MATH 5010 –Foundations of Statistical Theory and Probability

❖ Point Estimation 2

-Evaluating Estimators

Instructor: He Wang
Department of Mathematics
Northeastern University

Evaluating and Comparing Estimators

Statistical inference Question: Choosing among different point estimators when multiple methods are available, e.g., Method of Moments, MLE, Bayes methods, etc.

Some criteria for a "good" estimator, such as:

- Mean squared error (MSE)
- Unbiasedness
- Efficiency
- Sufficiency
- Consistency

Mean squared error (MSE)

Definition: The **Mean Squared Error (MSE)** of an estimator $W = W(X_1, \ldots, X_n)$ of a parameter θ is the function of θ defined by

$$MSE := E_\theta(W - \theta)^2$$

MSE measures the **average squared difference** between the estimator W and the parameter θ.

Definition: The **bias** of a point estimator W of a parameter θ is the difference between the expected value of $W\ and\ \theta$.

$$Bias_\theta(W) \coloneqq E_\theta[W] - \theta$$

An estimator is called **unbiased** if $E_\theta[W] = \theta$ for all θ.

Bias-Variance decomposition

$$E_\theta(W-\theta)^2 = Var_\theta(W) + (E_\theta W - \theta)^2 = Var_\theta(W) + [Bias_\theta(W)]^2$$

Remark: It is reasonable to consider other errors, for example,

$$E_\theta|W-\theta|$$

However, MSE is easier for computation and has the bias-variance decomposition.

Recall the results: Let $\{X_1, \ldots, X_n\}$ be a random sample from a population with mean μ and variance σ^2.

1.) $E[\bar{X}] = \mu$

2.) $Var(\bar{X}) = \frac{\sigma^2}{n}$

3.) $E[S^2] = \sigma^2$

So, both sample mean $\bar{X}$ and sample variance S^2 are unbiased estimators.

Example. Normal Distribution.

Suppose sample $X_1, \ldots, X_n \sim Normal(\mu, \sigma^2)$

The MSE of $\bar{X}$ and S^2 are:

$$\mathrm{E}(\bar{X} - \mu)^2 = \mathrm{Var}\,\bar{X} = \frac{\sigma^2}{n},$$

$$\mathrm{E}(S^2 - \sigma^2)^2 = \mathrm{Var}\,S^2 = \frac{2\sigma^4}{n-1}.$$

Example. Normal Distribution (MLE).

The MLE of σ^2 is

$$\widehat{\sigma^2}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n-1}{n}S^2$$

Then,

$$E\left[\widehat{\sigma^2}_{MLE}\right] = E\left[\frac{n-1}{n}S^2\right] = \frac{n-1}{n}\sigma^2$$

The variance of $\widehat{\sigma^2}_{MLE}$ is

$$Var\left(\widehat{\sigma^2}_{MLE}\right) = Var\left(\frac{n-1}{n}S^2\right) = \left(\frac{n-1}{n}\right)^2 Var(S^2) = \frac{2(n-1)\sigma^4}{n^2}$$

MSE of $\widehat{\sigma^2}_{MLE}$ is

$$MSE\left(\widehat{\sigma^2}_{MLE}\right) = E\left(\widehat{\sigma^2}_{MLE} - \sigma^2\right)^2 = \frac{2(n-1)\sigma^4}{n^2} + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4$$

So

$$MSE\left(\widehat{\sigma^2}_{MLE}\right) \leq MSE(S^2)$$

By trading off variance for bias, the MSE is improved.

Example: (Binomial Bayes Estimator)

Suppose random sample $X_i \sim Bernoulli(p)$ for $i = 1, \dots, n$ with unknown p.

The MSE of $\bar{X}$ as an estimator of p is

$$E_p[(\bar{X} - p)^2] = \mathrm{Var}_p(\bar{X}) = \frac{p(1-p)}{n}$$

Let $S = X_1 + \cdots + X_n$

$$\hat{p}_B := E[p|S] = \frac{\alpha + S}{\alpha + \beta + n}$$

$$E_p[(\hat{p}_B - p)^2] = Var_p \hat{p}_B + \left(Bias_p \hat{p}_B\right)^2$$

$$= Var_p\left(\frac{\alpha + S}{\alpha + \beta + n}\right) + \left(E\left(\frac{\alpha + S}{\alpha + \beta + n}\right) - p\right)^2$$

$$= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2$$

Choose $\alpha \ and \ \beta$ to make the MSE of $\hat{p}_B$ constant.

Choose $\alpha = \beta = \sqrt{\frac{n}{4}}$

$$\hat{p}_B = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}} \quad \text{and} \quad \mathrm{E}(\hat{p}_B - p)^2 = \frac{n}{4(n+\sqrt{n})^2}$$

Compare $\hat{p}_{MLE}$ and $\hat{p}_B$

For large n, $\hat{p}_{MLE}$ is better and for small n, $\hat{p}_B$ is better.

If there is a strong belief that p is close to $\frac{1}{2}$, $\hat{p}_B$ will be a good choice.

Best Unbiased Estimators

An estimator $\hat{\theta}$ of a parameter θ is called the **best unbiased estimator** if:

- **Unbiased**: $E\left(\hat{\theta}\right) = \theta$
- **Minimum Variance**: Among all unbiased estimators of θ, $\hat{\theta}$ has the lowest variance for all possible values of θ.

This estimator is also called Uniform Minimum Variance Unbiased Estimator (UMVUE)

The definition can be generalized to a Best Unbiased Estimator of $g(\theta)$.

Theorem: If W is a best unbiased estimator of θ, then W is unique.

Example: Poisson Unbiased Estimation.

Suppose random sample $X_i \sim Poisson(\lambda)$ for $i = 1, \ldots, n$ with unknown λ.

$$E_\lambda(\bar{X}) = \lambda$$

$$E_\lambda(S^2) = \lambda$$

Both $\bar{X}$ and S^2 are unbiased estimator for λ.

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

Compute $Var(S^2)$ next.

$$\mathrm{Var}(S^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right)$$

where, $\mu_4 \coloneqq E[(X_i - \mu)^4]$ is the fourth central moment, and $\sigma^2 = Var(X_i)$.

For Poisson distribution:

$$\mu = \lambda$$
$$\sigma^2 = \lambda$$
$$\mu_4 = \lambda + 3\lambda^2$$

So,

$$\operatorname{Var}(S^2) = \frac{1}{n}\left(\lambda + 3\lambda^2 - \frac{n-3}{n-1}\lambda^2\right)$$

$$= \frac{1}{n}\left(\lambda + \lambda^2\left[3 - \frac{n-3}{n-1}\right]\right)$$

$$= \frac{\lambda}{n} + \frac{2\lambda^2}{n-1}$$

$$MSE(S^2) = Var(S^2) = \frac{\lambda}{n} + \frac{2\lambda^2}{n-1}$$

So,

$$MSE(\bar{X}) < MSE(S^2)$$

Up to now, we can not claim that $\bar{X}$ is the best unbiased estimator, sine there are many other unbiased estimators.

Theorem(Cramér–Rao Inequality)

Let $\{X_1, \ldots, X_n\}$ be a sample from a population distribution with pdf $f(x|\theta)$

Let $W(\vec{X})$ be an estimator for $g(\theta)$ satisfying

$$\frac{d}{d\theta} E_\theta\left[W\left(\vec{X}\right)\right] = \int \frac{\partial}{\partial\theta}[W(\vec{x}) f(\vec{x}|\theta)] d\vec{x}$$

and

$$Var_\theta\left(W\left(\vec{X}\right)\right) > \infty$$

Then

$$Var_\theta\left(W\left(\vec{X}\right)\right) \geq \frac{\left(\frac{d}{d\theta} E_\theta\left[W\left(\vec{X}\right)\right]\right)^2}{E_\theta\left(\left(\frac{\partial}{\partial\theta} \log f\left(\vec{X}\middle|\theta\right)\right)^2\right)}$$

Equality hold if and only if

$$h(\theta)[W(\vec{x}) - g(\theta)] = \frac{\partial}{\partial\theta} \log L(\theta|\vec{x})$$

Proof: By Cauchy-Schwarz inequality: $[\mathrm{Cov}(X, Y)]^2 \leq (\mathrm{Var}\, X)(\mathrm{Var}\, Y)$

$$\mathrm{Var}\, X \geq \frac{[\mathrm{Cov}(X, Y)]^2}{\mathrm{Var}\, Y}$$

Corollary: Suppose the above theorem's assumptions are satisfied and iid.

$$Var_\theta\left(W(\vec{X})\right) \geq \frac{\left(\frac{d}{d\theta}E_\theta[W(\vec{X})]\right)^2}{nE_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right)}$$

$E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\vec{X}|\theta)\right)^2\right)$ is called the information number or Fisher information of the sample

Computation Lemma: For Exponential Family distributions,

$$E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\vec{X}|\theta)\right)^2\right) = -E_\theta\left(\frac{\partial^2}{\partial\theta^2}\log f(\vec{X}|\theta)\right)$$

See [CB, Lemma 7.3.11] for more general assumption for the lemma.

Example: Poisson Unbiased Estimation. (Conclusion)

Consider $W(\vec{X}) = \bar{X},\ So,\ E_\lambda(\bar{X}) = \lambda$ and $Var(\bar{X}) = \frac{\lambda}{n}$

$$\begin{aligned}
\mathrm{E}_\lambda\left(\left(\frac{\partial}{\partial\lambda}\log\prod_{i=1}^{n} f(X_i|\lambda)\right)^2\right) &= -n\mathrm{E}_\lambda\left(\frac{\partial^2}{\partial\lambda^2}\log f(X|\lambda)\right) \\
&= -n\mathrm{E}_\lambda\left(\frac{\partial^2}{\partial\lambda^2}\log\left(\frac{e^{-\lambda}\lambda^X}{X!}\right)\right) \\
&= -n\mathrm{E}_\lambda\left(\frac{\partial^2}{\partial\lambda^2}(-\lambda + X\log\lambda - \log X!)\right) \\
&= -n\mathrm{E}_\lambda\left(-\frac{X}{\lambda^2}\right) \\
&= \frac{n}{\lambda}.
\end{aligned}$$

So, by theorem, for any unbiased estimator, $Var_\theta\left(W(\vec{X})\right) \geq \frac{\lambda}{n}$

So, $W(\vec{X}) = \bar{X}$ is a best unbiased estimator of λ.

Example: Normal Distribution

Let $\{X_1, \ldots, X_n\}$ be a sample from a normal distribution $Normal(\mu, \sigma^2)$

Consider estimation of σ^2, where μ is unknown.

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

$$\begin{aligned}-\mathrm{E}\left(\frac{\partial^2}{\partial(\sigma^2)^2}\log f(X|\mu,\sigma^2)\bigg|\mu,\sigma^2\right) &= -\mathrm{E}\left(\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}\bigg|\mu,\sigma^2\right)\\ &= \frac{1}{2\sigma^4}.\end{aligned}$$

By **Cramér–Rao Inequality,** any unbiased estimator $W\ of\ \sigma^2$ must satisfy

$$Var(W(\vec{X})) \geq \frac{2\sigma^4}{n}$$

However, the sample variance S^2 can not attain the lower bound.

$$Var(S^2) = \frac{2\sigma^4}{n-1}$$

So, we don't know if S^2 is the best unbiased estimator.

Sufficiency and Unbiasedness

Rao-Blackwell Theorem:

Let $\{X_1, \dots, X_n\}$ be a sample from a population distribution with pdf $f(x|\theta)$

Let $W(\vec{X})$ be an estimator for $g(\theta)$.

Let $T(\vec{X})$ be a sufficient estimator for θ.

Then $\phi(T) := E[W|T]$ is a uniformly better unbiased estimator of $g(\theta), i.e.,$

$$E_\theta[\phi(T)] = g(\theta)$$

$Var_\theta\big(\phi(T)\big) \leq Var_\theta(W)$ for any θ

Proof:

Unbiasedness:

$$g(\theta) = E[W] = E[E(W|T)] = E[\phi(T)]$$

Uniformly better:

$$Var_\theta(W) = Var[E(W|T)] + E[Var(W|T)]$$

$$= Var[\phi(T)] + E[Var(W|T)]$$

$$\geq Var[\phi(T)]$$

Theorem: Suppose $E[W] = g(\theta)$.

$\boldsymbol{W}$ is the best unbiased estimator of if and only if W is uncorrelated with all unbiased estimators of zero.

Lehmann–Scheffé Theorem

Let $\{X_1, \dots, X_n\}$ be a sample from a population distribution with pdf $f(x|\theta)$

Let $T(\vec{X})$ be a complete sufficient estimator for θ.

Then, $\phi(T)$ is the best unbiased estimator of $E[\phi(T)]$

For example, if $W(\vec{X})$ be an estimator for $g(\theta)$.

Then $\phi(T) = E[W|T]$ is the unbiased estimator of $g(\theta)$.

Parameter	**Distribution**	**UMVUE**
Mean μ	Normal $N(\mu, \sigma^2)$ known σ^2	Sample mean $\bar{X}$
Mean μ	Normal $N(\mu, \sigma^2)$ unknown σ^2	Sample mean $\bar{X}$
Variance σ^2	Normal $N(\mu, \sigma^2)$ unknown μ	$\frac{1}{n-1}\sum(X_i - X)^2$
Parameter p	Bernoulli or Binomial	Sample proportion $\hat{p} = \frac{X}{n}$
Parameter λ	Poisson	Sample mean $\bar{X}$
Parameter θ	Uniform [0,θ]	$\frac{n+1}{n}\max(X_i)$

➢ Loss Function

MSE is a special case of the loss function in **decision theory.**

Let $\{X_1, \dots, X_n\}$ be a sample from a population distribution with pdf $f(x|\theta)$

Suppose $\theta \in \Theta$ the **parameter** (or, **state) space.**

Action space $\boldsymbol{\mathcal{A}}$ is the set of allowable decisions of θ based on observed Data: $\mathcal{D} = \{x_1, \dots, x_n\}$.

For example, for point estimators, $\boldsymbol{\mathcal{A}} = \boldsymbol{\Theta}$.

After an action/decision a is made, the **loss function** measures the distance from a to the true θ.

If the action is correct, the loss is minimum.

For example, **Absolute Error Loss** $\boldsymbol{L(\theta, a) = |a - \theta|}$

Squared Error Loss $\boldsymbol{L(\theta, a) = (a - \theta)^2}$

Variations can be constructed:

$$L(\theta, a) = \begin{cases} (a-\theta)^2 & \text{if } a < \theta, \\ 10(a-\theta)^2 & \text{if } a \geq \theta. \end{cases}$$

or

$$L(\theta, a) = \frac{(a-\theta)^2}{|\theta| + 1}.$$

More loss functions can be defined, e.g.

$$L(\theta, a) = \begin{cases} 0 & \text{If } a(x) = \theta \text{ (correct decision)} \\ 1 & \text{If } a(x) \neq \theta \text{ (incorrect decision)} \end{cases}$$

Decision rule is a function $\delta: \mathcal{X} \to \mathcal{A}$ that selects an action a given the observations $\mathcal{D}$

- Use loss/cost function to select which decision rule to use:

loss function $L: \Theta \times \mathcal{A} \to \mathbb{R}$

Let $\delta(\vec{x})$ be an estimator of θ.

The **risk function** is a function

$$R(\theta, \delta) := E_\theta[L(\theta, \delta(\vec{X}))]$$

The risk function is the average loss if the estimator $\delta(x)$ is used.

Given two estimators δ_1 and δ_2, we can compare their risk functions:

$R(\theta, \delta_1) < R(\theta, \delta_2)$ implies δ_1 is better than δ_2.

Mean Squared Error(MSE) of an estimator $\delta(\vec{X})$

$$MSE := E_\theta\left(\delta(\vec{X}) - \theta\right)^2$$

$$= E_\theta[L(\theta, \delta(\vec{X}))] \qquad \text{for } L(\theta, a) = (a - \theta)^2$$

$$= R(\theta, \delta)$$

MSE can has the Bias-Variance decomposition:

$$E_\theta(\delta - \theta)^2 = Var_\theta\left(\delta(\vec{X})\right) + \left(E_\theta\delta(\vec{X}) - \theta\right)^2 = Var_\theta\left(\delta(\vec{X})\right) + \left[Bias_\theta\left(\delta(\vec{X})\right)\right]^2$$

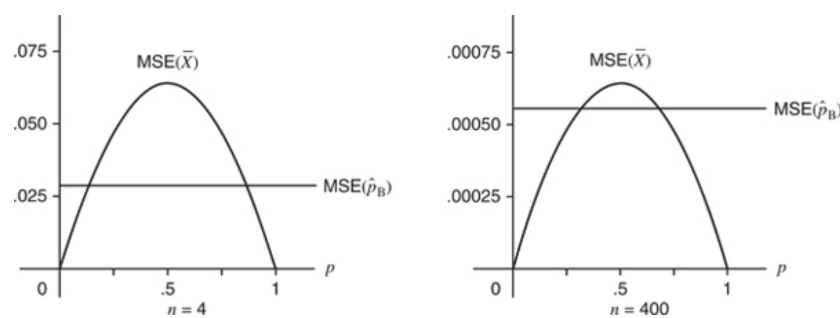
Example (Bernoulli)

Suppose random sample $X_i \sim Bernoulli(p)$ for $i = 1, \dots, n$ with unknown p.

Compare two estimators:

$$\hat{p}_{\mathrm{B}} = \frac{\sum_{i=1}^{n} X_i + \sqrt{n/4}}{n + \sqrt{n}} \qquad \text{and} \qquad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The risk functions are the MSE here:



Example (Normal Variance)

The Sample Variance S^2 is unbiased estimator of σ^2, i.e., $E(S^2) = \sigma^2$.

The Risk function (MSE) of S^2 is $Var(S^2) = \frac{2\sigma^4}{n-1}$

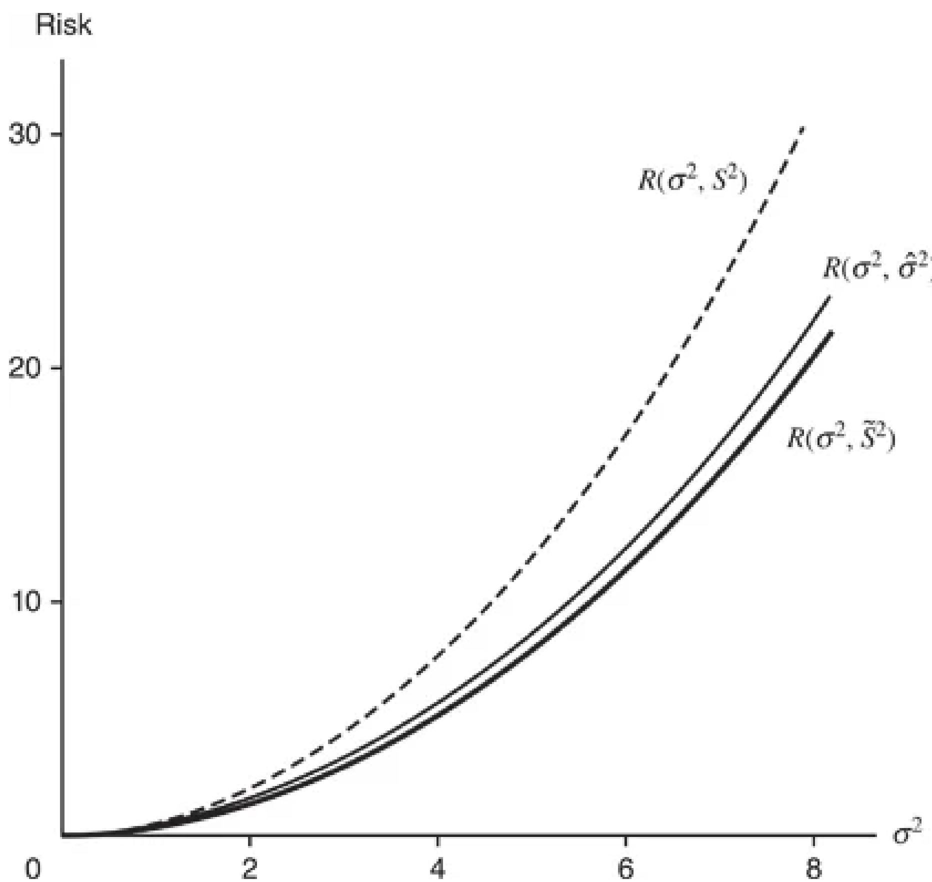
Consider estimators $\delta_b(\vec{X}) \coloneqq bS^2$.

The Risk function for $\delta_b(\vec{X})$

$$
\begin{aligned}
R((\mu,\sigma^2),\delta_b) &= \operatorname{Var} bS^2 + \left(\mathrm{E}bS^2 - \sigma^2\right)^2 \\
&= b^2 \operatorname{Var} S^2 + \left(b\mathrm{E}S^2 - \sigma^2\right)^2 \\
&= \frac{b^2 2\sigma^4}{n-1} + (b-1)^2\sigma^4 \\
&= \left[\frac{2b^2}{n-1} + (b-1)^2\right]\sigma^4.
\end{aligned}
$$

To minimize $R((\mu,\sigma^2),\delta_b)$, $b = \frac{n-1}{n+1}$. That is $\delta_b(\vec{X}) \coloneqq \frac{n-1}{n+1} S^2 = \frac{1}{n+1}\sum(X_i - \bar{X})^2 \coloneqq \tilde{S}^2$

Compare risk function of $S^2, \tilde{S}^2$ and $\hat{\sigma}^2_{MLE}$



Use different loss function

Stein's loss: $$L(\sigma^2, a) = \frac{a}{\sigma^2} - 1 - \log\frac{a}{\sigma^2},$$

The Risk function for estimators $\delta_b(\vec{X}) \coloneqq bS^2$.

$$\begin{aligned} R(\sigma^2, \delta_b) &= \mathrm{E}\left(\frac{bS^2}{\sigma^2} - 1 - \log\frac{bS^2}{\sigma^2}\right) \\ &= b\mathrm{E}\frac{S^2}{\sigma^2} - 1 - \mathrm{E}\log\frac{bS^2}{\sigma^2} \\ &= b - \log b - 1 - \mathrm{E}\log\frac{S^2}{\sigma^2}. \end{aligned}$$

To minimize the risk function, we need to choose $b = 1$

Bayes Risk:

Given a prior distribution $\pi(\theta)$, the Bayes Risk is

$$R_B(\theta, \delta) = \int_\Theta R(\theta, \delta)\pi(\theta)d\theta$$

$$= \int_\Theta \int_\mathcal{X} L\big(\theta, \delta(\vec{x})\big) f(\vec{x}|\,\theta)\,\pi(\theta)d\vec{x}d\theta$$

$$= \int_\mathcal{X} \int_\Theta L\big(\theta, \delta(\vec{x})\big)\pi(\theta|\vec{x})\,d\theta\;m(\vec{x})d\vec{x}$$

$\pi(\theta|\vec{x})$ is the posterior distribution of θ and $m(\vec{x})$ is the marginal distribution of $\vec{x}$

Posterior expected loss: expected value of the loss function with respect to the posterior distribution.

Criterion	Definition	Desirable Property	Notes
Unbiasedness	$E[\hat{\theta}] = \theta$	Estimator's expected value equals the true parameter	An unbiased estimator doesn't systematically over- or underestimate
Bias	$Bias(\hat{\theta}) = E[\hat{\theta}] - \theta$	Bias = 0 (for unbiased estimators)	Bias can be positive or negative
Variance	$Var(\hat{\theta}) = E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right]$	Lower is better	Measures spread of estimator values across samples
Mean Squared Error (MSE)	$MSE(\hat{\theta}) = E\left[\left(\hat{\theta} - \theta\right)^2\right]$	Lower is better	Combines bias and variance: $MSE = Var + Bias^2$
Efficiency	Relative measure of variance compared to best possible estimator	Lower variance among unbiased estimators	Often compared to the Cramér-Rao Lower Bound
Consistency	$\hat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$	Estimator converges to true value as sample size increases	A large-sample property
Sufficiency	Estimator captures all info in data about the parameter	Retains full information	Linked to data reduction and the factorization theorem
Robustness	Resistant to small deviations from assumptions	Less sensitive to outliers or model misspecification	Not always emphasized in classical theory, but important in practice

References:

- **Book 1. [CB] Statistical Inference**, by Casella, George, Berger, Roger L, 2nd edition
- **Book 2. [W]: All of Statistics: Larry Wasserman**
-

Online books:

https://www.probabilitycourse.com/