Relative Efficiency, Efficiency, and the Fisher Information

Guy Lebanon

April 19, 2006

In point estimation, we use an estimator $\hat{\theta}$ which is a function of the data $X_1, \ldots, X_n$, to estimate a parameter of interest θ. In a previous note, we saw that the expected squared error of an estimator decomposes as a sum of the bias squared and the variance

$$MSE(\hat{\theta}) = E((\hat{\theta}(X_1, \ldots, X_n) - \theta)^2) = \mathsf{Bias}^2(\hat{\theta}) + \mathsf{Var}\,(\hat{\theta}).$$

We can compare the quality of two estimators by looking at the ratio of their MSE. If the two estimators are unbiased this is equivalent to the ratio of the variances which is defined as the relative efficiency.

Definition 1. *The relative efficiency of two unbiased estimators $\hat{\theta}_1, \hat{\theta}_2$ is the ratio of their variances $\frac{\mathsf{Var}\,(\hat{\theta}_1)}{\mathsf{Var}\,(\hat{\theta}_2)}$.*

In general the relative efficiency is a function of θ, and so some estimators will have lower MSE than others for some values of θ but not for other values of θ. In some cases, however, the relative efficiency does not depend on θ and then in points at a clear advantage of one estimator over another in terms of the MSE.

Example: Consider two estimators for the parameter θ of a uniform distribution $U(0, \theta)$: $\hat{\theta}_1 = 2\bar{X}$ and $\hat{\theta}_2 = \frac{n+1}{n} X_{(n)}$ where $X_{(n)}$ is the maximum of the data $X_1, \ldots, X_n$. $\hat{\theta}_1$ is unbiased as $\mathsf{E}\,(\hat{\theta}_1) = 2\mathsf{E}\,(X_1) = 2\theta/2 = \theta$ and has variance

$$\mathsf{Var}\,(\hat{\theta}_1) = 4\mathsf{Var}\,(\bar{X}) = \frac{4}{n}\mathsf{Var}\,(X_1) = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

To find the expectation and variance of $\hat{\theta}_2$ first recall that the cdf of $X_{(n)}$ is

$$F_{X_{(n)}}(r) = P(X_i \leq n : \forall i) = \prod_i F_{X_i}(r) = (F_{X_1}(r))^n = \frac{r^n}{\theta^n}$$

and the pdf of $X_{(n)}$ is $f_{X_{(n)}}(r) = n\frac{r^{n-1}}{\theta^n}$. $\hat{\theta}_2$ is unbiased as

$$\mathsf{E}\,(\hat{\theta}_2) = \frac{n+1}{n}\frac{n}{\theta^2}\int_0^\theta r^n dr = \frac{n+1}{n} \cdot \frac{n}{n+1}\theta.$$

Using $\mathsf{E}\,(X_{(n)}^2) = \frac{n}{\theta^n}\int_0^\theta r^n r dr = \frac{n}{n+2}\theta^2$ we have that

$$\mathsf{Var}\,(\hat{\theta}_2) = \frac{(n+1)^2}{n^2}\mathsf{Var}\,(X_{(n)}) = \frac{(n+1)^2}{n^2}(\mathsf{E}\,(X_{(n)}^2) - (\mathsf{E}\,(X_{(n)}))^2) = \frac{(n+1)^2}{n^2}\theta^2\frac{n}{(n+2)(n+1)^2} = \frac{\theta^2}{n(n+2)}$$

and the relative efficiency is

$$\frac{\mathsf{Var}\,(\hat{\theta}_1)}{\mathsf{Var}\,(\hat{\theta}_2)} = \frac{\theta^2/3n}{\theta^2/(n(n+2))} = \frac{n+2}{3}$$

indicating that for $n > 1$, $\hat{\theta}_2$ has a lower variance.

The following theorem bounds the variance of estimators, and enables us to assert in some cases (when we find an estimator whose variance equals the lower bound) that we have an estimator with the lowest possible variance.

Definition 2. *For one dimensional parametric family with a pdf or pmf $f(x\,;\theta)$, $\theta \in \Theta \subset \mathbb{R}$, we define the Fisher information to be the following function*

$$I(\theta) = \mathsf{E}\left(\left(\frac{d}{d\theta}\log f(x\,;\theta)\right)^2\right).$$

Note that the above definition makes sense for both continuous and discrete RVs. The pdf or pmf, however, needs to be continuous and differentiable with respect to θ. An intuitive interpretation behind the Fisher information is that is serves as a quantity that determines the information that an observation X conveys with respect to estimating θ.

Theorem 1 (Cramer-Rao Lower Bound)**.** *For an arbitrary estimator*

$$\mathsf{Var}\,(\hat\theta) \geq \frac{(\frac{d}{d\theta}E(\hat\theta))^2}{nI(\theta)}.$$

Note that for unbiased estimators $\mathsf{E}\,(\hat\theta) = \theta$ and the numerator is 1.

Definition 3. *If $\hat\theta$ is an estimator whose variance achieves equality in the Cramer Rao lower bound (for all θ), it is called efficient.*

Note that an efficient estimator is an estimator with lowest possible variance. If it also has bias 0, it is the best estimator in terms of the MSE. Before we prove the theorem we establish several useful results (for a discrete RV replace integrals with sums)

$$\mathsf{E}\left(\frac{d}{d\theta}\log f(x\,;\theta)\right) = \int f(x\,;\theta)\frac{d}{d\theta}\log f(x\,;\theta)dx = \int \frac{d}{d\theta}f(x\,;\theta)dx = \frac{d}{d\theta}\int f(x\,;\theta)dx = \frac{d}{d\theta}1 = 0$$

$$I(\theta) = E\left(\left(\frac{d}{d\theta}\log f(x\,;\theta)\right)^2\right) - 0 = \mathsf{Var}\left(\frac{d}{d\theta}\log f(x\,;\theta)\right)$$

$$I(\theta) = -\int \frac{d^2}{d\theta^2}f(x\,;\theta)dx + \mathsf{E}\left(\left(\frac{d}{d\theta}\log f(x\,;\theta)\right)^2\right) = -\mathsf{E}\left(\frac{f(x\,;\theta)\frac{d^2}{d\theta^2}f(x\,;\theta) - (d/d\theta f(x\,;\theta))^2}{f(x\,;\theta)^2}\right)$$

$$= -\mathsf{E}\left(\frac{d^2}{d\theta^2}\log f(x\,;\theta)\right).$$

Proof. We proved previously that the correlation coefficient is in $[-1, 1]$ which implies that

$$\mathsf{Cov}\left(\hat\theta, \frac{d}{d\theta}\sum_i \log f(x_i\,;\theta)\right)^2 \leq \mathsf{Var}\,(\hat\theta)\ \mathsf{Var}\left(\frac{d}{d\theta}\sum_i \log f(x_i\,;\theta)\right) = \mathsf{Var}\,(\hat\theta)nI(\theta).$$

To complete the proof we need

$$\mathsf{Cov}\left(\hat\theta, \frac{d}{d\theta}\sum_i \log f(x_i\,;\theta)\right) = \mathsf{E}\left(\hat\theta\ \frac{d}{d\theta}\sum_i \log f(x_i\,;\theta)\right) - 0 = \int \prod_i f(x_i\,;\theta)\hat\theta\ \frac{d}{d\theta}\log\prod_i f(x_i\,;\theta)dx_1\cdots dx_n$$

$$= \int \hat\theta(x_1,\ldots,x_n)\ \frac{d}{d\theta}\prod_i f(x_i\,;\theta)\ dx_1\cdots dx_n = \frac{d}{d\theta}\mathsf{E}\,(\hat\theta)$$

□

There exists a multi-parameter definition of the Fisher information, which leads to a multi-parameter version of the Cramer-Rao lower bound, but we will not pursue it here.

Example: For the Bernoulli distribution we have $\log f(x\,;\theta) = x\log\theta + (1-x)\log(1-\theta)$ and

$$I(\theta) = -\mathsf{E}\left(\frac{d^2}{d\theta^2}\log f(x\,;\theta)\right) = \mathsf{E}\left(\frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}\right) = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}.$$

Since $\mathsf{E}\,(\bar X) = \theta$ and $\mathsf{Var}\,(\bar X) = n^{-1}\mathsf{Var}\,(X_1) = n^{-1}\theta(1-\theta)$, we have that $\hat\theta = \bar X$ is an efficient estimator.