# Consistency and Efficiency of Estimators

November 15, 2023

# The bias and unbiased estimators

In the previous lecture, we have seen an approach to estimation that is based on the likelihood of observed results. Next, we study general theory of estimation that is used to compare between different estimators and to decide on the most efficient one.

- Suppose that we are going to observe a value of a random vector $\mathbf{X}$. Let $\mathcal{X}$ denote the set of possible values $\mathbf{X}$ can take and, for $\mathbf{x} \in \mathcal{X}$, let $f(\mathbf{x}|\theta)$ denote the probability (or density) that $\mathbf{X}$ takes the value $\mathbf{x}$ where the parameter $\theta$ is some unknown element of the set $\Theta$.
- An estimator $\hat{\theta}$ is a procedure that for each possible value $\mathbf{x} \in \mathcal{X}$ specifies which element of $\Theta$ we should report as an estimate of $\theta$. When we observe $\mathbf{X} = \mathbf{x}$, we quote $\hat{\theta}(\mathbf{x})$ as our estimate of $\theta$. Thus $\hat{\theta}$ is a function of the random variable $\mathbf{X}$. Sometimes we write $\hat{\theta}(\mathbf{X})$ to emphasise this point.

# Bias and unbiased estimators

To evaluate the usefulness of an estimator $\hat{\theta} = \hat{\theta}(\mathbf{x})$ of $\theta$, examine the properties of the random variable $\hat{\theta} = \hat{\theta}(\mathbf{X})$.

- Given an estimator $\hat{\theta}$, we can calculate its expected value for each possible value of $\theta \in \Theta$.
- An estimator is said to be **unbiased** if this expected value is equal to $\theta$.
- If an estimator is unbiased then, by the Law of Large Numbers, we can conclude that by repeating the experiment an infinite number of times (never happens, not even twice:) with $\theta$ fixed and calculate the value of the estimator each time then the average of the estimator values will be exactly equal to $\theta$.

**Definition (Unbiased estimators)**

An estimator $\hat{\theta} = \hat{\theta}(\mathbf{X})$ is said to be unbiased for a parameter $\theta$ if it equals $\theta$ in expectation

$$E_\theta[\hat{\theta}(\mathbf{X})] = E_\theta(\hat{\theta}) = \theta.$$

Intuitively, an unbiased estimator is 'right on target'. In general, $Bias_\theta(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta$. $\square$

# Examples

The most well-known estimators are the sample mean and the sample variance

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

The strange factor $\frac{n}{n-1}$ is to force the unbiasedness of $S^2$ (Why?).

- Note that even if $\hat{\theta}$ is an unbiased estimator of $\theta$, $g(\hat{\theta})$ will generally not be an unbiased estimator of $g(\theta)$ unless $g$ is linear or affine.
- This limits the importance of the notion of unbiasedness. It might be at least as important that an estimator is accurate in the sense that its distribution is highly concentrated around $\theta$.
- For an arbitrary distribution, the estimator $S^2$ is an unbiased estimator of the variance of this distribution.
- Consider the estimator $S^2$ of variance $\sigma^2$ in the case of the normal distribution. Although $S^2$ is an unbiased estimator of $\sigma^2$, $S$ is not an unbiased estimator of $\sigma$.

# Mean Square Error

## Definition (Mean squared error)

The mean squared error of the estimator $\hat{\theta}$ is defined as $\mathrm{MSE}_\theta(\hat{\theta}) = E_\theta(\hat{\theta} - \theta)^2$. Given the same set of data, $\hat{\theta}_1$ is "better" than $\hat{\theta}_2$ if $MSE_\theta(\hat{\theta}_1) \leq MSE_\theta(\hat{\theta}_2)$ (uniformly better if true for all $\theta$). □

## Lemma (The MSE variance-bias tradeoff)

The MSE decomposes as $MSE_\theta(\hat{\theta}) = Var_\theta(\hat{\theta}) + Bias_\theta(\hat{\theta})^2$.

Proof.

$$\begin{aligned} MSE(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 = E\{\,[\,\hat{\theta} - E(\hat{\theta})\,] + [\,E(\hat{\theta}) - \theta\,]\}^2 = E[\hat{\theta} - E(\hat{\theta})]^2 + E[E(\hat{\theta}) - \theta]^2 \\ &\quad + 2\,\underbrace{E\left\{[\hat{\theta} - E(\hat{\theta})][E(\hat{\theta}) - \theta]\right\}}_{=0} \\ &= E[\hat{\theta} - \mathbb{E}(\hat{\theta})]^2 + E[E(\hat{\theta}) - \theta]^2 = Var(\hat{\theta}) + \underbrace{[E(\hat{\theta}) - \theta]^2}_{Bias(\hat{\theta})^2}. \end{aligned}$$

□

NOTE: This lemma implies that the mean squared error of an unbiased estimator is equal to the variance of the estimator.

# The minimal variance linear mean estimator

Let $X_1, \ldots, X_n$ be independent random variables with means $\mathbb{E}(X_i) = \mu$ and variances $\mathbb{V}ar(X_i) = \sigma_i^2$. Consider pooling the estimators of $\mu$ into a common estimator using the linear combination $\hat{\mu} = w_1 X_1 + w_2 X_2 + \cdots + w_n X_n$. We will see that the following is true

(i) The estimator $\hat{\mu}$ is unbiased if and only if $\sum w_i = 1$.

(ii) The estimator $\hat{\mu}$ has minimum variance among this class of estimators when the weights are inversely proportional to the variances $\sigma_i^2$.

(iii) The variance of $\hat{\mu}$ for optimal weights $w_i$ is $\mathbb{V}ar(\hat{\mu}) = 1/\sum_i \sigma_i^{-2}$.

# Minimum-Variance Unbiased Estimation

Getting a small MSE often involves a tradeoff between variance and bias. For unbiased estimators, the MSE obviously equals the variance, $MSE(\hat{\theta}) = Var(\hat{\theta})$, so no tradeoff can be made. One approach is to restrict ourselves to the subclass of estimators that are *unbiased* and *minimum variance*.

**Definition (Minimum-variance unbiased estimator)**

If an unbiased estimator of $g(\theta)$ has minimum variance (for all possible values of $\theta$) among all unbiased estimators of $g(\theta)$ it is called a uniformly minimum variance unbiased estimator (UMVUE or MVUE). □

We will develop a method of finding the MVUE when it exists. When such an estimator does not exist we will be able to find a lower bound for the variance of an unbiased estimator in the class of unbiased estimators, and compare the variance of our unbiased estimator with this lower bound.

# Assumptions

- We suppose throughout that we have a regular parametric model and further that $\Theta$ is an open subset of the line.
- The support set of the distributions does not depend on $\theta \in \Theta$.
- $\partial log\, p(x, \theta)/\partial\theta$ exists.
- Then, for a statistics $T$, we have the following change of integration with derivation

$$\frac{\partial}{\partial\theta} E_\theta(T) = \frac{\partial}{\partial\theta} \int_{\mathcal{X}} T(x) p(x|\theta)\, dx = \int_{\mathcal{X}} \frac{\partial}{\partial\theta} T(x) p(x|\theta)\, dx$$

In our discussion, we will work under these assumptions. We do not discuss in any detail when these assumptions are satisfied. What should be however remembered is that

- these assumptions are satisfied by exponential families in the one- and multiparameter cases,
- they are frequently valid both for the discrete and continuous distributions.

# The score function

## Definition (Score function)

For the (possibly vector valued) observation $X = x$ to be informative about $\theta$, the density must vary with $\theta$. If $f(x|\theta)$ is smooth and differentiable, then for finding MLE we have used zeros of the score function

$$S(\theta) = S(\theta|x) = \frac{\partial}{\partial\theta} \log p(x|\theta) \equiv \frac{\partial p(x|\theta)/\partial\theta}{p(x|\theta)}.$$

If differentiation wrt $\theta$ and integration wrt $x$ can be interchanged, as we assume, we have for $X$ distributed according to $p(x|\theta)$:

$$E_\theta\{S(\theta|X)\} = \int \frac{\partial p(x|\theta)/\partial\theta}{p(x|\theta)} p(x|\theta)dx = \int \partial p(x|\theta)/\partial\theta dx = \frac{\partial}{\partial\theta}\left\{\int p(x|\theta)dx\right\} = \frac{\partial}{\partial\theta} 1 = 0.$$

- Thus the score function has expectation zero.
- The score function $S(\theta|X)$, for a fixed $\theta$, is a random variable (but it is not a statistic, why?).
- We often drop explicit dependence on $X$ from the notation by simply writing $S(\theta)$.
- The negative derivative of the score function $I_{obs}(\theta) = -\partial S(\theta)/\partial\theta$ measures how concave down is the likelihood around value $\theta$.

# The Fisher Information

## Definition (Fisher information)

The Fisher information is defined as the average value of the minus derivative of the score function

$$I(\theta) \equiv -E_\theta \left( \frac{\partial}{\partial \theta} S(\theta) \right) = E_\theta I_{obs}(\theta).$$

The negative derivative of the score function $I_{obs}(\theta)$, which is a random variable dependent on $X$, is referred to as empirical or observed information about $\theta$. □

## Lemma

$$I(\theta) = Var_\theta \, S(\theta)$$

We note that $Var_\theta S = E_\theta S^2$ and

$$\frac{\partial S}{\partial \theta} = \frac{\partial^2 \log p}{\partial \theta^2} \quad = \quad \frac{\partial}{\partial \theta}\left[\frac{1}{p}\frac{\partial p}{\partial \theta}\right] = -\frac{1}{p^2}\left[\frac{\partial p}{\partial \theta}\right]^2 + \frac{1}{p}\frac{\partial^2 p}{\partial \theta^2} = -\left[\frac{\partial \log p}{\partial \theta}\right]^2 + \frac{1}{p}\frac{\partial^2 p}{\partial \theta^2} = -S^2 + \frac{1}{p}\frac{\partial^2 p}{\partial \theta^2}$$

If integration and differentiation can be interchanged

$$E_\theta \left[\frac{1}{p}\frac{\partial^2 p}{\partial \theta^2}\right] = \int_X \frac{\partial^2 p}{\partial \theta^2} dx = \frac{\partial^2}{\partial \theta^2} \int_X p dx = \frac{\partial^2}{\partial \theta^2} 1 \; = \; 0.$$

# The Cramér-Rao lower bound (CRLB)

**Theorem (Information Inequalit)**

*Let $\hat{\theta}$ be an unbiased estimator of $\theta$. Then $Var_\theta \hat{\theta} \geq 1/I(\theta)$.*

**Proof.**

Unbiasedness means that $\mathbb{E}_\theta(\hat{\theta}) = \int \hat{\theta}(x) p(x|\theta) dx = \theta$. Assume we can differentiate wrt $\theta$ under the integral, then $\int \frac{\partial}{\partial\theta} \left\{ \hat{\theta}(x) p(x|\theta) \right\} dx = \int \hat{\theta}(x) \frac{\partial}{\partial\theta} p(x|\theta)\, dx = 1$ since the estimator $\hat{\theta}(x)$ can't depend on $\theta$.
Now, $\frac{\partial p}{\partial\theta} = p \frac{\partial}{\partial\theta} (\log p)$, so that

$$\int \hat{\theta}(x) p \frac{\partial}{\partial\theta} (\log p)\, dx = E_\theta \left[ \hat{\theta}(x) \frac{\partial}{\partial\theta} (\log p) \right] = E_\theta(\hat{\theta} S) = 1.$$

We already know that the score function has expectation zero, $E_\theta(S) = 0$. Consequently $Cov_\theta(\hat{\theta}, S) = E_\theta(\hat{\theta} S) - E_\theta(\hat{\theta}) E_\theta(S) = E_\theta(\hat{\theta} S) = 1$. By Schwartz's inequality

$$1 = Cov_\theta(\hat{\theta}, S)^2 \leq Var_\theta(\hat{\theta}) Var_\theta(S)$$

This implies the Rao-Cramér inequality. □

# Comments

- Why **'information'**? – Variance measures the lack of knowledge. The reciprocal of the variance could be defined as the amount of information carried by the (possibly vector valued) random observation $X$ about $\theta$.
- We require that the ranges of the integrals do not depend on $\theta$. That is, the range of $x$, here $p(x|\theta)$, cannot depend on $\theta$. This second condition is violated for some density functions, i.e. the CRLB is not valid for the uniform distribution.
- We can have absolute assessment for unbiased estimators by comparing their variances to the CRLB. We can also assess biased estimators. If its variance is lower than CRLB then it can be indeed a very good estimate, although it is biased.
- In the iid case, i.e. $p(x|\theta) = p_1(x_1|\theta) \dots p_1(x_n|\theta)$, then $I(\theta) = nI_1(\theta)$, where the $I_1(\theta)$ is based on $p_1(x|\theta)$.

# Example

Consider IID case with

$$p_1(x_i|\mu) = \frac{1}{\mu}\exp\left(-\frac{1}{\mu}x_i\right).$$

and $p(x|\mu) = \left(\frac{1}{\mu}\right)^n \exp\left(-\frac{1}{\mu}\sum_{i=1}^{n} x_i\right)$. The score function, which is the partial derivative of $\log f$ wrt the unknown parameter $\mu$, is

$$S(\mu) = \frac{\partial}{\partial\mu}\log f = -\frac{n}{\mu} + \frac{1}{\mu^2}\sum_{i=1}^{n} x_i.$$

For $X \sim Exp(1/\mu)$, we have $E_\mu(X) = \mu$ implying

$$I(\mu) = -E_\mu\left\{\frac{\partial}{\partial\mu}\left(-\frac{n}{\mu} + \frac{1}{\mu^2}\sum_{i=1}^{n} X_i\right)\right\} = -\frac{n}{\mu^2} + \frac{2}{\mu^3}E_\mu\left\{\sum_{i=1}^{n} X_i\right\} = -\frac{n}{\mu^2} + \frac{2n\mu}{\mu^3}$$

Hence $CRLB = \frac{\mu^2}{n}$.
Let us propose $\hat{\mu} = \bar{X}$ as an unbiased estimator of $\mu$. For $X \sim Exp(1/\mu)$, we have $\mathbb{E}(X) = \mu = \sqrt{\mathbb{V}ar(X)}$, implying the unbiased estimator $\hat{\mu} = \bar{x}$ achieves its CRLB

$$\mathbb{V}ar(\hat{\mu}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}ar(X_i) = \frac{n\mu^2}{n^2} = \frac{\mu^2}{n}.$$

# Efficiency

## Definition (Efficiency )

Define the efficiency of an unbiased estimator $\hat{\theta}$ as

$$\mathrm{eff}(\hat{\theta}) \;=\; \frac{\mathrm{CRLB}}{\mathbb{V}ar(\hat{\theta})}\,,$$

where $\mathrm{CRLB} = 1/I(\theta)$ . Clearly $0 < \mathrm{eff}(\hat{\theta}) \leq 1$. An unbiased estimator $\hat{\theta}$ is said to be efficient if $\mathrm{eff}(\hat{\theta}) = 1$. □

## Definition (Asymptotic efficiency )

The *asymptotic efficiency* of an unbiased estimator $\hat{\theta}$ is the limit of the efficiency as $n \to \infty$. An unbiased estimator $\hat{\theta}$ is said to be asymptotically efficient if its asymptotic efficiency is equal to 1. □

# Examples

The following are examples where the efficiency or asymptotical efficiency can be easily demonstrated

- the MLE $\hat{\theta} = r/n$ for the binomial distribution that was considered.
- the MLE for the Poisson distribution is 100% efficient.
- the MLE $\hat{\theta}$ for the exponential distribution with parameter $\theta$ is $\tilde{\theta}$ asymptotically efficient.
- the MLE estimator of variance in the normal distribution is asymptotically efficient.

# Mutliparameter extension

The information matrix is the covariance of the score vector $\mathbf{S} = \partial \log p(X|\theta)/\partial \boldsymbol{\theta}$ function

$$\mathbf{I}(\boldsymbol{\theta}) = Cov_{\boldsymbol{\theta}}(S) = \left[E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial \theta_i} \log p(X|\boldsymbol{\theta}) \frac{\partial}{\partial \theta_j} \log p(X|\boldsymbol{\theta})\right)\right]_{i,j=1}^{d}$$

**Theorem**

*Suppose that $\widehat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$ and $\mathbf{I}(\boldsymbol{\theta})$ is non-singular.*

$$Cov_{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}} - \mathbf{I}(\boldsymbol{\theta})^{-1}$$

*is a positive definite matrix.*

# Canonical exponential family

Consider the canonical form of the exponential family

$$p(x|\boldsymbol{\theta}) = \exp\left(\sum_{j=1}^{d} T_j(x)\theta_j - A(\boldsymbol{\theta})\right) h(x)$$

Then

$$\mathbf{I}(\boldsymbol{\theta}) = Cov_{\boldsymbol{\theta}}\mathbf{T}(X) = \ddot{A}(\boldsymbol{\theta}).$$

# Asymptotics of estimators

We have already defined asymptotic efficiency of estimators, where asymptotics is with respect to the sample size converging to infinty. There are other properties that are desirable.

- Asymptotically unbiased $Bias_{\theta,n} \to 0$.
- Consistency $\hat{\theta}_n \to \theta$, where convergence is either in probability or with probability one (strong consistency).
- Asymptotic normality when $a_n(\hat{\theta} - \theta)$ converges to a standard normal distribution, for properly chosen non-random $a_n$ (can be a matrix in the multiparameter case).
- The asymptotic theory will be discussed later in the course. Here we only consider the consistency condition.

# Consistency

## Definition

An estimator $\hat{\theta}_n$ of a parameter $\theta$ is called consistent if it is convergent in probability to $\theta$ as $n \to \infty$, i.e. for each $\theta \in \Theta$

$$\lim_{n\to\infty} P_\theta(|\hat{\theta}_n - \theta| > \epsilon) = 0.$$

A sample mean $\bar{X}_n$ is a consistent estimator of the mean $\mu = E_\theta X$ since by the Law of Large Number it is always

$$\bar{X}_n \to EX.$$

An estimator is called uniformly consistent if

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} P_\theta(|\hat{\theta}_n - \theta| > \epsilon) = 0.$$

# Consistency through Chebyshev's inequality

- Recall that for an unbiased estimator $\hat{\theta}_n$:

$$P_\theta(|\hat{\theta}_n - \theta|\epsilon) \leq Var_\theta(\hat{\theta}_n)/\epsilon^2.$$

- Thus if one shows that for each $\theta$ we have that $Var_\theta(\hat{\theta}_n)$ converges to zero, then it yields consistency. If $\sup_{\theta\in\Theta} Var_\theta(\hat{\theta}_n)$ converges to zero, then it yields uniform consistency.
- Any efficient estimator based on an IID sample is consistent since

$$Var_\theta(\hat{\theta}_n) = \frac{1}{nI_1(\theta)} \to 0.$$

- Uniform consistency does not need to hold.

# Consistency for the exponential families

**Theorem**

*The MLE for an exponential family is consistent.*

- One of the reasons why the MLE for the exponential families is a good estimator.
- There will be more reasons given later.