Department of Economics M. Jansson
University of California at Berkeley

Point Estimation

Recall that a probability space (sometimes called a *probability model*) is a triple $(\Omega, \mathcal{B}, P)$, where Ω is a sample space, $\mathcal{B}$ is a σ-algebra of events (subsets of Ω) and P is a probability function (defined on $\mathcal{B}$). In statistics we need to be able to study several probability functions simultaneously.

Definition. A *statistical experiment* (sometimes called a *statistical model*) is a triple $(\Omega, \mathcal{B}, \mathcal{P})$, where Ω is a sample space, $\mathcal{B}$ is a σ-algebra of events and $\mathcal{P}$ is a collection of probability functions defined on $\mathcal{B}$; that is, $\mathcal{P}$ is a collection of probability functions such that $(\Omega, \mathcal{B}, P)$ is a probability space for each $P \in \mathcal{P}$.

Definition. A statistical model $(\Omega, \mathcal{B}, \mathcal{P})$ is *parametric* if $\mathcal{P}$ is of the form $\left\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\right\}$, where θ is a *parameter*, which takes on values in the *parameter space* Θ, a subset of $\mathbb{R}^k$.

For concreteness, we will only consider parametric statistical models. Moreover, we will assume that the elementary outcomes are vectors of real numbers and that these outcomes are realizations of a collection of *i.i.d.* random variables. That is, each outcome is of the form

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

where each x_i $(i = 1, \ldots, n)$ is a realization of a random variable X_i and the random variables $X_1, \ldots, X_n$ are a random sample from some distribution with cdf $F(\cdot|\theta)$, where $\theta \in \Theta$ is unknown. Under these assumptions, each of the probability functions P_θ appearing in the definition of a parametric statistical model is uniquely determined by the corresponding cdf $F(\cdot|\theta)$. In other words, there is a one-to-one correspondence between the collection $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and the associated family $\mathcal{F} = \{F(\cdot|\theta) : \theta \in \Theta\}$ of marginal cdfs, so we can (and typically will) specify a statistical model in terms of the latter.

Example. Suppose $X_i \sim i.i.d.\ Ber(p)$, where $p \in [0,1]$ is an unknown parameter. In this case, $\theta = p$, $\Theta = [0,1] \subseteq \mathbb{R}$ and

$$F(x|p) = \begin{cases} 0 & \text{for } x < 0 \\ 1-p & \text{for } 0 \leq x < 1 \\ 1 & \text{for } x \geq 1. \end{cases}$$

Example. Suppose $X_i \sim i.i.d.\ U[0,\theta]$, where $\theta > 0$ is an unknown parameter. In this case, $\Theta = \mathbb{R}_{++} \subseteq \mathbb{R}$ and

$$F(x|\theta) = \begin{cases} 0 & \text{for } x < 0 \\ x/\theta & \text{for } 0 \leq x < \theta \\ 1 & \text{for } x \geq \theta. \end{cases}$$

Example. Suppose $X_i \sim i.i.d.\ \mathcal{N}\left(\mu, \sigma^2\right)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters. In this case, $\theta = \left(\mu, \sigma^2\right)'$, $\Theta = \mathbb{R} \times \mathbb{R}_{++} \subseteq \mathbb{R}^2$ and

$$F\left(x|\mu, \sigma^2\right) = \int_{-\infty}^{x} \phi\left(t|\mu, \sigma^2\right) dt, \qquad x \in \mathbb{R},$$

where

$$\phi\left(t|\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(t-\mu\right)^2\right), \qquad t \in \mathbb{R}.$$

Definition. Let $X_1, \ldots, X_n$ be a random sample from a distribution with cdf $F\left(\cdot|\theta\right)$, where θ is an unknown parameter. A *point estimator* is any statistic $W\left(X_1, \ldots, X_n\right)$.

At this level of generality, a point estimator is just a random variable. A realized value of a (point) estimator is called a *(point) estimate*. The quantity we are trying to estimate (typically θ) is called the *estimand*. It is not required that the range of the estimator coincides with the range of the estimand; that is, $W\left(X_1, \ldots, X_n\right) \in \Theta$ is not required. On the other hand, an estimator $W\left(X_1, \ldots, X_n\right)$ of θ is a good estimator (only) if it is "close" to θ in some probabilistic sense and this will typically require $W\left(X_1, \ldots, X_n\right) \in \Theta$.

Casella and Berger (Sections 7.2.1-7.2.3) discuss three methods that can be used to generate estimators under quite general circumstances. We will cover two of these methods, the method of moments and the maximum likelihood procedure. Method of moments estimators are obtained by solving a system of equations, while maximum likelihood estimators are constructed by solving a maximization problem.

Suppose $X_1, \ldots, X_n$ is a random sample from a distribution with cdf $F\left(\cdot|\theta\right)$, where $\theta \in \Theta$ is an unknown scalar parameter. Let $\mu : \Theta \to \mathbb{R}$ be the function defined by

$$\mu\left(\theta\right) = \int_{-\infty}^{\infty} x dF\left(x|\theta\right), \qquad \theta \in \Theta.$$

As defined, $\mu\left(\theta\right)$ is the expected value of a random variable with cdf $F\left(\cdot|\theta\right)$. The true parameter value θ solves the equation

$$E\left(X\right) = \mu\left(\theta\right),$$

where X is a random variable with the same (marginal) distribution as X_i $\left(i = 1, \ldots, n\right)$. A *method of moments* estimator $\hat{\theta}$ solves the sample analogue of this equation, viz.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\left(\hat{\theta}\right).$$

To the extent that $\bar{X}$ is a good estimator of $E\left(X\right)$ (it turns out that it often is), one would expect $\hat{\theta}$ to be a good estimator of θ.

Example. Suppose $X_i \sim i.i.d.\ Ber\left(p\right)$, where $p \in [0,1]$ is an unknown parameter. In this case, $\theta = p$, $\Theta = [0,1]$ and

$$\mu\left(p\right) = p, \qquad 0 \leq p \leq 1.$$

Therefore, the method of moments estimator of p is

$$\hat{p} = \bar{X}.$$

Example. Suppose $X_i \sim i.i.d.\ U[0, \theta]$, where $\theta > 0$ is an unknown parameter. We have:

$$\mu(\theta) = \frac{\theta}{2}, \qquad \theta > 0.$$

The method of moments estimator $\hat{\theta}$ is found by solving the equation $\mu\left(\hat{\theta}\right) = \bar{X}$:

$$\mu\left(\hat{\theta}\right) = \frac{\hat{\theta}}{2} = \bar{X} \qquad \Leftrightarrow \qquad \hat{\theta} = 2\bar{X}.$$

As it turns out, this method of moments estimator is not a terribly good estimator. Notice that even though θ is unknown, we do know that $X_i > \theta$ is impossible when $X_i \sim U[0, \theta]$. It is possible to have $X_i > \hat{\theta}$ for some i (e.g. if $n = 3$ and $X_1 = X_2 = 1$ and $X_3 = 7$), so it seems plausible that a better estimator can be constructed.

In the case of a scalar parameter θ, the method of moments estimator is constructed by solving one (moment) equation in one unknown parameter. When θ is a k-dimensional parameter vector, $\theta = (\theta_1, \ldots, \theta_k)'$, the method of moments estimator of θ is constructed by solving k equations in the k unknown parameters $\theta_1, \ldots, \theta_k$.

Definition. Let $X_1, \ldots, X_n$ be a random sample from a distribution with cdf $F(\cdot|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$ is a vector of unknown parameters. For any $j = 1, \ldots, k$, let $\mu_j : \Theta \to \mathbb{R}$ be defined by

$$\mu_j(\theta) = \int_{-\infty}^{\infty} x^j dF(x|\theta), \qquad \theta \in \Theta.$$

A *method of moments* estimator $\hat{\theta}$ of θ solves the *estimating equations*

$$\frac{1}{n}\sum_{i=1}^{n} X_i^j = \mu_j\left(\hat{\theta}\right), \qquad j = 1, \ldots, k.$$

Example. Suppose $X_i \sim i.i.d.\ \mathcal{N}\left(\mu, \sigma^2\right)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters. In this case, $\theta = \left(\mu, \sigma^2\right)'$, $\Theta = \mathbb{R} \times \mathbb{R}_{++} \subseteq \mathbb{R}^2$ and the functions $\mu_1(\cdot)$ and $\mu_2(\cdot)$ are given by

$$\mu_1\left(\mu, \sigma^2\right) = \mu,$$
$$\mu_2\left(\mu, \sigma^2\right) = \sigma^2 + \mu^2.$$

Any solution $\left(\hat{\mu}, \hat{\sigma}^2\right)'$ to the equation

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mu_1\left(\hat{\mu}, \hat{\sigma}^2\right) = \hat{\mu}$$

satisfies

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}.$$

Using this relation, the equation

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mu_2\left(\hat{\mu}, \hat{\sigma}^2\right) = \hat{\sigma}^2 + \hat{\mu}^2$$

can solved for $\hat{\sigma}^2$ to yield

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

In all of the examples considered so far, there is a unique solution $\hat{\theta} \in \Theta$ to the system

$$\frac{1}{n}\sum_{i=1}^{n} X_i^j = \int_{-\infty}^{\infty} x^j dF\left(x|\hat{\theta}\right), \qquad j = 1, \ldots, k,$$

of estimating equations. It is not difficult to construct examples where the method of moments breaks down. In such cases, some variant of the method of moments may work.

Example. Suppose $X_i \sim i.i.d.\ \mathcal{N}\left(0, \sigma^2\right)$, where $\sigma^2 > 0$ is an unknown parameter. In this case, $\theta = \sigma^2$, $\Theta = \mathbb{R}_{++} \subseteq \mathbb{R}$ and the function $\mu_1'(\cdot)$ is given by

$$\mu_1\left(\sigma^2\right) = 0.$$

The equation

$$\bar{X} = \mu_1\left(\hat{\sigma}^2\right) = 0$$

has infinitely many solutions when $\bar{X} = 0$ and no solutions when $\bar{X} \neq 0$.

In contrast, the sample counterpart of the equation

$$E\left(X^2\right) = \mu_2\left(\sigma^2\right) = \sigma^2$$

has a unique solution:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Generalizing this example, let $X_1, \ldots, X_n$ be a random sample from a distribution with cdf $F(\cdot|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$ is a vector of unknown parameters. Even if we cannot find a unique solution $\hat{\theta}$ to the estimating equations

$$\frac{1}{n}\sum_{i=1}^{n} X_i^j = \int_{-\infty}^{\infty} x^j dF\left(x|\hat{\theta}\right), \qquad j = 1, \ldots, k,$$

we may be able to find functions $g_j : \mathbb{R} \to \mathbb{R}$ $(j = 1, \ldots, k)$ such that the system of equations

$$\frac{1}{n}\sum_{i=1}^{n} g_j\left(X_i\right) = \int_{-\infty}^{\infty} g_j\left(x\right) dF\left(x|\hat{\theta}\right), \qquad j = 1, \ldots, k,$$

has a unique solution $\hat{\theta} \in \Theta$. Estimators $\hat{\theta}$ constructed in this way are also called *method of moments* estimators.

Definition. Let $X = (X_1, \ldots, X_n)'$ be a discrete (continuous) n-dimensional random vector with joint pmf (pdf) $f_X(\cdot|\theta) : \mathbb{R}^n \to \mathbb{R}_+$, where $\theta \in \Theta$ is an unknown parameter vector. For any $x = (x_1, \ldots, x_n)'$, the *likelihood function* given x is the function $L(\cdot|x) : \Theta \to \mathbb{R}_+$ given by

$$L(\theta|x) = L(\theta|x_1, \ldots, x_n) = f_X(x|\theta), \qquad \theta \in \Theta.$$

The *log likelihood function* given x is the function $l(\cdot|x) : \Theta \to [-\infty, \infty)$ given by

$$l(\theta|x) = l(\theta|x_1, \ldots, x_n) = \log L(\theta|x), \qquad \theta \in \Theta.$$

When $X_1, \ldots, X_n$ is a random sample from a discrete (continuous) distribution with pmf (pdf) $f(\cdot|\theta)$, the likelihood function given $x = (x_1, \ldots, x_n)'$ is

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta), \qquad \theta \in \Theta,$$

while the log likelihood function given x is

$$l(\theta|x) = \sum_{i=1}^{n} \log f(x_i|\theta), \qquad \theta \in \Theta.$$

Definition. Let $X_1, \ldots, X_n$ be a random sample from a discrete (continuous) distribution with pmf (pdf) $f(\cdot|\theta)$, where $\theta \in \Theta$ is an unknown parameter vector. When $X = (X_1, \ldots, X_n)' = x$, a *maximum likelihood estimate* $\hat{\theta}(x)$ of θ satisfies

$$L\left(\hat{\theta}(x)|x\right) = \max_{\theta \in \Theta} L(\theta|x),$$

where $L(\cdot|x)$ is the likelihood function given x. The estimator $\hat{\theta}(X)$ is a *maximum likelihood estimator (MLE)* of θ.

Maximum likelihood estimators often enjoy favorable large sample properties. That result is related to the following fact, which in itself can be used to motivate the maximum likelihood estimator.

Theorem (Information Inequality; Ruud, Lemma D.2). *Let X be a discrete (continuous) random variable with pmf (pdf) f_0 and let f_1 be any other pmf (pdf). Then*

$$E\left(\log f_0\left(X\right)\right) \geq E\left(\log\left(f_1\left(X\right)\right)\right).$$

Remark. The information inequality is strict unless $P\left(f_0\left(X\right) = f_1\left(X\right)\right) = 1$.

Proof. The claim is that $E\left(\log\left(Y\right)\right) \leq 0$, where

$$Y = \begin{cases} f_1\left(X\right)/f_0\left(X\right) & \text{for } X \in \mathcal{X} \\ 0 & \text{for } X \notin \mathcal{X}. \end{cases},$$

where $\mathcal{X} = \{x : f_0\left(x\right) > 0\}$ is the *support* of X.

Recall the following implication of Jensen's inequality: If Y is a random variable with $P\left(Y \geq 0\right) = 1$, then

$$E\left(\log\left(Y\right)\right) \leq \log\left(E\left(Y\right)\right).$$

Now,

$$E\left(Y\right) = \sum_{x \in \mathcal{X}} \frac{f_1\left(x\right)}{f_0\left(x\right)} \cdot f_0\left(x\right) = \sum_{x \in \mathcal{X}} f_1\left(x\right) \leq \sum_{x \in \mathbb{R}} f_1\left(x\right) = 1$$

if X is discrete, while

$$E\left(Y\right) = \int_{\mathcal{X}} \frac{f_1\left(x\right)}{f_0\left(x\right)} \cdot f_0\left(x\right) dx = \int_{\mathcal{X}} f_1\left(x\right) dx \leq \int_{-\infty}^{\infty} f_1\left(x\right) dx = 1$$

if X is continuous. In both cases, $E\left(Y\right) \leq 1$ and it follows from Jensen's inequality that

$$E\left(\log\left(Y\right)\right) \leq \log\left(E\left(Y\right)\right) \leq \log\left(1\right) = 0,$$

as was to be shown. ■

Let $X_1, \ldots, X_n$ be a random sample from a discrete (continuous) distribution with pmf (pdf) $f\left(\cdot|\theta\right)$, where $\theta \in \Theta$ is unknown. It follows from the information inequality that

$$E_\theta\left(\log f\left(X|\theta\right)\right) \geq E_\theta\left(\log f\left(X|\theta^*\right)\right)$$

for any $\theta^* \in \Theta$, where $E_\theta\left(\cdot\right)$ denotes the expected value computed using the true (unknown) cdf $F\left(\cdot|\theta\right)$ of the random variable X. As a consequence, the true parameter value θ solves the problem of maximizing

$$E_\theta\left(\log f\left(X|\theta^*\right)\right)$$

with respect to $\theta^* \in \Theta$; that is,

$$E_\theta\left(\log f\left(X|\theta\right)\right) = \max_{\theta^* \in \Theta} E_\theta\left(\log f\left(X|\theta^*\right)\right).$$

The sample analogue of this problem is that of maximizing the *average log likelihood*

$$\frac{1}{n}\sum_{i=1}^{n} \log f\left(X_i|\theta^*\right)$$

with respect $\theta^* \in \Theta$. The average log likelihood is a strictly increasing function of $L\left(\theta^*|X_1,\ldots,X_n\right)$. Specifically,

$$\frac{1}{n}\sum_{i=1}^{n} \log f\left(X_i|\theta^*\right) = \log\left(\frac{1}{n}L\left(\theta^*|X_1,\ldots,X_n\right)\right).$$

Therefore, a maximum likelihood estimator $\hat{\theta}\left(X_1,\ldots,X_n\right)$ maximizes the average log likelihood with respect to θ^* :

$$\frac{1}{n}\sum_{i=1}^{n} \log f\left(X_i|\hat{\theta}\left(X_1,\ldots,X_n\right)\right) = \max_{\theta^* \in \Theta}\left(\frac{1}{n}\sum_{i=1}^{n} \log f\left(X_i|\theta^*\right)\right).$$

Example. Suppose $X_i \sim i.i.d.\ Ber\left(p\right)$, where $p \in [0,1]$ is an unknown parameter. Each X_i is discrete with pmf

$$\begin{aligned} f\left(x|p\right) &= \begin{cases} 1-p & \text{for } x=0 \\ p & \text{for } x=1 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} p^x\left(1-p\right)^{1-x} & \text{for } x \in \{0,1\} \\ 0 & \text{otherwise,} \end{cases} \\ &= p^x\left(1-p\right)^{1-x} \cdot 1\left(x \in \{0,1\}\right), \end{aligned}$$

where $0^0 = 1$ and $1\left(\cdot\right)$ is the indicator function. It suffices to consider the case where $x_i \in \{0,1\}$ for $i = 1,\ldots,n$, as the likelihood is zero for all other values of $x = \left(x_1,\ldots,x_n\right)'$.

The likelihood given x is

$$\begin{aligned} L\left(p|x\right) &= \prod_{i=1}^{n} f\left(x_i|p\right) = \prod_{i=1}^{n} p^{x_i}\left(1-p\right)^{1-x_i} \\ &= p^{\sum_{i=1}^{n} x_i}\left(1-p\right)^{n-\sum_{i=1}^{n} x_i}, \qquad p \in [0,1], \end{aligned}$$

while the log likelihood given x is

$$\begin{aligned} l(p|x) &= \sum_{i=1}^{n} \log f(x_i|p) = \sum_{i=1}^{n} (x_i \log p + (1-x_i)\log(1-p)) \\ &= \left(\sum_{i=1}^{n} x_i\right)\log p + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-p), \qquad p \in [0,1], \end{aligned}$$

where $0 \cdot \log 0 = 0$.

If $\sum_{i=1}^{n} x_i = 0$, then

$$L(p|x) = (1-p)^n$$

is a decreasing function of p and $p = 0$ maximizes $L(p|x)$ with respect to $p \in [0,1]$.

If $\sum_{i=1}^{n} x_i = n$, then

$$L(p|x) = p^n$$

and $p = 1$ maximizes $L(p|x)$ with respect to $p \in [0,1]$.

In intermediate cases where $0 < \sum_{i=1}^{n} x_i < n$, the maximum likelihood estimate can be found by solving the first-order condition for an interior maximum:

$$\begin{aligned} \frac{d}{dp} l(p|x)\Big|_{p=\hat{p}} &= \left(\sum_{i=1}^{n} x_i\right)\frac{1}{\hat{p}} - \left(n - \sum_{i=1}^{n} x_i\right)\frac{1}{1-\hat{p}} = 0 \\ &\Updownarrow \\ \hat{p} &= \frac{\sum_{i=1}^{n} x_i}{n}. \end{aligned}$$

This unique solution to the first-order condition is a maximizer because

$$\frac{d^2}{dp^2} l(p|x)\Big|_{p=\hat{p}} = -\left(\sum_{i=1}^{n} x_i\right)\frac{1}{\hat{p}^2} - \left(n - \sum_{i=1}^{n} x_i\right)\frac{1}{(1-\hat{p})^2} < 0.$$

Combining the results, we see that

$$\hat{p} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

is the maximum likelihood estimator of p. In this case, the maximum likelihood estimator coincides with the method of moments estimator.

Exercise. Verify that $p^* = p$ maximizes

$$\begin{aligned} E_p\left(\log f\left(X_i|p^*\right)\right) &= E_p\left(X_i \log p^* + (1-X_i)\log\left(1-p^*\right)\right) \\ &= p\cdot\log p^* + (1-p)\cdot\log\left(1-p^*\right) \end{aligned}$$

with respect to $p^* \in [0,1]$.

When a unique maximum likelihood estimator $\hat{\theta}$ of $\theta = (\theta_1, \ldots, \theta_k)'$ exists, it can usually be constructed by solving the *likelihood equations*

$$\frac{\partial}{\partial \theta_j} l\left(\theta|X_1, \ldots, X_n\right)\Big|_{\theta=\hat{\theta}} = 0, \qquad j = 1, \ldots, k,$$

and verifying that a second-order condition holds. For instance, if θ is a scalar parameter a unique solution $\hat{\theta}$ to

$$\frac{d}{d\theta} l\left(\theta|x_1, \ldots, x_n\right)\Big|_{\theta=\hat{\theta}} = 0$$

is a maximum likelihood estimate if $l\left(\cdot|x_1, \ldots, x_n\right)$ is twice differentiable, Θ is an interval (possibly unbounded) and

$$\frac{d^2}{d\theta^2} l\left(\theta|x_1, \ldots, x_n\right)\Big|_{\theta=\hat{\theta}} < 0.$$

Example. Suppose $X_i \sim$ *i.i.d.* $\mathcal{N}\left(\mu, \sigma^2\right)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters. The marginal pdf of X_i is $f\left(\cdot|\mu, \sigma^2\right)$, where

$$\begin{aligned} f\left(x|\mu, \sigma^2\right) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right) \\ &= \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right). \end{aligned}$$

The likelihood given $x = (x_1, \ldots, x_n)'$ is

$$\begin{aligned} L\left(\mu, \sigma^2|x\right) &= \prod_{i=1}^{n} f\left(x_i|\mu, \sigma^2\right) = \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}\left(x_i-\mu\right)^2\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right), \end{aligned}$$

while the log likelihood given x is

$$\begin{aligned} l\left(\mu,\sigma^2|x\right) &= \sum_{i=1}^{n}\log f\left(x_i|\mu,\sigma^2\right) \\ &= \sum_{i=1}^{n}\left(-\frac{1}{2}\log\left(2\pi\right)-\frac{1}{2}\log\left(\sigma^2\right)-\frac{1}{2\sigma^2}\left(x_i-\mu\right)^2\right) \\ &= -\frac{n}{2}\log 2\pi-\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2. \end{aligned}$$

The likelihood equations are:

$$\frac{\partial}{\partial\mu}l\left(\mu,\sigma^2|X_1,\ldots,X_n\right)\Big|_{\theta=\hat{\theta}}=\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}\left(X_i-\hat{\mu}\right)=0$$

and

$$\frac{\partial}{\partial\hat{\sigma}^2}l\left(\mu,\sigma^2|X_1,\ldots,X_n\right)\Big|_{\theta=\hat{\theta}}=-\frac{n}{2\hat{\sigma}^2}+\frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}\left(X_i-\hat{\mu}\right)^2,$$

where $\theta=\left(\mu,\sigma^2\right)'$ and $\hat{\theta}=\left(\hat{\mu},\hat{\sigma}^2\right)'$. The unique solution to these equations is

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X}, \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\hat{\mu}\right)^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2. \end{aligned}$$

The matrix

$$\frac{\partial^2}{\partial\theta\partial\theta'}l\left(\mu,\sigma^2|x\right)\Big|_{\theta=\hat{\theta}}=\left(\begin{array}{cc}\frac{\partial^2}{\partial\mu\partial\mu}l\left(\mu,\sigma^2|x\right) & \frac{\partial^2}{\partial\mu\partial\sigma^2}l\left(\mu,\sigma^2|x\right)\\ \frac{\partial^2}{\partial\sigma^2\partial\mu}l\left(\mu,\sigma^2|x\right) & \frac{\partial^2}{\partial\sigma^2\partial\sigma^2}l\left(\mu,\sigma^2|x\right)\end{array}\right)\Big|_{\theta=\hat{\theta}}=\left(\begin{array}{cc}-\frac{n}{\hat{\sigma}^2} & 0\\ 0 & -\frac{n}{2\hat{\sigma}^4}\end{array}\right)$$

is negative definite, so $\left(\hat{\mu},\hat{\sigma}^2\right)'$ is a local maximizer of $l\left(\mu,\sigma^2|x\right)$. In fact,

$$\lim_{|\mu|\to\infty}L\left(\mu,\sigma^2|x\right)=0$$

for any $\sigma^2>0$ and

$$\lim_{\sigma^2\to 0}L\left(\mu,\sigma^2|x\right)=\lim_{\sigma^2\to\infty}L\left(\mu,\sigma^2|x\right)=0$$

for any $\mu\in\mathbb{R}$, so $\left(\hat{\mu},\hat{\sigma}^2\right)'$ is the maximum likelihood estimator of $\left(\mu,\sigma^2\right)'$. Once again, the maximum likelihood estimator coincides with the method of moments estimator.

In the present case, the second-order condition can also be verified using univariate calculus (Casella and Berger, Example 7.2.11).

Exercise. Verify that $\left(\mu^*, \sigma^{*2}\right) = \left(\mu, \sigma^2\right)$ maximizes

$$\begin{aligned} & E_{(\mu,\sigma^2)}\left(\log f\left(X_i|\mu^*, \sigma^{*2}\right)\right) \\ = \; & E_{(\mu,\sigma^2)}\left(-\frac{1}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left(\sigma^{*2}\right) - \frac{1}{2\sigma^{*2}}\left(X_i - \mu^*\right)^2\right) \\ = \; & -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^{*2} - \frac{1}{2\sigma^{*2}}\left(\sigma^2 + \left(\mu - \mu^*\right)^2\right) \end{aligned}$$

with respect to $\left(\mu^*, \sigma^{*2}\right) \in \mathbb{R} \times \mathbb{R}_{++}$.

One case where the maximum likelihood estimator cannot be constructed by solving the likelihood equations is the following.

Example. Suppose $X_i \sim$ *i.i.d.* $U\left[0, \theta\right]$, where $\theta > 0$ is an unknown parameter. Each X_i is continuous with pdf

$$\begin{aligned} f\left(x|\theta\right) &= \begin{cases} 1/\theta & \text{for } 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases} \\ &= \frac{1}{\theta}1\left(0 \leq x \leq \theta\right). \end{aligned}$$

It suffices to consider the case where $x_i \geq 0$ for $i = 1, \ldots, n$, as the likelihood is zero for all other values of $x = \left(x_1, \ldots, x_n\right)'$.

The likelihood given x is

$$\begin{aligned} L\left(\theta|x\right) &= \prod_{i=1}^{n} f\left(x_i|\theta\right) = \prod_{i=1}^{n}\left(\frac{1}{\theta}1\left(0 \leq x_i \leq \theta\right)\right) \\ &= \frac{1}{\theta^n}\prod_{i=1}^{n} 1\left(0 \leq x_i \leq \theta\right) \\ &= \frac{1}{\theta^n}1\left(\max_{1\leq i\leq n} x_i \leq \theta\right), \qquad \theta > 0, \end{aligned}$$

where the third equality uses that fact that $x_i \geq 0$ for $i = 1, \ldots, n$.

The likelihood given x is zero for $\theta < \max_{1\leq i\leq n} x_i$ and is a decreasing function of θ for $\theta \geq \max_{1\leq i\leq n} x_i$. As a consequence, the maximum likelihood estimator of θ is

$$\hat{\theta} = \max_{1\leq i\leq n} X_i.$$

In this case, the maximum likelihood estimator is different from the method moments estimator (the latter is $2 \cdot \bar{X}$). Unlike the method of moments estimator, the maximum likelihood estimator has the property

that $\hat{\theta} \geq X_i$ for every i. On the other hand, since $\max_{1 \leq i \leq n} X_i$ is a lower bound on the true θ, $\hat{\theta}$ will tend to underestimate θ. Indeed, $P\left(\hat{\theta} \leq \theta\right) = 1$.

Exercise. Verify that $\theta^* = \theta$ maximizes

$$E_\theta \left(\log f\left(X_i|\theta^*\right)\right) = -\log \theta^* - \infty \cdot P_\theta \left(X_i > \theta^*\right)$$

with respect to $\theta^* > 0$, where $\infty \cdot 0 = 0$.

In the sense of the following definition, the maximum likelihood estimator $\max_{1 \leq i \leq n} X_i$ in the preceding example is the nth order statistic and is often denoted by $X_{(n)}$.

Definition. Let $X_1, \ldots, X_n$ be a random sample. The *order statistics* are the sample values placed in ascending order. They are denoted by $X_{(1)}, \ldots, X_{(n)}$.

Example. For any random sample $X_1, \ldots, X_n$, $X_{(1)} = \min_{1 \leq i \leq n} X_i$ and $X_{(n)} = \max_{1 \leq i \leq n} X_n$.

Remark. The pmf (pdf) of order statistics obtained from a discrete (continuous) distribution can be characterized using combinatorial arguments (Casella and Berger, Section 5.4).

Remark. Maximum likelihood estimators are *equivariant* in the sense that if $\hat{\theta}$ is a maximum likelihood estimator of θ, then $\tau\left(\hat{\theta}\right)$ is a maximum likelihood estimator of $\tau(\theta)$ for any function $\tau(\cdot)$ defined on Θ (Casella and Berger, Theorem 7.2.10).

Let $X_1, \ldots, X_n$ be a random sample from a distribution with cdf $F\left(\cdot|\theta\right)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$ is an unknown parameter vector. An estimator $\hat{\theta}$ of θ is called a *Z-estimator* if it is (implicitly) defined as a solution of a system of equations of the form

$$\frac{1}{n} \sum_{i=1}^{n} \psi_j \left(X_i, \hat{\theta}\right) = 0, \qquad j = 1, \ldots, k,$$

where each $\psi_j : \mathbb{R} \times \Theta \to \mathbb{R}$ is a function. Z-estimators are usually motivated by showing that θ is the only value of $\theta^* \in \Theta$ for which

$$E_\theta \psi_j \left(X, \theta^*\right) = 0 \qquad \forall j \in \{1, \ldots, k\}.$$

The leading special case is the method of moments estimator, which is a Z-estimator with

$$\psi_j \left(X_i, \theta^*\right) = X_i^j - \int_{-\infty}^{\infty} x^j dF\left(x|\theta^*\right), \qquad j = 1, \ldots, k,$$

or, more generally,

$$\psi_j \left(X_i, \theta^*\right) = g_j\left(X_i\right) - \int_{-\infty}^{\infty} g_j\left(x\right) dF\left(x|\theta^*\right), \qquad j = 1, \ldots, k.$$

An estimator $\hat{\theta}$ of θ is called an *M-estimator* if it is (implicitly) defined as

$$\frac{1}{n}\sum_{i=1}^{n} m\left(X_i, \hat{\theta}\right) = \max_{\theta^* \in \Theta} \frac{1}{n}\sum_{i=1}^{n} m\left(X_i, \theta^*\right),$$

where $m : \mathbb{R} \times \Theta \to \mathbb{R}$ is a function. M-estimators are usually motivated by showing that θ is the unique maximizer (with respect to $\theta^* \in \Theta$) of

$$E_\theta m\left(X, \theta^*\right).$$

The leading special case is the maximum likelihood estimator, which is an M-estimator with

$$m\left(X_i, \theta^*\right) = \log f\left(X_i | \theta^*\right),$$

where $f\left(\cdot|\theta^*\right)$ is the pmf/pdf of the cdf $F\left(\cdot|\theta^*\right)$.

Many M-estimators satisfy first-order conditions of the form

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta_j} m\left(X_i, \theta\right)\Big|_{\theta=\hat{\theta}}\right) = 0, \qquad j = 1, \ldots, k.$$

and can therefore be interpreted as Z-estimators with

$$\psi_j\left(X_i, \theta^*\right) = \frac{\partial}{\partial\theta_j} m\left(X_i, \theta\right)\Big|_{\theta=\theta^*}, \qquad j = 1, \ldots, k.$$

In particular, many maximum likelihood estimators can be interpreted as method of moments estimators with

$$g_j\left(X_i, \theta^*\right) = \frac{\partial}{\partial\theta_j} \log f\left(X_i, \theta\right)\Big|_{\theta=\theta^*}, \qquad j = 1, \ldots, k.$$

It will almost always be possible to find a set of functions $\left\{\psi_j : j = 1, \ldots, k\right\}$ such that θ is the only value of $\theta^* \in \Theta$ for which

$$E_\theta \psi_j\left(X, \theta^*\right) = 0 \qquad \forall j \in \{1, \ldots, k\}$$

or a function m such that θ is the unique maximizer (with respect to $\theta^* \in \Theta$) of

$$E_\theta m\left(X, \theta^*\right).$$

An important exception occurs when the model is not identified.

Definition. Let $(\Omega, \mathcal{B}, \{P_\theta : \theta \in \Theta\})$ be a parametric statistical model. A parameter value $\theta_1 \in \Theta$ is *identified* if there does not exist another parameter value $\theta_2 \in \Theta$ such that $P_{\theta_1} = P_{\theta_2}$. The model $(\Omega, \mathcal{B}, \{P_\theta : \theta \in \Theta\})$ is *identified* if every parameter value $\theta \in \Theta$ is identified.

In other words, a model is identified if knowledge of the true marginal cdf $F\left(\cdot|\theta\right)$ implies knowledge of the parameter θ. This is a very modest and reasonable requirement. Identification is a property of

the parameterization/specification of a statistical model. When the cdfs $\{F(\cdot|\theta) : \theta \in \Theta\}$ characterizing a statistical model are specified directly, identification usually holds. On the other hand, problems may arise when the observed sample is assumed to be generated by a transformation model.

Example. Suppose $X_1, \ldots, X_n$ is a random sample generated by the model

$$X_i = 1(X_i^* > 0) = \begin{cases} 0 & \text{for } X_i^* \leq 0 \\ 1 & \text{for } X_i^* > 0 \end{cases}, \qquad X_i^* \sim i.i.d. \mathcal{N}(\mu, \sigma^2),$$

where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters. In this case, $X_i \sim i.i.d.\ Ber(\Phi(-\mu/\sigma))$, where $\Phi(\cdot)$ is the cdf of the standard normal distribution:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt.$$

As a consequence, the marginal distribution of each X_i depends on $(\mu, \sigma^2)'$ only through μ/σ and any parameter value $(\mu_1, \sigma_1^2)'$ is unidentified. We can achieve identification by imposing an *identifying assumption* on the parameters. In this case, a natural identifying assumption is $\sigma = 1$.

Remark. If $\theta_1 \in \Theta$ is an unidentified parameter value, there is another parameter value $\theta_2 \in \Theta$ such that $F(\cdot|\theta_1) = F(\cdot|\theta_2)$, implying

$$E_{\theta_1} g(X) = \int_{-\infty}^{\infty} g(x)\, dF(x|\theta_1) = \int_{-\infty}^{\infty} g(x)\, dF(x|\theta_2) = E_{\theta_2} g(X)$$

for any function g. In particular, any solution θ^* to a system of equations of the form

$$E_{\theta_1} \psi_j(X, \theta^*) = 0 \qquad \forall j \in \{1, \ldots, k\}$$

will also be a solution to the following system of equations:

$$E_{\theta_2} \psi_j(X, \theta^*) = 0 \qquad \forall j \in \{1, \ldots, k\}.$$

Similarly, any maximizer (with respect to $\theta^* \in \Theta$) of

$$E_{\theta_1} m(X, \theta^*)$$

will also be a maximizer of

$$E_{\theta_2} m(X, \theta^*).$$

If one attempts to estimate the parameters of an unidentified model, unique method of moments (maximum likelihood) estimators typically cannot be found.

An estimator $W(X_1, \ldots, X_n)$ of θ is a good estimator (only) if it is "close" to θ in some probabilistic sense. We will use mean squared error as our measure of closeness.

Definition. The *mean squared error (MSE)* matrix of an estimator $\hat{\theta}$ of θ is the function (of θ) given by

$$MSE_\theta\left(\hat{\theta}\right) = E_\theta\left[\left(\hat{\theta}-\theta\right)\left(\hat{\theta}-\theta\right)'\right], \qquad \theta \in \Theta.$$

Definition. The *bias* of an estimator $\hat{\theta}$ of θ is

$$Bias_\theta\left(\hat{\theta}\right) = E_\theta\left(\hat{\theta}\right) - \theta, \qquad \theta \in \Theta.$$

An estimator $\hat{\theta}$ of θ is *unbiased* if

$$E_\theta\left(\hat{\theta}\right) = \theta \qquad \forall \theta \in \Theta.$$

Many results derived using MSE generalize to other measures of closeness. It is convenient to use MSE because it is analytically tractable and has a straightforward interpretation in terms of the variance and bias of the estimator $\hat{\theta}$. Specifically,

$$\begin{aligned}
MSE_\theta\left(\hat{\theta}\right) &= E_\theta\left[\left(\hat{\theta}-\theta\right)\left(\hat{\theta}-\theta\right)'\right] \\
&= E_\theta\left[\left(\hat{\theta}-E_\theta\left(\hat{\theta}\right)+E_\theta\left(\hat{\theta}\right)-\theta\right)\left(\hat{\theta}-E_\theta\left(\hat{\theta}\right)+E_\theta\left(\hat{\theta}\right)-\theta\right)'\right] \\
&= E_\theta\left[\left(\hat{\theta}-E_\theta\left(\hat{\theta}\right)\right)\left(\hat{\theta}-E_\theta\left(\hat{\theta}\right)\right)'\right] \\
&\quad + \left(E_\theta\left(\hat{\theta}\right)-\theta\right)\left(E_\theta\left(\hat{\theta}\right)-\theta\right)' \\
&\quad + E_\theta\left[\hat{\theta}-E_\theta\left(\hat{\theta}\right)\right]\left(E_\theta\left(\hat{\theta}\right)-\theta\right)' \\
&\quad + \left(E_\theta\left(\hat{\theta}\right)-\theta\right)E_\theta\left[\hat{\theta}-E_\theta\left(\hat{\theta}\right)\right]' \\
&= Var_\theta\left(\hat{\theta}\right) + Bias_\theta\left(\hat{\theta}\right)\cdot Bias_\theta\left(\hat{\theta}\right)'
\end{aligned}$$

because $E_\theta\left[\hat{\theta}-E_\theta\left(\hat{\theta}\right)\right]=0$ and $E_\theta\left(\hat{\theta}\right)-\theta = Bias_\theta\left(\hat{\theta}\right)$ is non-random. In particular,

$$MSE_\theta\left(\hat{\theta}\right) = Var_\theta\left(\hat{\theta}\right) + Bias_\theta\left(\hat{\theta}\right)^2$$

when θ is a scalar parameter.

Example. Suppose $X_i \sim$ *i.i.d.* $\mathcal{N}\left(\mu,\sigma^2\right)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters. Two estimators of σ^2 are the maximum likelihood estimator

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

and the sample variance

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

The sample variance satisfies

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

implying

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

and

$$Var\left(S^2\right) = \left(\frac{\sigma^2}{n-1}\right)^2 \cdot Var\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right) = \left(\frac{\sigma^2}{n-1}\right)^2 \cdot 2(n-1) = \frac{2}{n-1}\sigma^4.$$

In particular,

$$MSE_{(\mu,\sigma^2)}\left(S^2\right) = \frac{2}{n-1}\sigma^4.$$

Similarly,

$$E\left(\hat{\sigma}^2\right) = \frac{\sigma^2}{n}\cdot E\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right) = \frac{\sigma^2}{n}\cdot(n-1) = \frac{n-1}{n}\sigma^2$$

and

$$Var\left(\hat{\sigma}^2\right) = \left(\frac{\sigma^2}{n}\right)^2 \cdot Var\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right) = \left(\frac{\sigma^2}{n}\right)^2 \cdot 2(n-1) = \frac{2(n-1)}{n^2}\sigma^4,$$

so

$$\begin{aligned} MSE_{(\mu,\sigma^2)}\left(\hat{\sigma}^2\right) &= \frac{2(n-1)}{n^2}\sigma^4 + \left(\frac{1}{n}\sigma^2\right)^2 = \frac{2n-1}{n^2}\sigma^4 \\ &= \frac{2-1/n}{n}\sigma^4 < \frac{2}{n-1}\sigma^4 = MSE_{(\mu,\sigma^2)}\left(S^2\right). \end{aligned}$$

Unlike S^2, $\hat{\sigma}^2$ is biased. Nonetheless, its variance is so much smaller than that of S^2 that its MSE is smaller for all values of μ and σ^2.

In this example, the MSE ranking does not depend on the true value of the parameter(s). In spite of this we do not know whether an even better estimator exists. To answer that question, it might appear natural to look for an estimator that minimizes MSE uniformly in μ and σ^2. Unfortunately, such an estimator does not exist.

Example. As a competitor to $\hat{\sigma}^2$, consider the estimator $\tilde{\sigma}^2 = 1$. Evidently, $\tilde{\sigma}^2$ is a perfect estimator if σ^2 happens to equal unity, but is an inferior estimator for most other values of σ^2.

The point is that in order to find a uniformly (in the value of the parameters) best estimator, we need to impose certain restrictions on the class of estimators under consideration.

Definition. Let $\mathcal{W}$ be a class of estimators. An estimator $\hat{\theta}$ of θ is *efficient relative to* $\mathcal{W}$ if

$$MSE_\theta\left(\hat{\theta}\right) \leq MSE_\theta\left(W\right) \qquad \forall \theta \in \Theta$$

for every $W \in \mathcal{W}$.

Remark. When θ is a vector, the notation "$MSE_\theta\left(\hat{\theta}\right) \leq MSE_\theta\left(W\right)$" is shorthand for "the matrix $MSE_\theta\left(W\right) - MSE_\theta\left(\hat{\theta}\right)$ is positive semi-definite".

Example. Suppose $X_i \sim i.i.d.\ \mathcal{N}\left(\mu, \sigma^2\right)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters. The estimator (of σ^2) $\hat{\sigma}^2$ is efficient relative to $\mathcal{W} = \left\{\hat{\sigma}^2, S^2\right\}$.

Example. Suppose $X_i \sim i.i.d.\ \mathcal{N}\left(\mu, \sigma^2\right)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters. Consider the following class of estimators (of σ^2):

$$\mathcal{W} = \left\{\tilde{\sigma}_c^2 = \frac{1}{c}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 : c > 0\right\}.$$

The estimators

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \tilde{\sigma}_n^2$$

and

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

are both members of $\mathcal{W}$. It is not hard to show that $\tilde{\sigma}_{n+1}^2$ is efficient relative to $\mathcal{W}$.

A "natural" class of estimators with fairly general applicability is the class

$$\mathcal{W}_u\left(\theta\right) = \left\{W : E_\theta\left(W\right) = \theta \text{ and } Var_\theta\left(W\right) < \infty \text{ for every } \theta \in \Theta\right\}$$

of unbiased estimators of θ with finite variance. For unbiased estimators, the MSE is simply the variance.

Definition. An estimator $\hat{\theta} \in \mathcal{W}_u(\theta)$ of θ is a *uniform minimum variance unbiased (UMVU)* estimator of θ if $\hat{\theta}$ is efficient relative to $\mathcal{W}_u(\theta)$.

It turns out that UMVU estimators often exist. The Rao-Blackwell Theorem facilitates the search for UMVU estimators by showing that UMVU estimators can always be based on statistics that are sufficient in the sense of the following definition.

Definition. Let $X_1, \ldots, X_n$ be a random sample from a distribution with cdf $F(\cdot|\theta)$, where $\theta \in \Theta$ is unknown. A statistic $T = T(X_1, \ldots, X_n)$ is a *sufficient statistic for* θ if the conditional distribution of $(X_1, \ldots, X_n)'$ given T does not depend on θ.

Theorem (Rao-Blackwell Theorem; Casella and Berger, Theorem 7.3.17). *Let* $\hat{\theta} \in \mathcal{W}_u(\theta)$ *and let* T *be any sufficient statistic for* θ. *Then*

$$\tilde{\theta} = E_{X|T}\left(\hat{\theta}|T\right) \in \mathcal{W}_u(\theta)$$

and

$$Var_\theta\left(\tilde{\theta}\right) \leq Var_\theta\left(\hat{\theta}\right) \qquad \forall \theta \in \Theta.$$

Remark. The inequality $Var_\theta\left(\tilde{\theta}\right) \leq Var_\theta\left(\hat{\theta}\right)$ is strict unless $P_\theta\left(\tilde{\theta} = \hat{\theta}\right) = 1$.

Proof. The distribution of the estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ conditional on T does not depend on θ when T is sufficient. Therefore,

$$\tilde{\theta} = E_{X|T}\left(\hat{\theta}|T\right)$$

is a function of $T = T(X_1, \ldots, X_n)$ that does not depend on the true (unknown) value of θ. In particular, $\tilde{\theta}$ is an estimator.

The estimator $\tilde{\theta}$ is unbiased because

$$E_\theta\left(\tilde{\theta}\right) = E_\theta\left(E_{X|T}\left(\hat{\theta}|T\right)\right) = E_\theta\left(\hat{\theta}\right) = \theta,$$

where the second equality uses the law of iterated expectations and the last equality uses the fact that $\hat{\theta}$ is unbiased.

Applying the conditional variance identity, we have:

$$\begin{aligned} Var_\theta\left(\hat{\theta}\right) &= Var_\theta\left(E_{X|T}\left(\hat{\theta}|T\right)\right) + E_\theta\left(Var_{X|T}\left(\hat{\theta}|T\right)\right) \\ &\geq Var_\theta\left(E_{X|T}\left(\hat{\theta}|T\right)\right) = Var_\theta\left(\tilde{\theta}\right). \qquad \blacksquare \end{aligned}$$

Remark. With a little more effort, a proof of the relation $Var_\theta\left(\hat{\theta}\right) \geq Var_\theta\left(\tilde{\theta}\right)$ can be based on the conditional version of Jensen's inequality:

$$\begin{aligned} Var_\theta\left(\hat{\theta}\right) &= E_\theta\left(\left(\hat{\theta}-\theta\right)^2\right) = E_\theta\left(E_{X|T}\left(\left(\hat{\theta}-\theta\right)^2|T\right)\right) \\ &\geq E_\theta\left(\left(E_{X|T}\left(\hat{\theta}|T\right)-\theta\right)^2\right) = E_\theta\left(\left(\tilde{\theta}-\theta\right)^2\right) = Var_\theta\left(\tilde{\theta}\right), \end{aligned}$$

where the second equality uses the law of iterated expectations and the inequality uses the conditional version of Jensen's inequality. This method of proof is applicable whenever the measure of closeness is of the form $E_\theta\left(L\left(\hat{\theta},\theta\right)\right)$, where $L\left(\hat{\theta},\theta\right)$ is a convex function of $\hat{\theta}$:

$$\begin{aligned} E_\theta\left(L\left(\hat{\theta},\theta\right)\right) &= E_\theta\left(E_{X|T}\left(L\left(\hat{\theta},\theta\right)|T\right)\right) \\ &\geq E_\theta\left(L\left(E_{X|T}\left(\hat{\theta}|T\right),\theta\right)\right) = E_\theta\left(L\left(\tilde{\theta},\theta\right)\right). \end{aligned}$$

For instance, $\left|\hat{\theta}-\theta\right|$ is a convex function of $\hat{\theta}$ and therefore

$$\begin{aligned} E_\theta\left(\left|\hat{\theta}-\theta\right|\right) &= E_\theta\left(E_{X|T}\left(\left|\hat{\theta}-\theta\right||T\right)\right) \\ &\geq E_\theta\left(\left(\left|E_{X|T}\left(\hat{\theta}|T\right)-\theta\right|\right)\right) = E_\theta\left(\left|\tilde{\theta}-\theta\right|\right). \end{aligned}$$

It follows from the Rao-Blackwell Theorem that when looking for UMVU estimators, there is no need to consider estimators that cannot be written as functions of a sufficient statistic. Indeed, it suffices to look at estimators that are *necessary statistics* in the sense that they can be written as functions of every sufficient statistic. Here, the word "every" is crucial because any estimator is a function of $(X_1,\ldots,X_n)'$ and $(X_1,\ldots,X_n)'$ is always a sufficient statistic.

Of course, the usefulness of the Rao-Blackwell Theorem depends on the extent to which sufficient statistics of low dimension are available and easy to find. As it turns out, sufficient statistics of the same dimension as θ are available in many cases. In cases where determination of sufficient statistics by means of the definition is tedious, the following characterization of sufficiency may be useful.

Theorem (Factorization Criterion; Casella and Berger, Theorem 6.2.6). *Let $X_1,\ldots,X_n$ be a random sample from a discrete (continuous) distribution with cdf $F(\cdot|\theta)$, where $\theta\in\Theta$ is unknown. A statistic $T=T(X_1,\ldots,X_n)$ is a sufficient statistic for θ if and only if there exist functions $g(\cdot|\cdot)$ and $h(\cdot)$ such that $f_X(\cdot|\theta)$ is a pmf (pdf) of $(X_1,\ldots,X_n)'$, where*

$$f_X(x_1,\ldots,x_n|\theta) = g(T(x_1,\ldots,x_n)|\theta)\,h(x_1,\ldots,x_n)$$

for every $(x_1,\ldots,x_n)\in R^n$ and every $\theta\in\Theta$.

Example. For any random sample $X_1, \ldots, X_n$ from a discrete (continuous) distribution, two sufficient statistics are $(X_1, \ldots, X_n)'$ and $\left(X_{(1)}, \ldots, X_{(n)}\right)'$.

Example. Suppose $X_i \sim i.i.d.\ Ber(p)$, where $p \in [0,1]$ is an unknown parameter. The joint pmf of $(X_1, \ldots, X_n)'$ is

$$\begin{aligned} f_X(x_1, \ldots, x_n|p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \cdot 1(x_i \in \{0,1\}) \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \cdot \left(\prod_{i=1}^{n} 1(x_i \in \{0,1\})\right) \\ &= g\left(\sum_{i=1}^n x_i|p\right) \cdot h(x_1, \ldots, x_n), \end{aligned}$$

where

$$g(t|p) = p^t(1-p)^{n-t}$$

and

$$h(x_1, \ldots, x_n) = \prod_{i=1}^{n} 1(x_i \in \{0,1\}).$$

Therefore, $\sum_{i=1}^n X_i$ is a sufficient statistic for p.

The maximum likelihood (and method of moments) estimator

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n}$$

is a function of $\sum_{i=1}^n X_i$.

Example. Suppose $X_i \sim i.i.d.\ U[0,\theta]$, where $\theta > 0$ is an unknown parameter. The joint pdf of $(X_1, \ldots, X_n)'$ is

$$\begin{aligned} f_X(x_1, \ldots, x_n|\theta) &= \prod_{i=1}^{n} \left(\frac{1}{\theta} 1(0 \le x_i \le \theta)\right) \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} 1(0 \le x_i \le \theta) \\ &= \frac{1}{\theta^n} 1\left(x_{(n)} \le \theta\right) \cdot 1\left(x_{(1)} \ge 0\right) \\ &= g\left(x_{(n)}|\theta\right) \cdot h(x_1, \ldots, x_n), \end{aligned}$$

where

$$g\left(t|\theta\right)=\frac{1}{\theta^{n}}1\left(t\leq\theta\right)$$

and

$$h\left(x_{1},\ldots,x_{n}\right)=1\left(x_{(1)}\geq0\right).$$

Therefore, $X_{(n)}=\max_{1\leq i\leq n}X_{i}$ is a sufficient statistic for θ.

The method of moments estimator $\hat{\theta}_{MM}=2\bar{X}$ is unbiased but is not a function of $X_{(n)}$ (unless $n=1$) and therefore cannot be UMVU. On the other hand, the maximum likelihood estimator $\hat{\theta}_{ML}=X_{(n)}$ is based on $X_{(n)}$.

Example. Suppose $X_{i}\sim i.i.d.\ \mathcal{N}\left(\mu,\sigma^{2}\right)$, where $\mu\in\mathbb{R}$ and $\sigma^{2}>0$ are unknown parameters. In this case, the joint pdf of $\left(X_{1},\ldots,X_{n}\right)'$ is

$$\begin{aligned}
f_{X}\left(x_{1},\ldots,x_{n}|\mu,\sigma^{2}\right) &= \prod_{i=1}^{n}\left(2\pi\sigma^{2}\right)^{-1/2}\exp\left(-\frac{1}{2\sigma^{2}}\left(x_{i}-\mu\right)^{2}\right)\\
&= \left(2\pi\right)^{n/2}\left(\sigma^{2}\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^{2}}\sum_{i=1}^{n}\left(x_{i}^{2}+\mu^{2}-2\mu x_{i}\right)\right)\\
&= \left(2\pi\right)^{n/2}\left(\sigma^{2}\right)^{-n/2}\exp\left(-\frac{n\mu^{2}}{2\sigma^{2}}\right)\exp\left(\frac{\mu}{\sigma^{2}}\sum_{i=1}^{n}x_{i}-\frac{1}{2\sigma^{2}}\sum_{i=1}^{n}x_{i}^{2}\right)\\
&= g\left(\textstyle\sum_{i=1}^{n}x_{i},\sum_{i=1}^{n}x_{i}^{2}|\mu,\sigma^{2}\right)\cdot h\left(x_{1},\ldots,x_{n}\right),
\end{aligned}$$

where

$$g\left(t_{1},t_{2}|\mu,\sigma^{2}\right)=\left(2\pi\right)^{n/2}\left(\sigma^{2}\right)^{-n/2}\exp\left(-\frac{n\mu^{2}}{2\sigma^{2}}\right)\exp\left(\frac{\mu}{\sigma^{2}}t_{1}-\frac{1}{2\sigma^{2}}t_{2}\right)$$

and

$$h\left(x_{1},\ldots,x_{n}\right)=1.$$

Therefore, $\left(\sum_{i=1}^{n}X_{i},\sum_{i=1}^{n}X_{i}^{2}\right)'$ is a sufficient statistic for $\left(\mu,\sigma^{2}\right)'$.

The maximum likelihood (and method of moments) estimator of μ,

$$\hat{\mu}=\bar{X}=\frac{\sum_{i=1}^{n}X_{i}}{n},$$

is a function of $\left(\sum_{i=1}^{n}X_{i},\sum_{i=1}^{n}X_{i}^{2}\right)'$, as is the maximum likelihood (and method of moments) estimator of σ^{2},

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

A sufficient statistic T is particularly useful if unbiased estimators based on T are essentially unique in the sense that any two unbiased estimators based on T are equal with probability one. Indeed, if we can somehow find a sufficient statistic T such that unbiased estimators based on T are essentially unique, then any $\hat{\theta} \in \mathcal{W}_u(\theta)$ based on T is UMVU. The Lehmann-Scheffé Theorem establishes essential uniqueness of unbiased estimators based on a complete sufficient statistic T.

Definition. A sufficient statistic T for θ is *complete* if

$$E_\theta(g(T)) = 0 \qquad \forall \theta \in \Theta$$

implies

$$P_\theta(g(T) = 0) = 1 \qquad \forall \theta \in \Theta.$$

Theorem (Lehmann-Scheffé Theorem; Casella and Berger, Theorem 7.5.1). *Unbiased estimators based on complete sufficient statistics are essentially unique.*

Proof. Suppose $\hat{\theta} = \hat{\theta}(T)$ and $\tilde{\theta} = \tilde{\theta}(T)$ are unbiased estimators θ based on a sufficient statistic T. Then

$$E_\theta\left(\hat{\theta}(T) - \tilde{\theta}(T)\right) = E_\theta\left(\hat{\theta}(T)\right) - E_\theta\left(\tilde{\theta}(T)\right) = \theta - \theta = 0 \qquad \forall \theta \in \Theta$$

because $\hat{\theta}$ and $\tilde{\theta}$ are unbiased. If T is complete, then

$$P_\theta\left(\hat{\theta}(T) - \tilde{\theta}(T) = 0\right) = P_\theta\left(\hat{\theta}(T) = \tilde{\theta}(T)\right) = 1 \qquad \forall \theta \in \Theta. \qquad \blacksquare$$

Corollary. *If T is a complete sufficient statistic and $\hat{\theta} \in \mathcal{W}_u(\theta)$ is based on T, then $\hat{\theta}$ is a UMVU estimator of θ.*

Proof. Suppose there exists an estimator $\tilde{\theta} \in \mathcal{W}_u(\theta)$ such that $Var_{\theta^*}\left(\tilde{\theta}\right) < Var_{\theta^*}\left(\hat{\theta}\right)$ for some $\theta^* \in \Theta$. By the Rao-Blackwell theorem, $E\left(\tilde{\theta}|T\right) \in \mathcal{W}_u(\theta)$ is based on T and satisfies

$$Var_{\theta^*}\left(E\left(\tilde{\theta}|T\right)\right) \leq Var_{\theta^*}\left(\tilde{\theta}\right) < Var_{\theta^*}\left(\hat{\theta}\right).$$

In view of the Lehmann-Scheffé theorem, this is impossible because $Var_{\theta^*}\left(E\left(\tilde{\theta}|T\right)\right) < Var_{\theta^*}\left(\hat{\theta}\right)$ implies that $\hat{\theta}$ is not an essentially unique unbiased estimator of θ based on T. $\blacksquare$

Complete sufficient statistics can often be found if the family $\{f(\cdot|\theta) : \theta \in \Theta\}$ of pmfs/pdfs is an exponential family.

Definition. A family $\{f(\cdot|\theta) : \theta \in \Theta\}$ of pmfs/pdfs is called a *d-dimensional exponential family* if there exist functions $h : \mathbb{R} \to \mathbb{R}_+$, $c : \Theta \to \mathbb{R}_+$, $\eta_i : \Theta \to \mathbb{R}$ $(i = 1, \ldots, d)$ and $t_i : \mathbb{R} \to \mathbb{R}$ $(i = 1, \ldots, d)$ such that

$$f(x|\theta) = h(x)\, c(\theta) \exp\left(\sum_{i=1}^{d} \eta_i(\theta)\, t_i(x)\right), \qquad \forall x \in \mathbb{R}, \theta \in \Theta.$$

Example. Suppose $X_i \sim i.i.d.\ Ber(p)$, where $p \in (0,1)$. The marginal pmf satisfies

$$\begin{aligned} f(x|p) &= p^x (1-p)^{1-x} \cdot 1(x \in \{0,1\}) \\ &= \left(\frac{p}{1-p}\right)^x (1-p) \cdot 1(x \in \{0,1\}) \\ &= \exp\left(\log\left(\frac{p}{1-p}\right) \cdot x\right)(1-p) \cdot 1(x \in \{0,1\}) \\ &= h(x)\, c(p) \exp(\eta(p)\, t(x)), \end{aligned}$$

where

$$\begin{aligned} h(x) &= 1(x \in \{0,1\}), \\ c(p) &= 1-p, \\ \eta(p) &= \log\left(\frac{p}{1-p}\right), \\ t(x) &= x. \end{aligned}$$

Example. Suppose $X_i \sim i.i.d.\ \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters. In this case, the marginal pdf satisfies

$$\begin{aligned} f(x|\mu, \sigma^2) &= \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \\ &= (2\pi)^{1/2} \left(\sigma^2\right)^{-1/2} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2\right) \\ &= h(x)\, c(\mu, \sigma^2) \exp\left(\sum_{i=1}^{2} \eta_i(\mu, \sigma^2)\, t_i(x)\right), \end{aligned}$$

where

$$\begin{aligned} h(x) &= 1, \\ c(\mu, \sigma^2) &= (2\pi)^{1/2} (\sigma^2)^{-1/2} \exp\left(-\frac{\mu^2}{2\sigma^2}\right), \\ \eta_1(\mu, \sigma^2) &= \frac{\mu}{\sigma^2}, \\ t_1(x) &= x, \\ \eta_2(\mu, \sigma^2) &= -\frac{1}{2\sigma^2}, \\ t_2(x) &= x^2. \end{aligned}$$

Theorem (Casella and Berger, Theorem 6.2.25). *Let $X_1, \ldots, X_n$ be a random sample from a discrete (continuous) exponential family with pmf (pdf)*

$$f(x|\theta) = h(x) c(\theta) \exp\left(\sum_{i=1}^{d} \eta_i(\theta) t_i(x)\right), \qquad x \in \mathbb{R}, \theta \in \Theta.$$

The sufficient (for θ) statistic

$$T(X_1, \ldots, X_n) = \left(\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_d(X_i)\right)'$$

is complete if the set $\{(\eta_1(\theta), \ldots, \eta_d(\theta)) : \theta \in \Theta\}$ contains an open set.

Remark. A set $A \subseteq \mathbb{R}^d$ contains an open set if and only if we can find constants $a_1^L < a_1^U, \ldots, a_d^L < a_d^U$ such that

$$\left[a_1^L, a_1^U\right] \times \ldots \times \left[a_d^L, a_d^U\right] \subseteq A;$$

that is, a set $A \subseteq \mathbb{R}^d$ contains an open set if and only if it contains a d-dimensional rectangle.

Example. Suppose $X_i \sim$ *i.i.d. Ber* (p), where $p \in (0,1)$. The marginal pdf satisfies

$$f(x|p) = h(x) c(p) \exp(\eta(p) t(x)),$$

where

$$\begin{aligned} h(x) &= 1(x \in \{0,1\}), \\ c(p) &= 1-p, \\ \eta(p) &= \log\left(\frac{p}{1-p}\right), \\ t(x) &= x. \end{aligned}$$

The set

$$\{\eta(p) : p \in (0,1)\} = \left\{\log\left(\frac{p}{1-p}\right) : p \in (0,1)\right\} = \mathbb{R}$$

is open, so

$$\sum_{i=1}^{n} t(X_i) = \sum_{i=1}^{n} X_i$$

is a complete sufficient statistic.

The maximum likelihood estimator

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$$

is unbiased,

$$E_p(\hat{p}) = E_p(X_i) = p,$$

and is based on $\sum_{i=1}^{n} X_i$. Therefore, $\hat{p}$ is a UMVU estimator of p.

The conclusion is not affected if $\Theta = [0,1]$ is considered, as $Var_p(\hat{p}) = 0$ when $p \in \{0,1\}$.

Example. Suppose $X_i \sim i.i.d.\ \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters. In this case, the marginal pdf satisfies

$$f(x|\mu,\sigma^2) = h(x)\, c(\mu,\sigma^2) \exp\left(\sum_{i=1}^{2} \eta_i(\mu,\sigma^2)\, t_i(x)\right),$$

where

$$\begin{aligned} h(x) &= 1, \\ c(\mu,\sigma^2) &= (2\pi)^{1/2}(\sigma^2)^{-1/2}\exp\left(-\frac{\mu^2}{2\sigma^2}\right), \\ \eta_1(\mu,\sigma^2) &= \frac{\mu}{\sigma^2}, \\ t_1(x) &= x, \\ \eta_2(\mu,\sigma^2) &= -\frac{1}{2\sigma^2}, \\ t_2(x) &= x^2. \end{aligned}$$

The set

$$\left\{(\eta_1(\mu,\sigma^2),\eta_2(\mu,\sigma^2))' : \mu\in\mathbb{R},\sigma^2>0\right\} = \left\{\left(\frac{\mu}{\sigma^2},-\frac{1}{2\sigma^2}\right)' : \mu\in\mathbb{R},\sigma^2>0\right\} = \mathbb{R}\times(-\infty,0)$$

is open, so

$$\left(\sum_{i=1}^n t_1(X_i),\sum_{i=1}^n t_2(X_i)\right)' = \left(\sum_{i=1}^n X_i,\sum_{i=1}^n X_i^2\right)'$$

is a complete sufficient statistic.

The maximum likelihood estimator

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

of μ is unbiased and is based on $\left(\sum_{i=1}^n X_i,\sum_{i=1}^n X_i^2\right)'$ and $\bar{X}$ is therefore a UMVU estimator of μ.

An alternative UMVU estimator of μ is

$$\tilde{\mu} = \begin{cases} \bar{X} & \text{for } \bar{X}\neq 0 \\ 1 & \text{for } \bar{X}=0. \end{cases}$$

Therefore, $\bar{X}$ is only an essentially unique UMVU estimator of μ.

The sample variance

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

is unbiased and is based on $\left(\sum_{i=1}^n X_i,\sum_{i=1}^n X_i^2\right)'$ and S^2 is therefore a UMVU estimator of σ^2.

Outside the exponential family of distributions, we typically have to find complete sufficient statistics (if they exist) by applying the definition of completeness.

Example. Suppose $X_i \sim i.i.d.\ U[0,\theta]$, where $\theta > 0$ is an unknown parameter. The sufficient statistic $T = T(X_1,\ldots,X_n) = X_{(n)}$ is continuous with pdf $f(\cdot|\theta)$ given by (Casella and Berger, Example 6.2.23)

$$f(t|\theta) = nt^{n-1}\theta^{-n}1(0 \leq t \leq \theta), \qquad \theta > 0.$$

The statistic T is complete if the set $\{x \in \mathbb{R}_+ : g(x) \neq 0\}$ has (Lebesgue) measure zero whenever $g : \mathbb{R} \to \mathbb{R}$ is a function satisfying

$$E_\theta(g(T)) = \int_0^\theta g(t)\, nt^{n-1}\theta^{-n}dt = 0 \qquad \forall\theta > 0.$$

Given any such function g, let g^+ and g^- be the negative and positive parts of g, respectively; that is, let $g^+(t) = \max(0, g(t))$ and $g^-(t) = \max(0, -g(t))$. By assumption,

$$\int_0^\theta g^+(t)\, t^{n-1}dt = \int_0^\theta g^-(t)\, t^{n-1}dt \qquad \forall\theta > 0.$$

It can be shown (using the Radon-Nikodym theorem) that this implies that the set $\{x \in \mathbb{R}_+ : g^+(x) \neq g^-(x)\}$ has measure zero. Therefore, the set $\{x \in \mathbb{R}_+ : g(x) = g^+(x) - g^-(x) \neq 0\}$ has measure zero. In particular, $X_{(n)}$ is a complete sufficient statistic.

The maximum likelihood estimator $\hat{\theta}_{ML} = X_{(n)}$ is based on $X_{(n)}$ but is biased because

$$\begin{aligned} E_\theta\left(\hat{\theta}_{ML}\right) &= \int_0^\theta tf(t|\theta)\,dt = \int_0^\theta nt^n\theta^{-n}dt \\ &= \frac{n}{n+1}t^{n+1}\theta^{-n}\Big|_{t=0}^{\theta} \\ &= \frac{n}{n+1}\theta. \end{aligned}$$

On the other hand, the estimator

$$\hat{\theta} = \frac{n+1}{n}X_{(n)} = \frac{n+1}{n}\hat{\theta}_{ML}$$

is unbiased and is based on the complete sufficient statistic $X_{(n)}$. As a consequence, $\hat{\theta}$ is a UMVU estimator of θ.

In this example, we constructed a UMVU estimator of θ by finding "the" function $\hat{\theta}(\cdot)$ such that

$$E_\theta\left(\hat{\theta}(T)\right) = \theta \qquad \forall\theta \in \Theta.$$

In cases where an unbiased estimator $\hat{\theta}$ (not based on a complete sufficient statistic) has already been found, a UMVU estimator of θ can be found by "Rao-Blackwellization"; that is,

$$\tilde{\theta} = E\left(\hat{\theta}|T\right)$$

is UMVU if $\hat{\theta}$ is unbiased and T is a complete sufficient statistic.

Example. Suppose $X_i \sim i.i.d.\ U[0,\theta]$, where $\theta > 0$ is an unknown parameter. Since $E_\theta(X_i) = \theta/2$, an unbiased estimator of θ is

$$\hat{\theta} = 2X_1.$$

The conditional distribution of X_1 given $X_{(n)} = x_{(n)}$ is a mixture distribution. Specifically, X_1 equals $x_{(n)}$ with probability $1/n$ and is uniformly distributed on $\left[0, x_{(n)}\right]$ with probability $(n-1)/n$. As a consequence,

$$E\left(\hat{\theta}|X_{(n)}\right) = 2E\left(X_1|X_{(n)}\right) = 2\left(\frac{1}{n}X_{(n)} + \frac{n-1}{n}\cdot\frac{X_{(n)}}{2}\right) = \frac{n+1}{n}X_{(n)}$$

is a UMVU estimator of θ.

If a UMVU estimator does not exist or is hard to find, it is nice to have a benchmark against which all estimators can be compared. That is, it is nice to have a lower bound on the variance of any unbiased estimator.

Definition. An estimator $\hat{\theta} \in \mathcal{W}_u(\theta)$ of θ is *locally minimum variance unbiased (LMVU) at* θ_0 if

$$Var_{\theta_0}\left(\hat{\theta}\right) \leq Var_{\theta_0}(W)$$

for any $W \in \mathcal{W}_u(\theta)$.

Suppose an LMVU estimator exists at every $\theta_0 \in \Theta$ and let $V_L(\theta_0)$ denote the variance of the LMVU estimator at θ_0. The function $V_L(\cdot) : \Theta \to \mathbb{R}_+$ provides us with a lower bound on the variance of any estimator $\hat{\theta} \in \mathcal{W}_u(\theta)$. The bound is sharp in the sense that it can be attained for any $\theta_0 \in \Theta$. If the LMVU estimator (or its variance) is hard to find, we may nonetheless be able to find a nontrivial lower bound on the variance of any unbiased estimator of θ. A very useful bound, the Cramér-Rao bound, can be obtained by applying the covariance inequality, a corollary of the Cauchy-Schwarz inequality.

Theorem (Cauchy-Schwarz Inequality; Casella and Berger, Theorem 4.7.3). *If (X,Y) is a bivariate random vector, then*

$$|E(XY)| \leq E|XY| \leq \left(E\left(X^2\right)\right)^{1/2}\left(E\left(Y^2\right)\right)^{1/2}.$$

Corollary (Covariance Inequality; Casella and Berger, Example 4.7.4). *If (X,Y) is a bivariate random vector, then*

$$Cov\left(X,Y\right)^2 \leq Var\left(X\right)Var\left(Y\right).$$

Let $X_1,\ldots,X_n$ be a random sample from a discrete (continuous) distribution with pmf (pdf) $f\left(\cdot|\theta\right)$, where $\theta \in \Theta$ is an unknown parameter vector. Moreover, let $\hat{\theta} = \hat{\theta}\left(X_1,\ldots,X_n\right)$ be any estimator of θ and let $\psi = \psi\left(X_1,\ldots,X_n;\theta\right)$ be any vector-valued function of $\left(X_1,\ldots,X_n\right)'$ and θ. It follows from the (multivariate version of the) covariance inequality that

$$Var_\theta\left(\hat{\theta}\right) \geq Cov_\theta\left(\hat{\theta},\psi\right)Var_\theta\left(\psi\right)^{-1}Cov_\theta\left(\hat{\theta},\psi\right)'.$$

In general, the lower bound on the right hand side depends on $\hat{\theta}$ and therefore the inequality may not seem to be very helpful.

Remark. It can be shown that the function (of θ and $\hat{\theta}$) $Cov_\theta\left(\hat{\theta},\psi\right)$ depends on $\hat{\theta}$ only through $E_\theta\left(\hat{\theta}\right)$ if and only if

$$Cov_\theta\left(\psi,U\right) = 0 \qquad \forall\theta \in \Theta$$

whenever U is a random variable with $E_\theta\left(U\right) = 0$ and $Var_\theta\left(U\right) < \infty$ for every $\theta \in \Theta$.

In particular, the function (of θ and $\hat{\theta}$) $Cov_\theta\left(\hat{\theta},\psi\right)$ depends on $\hat{\theta}$ only through θ if and only if

$$Cov_\theta\left(\psi,U\right) = 0 \qquad \forall\theta \in \Theta$$

whenever U is a random variable with $E_\theta\left(U\right) = 0$ and $Var_\theta\left(U\right) < \infty$ for every $\theta \in \Theta$.

It turns out that under certain conditions (on $f\left(\cdot|\theta\right)$), $Cov_\theta\left(\hat{\theta},\psi\right)$ is independent of $\hat{\theta}$ when $\hat{\theta} \in \mathcal{W}_u\left(\theta\right)$ and ψ is the score function evaluated at θ.

Definition. Let $X_1,\ldots,X_n$ be a random sample from a discrete (continuous) distribution with pmf (pdf) $f\left(\cdot|\theta\right)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$ is an unknown parameter vector. The *score function* is the (random) function $S\left(\cdot|X_1,\ldots,X_n\right) : \Theta \to \mathbb{R}^k$ given by

$$S\left(\theta|X_1,\ldots,X_n\right) = \frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log f\left(X_i|\theta\right), \qquad \theta \in \Theta.$$

Theorem (Cramér-Rao Inequality; Casella and Berger, Theorem 7.3.9). *Let $X_1,\ldots,X_n$ be a random sample from a discrete (continuous) distribution with pmf (pdf) $f\left(\cdot|\theta\right)$, where $\theta \in \Theta$ is an unknown parameter vector. Moreover, let $\hat{\theta} \in \mathcal{W}_u\left(\theta\right)$. If*

$$E_\theta\left(S\left(\theta|X_1,\ldots,X_n\right)\right) = 0$$

and

$$E_\theta\left(\hat{\theta}\cdot S\left(\theta|X_1,\ldots,X_n\right)'\right)=I,$$

then

$$Var_\theta\left(\hat{\theta}\right)\geq Var_\theta\left(S\left(\theta|X_1,\ldots,X_n\right)\right)^{-1}$$

whenever $Var_\theta\left(S\left(\theta|X_1,\ldots,X_n\right)\right)$ *exists and is positive definite.*

Proof. Let $S\left(\theta\right)=S\left(\theta|X_1,\ldots,X_n\right)$. Under the stated assumptions,

$$\begin{aligned}Cov_\theta\left(\hat{\theta},S\left(\theta\right)\right) &= E_\theta\left(\hat{\theta}\cdot\left(S\left(\theta\right)-E_\theta\left(S\left(\theta\right)\right)\right)'\right)=E_\theta\left(\hat{\theta}\cdot S\left(\theta\right)'\right)\\ &= I,\end{aligned}$$

and it follows from the covariance inequality that

$$\begin{aligned}Var_\theta\left(\hat{\theta}\right) &\geq Cov_\theta\left(\hat{\theta},S\left(\theta\right)\right)Var_\theta\left(S\left(\theta\right)\right)^{-1}Cov_\theta\left(\hat{\theta},S\left(\theta\right)\right)'\\ &= Var_\theta\left(S\left(\theta\right)\right)^{-1}. \qquad \blacksquare\end{aligned}$$

The high-level assumptions

$$E_\theta\left(S\left(\theta|X_1,\ldots,X_n\right)\right)=0$$

and

$$E_\theta\left(\hat{\theta}\cdot S\left(\theta|X_1,\ldots,X_n\right)'\right)=I$$

both have natural interpretations. Suppose $X=\left(X_1,\ldots,X_n\right)'$ is continuous with joint pdf $f_X\left(\cdot|\theta\right)$ and let $T\left(X_1\ldots,X_n\right)$ be any statistic with $E_\theta\left|T\left(X_1\ldots,X_n\right)\right|<\infty$. If we can interchange the order of differentiation and differentiation, then

$$\begin{aligned}\frac{\partial}{\partial\theta'}E_\theta\left(T\left(X\right)\right) &= \frac{\partial}{\partial\theta'}\int_{\mathbb{R}^n}T\left(x\right)f_X\left(x|\theta\right)dx\\ &= \int_{\mathbb{R}^n}T\left(x\right)\frac{\partial}{\partial\theta'}f_X\left(x|\theta\right)dx\\ &= \int_{\mathbb{R}^n}T\left(x\right)\left(\frac{\partial}{\partial\theta'}\log f_X\left(x|\theta\right)\right)f_X\left(x|\theta\right)dx\\ &= E_\theta\left(T\left(X\right)S\left(\theta|X\right)'\right)\end{aligned}$$

where the last equality holds because

$$\begin{aligned} S(\theta|X_1,\ldots,X_n) &= \frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log f(X_i|\theta) \\ &= \frac{\partial}{\partial\theta}\log\left(\prod_{i=1}^{n} f(X_i|\theta)\right) \\ &= \frac{\partial}{\partial\theta}\log f_X(X_1,\ldots,X_n|\theta) \end{aligned}$$

when $X_1,\ldots,X_n$ is a random sample from a distribution with pdf $f(\cdot|\theta)$. Setting $h(X)=1$ and $h(X)=\hat{\theta}$, we obtain the relations

$$E_\theta\left(S(\theta|X_1,\ldots,X_n)\right)=0$$

and

$$E_\theta\left(\hat{\theta}\cdot S(\theta|X_1,\ldots,X_n)'\right)=I,$$

respectively.

Conditions under which we can interchange of the order of integration and differentiation are available (Casella and Berger, Section 2.4).

Lemma. *Let $X_1,\ldots,X_n$ be a random sample from a discrete (continuous) distribution with pmf (pdf) $f(\cdot|\theta)$, where $\theta\in\Theta$. Suppose*

(i) Θ is open.
(ii) The set $\{x\in\mathbb{R}: f(x|\theta)>0\}$ does not depend on θ.
(iii) For every $\theta\in\Theta$ there is a function $b_\theta:\mathbb{R}\to\mathbb{R}_+$ and a constant $\Delta_\theta>0$ such that

$$Var_\theta(b_\theta(X))<\infty$$

and

$$\left|\frac{f(x|\theta+\delta)-f(x|\theta)}{\delta}\right|<b_\theta(x)\qquad\forall x\in\mathbb{R}$$

whenever $|\delta|<\Delta_\theta$.

Then

$$\frac{\partial}{\partial\theta'}E_\theta(T(X))=E_\theta\left(T(X)\,S(\theta|X)'\right)$$

for any statistic $T(X_1\ldots,X_n)$ and any $\theta\in\Theta$.

Remark. Conditions (ii) and (iii) of the lemma hold whenever $f(x|\theta)$ is of the form

$$f(x|\theta) = h(x)\, c(\theta) \exp\left(\sum_{i=1}^{d} \eta_i(\theta)\, t_i(x)\right), \qquad x \in \mathbb{R}, \theta \in \Theta,$$

where each $\eta_i(\cdot)$ is differentiable.

The quantity

$$\mathcal{I}(\theta) = E_\theta\left(S(\theta|X_1,\ldots,X_n)\, S(\theta|X_1,\ldots,X_n)'\right)$$

is called the *information matrix,* or the *Fisher information.* Under the assumptions of Cramér-Rao Inequality, the Fisher information is

$$\mathcal{I}(\theta) = Var_\theta\left(S(\theta|X_1,\ldots,X_n)\right)$$

and $\mathcal{I}(\theta)^{-1}$ provides a lower bound on the variance of any estimator $\hat{\theta} \in \mathcal{W}_u(\theta)$.

The Fisher information is easy to compute when $X_1,\ldots,X_n$ is a random sample (Casella and Berger, Corollary 7.3.10):

$$\begin{aligned}
\mathcal{I}(\theta) &= E_\theta\left(\left(\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log f(X_i|\theta)\right)\left(\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log f(X_i|\theta)\right)'\right) \\
&= n \cdot E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)'\right),
\end{aligned}$$

where X is a random variable with pmf/pdf $f(\cdot|\theta)$. Moreover, if

$$\int_{-\infty}^{\infty}\frac{\partial^2}{\partial\theta\partial\theta'}f(x|\theta)\,dx = \frac{\partial^2}{\partial\theta\partial\theta'}\int_{-\infty}^{\infty}f(x|\theta)\,dx = 0,$$

then (Casella and Berger, Lemma 7.3.11)

$$\begin{aligned}
\mathcal{I}(\theta) &= n \cdot E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)'\right) \\
&= -n \cdot E_\theta\left(\frac{\partial^2}{\partial\theta\partial\theta'}\log f(X|\theta)\right).
\end{aligned}$$

If the conditions of the Cramér-Rao inequality are satisfied and it just so happens that an unbiased estimator attains the bound, then the estimator is UMVU. The Cramér-Rao inequality can therefore be used to establish optimality in some cases.

Example. Suppose $X_i \sim i.i.d.\ Ber(p)$, where $p \in (0,1)$ is unknown. When $x \in \{0,1\}$, we have:

$$\log f(x|p) = \log\left(p^x (1-p)^{1-x}\right) = x \cdot \log p + (1-x)\log(1-p),$$

and

$$\frac{\partial}{\partial p}\log f(x|p) = \frac{x}{p} - \frac{(1-x)}{1-p} = \frac{(1-p)x}{p(1-p)} + \frac{p(x-1)}{p(1-p)} = \frac{x}{p(1-p)} - \frac{1}{1-p}.$$

The conditions of the Cramér-Rao inequality are satisfied, so the Fisher information is

$$\begin{aligned}
\mathcal{I}(p) &= n \cdot Var_p\left(\frac{X_i}{p(1-p)} - \frac{1}{1-p}\right) \\
&= n \cdot \left(\frac{1}{p(1-p)}\right)^2 Var_p(X_i) \\
&= n \cdot \left(\frac{1}{p(1-p)}\right)^2 p(1-p) \\
&= \frac{n}{p(1-p)}.
\end{aligned}$$

The variance of any unbiased estimator of p is bounded from below by

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

Now, the maximum likelihood estimator

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}$$

of p satisfies

$$Var_p(\hat{p}) = Var_p\left(\bar{X}\right) = \frac{Var_p(X_i)}{n} = \frac{p(1-p)}{n}.$$

The maximum likelihood estimator attains the lower bound $\mathcal{I}(p)^{-1}$ and is therefore UMVU.

There are cases where the Cramér-Rao bound does not apply or fails to be sharp.

Example. Suppose $X_i \sim i.i.d.\ U[0,\theta]$, where $\theta > 0$ is an unknown parameter. When $0 < x < \theta$, we have:

$$\log f(x|\theta) = \log\frac{1}{\theta} = -\log\theta$$

and

$$\frac{\partial}{\partial\theta}\log f\left(x|\theta\right)=-\frac{1}{\theta}.$$

The conditions of the Cramér-Rao inequality are violated because

$$E_\theta\left(S\left(\theta|X_1,\ldots,X_n\right)\right)=n\cdot E_\theta\left(\frac{\partial}{\partial\theta}\log f\left(X_i|\theta\right)\right)=-\frac{n}{\theta}\neq 0.$$

We therefore cannot be sure that the Fisher information

$$\mathcal{I}\left(\theta\right)=n\cdot E_\theta\left(\left(-\frac{1}{\theta}\right)^2\right)=\frac{n}{\theta^2}$$

delivers a lower bound on the variance of unbiased estimators. In fact, the variance of UMVU estimator

$$\hat{\theta}=\frac{n+1}{n}X_{(n)}$$

of θ is (Casella and Berger, Example 7.3.13)

$$Var_\theta\left(\hat{\theta}\right)=\frac{1}{n\left(n+2\right)}\theta^2<\frac{1}{n}\theta^2=\mathcal{I}\left(\theta\right)^{-1}.$$

Example. Suppose $X_i\sim i.i.d.\ \mathcal{N}\left(\mu,\sigma^2\right)$, where $\mu\in\mathbb{R}$ and $\sigma^2>0$ are unknown parameters. We have:

$$\begin{aligned}\log f\left(x|\mu,\sigma^2\right) &= \log\left(\left(2\pi\sigma^2\right)^{-1/2}\exp\left(-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right)\right)\\ &= -\frac{1}{2}\log\left(2\pi\right)-\frac{1}{2}\log\left(\sigma^2\right)-\frac{1}{2\sigma^2}\left(x-\mu\right)^2,\\ \frac{\partial}{\partial\mu}\log f\left(x|\mu,\sigma^2\right)&=\frac{1}{\sigma^2}\left(x-\mu\right),\end{aligned}$$

and

$$\frac{\partial}{\partial\sigma^2}\log f\left(x|\mu,\sigma^2\right)=-\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}\left(x-\mu\right)^2.$$

The conditions of the Cramér-Rao inequality are satisfied, so the Fisher information is

$$\mathcal{I}\left(\mu,\sigma^2\right)=n\cdot\begin{pmatrix}\frac{1}{\sigma^2} & 0\\ 0 & \frac{1}{2\sigma^4}\end{pmatrix}.$$

because

$$Var_{(\mu,\sigma^2)}\left(\frac{\partial}{\partial\mu}\log f\left(X_i|\mu,\sigma^2\right)\right)=\frac{1}{\sigma^4}Var_{(\mu,\sigma^2)}\left(X_i-\mu\right)=\frac{1}{\sigma^2},$$

$$Var_{(\mu,\sigma^2)}\left(\frac{\partial}{\partial\sigma^2}\log f\left(X_i|\mu,\sigma^2\right)\right)=\frac{1}{\left(2\sigma^4\right)^2}Var_{(\mu,\sigma^2)}\left(\left(X_i-\mu\right)^2\right)=\frac{1}{2\sigma^4},$$

and

$$\begin{aligned}Cov_{(\mu,\sigma^2)}\left(\frac{\partial}{\partial\mu}\log f\left(X_i|\mu,\sigma^2\right),\frac{\partial}{\partial\sigma^2}\log f\left(X_i|\mu,\sigma^2\right)\right) &= \frac{1}{\sigma^2}\frac{1}{2\sigma^4}Cov_{(\mu,\sigma^2)}\left(\left(X_i-\mu\right),\left(X_i-\mu\right)^2\right)\\ &= 0.\end{aligned}$$

As a consequence, a lower bound on the covariance matrix of an unbiased estimator of $\left(\mu,\sigma^2\right)'$ is

$$\mathcal{I}\left(\mu,\sigma^2\right)^{-1}=\begin{pmatrix}\frac{\sigma^2}{n} & 0\\ 0 & \frac{2\sigma^4}{n}\end{pmatrix}.$$

In particular, no unbiased estimator of μ can have variance smaller than σ^2/n. This lower bound is attained by the maximum likelihood estimator $\hat{\mu}=\bar{X}$. The Cramér-Rao lower bound on the variance of unbiased estimators of σ^2 is $2\sigma^4/n$. The variance of the UMVU estimator S^2 is $2\sigma^4/\left(n-1\right)$ and the Cramér-Rao bound therefore cannot be attained.

The following theorem delivers a couple of useful implications of optimality.

Theorem. *If $\hat{\theta}_0$ is a UMVU estimator of θ and $\hat{\theta}_1\in\mathcal{W}_u\left(\theta\right)$, then*

$$Cov_\theta\left(\hat{\theta}_0,\hat{\theta}_1-\hat{\theta}_0\right)=0\qquad\forall\theta\in\Theta.$$

Proof. For any $a\in\mathbb{R}$, let $\hat{\theta}_a$ be the unbiased estimator of θ given by

$$\hat{\theta}_a=\left(1-a\right)\hat{\theta}_0+a\hat{\theta}_1=\hat{\theta}_0+a\left(\hat{\theta}_1-\hat{\theta}_0\right).$$

If $\hat{\theta}_0$ is UMVU, then $Var_\theta\left(\hat{\theta}_a\right)\geq Var_\theta\left(\hat{\theta}_0\right)$ for any $\theta\in\Theta$ and any $a\in\mathbb{R}$. For any $\theta\in\Theta$, $Var_\theta\left(\hat{\theta}_a\right)$ is a differentiable (indeed, a quadratic) function of a :

$$\begin{aligned}Var_\theta\left(\hat{\theta}_a\right) &= Var_\theta\left(\hat{\theta}_0+a\left(\hat{\theta}_1-\hat{\theta}_0\right)\right)\\ &= Var_\theta\left(\hat{\theta}_0\right)+a^2Var_\theta\left(\hat{\theta}_1-\hat{\theta}_0\right)+2a\cdot Cov_\theta\left(\hat{\theta}_0,\hat{\theta}_1-\hat{\theta}_0\right).\end{aligned}$$

Therefore, if $Var_\theta\left(\hat{\theta}_a\right)\geq Var_\theta\left(\hat{\theta}_0\right)$ for any $\theta\in\Theta$ and any $a\in\mathbb{R}$, then

$$\frac{d}{da}Var_\theta\left(\hat{\theta}_a\right)\Big|_{a=0}=2\cdot Cov_\theta\left(\hat{\theta}_0,\hat{\theta}_1-\hat{\theta}_0\right)=0\qquad\forall\theta\in\Theta.\qquad\blacksquare$$

Remark. The theorem also has a converse. Indeed, if $\hat{\theta}\in W_u\left(\theta\right)$ and

$$Cov_\theta\left(\hat{\theta}, \tilde{\theta} - \hat{\theta}\right) = 0 \qquad \forall \theta \in \Theta$$

for every $\tilde{\theta} \in W_u(\theta)$, then $\hat{\theta}$ is a UMVU estimator of θ because

$$\begin{aligned} Var_\theta\left(\tilde{\theta}\right) &= Var_\theta\left(\hat{\theta} + \tilde{\theta} - \hat{\theta}\right) \\ &= Var_\theta\left(\hat{\theta}\right) + Var_\theta\left(\tilde{\theta} - \hat{\theta}\right) + 2Cov_\theta\left(\hat{\theta}, \tilde{\theta} - \hat{\theta}\right) \\ &= Var_\theta\left(\hat{\theta}\right) + Var_\theta\left(\tilde{\theta} - \hat{\theta}\right) \\ &\geq Var_\theta\left(\hat{\theta}\right) \end{aligned}$$

whenever $Cov_\theta\left(\hat{\theta}, \tilde{\theta} - \hat{\theta}\right) = 0$.

Corollary (Casella and Berger, Theorem 7.3.20). *If $\hat{\theta}$ is a UMVU estimator of θ and U is a random variable with $E_\theta(U) = 0$ and $Var_\theta(U) < \infty$ for every $\theta \in \Theta$, then*

$$Cov_\theta\left(\hat{\theta}, U\right) = 0 \qquad \forall \theta \in \Theta.$$

Proof. Apply the theorem to $\hat{\theta}_0 = \hat{\theta}$ and $\hat{\theta}_1 = \hat{\theta} + U$. ■

Corollary. *If $\hat{\theta}$ is a UMVU estimator of θ and $\tilde{\theta} \in \mathcal{W}_u(\theta)$, then*

$$Var_\theta\left(\tilde{\theta} - \hat{\theta}\right) = Var_\theta\left(\tilde{\theta}\right) - Var_\theta\left(\hat{\theta}\right)$$

and

$$Cov_\theta\left(\hat{\theta}, \tilde{\theta}\right) = Var_\theta\left(\hat{\theta}\right)$$

for every $\theta \in \Theta$.

Proof. Now,

$$\begin{aligned} Cov_\theta\left(\hat{\theta}, \tilde{\theta}\right) &= Cov_\theta\left(\hat{\theta}, \tilde{\theta} - \hat{\theta} + \hat{\theta}\right) \\ &= Cov_\theta\left(\hat{\theta}, \tilde{\theta} - \hat{\theta}\right) + Var_\theta\left(\hat{\theta}\right) \\ &= Var_\theta\left(\hat{\theta}\right), \end{aligned}$$

where the last holds because $Cov_\theta\left(\hat{\theta}, \tilde{\theta} - \hat{\theta}\right) = 0$ in view of the theorem. Using this relation,

$$\begin{aligned} Var_\theta\left(\tilde{\theta} - \hat{\theta}\right) &= Var_\theta\left(\tilde{\theta}\right) + Var_\theta\left(\hat{\theta}\right) - 2Cov_\theta\left(\hat{\theta}, \tilde{\theta}\right) \\ &= Var_\theta\left(\tilde{\theta}\right) - Var_\theta\left(\hat{\theta}\right). \quad \blacksquare \end{aligned}$$

Corollary (Casella and Berger, Theorem 7.3.19). *UMVU estimators are essentially unique in the sense that if $\hat{\theta}$ is a UMVU estimator of θ and $\tilde{\theta} \in \mathcal{W}_u(\theta)$, then*

$$Var_\theta\left(\tilde{\theta}\right) > Var_\theta\left(\hat{\theta}\right)$$

unless $P_\theta\left(\tilde{\theta} = \hat{\theta}\right) = 1$.

Proof. If a random variable X has $Var(X) = 0$, then $P(X = E(X)) = 1$. Unbiasedness implies $E_\theta\left(\tilde{\theta} - \hat{\theta}\right) = E_\theta\left(\tilde{\theta}\right) - E_\theta\left(\hat{\theta}\right) = 0$ and it therefore suffices to show that $Var_\theta\left(\tilde{\theta}\right) > Var_\theta\left(\hat{\theta}\right)$ unless $Var_\theta\left(\tilde{\theta} - \hat{\theta}\right) = 0$. The stated result now follows from the previous corollary. $\blacksquare$