May 8, 2003

3.7 Efficiency of estimators. In this and the following two sections the distribution of the data is assumed to belong to a parametric family $\{P_\theta,\ \theta \in \Theta\}$, having densities $f(\theta, x)$.

The information inequality or Fréchet-Cramér-Rao lower bound, when Θ is an open interval in $\mathbb{R}$ and g is a differentiable real-valued function on Θ, is

$$\mathrm{var}_\theta(T_n) \ \geq \ g'(\theta)^2/(nI_1(\theta)),$$

where $I_1(\theta) \ := \ E_\theta((\partial f(\theta,x)/\partial\theta)^2)$, as was proved in Theorem 2.4.10 under some regularity conditions when T_n is an unbiased estimator of $g(\theta)$. But by Theorem 2.4.15, if $\log f(\theta, x)$ is C^1 in θ, the lower bound is attained for all θ only when the family of distributions is exponential of order 1 with $T(x)$ equal to the given estimator $T_n(x)$ where $x = (x_1, \dots, x_n)$. When this is true for one function $T(\cdot)$, the only other functions for which it holds are $aT(\cdot)+b$ where $a \neq 0$ and b are constants. So the only functions having unbiased estimators attaining the information inequality lower bound for all θ are $ag(\theta)+b$ where now a and b are any constants and g is the specific function $d\log K(\theta)/d\theta$ for which T is the unbiased estimator, by Corollary 2.5.9. Even for exponential families of order 1, unique unbiased, admissible estimators (for other functions) may be unsatisfactory, as in the example at the end of Sec. 2.5.

If the information inequality provided best possible lower bounds for mean-square errors only for estimating functions $ag(\theta)+b$ as just described, it would not be very useful. There is, however, an *asymptotic* lower bound,

$$\liminf_{n\to\infty} E_\theta([n^{1/2}(T_n - g(\theta))]^2) \ \geq \ g'(\theta)^2/I(\theta), \tag{3.7.1}$$

where $I(\theta) \equiv I_1(\theta)$, which is valid under rather general conditions, without unbiasedness, as will be shown here first for $g(\theta) \equiv \theta$, so $g'(\theta) \equiv 1$, in Theorem 3.7.3, then for more general g in Theorem 3.7.9. First, though, it will be seen that the bound may no longer hold for all θ:

Example. Let $X_1, X_2, \dots,$ be i.i.d. with a normal distribution $N(\mu, 1)$. Then the sample mean $\overline{X} = (X_1 + \cdots + X_n)/n$ is an unbiased estimator of μ which attains the information inequality lower bound for all μ. Let $T_n(X_1, \dots, X_n) \ := \ \overline{X}$ if $|\overline{X}| \geq n^{-1/4}$ and $T_n \ := \ 0$ if $|\overline{X}| < n^{-1/4}$. If $\mu \neq 0$, then the mean-square error of T_n will be asymptotic to $1/n$, in other words $nE_\mu(T_n - \mu)^2 \to 1$ as $n \to \infty$, as for $\overline{X}$. But if $\mu = 0$, the probability that $T_n = 0$ converges to 1, and $nE_0([T_n - 0]^2) \to 0$. In fact, for $\mu = 0$, $P(|\overline{X}| \geq n^{-1/4}) = P(|Z| \geq n^{1/4}) \leq 2e^{-\sqrt{n}/2}$ where Z is a $N(0,1)$ variable (RAP, Lemma 12.1.6), so $T_n = 0$ except with very small probability. Or, one can take $T_n = c\overline{X}$ for $|\overline{X}| < n^{-1/4}$ where $0 < |c| < 1$. Then for $\mu = 0$, $\sqrt{n}T_n$ is asymptotically $N(0, c^2)$ where $0 < c^2 < 1$.

A sequence of estimators which asymptotically attains the information inequality lower bound at a given θ is called "efficient" at that θ. A sequence with a smaller asymptotic variance, like the sequence in the last example at $\mu = 0$, is called "superefficient" at the given θ.

More complicated examples would show that without increasing the asymptotic variance of T_n for any μ, it could be made superefficient at some values of μ in a finite or countable set. It will be shown under some conditions below that as θ varies over an interval in $\mathbb{R}$, (3.7.1) will hold for almost all θ for Lebesgue measure. In other words, superefficiency can occur at most for θ in a set of Lebesgue measure 0. First, the assumptions will be listed.

Let $(X, \mathcal{B})$ be a measurable space (sample space). Suppose that the parameter space Θ is an open interval in $\mathbb{R}$. Let $\{P_\theta,\ \theta \in \Theta\}$, be a family of laws on $(X, \mathcal{B})$, dominated by a σ-finite measure ν on $(X, \mathcal{B})$, with as usual $f(\theta, x) \ := \ (dP_\theta/d\nu)(x)$. Assume that the densities can be chosen so that

(AV-1) There is a set $B \in \mathcal{B}$ such that for all θ, $f(\theta, x) > 0$ for all $x \in B$ and $f(\theta, x) = 0$ for all $x \notin B$.

So, $\{P_\theta,\ \theta \in \Theta\}$ is an equivalent family as defined in Sec. 2.4.

(AV-2) $f(\theta, x)$ is a C^2 function of θ, meaning that its first and second derivatives with respect to θ exist and are continuous at all θ in Θ, for all x.

For the family of laws $P_\theta = U[\theta, \theta+1]$ on $\mathbb{R}$, there exist (unbiased) estimators of θ with mean-square error of order $1/n^2$ (Sec. 2.4, Problem 3). Thus some regularity conditions (equivalence, differentiability in θ) cannot both just be removed.

Let $L(\theta, x) \ := \ \log f(\theta, x)$. Derivatives with respect to θ will be denoted by primes, so that $L'(\theta, x) \ := \ \partial L(\theta, x)/\partial\theta$, etc. Then by (AV-1) and (AV-2), $L(\theta, x)$ is a C^2 function of θ for any $x \in B$. The Fisher information $I(\theta) = E_\theta(L'(\theta, x)^2)$ as defined in Sec. 2.4. Note that if (AV-1) fails and the family is not equivalent, $L(\theta, x)$ can be $-\infty$ on a set of x which has P_θ probability 0 but positive P_ϕ probability for some $\phi \neq \theta$.

(AV-3) For all $\theta \in \Theta$, the Fisher information $I(\theta)$ exists with $0 < I(\theta) < \infty$, and $E_\theta(L'(\theta, x)) \ = \ 0$.

This last equation results if the equation $\int f(\theta, x) d\nu(x) = 1$ can be differentiated under the integral sign, multiplying and dividing by $f(\theta, x)$, as noted in Sec. 2.4.

(AV-4) $E_\theta(L''(\theta, x)) \ = \ -I(\theta)$ for all θ.

The latter equation follows if the differentiation under the integral sign just mentioned can be done also for the second derivative.

(AV-5) For any $\theta_0 \in \Theta$, there is a $\delta > 0$ and a $\mathcal{B}$-measurable function $M(x)$ such that $|L''(\theta, x)| \leq M(x)$ for all $x \in X$ and all θ with $|\theta - \theta_0| < \delta$, and $E_{\theta_0} M(x) < \infty$.

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. variables in X with distribution P_θ. For $n = 1, 2, \ldots,$ let X^n be the set of all ordered n-tuples $(x_1, \ldots, x_n)$ with $x_i \in X$ for each i. Let $\mathcal{B}^n$ be the product σ-algebra in X^n, i.e. the smallest σ-algebra of subsets of X^n making each coordinate projection $(x_1, \ldots, x_n) \ \mapsto x_i$ measurable for $i = 1, \ldots, n$. For any probability measure Q on $(X, \mathcal{B})$, let Q^n be the law on $(X^n, \mathcal{B}^n)$ for which the coordinates are i.i.d. (Q).

Let $\{T_n\}_{n\geq 1}$ be a sequence of estimators (statistics), so that for each n, T_n is measurable from X^n into Θ. It will be assumed that the T_n are consistent estimators of θ, at least in probability, and are asymptotically normal:

(AV-6) For each θ, there is a $v(\theta)$ with $0 < v(\theta) < \infty$ such that as $n \to \infty$, the distribution of $n^{1/2}(T_n - \theta)$ under P_θ^n converges to $N(0, v(\theta))$.

(AV-6) doesn't allow the example near the beginning of this section of superefficient estimation at 0 with $v(0) = 0$. To deal with this we could add to the estimator T_n an independent variable with distribution $N(0, \delta/n)$ for $\delta > 0$ and then let δ decrease to 0.

The asymptotic normality was proved in Section 3.6 under some conditions.

Let $\Pr_\theta$ denote probabilities for the distribution where $X_1, X_2, \ldots$ are i.i.d. with distribution P_θ.

If (AV-6) holds, then

$$v(\theta) \ \leq \ \liminf_{n\to\infty} E_\theta([n^{1/2}(T_n - \theta)]^2),$$

as follows. For each $K < \infty$, the function $\min(x^2, K)$ is bounded and continuous on $\mathbb{R}$, so

$$E_\theta \min(K, n(T_n - \theta)^2) \ \to \ \int \min(K, x^2) dN(0, v(\theta))(x).$$

Thus for each $K < \infty$

$$\liminf_{n\to\infty} E_\theta(n(T_n - \theta)^2) \ \geq \ \int \min(K, x^2) dN(0, v(\theta))(x).$$

Then let $K \to \infty$ and apply monotone convergence.

Thus, assuming asymptotic normality (AV-6), if

$$v(\theta) \ \geq \ 1/I(\theta) \tag{3.7.2}$$

holds, then so does (3.7.1) for $g(\theta) \ \equiv \ \theta$.

The next theorem will give an almost everywhere lower bound on efficiency of estimators of a 1-dimensional parameter. In the proof there will be a relationship between efficiency of estimators and tests (Lemma 3.7.4). Suppose, based on a sample of size n i.i.d. P_θ, we want to test a hypothesis $\theta = \theta_0$ against some alternatives ϕ_n depending on n, with a size that converges to some probability α other than 0 or 1. Let $\phi_n \ := \ \theta_0 + a_n$ for some numbers $a_n \downarrow 0$. To have a specific example in mind, suppose that $P_\theta = N(\theta, 1)$. Then if $a_n = o(1/\sqrt{n})$, the power of the tests will converge to α. If $1/\sqrt{n} = o(a_n)$, the power of the tests will converge to 1. An interesting case is where a_n is of the same order of magnitude as $1/\sqrt{n}$, specifically, $a_n = c/\sqrt{n}$ for some constant $c > 0$, giving a sequence of so-called "Pitman alternatives." The asymptotic power of a sequence of tests of θ_0 against ϕ_n gives a measure of the efficiency of the sequence of tests, called Pitman efficiency. We will not be dealing explicitly any further with Pitman efficiency, but Lemma 3.7.4 and its proof bring tests and the Neyman-Pearson lemma into a proof about efficiency of estimators. Under our assumptions, Lemma 3.7.6 will show that for Pitman alternatives, the power will converge to a limit larger than α.

3.7.3 Theorem. Under assumptions (AV-1) through (AV-6), (3.7.2) holds for almost all θ in the open interval Θ for Lebesgue measure.

Proof. First, the following will be helpful:

3.7.4 Lemma. Under the assumptions of Theorem 3.7.3, if $\theta_0 \in \Theta$ and for $\theta(n) := \theta_0 + n^{-1/2}$,

$$\liminf_{n\to\infty} \Pr_{\theta(n)}\{T_n < \theta(n)\} \ \le\ 1/2,$$

then (3.7.2) holds for $\theta = \theta_0$.

Remark. For a fixed θ, $\Pr_\theta(\sqrt{n}(T_n - \theta) < 0) \to 1/2$ as $n \to \infty$ by (AV-6).

Proof. Consider a likelihood ratio (Neyman-Pearson) test of θ_0 against the simple alternative $\theta(n)$ based on $X^{(n)} := (X_1, \dots, X_n)$. For any $\theta \in \Theta$, n and $X^{(n)}$, let

$$L_n(\theta, X^{(n)}) \ :=\ \textstyle\sum_{i=1}^n L(\theta, X_i), \quad I \ :=\ I(\theta_0), \ \text{ and}$$

$$K_n \ :=\ K_n(\theta_0, X^{(n)}) \ :=\ [L_n(\theta(n), X^{(n)}) - L_n(\theta_0, X^{(n)}) + I/2]/I^{1/2}.$$

Then K_n is a strictly increasing function (an affine function of the logarithm) of the likelihood ratio $R^{(n)}_{\theta(n),\theta_0}$ of $P^n_{\theta(n)}$ to $P^n_{\theta_0}$. In proving Lemma 3.7.4 another fact will be needed:

3.7.5 Lemma. As $n \to \infty$, the distribution of K_n under $P^n_{\theta_0}$ converges to $N(0,1)$.

Proof. By (AV-2) and Taylor's theorem with remainder,

$$L_n(\theta(n), X^{(n)}) \ =\ L_n(\theta_0, X^{(n)}) + n^{-1/2} L_n'(\theta_0, X^{(n)}) + (2n)^{-1} L_n''(\phi_n, X^{(n)})$$

where $\theta_0 < \phi_n < \theta(n)$ and ϕ_n depends on $X^{(n)}$, say $\phi_n := \phi_n(X^{(n)})$. Let $\xi_n := n^{-1}|L_n''(\phi_n, X^{(n)}) - L_n''(\theta_0, X^{(n)})|$.

Claim. $\xi_n \to 0$ a.s. for $P^\infty_{\theta_0}$ as $n \to \infty$.

Proof of Claim. For any $\delta > 0$ and $x \in X^n$, let

$$A(x, \delta) \ :=\ \sup\{|L''(\theta, x) - L''(\theta_0, x)| : \ \theta \in \Theta, \ |\theta - \theta_0| < \delta\}.$$

Let $m(\delta) := E_{\theta_0} A(\cdot, \delta)$. (AV-5) implies that $m(\delta) < \infty$ for δ small enough. (AV-1) and (AV-2) imply that L'' is continuous in θ for all x (in B), and (AV-5) gives dominated convergence, so $m(\delta) \downarrow 0$ as $\delta \downarrow 0$. Given $\varepsilon > 0$, take $\delta > 0$ such that $m(\delta) < \varepsilon$. Then for each $n > \delta^{-2}$ and all $X^{(n)}$, we have

$$\xi_n \ \le\ \frac{1}{n}\sum_{i=1}^n A(X_i, n^{-1/2}) \ \le\ \frac{1}{n}\sum_{i=1}^n A(X_i, \delta)$$

since $|\phi_n - \theta_0| < n^{-1/2}$. So by the strong law of large numbers, $\limsup_{n\to\infty} \xi_n \le \varepsilon$ a.s. Letting $\varepsilon \downarrow 0$ finishes the proof of the Claim. □

From (AV-4) and the strong law of large numbers, it follows that $n^{-1} L_n''(\theta_0, X^{(n)}) \to -I$ a.s. as $n \to \infty$. Then from the definition of K_n and the Taylor expansion,

$$\left| L_n(\theta(n), X^{(n)}) - L_n(\theta_0, X^{(n)}) - \frac{1}{\sqrt{n}} L_n'(\theta_0, X^{(n)}) + \frac{I}{2} \right| \ \le\ \xi_n + \frac{1}{2}\left| \frac{1}{n} L_n''(\theta_0, X^{(n)}) + I \right|$$

which converges to 0 a.s. Thus from the definition of K_n,

$$K_n = I^{-1/2}\left[\frac{1}{\sqrt{n}}L_n'(\theta_0, X^{(n)}) + O\left(\xi_n + \frac{1}{2}\left|\frac{1}{n}L_n''(\theta_0) + I\right|\right)\right]$$

where the O term goes to 0 a.s., so a.s.

$$K_n - (nI)^{-1/2}L_n'(\theta_0, X^{(n)}) \ \rightarrow \ 0.$$

Lemma 3.7.5 then follows from (AV-3), the central limit theorem (RAP, Sec. 9.5), and the fact that if U_n, V_n are random variables such that $|U_n - V_n| \rightarrow 0$ in probability, while the laws of V_n converge to some limit law Q, then U_n have the same limiting distribution (RAP, Lemma 11.9.4). □

Continuing with the proof of Lemma 3.7.4, let Φ be the standard normal distribution function. Let t be a constant and for each n let $C_n := C_{n,t} := \{X^{(n)} : K_n \geq t\}$. The next step is:

3.7.6 Lemma. For any $t \in \mathbb{R}$, $\Pr_{\theta_0}(C_{n,t}) \rightarrow 1 - \Phi(t)$ and $\Pr_{\theta(n)}(C_{n,t}) \rightarrow 1 - \Phi(t - I^{1/2})$ as $n \rightarrow \infty$.

Proof. The first statement is clear from Lemma 3.7.5. For the second, let H_n be the distribution function of K_n under $\Pr_{\theta_0}$. Then

$$1 - \Pr_{\theta(n)}(C_{n,t}) \ = \ \Pr_{\theta(n)}(K_n < t) \ = \ \int_{K_n<t} \exp(L_n(\theta(n), x))d\nu^n(x)$$

$$= \ \int_{K_n<t} \exp[L_n(\theta(n), x) - L_n(\theta_0, x)]dP_{\theta_0}^n(x)$$

$$= \ e^{-I/2}\int_{K_n<t} \exp(I^{1/2}K_n)dP_{\theta_0}^n \ = \ e^{-I/2}\int_{-\infty}^t \exp(I^{1/2}z)dH_n(z).$$

Let $o(1)$ denote (as always) any term that goes to 0 as $n \rightarrow \infty$. It follows from Lemma 3.7.5 and the Helly-Bray theorem (RAP, Theorem 11.1.2), since $z \mapsto \exp(I^{1/2}z)$ is bounded and continuous on $(-\infty, t]$ and Φ is continuous, that

$$\int_{-\infty}^t \exp(I^{1/2}z)dH_n(z) \ \rightarrow \ \int_{-\infty}^t \exp(I^{1/2}z)d\Phi(z).$$

Next, we have

$$e^{-I/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^t \exp\left(-\frac{z^2}{2} + \sqrt{I}z\right)dz = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^t \exp\left(-\frac{(z-\sqrt{I})^2}{2}\right)dz$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t-\sqrt{I}} \exp\left(-\frac{x^2}{2}\right)dx \ = \ \Phi(t - I^{1/2}).$$

Lemma 3.7.6 now follows. □

Now continuing with the proof of Lemma 3.7.4, for each n let

$$D_n \ := \ \{X^{(n)} : \ T_n(X^{(n)}) \geq \theta(n)\}.$$

Take any fixed constant $t > I^{1/2}$ and define $C_n \ := \ C_{n,t}$ as before. Then by Lemma 3.7.6, $\Pr_{\theta(n)}(C_n)$ converges to a limit less than 1/2, and $\limsup_{n\to\infty} \Pr_{\theta(n)}(D_n) \ \geq \ 1/2$ by the hypothesis of Lemma 3.7.4. So there is a sequence of positive integers, say $m_1 < m_2 < \dots$, such that

$$\Pr_{\theta(n)}(D_n) \ > \ \Pr_{\theta(n)}(C_n) \ \text{ for } \ n = m_1, m_2, \dots .$$

For each n, consider C_n and D_n as critical regions for testing the hypothesis θ_0 against the alternative $\theta(n)$. Since by the Neyman-Pearson Lemma (Theorem 1.1.3), C_n is an admissible critical region, by the statement just before Lemma 3.7.5, but D_n has larger power, it must also have larger size:

$$\Pr_{\theta_0}(D_n) > \Pr_{\theta_0}(C_n) \ \text{ for } \ n = m_1, m_2, \dots .$$

By (AV-6) and the definitions of $\theta(n)$ and D_n, recall that by Lemma 3.7.6, $\Pr_{\theta_0}(C_n) \to 1 - \Phi(t)$. Let $v \ := \ v(\theta_0)$. Then

$$\Pr_{\theta_0}(D_n) = P^n_{\theta_0}(T_n \geq \theta(n) = \theta_0 + n^{-1/2}) = P^n_{\theta_0}(\sqrt{n}(T_n - \theta_0) \geq 1),$$

which by (AV-6) converges as $n \to \infty$ to

$$N(0, v(\theta_0))([1, +\infty)) = N(0,1)([1/\sqrt{v}, +\infty)) = 1 - \Phi(v^{-1/2}) \ \geq \ 1 - \Phi(t)$$

(via $n = m_j$), so $v^{-1/2} \leq t$. Letting $t \downarrow I^{1/2}$, it follows that (3.7.2) holds for $\theta = \theta_0$, proving Lemma 3.7.4. □

Now continuing with the proof of Theorem 3.7.3, for any real θ let $f_n(\theta) \ := \ |\Pr_\theta(T_n < \theta) - \frac{1}{2}|$ if $\theta \in \Theta$, or 0 otherwise. To see that this is a measurable function of θ, it is enough to show that $\theta \mapsto \Pr_\theta(T_n < \theta)$ is a measurable function of $\theta \in \Theta$, or that $(\theta, x) \mapsto \Pr_\theta(T_n < x)$ is jointly measurable. In x, the function is nondecreasing and left-continuous. For each x and each positive integer m let $j(x,m)$ be the largest integer with $j(x,m) \leq 2^m x$. Then $j(x,m)/2^m \uparrow x$ and $\Pr_\theta(T_n < j(x,m)/2^m) \uparrow \Pr_\theta(T_n < x)$ for all x and θ. Each $j(\cdot, m)$ is clearly a measurable function of x. So it will be enough to show that $\theta \mapsto \Pr_\theta(T_n < y)$ is measurable for each fixed y. This is a special case of the property that the family of laws $\Pr_\theta$ is a measurable family as defined in Sec. 1.2. To see that it is in this case, we have that $f(\theta, x)$ is continuous in θ by (AV-2). Thus by Fatou's Lemma, for any measurable set $A \subset X^n$, and any convergent sequence $\theta_k \to \theta$ in Θ,

$$P^n_\theta(A) \ := \ \int_A \Pi^n_{j=1} f(\theta, X_j) d\nu^n(X^{(n)}) \ \leq \ \liminf_{k\to\infty} \int_A \Pi^n_{j=1} f(\theta_k, X_j) d\nu^n(X^{(n)}).$$

So $\theta \mapsto P^n_\theta(A)$ is a lower semicontinuous function: $\{\theta : \ P^n_\theta(A) \leq c\}$ is closed for any real c. This implies that $\theta \mapsto P^n_\theta(A)$ is Borel measurable as desired.

It follows from (AV-6) that for each θ, $\Pr_\theta(T_n < \theta) \to 1/2$ as $n \to \infty$. So, $0 \leq f_n(\theta) \leq 1/2$ and $f_n(\theta) \to 0$ as $n \to \infty$ for all θ. Let $g_n(\theta) := f_n(\theta + n^{-1/2})$ for any $\theta \in \Theta$. Then $0 \leq g_n \leq 1/2$ also. We next need another lemma:

3.7.7 Lemma. There is a set N of Lebesgue measure 0 and a sequence $n_1 < n_2 < \ldots$ such that for any $\theta \in \Theta$ with $\theta \notin N$, $\lim_{r\to\infty} g_{n_r}(\theta) = 0$.

Proof. $\int_{-\infty}^{\infty} g_n(\theta)d\Phi(\theta) = \int_{-\infty}^{\infty} f_n(\theta + n^{-1/2})d\Phi(\theta)$

$$= \int_{-\infty}^{\infty} f_n(\eta)\exp\left(-\frac{1}{2n} + \frac{\eta}{\sqrt{n}}\right)d\Phi(\eta) \to 0$$

as $n \to \infty$ by dominated convergence since $\exp(n^{-1/2}\eta) \leq e^\eta + 1$ for all η. Convergence in L^1 implies convergence in probability, which implies that there is a subsequence $g_{n_r} \to 0$ almost surely for $N(0,1)$ (RAP, Theorem 9.2.1), and so almost everywhere for Lebesgue measure. □

The function $\theta \mapsto I(\theta)$ is continuous since $L''(\cdot, x)$ is continuous by (AV-1) and (AV-2), we can apply (AV-4), and (AV-5) provides domination for the dominated convergence theorem. Thus $I(\cdot)$ is Borel measurable. To show that (3.7.2) holds for Lebesgue almost all θ it may be good to know that (3.7.2) holds for θ in a measurable set, which will follow from the next lemma. This lemma will also be applied in the multidimensional case. (AV-6) implies that $\liminf_{n\to\infty} nE((T_n - \theta)^2) \geq v(\theta)$, but the lim inf could be larger than $v(\theta)$, so a different approach to it is needed.

For any distribution function F and $0 < p < 1$, the p quantile of F is defined by $F^{\leftarrow}(p) := \inf\{x : F(x) \geq p\}$. Then $F^{\leftarrow}$ is a non-decreasing, left-continuous function of p.

3.7.8 Lemma. Under the assumptions (AV-2) and (AV-6), $v(\theta)$ in (AV-6) is a measurable function of θ.

Proof. Let $F_{n,\theta}(t) := \Pr_\theta(T_n \leq t)$ for $-\infty < t < \infty$. In the proof of Theorem 3.7.3, before Lemma 3.7.7, it is shown from (AV-2) that $(\theta, u) \mapsto \Pr_\theta(T_n < u)$ is jointly measurable. Taking $u = t + \frac{1}{k} \downarrow t$, it follows that $(\theta, t) \mapsto F_{n,\theta}(t)$ is jointly measurable. Restricting this function to t rational, we have for $0 < p < 1$ that

$$\theta \mapsto F_{n,\theta}^{\leftarrow}(p) = \inf\{q \in \mathbb{Q} : F_{n,\theta}(q) \geq p\}$$

is measurable. Then (AV-6) implies that $n(F_{n,\theta}^{\leftarrow}(3/4) - F_{n,\theta}^{\leftarrow}(1/2))^2 \to \Phi^{\leftarrow}(3/4)^2 v(\theta)$ as $n \to \infty$ where Φ is the standard normal distribution function. (Note that $\Phi^{\leftarrow}(1/2) = 0$. Here $\Phi^{\leftarrow}(3/4)^2 \doteq 0.455$.) The Lemma follows. □

Lemma 3.7.7 implies that $\liminf_{n\to\infty} g_n(\theta) = 0$ for almost all θ. By the definitions, this implies that the hypothesis, and so the conclusion, of Lemma 3.7.4 for θ in place of θ_0 holds for almost all θ, which finishes the proof of Theorem 3.7.3. □

Next, Theorem 3.7.3 will be extended to estimators of functions $g(\theta)$, by the delta-method. The factor $g'(\theta)^2$ is familiar from information inequalities (Section 2.4). Note that (AV-1) through (AV-5) don't mention any estimators T_n.

3.7.9 Theorem. Assume (AV-1) through (AV-5). Let g be a C^1 function: $\Theta \to \mathbb{R}$. Suppose that for each $\theta \in \Theta$, there is a $w(\theta) \geq 0$ such that for each θ with $g'(\theta) \neq 0$, $0 < w(\theta) < \infty$ and the distribution of $\sqrt{n}(T_n - g(\theta))$ under P_θ^n converges to $N(0, w(\theta))$. Then for Lebesgue almost all $\theta \in \Theta$, $w(\theta) \geq g'(\theta)^2/I(\theta)$.

Proof. For all $\theta \in \Theta$ such that $g'(\theta) = 0$, the inequality is trivial. Since g is C^1, the set where $g' \neq 0$ is open and thus a countable disjoint union of open intervals. Thus replacing Θ by a smaller interval as needed, we can assume that on the interval Θ, $g'(\theta) \neq 0$, and specifically $g'(\theta) > 0$. Then g is one-to-one and has a C^1 inverse g^{-1}. For a given $\theta \in \Theta$, we have

$$g^{-1}(g(\theta) + \phi) \;=\; \theta + (g^{-1})'(g(\theta))\phi + o(|\phi|)$$

as $\phi \to 0$. We have $T_n \to g(\theta)$ in probability for $\Pr_\theta$ as $n \to \infty$. Also, $(g^{-1})'(g(\theta)) \equiv 1/g'(\theta)$. Thus as $n \to \infty$,

$$\sqrt{n}(g^{-1}(T_n) - \theta) \;=\; \sqrt{n}(T_n - g(\theta))/g'(\theta) + o_p(1),$$

so the distribution of $\sqrt{n}(g^{-1}(T_n) - \theta)$ converges to $N(0, w(\theta)/g'(\theta)^2)$, where one can use e.g. RAP, Lemma 11.9.4. Thus (AV-6) holds for the estimators $g^{-1}(T_n)$ of θ and $v(\theta) := w(\theta)/g'(\theta)^2$. Then by Theorem 3.7.3, $w(\theta)/g'(\theta)^2 \geq 1/I(\theta)$ for Lebesgue almost all $\theta \in \Theta$, which proves the theorem. □

If A is a $k \times m$ matrix, then A' denotes its transpose, with $(A')_{ij} := A_{ji}$ for $i = 1, \ldots, m$, $j = 1, \ldots, k$. In particular, if x is a row vector $(x_1, \ldots, x_m)$ then x' is the corresponding column vector, and vice versa. In fact, elements of $\mathbb{R}^m$ will usually be taken as column vectors y, so that y' is the corresponding row vector. Matrix multiplication is written by juxtaposition. Thus for $x, y \in \mathbb{R}^m$, $x'y = \sum_{j=1}^m x_j y_j$ is the usual dot product $x \cdot y$. If $x, y \in \mathbb{R}^m$ and C is an $m \times m$ matrix, then $x'Cy$ is the number $\sum_{i,j=1}^m C_{ij} x_i y_j$.

The Fisher information for a single parameter extends to the Fisher information matrix for several parameters, defined as follows. Let Θ be an open set in $\mathbb{R}^m$. For $\theta := (\theta_1, \ldots, \theta_m)$, let

$$I(\theta)_{ij} \;:=\; E_\theta\left(\frac{\partial L(\theta, x)}{\partial \theta_i}\frac{\partial L(\theta, x)}{\partial \theta_j}\right) \tag{3.7.10}$$

if the partial derivatives exist and have finite variances. In passing, let's note here that

$$ds^2 \;:=\; \sum_{i,j=1}^m I(\theta)_{ij} d\theta_i d\theta_j$$

defines a Riemannian metric on the parameter space Θ which doesn't depend on the choice of parametrization. As in the development between (2.4.2) and (2.4.3), alternate forms of I_{ij} are

$$I(\theta)_{ij} \;=\; E_\theta\left(\left(\frac{\partial R_{\phi,\theta}}{\partial \phi_i}\frac{\partial R_{\phi,\theta}}{\partial \phi_j}\right)\Big|_{\phi=\theta}\right) \;=\; \int \frac{\partial f(\theta, x)}{\partial \theta_i}\frac{\partial f(\theta, x)}{\partial \theta_j}\frac{1}{f(\theta, x)} d\nu(x).$$

Bahadur (1964, p. 1550) pointed out that Theorem 3.7.3 extends to multidimensional parameter spaces. Such an extension might be considered non-trivial in view of the Stein phenomenon and James-Stein estimators (Section 2.7). But it turns out that the Stein phenomenon doesn't affect the asymptotic efficiency as $n \to \infty$. First, multidimensional forms of the assumptions (AV-1) through (AV-6) will be given.

(AC-1) Let Θ be an open set in a Euclidean space $\mathbb{R}^m$ and let $(X, \mathcal{B})$ be a sample space. Let $\{P_\theta,\ \theta \in \Theta\}$ be an equivalent family of laws on $(X, \mathcal{B})$, and let ν be an equivalent law, e.g. $\nu = P_\phi$ for some fixed ϕ, with $f(\theta, x) \ := \ (dP_\theta/d\nu)(x)$, and $f(\theta, x) > 0$ for all θ and x.

(AC-2) For all x, $f(\theta, x)$ is C^2 with respect to θ, so that the first and second partial derivatives of f with respect to θ exist and are continuous at all $\theta \in \Theta$ for all x.

(AC-3) Let $L(\theta, x) \ := \ \log f(\theta, x)$. For each $\theta \in \Theta$, the Fisher information matrix $I(\theta)$ as defined by (3.7.10) exists and is strictly positive definite. Also, $E_\theta(\nabla_\theta L(\theta, x)) = 0$ for the gradient of L.

(AC-4) $\{E_\theta \partial^2 L(\theta, x)/\partial\theta_i \partial\theta_j\}_{i,j=1}^m \ = \ -I(\theta)$ for all $\theta \in \Theta$.

(AC-5) (AV-5) holds for each of the second partial derivatives $\partial^2 L(\theta, x)/\partial\theta_i\partial\theta_j$ in place of L''.

(AC-6) T_n are estimators of $\theta \in \Theta$ such that for each θ, the distribution of $n^{1/2}(T_n - \theta)$ under $\Pr_\theta$ converges as $n \to \infty$ to some multivariate normal law $N(0, v(\theta))$ where $v(\theta)$ is a nonnegative definite symmetric matrix.

3.7.11 Theorem. Assume (AC-1) through (AC-6). Then for Lebesgue almost all $\theta \in \Theta$, $v(\theta) - I^{-1}(\theta)$ is nonnegative definite. Thus, $v(\theta)$ is positive definite.

Proof. We first prove a lemma.

3.7.12 Lemma. Let C be a symmetric, positive definite real $m \times m$ matrix and $0 \neq \eta \in \mathbb{R}^m$. Then

$$\min\{\phi' C\phi : \ \phi \in \mathbb{R}^m,\ \eta'\phi = 1\} \ = \ 1/(\eta' C^{-1}\eta)$$

and is attained for $\phi = C^{-1}\eta/(\eta' C^{-1}\eta)$.

Proof. Since C is positive definite, $\phi' C\phi \to +\infty$ as $|\phi| \to +\infty$. Thus by the Lagrange multiplier method, Appendix F, Theorem F.1 and Proposition F.2, $\phi \mapsto \phi' C\phi$ attains its minimum on the hyperplane $\{\phi : \ \eta'\phi = 1\}$ at some ϕ, and for each such ϕ, there is a $\lambda \in \mathbb{R}$ such that $\nabla_{\phi,\lambda}[\phi' C\phi + \lambda(\eta'\phi - 1)] = 0$. Thus $2C\phi + \lambda\eta = 0$, so $\phi = \lambda C^{-1}\eta/2$, $1 = \lambda\eta' C^{-1}\eta/2$, $\lambda = 2/(\eta' C^{-1}\eta)$, and $\phi = C^{-1}\eta/(\eta' C^{-1}\eta)$ as stated, so this ϕ gives the unique minimum. The value of the minimum is

$$\phi' C\phi \ = \ (C^{-1}\eta)'\eta/(\eta' C^{-1}\eta)^2 \ = \ \eta' C^{-1}\eta/(\eta' C^{-1}\eta)^2 \ = \ 1/(\eta' C^{-1}\eta),$$

proving the Lemma. □

Now continuing with the proof of Theorem 3.7.11, let's first see what we can infer from Theorem 3.7.3 about families with a 1-dimensional parameter. Let $\zeta \in \mathbb{R}^m$ and

$0 \neq \phi \in \mathbb{R}^m$. Since Θ is open in $\mathbb{R}^m$, the real t such that $\theta := \zeta + t\phi \in \Theta$ form an open set in $\mathbb{R}$, which is a countable union of open intervals. Consider the family with 1-dimensional parameter t for t in such an interval, $Q_t := Q_{\zeta,\phi;t}$ with densities $dQ_t(x)/dv = f_{\zeta,\phi}(t,x) := f(\zeta + t\phi, x)$ for $\zeta + t\phi \in \Theta$. Let $\eta \in \mathbb{R}^m$ satisfy $\eta'\phi = 1$. Then $\eta'(\zeta + t\phi) \equiv \eta'\zeta + t$. Under $\Pr_\theta$, $\eta'\sqrt{n}[T_n - (\zeta + t\phi)] = \sqrt{n}[\eta'(T_n - \zeta) - t]$ has distribution converging as $n \to \infty$ to $N(0, \eta' v(\zeta + t\phi)\eta)$. Let $K_{\zeta,\phi}(t)$ be the Fisher information of the family Q_t, a real-valued function of t. Taking the gradient $\nabla_\theta f(\theta, x)$, then letting $\theta = \zeta + t\phi$, we get

$$K_{\zeta,\phi}(t) = \phi' I(\theta)\phi \tag{3.7.13}$$

because

$$K_{\zeta,\phi}(t) = E_{\zeta+t\phi}[(\partial \log f(\zeta + t\phi, x)/\partial t)^2]$$

$$= E_{\zeta+t\phi}\left[\left(\frac{\phi'\nabla_\theta f(\zeta + t\phi, x)}{f(\zeta + t\phi, x)}\right)^2\right] = \phi' I(\zeta + t\phi)\phi.$$

By Theorem 3.7.3 applied to the family Q_t and to $\eta'(T_n - \zeta)$ as a sequence of estimators of t, we get for $\theta := \zeta + t\phi$

$$\eta' v(\theta)\eta \geq 1/[\phi' I(\theta)\phi] \tag{3.7.14}$$

for Lebesgue almost all t such that $\theta = \zeta + t\phi \in \Theta$.

Now, supposing heuristically for the moment that (3.7.14) holds for all ϕ such that $\eta'\phi = 1$, or at least for a countable dense set of such ϕ, we can take the supremum of the right side of (3.7.14) by taking the infimum of the denominator, which by Lemma 3.7.12 gives

$$\eta' v(\theta)\eta \geq \sup\{1/[\phi' I(\theta)\phi] : \eta'\phi = 1\} = \eta' I(\theta)^{-1}\eta. \tag{3.7.15}$$

This is the conclusion of Theorem 3.7.11 if we can prove it for λ^m-almost all θ.

Returning to the rigorous proof, let D_1 be a countable dense set in the unit sphere $S^{m-1} := \{y \in \mathbb{R}^m : |y| = 1\}$. For each $\eta \in D_1$, let E_η be a countable dense set in the hyperplane $\{\phi \in \mathbb{R}^m : \eta'\phi = 1\}$, the same relation between η and ϕ as in (3.7.14) above. For each $\eta \in D_1$, $\phi \in E_\eta$ and $\zeta \in \phi^\perp := \{\zeta \in \mathbb{R}^m : \zeta'\phi = 0\}$ such that $\theta = \zeta + t\phi \in \Theta$ for some real t, we get a one-parameter family Q_t as defined above.

Now $\theta \mapsto I(\theta)$ is continuous from Θ into $\mathbb{R}^{m^2}$ since f is C^2 in θ by (AC-2) and $f > 0$ by (AC-1), so $\log f$ is C^2 in θ. We use the form of $I(\theta)$ given by (AC-4), and we have local domination by (AC-5), so the dominated convergence theorem applies to give continuity of $I(\cdot)$.

Applying Lemma 3.7.8 to $\eta' v(\theta)\eta$ for suitable unit vectors η, namely the basis vectors $e_j := \{\delta_{ij}\}_{i=1}^m$ and $(e_i + e_j)/\sqrt{2}$, $i \neq j$, which we can assume are all in D_1, we see that the matrix elements of $v(\theta)$ and thus $v(\theta)$ itself are measurable functions of θ. Thus the set of all $\theta \in \Theta$ for which (3.7.14) holds for fixed η and ϕ is a measurable set. For ϕ fixed and ζ varying in $\phi^\perp$, $t \in \mathbb{R}$, by the Tonelli-Fubini theorem, we get that (3.7.14) holds for Lebesgue almost all $\theta \in \Theta$ for given η and ϕ.

Taking a countable union of sets of Lebesgue measure 0, we get that for any $\eta \neq 0$, for Lebesgue almost all $\theta \in \Theta$, we get (3.7.15) for $\phi \in E_\eta$, which suffices since E_η is dense in $\{\phi : \ \phi'\eta = 1\}$, namely

$$\eta' v(\theta)\eta \ \geq \ \sup\{1/[\phi' I(\theta)\phi] : \ \phi \in E_\eta\} \ = \ \eta' I(\theta)^{-1}\eta$$

by Lemma 3.7.12. Taking another countable union over $\eta \in D_1$, we get that $v(\theta) - I(\theta)^{-1}$ is nonnegative definite for Lebesgue almost all $\theta \in \Theta$. When this occurs, since $I(\theta)^{-1}$ is strictly positive definite, so is $v(\theta)$, proving the Theorem. □

NOTES

The idea that n times the variances of consistent and asymptotically normal estimators T_n should be asymptotically at least $1/I(\theta)$ goes back to Fisher. J. L. Hodges found the example as given after (3.7.1) where $1/I(\theta)$ is asymptotically attained for all $\theta \neq 0$ and the variance is smaller (asymptotically vanishing) for $\theta = 0$, a phenomenon called "superefficiency." The example was published in LeCam (1953) along with the first statement and proof that $n \cdot \mathrm{var}(T_n)$ is asymptotically bounded below by $1/I(\theta)$ for Lebesgue almost all θ. Bahadur (1964) stated and proved the version of that fact given in this section, Theorem 3.7.3. Lehmann (1983, Theorem 6.1.1) gave a statement, but not proof, of Bahadur's theorem. Theorem 3.7.3 benefited from both Bahadur's and Lehmann's expositions. Lehmann points out the step from Theorem 3.7.3 to Theorem 3.7.9 by the delta-method. For another proof of the multidimensional Theorem 3.7.11, see van der Vaart (1998), Theorem 8.9.

REFERENCES

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

Le Cam, Lucien (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. Calif. Publ. in Statist.* **1**, 277-330.

Lehmann, Erich (1983). *Theory of Point Estimation.* Wiley, New York.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.