness of $l$ doesn't matter but the scaling by $\sqrt{n}$ remains, to some extent, more a matter of tradition and mathematical convenience than good statistical commonsense.

The basic Hájek-LeCam theorem in this area is the striking inequality

$$\rho(\theta, \{T_n\}) \geqslant \int_{-\infty}^{\infty} l(y) N(0, I^{-1}(\theta))(dy). \qquad \ldots (11)$$

Equality holds above if $\sqrt{n}(T_n - \theta)$ is $A.N.(0, I^{-1}(\theta))$ uniformly on compacts and $l$ is say, continuous so that the integration by parts formula holds for any probability distribution function $F$,

$$\int_0^{\infty} l(y) dF(y) = \int_0^{\infty} (1 - F(y)) dl(y). \qquad \ldots (12)$$

Thus (11) does provide a means of comparing arbitrary estimates with the maximum likelihood estimate provided $\sqrt{n}(\hat{\theta}_n - \theta)$ is $A.N.(0, I^{-1}(\theta))$ uniformly on compacts and (12) holds. Moreover (11) rules out superefficiency in the sense that no estimate which is superefficient at $\theta$ can attain the lower bound in (11), for Hájek (1972) has proved that a necessary condition for attaining the lower bound for a nonconstant $l$ and with $\lim$ replaced by $\liminf$ above, is

$$\sqrt{n}(T_n - \theta) - V_n \xrightarrow{p} 0 \qquad \ldots (13)$$

where $V_n$ is as defined in the previous section, i.e., $T_n$ behaves essentially like $\hat{\theta}_n$ under $\theta$. In the special case where the observations are i.i.d. $N(\theta, 1)$, $T_n$ is the Hodges estimate and $\theta = \theta_0$, the left hand side of (11) is equal to $l(\infty)$, i.e., the behaviour is the worst possible.

It is interesting to note that relations like (13) had been proposed earlier by Rao (1963, pp. 193, 194) as definition of efficiency from the Fisherian point of view of approximating the score function. It is reassuring that in this case, at least his Definition 3C has good frequentist implications.

The inequality (11) can be proved in two closely related ways. We first sketch the original method due to LeCam (1953) since it seems to be the more natural one.

Fix $\theta$ and choose a smooth prior density $\pi(\theta)$ which is positive on $(\theta-\delta, \theta+\delta)$ and is zero elsewhere. Then,

$$\sup_{|\theta'-\theta|<\delta} R(\theta', T_n) \geqslant \inf_{T_n'} \int R(\theta', T_n')\pi(\theta')d\theta'. \qquad \ldots (14)$$

As most statisticians know, if $n \to \infty$, with probability tending to one the posterior is approximately normal with mean $\hat{\theta}_n$ (and variance $(d^2 \log L_n/d\theta^2)^{-1}$ evaluated at $\hat{\theta}_n$). So $\hat{\theta}_n$ is nearly Bayes for the loss functions under consideration. Hence hopefully the right hand side of (14) will converge to

$$\lim_{n\to\infty} \int R(\theta', \hat{\theta}_n)\pi(\theta')d\theta' = \iint l(y)N(0, I^{-1}(\theta'))(dy)\pi(\theta')d\theta' \qquad \ldots (15)$$

which will converge to the right hand side of (11) if $\delta \to 0$. Thus (15) constitutes the heart of the "proof" of (11).

The trouble with this proof is that justification of (15) needs fairly strong assumptions (stronger than the assumptions in Section 2) and the details are rather messy. The alternative, and more elegant, proof is due to Hájek but the version we sketch below is the substantially simplified form in which it occurs in Millar (1983); the basic ideas in the simplification are due to LeCam (1972).

To motivate the second method note that in the first proof one bounds the minimax risk by a Bayes risk, vide (14), and then takes the limit, vide (15). In the second method, the order of these two operations is reversed and to do this the very innovative notion of a limiting experiment is introduced. It is first proved that the limiting minimax risk is greater than or equal to the minimax risk of the limiting experiment and then the latter is calculated by a well-known Bayesian argument; in fact the computation of

the minimax risk of the limiting experiment turns out to be the well-known problem of finding a minimax estimate for a normal mean with known variance.

In view of Proposition 2.3 of Millar a limiting experiment may be defined as follows. Consider a family of probability measures $\{Q_\delta^n, \delta \in \Theta\}$ (on some measurable space) such that each $\{Q_\delta^n\}$ is contiguous to $\{Q_{\delta_0}^n\}$, $\delta_0$ being a fixed element of $\Theta$. Let $\{Q_\delta, \delta \in \Theta\}$ be another family of probability measures with the same index set $\Theta$ (but possibly and usually on a different measurable space) such that $Q_\delta$ is dominated by $Q_{\delta_0}$. Then $\{Q_\delta, \delta \in \Theta\}$ is the limiting experiment for $\{Q_\delta^n, \delta \in \Theta\}$ if the $Q_{\delta_0}^n$-distribution of the $k$-tuple $\left(\frac{dQ_{\delta_i}^n}{dQ_{\delta_0}^n}, i=1, \ldots, k\right)$ converges weakly to the $Q_{\delta_0}$-distribution of $\left(\frac{dQ_{\delta_i}}{dQ_{\delta_0}}, i=1, \ldots, k\right)$ for all choice of $k$ and $\delta_1, \ldots, \delta_k \in \Theta$. Roughly speaking, the likelihoods in the limiting experiment provide a clue to how the likelihoods of the $n$-stage experiment behave for large $n$. To apply this definition in the present context, fix $A > 0$ and define $Q_\delta^n \equiv P_{\theta+\delta/\sqrt{n}}^n$, $|\delta| \leqslant A$. Then easy calculation shows the limiting experiment is $\{N(\delta, I^{-1}(\theta)), |\delta| \leqslant A\}$. It follows by a general result for limiting experiments that

$$\liminf_{T_n} \sup_{|\theta'-\theta| \leqslant A/\sqrt{n}} R(\theta', T_n) \geqslant \inf_{T} \sup_{|\delta| \leqslant A} \int l(T(y)-\delta) N(\delta, I^{-1}(\theta))(dy). \quad \ldots \quad (16)$$

But the limit of the right hand side of (16) as $A \to \infty$ is $\int l(y) N(0, I^{-1}(\theta))(dy)$ i.e., $T(y) = y$ is the minimax estimate in the limiting experiment, which is proved by Bayesian arguments or invariance considerations. Hence making $A \to \infty$ in (16) one gets

$$\lim_{\delta \downarrow 0} \liminf_{|\theta'-\theta| < \delta} \sup R(\theta', T_n) \geqslant \lim_{A \to \infty} \liminf_{n \to \infty} \inf_{T_n} \sup_{|\theta'-\theta| \leqslant \frac{A}{\sqrt{n}}} R(\theta', T_n)$$

$$\geqslant \int l(y) N(0, I^{-1}(\theta))(dy). \quad \ldots \quad (17)$$

Fabian and Hannan (1982) refer to the second inequality in (17) as the sharper form of the Hájek-LeCam inequality and make interesting comments on the attainability of these bounds. Ibragimov and Khasminskii (1981) as well as Millar (1983) make significant applications to non-parametric problems. A slightly different version of the LAMN theory due to Stone is available in Lehmann (1983); it has the pedagogical advantage of being presentable in an elementary way but the criterion used is less appealing.

## 6. Miscellaneous

We conclude with two general remarks, the first being concerned with those aspects of efficiency that we have either ignored or stressed insufficiently and the second with various extensions.

*Remark* 1 : To a Bayesian the most important property of $\hat{\theta}_n$ is its approximate Bayes property. Indeed this means that in an estimation problem for a wide variety of priors and a wide variety of loss functions, the Bayes estimate is often independent of both the prior and the loss function. Pitman has called the Glivenko-Cantelli theorem the existence theorem for statistics. Many statisticians will feel the same way about the asymptotic normality of the posterior. I have not so far given any references for this result partly because like many basic results it has a long history and many names associated with it. It can be traced back to Laplace, more modern contributors are Bernstein and von Mises, vide Borwanker *et al.* (1971), and Kolmogorov, vide Wolfowitz (1953).

This approximate Bayes property of $\hat{\theta}_n$ may also be thought of as the main reason why $\hat{\theta}_n$ is efficient from the frequentist point of view. An excellent early heuristic paper on maximum likelihood estimation along these lines is Wolfowitz (1953). The same point of view underlies Weiss and Wolfowitz (1974).

It may be pointed out that most of the results on asymptotic normality of posteriors—for example Johnson (1970) or Borwanker, Prakasa Rao and Kallianpur (1971) to give two recent references—treat of the case where the true distribution is that holding under $\theta$, $\theta$ being a fixed point in whose neighbourhood the prior density $\pi(\theta')$ is positive and continuous. However both from the Bayesian and the frequentist point of view this is not always satisfactory. From a Bayesian point of view it seems much more natural to study the asymptotic behaviour of the posterior when $\theta$ is a random variable with $\pi(\theta')$ as its density. From a frequentist point of view one needs to compare the integrated risk of the Bayes estimate and the integrated risk of $\hat{\theta}_n$. Such studies were initiated in LeCam (1953, 1958) and continued in Ghosh, Joshi and Sinha (1982), where asymptotic expansions are also considered.

It is interesting to note that though the approximate Bayes role of $\hat{\theta}_n$ is so fundamental, it has not played any important role in the study of efficiency except in LeCam (1953) and Wolfowitz (1953) and, to some extent, Weiss and Wolfowitz (1974). Indeed though most of the lower bounds were

motivated by the need to justify $\hat{\theta}_n$, $\hat{\theta}_n$ itself does not play any important role in their derivation. This is partly because the lower bounds have been developed under conditions which are much weaker than those assumed to prove asymptotic Bayes property of $\hat{\theta}_n$.

As a matter of fact the lower bounds are proved under the so called LAN conditions due to LeCam (1960) which are much weaker than even those assumed to prove asymptotic normality of $\hat{\theta}_n$. The two main tools in this area, both due to LeCam, are contiguity and the LAN conditions. A brief description of the latter seems to be in order. The family $\{P_\theta^n, \theta \in \Theta, n \geqslant 1\}$ satisfies the LAN condition at $\theta$ if the log-likelihood for contiguous alternatives $\log\{dP^n_{\theta+\delta_n/\sqrt{n}}/dP^n_\theta\}$ can be written as a quadratic

$$\delta_n W_n - \tfrac{1}{2}\delta_n^2 I(\theta) + o_p(1)$$

where $W_n$ is $A.N.(0, I(\theta))$, $\forall$ bounded $\delta_n$. All the results surveyed above hold provided only that the LAN condition is valid for all $\theta$. Independence or identical distribution is not needed. In particular many Markovian examples can be handled in this way, vide Roussas (1972).

Getting sufficient condition for $\hat{\theta}_n$ to attain these lower bounds boils down to getting a sufficient condition for its uniform asymptotic normality on compacts. Such problems are best tackled by getting Berry-Esseen bounds for $\hat{\theta}_n$. The best possible rate for the one-parameter case is given by Pfanzagl (1971), slightly worse rates for the multiparameter case are given in Pfanzagl (1973a). A much simpler treatment of the multiparameter case is available in Bhattacharya and Ghosh (1978).

*Remark 2:* The extension to the multiparameter case is, generally speaking, straightforward. All the lower bounds and the Hájek-Inagaki convolution theorem remain valid, the definition of bowl-shaped loss in the multiparameter case is given in Hájek (1970). The only result which does not hold—and is, thus, a reminder of the Stein phenomenon—is the necessary condition (13) for dimension greater than two. On the other hand Pfanzagl (1980, pp. 26–27) makes an interesting remark on the Hájek-Inagaki convolution theorem which shows that in a certain sense the Stein phenomenon is absent. See also a curious example due to Takeuchi, reproduced in Pfanzagl (1980, p. 28).

Of other extensions we mention only two. Weiss and Wolfowitz (1974) have studied the efficiency of the maximum probability estimate for many non-regular cases. The success of their programme suggests that in many non-regular cases, as in the regular case, the posterior doesn't depend

much on the prior but the Bayes estimate depends on the loss function chosen. An analogue of the Bernstein-von Mises theorem is worth investigating The examples in Weiss and Wolfowitz (1974) are interesting but their argument for their basic result, though elegant and tantalisingly simple, isn't very illuminating. Possibly their main result as well as the examples can be profitably studied from the Hájek-LeCam-Millar point of view as outlined in Section 5.

The other substantial extension has been to the so called LAMN (locally asymptotically mixed normal) case which generalises the LAN condition by allowing a random quadratic term. Many "non-ergodic" Markovian examples belong to this class. A penetrating analysis of all aspects of estimation under the LAMN condition is presented in Jegannathan (1980) and Swensen. Much of Jegannathan's work can be found in recent issues of *Sankhyā A*. An excellent recent monograph is Basawa and Scott (1983).

*Acknowledgement.* I would like to thank Mr. T. Samanta who read the paper carefully, weeding out several errors.

## References

Akahira, M. and Takeuchi, K. (1976): On the second order asymptotic efficiencies of estimators. *Proceedings of the Third Japan-USSR Symposium on Probability Theory* (G. Maruyama and Y. V. Prokhorov, eds.). *Lecture Notes in Mathematics*, Springer-Verlag, Berlin.

Bahadur, R. R. (1964): On Fisher's bound for asymptotic variances. *Ann. Math. Statist.*, **35**, 1545-1552.

Basawa, I. V. and Scott, D. S. (1983): Asymptotic optimal inference for non-ergodic models. *Lecture Notes in Statistics*, Springer-Verlag, New York.

Bhattacharya, R. N. and Ghosh, J. K. (1978): On the validity of the formal Edgeworth expansion. *Ann. Statist.*, **6**, 434-451.

Bickel, P. J., Chibisov, D. M. and Van Zwet, W. R. (1981): On efficiency of first and second order. *International Statistical Review*, **49**, 169-175.

Bickel, P. J., Goetze, F. and Van Zwet, W. R. (1984): A simple analysis of third order efficiency of estimates. To appear in the *Proceedings of the Neyman-Kiefer Symposium*.

Borwanker, J. D., Kallianpur, G. and Prakasa Rao, B. L. S. (1971): The Bernstein-von Mises theorem for Markov processes. *Ann. Math. Statist.*, **42**, 1241-1253.

Efron, B. (1975): Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.*, **3**, 1189-1242.

Fabian, V. and Hannan, J. (1982): On estimation and adaptive estimation for locally asymptotically normal families. *Z. Warsch. Verw. Gebite*, **59**, 459-478.

Ghosh, J. K. (1983): Review of "Approximation Theorems of Mathematical Statistics", by R. J. Serfling. *Jour. Amer. Statist. Assoc.* **78**, 731-732.

Ghosh, J. K. and Subramanyam, K. (1974): Second order efficiency of maximum likelihood estimators. *Sankhyā*, Ser. A, **36**, 325-358.

GHOSH, J. K., JOSHI, S. M. and SINHA, B. K. (1982): Expansions for posterior probability and integrated Bayes risk. *Proc. III Purdue Symp. on Decision Theory and Related Topics*, Vol. II, 403-456 (J. Berger and S. Gupta, Eds.), Academic Press, New York.

HÁJEK, J. (1970): A characterisation of limiting distributions of regular estimates. *Z. Warsch. Verw. Gebiete*, **14**, 323-330.

——— (1972): Local asymptotic minimax and admissibility in estimation, *Proc. 6th Berkeley Symp.*, Vol. 1, 175-194 (L. M. LeCam, J. Neyman and E. J. Scott, eds.) University of California Press, Berkeley and Los Angles.

HÁJEK, J. and SIDÁK (1967): *Theory of Rank Tests*, Academic Press, New York.

IBRAGIMOV, I. A. and Khasminski, R. Z. (1981): *Statistical Estimation: Asymptotic Theory*. Springer, New York.

INAGAKI, N. (1970): On the limiting distribution of a sequence of estimators with uniformity property. *Ann. Inst. Statist. Math.*, **22**, 1-13.

JEGANNATHAN, P. (1980): Asymptotic theory of estimation when the limit log likelihood ratios is mixed normal. Ph. D. Thesis submitted to Indian Statistical Institute.

KALLIANPUR, G. (1963): Von Mises functionals and maximum likelihood estimation. *Sankhyā*. Ser A, **1**. 149-158.

KALLIANPUR, G. and RAO, C. R. (1955): On Fisher's lower bound to asymptotic variance of a consistent estimate. *Sankhyā*, Ser. A, **15**, 331-342.

LECAM, L. (1953): On some asymptotic properties of maximum likelihood and Bayes estimates. *Univ. Calif. Publ. Statist.*, **1**, 277-330.

——— (1958): Les Proprietes asymptotiques des solutions de Bayes. *Pub. Inst. Statist Univ., Paris*, **3**, 17-35.

——— (1960): Locally asymptotically normal families of distributions. *Univ. Calif. Publ. Statist.*, **3**, 37-98.

——— (1972): Limits of experiments. *Proc. 6th Berkeley Symp.*, Vol. 1, 245-261 (L. M. LeCam, J. Neyman and E. L. Scott, eds.) University of California Press, Berkeley and Los Angeles.

LEHMANN, E. L. (1959): *Testing Statistical Hypotheses*, Wiley, New York.

LEHMANN, E. L. (1983): *Theory of Point Estimation*, Wiley, New York.

MILLAR, P. W. (1983): The minimax principle in asymptotic theory. Ecole d'Ele de Probabilities de Saint-Flour XI (P. L. Honnequin, ed.), *Lecture Notes in Mathematics* Springer-Verlag, Berlin.

PFANZAGL, J. (1970): On the asymptotic efficiency of median unbiased estimates. *Ann. Math. Statist.*, **41**, 1500-9.

——— (1971): The Berry-Esseen bound for minimum contrast estimates. *Metrika*, **17**, 82-91.

——— (1973): Asymptotic expansions related to minimum contrast estimators. *Ann. Statist.*, **1**, 993-1026; Corrections **2**, 1367-1368.

——— (1973a): The accuracy of the normal approximation for estimates of vector parameters. *Z. Wahsch. Verw. Gebiete*, **25**, 171-198.

——— (1975): On asymptotically complete classes. *Statistical Inference and Related Topics*. Vol. 2, 1-43, (M. L. Puri, ed.), Academic Press, New York.

PFANZAGL, J. (1980) : Asymptotic expansions in parametric statistical theory. In : Developments in Statistics, Vol. 3 (P. R. Krishnaiah, ed.), 1-97, Academic Press, New York.

PRATT, J. W. (1976) : F. Y. Edgeworth and R. A. Fisher on the efficiency of maximum likelihood estimation. *Ann. Statist.*, **4**, 501-514.

RAO, C. R. (1961) : Asymptotic efficiency and limiting information. *Proc. 4th Berkeley Symp.*, Vol. 1, 531-545 (J. Neyman, ed.) University of California Press, Berkeley.

——— (1962) : Efficient estimates and optimum inference procedures in large samples. *J. Roy. Statist. Soc.*, Ser. B, **24**, 46-72.

——— (1963) : Criteria of estimation in large samples. *Sankhyā*, Ser. A, **25**, 189-206.

——— (1965) : *Linear Statistical Inference and Its Applications*, Wiley, New York.

ROUSSAS, G. C. (1972) : *Contiguity of Probability Measures : Applications in Statistics*, Cambridge University Press, Cambridge.

SCHMETTERER, L. (1966) : On the asymptotic efficiency of estimates. *Research Papers in Statistics Festschrift for J. Neyman*, 301-317. (F. N. David, ed.) Wiley, New York.

SERFLING, R. J. (1980) : *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

WEISS, L. and WOLFOWITZ, J. (1974) : Maximum probability estimators and related topics. *Lecture notes in Mathematics*,. Springer-Verlag. Berlin.

WALKER, A. M. (1963) : A note on the asymptotic efficiency of an asymptotically normal estimator sequence. *J, Roy Statist. Soc.* Ser. B. **25**, 195-208.

WOLFOWITZ. J. (1953) : The method of maximum likelihood and the Wald theory of decision functions. *Indag. Mathematicae*, **15**, 114-119.

——— (1965) : Asymptotic efficiency of the maximum likelihood estimator. *Theor. Probability Appl.*, **10**, 247-60.

*Paper received : February, 1985.*