3 Point Estimation

3.1 Estimation Methods

Definition 1 (M-Estimator and Z-Estimator)**.** Suppose that data is X. An M-estimator is an estimator that maximizes some function $M_n(\theta \mid X)$ over Θ. An Z-estimator is an estimator that solves some equation $\Psi_n(\theta \mid X) = 0$.

Some interesting examples include the method of moments, maximum likelihood estimator, least squares, etc.

Definition 2 (Method of Moments)**.** Let X_1, ..., X_n be a sample from a distribution P_θ with parameter $\theta \in \Theta$. The method of moments estimates θ by solving the system of equations

$$\frac{1}{n}\sum_{i=1}^{n} f_j(X_i) = \mathrm{E}\left[f_j(X_i) \mid \theta\right], \quad j = 1, ..., k.$$

The generalized method of moments estimates θ by approximating the solution.

Example 1. Let X_1, ..., X_n be i.i.d. random variable from $N\left(\mu, \sigma^2\right)$.

1. To estimate μ, we solve $\bar{X} = \mu$.
2. To estimate σ^2, we solve $\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \sigma^2 + \mu^2$, yielding $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$.

Definition 3 (MLE)**.** Suppose that data vector X has density $p(x \mid \theta)$. The likelihood function of θ is $L(\theta \mid x) = p(x \mid \theta)$. An estimator $\hat{\theta}(X)$ is called maximum likelihood estimator (MLE) of θ, if

$$L\left(\hat{\theta}\right) = \max_{\theta \in \Theta} L(\theta).$$

Example 2. Find the MLE of θ. Let X_1, X_2, ..., X_n be iid from $N\left(\mu, \sigma^2\right)$. The likelihood is

$$L\left(\mu, \sigma^2\right) = \exp\left\{-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right\}.$$

The maximizer is $\mu = \bar{X}$ and $\sigma^2 = n^{-1}\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2$.

Example 3 (MLE may not exist)**.** Consider a Gaussian mixture

$$p(x \mid \theta) = \rho\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} + (1-\rho)\frac{1}{\sqrt{2\pi\tau^2}}\exp\left\{-\frac{(x-\nu)^2}{2\tau^2}\right\},$$

where $\theta = \left(\mu, \sigma^2, \nu, \tau^2, \rho\right)$. The likelihood function satisfies

$$\prod_{i=1}^{n} p(x_i \mid \theta) \geq \frac{\rho}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x_1-\mu)^2}{2\sigma^2}\right\}\left[\prod_{i=2}^{n}\frac{1-\rho}{\sqrt{2\pi\tau^2}}\exp\left\{-\frac{(x_i-\nu)^2}{2\tau^2}\right\}\right].$$

For the RHS, we let $\mu = x_1$ and $\sigma^2 \to 0$. Then corresponding supremum of RHS is ∞. Hence, the MLE does not exist.

Example 4. We can also specify a loss function Q and minimize the loss function. For example,

$$Q = \sum_{i=1}^{n} (y_i - \theta_1 - \theta_2 x_i)^2$$

$$Q = \sum_{i=1}^{n} (y_i - \theta_1 - \theta_2 x_i) \left[\tau - 1 (y_i - \theta_1 - \theta_2 x_i < 0)\right], \quad \text{known } \tau,$$

$$Q = \sum_{i=1}^{n} \max \{0, \ 1 - y_i (\theta_1 + \theta_2 x_i)\}.$$

The M-estimator and Z-estimator are often related.

Example 5. The method of moment estimator is a Z-estimator

$$n^{-1} \sum_{i=1}^{n} f(X_i) - \mathrm{E}\left[f_j(X_i) \mid \theta\right] f = 0.$$

If the likelihood function is differentiable, then MLE is a Z-estimator

$$n^{-1} \sum_{i=1}^{n} \frac{\partial \log p (X_i \mid \theta)}{\partial \theta} = 0.$$

3.2 Maximum Likelihood

A very useful property of the MLE is its invariance property. Let $\gamma = g(\theta)$, not necessarily one-to-one. We define MLE as $\hat{\gamma}$ that maximizes

$$L^*(\gamma) = \sup_{\{\theta : g(\theta) = \gamma\}} L(\theta).$$

Theorem 1. *For any function $g(\theta)$, the MLE of $g(\theta)$ is $g\left(\hat{\theta}_{MLE}\right)$.*

Example 6. Let X_1, ..., X_n be i.i.d. random variable from Bernoulli (θ). The MLE of θ is $\hat{\theta} = \bar{X}$. The MLE of $\gamma = \theta / (1 - \theta)$ is $\hat{\gamma} = \bar{X} / \left(1 - \bar{X}\right)$.

Definition 4 (Score Function)**.** Suppose that the log-likelihood $\ell(\theta \mid x) = \log p(x \mid \theta)$ is well defined and the derivative with respect to θ exists. For every $x \in \mathcal{X}$, the score function is defined to be

$$V(\theta; x) = \frac{\partial \ell (\theta \mid x)}{\partial \theta}.$$

We need introduce following regularity conditions. Let $p(x \mid \theta)$ be the density.

R1 The distributions $\{\mathrm{P}_\theta : \theta \in \Theta\}$ have a common support, so that the set $\mathcal{X} = \{x : p(x \mid \theta) > 0\}$ is independent of θ.

R2 The dimension of θ is k and the parameter space $\Theta \subseteq \mathbb{R}^k$ is an open set.

R3 For any $x \in \mathcal{X}$ and all $\theta \in \Theta$, the partial derivatives $\frac{\partial p(x|\theta)}{\partial \theta_j}$ exist and satisfy

$$\frac{\partial}{\partial \theta} \int_{\mathcal{X}} p(x \mid \theta) \, d\mu(x) = \int_{\mathcal{X}} \frac{\partial p(x \mid \theta)}{\partial \theta} d\mu(x).$$

R4 For any $x \in \mathcal{X}$ and all $\theta \in \Theta$, the partial derivatives $\frac{\partial^2 p(x|\theta)}{\partial\theta_i \partial\theta_j}$ exist and satisfy

$$\frac{\partial^2}{\partial\theta\partial\theta^T} \int_{\mathcal{X}} p(x \mid \theta) \, d\mu(x) = \int_{\mathcal{X}} \frac{\partial^2 p(x \mid \theta)}{\partial\theta\partial\theta^T} d\mu(x).$$

Theorem 2. *Under the regularity conditions R1, R2, and R3, we have*

$$E\left[\frac{\partial \ell(\theta \mid X)}{\partial\theta} \mid \theta\right] = 0, \text{ for all } \theta \in \Theta,$$

where $\ell(\theta \mid X) = \log p(X \mid \theta)$ *and the expectation is taken to the distribution where the probability function of* X *is* $p(X \mid \theta)$.

Definition 5 (Fisher Information)**.** Suppose that the conditions R1, R2, and R3 are satisfied. The Fisher information is defined to be

$$\mathcal{I}(\theta) = \text{Cov}\left[\frac{\partial \ell(\theta \mid X)}{\partial\theta}\right] = \text{Cov}\left[\frac{\partial \ell(\theta \mid X)}{\partial\theta}\left(\frac{\partial \ell(\theta \mid X)}{\partial\theta}\right)^T\right],$$

as a $k \times k$ matrix, where the (i,j)th element of $\mathcal{I}(\theta)$ is

$$\text{Cov}\left[\frac{\partial \ell(\theta \mid X)}{\partial\theta_i}, \frac{\partial \ell(\theta \mid X)}{\partial\theta_j}\right].$$

Theorem 3 (Fisher Information, Equivalent Form)**.** *Under the regularity conditions R1, R2, R3, and R4, then*

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2 \ell(\theta \mid X)}{\partial\theta\partial\theta^T}\right],$$

where $\ell(\theta \mid X) = \log p(X \mid \theta)$, $\frac{\partial^2 \ell(\theta|X)}{\partial\theta\partial\theta^T}$ *is the Hessian matrix, and the expectation is taken to the distribution where the probability function of* X *is* $p(X \mid \theta)$.

The Fisher information $\mathcal{I}(\theta)$ is often called the expected information. The observed information is

$$J(\theta) = -\frac{\partial^2 \ell(\theta \mid X)}{\partial\theta\partial\theta^T}.$$

Example 7. Let X_1, ..., X_n be an i.i.d. sample from $N(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2) \in \mathbb{R} \otimes \mathbb{R}_+$. The log-likelihood is

$$\ell(\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Then,

$$\ell' = \frac{\sum_{i=1}^{n}(x_i - \mu)}{\sigma^2},$$
$$\ell'' = -\frac{n}{\sigma^2}.$$

Hence,

$$V(\ell') = V\left[\frac{\sum_{i=1}^{n}(x_i - \mu)}{\sigma^2}\right] = \frac{n}{\sigma^2},$$
$$-E[\ell''] = \frac{n}{\sigma^2}.$$

The Fisher information is

$$\begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}.$$

Definition 6 (Kullback-Leibler Divergence)**.** Suppose that P and Q are two probability measures with densities p and q, respectively. The Kullback-Leibler divergence between them is

$$\mathrm{KL}(P,Q) \quad = \quad \int \log\left[\frac{p(x)}{q(x)}\right] p(x)\, d\mu(x)\,.$$

Theorem 4 (Kullback-Leibler Inequality)**.** *The Kullback-Leibler divergence satisfies* $KL(P,Q) \geq 0$ *where the equality holds if and only if* $p = q$ *almost everywhere (under* μ*).*

Proof. Note that $-\log(\cdot)$ is strictly convex. Then, Jensen's inequality implies that

$$\begin{aligned}
\int\limits_{p(x)>0} -\log\left[\frac{q(x)}{p(x)}\right] p(x)\, d\mu(x) \quad &= \quad \mathrm{E}\left[-\log\left[\frac{q(x)}{p(x)}\right] \mid p(x)\right] \\
&\geq \quad -\log\left\{\mathrm{E}\left[\frac{q(x)}{p(x)} \mid p(x)\right]\right\} = -\log\left\{\int\limits_{p(x)>0} \frac{q(x)}{p(x)} p(x)\, d\mu(x)\right\} \\
&= \quad -\log\left\{\int\limits_{p(x)>0} q(x)\, d\mu(x)\right\} \\
&\geq \quad -\log\left\{\int\limits_{\mathcal{X}} q(x)\, d\mu(x)\right\} = 0.
\end{aligned}$$

The equality of Jensen's inequality holds if and only if $\frac{q(x)}{p(x)} =$ constant with probability 1 under density $p(x)$. Since both are density functions, such constant must be 1. $\square$

Note that

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} p(x_i \mid \theta)\right] \quad = \quad \mathrm{E}\left[p(x_i \mid \theta)\right].$$

We expect the MLE minimizes the Kullback-Leibler divergence to the truth.

3.3 UMVUE

Definition 7 (Unbiased Estimator)**.** The bias of the estimator $\delta(X)$ of $g(\theta)$ is

$$\mathrm{Bias}(T, g(\theta)) \quad = \quad \mathrm{E}[T] - g(\theta)\,.$$

The estimator is unbiased for $g(\theta)$ if $\mathrm{Bias}(T, g(\theta)) = 0$ for all $\theta \in \Theta$.

Example 8. Suppose that $\mu = \mathrm{E}(X) < \infty$ and $\sigma^2 < \infty$. Then, $\mathrm{E}\left[\bar{X} \mid \mu\right] = \mu$, unbiased estimator $\forall \mu$. But $\bar{X}^2$ is not an unbiased estimator of μ^2, since

$$\mathrm{E}\left[\left(\bar{X}\right)^2 \mid \mu\right] \quad = \quad \mathrm{Var}\left[\bar{X} \mid \mu\right] + \left(\mathrm{E}\left[\bar{X} \mid \mu\right]\right)^2 = \frac{\sigma^2}{n} + \mu^2 \neq \mu^2.$$

But the bias is low for large enough n.

Definition 8 (UMVUE)**.** An unbiased estimator $\delta(X)$ of $g(\theta)$ is uniformly minimum variance unbiased (UMVUE) if $\mathrm{Var}[\delta(X) \mid \theta] - \mathrm{Var}[\delta^*(X) \mid \theta] \leq 0$, $\forall \theta \in \Theta$, for any other unbiased estimator $\delta^*(X)$.

Theorem 5 (Rao-Blackwell Theorem)**.** *Let* T *be a sufficient statistic for* $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$*. Let* δ *be an unbiased estimator of* $g(\theta)$*. Define* $\eta(T) = E[\delta(X) \mid T]$*. Then,* $\eta(T)$ *does not depend on* θ*. Furthermore, for all* $\theta \in \Theta$*, we have* $E[\eta(T)] = g(\theta)$ *and* $Cov[\delta] - Cov[\eta] \geq 0$ *(positive semi-definite matrix).*

Proof. Since T is a sufficient statistic, $p(x \mid T)$ does not depend on θ. Then, $\hat{\gamma}(T)$ does not involve θ and we can use it as an estimator. Note that

$$\mathrm{E}[\eta(T)] = \mathrm{E}\{\mathrm{E}[\delta(X) \mid T]\} = \mathrm{E}[\delta(X)] = g(\theta),$$

since $\eta(T) = \mathrm{E}[\delta(X) \mid T]$ by definition. Hence, $\eta(T)$ is also unbiased for $g(\theta)$.

Further,

$$\begin{aligned}
\mathrm{Cov}[\delta] &= \mathrm{E}\left[(\delta - g(\theta))(\delta - g(\theta))^T\right] \\
&= \mathrm{E}\left[(\delta - \eta + \eta - g(\theta))(\delta - \eta + \eta - g(\theta))^T\right] \\
&= \mathrm{E}\left[(\delta - \eta)(\delta - \eta)^T\right] + \mathrm{E}\left[(\delta - \eta)(\eta - g(\theta))^T\right] + \mathrm{E}\left[(\eta - g(\theta))(\delta - \eta)^T\right] + \mathrm{E}\left[(\eta - g(\theta))(\eta - g(\theta))^T\right].
\end{aligned}$$

Since

$$\begin{aligned}
\mathrm{E}\left[(\delta - \eta)(\eta - g(\theta))^T \mid T\right] &= \mathrm{E}[\delta - \eta \mid T](\eta - g(\theta))^T = (\mathrm{E}[\delta \mid T] - \eta)(\eta - g(\theta))^T = 0, \\
E(\tilde{\gamma} - \hat{\gamma} \mid T) &= E(\tilde{\gamma} \mid T) - E(\hat{\gamma} \mid T) = E(\tilde{\gamma} \mid T) - \hat{\gamma} = 0
\end{aligned}$$

by definition of $\eta(T)$, we have

$$\mathrm{E}\left[(\delta - \eta)(\eta - g(\theta))^T\right] = \mathrm{E}\left\{\mathrm{E}\left[(\delta - \eta)(\eta - g(\theta))^T \mid T\right]\right\} = 0.$$

Likewise, $\mathrm{E}\left[(\eta - g(\theta))(\delta - \eta)^T\right] = 0$. Hence,

$$\mathrm{Cov}[\delta] = \underbrace{\mathrm{E}\left[(\delta - \eta)(\delta - \eta)^T\right]}_{\geq 0} + \underbrace{\mathrm{E}\left[(\eta - g(\theta))(\eta - g(\theta))^T\right]}_{=\mathrm{Cov}[\eta]}.$$

□

An issue is that if we have $\eta(T) = \mathrm{E}[\delta(X) \mid T]$. Now we consider another sufficient statistic S, but $S \neq T$. Let $\mathrm{E}[\eta(T) \mid S]$. Can we improve $\eta(T)$? Consider the special case where T is minimal sufficient and S is any sufficient statistic. Then, T is a function of S, such as $T = T(S)$. Then,

$$\mathrm{E}[\eta(T) \mid S] = \mathrm{E}[\eta(T(S)) \mid S] = \eta(T(S)) = \eta(T),$$

that is, no further improvements. But this discussion only means that we cannot improve δ further. It does not mean that we cannot improve another estimator δ^* such that $\mathrm{E}[\delta^*(X) \mid T]$ is better than $\mathrm{E}[\delta(X) \mid T]$.

In order to make sure no further improvements can be done and to find the UMVUE, we can use the following theorem.

Theorem 6 (Lehmann-Scheffé Theorem)**.** *Let T be a complete and sufficient statistic for a parameter θ. Let $\delta(X)$ be any unbiased estimator of $g(\theta)$. Then $E[\delta(X) \mid T]$ is the unique UMVUE of $g(\theta)$. [The theorem is stated for variance, it can be easily extended to convex loss functions. We will consider it in decision theory part.]*

Proof. Let $\delta(X)$ be any unbiased estimator of $g(\theta)$ and define $\eta(T) = \mathrm{E}[\delta(X) \mid T]$. Since T is a function of X, we have the law of iterated expectation

$$\begin{aligned}
\mathrm{E}[\eta(T) \mid \theta] &\overset{\text{definition of } \eta}{=} \mathrm{E}[\mathrm{E}[\delta(X) \mid T] \mid \theta] \\
&= \mathrm{E}[\mathrm{E}[\delta(X) \mid T, \theta] \mid \theta] = \mathrm{E}[\delta(X) \mid \theta] = g(\theta),
\end{aligned}$$

where the first equality in the second line holds since T is sufficient [the distribution of $\delta(X) \mid T$ does not depend on θ, so conditioning on θ or not are the same.]. This means that $\eta(T)$ is also an unbiased estimator of $g(\theta)$. By the Rao-Blackwell theorem, we know that $\mathrm{Cov}(\eta(T)) - \mathrm{Cov}(\delta(X)) \leq 0$ for any unbiased estimator $\delta(X)$ of $g(\theta)$.

Suppose that $\eta^*(T)$ is another unbiased estimator of $g(\theta)$. Then,

$$\mathrm{E}[\eta(T) - \eta^*(T) \mid \theta] = 0, \quad \forall \theta \in \Theta.$$

Since both η and η^* are statistics, we can define $h(T) = \eta(T) - \eta^*(T)$. This means that, by completeness of T, we must have $\mathrm{P}[\eta(T) - \eta^*(T) \mid \theta] = 1$. $\square$

Example 9. Consider X_1, ..., X_n from $\mathrm{Bernoulli}(\theta)$. Note that

$$p(X \mid \theta) = \prod_{i=1}^{n} \theta^{X_i} (1-\theta)^{1-X_i} = \exp\left\{\sum_i X_i \log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right\}$$

Hence, $T = \sum_i X_i$ is sufficient and complete. Note that $\mathrm{E}\left[n^{-1}T \mid \theta\right] = \theta$. Hence, $\bar{X}$ is the unique UMVUE of θ.

The role of completeness in the theorem is important. Let T be a minimal sufficient statistic. Let Z_1 be an unbiased estimator of θ. Then, the Rao-Blackwell theorem says that, if $\eta_1(T) = \mathrm{E}[Z_1 \mid T]$, then $\mathrm{Var}[\eta_1(T) \mid \theta] \leq \mathrm{Var}[Z_1 \mid \theta]$. Since T is minimal sufficient, it is a function of any other sufficient statistic. Hence, for any sufficient statistic S, we have

$$\mathrm{E}[\eta_1(T) \mid S] = \mathrm{E}[\eta_1(T(S)) \mid S] = \eta_1(T(S)) = \eta_1(T),$$

that is, conditional on any sufficient statistic will not further improve $\eta_1(T)$ [the best estimator that we can derive from Z_1].

Suppose that T is not complete. Consider another unbiased estimator Z_2. Then, we can obtain $\eta_2(T) = \mathrm{E}[Z_2 \mid T]$ that is the best estimator that we can derive from Z_2. Consider a new estimator

$$U = \frac{1}{2}[\eta_1(T) + \eta_2(T)].$$

Then,

$$\begin{aligned} \mathrm{Var}[U \mid \theta] &= \frac{1}{4}\mathrm{Var}[\eta_1(T) \mid \theta] + \frac{1}{2}\mathrm{Cov}[\eta_1(T), \eta_2(T) \mid \theta] + \frac{1}{4}\mathrm{Var}[\eta_2(T) \mid \theta] \\ &= \frac{1}{4}\underbrace{\mathrm{Var}[\eta_1(T) \mid \theta]}_{\equiv v_1} + \frac{\rho}{2}\sqrt{\underbrace{\mathrm{Var}[\eta_1(T) \mid \theta]}_{\equiv v_1}\underbrace{\mathrm{Var}[\eta_2(T) \mid \theta]}_{\equiv v_2}} + \frac{1}{4}\underbrace{\mathrm{Var}[\eta_2(T) \mid \theta]}_{\equiv v_2}, \end{aligned}$$

where ρ is the correlation between random variables $\eta_1(T)$ and $\eta_2(T)$.

1. Suppose that we can find an estimator Z_2 such that $v_1 \neq v_2$. Without loss of generality, we assume $v_1 < v_2$. Hence, together with $\rho \leq 1$, we have

$$\begin{aligned} \mathrm{Var}[U \mid \theta] &= \frac{1}{4}(v_1 + v_2) + \frac{\rho}{2}\sqrt{v_1 v_2} \\ &< \frac{1}{4}(v_2 + v_2) + \frac{1}{2}\sqrt{v_2^2} = v_2. \end{aligned}$$

   This means that $\eta_2(T) = \mathrm{E}[Z_2 \mid T]$ is the best that we can do if we start with Z_2. But we cannot guarantee that $\eta_2(T)$ is universally the best.

2. Even though we find an estimator Z_2 such that $v_1 = v_2$, we cannot guarantee that there is no Z_3 that makes $v_3 < v_1 = v_2$. For example, if $v_1 = v_2$, then

$$\mathrm{Var}[U \mid \theta] = \frac{1}{2}(1+\rho)v_1.$$

   If $\rho < 1$, then $\mathrm{Var}[U \mid \theta] < v_1 = v_2$.

Example 10. Consider X_1, ..., X_n from $N\left(\mu,\sigma^2\right)$, where both μ and σ^2 are unknown. Note that

$$\begin{aligned} p\left(X \mid \theta\right) &= \exp\left\{-\frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right\}\frac{1}{\left(2\pi\right)^{n/2}} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}X_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n}X_i - \left[\frac{n\mu^2}{2\sigma^2}+\frac{n}{2}\log\left(\sigma^2\right)\right]\right\}\frac{1}{\left(2\pi\right)^{n/2}} \end{aligned}$$

Hence, $\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ are sufficient, minimal sufficient, and complete, so as $\left(\bar{X}, \sum_{i=1}^n X_i^2\right)$.

1. To estimate σ^2, we note that

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

   as a function of $\left(\bar{X}, \sum_{i=1}^n X_i^2\right)$, and

$$\mathrm{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 \mid \theta\right] = \frac{1}{n-1}\mathrm{E}\left[\sum_{i=1}^{n}X_i^2 - n\left(\bar{X}\right)^2 \mid \theta\right] = \frac{n\left(\sigma^2+\mu^2\right)-n\left(\sigma^2/n+\mu^2\right)}{n-1} = \sigma^2.$$

   Hence, $\frac{1}{n-1}\sum_{i=1}^n\left(X_i-\bar{X}\right)^2$ is the UMVUE.

2. To estimate μ, we note that $\mathrm{E}\left[\bar{X} \mid \theta\right] = \mu$. Hence, $\bar{X}$ is the UMVUE of μ.

3. To estimate μ^2, we note that

$$\mathrm{E}\left[\left(\bar{X}\right)^2 - \frac{1}{n\left(n-1\right)}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 \mid \theta\right] = \frac{\sigma^2}{n} + \mu^2 - \frac{\sigma^2}{n} = \mu^2.$$

   Hence, the UMVUE is $\left(\bar{X}\right)^2 - \frac{1}{n(n-1)}\sum_{i=1}^n\left(X_i-\bar{X}\right)^2$.

3.4 Efficiency

Definition 9 (Regular Estimator)**.** Let $\mathcal{A} = \{x : p\left(x \mid \theta\right) > 0\}$ be the common support of the probability measures P_θ of the underlying family $\mathcal{P}$. The estimator T is a regular estimator if

$$\frac{\partial}{\partial\theta}\int_{\mathcal{A}} T\left(x\right)L\left(\theta;x\right)d\mu\left(x\right) = \int_{\mathcal{A}} T\left(x\right)\frac{\partial L\left(\theta;x\right)}{\partial\theta}d\mu\left(x\right).$$

Theorem 7 (Cramer-Rao lower bound)**.** *Let $X = \left(X_1,...,X_n\right)$ be a sample from $\{P_\theta, \theta\in\Theta\}$, where Θ is an open set in $\mathbb{R}^k$. Suppose that*

1. *the joint distribution of X has a density $p\left(x \mid \theta\right)$ with respect to a measure μ for all $\theta\in\Theta$,*

2. *$p\left(x \mid \theta\right)$ is differentiable as a function of θ and*

$$\frac{\partial}{\partial\theta}\int p\left(x \mid \theta\right)d\mu\left(x\right) = \int \frac{\partial p\left(x \mid \theta\right)}{\partial\theta}d\mu\left(x\right).$$

3. *$T\left(X\right)$ is a regular estimator with $E\left[T\left(X\right) \mid \theta\right] = g\left(\theta\right)$, where $g\left(\theta\right)$ is a differentiable function of θ,*

Then,

$$Var\left(T\left(X\right)\right) \geq \frac{\partial g\left(\theta\right)}{\partial\theta^T}\left[\mathcal{I}\left(\theta\right)\right]^{-1}\left(\frac{\partial g\left(\theta\right)}{\partial\theta^T}\right)^T,$$

where the Fisher information

$$\mathcal{I}\left(\theta\right) = E\left[\frac{\partial\log p\left(x\mid\theta\right)}{\partial\theta}\frac{\partial\log p\left(x\mid\theta\right)}{\partial\theta^T}\mid\theta\right]$$

is assumed to be positive definite for all $\theta\in\Theta$.

Definition 10 (Efficiency)**.** The efficiency of an unbiased estimator T is the ratio of its variance and the Cramér-Rao lower bound, that is,

$$e\left(T,\theta\right) = \frac{\left[g'\left(\theta\right)\right]^2/\mathcal{I}\left(\theta\right)}{\mathrm{Var}\left(T\right)}.$$

An unbiased estimator which attains the Cramér-Rao lower bound is called an efficient estimator. An efficient estimator is also a UMVUE.

Example 11. Consider the iid Bernoulli example again. The log-likelihood is

$$\log p\left(x\mid\theta\right) = \sum_i X_i\log\theta+\left(n-\sum_i X_i\right)\log\left(1-\theta\right).$$

Consider the statistic $T=\bar{X}$. Then,

$$\begin{aligned}\frac{d\mathrm{E}\left[T\left(X\right)\mid\theta\right]}{d\theta} &= \frac{d}{d\theta}\theta=1,\\ \mathrm{E}\left[\left(\frac{\partial\log p\left(x\mid\theta\right)}{\partial\theta}\right)^2\mid\theta\right] &= \mathrm{E}\left[\left(\frac{n\bar{X}}{\theta}-\frac{n-n\bar{X}}{1-\theta}\right)^2\mid\theta\right]=\mathrm{E}\left[\left(\frac{n\left(\bar{X}-\theta\right)}{\theta\left(1-\theta\right)}\right)^2\mid\theta\right]\\ &= \frac{n}{\theta\left(1-\theta\right)}.\end{aligned}$$

Hence, $\bar{X}$ is the UMVUE of θ since $\mathrm{E}\left[\bar{X}\mid\theta\right]=\theta$ and $\mathrm{Var}\left[\bar{X}\mid\theta\right]=\theta\left(1-\theta\right)/n$ attains the Cramér-Rao lower bound.

Example 12. Let X_1, ..., X_n be i.i.d. random variable from $N\left(\mu,\sigma^2\right)$ where $\theta=\left(\mu,\sigma^2\right)$. We have shown that, by the Lehmann-Scheffé Theorem, $\left(\bar{X},S^2\right)$ is the UMVUE. However,

$$\mathrm{Var}\begin{bmatrix}\bar{X}\\ S^2\end{bmatrix} = \begin{bmatrix}\frac{\sigma^2}{n} & 0\\ 0 & \frac{2\sigma^4}{n-1}\end{bmatrix}.$$

We have also shown that

$$\mathcal{I}\left(\theta\right) = \begin{bmatrix}\frac{n}{\sigma^2} & 0\\ 0 & \frac{n}{2\sigma^4}\end{bmatrix}.$$

Hence, $\mathrm{Var}\begin{bmatrix}\bar{X}\\ S^2\end{bmatrix}-\mathcal{I}^{-1}\left(\theta\right)\geq 0$. Hence, the UMVUE is not necessarily efficient.

Corollary 1. *Suppose that the assumptions in Cramér-Rao lower bound (Theorem 7) hold, and that R4 holds. If T is a regular unbiased estimator for $\gamma = g(\theta)$ and $\frac{\partial g(\theta)}{\partial \theta^T}$ is invertible, then the Cramér-Rao lower bound is attained if and only if*

$$A(\theta)[T(x) - g(\theta)] = V(\theta; x),$$

for some function $A(\theta)$.

Example 13. Let X_1, ..., X_n be i.i.d. random variable from Bernoulli(θ) where $\theta \in (0,1)$. Then,

$$p(x \mid \theta) = \exp\left\{\sum_{i=1}^{n} x_i \log\theta + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-\theta)\right\}$$

and

$$\frac{d \log p(x \mid \theta)}{d\theta} = \frac{n}{\theta(1-\theta)}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \theta\right).$$

Hence $\bar{X}$ is the efficient estimator of θ.

3.5 Mean Squared Error

Definition 11 (Mean Squared Error)**.** Let $\mathcal{P} = \{\mathrm{P}_\theta : \theta \in \Theta\}$ be a statistical model for a random variable X on $\mathcal{X}$, and $g : \Theta \to \Gamma$ a function, and $T : \mathcal{X} \to \Gamma$ an estimator for $\gamma = g(\theta)$. The mean squared error (MSE) of T is given by

$$\mathrm{MSE}(T, g(\theta)) = \mathrm{E}\left\{[T - g(\theta)]^T [T - g(\theta)]\right\}.$$

The bias-variance decomposition means that we can decompose MSE as

$$\mathrm{MSE}(T, g(\theta)) = [\mathrm{Bias}(T,\theta)]^T \mathrm{Bias}(T,\theta) + \mathrm{Var}(T).$$

This decomposition holds since

$$\begin{aligned}
\mathrm{MSE}(T, g(\theta)) &= \mathrm{E}\left[(T - \mathrm{E}(T) + \mathrm{E}(T) - g(\theta))^T (T - \mathrm{E}(T) + \mathrm{E}(T) - g(\theta))\right] \\
&= \mathrm{E}\left[(T - \mathrm{E}(T))^T (T - \mathrm{E}(T))\right] + 2\mathrm{E}\left[(T - \mathrm{E}(T))^T (\mathrm{E}(T) - g(\theta))\right] \\
&\quad + \mathrm{E}\left[(\mathrm{E}(T) - g(\theta))^T (\mathrm{E}(T) - g(\theta))\right] \\
&= \mathrm{tr}\Bigg\{\underbrace{\mathrm{E}\left[(T - \mathrm{E}(T))(T - \mathrm{E}(T))^T\right]}_{=\mathrm{Var}(T)}\Bigg\} + \underbrace{(\mathrm{E}(T) - g(\theta))^T (\mathrm{E}(T) - g(\theta))}_{[\mathrm{Bias}(T,\theta)]^T \mathrm{Bias}(T,\theta)}.
\end{aligned}$$

Alternatively, we can compute

$$\begin{aligned}
\mathrm{E}\left\{[T - g(\theta)][T - g(\theta)]^T\right\} &= \mathrm{E}\left[(T - \mathrm{E}(T) + \mathrm{E}(T) - g(\theta))(T - \mathrm{E}(T) + \mathrm{E}(T) - g(\theta))^T\right] \\
&= \mathrm{E}\left[(T - \mathrm{E}(T))(T - \mathrm{E}(T))^T\right] + \mathrm{E}\left[(\mathrm{E}(T) - g(\theta))(T - \mathrm{E}(T))^T\right] \\
&\quad + \mathrm{E}\left[(T - \mathrm{E}(T))(\mathrm{E}(T) - g(\theta))^T\right] + (\mathrm{E}(T) - g(\theta))(\mathrm{E}(T) - g(\theta))^T \\
&= \mathrm{Var}(T) + \mathrm{Bias}(T,\theta)[\mathrm{Bias}(T,\theta)]^T.
\end{aligned}$$

This corresponds to the bias-variance trade-off: a complicated model typically has a low bias but a large bias, whereas a simple model typically has a large bias but a low variance.

- Large bias means that in darts, all darts are far away from the bullseye.
- Small variance means that all darts landed very concentrated.
- Darts spread everywhere on the dartboard can have a larger MSE comparing to concentrated at certain region but never reach the bullseye.

Example 14. Let X_1, ..., X_n be i.i.d. random variable from Uniform $(0, \theta)$. The MLE $X_{(n)} = \max\{X_1, ..., X_n\}$ is not an unbiased estimator, since

$$\mathrm{E}\left[X_{(n)} \mid \theta\right] = \frac{n}{n+1}.$$

An unbiased estimator is

$$\frac{n+1}{n} X_{(n)}.$$

But its variance is

$$\mathrm{Var}\left[\frac{n+1}{n} X_{(n)} \mid \theta\right] = \frac{\theta^2}{n(n+2)}.$$

The MSE of the MLE satisfies

$$\mathrm{MSE}\left(X_{(n)}, \theta\right) = \frac{2\theta^2}{(n+1)^2(n+2)} < \frac{\theta^2}{n(n+2)} = \mathrm{Var}\left[\frac{n+1}{n} X_{(n)} \mid \theta\right].$$