Topic IV:
Point Estimation

Wei You

香港科技大學
THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

Fall, 2023

Introduction

The statistical model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ is unknown up to a parameter θ.

- After observing the random samples $\boldsymbol{X} = (X_1, \ldots, X_n)$, we want to estimate θ.
- A point estimator is a statistic $T(\boldsymbol{X})$ that aims to estimate the unknown parameter θ of the given statistical model $\mathcal{P}$.
- An estimator is a random variable.

A generic estimator for the parameter θ is usually denoted as $\widehat{\theta}$.

A Remark on the Notations

- Regular letters, e.g. x, Y, θ, denote scalars.
- Bold face letters, e.g. $\boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{\theta}$, denote vectors.
- Lower-case letters, e.g. $\boldsymbol{x}, y, \beta$, denote deterministic numbers.
- Upper-case letters, e.g. $X, \boldsymbol{Y}$, denote random variables/vectors.

Point Estimators

Example: Some common estimators.

Parameter	Estimator
$\theta = \mathbb{E}[X]$	$\bar{X}$
$\theta = \mathrm{Var}(X)$	S^2
$\theta = \mathrm{Cov}(Y_1, Y_2)$	$\frac{1}{n-1}\sum_{i=1}^{n}(Y_{i,1} - \bar{Y}_1)(Y_{i,2} - \bar{Y}_2)$
$\theta = \rho(Y_1, Y_2)$	$\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{Y_{i,1}-\bar{Y}_1}{S_{Y_1}}\right)\left(\frac{Y_{i,2}-\bar{Y}_2}{S_{Y_2}}\right)$

Error

To evaluate a estimator, it may be tempting to use the error:

Error

Suppose T is used to estimate θ. Then $T - \theta$ is the error of T.

However, there could be problems. For example

- For continuous distributions, $\mathbb{P}_\theta(T(\boldsymbol{X}) = \theta) = 0$.
- For Bernoulli random sample with sample size n, if θ is not one of j/n, then $\bar{X}$ can never equal θ.
- The value of the error is itself a random variable. Thus may take large values even only with a small probability.

Bias

The concept of bias gets rid of the randomness in the error.

Bias

Suppose T is used to estimate θ. Then $\mathbb{E}_\theta[T - \theta]$ is the bias of T.

Biasedness evaluates the "systematic error" of an estimator.

- T is unbiased if $\mathbb{E}_\theta[T] = \theta$ for all $\theta \in \Theta$.
- T is negatively biased if $\mathbb{E}_\theta[T] \leq \theta$ for all $\theta \in \Theta$.
- T is positively biased if $\mathbb{E}_\theta[T] \geq \theta$ for all $\theta \in \Theta$.

A biased estimator does not have to be negatively/positively biased. It may be negatively biased for some θ and positively biased for others.

Bias and Mean Squared Error

It is possible that the error $T(x) - \theta$ can have large positive and negative error, although these errors can cancel each other in the expectation.

In this case, it may be preferable to use

Mean squared error

Suppose T is used to estimate θ. Then $\mathbb{E}_\theta[(T-\theta)^2]$ is the mean squared error of W.

Decomposition of MSE

$$\text{MSE}_\theta T = \text{Var}_\theta(T) + \text{Bias}_\theta^2 T.$$

- The derivation: $\mathbb{E}[(T-\theta)^2] = \text{Var}(T-\theta) + \mathbb{E}[T-\theta]^2 = \text{Var}[T] + Bias^2(T)$.
- For many estimators, it is usually the case that one estimator has lower bias but higher variance. In minimizing MSE, we face the bias-variance trade-off.

Other Evaluations

In addition to the error, the bias, and the MSE, there are others that are frequently used.

- Mean absolute error: $\mathbb{E}_\theta[|T-\theta|]$.
- $\mathbb{P}(|T-\theta|>\varepsilon)$ for a fixed ε.

The most commonly used evaluation is MSE, because a smooth function is mathematically easy to handle. Others may be preferred in some cases

- Mean absolute error usually results in more robust estimation.
- $\mathbb{P}(|T-\theta|>\varepsilon)$ does not implicitly assume that the distribution has finite moments.

Efficiency

For unbiased estimators, the one with less variance has less mean squared error.

Efficiency

If U and V are unbiased estimators of θ. Then

- U is more efficient than V if $\mathrm{Var}(U) \leq \mathrm{Var}(V)$.
- The relative efficiency of U with respect to V is

$$\frac{\mathrm{Var}(V)}{\mathrm{Var}(U)}.$$

If the relative efficiency (of U w.r.t. V) is larger than 1, then U is more efficient than V.

Asymptotic Properties

- The sequence of estimators $T_n(\boldsymbol{X}_n)$ is asymptotically unbiased if $\lim_{n\to\infty} \mathbb{E}[T_n] = \theta$.
- Consistency: T_n converges to θ in probability, i.e., $\mathbb{P}(|T_n - \theta| > \epsilon) \to 0$.
- For asymptotic unbiased estimators U_n and V_n, the asymptotic relative efficiency of U to V is $\lim_{n\to\infty} \frac{\mathrm{Var}(V_n)}{\mathrm{Var}(U_n)}$.

Consistency

If the MSE of T_n converges to 0 as $n \to \infty$, then T_n is consistent.

Proof. For any $\varepsilon > 0$, apply Markov's inequality to $|T_n - \theta|^2$, we have

$$\mathbb{P}(|T_n - \theta| > \varepsilon) = \mathbb{P}(|T_n - \theta|^2 > \varepsilon^2) \leq \frac{\mathbb{E}[(T_n - \theta)^2]}{\varepsilon^2} \to 0.$$

So T_n is consistent.

Common Estimators

- Sample mean $\bar{X}$ is unbiased and consistent for $\theta = \mathbb{E}[X]$ (LLN). Special cases:
  - $X = \mathbb{1}_A$. Then $\bar{X}$ estimates $\mathbb{P}(A)$.
  - $X = \mathbb{1}_{Y \leq y}$. Then $\bar{X}$ estimates $F_Y(y)$.
- Sample variance S^2 is unbiased and consistent for $\sigma^2 = \mathrm{Var}(X)$ (LLN).
- Sample covariance is unbiased and consistent for $\mathrm{Cov}(Y_1, Y_2)$.

A Poisson Example

Example: Poisson(λ). We have the following potential estimators:

- Sample mean: $\mathbb{E}[\bar{X}] = \lambda$.
- Sample variance: $\mathbb{E}[S_n^2] = \lambda$.

Both are unbiased and consistent. Which one is better?

- The MSE of $\bar{X}$

$$\mathrm{Var}(\bar{X}) = \frac{\lambda}{n}.$$

- The MSE of S_n^2

$$\mathrm{Var}(S_n^2) = \frac{1}{n}\left(\mathbb{E}[(X-\lambda)^4] - \mathbb{E}[(X-\lambda)^2]^2\frac{n-3}{n-1}\right) = \frac{\lambda}{n}(1+2\lambda\frac{n}{n-1}).$$

  This one needs tedious calculation so the steps are omitted.

The asymptotic relative efficiency of $\bar{X}$ to S_n^2 is $1+2\lambda > 1$. So $\bar{X}$ is strictly better.

Method of Moments

Method of moments is a classical way to construct estimators.

- Key Idea: set the sample moments equal to population moments, and the later ones are functions of the unknown parameter $\boldsymbol{\theta}$.
- Let $X_1, \ldots, X_n$ be a sample from $f(x|\theta_1, \ldots, \theta_k)$

$$\mathbb{E}[X] = m_1(\theta_1, \ldots, \theta_k) \approx \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\ldots$$

$$\mathbb{E}[X^k] = m_k(\theta_1, \ldots, \theta_k) \approx \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

- In general, we use the first kth moments to construct k equations and solve for the k unknown parameters.

Method of Moments

Example: Normal $N(\mu, \sigma^2)$

$$\mathbb{E}[X] = \mu = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

$$\mathbb{E}[X^2] = \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

So the estimators are

$$\widehat{\mu} = \bar{X}$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Example: Binomial (k, p)

$$\mathbb{E}[X] = kp = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

$$\mathbb{E}[X^2] = kp(1-p) + k^2p^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

So the estimators are

$$\widehat{k} = \frac{\bar{X}^2}{\bar{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \in \mathbb{Z}_+ ?$$

$$\widehat{p} = \frac{\bar{X}}{\widehat{k}}$$

Remarks

- In its general form, we can use k functions of the population averages:

$$\mathbb{E}[g_i(X)] = m_i(\boldsymbol{\theta}), 1 \leq i \leq k,$$

and equate them to the sample averages

$$\frac{1}{n}\sum_{j=1}^{n} g_i(X_j), 1 \leq i \leq k.$$

Example: Lognormal distirbution. $X \sim e^{\mu+\sigma Z}$, where Z is standard normal.

- Simple to construct, usually easy to calculate, does not need too much distribution information.
- It sometimes give estimates outside the domain, like $\widehat{k}$ in the last slide.
- It is quite arbitrary depending on which moments (or g_i functions) to use.
- Generally inferior to the Maximum Likelihood Estimators.

Maximum Likelihood Estimators

One of the most important estimators in statistics is the Maximum Likelihood Estimators (MLE).

- It traces back to Gauss for the estimation of parameters of a Normal distribution. Its general form is due to the British statistician R.A. Fisher.
- The magnificent intuitive of Gauss and Fisher is that: a proper estimator of the true parameter is the one that makes the given observation $\boldsymbol{x}$ most likely to occur.

The likelihood function $L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\theta)$ is a function of the unknown $\theta \in \Theta$, which depends on the observed sample $\boldsymbol{x}$. It measures how likely an experiment produces $\boldsymbol{x}$ as a sample under θ.

It is exactly the joint density of $\boldsymbol{X}$, but is understood as a function of θ.

Maximum likelihood estimator

The MLE $\widehat{\theta}(\boldsymbol{X}) = \arg\max_{\theta\in\Theta} L(\theta|\boldsymbol{X})$ is the parameter $\theta \in \Theta$ that maximizes the likelihood function for a given sample $\boldsymbol{X}$.

We assume that there exist a unique maximizer.

The MLE $\widehat{\theta}$ is a statistic.

- It is understood as $\widehat{\theta}(\boldsymbol{x})$ when $\boldsymbol{X} = \boldsymbol{x}$.
- In calculating the MLE, we ususally first compute

$$\widehat{\theta}(\boldsymbol{x}) = \arg\max_{\theta\in\Theta} L(\theta|\boldsymbol{x})$$

- Then plug-in $\boldsymbol{X}$ to obtain the random variable $\widehat{\theta}(\boldsymbol{X})$.
- Sufficiency and MLE: If $T(\boldsymbol{X})$ is sufficient for θ and if MLE exists, then there exists an MLE that is a function of T. **Proof:** the likelihood function is of the form $\log g(T(\boldsymbol{x})|\theta) + \log h(\boldsymbol{x})$. To maximize over θ, only the first term is involved.

The Calculation of the MLE

$$\widehat{\theta}(\boldsymbol{x}) = \arg\max_{\theta \in \Theta} L(\theta|\boldsymbol{x}).$$

Closed-form solution:

- The optimization can be solved by letting $\frac{\partial}{\partial \theta_i} L(\theta|\boldsymbol{x}) = 0$ if θ is k-dimensional.
- Check the Hessian matrix at $\widehat{\theta}(\boldsymbol{x})$. If the Hessian is negative semi-definite, then it is a local maximum.
- Equivalently, $\widehat{\theta}(\boldsymbol{x}) = \arg\max_{\theta \in \Theta} \log L(\theta|\boldsymbol{x})$ and the optimization can be solve by setting $\frac{\partial}{\partial \theta_i} \log L(\theta|\boldsymbol{x}) = 0$. Maximizing $\log L(\theta|\boldsymbol{x})$ is usually much easier. (Conveniently, most common probability distributions – in particular the exponential family – are logarithmically concave.)

The function $l(\theta|\boldsymbol{x}) \equiv \log L(\theta|\boldsymbol{x})$ is called the log-likelihood function.

The Calculation of the MLE

Numerical solution:

- If a closed-form solution is not available, we usually apply the Newton's method, or gradient descent, to numerically calculate the MLE.
- For such algorithms, the MoM may be used to provide an initial solution.

Maximum Likelihood Estimator – Normal

Example: Normal(μ, σ^2). Likelihood function

$$L(\mu, \sigma|\boldsymbol{x}) = f(\boldsymbol{x}|\mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[\frac{-\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right]$$

Take logarithm

$$\log L(\mu, \sigma|\boldsymbol{x}) = -n\log(\sqrt{2\pi}\sigma) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}$$

Partial derivatives

$$\frac{\partial}{\partial\mu}\log L(\mu, \sigma|\boldsymbol{x}) = -\frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2}, \qquad \frac{\partial}{\partial\sigma}\log L(\mu, \sigma|\boldsymbol{x}) = -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma^3}$$

MLE for Normal Sample

$$\widehat{\mu} = \bar{X} \quad \text{and} \quad \widehat{\sigma} = \left[\frac{\sum_{i=1}^n (X_i - \widehat{\mu})^2}{n}\right]^{\frac{1}{2}} \neq S.$$

Maximum Likelihood Estimator – Bernoulli

Example: Bernoulli(p). Likelihood function

$$L(p|\boldsymbol{x}) = p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}\dots p^{x_n}(1-p)^{1-x_n} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$$

Take logarithm

$$\log L(p|\boldsymbol{x}) = \sum_{i=1}^n x_i \log(p) + \Big(n - \sum_{i=1}^n x_i\Big)\log(1-p)$$

Differentiate

$$\frac{d}{dp}\log L(p|\boldsymbol{x}) = \frac{1}{p}\sum_{i=1}^n x_i - \frac{1}{1-p}\Big(n - \sum_{i=1}^n x_i\Big)$$

MLE for Bernoulli Sample

$$\widehat{p} = \bar{X}$$

Maximum Likelihood Estimator – Poisson

Example: Poisson(λ). Likelihood

$$L(\lambda|\boldsymbol{x}) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\frac{e^{-\lambda}\lambda^{x_2}}{x_2!}\cdots\frac{e^{-\lambda}\lambda^{x_n}}{x_n!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{x_1!\ldots x_n!}$$

Take the log

$$\log L(\lambda|\boldsymbol{x}) = -n\lambda + \sum_{i=1}^n x_i \log(\lambda) - \log(x_1!\ldots x_n!)$$

Differentiate

$$\frac{d}{d\lambda}\log L(\lambda|\boldsymbol{x}) = -n + \frac{1}{\lambda}\sum_{i=1}^n x_i$$

MLE for Poisson Sample

$$\widehat{\lambda} = \bar{X}$$

Maximum Likelihood Estimator – Uniform

Example: Uniform$(0, \theta)$. Likelihood

$$l(\theta|\boldsymbol{x}) = \frac{1}{\theta^n}\mathbb{1}(0 \leq X_{(1)} \leq X_{(n)} \leq \theta)$$

MLE for Uniform Sample

$$\widehat{\theta} = \max(X_1, \ldots, X_n) = X_{(n)}$$

MoM uses $\widehat{\theta} = 2\bar{X}$, note that it is possible that some $x_i > 2\bar{X}$.

The two estimators are very different.

Invariance Property of MLEs

Suppose we re-parametrize the distribution. Let $\lambda = h(\theta)$ be a new parameter in the space $\Lambda = h(\Theta)$.

Example: For exponential distribution, we can use the rate λ or the mean $\mu = 1/\lambda$.

Theorem (Invariance of MLE)

Suppose $h(\cdot)$ is one-to-one. If $\widehat{\theta}$ is the MLE of θ, then for any function $h(\cdot)$, the MLE of $h(\theta)$ is $h(\widehat{\theta})$.

Example: for exponential distribution, we can show that the MLE for λ is $\widehat{\lambda} = 1/\bar{X}$; thus the MLE for μ is $\widehat{\mu} = \bar{X}$.

What if $h(\cdot)$ is not one-to-one? For example, $h(x) = x^2$.

- Given a η, there may be multiple θ such that $h(\theta) = \eta$. In this caes, η does not uniquely determine the distribution and the likelihood is not well defined.

When if $h(\cdot)$ is not one-to-one:

- To avoid such difficulty, we define the *induced likelihood function*

$$L^*(\eta|\boldsymbol{x}) = \sup_{\{\theta:h(\theta)=\eta\}} L(\theta|\boldsymbol{x}).$$

The $\widehat{\eta}$ that maximizes $L^*(\eta|\boldsymbol{x})$ will be called the MLE of η.

Theorem (Invariance of MLE)

If $\widehat{\theta}$ is the MLE of θ, then for any function $h(\cdot)$, the MLE of $h(\theta)$ is $h(\widehat{\theta})$.

$$L^*(\widehat{\eta}|\boldsymbol{x}) = \sup_{\eta} \sup_{\{\theta:h(\theta)=\eta\}} L(\theta|\boldsymbol{x}) = \sup_{\theta} L(\theta|\boldsymbol{x}) = L(\widehat{\theta}|\boldsymbol{x}) = \sup_{\{\theta:h(\theta)=h(\widehat{\theta})\}} L(\theta|\boldsymbol{x}) = L^*(h(\widehat{\theta})|\boldsymbol{x}).$$

- Implication: If $\bar{X}$ is the MLE for θ, then $\bar{X}^2$ is the MLE for θ^2. If $\widehat{\sigma}$ is the MLE for the standard deviation, then $\widehat{\sigma}^2$ is the MLE for the variance.

Bias-Variance Trade-Off – Normal

- Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$. The sample mean and sample variance $\bar{X}$ and S^2 are unbiased for μ and σ^2. For MLE, $\mu = \bar{X}$ and $\widehat{\sigma}^2 = \frac{n-1}{n} S^2$.
- MSE of S^2: since $\mathrm{Var}(\chi^2_{n-1}) = 2(n-1)$, we have

$$\mathbb{E}[(S^2 - \sigma^2)^2] = \mathrm{Var}(S^2) = \mathrm{Var}(\sigma^2 \chi^2_{n-1})/(n-1)^2 = \frac{2\sigma^4}{n-1}.$$

- MSE of $\widehat{\sigma}^2$: Using the variance-bias decomposition of MSE

$$\mathbb{E}[(\widehat{\sigma}^2 - \sigma^2)^2] = \frac{(n-1)^2}{n^2}\mathrm{Var}(S^2) + \left(\frac{\sigma^2}{n}\right)^2 = \frac{2n-1}{n^2}\sigma^4 < \frac{2\sigma^4}{n-1}.$$

- By trading off variance for bias, the MSE is improved.
- Both S^2 and $\widehat{\sigma}^2$ are asymptotically unbiased and consistent, and the relative efficiency is 1.

Bias-Variance Trade-Off – Uniform

Suppose $X_1, \ldots, X_n$ are sampled from Uniform$(0, \theta)$.

- Estimator 1 (MoM)

$$d_1(\mathbf{X}) = 2\frac{\sum_{i=1}^n X_i}{n}$$

- Estimator 2 (MLE)

$$d_2(\mathbf{X}) = \max_i X_i$$

Compare:

$$\begin{aligned}
\text{Bias}_\theta d_1(\mathbf{X}) &= 0 \\
\text{MSE}_\theta d_1(\mathbf{X}) &= \text{Var}(d_1(\mathbf{X}, \theta)) + 0^2 \\
&= \frac{4}{n^2}\left(\text{Var}(X_1) + \ldots + \text{Var}(X_n)\right) = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}
\end{aligned}$$

Bias-Variance Trade-Off – Uniform

$d_2(\mathbf{X})$ is an order Statistic.

$$f_{d_2}(x) = \frac{n}{\theta^n} x^{n-1} \mathbb{1}_{0 \le x \le \theta}.$$

So

$$\mathbb{E}[d_2(\mathbf{X})] = \frac{n}{n+1}\theta \qquad \mathrm{Var}(d_2(\mathbf{X})) = \frac{n\theta^2}{(n+2)(n+1)^2}$$

Hence

$$\begin{aligned} \mathsf{Bias}_\theta d_2(\mathbf{X}) &= -\frac{1}{n+1}\theta \\ \mathsf{MSE}_\theta d_2(\mathbf{X}) &= \frac{n\theta^2}{(n+2)(n+1)^2} + \frac{1}{(n+1)^2}\theta^2 \\ &= \frac{2\theta^2}{(n+1)(n+2)} < \frac{\theta^2}{3n} = \mathsf{MSE}_\theta d_1(\mathbf{X}) \end{aligned}$$

Comparing of d_1 and d_2

By CLT

$$2\bar{X} \approx \frac{1}{\sqrt{n}} N(0, \theta^2/3)$$

From Topic 2, we have

$$X_{(n)} \approx \theta - \theta \frac{\mathrm{Exp}(1)}{n}.$$

- d_1 is unbiased; has MSE $\frac{\theta^2}{3n}$; is consistent (WLLN).
- d_2 is negatively biased; has MSE $\frac{2\theta^2}{(n+1)(n+2)}$; is consistent and asymptotically unbiased.
- Asymptotically d_2 is more efficient than d_1: n^{-2} versus n^{-1}.
- Those MSE decreases on the order of n^{-2} is called super efficient. Recall that the MSE of sample mean (which is a decent estimator) only decreases at n^{-1}.
- Unbiased estimator can be very bad: let $d_3(\boldsymbol{X}) = (n+1)X_{(1)}$. Then $\mathbb{E}[d_3] = \theta$, but $\mathrm{Var}(d_3) = \frac{n}{n+2}\theta^2$. It is not even consistent.

Best Unbiased Estimators

If an estimator is unbiased, it delivers correct estimation on average.

It would then be nice if the estimator has low variance.

- When unbiased, this is equivalent to minimizing the mean squared error (MSE).
- In general, the variance (MSE) depends on θ.
- If $\mathrm{Var}_\theta(W_1) \leq \mathrm{Var}_\theta(W_2)$ for all $\theta \in \Theta$, then W_1 is uniformly better than W_2.

UMVUE

If an unbiased estimator W^* is uniformly better than any other unbiased estimators, then W^* is called an uniform minimum variance unbiased estimator (UMVUE) of θ.

UMVUE does not always exist.

Example: Poisson. If $\boldsymbol{x}$ is a sample from Poisson(λ), then

$$\mathbb{E}_\lambda \bar{X} = \lambda, , \qquad \mathbb{E}_\lambda S^2 = \lambda.$$

We have shown that $\mathrm{Var}_\lambda(\bar{X}) \leq \mathrm{Var}_\lambda(S^2)$. So S^2 can not be UMVUE.

- Finding an UMVUE is difficult: there can be infinitely many unbiased estimators. For example,

  $$W_a(\bar{X}) = a\bar{X} + (1-a)S^2$$

  is unbiased for all $a \in [0, 1]$.

Before we discuss possible ways to find a UMVUE, let's first consider the following question

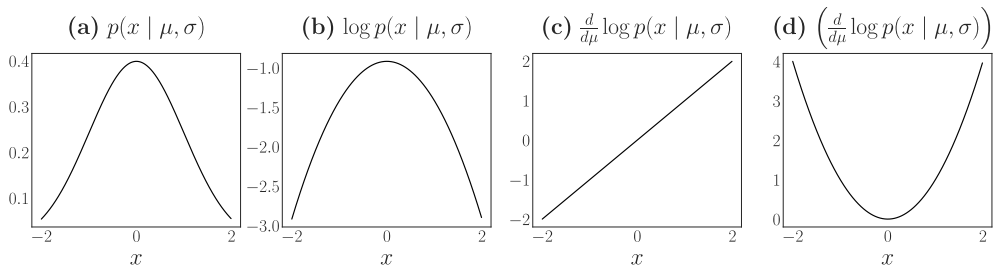
Are we fundamentally limited by the finite amount of data available? I.e, is it possible to achieve arbitrarily small variance with random samples of finite size?

Fisher Information – Intuition

In point estimation, we obtain information about θ from a random sample.

- **Question**: how much information can a random sample provide?
- **Key point**: If the amount of information in the random sample is bounded, then the efficiency of our estimators would be upper bounded.

Example: Consider a random variable X from $N(\mu, \sigma^2)$. We plot the log-likelihood $\log(f(X|\mu))$ as functions of μ.



Observation

$$\begin{aligned}\text{If } \theta \text{ were the true value} &\Rightarrow \text{large } L(\theta|X) \equiv f(X|\theta)\\ &\Rightarrow \theta \text{ near the peak of } l(\theta|X) \equiv \log(f(X|\theta))\\ &\Rightarrow \text{the derivative of } l(\theta|X) \text{ close to zero.}\end{aligned}$$

Score

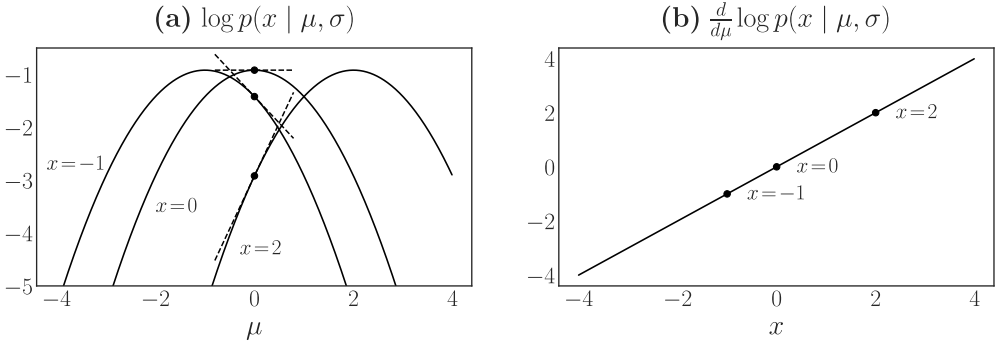
The gradient of the log-likelihood function is called the score function

$$s(\boldsymbol{\theta}|X) = \frac{\partial \log L(\boldsymbol{\theta}|X)}{\partial \boldsymbol{\theta}} = \frac{\partial l(\boldsymbol{\theta}|X)}{\partial \boldsymbol{\theta}}.$$

- The score measures the steepness of the log-likelihood.

If $|s(\mu|X)|$ is large, the random variable X provides more information about μ.

- To understand, $s(\mu|X)$ is the slope of the log-likelihood function, measuring the how sensitive the log-likelihood function is to the change of μ. (High sensitivity $\Rightarrow$ information is more precise.)



- To measure the information, we use

$$\mathrm{Var}_{\mu^*}(s(\mu^*|X)) = \mathrm{Var}_{X\sim N(\mu^*,\sigma^2)}\left(\frac{\partial l(\theta|X)}{\partial \mu}\bigg|_{\mu=\mu^*}\right)$$

- Higher variance above implies that the absolute value of the score is often high. Hence more information is carried.

Fisher Information

The Fisher information is defined as

$$\mathcal{I}(\theta) = \mathrm{Var}_\theta(s(\theta|X)) = \mathrm{Var}_\theta\left(\frac{\partial l(\theta|X)}{\partial \theta}\right).$$

Alternative Expressions of Fisher Information

Assuming that f is differentiable in θ and we can exchange the order of differentiation and integration,

$$\int \frac{\partial}{\partial \theta} f(x|\theta) dx = \frac{\partial}{\partial \theta} \int f(x|\theta) dx = 0$$

Hence

$$\mathbb{E}_\theta[s(\theta|X)] = \mathbb{E}_\theta\left[\frac{\partial l(\theta|X)}{\partial \theta}\right] = \int \frac{\partial l(\theta|x)}{\partial \theta} f(x|\theta) dx = \int \frac{\frac{\partial}{\partial \theta} f(x|\theta)}{f(x|\theta)} f(x|\theta) dx = 0.$$

Alternative expression I for Fisher information

$$\mathcal{I}(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial l(\theta|X)}{\partial \theta}\right)^2\right] = \int \left(\frac{\partial}{\partial \theta} \log f(x|\theta)\right)^2 f(x|\theta) dx.$$

Assuming that f is twice differentiable in θ and we can exchange the order of differentiation and integration,

$$\int \frac{\partial^2}{\partial \theta^2} f(x|\theta) dx = \frac{\partial^2}{\partial \theta^2} \int f(x|\theta) dx = 0.$$

Also, notice that

$$\frac{\partial^2 l(\theta|x)}{\partial \theta 2} = \frac{\partial}{\partial \theta} \frac{\partial l(\theta|x)}{\partial \theta} = \frac{\partial}{\partial \theta} \left[\frac{\frac{\partial}{\partial \theta} f(x|\theta)}{f(x|\theta)} \right] = \frac{\frac{\partial^2}{\partial \theta^2} f(x|\theta)}{f(x|\theta)} - \left[\frac{\partial l(\theta|x)}{\partial \theta} \right]^2$$

Alternative expression II for Fisher information

$$\mathcal{I}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2 l(\theta|X)}{\partial \theta^2} \right] = -\int \left[\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right] f(x|\theta) dx.$$

- This is usually the most convenient one to use.

To interpret

$$\mathcal{I}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2 l(\theta|X)}{\partial \theta^2} \right] = -\int \left[\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right] f(x|\theta) dx.$$

Note that

$$-\frac{\partial^2 l(\theta|X)}{\partial \theta^2}$$

measures the curvature of the loglikelihood function $l(\theta|X)$ at θ. The higher the (expected) curvature, the more information the random variable X provides about θ.

Remarks – Random Sample

We've looked at Fisher information in a random variable X. Now, let's consider the amount of information a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ provides fot θ.

In this case, Fisher information $\mathcal{I}_n(\theta)$ in $\boldsymbol{X}$ is defined as

$$\mathcal{I}_n(\theta) = \mathsf{Var}_\theta\left(s(\theta|\boldsymbol{X})\right) = \mathbb{E}_\theta\left[\left(\frac{\partial l(\theta|\boldsymbol{X})}{\partial \theta}\right)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial^2 l(\theta|\boldsymbol{X})}{\partial \theta^2}\right].$$

Using the i.i.d. assumption, $L_n(\theta|\boldsymbol{X}) = \prod_i f(X_i|\theta)$ and $l(\theta|\boldsymbol{X}) = \sum_i l(\theta|X_i)$. We can show that

$$\mathbb{E}_\theta\left[s(\theta|\boldsymbol{X})\right] = \mathbb{E}_\theta\left[\frac{\partial l(\theta|\boldsymbol{X})}{\partial \theta}\right] = \mathbb{E}_\theta\left[\sum_{i=1}^n \frac{\partial l(\theta|X_i)}{\partial \theta}\right] = 0.$$

$$\mathcal{I}_n(\theta) = -\mathbb{E}_\theta\left[\sum_{i=1}^n \frac{\partial^2 l(\theta|X_i)}{\partial \theta^2}\right] = n\mathcal{I}(\theta).$$

Example: Normal(μ, σ^2) random variable with unknown $\theta = \mu$ and known σ.

$$-\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) = \frac{1}{\sigma^2}$$

Hence

$$\mathcal{I}(\mu) = \frac{1}{\sigma^2}.$$

Example: Normal(μ, σ^2) random sample of size n with unknown $\theta = \mu$ and known σ.

$$\mathcal{I}_n(\mu) = n\mathcal{I}(\mu) = \frac{n}{\sigma^2}.$$

Example: Normal(μ, σ^2) random variable with unknown $\theta = \sigma$ and known μ.

$$-\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) = -\frac{1}{\sigma^2} + 3\frac{(x-\mu)^2}{\sigma^4}$$

Hence

$$\mathcal{I}(\sigma) = \frac{2}{\sigma^2}.$$

Example: Normal(μ, σ^2) random variable with unknown $\theta = \sigma^2$ and known μ.

$$-\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) = -\frac{1}{2\theta^2} + \frac{(x-\mu)^2}{\theta^3}$$

Hence

$$\mathcal{I}(\sigma^2) = \frac{1}{2\sigma^4}.$$

Remarks – Matrix Form

Fisher Information (matrix form)

The Fisher information matrix for $\boldsymbol{\theta} \in \mathbb{R}^k$ is defined as

$$\begin{aligned}\mathcal{I}(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{\theta}}\left[\left(\frac{\partial l(\boldsymbol{\theta}|X)}{\partial \boldsymbol{\theta}}\right)^T \left(\frac{\partial l(\boldsymbol{\theta}|X)}{\partial \boldsymbol{\theta}}\right)\right] \\ &= -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 l(\boldsymbol{\theta}|X)}{\partial \boldsymbol{\theta}^2}\right]\end{aligned}$$

Here

$$\frac{\partial^2 l(\boldsymbol{\theta}|X)}{\partial \boldsymbol{\theta}^2} = \left[\frac{\partial^2 l(\boldsymbol{\theta}|X)}{\partial \theta_i \partial \theta_j}\right] \in \mathbb{R}^{k\times k}$$

is the Hessian matrix of $l(\boldsymbol{\theta}|X)$ w.r.t. $\boldsymbol{\theta}$.

Cramér-Rao Bound

Now we return to the point estimators

- We have seen that the information in the random sample is characterized by Fisher information.
- With the finite amount of information, it is natural to believe that we cannot do infinitely good in estimation, that is, the variance of our estimators must be somehow lower bounded by the amount of information.

This heuristic is supported by the Cramér-Rao inequality.

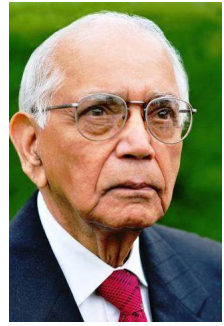


Figure: Rao, student of Fisher, recently passed away in August 2023 (aged 102).

The Fisher information is the variance of the score function evaluated at the true parameter

$$\mathcal{I}(\theta) = \mathrm{Var}_\theta(s(\theta|X)).$$

If we know the covariance between the estimator $\widehat{\theta}$ and the score $s(\theta|\boldsymbol{X})$, then we can use Cauchy-Schwartz inequality to obtain a lower bound of $\mathrm{Var}(\widehat{\theta})$.

Cauchy-Schwartz inequality

$$\Big(\mathrm{Cov}_\theta(\widehat{\theta}, s(\theta|\boldsymbol{X}))\Big)^2 \leq \mathrm{Var}_\theta(\widehat{\theta})\mathrm{Var}_\theta(s(\theta|\boldsymbol{X})) = \mathrm{Var}_\theta(\widehat{\theta})n\mathcal{I}(\theta).$$

At first glance, this is useless, because $\mathrm{Cov}(\widehat{\theta}, s(\theta|\boldsymbol{X}))$ may depend on $\widehat{\theta}$. Nevertheless, let's calculate

$$\mathrm{Cov}(\widehat{\theta}, s(\theta|\boldsymbol{X})).$$

Now, let $m(\theta)$ be the expectation of $\widehat{\theta}$

$$\mathbb{E}_\theta[\widehat{\theta}] = m(\theta).$$

Recall that the expectation of the score evaluated at the true parameter is zero

$$\mathbb{E}_\theta\left[s(\theta|\boldsymbol{X})\right] = 0, \text{ for all } \theta.$$

Hence

$$\begin{aligned}
\text{Cov}\left(\widehat{\theta}, s\left(\theta|\boldsymbol{X}\right)\right) = \mathbb{E}_\theta\left[\widehat{\theta}s(\theta|\boldsymbol{X})\right] &= \int \widehat{\theta}(\boldsymbol{x})\frac{\partial l(\theta|\boldsymbol{x})}{\partial\theta} f(\boldsymbol{x}|\theta)d\boldsymbol{x} \\
&= \int \widehat{\theta}(\boldsymbol{x})\frac{\frac{\partial f(\boldsymbol{x}|\theta)}{\partial\theta}}{f(\boldsymbol{x}|\theta)} f(\boldsymbol{x}|\theta)d\boldsymbol{x} \\
&= \int \widehat{\theta}(\boldsymbol{x})\frac{\partial f(\boldsymbol{x}|\theta)}{\partial\theta} d\boldsymbol{x} \\
\text{(assume)} &= \frac{\partial}{\partial\theta}\int \widehat{\theta}(\boldsymbol{x}) f(\boldsymbol{x}|\theta)d\boldsymbol{x} = m'(\theta)
\end{aligned}$$

Cramér-Rao inequality

Consider an estimator $\widehat{\theta}(\boldsymbol{X})$ (based on i.i.d. sample $\boldsymbol{X}$) with

$$\mathbb{E}_\theta[\widehat{\theta}] = m(\theta).$$

Under the assumption that

$$m'(\theta) = \int \widehat{\theta}(\boldsymbol{x}) \frac{\partial}{\partial \theta} f(\boldsymbol{x}|\theta) d\boldsymbol{x}, \quad \text{and} \quad \mathrm{Var}_\theta \widehat{\theta} < \infty,$$

we have

$$\mathrm{Var}_\theta(\widehat{\theta}) \geq \frac{[m'(\theta)]^2}{n\mathcal{I}(\theta)}, \text{ for all } \theta.$$

For unbiased estimators, $m'(\theta) = 1$, hence

Cramér-Rao inequality for unbiased estimators

Consider an unbiased estimator $T(\boldsymbol{X})$ of θ (based on i.i.d. sample $\boldsymbol{X}$). Then under suitable assumptions,

$$\mathrm{Var}_\theta T(\boldsymbol{X}) \geq \frac{1}{n\mathcal{I}(\theta)}, \text{ for all } \theta.$$

Remarks

- Cramér and Rao independently developed the inequality in the 1940's.
- As the information $\mathcal{I}(\theta)$ increase, the variance of the estimator decreases. Hence, the bound is also called the information bound.
- $\frac{1}{n\mathcal{I}(\theta)}$ is called the Cramér-Rao lower bound (CRLB): under suitable conditions, no unbiased estimator of the parameter θ based on a i.i.d. sample of size n can have a smaller variance.

If we have found a estimator that achieves CRLB, then we have found the UMVUE.

The Inequality Fails: Uniform

Example: Let $\boldsymbol{x}$ be a sample from Uniform $(0, \theta)$, the Fisher information number

$$\mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \right] = \frac{1}{\theta^2}$$

So the CRLB of the MSE of any unbiased estimator W is $\mathrm{Var}_\theta W \geq \frac{\theta^2}{n}$.
We have seen that $\mathbb{E}[d_2(\mathbf{X})] = \frac{n}{n+1}\theta$, so $\widehat{\theta} = \frac{n+1}{n} d_2(\boldsymbol{X})$ is unbiased.

$$\mathrm{Var}_\theta(\widehat{\theta}) = \frac{(n+1)^2}{n^2} \frac{n\theta^2}{(n+2)(n+1)^2} = \frac{1}{n(n+2)}\theta^2 \leq \frac{\theta^2}{n}$$

What is the problem? The condition of the Cramér-Rao inequality is not satisfied, i.e., we cannot interchange the order of $\int$ and $\partial/\partial\theta$. In general, if the range of the PDF depends on the parameter, the theorem will not be applicable.

Examples

Example: Bernoulli(p) estimators.

Note that

$$\frac{\partial^2 l(p|x)}{\partial p^2} = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}.$$

The Fisher information is

$$\mathcal{I}(p) = \frac{n}{p(1-p)}.$$

- Consider $\widehat{\theta} = \bar{X}$. We have $\mathbb{E}[\bar{X}] = p$ and $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n = p(1-p)/n$. So $\bar{X}$ is UMVUE.

Example: Poisson(λ) estimators.

$$\mathbb{E}_\lambda\left[\left(\frac{\partial}{\partial\lambda}\log\prod_{i=1}^{n}f(X_i|\lambda)\right)^2\right] = -n\mathbb{E}_\lambda\left[\frac{\partial^2}{\partial\lambda^2}\log f(X|\lambda)\right]$$

$$= -n\mathbb{E}_\lambda\left[\frac{\partial^2}{\partial\lambda^2}\log\left(\frac{e^{-\lambda}\lambda^X}{X!}\right)\right] = -n\mathbb{E}_\lambda\left[-\frac{X}{\lambda^2}\right] = \frac{n}{\lambda}$$

So by Cramér-Rao, for any unbiased W

$$\mathrm{Var}_\lambda W \geq \frac{\lambda}{n}$$

Since the estimator $\bar{X}$ satisfies $\mathrm{Var}_\lambda\bar{X} = \frac{\lambda}{n}$, it is UMVUE.

Example: Normal $N(\mu, \sigma^2)$.

- For estimator of μ, CR states that $\mathrm{Var}(\widehat{\theta}) \geq \sigma^2/n$. The sample mean $\bar{X}$ attains the lower bound and is UMVUE.
- For estimator of σ^2, CR states that

$$\frac{\partial^2}{\partial(\sigma^2)^2} \log\left(\frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\right) = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}$$

$$-\mathbb{E}_{\mu,\sigma^2}\left[\frac{\partial^2}{\partial(\sigma^2)^2} \log f(X|\mu,\sigma^2)\right] = -\mathbb{E}_{\mu,\sigma^2}\left[\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}\right] = \frac{1}{2\sigma^4}$$

  So the lower bound is $\frac{2\sigma^4}{n}$. If μ is known, then $\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ is UMVUE.
- When μ is unknown, $\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}$. Can we do better?

Attainment of Cramér-Rao Lower Bound

- From the proof of CR, we used Cauchy-Schwartz inequality

$$\mathrm{Var}(\widehat{\theta}) \geq \frac{\mathrm{Cov}(\widehat{\theta}, s(\theta|\boldsymbol{X}))^2}{\mathrm{Var}(s(\theta|\boldsymbol{X}))}$$

  for an *unbiased* $\widehat{\theta}$, the lower bound is attained if and only if

$$\widehat{\theta} = \theta + a(\theta)s(\theta|\boldsymbol{X}).$$

  In particular, $a(\theta)$ must be deterministic.

- In the normal example, $s(\sigma^2|\boldsymbol{X}) = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (X_i-\mu)^2}{2\sigma^4}$. It requires $\widehat{\theta} = \sigma^2 + a(\sigma^2)s(\sigma^2|\boldsymbol{X})$.
- If μ is unknown, no estimator can be of that form. In other words, the lower bound of CR cannot be attained.

Use Sufficient Statistics to Construct a Better Estimator

Let W be an estimator of $\tau(\theta)$ and T be a sufficient statistic for θ. The statistic $\phi(T) = \mathbb{E}[W|T]$ is called the *Rao-Blackwell estimator* for $\tau(\theta)$.

Theorem (Rao-Blackwell)

If W is an unbiased estimator of $\tau(\theta)$, then

$$\mathbb{E}_\theta[\phi(T)] = \tau(\theta) \qquad \mathrm{Var}_\theta(\phi(T)) \leq \mathrm{Var}_\theta(W)$$

Proof:

$$\tau(\theta) = \mathbb{E}_\theta W = \mathbb{E}_\theta[\mathbb{E}(W|T)] = \mathbb{E}_\theta[\phi(T)]$$

$$\begin{aligned} \mathrm{Var}_\theta W &= \mathrm{Var}_\theta[\mathbb{E}(W|T)] + \mathbb{E}_\theta[\mathrm{Var}(W|T)] \\ &= \mathrm{Var}_\theta[\phi(T)] + \mathbb{E}_\theta[\mathrm{Var}(W|T)] \\ &\geq \mathrm{Var}_\theta[\phi(T)] \end{aligned}$$

Why Sufficiency Is Needed

- For any statistic Y, we can always consider $\phi(Y) = \mathbb{E}[W|Y]$. Then $\mathbb{E}[\mathbb{E}[W|Y]] = \mathbb{E}[W] = \theta$ and $\mathrm{Var}_\theta(Y) \leq \mathrm{Var}_\theta(W)$. Does it mean we can always improve an unbiased estimator?
- Let X_1, X_2 be i.i.d. $\mathcal{N}(\theta, 1)$. We know $\bar{X} = (X_1 + X_2)/2$ is sufficient for θ. Let $\phi'(X_1) = \mathbb{E}[\bar{X}|X_1]$. Then we can show $\mathrm{Var}(\phi'(X_1)) < \mathrm{Var}(\bar{X})$. However, $\phi'(X_1) = (X_1 + \theta)/2$ is not a statistic (estimator). Hence, it does not contradict the Rao-Blackwell Theorem.
- Sufficiency implies that $\phi(T)$ is indeed a statistic (does not depend on θ). Becaues conditional on T, the sample (and hence the statistic) does not depend on θ.
- Sufficiency and MLE: If $T(\boldsymbol{X})$ is sufficient for θ and if MLE exists, then there exists an MLE that is a function of T. **Proof:** the likelihood function is of the form $\log g(T(\boldsymbol{x})|\theta) + \log h(\boldsymbol{x})$. To maximize over θ, only the first term is involved.

Completeness and UMVUE

- For each unbiased estimator, Rao-Blackwell gives a better one.
- But for another unbiased estimator W', RB does not tell us whether $\mathrm{Var}(\phi(T)) \leq \mathrm{Var}(W')$ or not. How to find the a UMVUE?

Theorem (Lehmann-Scheffè)

- *Let T be a complete sufficient statistic for a parameter θ*
- *Let $\phi(T)$ be any estimator based only on T.*

Then $\phi(T)$ is UMVUE for $\mathbb{E}_\theta[\phi(T)]$.

- Proof: Let W be another unbiased estimator for $\mathbb{E}[\phi(T)]$. Consider $\mathbb{E}[W|T]$ and let $g(T) = \phi(T) - \mathbb{E}[W|T]$. We have $\mathbb{E}[g(T)] = 0$ and thus by completeness, $\phi(T) = \mathbb{E}[W|T]$. By R-B, $\phi(T)$ is uniformly better than W.

Uniqueness of Best Unbiased Estimator

Theorem

UMVUE is unique (if it exists).

Suppose there are two UMVUE W_1 and W_2, consider $W^* = (W_1 + W_2)/2$

$$\begin{aligned}\mathrm{Var}_\theta(W^*) &= \frac{1}{4}\mathrm{Var}_\theta(W_1) + \frac{1}{4}\mathrm{Var}_\theta(W_2) + \frac{1}{2}\mathrm{Cov}_\theta(W_1, W_2) \\ &\leq \frac{1}{4}\mathrm{Var}_\theta(W_1) + \frac{1}{4}\mathrm{Var}_\theta(W_2) + \frac{1}{2}\left[\mathrm{Var}_\theta(W_1)\mathrm{Var}_\theta(W_2)\right]^{1/2} \qquad \text{(Cauchy-Schwarz)} \\ &= \mathrm{Var}_\theta(W_1)\end{aligned}$$

Strict inequality cannot hold since otherwise W^* is a better one. This means C-S holds in equality, hence $W_1 = a(\theta)W_2 + b(\theta)$. Using $\mathrm{Var}_\theta(W_1) = \mathrm{Cov}_\theta(W_1, W_2) = a(\theta)\mathrm{Var}_\theta(W_1)$, we have $a(\theta) = 1$. Since both W_1 and W_2 are unbiased, we have $b(\theta) = 0$.

Example – Direct application of L-S

Finding UMVUE becomes easy if you have complete sufficient statistics.

Method 1: find directly a function ϕ such that $\mathbb{E}[\phi(T)] = \theta$, where T is a S-C statistic.

Example: Bernoulli sample variance.

- For Bernoulli(p) samples $X_1, \ldots, X_n$, $\bar{X}$ is sufficient and complete. The sample variance S^2 is unbiased for $p(1-p)$.
- Because $X_i^2 = X_i$ for Bernoulli, we have

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = \frac{n}{n-1}\bar{X}(1-\bar{X}),$$

it is UMVUE.

This works when CR is unattainable.

Example: For $\mathcal{N}(\mu, \sigma^2)$ with unknown μ

- S^2 is UMVUE for σ^2 because S^2 is unbiased and is a function of $(\bar{X}, S^2)$.

This works when the condition in CR is not satisfied.

Example: Uniform $(0, \theta)$. Let $Y = \max\{X_1, \ldots, X_n\}$. Recall that CR lower bound cannot be used here.

- Y is complete and sufficient, and $\frac{n+1}{n}Y$ is unbiased, it is UMVUE.

Example: Uniform $(0, \theta)$. What if we are looking at $\tilde{\theta} = g(\theta)$, where g is differentiable?

Recall the pdf of $X_{(n)}$ is $n\theta^{-n}x^{n-1}\mathbb{1}(0 \leq x \leq \theta)$. If $h(X_{(n)})$ is unbiased for $\tilde{\theta}$, then

$$\theta^n g(\theta) = n \int_0^\theta h(x)x^{n-1}dx, \text{ for all } \theta > 0.$$

Take derivatives w.r.t. θ on both sides

$$n\theta^{n-1}g(\theta) + \theta^n g'(\theta) = nh(\theta)\theta^{n-1}.$$

Hence,

$$h(X_{(n)}) = g(X_{(n)}) + n^{-1}X_{(n)}g'(X_{(n)})$$

is UMVUE.

Example – The Conditioning Method

Method 2: Find any unbiased estimator W; then $\mathbb{E}[W|T]$ is UMVUE.

For minimal exponential families of full-rank, finding UMVUE is not an issue. Although $\mathbb{E}[W|T]$ does not have a tractable form in many cases.

Example – The Conditioning Method

Example: Consider Binomial(k, θ). Want to estimate

$$\tau(\theta) = \mathbb{P}_\theta(X = 1) = k\theta(1-\theta)^{k-1}.$$

Let $X_1, \ldots, X_n$ be a sample, then $\sum_{i=1}^n X_i \sim$Binomial(kn, θ) is complete and sufficient.

- First step, find an unbiased estimator.

$$h(X_1) = \mathbb{1}(X_1 = 1) \quad \Rightarrow \quad \mathbb{E}_\theta[h(X_1)] = k\theta(1-\theta)^{k-1} \quad \text{unbiased!}$$

- Second step, compute $\phi(\sum_{i=1}^n X_i) = \mathbb{E}\left[h(X_1) | \sum_{i=1}^n X_i\right]$, the conditional probability of $X_1 = 1$ given $\sum_{i=1}^n X_i$.

Cont.

Suppose we observe $\sum_{i=1}^n X_i = t$

$$\phi(t) = \mathbb{E}\left(h(X_1)\middle|\sum_{i=1}^n X_i = t\right) = \mathbb{P}\left(X_1 = 1\middle|\sum_{i=1}^n X_i = t\right)$$
$$= \frac{\mathbb{P}_\theta(X_1 = 1, \sum_{i=1}^n X_i = t)}{\mathbb{P}_\theta(\sum_{i=1}^n X_i = t)} = \frac{\mathbb{P}_\theta(X_1 = 1)\mathbb{P}_\theta(\sum_{i=2}^n X_i = t-1)}{\mathbb{P}_\theta(\sum_{i=1}^n X_i = t)}$$

Then is it straightforward to compute (note there is no θ due to sufficiency!)

$$\phi(t) = \frac{\binom{k}{1}\binom{k(n-1)}{t-1}}{\binom{kn}{t}}$$

Plugging in $t = \sum_{i=1}^n X_i$, we have found the UMVUE.

A Nonparametric Example – Empirical CDF

Let $(X_1, \ldots, X_n)$ be i.i.d. from an unknown CDF F. Suppose that the parameter of interest is $\theta = 1 - F(t)$ for a fixed t. If F is not in a parametric family, then a nonparametric estimator of $F(t)$ is the empirical CDF.

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq t)$$

Note that the random variable $nF_n(t) \sim \text{Binomial}(n, F(t))$. It is easy to see that $F_n(t)$ is an unbiased estimator of $F(t)$ with $\text{MSE}_{F(t)} F_n(t) = F(t)(1 - F(t))/n$. Hence $U(\boldsymbol{X}) = 1 - F_n(t)$ is unbiased for θ with the same MSE.

A Nonparametric Example – Empirical CDF

In last lecture, we have seen that, when Θ is all distributions with a density function, the empirical CDF is sufficient and complete.

As a result, $U(\boldsymbol{X}) = 1 - F_n(t)$ is an UMVUE for θ and the empirical CDF is an UMVUE for the CDF F.

A Nonparametric Example – Empirical CDF

Suppose now we have further information on F, say it is the CDF of Exp$(1/\theta)$ for some $\theta > 0$. We can use this piece of information to improve the MSE of $U(\boldsymbol{X})$.

We know that $\bar{X}$ is sufficient and complete for θ hence also for $e^{-t/\theta} = 1 - F(t)$, using the L-S theorem, the estimator $T(\boldsymbol{X}) = \mathbb{E}[U(\boldsymbol{X}) \mid \bar{X}]$, which is also unbiased, is UMVUE and is better than $U(\boldsymbol{X})$ in terms of MSE.

Here, the set of distribution is restricted to exponential distributions, and empirical CDF is no longer complete.

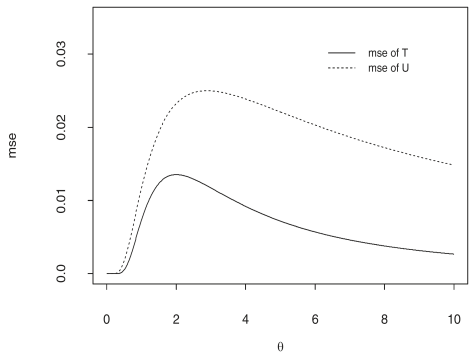


Figure: Cover illustration of Jun Shao's book.

Summary

Usual method to find UMVUE

1. Check if a estimator attains the CRLB; see S.49-51.
   - Not an effective way.
   - But CRLB can be used to assess the performance of an estimator.
2. Directly solving for ϕ; see S.57-59.
   - Try some function h and see if $\mathbb{E}[h(T)]$ is related to θ.
   - May require some tricks, e.g. by comparing coefficient in polynomials or differentiating a integral.
3. Conditioning on a sufficient and complete statistic; see S.61-62.
   - Need to work out the conditional expectation. Usually not tractable.
   - By uniqueness of UMVUE, it doesn't matter which unbiased estimator you use; pick the one that makes the conditional expectation as easy as possible.

Reading Materials

Same level

- Robert W. Keener, Theoretical Statistics, Chapter 4. Especially Section 4.5 for another introduction of Fisher information.
- Casella and Berger, Statistical Inference, Chapter 7.

Advanced

- Jun Shao, Mathematical Inference, Section 2.3, 2.4.1, 3.1.
  - See Section 3.1.2 for the case where a sufficient and complete statistic is not available.