CHAPTER 7

Small Sample Efficiency

7.1. Introduction. There is no small sample analogue of third order efficiency, but the same ideas can be used to derive exact results for small samples. Bickel, Götze and van Zwet (1985) use a Bayesian argument and perturbation of loss functions to prove the old result that $\bar{x}$ is a minimum variance unbiased estimate for a normal mean, without using the Cramér–Rao inequality or the Rao–Blackwell theorem. (The result is slightly weaker in that there is a restriction on the estimates considered.) We reproduce this argument in Section 7.4.

Consider the following scenario. You have a practical problem, the sample size n is reasonably large, say, $n = 50$, you know $\hat{\theta}$ is not only FOE but TOE, and simulations show asymptotic normality is well borne out by simulations. Out of this clear, blue sky appears an estimate T which is always at least 80% as efficient as $\hat{\theta}$, most of the time more efficient and occasionally more than *30 times more efficient than* $\hat{\theta}$. That this nightmare (to people who have come to adore the mle) can be a reality is documented in Khedr and Katti (1982).

Note that not only higher order optimality of $\hat{\theta}$, but even first order optimality of $\hat{\theta}$ seems to be in doubt. What has gone wrong?

It is a weakness of all asymptotic theories of optimality that they apply only to sequences of estimates and are silent about what can happen with a particular n (even if large) and a particular estimate. The situation is similar to that with asymptotic expansions which may be correct to $o(n^{-1})$, but may be pretty bad for a particular n, even if n is large. (This is not the case in the above example since simulation of the mean square of $\hat{\theta}$ seems to agree well with its asymptotic value.) This can happen for the trivial reason that a term $An^{\ 3/2}$ is $o(n^{\ 1})$, but may be large for n as large as 50 because the constant A is large. For most asymptotic expansions, we do not have good bounds on the constant A, but accuracy can be and often is checked by simulation. Also, one has the Berry–Esseen bound for asymptotic normality for $\sqrt{n}(\bar{x} - \theta)$

with a sharp bound on the associated constant. This may be used to get reasonably good, though conservative, bounds for at least asymptotic normality of $\sqrt{n}(\hat{\theta} - \theta)$. Simulation is no answer in the problem with asymptotic optimality because there are infinitely many estimates to worry about. So something like a (conservative) lower bound to the mean square of an estimate is needed. The Cramér–Rao bound will not do because it either requires unbiasedness or it provides a lower bound depending on the bias of the estimate, and the potentially much better estimates than $\hat{\theta}$ are likely to have an unknown bias $b(\theta)$, where $|b(\theta)|$ itself is small but its derivative $b'(\theta)$ is large in magnitude. In Section 7.3 we will develop a lower bound to the local minimax risk over $(\theta - \delta, \theta + \delta)$. If this lower bound is close to $(nI(\theta))^{-1}$ (which is approximately the risk of the mle for all θ, it would be impossible to have the phenomenon of an estimate which is much better in $(\theta_0 - \delta, \theta_0 + \delta)$ for some θ_0. Ideally, it should be possible to improve the bound at each θ_0 by confining to estimates whose risk at all θ is no worse than, say, a certain specified function of θ, which may be a specified fraction (like 80%) of $(nI(\theta))^{-1}$. However, that seems analytically intractable.

The bound in Section 7.3 is obtained by an application of a Cramér–Rao type lower bound to the (integrated) Bayes risk and a variational result. The lower bound to Bayes risk is due to Borovkov and Sakhanienko (1980). This will be the subject of Section 7.2. Improvements have been obtained recently by Brown and Gajek (1990) and, in one special case, by DasGupta and Vidakovic; see Vidakovic (1992). (See also Bobrovosky, Mayer-Wolf and Zakai (1987).) The variational result is used in Bickel (1981), and Levit (1980) and, in a more general form, is owing to Huber (1974) (it was rediscovered by Ghosh and Bhattacharya in 1983, and it is probably an old result in calculus of variations). The results of Bickel (1981) and Levit (1980), in the context of estimating a normal mean knowing that it lies in a specified interval, has far reaching implications in much recent work on asymptotic or approximate minimaxity.

The multiparametric extension in Section 7.2 and the lower bound in Section 7.3 are joint unpublished work of Ghosh and Joshi. The lower bound of Section 7.3, along with the variational result and the multiparameter extension with Joshi, was part of one of three examples presented at the Neyman–Kiefer Conference in Berkeley in 1983 by Ghosh, but has not been published before. [Of the other two examples presented at the Neyman–Kiefer Conference, one was published as part of DasGupta and Ghosh (1983) and the other appears in Ghosh and Sen (1985).]

Some problems are posed in the final section.

7.2. A lower bound to Bayes risk.

The following inequality on the (integrated) Bayes risk is due to Borovkov and Sakhanienko (1980). Let X_i's be n i.i.d. r.v.'s with density $p(x_1, x_2, \dots, x_n | \theta)$ satisfying the regularity conditions of Chapter 1.

THEOREM 7.1. *Under regularity conditions,*

$$R(\pi) \geq n^{-1} \int_a^b \frac{\pi(\theta)}{I(\theta)}\, d\theta - n^{-2} \int_a^b \left\{ \frac{d}{d\theta}(\pi I^{-1}) \right\}^2 \frac{1}{\pi}\, d\theta, \tag{7.1}$$

where $[a, b]$ *is the support of* π, I *is positive and continuously differentiable on* $[a, b]$, π *is continuously differentiable on* $[a, b]$ *and* $\pi = 0$ *at* a *and* b.

The theorem is proved under weaker assumptions in Borovkov and Sakhanienko (1980). It is derived from the Carmér–Rao inequality in Brown and Gajek (1990).

PROOF OF THEOREM 7.1. Assume without loss of generality the second integral in (7.1) is finite. Let

$$G = \frac{1}{I(\theta)} \frac{d}{d\theta} \log(\pi(\theta) p(x_1, x_2, \dots, x_n | \theta)/I(\theta)). \tag{7.2}$$

Check that

$$G = \frac{1}{I} \frac{d}{d\theta}\left(\log \frac{\pi}{I}\right) + \frac{1}{I}\left(\frac{d \log p}{d\theta}\right) \tag{7.3}$$

and the first term in (7.3) is

$$\left(\frac{d}{d\theta} \frac{\pi}{I}\right) \frac{1}{\pi}, \tag{7.4}$$

which is square integrable with respect to π under the assumption made at the beginning of the proof. The second term in G in (7.3) is square integrable with respect to $\pi \otimes P_\theta$, so G is square integrable.

Let

$$B = \text{the Bayes estimate, that is, } E(\theta | X_1, X_2, \dots, X_n). \tag{7.5}$$

By (7.3, (7.4) and $\pi(a) = \pi(b) = 0$,

$$E(G) = \int_a^b \frac{d}{d\theta}\left(\frac{\pi}{I}\right) + E\left(E\left(\frac{1}{I} \frac{d \log p}{d\theta} \middle| \theta\right)\right) = 0. \tag{7.6}$$

Similarly,

$$\begin{aligned} E(G^2) &= \int_a^b \left(\frac{d}{d\theta} \frac{\pi}{I}\right)^2 \frac{1}{\pi} + n \int_a^b \frac{\pi}{I}\, d\theta \\ &\underset{\text{def}}{=} A_1 + nA_2, \end{aligned} \tag{7.7}$$

$$\begin{aligned} \operatorname{Cov}(B, G) &= \int \cdots \int B(x_2, x_2, \dots, x_n) \\ &\quad \times \frac{d}{d\theta}\left(\frac{\pi}{I} \cdot p(x_1, x_2 \cdots x_n | \theta)\right) d\theta\, dx_1 \cdots dx_n \\ &= 0. \end{aligned} \tag{7.8}$$

Also integrating by parts,

$$\operatorname{Cov}(\theta, G) = \int_a^b \theta\left(\frac{d}{d\theta}\frac{\pi}{I}\right) d\theta = -\int_a^b \frac{\pi}{I}\, d\theta = -A_2, \tag{7.9}$$

so

$$\operatorname{Cov}(B - \theta, G) = -A_2 \tag{7.10}$$

and by (7.6), (7.7) and (7.10),

$$\begin{aligned} R(\pi) = E(B - \theta)^2 &\geq A_2^2(A_1 + nA_2)^{-1} \\ &= \frac{A_2}{n}\left(1 + \frac{A_1}{nA_2}\right)^{-1} \geq \frac{A_2}{n}\left(1 - \frac{A_1}{nA_2}\right) \end{aligned} \tag{7.11}$$

since $(1 + x)^{-1} \geq 1 - x$ for $x \geq 0$. This completes the proof. □

If one has a k-dimensional parameter, $\theta = (\theta_1, \theta_2, \dots, \theta_k)$, let $[I_{ij}(\theta)]$ be the information matrix, assumed nonsingular, and let $[I^{ij}(\theta)]$ be the inverse. To get a multiparameter analogue of Theorem 7.1, we assume the same regularity conditions on p, take the support of π to be compact, $\pi = 0$ on the boundary of its support, π is continuously differentiable on its support and the information matrix is positive definite and continuously differentiable on the support of π. Let B_1 be the Bayes estimate of θ_1, that is, $E(\theta_1 | x_1, x_2, \dots, x_n)$, and let $R_1(\pi) = E(B - \theta_1)^2$ be the Bayes risk. Then we have the following result due to Ghosh and Joshi (unpublished); see also Prakasa Rao (1992).

THEOREM 7.2. *Under the conditions stated above,*

$$\begin{aligned} R_1(\pi) \geq n^{-1} \int \cdots \int I^{11}(\theta)\pi(\theta)\, d\theta \\ - n^{-2} \sum_i \int \cdots \int \left\{\frac{\partial}{\partial\theta_i}(\pi I^{1i})\right\}^2 \frac{1}{\pi}\, d\theta. \end{aligned} \tag{7.12}$$

PROOF. Let

$$\begin{aligned} G &= \sum I^{1i} \frac{\partial}{\partial\theta_i} \log(\pi p I^{1i}) \\ &= \sum \frac{1}{\pi}\frac{\partial}{\partial\theta_i}(\pi I^{1i}) + \sum I^{1i} \frac{\partial \log p}{\partial\theta_1}. \end{aligned} \tag{7.13}$$

Proceeding in an identical way, we get

$$E(G) = 0, \tag{7.14}$$

$$\operatorname{Cov}(B_1 - \theta_1, G) = -\int \cdots \int I_{(\theta)}^{11} \pi(\theta)\, d\theta. \tag{7.15}$$

Also,

$$\operatorname{Var}\left(\sum I^{1i} \frac{\partial \log p}{\partial\theta_i} \middle| \theta\right) = nI^{11}(\theta). \tag{7.16}$$

Using (7.13) and (7.16),

$$(7.17)\quad \operatorname{Var}(G) = \sum_i \int \cdots \int \left(\frac{\partial}{\partial \theta_i} \pi I^{1i} \right)^2 \frac{1}{\pi}\, d\theta + n \int \cdots \int I^{11}(\theta) \pi(\theta)\, d\theta.$$

The inequality (7.12) now follows, exactly as in the proof of Theorem 7.1, by an application of the Cauchy–Schwarz inequality and the elementary fact $(1 + x)^{-1} \geq 1 - x$ for $x \geq 0$. □

7.3. A lower bound to the local minimax risk. Consider the setup for Theorem 7.1 and let the interval (a, b) be $(\theta_0 + \delta, \theta_0 + \delta)$. Then the local minimax risk at θ_0 is, for all π satisfying the conditions of Theorem 7.1,

$$\inf_T \sup_{\theta \in (\theta_0 - \delta,\, \theta_0 + \delta)} E\{(T - \theta)^2 | \theta\} \geq R(\pi)$$

$$(7.18)\qquad \geq \frac{A_2}{n}\left(1 - \frac{A_1}{A_2 n}\right)$$

$$\geq \frac{1}{n\bar{I}} - \frac{1}{\bar{I}\underline{I}n^2} \int_{\theta_0 - \delta_0}^{\theta_0 + \delta} \left(\frac{g'(\theta)}{g(\theta)} \right)^2 d\theta$$

where $\bar{I}(\theta_0)$ and $\underline{I}(\theta_0)$ are the maximum and minimum of $I(\theta)$ over $[\theta_0 - \delta, \theta_0 + \delta]$ and

$$(7.19)\qquad g(\theta) = \left\{ \frac{\pi(\theta)}{I(\theta)} \right\} \Bigg/ \int_{\theta_0 - \delta}^{\theta_0 + \delta} \frac{\pi(\theta)}{I(\theta)}\, d\theta$$

is a probability density.

To get the best bound of this kind, we maximize it with respect to π and make use of the variational result

$$(7.20)\qquad \inf_{\pi(\cdot)} \int_0^1 \frac{(\pi'(\theta))^2}{\pi(\theta)}\, d\theta = 4\pi^2.$$

(The infimum is over twice differentiable π with π equal to zero at $\theta = 0, 1$.)

Using (7.20) in (7.18) we get, finally,

$$(7.21)\quad \inf_T \sup_{\theta \in (\theta_0 - \delta,\, \theta_0 + \delta)} E\{(T - \theta)^2 | \theta\} \geq \frac{1}{n\bar{I}(\theta_0)} - \frac{\pi^2}{n^2 \bar{I}(\theta_0) \underline{I}(\theta_0) \delta^2}.$$

It is clear from (7.21) that if $(\bar{I}(\theta_0))^{-1}$ is close to $(I(\theta_0))^{-1}$, no estimate can do much better than $\hat{\theta}$, provided $E\{(\hat{\theta} - \theta)^2 | \theta\}$ is well approximated by $(nI(\theta))^{-1}$ and $n\underline{I}(\theta_0)\delta^2$ is moderately large. These conditions are likely to be violated if I is small in $(\theta_0 - \delta, \theta_0 + \delta)$. In any case, for given $I(\cdot)$ and $\delta > 0$ and $\varepsilon > 0$, one can calculate the smallest value of n, say n_0, for which

$$(7.22)\qquad \pi^2 / n\underline{I}(\theta_0)\delta^2 < \varepsilon.$$

Then for $n \geq n_0$,

$$(7.23)\qquad \inf_T \sup_{\theta \in (\theta_0 - \delta,\, \theta_0 + \delta)} E\{(T - \theta)^2 | \theta\} \geq \frac{1}{n\bar{I}(\theta_0)}(1 - \varepsilon).$$

For a suitable δ, ε, the value of n_0 can be used as a diagnostic tool for deciding whether the asymptotic lower bound $(nI(\theta))^{-1}$ is usable for a given n.

The same sort of thing can be done for the multiparameter case, using as support of π a k-dimensional rectangle and a π under which $\theta_1, \theta_2, \ldots, \theta_k$ are independent. In any case, it is clear from Theorem 7.2 that the optimality of $\hat{\theta}$ may be doubtful at θ_0 if the information matrix is nearly singular in $[\theta_0 - \delta, \theta_0 + \delta]$.

We end this section by sketching a proof of the variational result (7.20).

Put $q = \{\pi(\theta)\}^{1/2}$. Then,

$$\int_0^1 \frac{(\pi')^2}{\pi}\, d\theta = 4\int_0^1 (q')^2 = -4\int_0^1 qq'' = -4\langle q, Lq\rangle, \tag{7.24}$$

where $\langle f, g\rangle = \int fg\, d\theta$ and L is the differential operator defined by $Lf = f''$.

A pair (λ, f) is an eigenvalue and eigenfunction of L if

$$f'' = \lambda f. \tag{7.25}$$

The only possible values of λ are $-(k\pi)^2$, $k = 1, 2, 3, \ldots$, and the solution of (7.25) with $\lambda = -(k\pi)^2$ is

$$f_k = \sin k\pi\theta. \tag{7.26}$$

The set $\{f_k, k \geq 1\}$ is complete for the family of square integrable functions on $[0, 1]$, vanishing at the endpoint.

We wish to minimize (7.24) subject to $q(\theta) = 0$ at $\theta = 0, 1$, and

$$\int_0^1 q^2\, d\theta = 1. \tag{7.27}$$

Let q have the expansion

$$q = \sum a_k f_k. \tag{7.28}$$

Then we have to minimize

$$-4\langle q, Lq\rangle = 4\sum (k\pi)^2 a_k^2 \|f_k\|^2 = 4\sum (k\pi)^2 w_k \tag{7.29}$$

subject to

$$\sum w_k = \sum a_k^2 \|f_k\|^2 = 1, \tag{7.30}$$

where $\|f\|^2 = \int_0^1 f^2\, d\theta$.

It follows that the minimum value of (7.29) is obtained when $w_1 = 1$, $w_k = 0$ if $k > 2$. Thus

$$\inf \int_0^1 \frac{(\pi'(\theta))^2}{\pi(\theta)}\, d\theta = 4\pi^2$$

and the minimizing $\pi(\theta)$ is const. $\sin^2 \pi\theta$.

7.4. A "third order efficiency" proof that $\bar{X}$ is the minimum variance unbiased estimate of the normal mean. Let $X_1, X_2, \dots, X_n$ be i.i.d. $N(\theta, 1)$. We consider unbiased estimates T of θ satisfying

$$E(T^2|\theta) < \infty \qquad \forall\, \theta \tag{7.31}$$

and

$$\int_{-\infty}^{\infty} E(T^2|\theta) e^{-\alpha\theta^2}\, d\theta < \infty \quad \text{for some } \alpha > 0. \tag{7.32}$$

Condition (7.31) implies, by a property of exponential distributions, that $E\{(T - \theta)^2|\theta\}$ is continuous in θ.

We want to show $\bar{X}$ is the minimum variance unbiased estimate (MVUE) in the class of all unbiased T satisfying (7.31) and (7.32).

Consider a conjugate prior π under which θ is $N(\mu, \tau^2)$. The posterior mean is

$$E(\theta|X_1, X_2, \dots, X_n) = \lambda\mu + (1 - \lambda)\bar{X}, \tag{7.33}$$

where $\lambda = (n\tau^2 - 1)^{-1}$. This is the Bayes estimate under the squared error loss. We now consider a perturbed loss

$$L(\theta, a) = (\theta - a)^2 - 2B(\theta - c)(\theta - a) \tag{7.34}$$

and note that one can choose B and c to ensure that the Bayes estimate under this loss is $\bar{X}$ itself. [In fact $B = \lambda/(1 - \lambda)$ and $c = \mu$.] Also for an unbiased T,

$$E\{(T - \theta)^2|\theta\} = E(L(\theta, T)|\theta). \tag{7.35}$$

Hence

$$\begin{aligned} \int E\{(\bar{X} - \theta)^2|\theta\}\pi(\theta)\, d\theta &= \int E\{L(\theta, \bar{X})|\theta\}\pi(\theta)\, d\theta \\ &\le \int E\{L(\theta, T)|\theta\}\pi(\theta)\, d\theta \\ &= \int E\{(T - \theta)^2|\theta\}\pi(\theta)\, d\theta. \end{aligned} \tag{7.36}$$

Now making $\tau \to 0$, conclude, using (7.32) and continuity of $E\{(T - \theta)^2|\theta)$,

$$E\{(\bar{X} - \mu)^2|\mu\} \le E\{(T - \mu)^2|\mu\}. \tag{7.37}$$

Since this is true for all μ, $\bar{X}$ is MVUE among T's.

7.5. Problems. As indicated in Section 7.3, on the basis of Theorem 7.2, optimality of mle is doubtful when the information matrix is nearly singular. Can one exhibit an example with nearly singular information matrix where one can exhibit a much better estimate without having to go through simulations as in Khedr and Katti (1982)? In the same vein, but more precisely, one can ask if, for the multivariate normal, or more generally a multiparamter exponential density, one can do much better than a minimum variance unbiased estimate of the population mean if there is a near singularity of the information matrix.

It is known that near singularity can cause problems in deriving the asymptotic distribution of the mle for the independent, nonidentically distributed random variables. For an example in a reliability growth model, see Bhattacharya and Ghosh (1991). It would be nice to have a general theorem of which the result in reliability growth is a special case.

The proof of posterior normality and expansions would also be in trouble in the situation described in the previous paragraph. That too needs attention.