EFFICIENCY AND THE CRAMÉR-RAO LOWER BOUND

ROD MCBANE AND ADAM DRAKE

1. Efficiency

Definition 1.1. The relative efficiency of an estimator is a term used to compare two estimators where a measure of relative efficiency is needed. For two unbiased estimators T_1 and T_2 of θ, the efficiency of T_1 relative to T_2 is defined to be

$$e(T_1, T_2) = \frac{Var(T_2)}{Var(T_1)} = \frac{mseT_2}{mseT_1} = \frac{E[(T_2-\theta)^2]}{E[(T_1-\theta)^2]} \tag{1}$$

and T_2 is more efficient than T_1 if $Var(T_2) < Var(T_1)$.

Since the mean squared error (mse) $E(d-\theta)^2$ becomes the variance of d when $E(d) = \theta$, the basis for comparing unbiased estimators is the variance of those estimators.

Example 1.1. For $X_1, \ldots, X_n$ a random sample from $U(0, \theta)$ with $Y_1, \ldots, Y_n$ the corresponding ordered sample and $T_1 = 2\bar{X}, T_2 = \frac{n+1}{n} Y_n$ unbiased estimates of θ, find $e(T_1, T_2)$.

First, note that $Var(T_1) = Var(2\bar{X}) = 4Var(\bar{X}) = \frac{4Var(X_i)}{n} = \frac{4\theta^2}{12n} = \frac{\theta^2}{3n}$

Finding $Var(T_2)$ requires that we find $E(Y_n^2)$

$$E(Y_n^2) = \int_0^\theta y^2 \frac{n}{\theta^n} y^{n-1} dy = \frac{n}{\theta^2}\frac{\theta^{n+2}}{n+2} = \frac{n}{n+2}\theta^2$$

Then we have that

$$Var(Y_n) = \frac{n}{n+2}\theta^2 - \frac{n^2\theta^2}{(n+1)^2} = \frac{n\theta^2}{(n+1)^2(n+2)}$$

So then

$$Var(T_2) = \frac{(n+1)^2}{n^2}\frac{n\theta^2}{(n+1)^2(n+2)} = \frac{\theta^2}{n(n+2)}$$

Now since these estimates are unbiased we can use (1) to determine that

$$e(T_1, T_2) = \frac{Var(T_2)}{Var(T_1)} = \frac{\theta^2}{n(n+2)} \frac{3n}{\theta^2} = \frac{3}{n+2}$$

and this is less than 1 for $n > 1$ so T_2 is more efficient than T_1.

Definition 1.2. The score or score function is the partial derivative, with respect to some parameter θ, of the logarithm (commonly the natural logarithm) of the likelihood function. If the observation is X and its likelihood is $L(\theta, X)$, then the score V can be found through the chain rule:

$$V = \frac{\partial}{\partial\theta} log L(\theta, X) = \frac{1}{L(\theta, X)} \frac{\partial L(\theta, X)}{\partial\theta} \tag{2}$$

Note that V is a function of θ and the observation X, so that, in general, it is not a statistic and that the mean of the score function is 0.

Proposition 1.1. *The mean of the score function is 0.*

Proof. First, we rewrite the definition of expectation, using the fact that the probability density function is just $L(\theta; x)$, which is conventionally denoted by $f(x; \theta)$. The corresponding cumulative distribution function is denoted $F(x; \theta)$. With this change of notation and writing $f'_\theta(x; \theta)$ for the partial derivative with respect to θ,

$$\mathbb{E}[V|\theta] = \int_{[0,1]} \frac{f'_\theta(x;\theta)}{f(x;\theta)} dF(x;\theta) = \int_X \frac{f'_\theta(x;\theta)}{f(x;\theta)} f(x;\theta)dx = \int_X \frac{\partial f(x;\theta)}{\partial\theta} dx$$

where the integral runs over the whole of the probability space of X and a prime denotes partial differentiation with respect to θ. If certain differentiability conditions are met, the integral may be rewritten as

$$\frac{\partial}{\partial\theta} \int_X f(x;\theta)dx = \frac{\partial}{\partial\theta} 1 = 0$$

□

Thus, the expected value of the score is zero. In other words, repeatedly sampling from some distribution and calculating the score with the true θ would result in the mean value of the scores tending to zero as the number of repeat samples approached infinity.

Definition 1.3. The Fisher Information is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ upon which the likelihood function of θ, $L(\theta) = f(X; \theta)$, depends. Since the expectation of the score is zero, the variance is simply the second moment of the score, which is the derivative of the log of the likelihood function with respect to θ. We can write this as

$$\mathcal{I}(\theta) = E\left\{\left[\frac{\partial}{\partial\theta} ln f(X;\theta)\right]^2 |\theta\right\}$$

which implies $0 \leq \mathcal{I}(\theta) < \infty$. The Fisher information is thus the expectation of the squared score.

Now that we have developed the notion of relative efficiency we can choose between two estimators but we have no assurance that the prevailing estimator is particularly good. For this we turn to the Cramér-Rao Lower Bound which tells us that there is not another estimator whose variance (or mse) is smaller than the two considered.

2. The Cramér-Rao Lower Bound

Using V as defined in (2) and noting that $E[V] = 0$ so $Var(V) = E[V^2]$ and

$$cov(V,T) = E[VT] = E[T\frac{\partial}{\partial\theta} log f(X,\theta)] = \int \dots \int t(x) \frac{\frac{\partial}{\partial\theta} f(x,\theta)}{f(x,\theta)} f(x,\theta) dx$$

Then assuming regularity conditions where the interchange of integration and differentiation is permitted and on the existence and integrability of any needed partial derivative we have:

$$\int \dots \int t(x) \frac{\frac{\partial}{\partial\theta} f(x,\theta)}{f(x,\theta)} f(x,\theta) dx = \frac{\partial}{\partial\theta} E[T] = \frac{\partial}{\partial\theta}[\theta + b_T(\theta)] = 1 + b'_T(\theta)$$

Here b_T is the bias of the estimator which is equal to $E[T] - \theta$.

Now recall that the absolute value of the correlation coefficient, for measuring correlation between any two variables is less than or equal to 1 and that

$$\rho_{V,T} = \frac{cov(V,T)}{\sigma_V \sigma_T}$$

so then we have

$$[cov(V,T)]^2 \leq Var(V)Var(T)$$

or

$$Var(T) \geq \frac{[cov(V,T)]^2}{Var(V)} = \frac{[1 + b'_T(\theta)]^2}{\mathcal{I}_X(\theta)}$$

Now in the case of an unbiased estimator, where $b_T = 0$ and therefore all derivatives of b_T are zero we have

$$Var(T) = mse(T) \geq \frac{1}{\mathcal{I}_X(\theta)}$$

which is known as the Cramèr-Rao lower bound, the minimum variance bound (MVB), or sometimes the Information Inequality.

Definition 2.1. The absolute efficiency of an unbiased estimator T of a parameter θ when the Cramér-Rao inequality is valid is

$$e(T) = \frac{1/\mathcal{I}(\theta)}{Var(T)}$$

if this ratio is independent of θ. When $e(T) = 1$ then T is said to be efficient.

2.1. **Regularity Conditions.** The bound relies on two weak regularity conditions on the PDF $f(x;\theta)$ and the estimator $T(X)$:

- The Fisher information is always defined, that is, for all x such that $f(x;\theta) > 0$ we have that
$$\frac{\partial}{\partial\theta} lnf(x;\theta)$$
exists and is finite.
- The operations of integration with respect to x and differentiation with respect to θ can be interchanged in the expectation of T, that is:
$$\frac{\partial}{\partial\theta}\left[\int T(x)f(x;\theta)dx\right] = \int T(x)\left[\frac{\partial}{\partial\theta}f(x;\theta)\right]dx$$
whenever the right-hand side is finite.

This condition can often be confirmed by using the fact that integration and differentiation can be swapped when either of the following cases hold:

(1) The function $f(x;\theta)$ has bounded support in x and the bounds do not depend on θ;

(2) The function $f(x;\theta)$ has infinite support, is continuously differentiable, and the integral converges uniformly for all θ.

Example 2.1. Suppose that $\underset{\sim}{X} = (X)$ is a single observation from $Bin(m,p)$ where m is known. The PMF is given by

$$f(x,p) = \binom{m}{x} p^x (1-p)^{m-x}$$

where $x = 0, 1, \ldots, m$.

Note that the range of X depends on m but not on the unknown parameter p. Also, the sample size is $n = 1$.

Since the range of X does not depend on the unknown parameter p which we wish to estimate, we can proceed to compute and use the Cramér-Rao lower bound for unbiased estimators:

$$logf(x,p) = log\binom{m}{x} + xlogp + (m-x)log(1-p)$$

$$\frac{\partial}{\partial p}logf(x,p) = \frac{x}{p} - \frac{m-x}{1-p} = \frac{x-mp}{p(1-p)}$$

$$\left(\frac{\partial}{\partial p}logf(x,p)\right)^2 = \frac{(x-mp)^2}{p^2(1-p)^2}$$

Thus

$$E\left(\left(\frac{\partial}{\partial p}logf(x,p)\right)^2\right) = \frac{E[X-mp]^2}{p^2(1-p)^2} = \frac{Var(X)}{p^2(1-p)^2} = \frac{m}{p(1-p)}$$

If follows that for any unbiased estimator, $g(\underline{X})$, for p, we have

$$Var(g(\underline{X})) \geq \frac{1}{\frac{m}{p(1-p)}} = \frac{p(1-p)}{m}$$

Now consider the estimator $g_1(\underline{X}) = \frac{X}{m}$.

$$E[g_1(\underline{X})] = \frac{E[X]}{m} = \frac{mp}{m} = p$$

so $g_1(\underline{X})$ is an unbiased estimator of p. To address the issue of whether or not it's the most efficient unbiased estimator for p we compute the variance of g_1 and compare it to the Cramér-Rao lower bound we calculated above.

$$Var(g_1(\underline{X})) = Var\left(\frac{X}{m}\right) = \frac{Var(X)}{m^2} = \frac{mp(1-p)}{m^2} = \frac{p(1-p)}{m}$$

Since $Var(g_1)$ equals the Cramér-Rao lower bound, we can conclude that $g_1(\underline{X})$ is the most efficient unbiased estimator for p.

While it is useful to know how to verify the optimality of estimators it is also useful to understand when the Cramér-Rao lower bound can be obtained.

Example 2.2. First we recall the definition of the score function above, and that when V is a linear function of T we have that $\rho_{V,T} = \pm 1$ we have that $V = A(T-\theta)$ where A is independent of the observations but may be a function of θ, so we will write it as $A(\theta)$. So the condition for the MVB to be attained is that the statistic $T = t(X_1, \ldots, X_n)$ satisfies

$$\frac{\partial logL(\theta)}{\partial\theta} = A(\theta)(T - E[T])$$

Now consider the problem of estimating the variance θ of a normal distribution with known mean μ, based on a sample size n.

The likelihood is

$$L(\theta) = (2\pi)^{-n/2}\theta^{-n/2}e^{-\Sigma(x_i-\mu)^2/2\theta}$$
$$logL(\theta) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log\theta - \Sigma_{i=1}^{n}(x_i-\mu)^2/2\theta$$
$$\frac{\partial logL(\theta)}{\partial\theta} = -\frac{n}{2\theta} + \frac{\Sigma(x_i-\mu)^2}{2\theta^2} = \frac{n}{2\theta^2}\left(\frac{\Sigma(x_i-\mu)^2}{n} - \theta\right)$$

where $T = t(X_1, \ldots, X_n) = \Sigma_{i=1}^{n}(x_i-\mu)^2/n$. So, using this as the estimate of θ, the MVB is achieved and it is $2\theta^2/n$.

Note that $E[\Sigma(X_i-\mu)^2] = \Sigma E[X_i-\mu)^2 = nVar(X_i) = n\theta$ so T is an unbiased estimator of θ. Also, $\Sigma(X_i-\mu)^2/\theta \sim \chi_n^2$ so has variance $2n$. Hence,

$$Var(T) = Var\left(\frac{\theta}{n}\frac{\Sigma(X_i-\mu)^2}{\theta}\right) = \frac{\theta^2}{n^2}2n = \frac{2\theta^2}{n}$$

Example 2.3. Consider the problem where we have a random sample $X_1, \ldots, X_n$ from a Poisson distribution with parameter θ and we wish to find the CRLB for the variance of an unbiased estimator of θ, and identify the estimator that has this variance.

Now for $f(x,\theta) = e^{-\theta}\theta^x/x!$, the likelihood of the sample is

$$L(\theta; x_1, \ldots, x_n) = \frac{e^{-n\theta}\theta\Sigma x_i}{\prod_{i=1}^{n}(x_i)!}$$
$$logL(\theta; x_1, \ldots, x_n) = -n\theta + \Sigma x_i log\theta - logK$$
$$\frac{\partial logL(\theta)}{\partial\theta} = -n + \frac{\Sigma x_i}{\theta} = \frac{n\theta + n\bar{x}}{\theta} = \frac{n}{\theta}[\bar{x} - \theta] = A(\theta)[T - \theta]$$

where $T(X_i) = \bar{X}$ is the statistic. This is in the correct form for the MVB to be attained (see Section 4, item 4) and it is $1/A(\theta) = \theta/n$. We note that $\bar{X}$ is an estimator which has variance θ/n.

3. Appendix A

- Underwater Navigation From Experimental Data The fundamental limitations in navigation uncertainty can be described in terms of the Cramér-Rao lower bound, which is interpreted in terms of the inertial navigation system (INS) error, the sensor accuracy and the terrain map excitation. Hence, the Cramér-Rao lower bound can be interpreted and used in design for INS systems, sensor performance or if these are given, how much terrain or depth excitation that is needed for use in positioning and navigation. `http://www.control.isy.liu.se/~fredrik/reports/03icasspkarlsson.pdf`
- An estimator for the non-Gaussianity of CMB anisotropies We find that our estimator is optimal, where optimality is defined by saturation of the Cramer Rao bound, if noise is homogeneous. `http://arxiv.org/abs/astro-ph/0701921`
- Error in parameter-estimation performance of gravitational wave measurements The Fisher-matrix formalism is used routinely in the literature on gravitational-wave detection to characterize the parameter-estimation performance of gravitational-wave measurements, given parametrized models of the waveforms, and assuming detector noise of known colored Gaussian distribution. `http://arxiv.org/abs/gr-qc/0703086`
- Fisher information gives smoother probability density functions than PDFs obtained using maximum entropy Reconstructing probability densities inherent in securities pricing. Section 3 of this paper appears to be very useful: In this section we use the Fisher information approach to motivate practi- cal computational approaches to the extraction of probability densities and related quantities from nancial observables. In each case we contrast the results obtained by this method with those of maximum entropy. `http://arxiv.org/abs/cond-mat/0302579`

4. References

(1) **Introduction to Statistical Theory**
Hoel, Port, Stone
Haughton Mifflin 1971

(2) **Statistical Inference, 2nd ed.**
Casella, Berger
Duxbury 2002

(3) **Mathematical Statistics with Applications, 7th ed.**
Wackerly, Mendenhall, Scheaffer
Duxbury 2008

(4) **Statistical Theory**
Bernard William Lindgren
CRC Press

(5) `http://en.wikipedia.org/wiki/Score_%28statistics%29`

(6) `http://en.wikipedia.org/wiki/Cramer-Rao`

(7) `http://en.wikipedia.org/wiki/Fisher_information`

(8) `http://mcs.une.edu.au/~stat354/notes`