Notes on the Cramér-Rao Inequality

Kimball Martin

February 8, 2012

Suppose X is a random variable with pdf $f_X(x;\theta)$, θ being an unknown parameter. Let $X_1, \dots, X_n$ be a random sample and $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$. We've seen that $E(\hat{\theta})$, or rather $E(\hat{\theta}) - \theta$, is a measure of how biased $\hat{\theta}$ is. We've also seen that $Var(\hat{\theta})$ provides a measure of efficiency, i.e., the smaller the variance of $\hat{\theta}$, the more likely $E(\hat{\theta})$ will provide an accurate estimate of θ.

Given a specific unbiased estimator $\hat{\theta}$, how do we know if it is the best (most efficient, i.e., smallest variance) one, or if there is a better one? A key tool in understanding this question is a theoretical lower bound on how small $Var(\hat{\theta})$ can be. This is the Cramér-Rao Inequality.

From now on, we **assume** X is continuous and θ is a single real parameter (i.e., there is only one unknown). We will also **assume** the range of X does not depend on θ. To be more precise, we will assume there exist $a, b \in \mathbb{R} \cup \{\pm\infty\}$ independent of θ such that

$$\begin{cases} f_X(x;\theta) > 0 & \text{if } a < x < b \\ f_X(x;\theta) = 0 & \text{if } x < a \text{ or } x > b. \end{cases}$$

For instance if $(a, b) = (-\infty, \infty)$, then we are assuming $f_X(x;\theta) > 0$ for all θ and all real x. Things like the normal distribution on $\mathbb{R}$ and the exponential distribution on $[0, \infty)$ satisfy these conditions. An example which does not satisfy this regularity condition is X uniform on $[0, \theta]$, because then we would need to take $a = 0$ and $b = b(\theta) = \theta$, which is dependent upon θ.

To discuss the Cramér-Rao Inequality, it will be helpful to introduce a bit more notation and terminology. As a bit of motivation, we've already seen in the maximum likelihood method, it is sometimes useful to work with the function $\ln f_X(x;\theta)$ (the natural log of the likelihood function $L(\theta) = \prod f_X(X_i;\theta)$ becomes $\ln L(\theta) = \sum \ln f_X(X_i, \theta)$).

Definition 1. *We call $V(\theta) = V_X(\theta) = \ln f_X(X;\theta)$ the* **score** *with respect to θ. The number*

$$I(\theta) = E(V'(\theta)^2)$$

is called **Fisher's information number**.

What does the above mean?

Example 1. *Suppose X is exponential with pdf $f_X(x;\theta) = \theta e^{-\theta x}$ for $x \geq 0$. As usual, θ is an unknown parameter. Then*

$$V(\theta) = \ln \theta e^{-\theta X} = \ln \theta + \ln e^{-\theta X} = \ln \theta - \theta X.$$

For a given value of θ, the score gives a random variable. It makes sense to talk about its expectation value. Similarly, the derivative with respect to θ also gives a random variable for each θ. The square of the expectation value of the derivative is Fisher's information number, which is a function of θ. Namely, in our example,

$$V'(\theta) = \frac{d}{d\theta} V(\theta) = \frac{1}{\theta} - X$$

so

$$I(\theta) = E(V'(\theta)^2) = E\left((\frac{1}{\theta} - X)^2\right) = E(\frac{1}{\theta^2} - \frac{2X}{\theta} + X^2) = \frac{1}{\theta^2} - \frac{2}{\theta}E(X) + E(X^2).$$

From probability we know $E(X) = \frac{1}{\theta}$ *and* $E(X^2) = Var(X) + E(X)^2 = \frac{1}{\theta^2} + \frac{1}{\theta^2} = \frac{2}{\theta^2}$. *Thus*

$$I(\theta) = \frac{1}{\theta^2} - \frac{2}{\theta} \cdot \frac{1}{\theta} + \frac{2}{\theta^2} = \frac{1}{\theta^2}.$$

Remark. Since the logarithm only makes sense for a positive argument, the score only makes sense when $f_X(X;\theta) \neq 0$. We also require $f_X(X;\theta)$ to be differentiable with respect to θ, which we assume. (This means we need $f_X(x;\theta) > 0$ for all x, θ.)

Lemma 1. *If* $V'(\theta)$ *is continuous (except possibly at finitely many points), then* $E(V'(\theta)) = 0$.

Proof. Note

$$V'(\theta) = \frac{\frac{\partial}{\partial\theta} f_X(x;\theta)}{f_X(x;\theta)}. \tag{1}$$

Thus, by definition

$$E(V'(\theta)) = \int_a^b V'_x(\theta) f_X(x;\theta) dx = \int_a^b \frac{\partial}{\partial\theta} f_X(x;\theta) dx = \frac{\partial}{\partial\theta} \int_a^b f_X(x;\theta) dx = \frac{\partial}{\partial\theta} 1 = 0.$$

Here we used Leibnitz's rule, which says one can interchange order of differentiation and integration with respect to independent variables assuming continuous partial derivatives. □

Corollary 1. $I(\theta) = Var(V'(\theta))$.

Proof.

$$I(\theta) = E(V'(\theta)^2) = Var(V'(\theta)) + E(V'(\theta))^2 = Var(V'(\theta)).$$

□

Thus Fisher's information number is the variance of the derivative of the score, i.e., the variance of the logarithmic derivative of the pdf $f_X(X;\theta)$ (cf. (1)). The logarithmic derivative is often a useful quantity to work with mathematically. For us, the point is that $I(\theta)$ appears in the Cramér-Rao bound. I'm sure you're anxious to get to this bound, now that I've hyped it up so much, but permit me one more

Lemma 2. *Assume* $V(\theta)$ *has continuous first and second derivatives. Then* $I(\theta) = -E(V''(\theta))$.

Exercise 1. *(*for grad students) Prove Lemma 2.*

The point is this often gives a simpler way to compute $I(\theta)$.

Example 2. *Returning to our example above of the exponential distribution,*

$$V''(\theta) = -\frac{1}{\theta^2}.$$

Since there is no dependence on X, *we could more quickly compute the Fisher information as*

$$I(\theta) = -E(V''(\theta)) = -V''(\theta) = \frac{1}{\theta^2}.$$

Theorem 1. (Cramér-Rao Inequality.) *Assume* $V(\theta)$ *has continuous first derivative (except possibly at finitely many points). Then for any unbiased estimator* $\hat{\theta}$,

$$Var(\hat{\theta}) \geq \frac{1}{nI(\theta)}.$$

This is the desired theoretical bound on how efficient an estimator can be. The theorem is in fact valid under weaker assumptions (see the text), i.e., $V(\theta)$ does not need to be differentiable everywhere, but we assume this for simplicity.

To prove this result, first we need a little material from Section 11.4 of the text.

Definition 2. *If X and Y are random variables, their* **covariance** *is*

$$Cov(X,Y) = E(XY) - E(X)E(Y).$$

Note $Cov(X,X) = var(X)$. Also note if X and Y are independent, then $Cov(X,Y) = E(X)E(Y) - E(X)E(Y) = 0$ so covariance measures how dependent X and Y are.

Lemma 3. $|Cov(X,Y)| \leq \sqrt{Var(X)Var(Y)}$.

Proof. Compute

$$Var(X \pm Y) = E((X \pm Y)^2) - E(X \pm Y)^2 = Var(X) \pm 2Cov(X,Y) + Var(Y).$$

Since this is always ≥ 0, we have

$$Cov(X,Y) \leq \frac{Var(X) + Var(Y)}{2}.$$

Applying this inequality to the normalized random variables $X' = \frac{X-\mu_X}{\sigma_X}$ and $Y' = \frac{Y-\mu_Y}{\sigma_Y}$ gives

$$Cov(X',Y') \leq \frac{Var(X') + Var(Y')}{2}.$$

Note $E(X') = E(Y') = 0$ and $Var(X') = Var(Y') = 1$, so we have

$$Cov(X',Y') = E(X'Y') \leq 1.$$

Since

$$\begin{aligned} E(X'Y') = \frac{E(XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y)}{\sigma_X\sigma_Y} &= \frac{E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X\mu_Y}{\sigma_X\sigma_Y} \\ &= \frac{E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y)}{\sqrt{Var(X)Var(Y)}} \\ &= \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}, \end{aligned}$$

we are done. □

Now we have all we need to piece together a proof of the theorem, at least when $n = 1$.

Proof. (of Theorem when $n = 1$) Here $\hat{\theta} = \hat{\theta}(X_1)$ is just a function of X_1, so we may think of it as a function of X. Observe

$$Cov(V'(\theta), \hat{\theta}) = E(V'(\theta) \cdot \hat{\theta}) - E(V'(\theta))E(\hat{\theta}) = E(V'(\theta) \cdot \hat{\theta})$$

by Lemma 1. By Lemma 3, we have

$$|E(V'(\theta) \cdot \hat{\theta})| = |Cov(V'(\theta), \hat{\theta})| \leq \sqrt{Var(V'(\theta))Var(\hat{\theta})}.$$

Hence

$$Var(\hat{\theta}) \geq \frac{|E(V'(\theta) \cdot \hat{\theta})|^2}{Var(V'(\theta))} = \frac{|E(V'(\theta) \cdot \hat{\theta})|^2}{I(\theta)},$$

where we used Corollary 1 for the equality on the right. Thus to prove the theorem, it suffices to show

$$|E(V'(\theta)\cdot\hat{\theta})|^2 = 1.$$

Note by (1) and the definition of expected value,

$$E(V'(\theta)\cdot\hat{\theta}) = \int_a^b \frac{\partial}{\partial\theta} f_X(x;\theta)\cdot\hat{\theta}(x)dx = \frac{\partial}{\partial\theta}\int_a^b f_X(x;\theta)\cdot\hat{\theta}(x)dx = \frac{\partial}{\partial\theta}E(\hat{\theta}) = \frac{\partial}{\partial\theta}\theta = 1.$$

Here we used Leibnitz's rule again in the middle, and in the next to the last step we used fact that $\hat{\theta}$ is unbiased. □

The proof for $n > 1$ is similar, but enters into a little unfamiliar territory that we have carefully sidestepped until now. Namely, we have only considered random variables defined on a 1-dimensional sample space. For example, let us suppose for concreteness X is defined on the sample space $(0,\infty)$. In other words, X is a function from $(0,\infty)$ to $\mathbb{R}$. We also looked at new random variables defined as functions of X, e.g., X^2 or $3X$ or X^2+X+1. These are still functions from $(0,\infty)$ to $\mathbb{R}$, and one can determine their pdfs from that of X.

Then we took a random sample $X_1,\dots X_n$ of X and considered a statistic (e.g., estimator) $\hat{\theta} = \hat{\theta}(X_1,\dots,X_n)$. For example $\hat{\theta} = \frac{1}{n}\sum X_i$. Since this is a sum of random variables (like $3X = X+X+X$ or X^2+X which we've looked at before), we also said this is a random variable. Then we computed its expected value, say, as

$$E(\hat{\theta}) = \frac{1}{n}\sum E(X_i) = \frac{1}{n}\sum E(X) = E(X).$$

Well, the jig is up. While $\hat{\theta}$ is a sum of random variables, there's is a qualitative difference between something like X_1+X_2 and $X+X$. Even though X_1 and X_2 have the same pdf as X, they are distinct measurement (they are independent!). Since X_1 and X_2 are independent functions from $(0,\infty)$ to $\mathbb{R}$, their sum

$$X_1+X_2 : (0,\infty)\times(0,\infty)\to\mathbb{R}$$

needs to be viewed as a function of $(0,\infty)\times(0,\infty)$, not $(0,\infty)$. If I think of X_1+X_2 as a function of just one parameter in $(0,\infty)$ then I won't be able to have $X_1 = 1$ and $X_2 = 3$, since X_1 will always equal X_2.

What is means for X_1 and X_2 to have identical pdfs, is that they really are the same function from $(0,\infty)\to\mathbb{R}$. What distinguishes them as independent events is that I think of them as living on two different sample spaces S_1 and S_2, which just both happen to be represented by $(0,\infty)$. Hence X_1+X_2 is not a random variable in the sense we defined earlier (namely, a 1-dimensional sample space). However, we can (and at least I will) think of X_1+X_2 as being a single random variable, only defined on a 2-dimensional sample space $(0,\infty)\times(0,\infty)$.

Since X_1 and X_2 are independent, we could still compute expectation values and variance for *any specific simple example* of an estimator $\hat{\theta}(X_1,X_2)$ without resorting to thinking of $\hat{\theta}(X_1,X_2)$ as a random variable on a 2-dimensional sample space. For instance if $\hat{\theta}(X_1,X_2) = \frac{1}{4}(X_1+X_2)^2$, then

$$E(\hat{\theta}) = \frac{1}{4}E((X_1+X_2)^2) = \frac{1}{4}E(X_1^2)+\frac{1}{2}E(X_1)E(X_2)+\frac{1}{4}E(X_2)^2 = \frac{1}{2}E(X^2)+\frac{1}{2}E(X)^2 = \frac{1}{2}Var(X)+E(X)^2,$$

and thus one can reduce $E(\hat{\theta})$ to computing things like $E(X^k)$.

However, if one wants to think about things in any generality, then a more sophisticated point of view needs to be considered. Recall the result that, given a nice function $h:\mathbb{R}\to\mathbb{R}$ we can compute the expected value of the continuous random variable $h(X)$ as

$$E(h(X)) = \int_{-\infty}^{\infty} h(X)f_X(x)dx.$$

For instance if $h(X) = X^2 + 3X + 1$, then

$$E(h(X)) = \int_{-\infty}^{\infty} (X^2 + 3X + 1) f_X(x) dx.$$

Now what happens if we try the same thing something like our previous example $\hat{\theta}(X_1, X_2) = \frac{1}{4}(X_1+X_2)^2$. It doesn't make sense to write

$$E(\hat{\theta}) = \int_{-\infty}^{\infty} \hat{\theta} f_X(x;\theta) dx$$

because $\hat{\theta}$ is now a function of two *independent* variables X_1 and X_2, so a single integral won't cut it. Instead, we need to look at a double integral. Here, what should be true is

$$E(\hat{\theta}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{\theta}(x_1, x_2) f_{X_1}(x_1;\theta) f_{X_2}(x_2;\theta) dx_1 dx_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_1 + x_2)^2 f_X(x_1;\theta) f_X(x_2;\theta) dx_1 dx_2.$$

This is in fact correct—we will skip the details, but they can be found in more advanced probability and statistics texts, along with much more general statements—and is essentially Theorem 3.9.1 in our text. Precisely, we state what we need as the following

Proposition 1. *Let $X_1, \ldots, X_n$ be a random sample for a continuous random variable X, and consider a function $h : \mathbb{R}^n \to \mathbb{R}$ which gives a continuous random variable $h(X_1, \ldots, X_n)$ on a n-dimensional sample space. Then*

$$E(h(X_1, \ldots, X_n)) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, \ldots, x_n) f_X(x_1;\theta) \cdots f_X(x_n;\theta) dx_1 \cdots dx_n.$$

Exercise 2. *Suppose X is uniform given with pdf $f_X(x;\theta) = \frac{1}{\theta}$ for $0 < x < \theta$, where $\theta > 0$. Consider the estimator $\hat{\theta}(X_1, X_2) = \frac{1}{4}(X_1 + X_2)^2$ for a random sample of size 2.*

(i) Compute $E(\hat{\theta})$ by reducing to calculations of terms of $E(X^k)$ as discussed above.

(ii) Compute $E(\hat{\theta})$ using the proposition above (i.e., Theorem 3.9.1 in the text), and check it agrees with your answer for (i).

In fact, there are fairly simple estimators, where the above proposition is the easiest way to calculate expected values. See, for example, Example 5.4.5 in the text. Here is another example where you should use the above proposition.

Exercise 3. *Let X and X_1, X_2 be as in the previous exercise, and consider the random variable $h(X_1, X_2) = \sqrt{X_1 + X_2}$. Compute $E(h(X_1, X_2))$. Is this a reasonable estimator?*

Using the above proposition, we can now give a proof of the Cramér-Rao inequality for an arbitrary sample size n.

Proof. (of Theorem for general n) Let

$$\mathbf{V}'(\theta) = \sum_{i=1}^{n} V'_{X_i}(\theta) = \sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f_X(x_i;\theta)}{f_X(x_i;\theta)}$$

One easily sees from the product rule

$$\mathbf{V}'(\theta) = \frac{\frac{\partial}{\partial\theta}[f_X(x_1;\theta) \cdots f_X(x_n;\theta)]}{f_X(x_1;\theta) \cdots f_X(x_n;\theta)}. \tag{2}$$

For any θ, this gives a random variable with an n-dimensional sample space.

By Lemma 1 and independence we still have

$$E(\mathbf{V}'(\theta)) = \sum_{i=1}^{n} E(V'_{X_i}(\theta)) = nE(V'_X(\theta)) = 0.$$

Thus, as in the $n = 1$ case,

$$Cov(\mathbf{V}'(\theta), \hat{\theta}) = E(\mathbf{V}'(\theta) \cdot \hat{\theta}) - E(\mathbf{V}'(\theta))E(\hat{\theta}) = E(\mathbf{V}'(\theta) \cdot \hat{\theta}).$$

By Lemma 3, we have

$$|E(\mathbf{V}'(\theta) \cdot \hat{\theta})| = |Cov(\mathbf{V}'(\theta), \hat{\theta})| \leq \sqrt{Var(\mathbf{V}'(\theta))Var(\hat{\theta})}.$$

By Corollary 1, i.e., $I(\theta) = Var(V'_X(\theta))$, we see ote

$$Var(\mathbf{V}'(\theta)) = \sum_{i=1}^{n} Var(V'_{X_i}(\theta)) = nI(\theta).$$

Hence

$$Var(\hat{\theta}) \geq \frac{|E(\mathbf{V}'(\theta) \cdot \hat{\theta})|^2}{Var(\mathbf{V}'(\theta))} = \frac{|E(\mathbf{V}'(\theta) \cdot \hat{\theta})|^2}{nI(\theta)},$$

where we used Corollary 1 for the equality on the right. Thus to prove the theorem, it suffices to show

$$|E(\mathbf{V}'(\theta) \cdot \hat{\theta})|^2 = 1.$$

Note Proposition 1 and (2) imply,

$$\begin{aligned}
E(\mathbf{V}'(\theta) \cdot \hat{\theta}) &= \int_a^b \cdots \int_a^b \left(\sum \frac{\frac{\partial}{\partial \theta} f_X(x_i;\theta)}{f_X(x_i;\theta)} \right) f_X(x_1;\theta) \cdots f_X(x_n;\theta) \hat{\theta}(x_1, \ldots, x_n) dx_1 \cdots dx_n \\
&= \int_a^b \cdots \int_a^b \frac{\partial}{\partial \theta} [f_X(x_1;\theta) \cdots f_X(x_n;\theta)] \hat{\theta}(x_1, \ldots, x_n) dx_1 \cdots dx_n \\
&= \frac{\partial}{\partial \theta} \int_a^b \cdots \int_a^b f_X(x_1;\theta) \cdots f_X(x_n;\theta) \hat{\theta}(x_1, \ldots, x_n) dx_1 \cdots dx_n \\
&= \frac{\partial}{\partial \theta} E(\hat{\theta}) = \frac{\partial}{\partial \theta} \theta = 1.
\end{aligned}$$

□