March 22, 2003

2.4 Lower bounds on mean squared errors: information inequalities. When we consider a parametrized family $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ of laws, we will always assume that $P_\theta \neq P_\phi$ for $\theta \neq \phi$.

If $T = T(X_1, ..., X_n)$ is a statistic and $X_1, \ldots, X_n$ are i.i.d. (P_θ) then (as before) we let

$$E_\theta T \ := \ \int \cdots \int T(x_1, \ldots, x_n) dP_\theta(x_1) \cdots dP_\theta(x_n),$$

or if T is a function on the basic sample space ($n = 1$) then $E_\theta T = \int T(x) dP_\theta(x)$. Correspondingly, variances and covariances of real-valued statistics T and U are defined by

$$\mathrm{var}_\theta T \ := \ E_\theta(T^2) - (E_\theta T)^2, \quad \mathrm{cov}_\theta(T, U) \ := \ E_\theta(TU) - (E_\theta T)(E_\theta U)$$

when the integrals converge, with $\mathrm{var}_\theta T \ := \ +\infty$ if $E_\theta(T^2) = +\infty$. For squared-error loss $L(\theta, T) = (T - g(\theta))^2$, if T is an unbiased estimator of $g(\theta)$, then the mean squared-error loss equals $\mathrm{var}_\theta T$. On the other hand, trivial constant estimators as mentioned in the last section will have variance 0 without being good estimators except for special parameter values. Inequalities of the type in this section were first found for unbiased estimators, but there will be a form (Theorem 2.4.12) which applies to estimators that may have a bias.

We are looking for lower bounds for variances of unbiased estimators T of functions $g(\theta)$. Suppose first that T is an unbiased estimator of θ. Then for any constants a and b, $a+bT$ is an unbiased estimator of $h(\theta) = a+b\theta$, with $\mathrm{var}_\theta(a+bT) = b^2 \mathrm{var}_\theta T$. This variance doesn't depend on a and is proportional to b^2 where b is the derivative of h (everywhere, in this simple case). Or more generally, if T is an unbiased estimator of $g(\theta)$ then $a + bT$ is an unbiased estimator of $a + bg(\theta)$ and $\mathrm{var}_\theta(a+bT) = b^2 \mathrm{var}_\theta T$. Thus it seems natural that (lower) bounds for the variances of unbiased estimators should be proportional to $g'(\theta)^2$, as they will be.

Also, recall that for n i.i.d. observations, the sample mean $\overline{X}$ as an estimator for an unknown mean μ is unbiased and has a variance equal to σ^2/n where σ^2 is the variance of one observation. So we can anticipate that lower bounds for the variance of an unbiased estimator of $g(\theta)$ based on n i.i.d. observations should be of the form $u(\theta)g'(\theta)^2/n$ for some function $u(\theta)$. This will also turn out to be true (Theorem 2.4.10), so we have to find suitable functions $u(\theta)$, which are most often written as $1/I(\theta)$ where $I(\cdot)$ is the so-called Fisher information, to be defined.

A family $\mathcal{P}$ of probability measures will be called *equivalent* if any two laws P and Q in the family are equivalent, in other words for any measurable set B, $P(B) = 0$ if and only if $Q(B) = 0$. Then $\{P_\theta,\ \theta \in \Theta\}$ will be equivalent if for some σ-finite measure v, P_θ is equivalent to v for all $\theta \in \Theta$. Conversely, if $\mathcal{P}$ is equivalent, we can take any member of $\mathcal{P}$ as v. Let the density (Radon-Nikodym derivative) be $f(\theta, x) \ := \ (dP_\theta/dv)(x)$. Then $f(\theta, x) > 0$ for v-almost all x and for P_ϕ-almost all x for each $\phi \in \Theta$. The likelihood ratio $R_{\phi,\theta} \ := \ R_{P_\phi/P_\theta} \ = \ f(\phi, x)/f(\theta, x)$ will be defined, with $0 < R_{\phi,\theta} < \infty$ for almost all x in the same sense; $0/0$ will be defined as 0 in this case. Here is a first lower bound on variances of unbiased estimators. Note that in it, there is no restriction on the parameter space Θ, which could be an arbitrary set.

2.4.1 Theorem. Suppose T is an unbiased estimator of a real function $g(\theta)$ for an equivalent family $\{P_\theta,\ \theta \in \Theta\}$. Then

$$\mathrm{var}_\theta T \ \geq \ \sup\{(g(\phi) - g(\theta))^2/\mathrm{var}_\theta R_{\phi,\theta} :\ \phi \in \Theta,\ \phi \neq \theta\}.$$

Note. The conclusion of the theorem holds trivially if $\mathrm{var}_\theta T = +\infty$ or if $\mathrm{var}_\theta R_{\phi,\theta} \ \equiv \ +\infty$ for all $\phi \neq \theta$. So the theorem has content if and only if both $E_\theta(T^2) < \infty$ and $\mathrm{var}_\theta R_{\phi,\theta} < \infty$ for at least one value of $\phi \neq \theta$. For $\phi \neq \theta$, since $P_\theta \neq P_\phi$, $R_{\phi,\theta}$ is non-constant with respect to P_θ, so its variance is non-zero.

Proof. Since T is unbiased, $\int T(x) f(\phi, x) dv(x) = g(\phi)$ for all ϕ, and

$$\int T(x) \frac{f(\phi, x) - f(\theta, x)}{f(\theta, x)} f(\theta, x) dv(x) \ = \ g(\phi) - g(\theta),$$

$$\begin{aligned} \mathrm{cov}_\theta(T, R_{\phi,\theta}) \ &= \ \int (T(x) - g(\theta)) \left(\frac{f(\phi, x)}{f(\theta, x)} - 1 \right) dP_\theta(x) \\ &= \ g(\phi) - 2g(\theta) + g(\theta) \ = \ g(\phi) - g(\theta). \end{aligned}$$

Then by the Cauchy-Bunyakovsky-Schwarz inequality (RAP, 5.1.4),

$$\mathrm{var}_\theta T \ \geq \ (g(\phi) - g(\theta))^2/\mathrm{var}_\theta R_{\phi,\theta},$$

where $\mathrm{var}_\theta R_{\phi,\theta} > 0$ for $\theta \neq \phi$; then take the supremum over $\phi \neq \theta$. □

In the rest of this section, Θ is an open interval in $\mathbb{R}$. Often, the function $g(\theta)$ to be estimated is just θ. Then g has the derivative $g' \ \equiv \ 1$, so that all the further facts in this section in terms of $g'(\theta)$ simplify.

2.4.2 Theorem. Assume that T is an unbiased estimator of $g(\theta)$ for an equivalent family $\{P_\theta,\ \theta \in \Theta\}$, Θ is an open interval in $\mathbb{R}$, g has a derivative at θ and as $\phi \to \theta$, for some $J(\theta)$, $(\mathrm{var}_\theta R_{\phi,\theta})/(\phi - \theta)^2 \to J(\theta)$. Then if $g'(\theta) \neq 0$ or $J(\theta) > 0$,

$$\mathrm{var}_\theta T \ \geq \ g'(\theta)^2/J(\theta).$$

Proof. In Theorem 2.4.1, divide numerator and denominator by $(\phi - \theta)^2$ and let $\phi \to \theta$. If $J(\theta) = 0$, $\mathrm{var}_\theta T$ must be $+\infty$, so the conclusion follows. □

Note that for any $\phi \neq \theta$, $\mathrm{var}_\theta R_{\phi,\theta}/(\phi - \theta)^2 \ = \ E_\theta(((R_{\phi,\theta} - 1)/(\phi - \theta))^2)$ and $R_{\theta,\theta} \ \equiv \ 1$. Suppose that in $J(\theta)$, the limit as $\phi \to \theta$ can be interchanged with the integral E_θ, so that the integrands converge. Their limit is then the square of a partial derivative, $(\partial R_{\phi,\theta}/\partial\phi)|_{\phi=\theta})^2$, which can also be written as

$$\left(\frac{\partial f(\theta, x)/\partial\theta}{f(\theta, x)} \right)^2 \ = \ \left(\frac{\partial \log f(\theta, x)}{\partial\theta} \right)^2.$$

The quantity $\partial \log f(\theta, x)/\partial\theta$ is known as the *score function.* If the derivatives in the last display exist for almost all x, then the quantity

$$I(\theta) \ := \ E_\theta((\partial \log f(\theta, x)/\partial\theta)^2) \ = \ \int (\partial f(\theta, x)/\partial\theta)^2/f(\theta, x) \ dv(x)$$

is called the *information* of the family $\{P_\theta\}$ at θ. It is by no means the same as the "information" studied in information theory. $I(\theta)$ is often called the *Fisher* information. Fisher made good use of it, but it was originally due to Edgeworth, see the Notes.

A famous inequality, $\mathrm{var}_\theta T \ \geq \ g'(\theta)^2/I(\theta)$, then follows from the interchange of limits. One set of sufficient conditions for the interchange will imply that the identities

$$1 \ \equiv \ \int f(\theta, x) dv(x) \quad \text{and} \quad g(\theta) \ \equiv \ \int T(x) f(\theta, x) dv(x)$$

can be differentiated with respect to θ under the integral sign, as follows:

2.4.3 Information inequality. Let T be an unbiased estimator of a function $g(\cdot)$ on an open interval Θ for an equivalent family $\{P_\theta, \ \theta \in \Theta\}$. For a given value of θ, assume that $\partial f(\theta, x)/\partial\theta$ exists for almost all x, $I(\theta) > 0$ and that

(2.4.4) $$\int (|T(x)| + 1) \left| \frac{\partial f(\theta, x)}{\partial\theta} \right| dv(x) \ < \infty,$$

(2.4.5) $$0 \ = \ \int \frac{\partial f(\theta, x)}{\partial\theta} \ dv(x), \quad \text{and}$$

(2.4.6) $$g'(\theta) \ = \ \int T(x) \frac{\partial f(\theta, x)}{\partial\theta} \ dv(x).$$

Then

$$\mathrm{var}_\theta T \ \geq \ g'(\theta)^2/I(\theta).$$

Notes. The information inequality has been called the Cramér-Rao inequality, but Fréchet found it earlier and Darmois also played a part (see the Notes). Existence and finiteness of the Lebesgue integrals in both (2.4.5) and (2.4.6) is equivalent to (2.4.4).

Proof. Multiplying (2.4.5) by $g(\theta)$ and subtracting from (2.4.6) gives

$$g'(\theta) \ = \ E_\theta((T(x) - g(\theta))\partial \log f(\theta, x)/\partial\theta).$$

If $\mathrm{var}_\theta T = +\infty$ or $I(\theta) = +\infty$, the inequality holds trivially since $g'(\theta)$ is finite. If $\mathrm{var}_\theta T$ and $I(\theta)$ are both finite, then the Cauchy-Bunyakovsky-Schwarz inequality can be applied as in the proof of Theorem 2.4.1 to get $g'(\theta)^2 \ \leq \ I(\theta)\mathrm{var}_\theta T$. □

Most books state the information inequality under assumptions such as those of Theorem 2.4.3. Exchanging differentiation with an integral as in (2.4.5) and (2.4.6) may seem a plausible and reasonable kind of hypothesis. But an example will be given below (Proposition 2.4.13) showing that assumption (2.4.6) may not hold even when (2.4.4) does, and where each derivative and integral in (2.4.6) is well-defined and finite. So let's see how (2.4.4) can be strengthened enough to imply (2.4.5) and (2.4.6), by way of the notion of uniform integrability (RAP, Section 10.3). A set $\mathcal{F}$ of integrable functions on a probability space $(X, \mathcal{S}, \mu)$ is *uniformly integrable* iff

$$\lim_{M\to\infty} \sup\{E|f|1_{\{|f|>M\}} :\ f \in \mathcal{F}\} \ = \ 0.$$

This will hold if (but not only if) there is an integrable function g with $|f| \leq g$ for all $f \in \mathcal{F}$.

2.4.7 Theorem. Assume that for a given θ, $\partial f(\theta, x)/\partial\theta$ exists for almost all x and there is a $\delta > 0$ such that the functions

$$(|T(x)| + 1)(f(\phi, x) - f(\theta, x))/(\phi - \theta) \ \text{ for } \ 0 < |\phi - \theta| < \delta$$

are uniformly integrable for v, or equivalently the functions

$$(|T(x)| + 1)(R_{\phi,\theta} - 1)/(\phi - \theta) \ \text{ for } \ 0 < |\phi - \theta| < \delta$$

are uniformly integrable with respect to P_θ. Then (2.4.4), (2.4.5) and (2.4.6) all hold.

Proof. The conditions follow from convergence of integrals of pointwise convergent, uniformly integrable functions (RAP, Theorem 10.3.6). □

Theorems 2.4.3 and 2.4.7 have been stated for one unbiased estimator T, but the information inequality has usually been stated as applying to all unbiased estimators, with hypotheses (2.4.4) and (2.4.6) assumed for all such estimators of $g(\theta)$. If attention is restricted to estimators measurable for a Lehmann-Scheffé sufficient σ-algebra, the unbiased estimator (if it exists) is unique by Theorem 2.3.5 and the results of this section are not needed to choose between unbiased estimators (although Theorem 2.4.12 could be helpful for other estimators). Otherwise, there can be a large family of different unbiased estimators of $g(\theta)$. So it may not really be clear what it means, in terms of the family of laws P_θ, that (2.4.4) and (2.4.6) hold for all unbiased estimators of g. An alternate sufficient condition will be stated just in terms of g and the family $\{P_\theta,\ \theta \in \Theta\}$:

2.4.8 Theorem. The information inequality holds for a given θ for every unbiased estimator T of $g(\cdot)$, if: Θ is an open interval in $\mathbb{R}$, $\{P_\theta,\ \theta \in \Theta\}$ is an equivalent family, g has a non-zero derivative at θ, $\partial f(\theta, x)/\partial\theta$ exists for almost all x, and there is a $\delta > 0$ such that the set of functions $((R_{\phi,\theta} - 1)/(\phi - \theta))^2$ for $0 < |\phi - \theta| < \delta$ is uniformly integrable for P_θ.

Proof. Applying Theorem 2.4.2, we have

$$J(\theta) \ = \ \lim_{\phi\to\theta}(\mathrm{var}_\theta R_{\phi,\theta})/(\phi - \theta)^2 \ = \ \lim_{\phi\to\theta} E_\theta([(R_{\phi,\theta} - 1)/(\phi - \theta)]^2)$$

$$= \ E_\theta((\partial f(\theta,x)/\partial\theta)^2/f(\theta,x)^2) \ = \ I(\theta),$$

again by convergence of integrals of pointwise convergent, uniformly integrable functions (RAP, Theorem 10.3.6) and the assumptions. □

If $x = (x_1, \ldots, x_n)$ for i.i.d. x_i, then $I(\theta)$ is n times the information for one observation:

2.4.9 Theorem. Suppose that $x = (x_1, \ldots, x_n)$ where x_i are i.i.d. with distribution having density $f_1(\theta, x_1)$ with respect to v, so that $dP_\theta^n/dv^n \ = \ f(\theta,x) \ = \ \Pi_{1\le j\le n} f_1(\theta, x_j)$. Also assume the hypotheses on f and $R_{\phi,\theta}$ in Theorem 2.4.8 hold for f_1 and $f_1(\phi,\cdot)/f_1(\theta,\cdot)$ respectively. Then $I(\theta) = nI_1(\theta)$ where $I_1(\theta) \ := \ E_\theta((\partial \log f_1(\theta,x_1)/\partial\theta)^2)$.

Proof. We have $\log f(\theta,x) = \sum_{j=1}^n \log f_1(\theta,x_j)$, and

$$I(\theta) \ = \ nE_\theta((\partial \log f_1(\theta,x_1)/\partial\theta)^2) + n(n-1)(E_\theta \partial \log f_1(\theta,x_1)/\partial\theta)^2.$$

Now since uniform square-integrability implies uniform integrability,

$$E_\theta \partial \log f_1(\theta,y)/\partial\theta \ = \ \int \lim_{\phi\to\theta} (f_1(\phi,y) - f_1(\theta,y))((\phi-\theta)f_1(\theta,y))^{-1} f_1(\theta,y)dv(y)$$

$$= \ \lim_{\phi\to\theta}(\phi-\theta)^{-1} \int f_1(\phi,y) - f_1(\theta,y)dv(y) \ = \ 0.$$

Thus $I(\theta) = nI_1(\theta)$. □

For Theorem 2.4.9 to be useful, it will be helpful to know that the information inequality holds for the case of n i.i.d. observations under hypotheses on the densities $f_1(\theta,x)$ of individual variables. One such fact is as follows:

2.4.10 Theorem. Under the conditions of Theorem 2.4.9, if $T = T(x_1, \ldots, x_n)$ is an unbiased estimator of $g(\theta)$ and $g'(\theta) \neq 0$ exists, then

$$\mathrm{var}_\theta T \ \geq \ g'(\theta)^2/(nI_1(\theta)).$$

Proof. The uniform integrability condition in Theorem 2.4.8 extends to more than one variable as follows. First, for $n = 2$,

$$R_{\phi,\theta}(X_1)R_{\phi,\theta}(X_2) - 1 \ = \ R_{\phi,\theta}(X_1)(R_{\phi,\theta}(X_2) - 1) + (R_{\phi,\theta}(X_1) - 1). \tag{2.4.11}$$

To show that a class of functions of the form $(f + g)^2$ is uniformly integrable for f in a class $\mathcal{F}$ and g in a class $\mathcal{G}$, noting that $(f+g)^2 \le 2f^2 + 2g^2$, it is enough to show that the sets of functions f^2 and g^2 are uniformly integrable. Dividing each term on the right of (2.4.11) by $\phi - \theta$ and squaring, the latter term is uniformly integrable for $0 < |\phi-\theta| < \delta$ by assumption. For the former term, using independence of X_1 and X_2, it will be enough to show that the $R_{\phi,\theta}(X_1)^2$ are uniformly integrable, or equivalently that $(R_{\phi,\theta}(X_1)-1)^2$ are. This is clear on multiplying and dividing by $(\phi-\theta)^2$, which is less than δ^2. The uniform integrability in Theorem 2.4.8 then extends to $n > 2$ by induction, so the information inequality holds and the form of $I(\theta)$ is given by Theorem 2.4.9. □

Note. If the hypotheses of Theorems 2.4.3 and 2.4.7 hold for unbiased estimators $T(x)$, which can be viewed as $T(X_1)$, then they do not necessarily follow for unbiased estimators $T(X_1, \ldots, X_n)$. In fact, often the set of functions $g(\theta)$ that have unbiased estimators depends on n, e.g. for binomial distributions, Section 2.2, Problem 5. So to apply these theorems to $n > 1$ we would need to check their hypotheses for $x = (x_1, \ldots, x_n)$ rather than only for $n = 1$.

Examples. (1) Let $f_1(\theta, x)$ be the normal $N(\theta, 1)$ density and $g(\theta) = \theta$. Then by Theorem 2.4.10, $\mathrm{var}_\theta T \geq 1/n$ for any unbiased estimator $T(X_1, \ldots, X_n)$ of θ. This variance is attained by $T := \overline{X}$, for any θ, so $\overline{X}$ is a "uniformly minimum-variance unbiased estimator."
(2) Let v be counting measure on the nonnegative integers and let P_θ be the Poisson law with parameter θ, $P_\theta(j) = e^{-\theta}\theta^j/j!$, $j = 0, 1, \ldots$. Let $g(\theta) \equiv \theta$. Then $I_1 = E_\theta((j\theta^{-1} - 1)^2) = 1/\theta$ and Theorem 2.4.10 gives $\mathrm{var}_\theta T \geq \theta/n$ for any unbiased estimator T. Again, this minimum variance is attained by the unbiased estimator $\overline{X}$ for all θ.
(3) The information inequality lower bound cannot always be attained. For normal measures $N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma > 0$, $s^2 := \sum_{i=1}^n (X_i - \overline{X})^2/(n-1)$ is an unbiased estimator of σ^2 with variance $2\sigma^4/(n-1)$ while $I_1(\sigma^2) = 1/(2\sigma^4)$, so the lower bound given by Theorem 2.4.10 is $2\sigma^4/n$. It will be shown in the next section that $2\sigma^4/(n-1)$ is the smallest variance actually attainable.

The requirement that an estimator is unbiased can be restrictive, and as we saw in Sec. 2.2, can force a bad choice of estimator. The inequalities proved earlier in this section can be adapted to give bounds for mean-square errors for more general estimators as follows.

Let T be a statistic used as an estimator of a function $g(\theta)$. Let $b(\theta) := E_\theta T - g(\theta)$ for all θ. Then $b(\theta)$ is called the *bias* at θ and is 0 for all θ if and only if T is an unbiased estimator of g. In general, as long as $E_\theta|T| < \infty$ for all θ, T will always be an unbiased estimator of $(g + b)(\theta)$, and so:

2.4.12 Theorem. If sufficient conditions for the information inequality hold for $g + b$ in place of g, then for all θ,

$$E_\theta((T - g(\theta))^2) \geq \frac{(g+b)'(\theta)^2}{I(\theta)} + b(\theta)^2.$$

Proof. For any random variable Y with mean μ and any c, we have $E((Y - c)^2) = \mathrm{var}(Y) + (c - \mu)^2$, so the Theorem follows. □

The hypotheses of Theorem 2.4.2 can be weakened as follows. Let

$$J_-(\theta) := \liminf_{y \to \theta}(\mathrm{var}_\theta R_{y,\theta})/(y - \theta)^2 \quad \text{and}$$
$$S(\theta) := \limsup_{y \to \theta} |g(y) - g(\theta)|/|y - \theta|.$$

Then if either J_- or S is a limit as well as a lim inf or lim sup respectively, and at least one is not zero, it will follow that $\mathrm{var}_\theta T \geq S^2(\theta)/J_-(\theta)$. In particular, if $g'(\theta) \neq 0$ exists, then $\mathrm{var}_\theta T \geq g'(\theta)^2/J_-(\theta)$.

The information $I(\theta)$ equals $J_-(\theta)$ if in the definition of $J_-(\theta)$, the lim inf is a limit $J(\theta)$ *and* the limit can be interchanged with the integral sign, as it can be under conditions treated above.

If $\partial f(\theta, x)/\partial\theta$ exists for v-almost all x, then $I(\theta)$ is defined (possibly $+\infty$) and $I(\theta) \leq J_-(\theta)$ by Fatou's Lemma (RAP, 4.3.3). Here $\mathrm{var}_\theta T \geq g'(\theta)^2/I(\theta)$ may not hold without further hypotheses:

2.4.13 Proposition. There exist densities $f(\theta, x)$ with respect to Lebesgue measure on $\mathbb{R}$ defined for $-1 < \theta < 1$ such that $f(\cdot, \cdot)$ is jointly C^∞ (infinitely differentiable) in both its variables, with $f(\theta, x) > 0$ and $\partial f(\theta, x)/\partial\theta|_{\theta=0} = 0$ for all x, $I(0) = 0$, and $J(0) = +\infty$. Also, x is an unbiased estimator of θ, $E_\theta x \equiv \theta$, and $\mathrm{var}_0 x = 1$. Thus the information inequality $\mathrm{var}_0 x \geq 1/I(0)$ fails.

Proof. Let f be a nonnegative C^∞ function which is even ($f(x) \equiv f(-x)$) and has compact support and $\int_{-\infty}^\infty f(x)dx = 1$, such as, for the suitable normalizing constant c,

$$f(x) = \begin{cases} c \cdot \exp(-(1-x)^{-2} - (1+x)^{-2}), & \text{for } -1 < x < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Then $\int_{-\infty}^\infty x f(x)dx = 0$. Let h be the standard normal $N(0, 1)$ density. Let

$$f(\theta, x) = \begin{cases} (1-\theta^2)h(x) + \theta^2 f(x - \theta^{-1}), & \text{for } 0 < |\theta| < 1 \\ h(x), & \text{for } \theta = 0. \end{cases}$$

Then $f(\theta, x) > 0$ for all x and $|\theta| < 1$. Since h and f both have mean 0, the mean $\int_{-\infty}^\infty x f(\theta, x)dx$ is 0 for $\theta = 0$ and $\theta^2\theta^{-1} = \theta$ otherwise, so x is an unbiased estimator of θ. For x in any bounded interval, $f(\theta, x) = (1-\theta^2)h(x)$ for θ in a neighborhood of 0, specifically, for $|x| \leq M$ and $|\theta| < 1/(M+1)$. Since $f(\theta, x)$ is clearly C^∞ in x and θ for θ outside a neighborhood of 0, it is in fact jointly C^∞ for all x and for $-1 < \theta < 1$, with $\partial f(\theta, x)/\partial\theta|_{\theta=0} = 0$ for all x. So $I(0) = 0$. For $y \neq 0$,

$$\begin{aligned} \mathrm{var}_0 R_{y,0} &= \int_{-\infty}^\infty f(y, x)^2 f(0, x)^{-1} dx - 1 \\ &= -2y^2 + y^4 + 2y^2(1-y^2) + y^4 \int_{-\infty}^\infty f(x - y^{-1})^2 h^{-1}(x)dx \\ &= y^4 \left[-1 + \int_{-\infty}^\infty f(x - y^{-1})^2 \exp(x^2/2)(2\pi)^{1/2} dx\right]. \end{aligned}$$

The latter integral goes to $+\infty$ as $y \to 0$, as $\exp(y^{-2}/2)$ or faster, so $J(0) = +\infty > I(0) = 0$. The rest follows. $\square$

So, existence of integrals involving $\partial f(\theta, x)/\partial\theta$ does not guarantee that limits can be interchanged with integrals, and the uniform integrability conditions in Theorems 2.4.7 and 2.4.8 can't simply be removed. In the example in the last proof, letting τ^2 be the variance of the law with density f,

$$\mathrm{var}_\theta x = E_\theta(x^2) - \theta^2 = (1-\theta^2) \cdot 1 + \theta^2(\theta^{-2} + \tau^2) - \theta^2 = 2 - (2-\tau^2)\theta^2$$

for $\theta \neq 0$ and 1 for $\theta = 0$. So the variance of x is discontinuous at $\theta = 0$.

Suppose we have another parametrization of a family $\{P_\theta,\ \theta \in \Theta\}$ where $Q_\psi = P_{\theta(\psi)}$ and that we want to estimate $g(\theta) = g(\theta(\psi))$. Then we have

2.4.14 Theorem. If $\psi \mapsto \theta(\psi)$ is differentiable with a non-zero derivative then the information inequality lower bound for var T is the same for the parametrization by ψ as for the parametrization by θ.

Proof. In the change from parameter θ to parameter ψ in the information inequality, by the chain rule, both numerator and denominator are multiplied by $\theta'(\psi)^2 > 0$, not changing the bound. □

The information inequality is most useful in cases where there exists some unbiased estimator T whose variance attains the lower bound given in 2.4.3 for all θ. It turns out that under some regularity conditions (stronger than those needed for the information inequality itself), the bound is attained only for densities of a certain "exponential" form, which will be studied more generally and in more detail in the next section. Recall that a function is called C^1 if it is everywhere differentiable with a continuous derivative.

2.4.15 Theorem. Assume the hypotheses of Theorem 2.4.3 for all θ in an open interval Θ and that $0 < \mathrm{var}_\theta T < \infty$ for all θ and $\partial \log f(\theta, x)/\partial\theta$ is continuous in θ for almost all x. Then the information inequality becomes an equation for all θ if and only if there exist C^1 functions $c(\cdot)$ and $d(\cdot)$ of θ and a measurable function h of x such that for all θ and almost all x,

$$f(\theta, x) \ = \ c(\theta)h(x)\exp(d(\theta)T(x)).$$

Proof. By the assumptions, $I(\theta)\mathrm{var}_\theta T > 0$ and g is everywhere differentiable on the open interval Θ, so it is continuous. The proof of Theorem 2.4.3 gives $g'(\theta)^2 \ \leq \ (\mathrm{var}_\theta T)I(\theta)$, which must become an equation since the information inequality does. So, for each θ, the functions $T - g(\theta)$ and $\partial \log f/\partial\theta$ must be proportional (in the proof of RAP, 5.3.3, $b^2 - 4ac = 0$ implies $\|f + tg\|^2 = 0$ for some t). So for each θ, there is an $a(\theta)$ such that $\partial \log f(\theta, x)/\partial\theta \ = \ a(\theta)(T(x) - g(\theta))$ for almost all x. Since $\mathrm{var}_\theta T > 0$ there is a set of x of positive measure on which $T(x) \neq g(\theta)$, so $a(\theta)$ is uniquely determined. For the same reason, there must exist some number c such that $T(x) > c$ and $T(x') < c$ for x and x' in sets A, B of positive measure respectively, where also for $y = x$ or x', $\partial \log f(\theta, y)/\partial\theta$ is continuous in θ. Thus $\partial \log f(\theta, x)/\partial\theta - \partial \log f(\theta, x')/\partial\theta$ is continuous in θ for any such x, x'. For any given θ, the difference equals $a(\theta)[T(x) - T(x')]$ for almost all $x \in A$ and $x' \in B$. Taking any convergent sequence $\theta_j \to \theta_0$ of values of θ, we have the equality for almost all $x \in A$ and $x' \in B$ for all $\theta_j,\ j \geq 0$. Thus $a(\theta_k) \to a(\theta_0)$. A real-valued function of a real variable, continuous along any sequence, is continuous, so $a(\cdot)$ is continuous. We can then take an indefinite integral to get $\log f(\theta, x) \ = \ d(\theta)T(x) + u(x) - j(\theta)$ for some measurable function $u(x)$ and C^1 functions $d(\theta)$ and $j(\theta)$. Taking the exponential of both sides finishes the proof in one direction. Conversely, when functions are proportional, the Cauchy-Bunyakovsky-Schwarz inequality always becomes an equation. □

On the regularity conditions needed for Theorem 2.4.15, see the Notes.

PROBLEMS

1. Consider the family of exponential distributions with densities $f_c(x) = e^{-x/c}/c$ for $x \geq 0$ and 0 for $x < 0$, where $0 < c < \infty$. For n observations, show that $\overline{X}$ is an unbiased estimator of c with variance c^2/n and that this attains the minimum possible variance given by the information inequality.
2. Suppose that the exponential distributions are parametrized by $\lambda = 1/c$ instead of c, so that the densities are $h_\lambda(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, and that we want to estimate λ, or in other words we want to estimate the function $g(c) = 1/c$ for the parametrization in Problem 1. Show that for $n = 1$ there is no unbiased estimator of λ, while for $n > 1$, $(n-1)/(n\overline{X})$ is such an estimator. Compare its variance to the information inequality lower bound. Hint: if $X_1, \dots, X_n$ are i.i.d. with standard exponential density $f_1 \equiv h_1$, then $X_1 + \cdots X_n$ has a gamma density $x^{n-1}e^{-x}/(n-1)!$ for $x \geq 0$ and 0 for $x < 0$. Also, cX_i are i.i.d. with density f_c.
3. For a fixed $h > 0$ consider the family $\mathcal{U}_h$ of all uniform distributions on $[\theta, \theta + h]$ for $\theta \in \mathbb{R}$, as in Problem 3 of Sec. 2.3. Show that for estimating θ with squared-error loss $(T - \theta)^2$, there exists an unbiased estimator with variance of the order $1/n^2$ as the sample size $n \to \infty$, which is smaller than can happen when the information inequality applies. Explain why the information inequality fails to apply in this case. Hint: for $\theta = 0$ and $h = 1$, show that the probability that $X_{(1)} \leq x$ is $1 - (1-x)^n$ and then that $EX_{(1)} = 1/(n+1)$.
4. Evaluate the information $I(\theta)$ in the following cases.
   (a) Binomial distribution $B(n,p)$ for n fixed, $0 < p < 1$.
   (b) geometric distribution $P(k) = (1-p)^{k-1}p$ for $k = 1, 2, \dots$.
5. Evaluate the information of $N(0, \sigma^2)$, taking the parameter θ to be (a) σ, (b) σ^2.

NOTES

Theorem 2.4.1 is due to Hammersley (1950) and was rediscovered by Chapman and Robbins (1951). The notion of information $I(\theta)$ originated with Edgeworth (1908,1909) and was developed by Fisher (1922 and later papers), see Savage (1976).

The information inequality (2.4.3), $\mathrm{var}_\theta T \geq g'(\theta)^2/I(\theta)$, was first found by Fréchet (1943) and extended by Darmois (1945). It was rediscovered by C. R. Rao (1945) and Cramér (1946a; 1946b, pp. 475-476) and has been widely known as the "Cramér-Rao" inequality. In view of the contributions of Fréchet and Darmois, L. J. Savage (1954) proposed the name "information inequality." Rao (1945, equ. (3.2)) did not actually state regularity conditions adequate to justify his interchange of limits. Cramér did, but in a special case where not only $g(\theta) \equiv \theta$ but $T(x) \equiv x$.

Joshi (1976) gives an example of a location family, so that $f(\theta, x) \equiv f(x - \theta)$, and an estimator for which the information inequality becomes an equation for all θ, $-\infty < \theta < \infty$, but which does not have the exponential form given in Theorem 2.4.15. The given $f(\cdot)$ is not continuous, having some jumps, so for no x is $f(\cdot, x)$ everywhere differentiable with respect to θ, and the hypothesis of 2.4.15 fails although for each x, the density is smooth with respect to θ except for a few jumps. See Joshi (1976) for details.

The notes for this section are largely based on those in Lehmann (1983, p. 145).

REFERENCES

Chapman, D. G., and Robbins, H. (1951). Minimum variance estimation without regularity assumptions. *Ann. Math. Statist.* **22**, 581-586.

Cramér, Harald (1946a). A contribution to the theory of statistical estimation. *Skand. Aktuarietidskr.* **29**, 85-94.

Cramér, H. (1946b). *Mathematical Methods of Statistics.* Princeton University Press.

Darmois, G. (1945). Sur les lois limites de la dispersion de certaines estimations. *Rev. Inst. Internat. Statist.* **13**, 9-15.

Edgeworth, F. Y. (1908,1909). On the probable errors of frequency constants. *J. Royal Statist. Soc.* **71**, 381-397, 499-512, 651-678; **72**, 81-90.

Fisher, Ronald A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London Ser. A* **222**, 309-368.

Fréchet, Maurice (1943). Sur l'extension de certaines évaluations statistiques au cas de petits échantillons. *Rev. Inst. Internat. Statist.* **11**, 182-205.

Hammersley, J. M. (1950). On estimating restricted parameters. *J. Roy. Statist. Soc. (Ser. B)* **12**, 192-240.

Joshi, V. M. (1976). On the attainment of the Cramér-Rao lower bound. *Ann. Statist.* **4**, 998-1002.

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

Rao, C. Radhakrishna (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37**, 81-91.

Savage, L. J. (1954, 1972). *The Foundations of Statistics.* Wiley, New York; Rev. Ed. Dover.