Derivation of the Cramér-Rao Bound

Ryan D. Reece

October 1, 2009

Abstract

I give a pedagogical derivation of the Cramér-Rao Bound, which gives a lower bound on the variance of estimators used in statistical point estimation, commonly used to give numerical estimates of the systematic uncertainties in a measurement.

1 Derivation

For estimators $\hat{\theta}_i$ of parameters θ_i in a given model with liklihood function L, the bias of these estimators b_i is defined as

$$b_i \equiv \mathrm{E}[\hat{\theta}_i(x) - \theta_i] \equiv \int dx\, L(x,\theta)\, (\hat{\theta}_i(x) - \theta_i)$$

Note that the estimators $\hat{\theta}_i$ depend on the observable x, and the likelihood function L depends on x and the paramters of the model θ. For ease of notation these dependencies will now be left implicit. The observable could be a tuple of many measurements $x = \{x_1, x_2, \ldots x_N\}$, in which case the integral over x actual denotes integrals over each independent measurement.

$$\int dx = \int dx_1 \int dx_2 \cdots \int dx_N$$

In that case, the likelihood function is the joint likelihood function, which is a product of the likelihood functions for each measurement.

$$L = \prod_{i=1}^{N} L_i$$

Taking the derivative of the bias with respect to its corresponding parameter gives

$$\frac{\partial b_i}{\partial \theta_i} = \int dx\, (\hat{\theta}_i - \theta_i) \underbrace{\frac{\partial L}{\partial \theta_i}}_{L\,\frac{\partial \ln L}{\partial \theta_i}} - \underbrace{\int dx\, L}_{1}$$

$$1 + \frac{\partial b_i}{\partial \theta_i} = \int dx\, (\hat{\theta}_i - \theta_i)\, L\, \frac{\partial \ln L}{\partial \theta_i}$$

$$\left(1 + \frac{\partial b_i}{\partial \theta_i}\right)\left(1 + \frac{\partial b_j}{\partial \theta_j}\right) = \left[\int dx\, (\hat{\theta}_i - \theta_i)\, L\, \frac{\partial \ln L}{\partial \theta_i}\right]\left[\int dx\, (\hat{\theta}_j - \theta_j)\, L\, \frac{\partial \ln L}{\partial \theta_j}\right]$$

Then we use the Cauchy-Schwarz inequality:

$$\left|\int dx\, f(x)\, g(x)\right|^2 \leq \int dx\, |f(x)|^2 \cdot \int dx\, |g(x)|^2$$

$$\left(1 + \frac{\partial b_i}{\partial \theta_i}\right)\left(1 + \frac{\partial b_j}{\partial \theta_j}\right) \leq \left[\int dx\, L\, (\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j)\right]\left[\int dx\, L\, \frac{\partial \ln L}{\partial \theta_i}\, \frac{\partial \ln L}{\partial \theta_j}\right]$$

The first integral is the covariance of the estimators.

$$V_{ij} \equiv \mathrm{Cov}[\hat{\theta}_i, \hat{\theta}_j] \equiv \int dx\, L\, (\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j)$$

The second integral is defined as the *Fisher information matrix*.

$$I_{ij} \equiv \mathrm{E}\left[\frac{\partial \ln L}{\partial \theta_i}\, \frac{\partial \ln L}{\partial \theta_j}\right] \equiv \int dx\, L\, \frac{\partial \ln L}{\partial \theta_i}\, \frac{\partial \ln L}{\partial \theta_j}$$

A more convenient and equivalent expression for I_{ij} can be derived as follows.

Consider the following.

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial^2 \ln L}{\partial\theta_i\,\partial\theta_j}\right] &= \mathrm{E}\left[\frac{\partial}{\partial\theta_i}\left(\frac{1}{L}\,\frac{\partial L}{\partial\theta_j}\right)\right] \\
&= \mathrm{E}\left[-\frac{1}{L^2}\,\frac{\partial L}{\partial\theta_i}\,\frac{\partial L}{\partial\theta_j} + \frac{1}{L}\,\frac{\partial}{\partial\theta_i}\,\frac{\partial L}{\partial\theta_j}\right] \\
&= -\mathrm{E}\left[\frac{\partial \ln L}{\partial\theta_i}\,\frac{\partial \ln L}{\partial\theta_j}\right] + \mathrm{E}\left[\frac{1}{L}\,\frac{\partial}{\partial\theta_i}\left(L\,\frac{\partial \ln L}{\partial\theta_j}\right)\right] \\
&= -I_{ij} + \int dx\; L\frac{1}{L}\,\frac{\partial}{\partial\theta_i}\left(L\,\frac{\partial \ln L}{\partial\theta_j}\right) \\
&= -I_{ij} + \frac{\partial}{\partial\theta_i}\int dx\; L\,\frac{\partial \ln L}{\partial\theta_j} \\
&= -I_{ij} + \frac{\partial}{\partial\theta_i}\int dx\; L\,\frac{1}{L}\,\frac{\partial L}{\partial\theta_j} \\
&= -I_{ij} + \frac{\partial}{\partial\theta_i}\,\frac{\partial}{\partial\theta_j}\int dx\; L \\
&= -I_{ij} + \frac{\partial}{\partial\theta_i}\,\frac{\partial}{\partial\theta_j}(1) = -I_{ij}
\end{aligned}$$

Therefore

$$I_{ij} \equiv \mathrm{E}\left[\frac{\partial \ln L}{\partial\theta_i}\,\frac{\partial \ln L}{\partial\theta_j}\right] \equiv \mathrm{E}\left[-\frac{\partial^2 \ln L}{\partial\theta_i\,\partial\theta_j}\right]$$

One can see that the Fisher information matrix measures the curvature of the likelihood function in parameters space, averaged over the possible observed data. Intuitively, this means the larger the value of I_{ij}, the more sharply pronounced the likelihood function is in that region, and the more sensitive the experiment is to the parameters of the model.

Plugging these expressions back into the inequality we derived by using the Cauchy-Schwarz inequality, we have the general expression for the *Cramér-Rao bound*.

$$\boxed{V_{ij} \geq \frac{\left(1+\frac{\partial b_i}{\partial\theta_i}\right)\left(1+\frac{\partial b_j}{\partial\theta_j}\right)}{\mathrm{E}\left[-\frac{\partial^2 \ln L}{\partial\theta_i\,\partial\theta_j}\right]}} \tag{1}$$

It is often the case that the bias of a well chosen estimator is asymtotically zero in the large sample limit. In which case, the bound for the covariance

matrix of unbiased estimators simplifies to the following.

$$V_{ij} \geq \left(\mathrm{E}\left[-\frac{\partial^2 \ln L}{\partial\theta_i \, \partial\theta_j} \right] \right)^{-1}$$

The diagonal elements of which give the variance of the estimators.

$$V_{ii} = \sigma^2_{\hat{\theta}_i} \geq \left(\mathrm{E}\left[-\frac{\partial^2 \ln L}{\partial\theta_i^2} \right] \right)^{-1}$$

The *efficiency* of an unbiased estimator is defined as

$$\varepsilon(\hat{\theta}_i) \equiv \frac{\left(\mathrm{E}\left[-\frac{\partial^2 \ln L}{\partial\theta_i^2} \right] \right)^{-1}}{\sigma^2_{\hat{\theta}_i}}$$

The efficiency is therefore less than or equal to one, and equal to one in the case that the Cramér-Rao bound becomes an equality.

If the estimators are unbiased and efficient ($\varepsilon = 1$), it is evident that the covariance matrix of the estimators is given by the following.

$$V_{ij} = \left(\mathrm{E}\left[-\frac{\partial^2 \ln L}{\partial\theta_i \, \partial\theta_j} \right] \right)^{-1}$$

It can be shown that in the large sample limit, Maximum Likelihood Estimators (MLE) are asymptotically unbiased and efficient.

Since integrating the second derivatives of the likelihood function over all possible observables is often prohibitive in practice, one often estimates this expectation value by the numerically determined second derivatives evaluated at the point of maximum likelihood.

$$V_{ij} \approx \left(-\left. \frac{\partial^2 \ln L}{\partial\theta_i \, \partial\theta_j} \right|_{\theta=\hat{\theta}} \right)^{-1}$$

Therefore, the variance of the estimators can be estimated by

$$\boxed{\sigma^2_{\hat{\theta}_i} \approx \left(-\left. \frac{\partial^2 \ln L}{\partial\theta_i^2} \right|_{\theta_i=\hat{\theta}_i} \right)^{-1}} \tag{2}$$

2 Example

The purpose of equation 2 is to estimate the variance of an estimator when an analytic calculation is not practical. In this example, however, we will study a case where an analytic calculation of the variance *is* trivial such that we make the validity of equation 2 apparent.

Consider an experiment with N repeated measurements that are Gaussian distributed. The likelihood function is therefore

$$L = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

The MLE for the mean, μ, can be found by maximizing the likelihood function, or equivalently, its natural logrhythm.

$$\ln L = -N \ln(\sigma\sqrt{2\pi}) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$0 = \frac{\partial \ln L}{\partial \mu} = \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2}$$

$$\Rightarrow \qquad \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Therefore the MLE of μ is just the mean of the sample, as one might expect. Calculating the second derivative of the likelihood gives

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{N}{\sigma^2}$$

Therefore, equation 2 gives the following for the variance of this estimator

$$\sigma^2_{\hat{\mu}} = \frac{\sigma^2}{N} \tag{3}$$

which can easily be shown to be the variance of the sample mean of *any* distribution as follows.

Let

$$\bar{x} \equiv \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\mathrm{E}\left[\bar{x}\right] = \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}\left[x_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mu = \mu$$

Thefore, $\bar{x}$ is an unbiased estimator of the mean. Now we calculate the variance.

$$\begin{aligned}
\mathrm{V}\left[\bar{x}\right] &= \mathrm{E}\left[\bar{x}^2\right] - \left(\mathrm{E}\left[\bar{x}\right]\right)^2 \\
&= \mathrm{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)\left(\frac{1}{N}\sum_{j=1}^{N} x_j\right)\right] - \mu^2 \\
&= \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathrm{E}\left[x_i^2\right] + \sum_{i\neq j}^{N}\mathrm{E}\left[x_i\, x_j\right]\right) - \mu^2
\end{aligned}$$

For the first sum, note that

$$\mathrm{V}\left[x\right] = \sigma^2 = \mathrm{E}\left[x^2\right] - \left(\mathrm{E}\left[x\right]\right)^2 = \mathrm{E}\left[x^2\right] - \mu^2$$

$$\Rightarrow \qquad \mathrm{E}\left[x^2\right] = \mu^2 + \sigma^2$$

For the second sum, recall that the individual measurements x_i are made independently, therefore

$$\mathrm{E}\left[x_1\, x_2\right] = \mathrm{E}\left[x_1\right]\ \mathrm{E}\left[x_2\right] = \mu^2$$

Therefore

$$\mathrm{V}\left[\bar{x}\right] = \frac{1}{N^2}\left(\mu^2 + \sigma^2 + (N^2 - N)\mu^2\right) - \mu^2 = \frac{\sigma^2}{N}$$

in agreement with the variance we calculated using the Cramér-Rao bound (equation 3), implying that $\hat{\mu}$ is indeed an efficient estimator.