INFORMATION INEQUALITIES

When we consider a parametric 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 $T = T(X_1, \ldots, X_n)$ 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 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 10) 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 7), 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.

Theorem 1. *Suppose T is an unbiased estimator of a real function $g(\theta)$ for an equivalent family $\{P_\theta,\ \theta \in \Theta\}$. Then*

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

Note. The conclusion of the theorem holds trivially if $\operatorname{Var}_\theta T = +\infty$ or if $\operatorname{Var}_\theta R_{\phi,\theta} \equiv +\infty$ for all $\phi \neq \theta$. So the theorem has content if and only if both $E_\theta(T^2) < \infty$ and $\operatorname{Var}_\theta R_{\phi,\theta} < \infty$ 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

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

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

Then by the Cauchy–Bunyakovsky–Schwarz inequality (e.g. Dudley [2002, 5.1.4]),

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

where $\operatorname{Var}_\theta R_{\phi,\theta} > 0$ 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.

Theorem 2. *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 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, and 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:

Theorem 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*

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

$$0 \;=\; \int \frac{\partial f(\theta,x)}{\partial\theta}\; dv(x), \tag{2}$$

and

$$g'(\theta) \;=\; \int T(x)\frac{\partial f(\theta,x)}{\partial\theta}\,dv(x). \tag{3}$$

Then

$$\operatorname{Var}_\theta T \;\geq\; g'(\theta)^2/I(\theta). \tag{4}$$

Note. 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) and (3) is equivalent to (1), since the integrands are measurable functions of x.

Proof. Multiplying (2) by $g(\theta)$ and subtracting from (3) gives

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

If $\operatorname{Var}_\theta T = +\infty$ or $I(\theta) = +\infty$, the inequality holds trivially since $g'(\theta)$ is finite. If $\operatorname{Var}_\theta T$ and $I(\theta)$ are both finite, then the Cauchy–Bunyakovsky–Schwarz inequality can be applied as in the proof of Theorem 1 to get $g'(\theta)^2 \;\leq\; I(\theta)\operatorname{Var}_\theta T$. □

When n i.i.d. observations are taken, the Fisher information, if it exists, is multiplied by n, as follows.

Proposition 4. *Let $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$ be an equivalent family where Θ is an open interval in $\mathbb{R}$. Suppose that the Fisher information $I_1(\theta) := I(\theta)$ exists and is finite and that (2) holds for each $\theta\in\Theta$. Then for the family $\mathcal{P}^n := \{{P_\theta}^n : \theta\in\Theta\}$, the Fisher information $I_n(\theta)$ exists and equals $nI_1(\theta)$ for each θ.*

Proof. Let v be a σ-finite dominating measure for $\mathcal{P}$, e.g. a member of $\mathcal{P}$, and $f(\theta,x) := (dP_\theta/dv)(x)$. Then $\mathcal{P}^n$ is dominated by v^n, with likelihood functions $f^{(n)}(\theta,(X_1,...,X_n)) = \Pi_{j=1}^n f(\theta,X_j)$, so

$$\log f^{(n)}(\theta,(X_1,...,X_n)) = \sum_{j=1}^n \log f(\theta,X_j).$$

By the assumptions, $\partial\log f(\theta,X_j)/\partial\theta$, $j=1,...,n$, are i.i.d. variables for ${P_\theta}^n$ with mean 0 and finite variance $I_1(\theta)$, so the conclusion follows. □

The information inequality is usually stated under assumptions such as those of Theorem 3. Exchanging differentiation with an integral as in (2) and (3) may seem a plausible and reasonable kind of hypothesis. But an example will be given below (Proposition 11) showing that assumption (3) may not hold even when (1) does, and where each

derivative and integral in (3) is well-defined and finite. So let's see how (1) can be strengthened enough to imply (2) and (3), by way of the notion of uniform integrability (Dudley [2002, 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 $|f| \le g$ for all $f \in \mathcal{F}$.

Theorem 5. *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*

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

are uniformly integrable for v, or equivalently the functions

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

are uniformly integrable with respect to P_θ. Then (1), (2) and (3) all hold.

Proof. The conditions follow from convergence of integrals of pointwise convergent, uniformly integrable functions (Dudley [2002, Theorem 10.3.6]). □

Theorems 3 and 5 have been stated for one unbiased estimator T, but the information inequality has usually been stated as applying to all unbiased estimators, with hypotheses (1) and (3) assumed for all such estimators of $g(\theta)$. There can in general be many 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 (1) and (3) hold for all unbiased estimators of g. An alternate sufficient condition will be stated just in terms of g and the family $\{P_\theta, \ \theta \in \Theta\}$:

Theorem 6. *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, we have

$$\begin{aligned} J(\theta) \ &= \ \lim_{\phi\to\theta}(\mathrm{Var}_\theta\, R_{\phi,\theta})/(\phi - \theta)^2 \ = \ \lim_{\phi\to\theta} E_\theta([(R_{\phi,\theta} - 1)/(\phi - \theta)]^2) \\ &= \ E_\theta((\partial f(\theta, x)/\partial\theta)^2/f(\theta, x)^2) \ = \ I(\theta), \end{aligned}$$

again by convergence of integrals of pointwise convergent, uniformly integrable functions (Dudley [2002, Theorem 10.3.6]) and the assumptions. □

The information inequality extends to n i.i.d. observations under uniform integrability assumptions:

Theorem 7. *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 6 hold for f_1 and $f_1(\phi,\cdot)/f_1(\theta,\cdot)$ respectively. If $T = T(x_1, \ldots, x_n)$ is an unbiased estimator of $g(\theta)$ and $g'(\theta) \neq 0$ exists, then*

$$\operatorname{Var}_\theta T \;\ge\; g'(\theta)^2/(nI_1(\theta)).$$

Proof. The uniform integrability condition in Theorem 6 extends to more than one variable as follows. First, for $n = 2$,

$$(5)\quad R_{\phi,\theta}(X_1)R_{\phi,\theta}(X_2) - 1 \;=\; R_{\phi,\theta}(X_1)(R_{\phi,\theta}(X_2) - 1) + (R_{\phi,\theta}(X_1) - 1).$$

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 (5) 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 6 then extends to $n > 2$ by induction, so the information inequality holds and the form of $I(\theta)$ is given by Proposition 4. □

Note. If the hypotheses of Theorems 3 and 5 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. 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 $n = 1$.

Example 8. (1) Let $f_1(\theta, x)$ be the normal $N(\theta, 1)$ density and $g(\theta) = \theta$. Then by Theorem 7, $\operatorname{Var}_\theta T \ge 1/n$ 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 7 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 7 is $2\sigma^4/n$. According to the following Proposition 9, $2\sigma^4/(n-1)$ is the smallest variance actually attainable.

If we drop the restriction that an estimator $T(X)$ of the variance σ^2 be unbiased, but consider estimators $T(X) = a_n \sum_{j=1}^n (X_j - \overline{X})^2$, we know that for normal distributions, $a_n = \frac{1}{n+1}$ gives smallest mean-square error, and that the factor $a_n = \frac{1}{n-1}$ gives an estimator inadmissible for general distributions with $E(X_1^4) < \infty$ (Yatracos's theorem), where $a_n = (n+2)/(n(n+1))$ always gives smaller risk. But, if we do restrict to unbiased estimators, then for normal distributions, $a_n = 1/(n-1)$ gives smallest risk. This is not surprising, as for example, for $n \geq 3$, $T(X) = \frac{1}{n-2} \sum_{j=1}^{n-1} (X_j - \overline{X})^2$ is unbiased but inferior (it doesn't use X_n).

A fuller proof of the following was given in the OCW 2003 notes, Section 2.5, depending on the notion of Lehmann–Scheffé statistic or σ-algebra (Section 2.3).

Proposition 9. *Let $X_1, \ldots, X_n$ be i.i.d. with law $N(\mu, \sigma^2)$ and $n \geq 2$. Then $s^2 := (n-1)^{-1} \sum_{j=1}^n (X_j - \overline{X})^2$ has, among unbiased estimators of σ^2, smallest risk for squared-error loss, for all (μ, σ^2). The risk of s^2 is $2\sigma^4/(n-1)$.*

Proof. (Sketch) We know that s^2 is an unbiased estimator of σ^2. Let $\mathcal{S}$ be the smallest σ-algebra for which T_1 and T_2 are measurable in this case. Then $\mathcal{S}$ is sufficient. It is actually minimal sufficient. Since $s^2 = (n-1)^{-1}(T_1 - T_2^2/n)$, s^2 is $\mathcal{S}$-measurable. Then it follows that s^2 has minimum risk for squared-error loss (which is convex).

To find the variance of s^2, first note that if X has distribution $N(0, \sigma^2)$, then $EX^4 = 3\sigma^4$, by integration by parts or the moment generating function. Also, $Es^2 = \sigma^2$ and we can assume $m = 0$. Make an orthogonal change of coordinates from $(X_1, \ldots, X_n)$ to $(Y_1, \ldots, Y_n)$ where the Y_1 axis is in the direction of $(1, 1, \ldots, 1)$, so that $Y_1 = n^{1/2}\overline{X}$.

Then the Y_j are i.i.d. $N(0,\sigma^2)$ and $s^2 = (n-1)^{-1}\sum_{j=2}^n Y_j^2$. So

$$E((s^2)^2) \;=\; (n-1)^{-2}\sigma^4[3(n-1)+(n-1)(n-2)] \;=\; (n+1)\sigma^4/(n-1),$$

and $\mathrm{Var}(s^2) = 2\sigma^4/(n-1)$. □

The requirement that an estimator is unbiased can be restrictive, and as we have seen, 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. If T is not necessarily unbiased, the mean-square error of T as an estimator of g is

(6)
$$E_\theta\left[(T-g(\theta))^2\right] = E_\theta\left[(T - E_\theta T + E_\theta T - g(\theta))^2\right] = \mathrm{Var}_\theta\, T + b(\theta)^2.$$

Classically, up through the 1940's or perhaps 1950's, the usual approach to minimizing the mean-square error was to look for unbiased estimators T and then minimize their variance. In more recent decades, it was realized that a small bias $b(\theta)$ is not necessarily harmful, because when squared it becomes very small, and it may give us freedom to reduce the variance and total mean-square error. So, there is a "bias-variance tradeoff," now a frequently used phrase, where the focus is more directly on minimizing mean-square error, by estimators T that may be biased. Theorem 4 of the "Bayes estimates" handout shows that Bayes estimators, which minimize average mean-square error with respect to a prior, are virtually never unbiased.

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 by equation (6) we have:

Theorem 10. *If sufficient conditions for the information inequality hold for $g+b$ in place of g, then for all θ,*

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

The hypotheses of Theorem 2 can be weakened as follows. Let

$$J_-(\theta) \quad := \quad \liminf_{y\to\theta} (\mathrm{Var}_\theta\, R_{y,\theta})/(y-\theta)^2$$

and

$$S(\theta) \quad := \quad \limsup_{y\to\theta} |g(y)-g(\theta)|/|y-\theta|.$$

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 $\operatorname{Var}_\theta T \geq S^2(\theta)/J_-(\theta)$. So, if $g'(\theta) \neq 0$ exists, then $\operatorname{Var}_\theta T \geq g'(\theta)^2/J_-(\theta)$.

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 (Dudley [2002, 4.3.3]) Here $\operatorname{Var}_\theta T \geq g'(\theta)^2/I(\theta)$ may not hold without further hypotheses:

Proposition 11. *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 $\operatorname{Var}_0 x = 1$. Thus the information inequality $\operatorname{Var}_0 x \geq 1/I(0)$ 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,

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

Then $\int_{-\infty}^{\infty} xf(x)dx = 0$. 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 real x and $|\theta| < 1$. Since h and f both have mean 0, the mean $\int_{-\infty}^{\infty} xf(\theta, x)dx$ 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 $y \neq 0$,

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

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. □

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 5 and 6 can't simply be removed. In the example in the last proof, letting τ^2 be the variance of the law with density f,

$$\operatorname{Var}_\theta x \;=\; E_\theta(x^2) - \theta^2 \;=\; (1-\theta^2)\cdot 1 + \theta^2(\theta^{-2}+\tau^2) - \theta^2 \;=\; 2 - (2-\tau^2)\theta^2$$

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

Suppose we have another parameterization 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

Theorem 12. *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 parameterization by ψ as for the parameterization 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 Theorem 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 in exponential families. Recall that a function is called C^1 if it is everywhere differentiable with a continuous derivative.

Theorem 13. *Assume the hypotheses of Theorem 3 for all θ in an open interval Θ and that $0 < \operatorname{Var}_\theta T < \infty$ 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, $I(\theta)\operatorname{Var}_\theta T > 0$ and g is everywhere differentiable on the open interval Θ, so it is continuous. The proof of Theorem 3 gives $g'(\theta)^2 \le (\operatorname{Var}_\theta T)I(\theta)$ via the Cauchy–Bunyakovsky–Schwarz inequality, 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 (e.g. in the proof of Dudley [2002, 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 $\operatorname{Var}_\theta T > 0$ there is a set of x of positive measure on which $T(x) \ne g(\theta)$, 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 $\theta_j,\ j \ge 0$. 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. □

NOTES

Theorem 1 is due to Hammersley (1950). Chapman and Robbins (1951) rediscovered it. The notion of information $I(\theta)$ originated with Edgeworth (1908,1909). Fisher (1922 and later papers) developed it, see Savage (1976).

The information inequality (3), $\operatorname{Var}_\theta T \ge g'(\theta)^2/I(\theta)$, was first found by Fréchet (1943) and extended by Darmois (1945). It was rediscovered by C. R. Rao (1945) and Cramér (1945, pp. 475-476; 1946) and had 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, eq. (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 13. 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 Theorem 13 fails although for each x, the density is smooth with respect to θ except for a few jumps. See Joshi (1976) for details.

These notes are largely based on those in Lehmann (1983, p. 145).