Foundations of Statistical Inference

J. Berestycki & D. Sejdinovic

Department of Statistics
University of Oxford

MT 2019

Lecture 3 - Estimators, Minimum Variance Unbiased Estimators and the Cramér-Rao Lower Bound.

Estimators

Definition

A point estimate for θ is a statistic of the data.

$$\widehat{\theta} = \widehat{\theta}(x) = t(x_1, \ldots, x_n).$$

An interval estimate is a set valued function $C(X) \subseteq \Theta$ such that $\theta \in C(X)$ with a specified probability.

Definition (Maximum likelihood estimation)

If $L(\theta)$ is differentiable and there is a unique maximum in the interior of $\theta \in \Theta$, then the MLE $\widehat{\theta}$ is the solution of

$$\frac{\partial}{\partial \theta} L(\theta; x) = 0 \quad \text{or} \quad \frac{\partial}{\partial \theta} \ell(\theta) = 0,$$

where $\ell(\theta) = \log L(\theta; x)$.

Estimators

Let $\hat{\theta}$ be the MLE (or a mle) and $\eta = g(\theta)$ where $g : \Theta \to \mathcal{S}$ is a C^1 diffeomorphism (invertible, continuously differentiable and its inverse is continuously differentiable). The likelihood in the parametrization $\eta = g(\theta)$ is written as

$$\tilde{L}(\eta) = L(g^{-1}(\eta))$$

and any MLE $\widehat{\eta}$ can be written as

$$\widehat{\eta} = g(\widehat{\theta}), \quad \widehat{\theta} = MLE$$

We say that the MLE is invariant by reparametrization. exercise

Lemma 2 : MLEs and exponential families

Consider a k-dimensional exponential family in canonical form

$$L(\theta; x) = \exp\left\{\sum_{j=1}^{k} \phi_j \left(\sum_{i=1}^{n} B_j(x_i)\right) - nD(\boldsymbol{\phi}) + \sum_{i=1}^{n} C(x_i)\right\}.$$

Let $T_j(X) = \sum_{i=1}^{n} B_j(X_i)$, $j = 1, \ldots, k$. If the realized data are $X = x$, then the statistics evaluated on the data are $T_j(x) = t_j$.

Theorem

The MLEs of $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)$ are the solutions of

$$t_j = n\frac{\partial D(\boldsymbol{\phi})}{\partial \phi_j} (= \mathbb{E}_X(B_j; \boldsymbol{\phi}))\ j = 1, \ldots, k.$$

when they exist.
If $\{\boldsymbol{\phi}; \int e^{\boldsymbol{\phi}^t B(x)} e^{C(x)} dx < +\infty\}$ is open and there is at least one solution then it is the unique MLE

Proof

$$\ell = \log L = \text{const} + \sum_{j=1}^{k} \phi_j t_j - nD(\boldsymbol{\phi})$$

$$\Rightarrow \qquad \frac{\partial}{\partial \phi_j}\ell = t_j - n\frac{\partial}{\partial \phi_j}D(\boldsymbol{\phi})$$

However, since $\mathbb{E}_X[B_i(X)] = \frac{\partial}{\partial \phi_i}D(\boldsymbol{\phi})$ and $T_j(X) = \sum_{i=1}^{n} B_j(X_i)$ we know that

$$\mathbb{E}_X[T_j] = n\frac{\partial}{\partial \phi_j}D(\boldsymbol{\phi}), \text{ so}$$

$$\frac{\partial}{\partial \phi_j}\ell = t_j - \mathbb{E}_X(T_j; \boldsymbol{\phi}) = 0$$

is equivalent to $t_j = \mathbb{E}_X(T_j; \boldsymbol{\phi})$.

Bias, Variance, Mean Squared Error

Definition

- The quadratic risk or Mean Squared Error (MSE) of T_n is

$$\text{MSE}(T_n;\theta) = \mathbb{E}_X\left[(T_n - g(\theta))^2;\theta\right] = \int_{\mathcal{X}} (T_n(x) - g(\theta))^2 f(x;\theta)dx$$

  (also Quadratic loss function)

- bias of T_n :

$$\text{bias}(T_n;\theta) = \mathbb{E}_X\left[T_n;\theta\right] - g(\theta)$$

Bias - Variance decomposition and unbiased estimator

$$\text{MSE}(T_n;\theta) = \underbrace{V_X(T_n;\theta)}_{\text{variance}} + \underbrace{\left(\mathbb{E}_X\left[T_n;\theta\right] - g(\theta)\right)^2}_{\text{bias}^2}$$

A statistic $T_n = T(X_1, \ldots, X_n)$ is unbiased for a function $g(\theta)$ iff

$$\text{bias}(T_n;\theta) = 0, \quad \text{for all} \quad \theta \in \Theta$$

$$\text{iff} \quad \mathbb{E}_X(T_n;\theta) = \int_{\mathcal{X}} T_n(x) f(x;\theta) dx = g(\theta), \quad \text{for all } \theta \in \Theta.$$

Example 10 $N(\mu, \sigma^2)$. $\widehat{\mu} = \bar{X}$ and $S^2 = (n-1)^{-1} \sum_{i=1}^n \left(X_i - \bar{X}\right)^2$ are unbiased estimates of μ and σ^2.

Consistency

Ultimately, if infinite number of observations (or infinite information in the data), we expect *to know* θ : This is the notion of consistency
There are multiple modes of consistency

Definition

Consistencies
T_n is a consistent estimator of $g(\theta)$ in probability iff

$$\forall \epsilon > 0, P(|T_n - g(\theta)| > \epsilon) \to 0 \text{ as } n \to \infty.$$

T_n is a consistent estimator in MSE iff

$$\text{MSE}(T_n; \theta) \to 0 \text{ as } n \to \infty.$$

If T_n is consistent in MSE for $g(\theta)$ then it is consistent in Probability

Finite distant analysis : optimality?

- Aim of the game : minimize the risk MSE at a given n

$MSE(T_n;\theta) = E_X([T_n - g(\theta)]^2;\theta)$: function of θ

Definition

$T_{n,1}$ is a better estimate than $T_{n,2}$ in quadratic mean if

$$\forall \theta \in \Theta; \quad MSE(T_{n,1};\theta) \leq MSE(T_{n,2};\theta)$$

Problem : It is not possible to find an estimator better than any other estimators . For example, if we choose the estimator $\widehat{\theta} = \theta_0$ then this has MSE=0 when $\theta = \theta_0$, so no other estimator can be uniformly best unless it has zero MSE everywhere.

Uniformly Minimum Variance Unbiased Estimators (UMVUE)

Since we cannot find an optimal estimator : we reduce the class of possible estimators

Definition

UMVUE T_n is the minimum variance unbiased estimator (MVUE) for $g(\theta)$ iff

- T_n is unbiased : $\text{bias}(T_n, \theta) = 0$ for all $\theta \in \Theta$
- For all T_n' unbiased, $V_X(T_n; \theta) \leq V_X(T_n'; \theta)$ for all $\theta \in \Theta$

Theorem : Cramér-Rao inequality (and bound).

Theorem

If T_n is an unbiased estimator of $g(\theta)$ with $\theta \in \Theta \subset \mathbb{R}$ then subject to certain regularity conditions on $f(x;\theta)$, we have

$$Var(T_n) \geq \frac{g'(\theta)}{I_\theta}.$$

where I_θ, the Fisher information, is given by

$$I_\theta = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \ell(\theta) \right]$$

Comment If an estimator achieves the bound then it is UMVUE. There is no guarantee that the bound will be attainable. In many cases it is attainable asymptotically. Intuitively, the more 'information' we have about θ, the larger I_θ will be and lowest possible variance of the estimator will be smaller.

If $g(\theta) = \theta$

Theorem

If $\widehat{\theta}$ is an unbiased estimator of θ, then subject to certain regularity conditions on $f(x;\theta)$, we have

$$Var(\widehat{\theta}) \geq I_\theta^{-1}.$$

Regularity conditions for CRLB

- First order derivative : For all h integrable wrt $f(x;\theta), \theta \in \Theta$

$$\frac{\partial \int h(x) f(x;\theta) dx}{\partial \theta} = \int h(x) \frac{\partial f(x;\theta)}{\partial \theta} dx$$

- 2nd order derivative : For all h integrable wrt $f(x;\theta), \theta \in \Theta$

$$\frac{\partial^2 \int h(x) f(x;\theta) dx}{\partial \theta^2} = \int h(x) \frac{\partial^2 f(x;\theta)}{\partial \theta^2} dx$$

- Θ is an open set

In order to prove the CRLB we will need to use a few results.

Lemma (Variance-Covariance inequality)

Let U and V be scalar rv. Then

$$cov(U, V)^2 \leq var(U)var(V)$$

with equality if and only if $U = aV + b$ for constants and $a \neq 0$.

Simplified form of Fisher Information

The Fisher Information I_θ, which is used in the Cramér-Rao lower bound, can be expressed in two different forms.

Lemma

Under regularity conditions

$$I_\theta = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right] = \mathbb{E}_\theta\left[\left(\frac{\partial\ell}{\partial\theta}\right)^2\right] = Var[S(X;\theta)],$$

where the score function $s(x;\theta)$ is defined as

$$S(x;\theta) = \frac{\partial}{\partial\theta}\ell(\theta) = \frac{f'(x;\theta)}{f(x;\theta)}$$

Simplified form of Fisher Information

The Fisher Information I_θ, which is used in the Cramér-Rao lower bound, can be expressed in two different forms.

Lemma

Under regularity conditions

$$I_\theta = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right] = \mathbb{E}_\theta\left[\left(\frac{\partial\ell}{\partial\theta}\right)^2\right] = \mathit{Var}[S(X;\theta)],$$

where the score function $s(x;\theta)$ is defined as

$$S(x;\theta) = \frac{\partial}{\partial\theta}\ell(\theta) = \frac{f'(x;\theta)}{f(x;\theta)}$$

Lemma 3 - Proof

We need to prove $-\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right] = \mathbb{E}\left[\left(\frac{\partial\ell}{\partial\theta}\right)^2\right]$.

$$\begin{aligned}
\frac{\partial^2\ell}{\partial\theta^2} &= \frac{\partial}{\partial\theta}\left\{\frac{1}{L}\frac{\partial L}{\partial\theta}\right\} \qquad \left[\text{since } \frac{\partial\ell}{\partial\theta} = \frac{1}{L}\frac{\partial L}{\partial\theta}\right] \\
&= -\frac{1}{L^2}\left(\frac{\partial L}{\partial\theta}\right)^2 + \frac{1}{L}\frac{\partial^2 L}{\partial\theta^2} \\
&= -\left(\frac{\partial\ell}{\partial\theta}\right)^2 + \frac{1}{L}\left(\frac{\partial^2 L}{\partial\theta^2}\right)
\end{aligned}$$

The second term has expectation zero because

$$\mathbb{E}\left[\frac{1}{L}\left(\frac{\partial^2 L}{\partial\theta^2}\right)\right] = \int \frac{1}{L}\frac{\partial^2 L}{\partial\theta^2} L dx = \int \frac{\partial^2 L}{\partial\theta^2} dx = \frac{\partial^2}{\partial\theta^2}\int L dx = 0$$

The alternative form $I_\theta = \mathsf{Var}[S(X;\theta)]$ follows from $\mathbb{E}\left[\frac{\partial\ell}{\partial\theta}\right] = 0$.

Proof of the CRLB : case $g(\theta) = \theta$

We consider only unbiased estimators, so we have

$$\mathbb{E}(\widehat{\theta}) = \int_{\chi} \widehat{\theta}(x) L(\theta; x) dx = \theta$$

Differentiate both sides w.r.t. θ

$$\int_{\chi} \widehat{\theta} \frac{\partial L}{\partial \theta} dx = 1$$

Now

$$\frac{\partial L}{\partial \theta} = L \frac{\partial \ell}{\partial \theta}$$

so

$$1 = \int_{\chi} \widehat{\theta} \frac{\partial \ell}{\partial \theta} L dx = \mathbb{E}\left[\widehat{\theta} \frac{\partial \ell}{\partial \theta}\right]$$

Proof of the CRLB : case $g(\theta) = \theta$

We consider only unbiased estimators, so we have

$$\mathbb{E}(\widehat{\theta}) = \int_{\chi} \widehat{\theta}(x) L(\theta; x) dx = \theta$$

Differentiate both sides w.r.t. θ

$$\int_{\chi} \widehat{\theta} \frac{\partial L}{\partial \theta} dx = 1$$

Now

$$\frac{\partial L}{\partial \theta} = L \frac{\partial \ell}{\partial \theta}$$

so

$$1 = \int_{\chi} \widehat{\theta} \frac{\partial \ell}{\partial \theta} L dx = \mathbb{E}\left[\widehat{\theta} \frac{\partial \ell}{\partial \theta}\right]$$

Proof of the CRLB : case $g(\theta) = \theta$

We consider only unbiased estimators, so we have

$$\mathbb{E}(\widehat{\theta}) = \int_{\chi} \widehat{\theta}(x) L(\theta; x) dx = \theta$$

Differentiate both sides w.r.t. θ

$$\int_{\chi} \widehat{\theta} \frac{\partial L}{\partial \theta} dx = 1$$

Now

$$\frac{\partial L}{\partial \theta} = L \frac{\partial \ell}{\partial \theta}$$

so

$$1 = \int_{\chi} \widehat{\theta} \frac{\partial \ell}{\partial \theta} L dx = \mathbb{E}\left[\widehat{\theta} \frac{\partial \ell}{\partial \theta}\right]$$

Proof of the CRLB

Now we use the inequality that for two random variables U, V

$$\mathsf{Cov}[U, V]^2 \leq \mathsf{Var}[U]\mathsf{Var}[V]$$

with $U = \widehat{\theta}$, $V = \frac{\partial \ell}{\partial \theta}$. We know $\mathsf{Var}[\frac{\partial \ell}{\partial \theta}] = I_\theta$. Must show $\mathsf{Cov}[U, V] = 1$.

$$\mathsf{Cov}[U, V] = \mathbb{E}[UV] - \mathbb{E}[U]\mathbb{E}[V], \qquad \mathbb{E}[U] = \theta, \qquad \mathbb{E}\left[\widehat{\theta}\frac{\partial \ell}{\partial \theta}\right] = 1$$

$$\mathbb{E}[V] = \int_\chi \frac{\partial \ell}{\partial \theta} L dx = \int_\chi \frac{\partial L}{\partial \theta} dx = \frac{\partial}{\partial \theta}\left[\int_\chi L dx\right] = \frac{\partial}{\partial \theta}[1] = 0$$

$$\mathsf{Var}[\widehat{\theta}] = \mathsf{Var}[U] \geq \frac{\mathsf{Cov}[U, V]^2}{\mathsf{Var}[V]} = \frac{1^2}{I_\theta} = I_\theta^{-1}$$

Exercise: generalize to other $g(\theta)$

Information in a sample of size n.

If we have n iid observations then

$$f(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

and the Fisher information is

$$I_n(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right] = -\int \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f(x_i;\theta) f(x;\theta) dx = n i_1(\theta).$$

That is, $i_1(\theta)$ is calculated from the density as

$$i_1(\theta) = -\int \frac{\partial^2}{\partial\theta^2} \log f(x;\theta) f(x;\theta) dx$$

Question Under what conditions will we be able to attain the Cramér-Rao bound and find a MVUE?

Corollary (1)

There exists an unbiased estimator $\widehat{\theta}$ which attains the CR lower bound (under regularity conditions) if and only if

$$S(x,\theta) = \frac{\partial \ell}{\partial \theta} = I_\theta(\widehat{\theta} - \theta)$$

Proof In the CR proof

$$\mathrm{Cov}[U,V]^2 \leq \mathrm{Var}[U]\mathrm{Var}[V]$$

and the lower bound is attained if and only equality is achieved. If $U = \widehat{\theta}, V = \frac{\partial \ell}{\partial \theta}$, the equality occurs when $\frac{\partial \ell}{\partial \theta} = c + d\widehat{\theta}$, where c, d are constants.
$\mathbb{E}[V] = 0$ so $c = -d\theta$ and $\frac{\partial \ell}{\partial \theta} = d(\widehat{\theta} - \theta)$.

Multiply by $\partial\ell/\partial\theta$ and take expectations.

$$\mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] \quad = \quad d\mathbb{E}\left[\frac{\partial \ell}{\partial \theta}\widehat{\theta}\right] - d\theta\mathbb{E}\left[\frac{\partial \ell}{\partial \theta}\right] = d \times 1 - 0$$

The LHS is I_θ so we have $d = I_\theta$ and

$$\frac{\partial \ell}{\partial \theta} = I_\theta(\widehat{\theta} - \theta)$$

Question What is the relationship between the CRLB and exponential families?

Corollary (2)

If there exists an unbiased estimator $\widehat{\theta}(X)$ which attains the CR lower bound (under regularity conditions) it follows that X must be in an exponential family

Proof Taking $n = 1$

$$\frac{\partial \log f(x;\theta)}{\partial \theta} = \frac{\partial \ell}{\partial \theta} = I_\theta(\widehat{\theta} - \theta)$$

and

$$\log f(x;\theta) = \widehat{\theta} A(\theta) - D(\theta) + C(x)$$

which is an exponential family form. The constant of integration $C(x)$ is a function of x.

Question What is the relationship between the CRLB and MLEs?

Corollary (3)

Suppose $\widetilde{\theta}(X)$ is an unbiased estimator that attains the CRLB, and so is a MVUE. Suppose that the MLE $\widehat{\theta}$ is a solution to $\partial\ell/\partial\theta = 0$ (so, not on boundary). Then $\widetilde{\theta} = \widehat{\theta}$.

i.e. if the CRLB is attained then it is generally the MLE that attains it.

Proof $\widetilde{\theta}$ must satisfy $\frac{\partial \ell}{\partial \theta} = I_\theta(\widetilde{\theta} - \theta)$.
Setting $\frac{\partial \ell}{\partial \theta} = 0$ and solving will give the MLE $\widehat{\theta}$.
Since $I_\theta > 0$ (in all but exceptional circumstances), this gives $\widetilde{\theta} = \widehat{\theta}$.

Question Do all MLEs attain the CRLB?
No, because not all MLEs are unbiased.

Example 11

Let $X_1, \ldots, X_n$ be a random sample from $\mathsf{N}(\mu, \sigma^2)$.

Then we know the MLEs are $\widehat{\mu} = \bar{X}$, $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$.

Exercise $\widehat{\mu}$ is unbiased, but $\widehat{\sigma^2}$ is biased. CRLBs are $1/I_\mu = \sigma^2$ and $1/I_{\sigma^2} = 2\sigma^4/n$.

$\mathsf{Var}(\widehat{\mu}) = \sigma^2/n$ which equals the CRLB so is MVUE.

$\mathsf{Var}(\widehat{\sigma^2}) = 2(n-1)\sigma^4/n^2$ is less than the CRLB. But $\widehat{\sigma^2}$ is biased.

The sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is unbiased and has variance $2\sigma^4/(n-1)$ which is larger than the CRLB.
Question Is S^2 a MVUE?

Efficiency

Definition

The (Bahadur) efficiency of an estimator $\widetilde{\theta}$ is defined as a comparison of the variance of $\widetilde{\theta}$ with the CR bound I_θ^{-1}. That is

$$\text{Efficiency of } \widetilde{\theta} = \frac{I_\theta^{-1}}{\text{Var}[\widetilde{\theta}]} = \frac{1}{I_\theta \text{Var}[\widetilde{\theta}]}$$

The asymptotic efficiency is the limit as $n \to \infty$.

There are similar definitions for the relative efficiency of two estimators.

Asymptotic normality of MLE

Revision from Part A Statistics As the sample size $n \to \infty$, the MLE

$$\widehat{\theta} \approx N(\theta, I_\theta^{-1}).$$

This is a powerful and general result. Assuming the usual regularity conditions hold then it tells us that the MLE has the following properties

1. it is asymptotically unbiased
2. it is asymptotically efficient i.e. it attains the CRLB asymptotically.
3. it has a normal distribution asymptotically.

Extensions to the Cramér-Rao inequality

1. If $\widehat{\theta}$ is an estimator with bias $b(\theta) = \text{bias}(\widehat{\theta})$, then

$$\text{Var}[\widehat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 I_\theta^{-1}$$

2. If $\widehat{g}(x)$ is an unbiased estimator for $g(\theta)$, then

$$\text{Var}[\widehat{g}(X)] \geq \left(\frac{\partial g}{\partial \theta}\right)^2 I_\theta^{-1}.$$

Proof Begin with $\mathbb{E}_\theta(\widehat{\theta}(X)) = \theta + b(\theta)$ (in 1.) and $\mathbb{E}_\theta(\widehat{g}(X)) = g(\theta)$ (in 2.). Differentiate both sides and proceed as above to find $\text{Cov}[U, V] = (1 + \partial b/\partial\theta)$ (in 1.) and $\text{Cov}[U, V] = \partial g/\partial\theta$ (in 2., with $U = \widehat{g}$). The bound is against $\text{Cov}[U, V]^2$ which leads to the results above.

Fisher Information for a d-dimensional parameter

Information matrix:

$$I_{ij} = \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\frac{\partial \ell}{\partial \theta_j}\right] = -\mathbb{E}\left[\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}\right]$$

under regularity conditions. The CR inequality is

$$\mathsf{Var}(\widehat{\theta}_i) \geq [I^{-1}]_{ii},\ i = 1, \ldots, d.$$

Exercise: verify that we have already proved $\mathsf{Var}(\widehat{\theta}_i) \geq [I_{ii}]^{-1}$. Note that $[I^{-1}]_{ii} \geq [I_{ii}]^{-1}$ (GJJ) so bound above is stronger.

Exercise For an Exponential family in canonical form,

$$I_{ij} = -\frac{\partial^2}{\partial \phi_i \partial \phi_j} nD(\phi).$$