The CRAMÉR-RAO INEQUALITY

Maarten Jansen [1], Université libre de Bruxelles, Belgium

Gerda Claeskens [2] , KU Leuven, Belgium

1 The Cramér-Rao lower bound

The Cramér-Rao inequality gives a lower bound for the variance of an unbiased estimator of a parameter. It is named after work by Cramér (1946) and Rao (1945). The inequality and the corresponding lower bound in the inequality are stated for various situations. We will start with the case of a scalar parameter and independent and identically distributed random variables $X_1, \ldots, X_n$, with the same distribution as X.

Denote $\boldsymbol{X} = (X_1, \ldots, X_n)$ and denote the common probability mass function or probability density function of X at a value x by $f(x;\theta)$ where $\theta \in \Theta$, which is a subset of the real line $\mathbb{R}$ and $x \in \mathbb{R}$. Denote the support of X by R, that is, $R = \{x : f(x;\theta) > 0\}$.

Assumptions.

(i) The partial derivative $\frac{\partial}{\partial\theta} \log f(x;\theta)$ exists for all $\theta \in \Theta$ and all $x \in R$ and it is finite. This is equivalent to stating that the Fisher information value $I_X(\theta) = E\left[\left(\frac{\partial}{\partial\theta} \log f(X;\theta)\right)^2\right]$ is well defined, for all $\theta \in \Theta$.

(ii) The order of integration and differentiation in $\int \frac{\partial}{\partial\theta} \log f(x;\theta)dx$ is interchangeable. If the support of X, that is, the set R, is finite, then the interchangeability is equivalent with the condition that the support does not depend on θ. A counter-example on uniformly distributed random variables is elaborated below.

[1] Maarten Jansen is Professor at the Departments of Mathematics and Computer Science of the Université libre de Bruxelles (Belgium). He is Elected member of the International Statistical Institute and author of three books: *Wavelets from a statistical perspective* (CRC Press, 2022), *Second generation wavelets and applications* (with P. Oonincx, Springer Verlag, 2005) and *Noise reduction by wavelet thresholding* (Springer Verlag, 2001), and over 30 journal papers.

[2] Gerda Claeskens is Professor at ORStat, Faculty of Economics and Business and at the Leuven Statistics Research Center of the KU Leuven (Belgium). She is Elected member of the International Statistical Institute, fellow of the American Statistical Association (2019) and the Institute of Mathematical Statistics (2014), and recipient of the Noether Young Scholar Award (2004). She is the author of about 90 journal papers and of the book *Model selection and model averaging* (with N.L. Hjort, Cambridge University Press, 2008). Currently she is Associate editor of the *Journal of the American Statistical Association*, *TEST*, and *International Statistical Review.*

The Cramér-Rao inequality

Under assumptions (i) and (ii), if $\hat{\theta} = g(\boldsymbol{X})$ is an unbiased estimator of θ, this means that $E[\hat{\theta}] = \theta$, then

$$\mathrm{var}(\hat{\theta}) \geq 1/\left[n \cdot I_X(\theta)\right].$$

The lower bound in this inequality is called the Cramér-Rao lower bound.

The proof starts by realizing that the correlation of the score $V = \frac{\partial}{\partial\theta} \sum_{i=1}^{n} \log f_X(X_i;\theta)$ and the unbiased estimator $\hat{\theta}$ is bounded above by 1. This implies that $\left(\mathrm{var}(V) \cdot \mathrm{var}(\hat{\theta})\right)^{1/2} \geq \mathrm{cov}(V, \hat{\theta})$. The assumptions are needed to prove that the expected score $E(V)$ is zero. This implies that the covariance $\mathrm{cov}(V, \hat{\theta}) = 1$, from which the stated inequality readily follows.

A second version of the Cramér-Rao inequality holds if we estimate a functional $\kappa = H(\theta)$. Under assumptions (i) and (ii), if $\boldsymbol{X}$ is a sample vector of independent observations from random variable X with density function $f(x;\theta)$ and $\hat{\kappa} = h(\boldsymbol{X})$ is an unbiased estimator of $H(\theta)$, such that the first derivative $\frac{dH(\theta)}{d\theta}$ exists and is finite for all θ, then

$$\mathrm{var}(\hat{\kappa}) \geq \left[\frac{dH(\theta)}{d\theta}\right]^2 / \left[n \cdot I_X(\theta)\right].$$

Similar versions of the inequality can be phrased for observations that are independent but not identically distributed.

In the case of a vector parameter $\boldsymbol{\theta}$, the variance of the single parameter estimator $\mathrm{var}(\hat{\theta})$ is replaced by the covariance matrix of the estimator vector $\Sigma_{\widehat{\boldsymbol{\theta}}}$. This matrix is bounded by a matrix expression containing the inverse of the Fisher information matrix, where bounded means that the difference between the covariance matrix and its "upper bound" is negative semidefinite matrix.

The Cramér-Rao inequality is important because it states what the best attainable variance is for unbiased estimators. Estimators that actually attain this lower bound are called efficient. It can be shown that maximum likelihood estimators asymptotically reach this lower bound, hence are asymptotically efficient.

2 Cramér-Rao and UMVUE

If $\boldsymbol{X}$ is a sample vector of independent observations from the random variable X with density function $f_X(x;\theta)$ and $\hat{\theta} = g(\boldsymbol{X})$ is an unbiased estimator of θ, then $\mathrm{var}(\hat{\theta}) = 1/\left[n \cdot I_X(\theta)\right] \Leftrightarrow \hat{\theta} = aV + b$ with probability one, where V is the score and a and b are some constants. This follows from the proof of the Cramér-Rao inequality: the lower bounded is reached if the correlation between the score and the estimator is one. This implies that $\mathrm{var}\left(\frac{V}{\sigma_V} + \frac{\hat{\theta}}{\sigma_{\hat{\theta}}}\right) = 0 \Rightarrow \frac{V}{\sigma_V} + \frac{\hat{\theta}}{\sigma_{\hat{\theta}}} = c$ almost surely for some constant c. We here used the notation σ_X to denote the standard deviation of a random variable X.

The coefficients a and b may depend on θ, but $\hat{\theta}$ should be observable without knowing θ.

If a and b exist such that $\hat{\theta}$ is unbiased and observable, then $\hat{\theta}$ has the smallest possible variance among all unbiased estimators: it is then certainly the uniformly minimum variance unbiased estimator (UMVUE).

It may, however, be well possible that no a and b can be found. In that case, the UMVUE, if it exists, does not reach the Cramér-Rao lower bound. In that case, the notion of *sufficiency* can be used to find such UMVUE.

Counter example: estimators for the upperbound of uniform data

Let $X \sim \text{unif}[0,a]$, so $f_X(x) = \frac{1}{a}I(0 \leq x \leq a)$, where $I(c \leq x \leq d)$ is the *indicator* function of the interval $[c,d]$. We want to estimate a. The maximum likelihood estimator (MLE) is $\hat{a}_{\text{MLE}} = \max\limits_{i=1,\dots,n} X_i$, which is biased. Define $\hat{a}_u = \frac{n}{n-1}\hat{a}_{\text{MLE}}$, which is unbiased. The method of moments leads to an estimator $\hat{a}_{\text{MME}} = 2\overline{X}$, which is also unbiased. The score is $V_i = \frac{\partial}{\partial a}\log f_X(X_i;a) = -\frac{1}{a}$. This is a constant (so, not a random variable), whose expected value is of course *not zero*. This is because the partial derivative and expectation cannot be interchanged, as the boundary of the support of X depends on a. As a consequence, the Cramér-Rao lower bound is *not* valid here. We can verify that $\text{var}(\hat{a}_{\text{MLE}}) = \frac{n}{(n+2)(n+1)^2}a^2$ and $\text{var}(\hat{a}_u) = \frac{1}{n(n+2)}a^2$. This is (for $n \to \infty$) one order of magnitude smaller than $\text{var}(\hat{a}_{\text{MME}}) = \frac{1}{3n}a^2$ and also one order of magnitude smaller than what you would expect for an unbiased estimator if the Cramér-Rao inequality would hold.

3 A Bayesian Cramér-Rao bound

It should be noted that biased estimators can have variances below the Cramér-Rao lower bound. Even the MSE (mean squared error), which equals the sum of the variance and the squared bias can be lower than the Cramér-Rao lower bound (and hence lower than any unbiased estimator could attain). A notable example in this respect is Stein's phenomenon on shrinkage rules (Efron and Morris, 1977).

In practice, large classes of estimators, for example most nonparametric estimators, are biased. An inequality that is valid for biased or unbiased estimators is due to van Trees (1968, p. 72), see also Gill and Levit (1995) who developed multivariate versions of the inequality.

We assume that the parameter space Θ is a closed interval on the real line and denote by g some probability distribution on Θ with density $\lambda(\theta)$ with respect to the Lebesgue measure. This is where the Bayesian flavor enters. The θ is now treated as a random variable with density λ. We assume that λ and $f(x;\cdot)$ are absolutely continuous and that λ converges to zero at the endpoints of the interval Θ. Moreover we assume that $E[\frac{\partial}{\partial\theta}\log f(X;\theta)] = 0$. We denote $I(\lambda) = E[\{\log\lambda(\theta)\}^2]$ and have that $E[I_X(\theta)] = \int I_X(\theta)g(\theta)d\theta$. Then, for an estimator $\hat{\theta} = \hat{\theta}(X)$, it holds that

$$E[\{\hat{\theta} - \theta\}^2] \geq \frac{1}{E[I_X(\theta)] + I(\lambda)}.$$

A second form of this inequality is obtained for functionals $\kappa = H(\theta)$. Under the above assumptions, for an estimator $\hat{\kappa} = h(\boldsymbol{X})$ of $H(\theta)$, such that the first derivative $\frac{dH(\theta)}{d\theta}$ exists and is finite for all θ,

$$E[\{\hat{\kappa} - H(\theta)\}^2] \geq \frac{\{E[\frac{d}{d\theta}H(\theta)]\}^2}{E[I_X(\theta)] + I(\lambda)}.$$

References

[1] Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press.

[2] Efron, B. and Morris, C. (1977). *Stein's paradox in statistics.* Scientific American, vol. 236, pp. 119–127.

[3] Gill, R.D. and Levit, B.Y. (1995). Applications of the van Trees inequality: a Bayesian Cramér-Rao bound, *Bernoulli*, **1**, no. 1–2, 59–79.

[4] Rao, C. (1945). Information and the accuracy attainable in the estimation of statistical parameters, *Bulletin of the Calcutta Mathematical Society*, **37**, 81–89.

[5] van Trees, H.L. (1968). *Detection, estimation and modulation theory: part I*, Wiley.