Estimating the Mean from Data: Introduction to Estimation Theory

Motonobu Kanagawa

Introduction to Statistics, EURECOM

March 4, 2024

Outline

Estimation of the Mean

- Let X be a random variable taking values in $\mathbb{R}$ with probability distribution P.

(Note: In the previous lecture, P is used to denote the distribution of the underlying probability space, but here P denotes the distribution of X).

- The mean (or the expected value) of X is defined by

$$\mu := \mathbb{E}_{X \sim P}[X] = \int x \, dP(x) \in \mathbb{R}.$$

Assume that we don't know P, and thus we don't know μ.

Estimation of the Mean

- Assume instead that we are given some **data**:

$$X_1, \ldots, X_n \in \mathbb{R}$$

- These are assumed to be random variables taking values in $\mathbb{R}$.

- The task of mean estimation is estimating the unknown mean μ from the data $X_1, \ldots, X_n$.

- This is one of the most ubiquitous and fundamental problems in statistics.

- In this lecture, we look at this problem in details.

Motivation 1: Relation to Many Problems

Many problems can be formulated as estimation of the mean.

Examples:

- Monte Carlo: Simulation-based mean estimation.
- Design of experiments: Average treatment (causal) effect.
- Regression: Estimation of the conditional mean.
- Supervised machine learning:
  - Risk = the mean of a loss function.
  - Stochastic gradient = approximation of the expected gradient.

Motivation 2: Different Statistical Approaches

Mean estimation can be used for illustrating different approaches.

- The “frequentist” approach - maximum likelihood estimation.

- The “Bayesian” approach - posterior inference.

- The “empirical Bayes” - the mixed approach.

Motivation 3: Key Notions

We can learn key notions in statistics.

- Estimator and consistency.

- Bias-variance decomposition/trade-off

- Law of large numbers and the central limit theorem.

Most importantly,

- The key is how data are generated/obtained.

Is the Empirical Average a Good Approach?

A standard approach is to take the empirical average of data points:

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

In this lecture, we will address questions like:

- When is the empirical average a good estimate, and when is it not?
- When can we justify the use of the empirical average?
- What conditions do we need for the data $X_1, \ldots, X_n$?

Outline

Population and Data

In the mean estimation problem, we have two kinds of random variables:

[Population] Random variable X represents the hypothetical population of interest, with P being its probability distribution.

[Data] Random Variables $X_1, \dots, X_n$ represent the given data.

- The data $X_1, \dots, X_n$ are assumed to provide information about the population random variable X (or its distribution P).

- Otherwise, we cannot estimate the population mean $\mu = \mathbb{E}_{X \sim P}[X]$ from the data $X_1, \dots, X_n$.

- Therefore, how the data are generated/obtained becomes very important.

Example: Estimating the Average Income in France

- Assume that $X \in \mathbb{R}$ represents the income of a randomly sampled French person, with P being its distribution.

- The population mean $\mu = \mathbb{E}_{X \sim P}[X]$ represents the average income of French people.

- The data $X_1, \ldots, X_n$ are the incomes of n French people randomly selected from the French population.

- Then, is the empirical average

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

a good estimate of the true average income $\mu = \mathbb{E}_{X \sim p}[X]$?

Example: Estimating the Average Income in France

- Assume that data $X_1, \ldots, X_n$ are the incomes of randomly sampled French persons in French Riviera.

- Then, the empirical average

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

would be higher than the average income of the French population.

Example: Estimating the Average Income in France

- Assume that the data $X_1, \ldots, X_n$ are the incomes of randomly sampled French people between age 20 and 30.

- Then, the empirical average

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

would give an estimate lower than the true average income.

The Data Generating Process Matters

These examples indicate that how the data are generated/obtained strongly affects the validity of the empirical average.

- We need to make sure that data $X_1, \ldots, X_n$ are sampled from the same population as that of the target random variable $X \sim P$.

- This requirement is mathematically formulated by assuming that random variables $X_1, \ldots, X_n$ are independently and identically distributed (i.i.d.) with $X \sim P$.

Independently and Identically Distributed (i.i.d.)

Recall that random variables $X_1, \ldots, X_n$ are i.i.d. with a random variable $X \sim P$ if they satisfy the following:

- Independence:

  - X_i and X_j are independent for all $i \neq j$.
  - X_i and X are independent for all $i = 1, \ldots, n$;
    - Recall that X represents the hypothetical population (e.g., randomly selected French person).

- Identity:

  - X_i follows the same probability distribution P of X (for all $i = 1, \ldots, n$).

We often write $X_1, \ldots, X_n \sim P$ $(i.i.d.)$.

See also the lecture slides on Probability Theory.

Outline

Preliminaries

Before going further, we collect here some key properties of Expectation and Variance of random variables.

Some Key Properties of Expectation

- For any real-valued random variable X and a constant $c \in \mathbb{R}$, we have

$$\mathbb{E}[cX] = c\mathbb{E}[X].$$

- For any real-valued random variables X and Y, we have

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

- If X and Y are independent,

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

Variance of a Random Variable

In statistics, the variance of a random variable plays a key role.

- Let X be a real-valued random variable with probability distribution P.

Then the variance of X is defined by

$$\mathbb{V}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2] = \int (x - \mathbb{E}[X])^2 dP(x) \geq 0.$$

- Note that the mean $\mathbb{E}[X] \in \mathbb{R}$ is a constant.

Some Key Properties of Variance

Let X be a real-valued random variable.

Then we have

$$\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Proof:

$$\begin{aligned}
\mathbb{V}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\
&= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2] \\
&= \mathbb{E}[X^2] - 2(\mathbb{E}[X])^2 + (\mathbb{E}[X])^2 \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2
\end{aligned}$$

Some Key Properties of Variance

Let X be a real-valued random variable. Then for any constant $c \in \mathbb{R}$, we have

$$\mathbb{V}[cX] = c^2\mathbb{V}[X].$$

Proof:

$$\begin{aligned}\mathbb{V}[cX] &:= \mathbb{E}[(cX - \mathbb{E}[cX])^2] \\ &= c^2\mathbb{E}[(X - \mathbb{E}[X])^2] = c^2\mathbb{V}[X].\end{aligned}$$

In particular, by setting $c = 1/n$, we have

$$\mathbb{V}\left[\frac{X}{n}\right] = \frac{1}{n^2}\mathbb{V}[X].$$

Some Key Properties of Variance

Let X and Y be real-valued random variables.

If X and Y are independent, then

$$\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y].$$

Proof:

$$\begin{aligned}
&\mathbb{V}[X+Y] := \mathbb{E}[(X+Y-\mathbb{E}[X+Y])^2] \\
&= \mathbb{E}[(X-\mathbb{E}[X]+Y-\mathbb{E}[Y])^2] \\
&= \mathbb{E}[(X-\mathbb{E}[X])^2 + 2(X-\mathbb{E}[X])(Y-\mathbb{E}[Y]) + (Y-\mathbb{E}[Y])^2] \\
&= \mathbb{E}[(X-\mathbb{E}[X])^2] + 2\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] + \mathbb{E}[(Y-\mathbb{E}[Y])^2] \\
&= \mathbb{E}[(X-\mathbb{E}[X])^2] + \mathbb{E}[(Y-\mathbb{E}[Y])^2] = \mathbb{V}[X] + \mathbb{V}[Y],
\end{aligned}$$

where we used

$$\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[(X-\mathbb{E}[X])]\mathbb{E}[(Y-\mathbb{E}[Y])] = 0,$$

which follows from the independence of X and Y.

Some Key Properties of Variance

By recursive applications of the previous result, we have the following useful result:

Let $X_1, X_2, \ldots, X_n$ are independent real-valued random variables (note: they don't necessary identically distributed).

Then we have

$$\mathbb{V}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{V}[X_i]$$

Corollary:

- Let $X_1, \ldots, X_n$ be independent real-valued random variables.
- Let $c_1, \ldots, c_n \in \mathbb{R}$ be constants.

Then

Weighted average

$$\mathbb{V}[\sum_{i=1}^{n} c_i X_i] = \sum_{i=1}^{n} \mathbb{V}[c_i X_i] = \sum_{i=1}^{n} c_i^2 \mathbb{V}[X_i].$$

Some Key Properties of Variance

In particular, assuming that $X_1, \ldots, X_n$ are i.i.d. with a random variable X, and setting $c_i := 1/n$, we have

$$\mathbb{V}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i] = \frac{1}{n}\mathbb{V}[X].$$

n times the same variance because iid

- Thus, the variance of the empirical average $\frac{1}{n}\sum_{i=1}^{n} X_i$ is n times smaller than the variance of X.

- By taking the average over independent observations, the variance can be reduced.

Outline

Estimators and Estimates

In statistics, the procedure of estimating a quantity of interest is formulated as a function of data.

- This function is called an estimator.

- The output from the estimator is called an estimate.

Estimators and Estimates

- Let $\theta^* \in \Theta$ be an unknown quantity of interest that we want to estimate (Θ is an appropriate set) (θ^* is also called an estimand).

- Assume that we are given some data D_n of size $n \in \mathbb{N}$ of the form

$$D_n := (X_1, \ldots, X_n) \in \mathcal{X}^n$$ (X times, Y times, ...)

where each $X_i \in \mathcal{X}$ is a random variable ($\mathcal{X}$ is a measurable space.).

Definition: a map

$$F_n : \mathcal{X}^n \to \Theta$$

is called an estimator (of θ^*).

- The estimator should be designed so that the estimate will be close to θ^*.

- $\hat{\theta}_n := F_n(D_n)$ is called an estimate (of θ^*).

Estimators and Estimates: Mean Estimation

Let's consider the mean estimation problem as an example.

The quantity of interest is the mean of the random variable $X \sim P$:

$$\theta^* := \mu := \mathbb{E}[X] \in \mathbb{R} =: \Theta.$$

Assume that n random variables $X_1, \ldots, X_n$ are given as data:

$$D_n = (X_1, \ldots, X_n) \in \mathcal{X}^n, \quad \mathcal{X} := \mathbb{R}.$$

Then one can define an estimator $F_n : \mathcal{X}^n \to \Theta$ of the mean θ^* by

$$F_n(D_n) := \frac{1}{n} \sum_{i=1}^{n} X_i = \hat{\mu} =: \hat{\theta}.$$

i.e., the empirical average of $X_1, \ldots, X_n$.

Which Estimator Should We Choose?

Note that the empirical average is not the only choice.

For instance, we can define various estimators for the mean estimation problem; e.g.,

1. $F_n(D_n) := (X_1 + \cdots + X_n)/n$. empirical average
2. $F_n(D_n) := X_1$ (i.e., discarding $X_2, \ldots, X_n$).
3. $F_n(D_n) := 0$ (i.e., always outputs constant 0, no matter what D_n is).
4. $F_n(D_n) := c_0 + c_1 X_1 + \cdots + c_n X_n$ for some $c_0, c_1, \ldots, c_n \geq 0$.
   or define weighted average

- Which estimator should we choose?

- When is the empirical average a good choice, and when is it not?

(Actually we'll see that the empirical average is not always a good choice).

Which Estimator Should We Choose?

To investigate these questions, we need to introduce criteria for comparing different estimators.

Outline

Mean Square Error (MSE)

To discuss the quality of a statistical estimator, we need a certain error criterion.

Here we consider the mean square error (MSE), one of the most standard criteria.

- Let $\theta^* \in \Theta \subset \mathbb{R}$ be the unknown quantity of interest.

— We assume $\Theta \subset \mathbb{R}$ for simplicity, but the following argument also holds for more general situations.

- Consider an estimator $F_n : \mathcal{X}^n \to \Theta$ such that

$$\hat{\theta}_n := F_n(D_n) \in \Theta, \quad D_n := (X_1, \ldots, X_n) \in \mathcal{X}^n.$$

- Note that the estimate $\hat{\theta}_n = F_n(D_n) = F_n((X_1, \ldots, X_n))$ is a random variable, since $X_1, \ldots, X_n$ are random variables.

Mean Square Error (MSE)

- Then we can consider the squared error between the target θ^* and estimate $\hat{\theta}_n$:

$$(\hat{\theta}_n - \theta^*)^2 = (F_n(D_n) - \theta^*)^2.$$

- This error is also a random variable, because the estimate $\hat{\theta}_n = F_n(D_n)$ is a random variable.

- Then the mean square error (MSE) of the estimator F_n is defined as the expectation of the squared error:

$$\mathbb{E}[(\hat{\theta}_n - \theta^*)^2] = \mathbb{E}[(F_n(D_n) - \theta^*)^2]$$

where the expectation is with respect to the data $D_n = (X_1, \ldots, X_n)$.

- The MSE quantifies how the estimate $\hat{\theta}_n$ is close to (or far from) the target θ^* on average.

Mean Square Error (MSE)

Note that the MSE

$$\mathbb{E}[(\hat{\theta}_n - \theta^*)^2] = \mathbb{E}[(F_n(D_n) - \theta^*)^2]$$

depends on

1. the target quantity θ^*
2. the estimator F_n
3. the distribution of the data $X_1, \ldots, X_n$

By theoretically studying the MSE, we can study

- which estimator F_n is good for estimating the target θ^*,
- when the data $X_1, \ldots, X_n$ are distributed in an assumed way.

Probabilistic Error Bound from MSE

- A general fact: For any non-negative real-valued random variable Z, Markov's inequality states that

$$\Pr(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}, \quad \forall c > 0.$$

- By setting $Z := (\hat{\theta}_n - \theta^*)^2$, we then have

$$\Pr((\hat{\theta}_n - \theta^*)^2 \geq c) \leq \frac{\mathbb{E}[(\hat{\theta}_n - \theta^*)^2]}{c}, \quad \forall c > 0.$$

- Thus, if the MSE $\mathbb{E}[(\hat{\theta} - \theta^*)^2]$ is small, then the probability of

$$(\hat{\theta}_n - \theta^*)^2 > c$$

becomes small for any $c > 0$.

Bias-Variance Decomposition

- The following is a very important result concerning the MSE.

Theorem: The MSE can be decomposed into the bias and the variance of the estimator, as follows:

$$\mathbb{E}[(\hat{\theta}_n - \theta^*)^2] = \underbrace{\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2]}_{Variance} + (\underbrace{\mathbb{E}[\hat{\theta}_n] - \theta^*}_{Bias})^2$$

This is called the bias-variance decomposition.

- The bias of the estimator $F_n : \mathcal{X}^n \to \Theta$ is defined as the difference between the expectation of the estimate $\mathbb{E}[\hat{\theta}_n]$ and the target θ^*:

$$\mathbb{E}[\hat{\theta}_n] - \theta^* = \mathbb{E}[F_n(D_n)] - \theta^*.$$

where the expectation is with respect to the data $D_n = (X_1, \ldots, X_n)$.

Bias-Variance Decomposition

$$\mathbb{E}[(\hat{\theta}_n - \theta^*)^2] = \underbrace{\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2]}_{Variance} + (\underbrace{\mathbb{E}[\hat{\theta}_n] - \theta^*}_{Bias})^2$$

- The variance of the estimator $F_n : \mathcal{X}^n \to \Theta$ is defined as

$$\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2] = \mathbb{E}[(F_n(D_n) - \mathbb{E}[F_n(D_n)])^2].$$

  - i.e., the average deviation of the estimate $\hat{\theta}_n := F_n(D_n)$ from its mean $\mathbb{E}[\hat{\theta}_n]$.
  - Recall again that the estimate $\hat{\theta}_n$ is a random variable.

- To make the mean-square error small, both the bias and variance need to be small!

Proof of Bias-Variance Decomposition

- The mean square error can be expanded as

$$
\begin{aligned}
&\mathbb{E}[(\hat{\theta}_n - \theta^*)^2] \\
=&\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n] + \mathbb{E}[\hat{\theta}_n] - \theta^*)^2] \\
=&\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2] + \mathbb{E}[(\mathbb{E}[\hat{\theta}_n] - \theta^*)^2] + 2\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])(\mathbb{E}[\hat{\theta}_n] - \theta^*)] \\
=&\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2] + (\mathbb{E}[\hat{\theta}_n] - \theta^*)^2,
\end{aligned}
$$

adding

1 Term

1 Term

disappea

where the last line follows from $\mathbb{E}[\hat{\theta}_n]$ being a constant:

$$
\begin{aligned}
\mathbb{E}[(\mathbb{E}[\hat{\theta}_n] - \theta^*)^2] &= (\mathbb{E}[\hat{\theta}_n] - \theta^*)^2, \\
\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])(\mathbb{E}[\hat{\theta}_n] - \theta^*)] &= \mathbb{E}\left[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])\right](\mathbb{E}[\hat{\theta}_n] - \theta^*) = 0.
\end{aligned}
$$

Constant can be extracted

Remarks on the Bias-Variance Decomposition

- The bias-variance decomposition holds under a very generic situation.

  - This is because the proof does not require any assumption about the joint distribution of the data $X_1, \ldots, X_n$ (essentially).
  - The only assumption is that the MSE is finite.

- Thus, for instance, we can consider cases like:

  - where $X_1, \ldots, X_n$ are not independently distributed
  - where $X_1, \ldots, X_n$ are not identically distributed.

- By considering a different setting for the distribution of the data $X_1, \ldots, X_n$, we can study when a certain estimator is a good choice, when it is not.

- This is done by analyzing the bias and variance of the estimator.

Bias-Variance Decomposition: Multivariate Case

- Let $\theta^* \in \Theta \subset \mathbb{R}^d$ be the quantity of interest.

- Let $\hat{\theta}_n$ be any estimate of θ^* (you can just think of $\hat{\theta}_n$ as a random variable in $\mathbb{R}^d$).

- Define the mean square error by

$$\mathbb{E}[\|\hat{\theta}_n - \theta^*\|^2],$$

where $\|\cdot\|$ is the norm of $\mathbb{R}^d$.

Theorem. - Assume that

$$\|\mathbb{E}[\hat{\theta}_n]\| < \infty, \quad \mathbb{E}[\|\hat{\theta}_n\|^2] < \infty.$$

Then the following bias-variance decomposition holds:

$$\mathbb{E}[\|\hat{\theta}_n - \theta^*\|^2] = \underbrace{\mathbb{E}[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2]}_{\textit{Variance}} + \|\underbrace{\theta^* - \mathbb{E}[\hat{\theta}_n]}_{\textit{Bias}}\|^2$$

Bias-Variance Decomposition: Multivariate Case

Exercise: Prove the above bias-variance decomposition.

Hint: for any $a, b \in \mathbb{R}^d$,

$$\|a - b\|^2 = \langle a - b, a - b \rangle$$

where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathbb{R}^d$.

Outline

Mean Estimation Problem: Setup

- Now consider the mean estimation problem.

- Let $X \sim P$ be the random variable of interest, whose mean

$$\mu_P := \mathbb{E}[X] = \int x \, dP(x)$$

is the estimand.

- To deal with a generic situation, we assume that i.i.d. data $X_1, \ldots, X_n$ are generated from a probability distribution Q, which can be different from P:

$$X_1, \ldots, X_n \sim Q, i.i.d.$$

- Let $Y \sim Q$ be a random variable, with distribution Q;

- Then $X_1, \ldots, X_n$ are i.i.d. with Y.

Bias-Variance Decomposition in Mean Estimation

- Assume that the mean and the variance of $Y \sim Q$ are finite:

$$
\begin{aligned}
|\mu_Q| < \infty, &\quad \mu_Q := \mathbb{E}_{Y\sim Q}[Y] \\
\sigma_Q^2 < \infty, &\quad \sigma_Q^2 := \mathbb{V}_{Y\sim Q}[Y] := \mathbb{E}_{Y\sim Q}[(Y-\mu_Q)^2].
\end{aligned}
$$

Theorem: The mean square error of the empirical average estimator

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

is given by

$$
\begin{aligned}
\mathbb{E}[(\hat{\mu}-\mu_P)^2] &= \mathbb{E}[(\hat{\mu}-\mathbb{E}[\hat{\mu}])^2] + (\mathbb{E}[\hat{\mu}]-\mu_P)^2 \\
&= \frac{\sigma_Q^2}{n} + (\mu_Q-\mu_P)^2.
\end{aligned}
$$

Proof: Bias-Variance Decomposition in Mean Estimation

Proof:

- The first identity follows from the bias-variance decomposition.

- Thus, we show the second identity.

Variance term.

Because $X_1, \ldots, X_n$ are i.i.d. with $Y \sim Q$, the variance term can be expressed as

$$\mathbb{E}[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2] = \mathbb{V}[\hat{\mu}] = \mathbb{V}[\frac{1}{n}\sum_{i=1}^{n} X_i]$$
$$= \sum_{i=1}^{n} \mathbb{V}[\frac{1}{n}X_i] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}[X_i] = \frac{1}{n}\mathbb{V}[Y] = \frac{\sigma_Q^2}{n}.$$

Proof: Bias-Variance Decomposition in Mean Estimation

Bias term. On the other hand,

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mathbb{E}[Y] = \mu_Q.$$

Therefore, the bias term is

$$(\mathbb{E}[\hat{\mu}] - \mu_P)^2 = (\mu_Q - \mu_P)^2.$$

$\square$

Interpretation of the Bias-Variance Decomposition

We proved the bias-variance decomposition:

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = \frac{\sigma_Q^2}{n} + (\mu_Q - \mu_P)^2.$$

Let's study what this means.

- The bias of the estimator $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is

$$\mathbb{E}[\hat{\mu}] - \mu_P = \mu_Q - \mu_P$$

i.e., the difference between

- the mean μ_Q of the data distribution Q, and
- the mean μ_P of the target distribution P.

Interpretation of the Bias-Variance Decomposition

Therefore,

- if the data $X_1, \ldots, X_n$ are independently generated from a distribution Q, and
- if the mean μ_Q of Q is different from the mean μ_P of the target random variable $X \sim P$,

then the use of the empirical average

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

causes a non-zero bias, $\mu_Q - \mu_P \neq 0$.

Note that in this case, since $(\mu_Q - \mu_P)^2 > 0$, the mean square error

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = (\mu_Q - \mu_P)^2 + \frac{\sigma_Q^2}{n} \geq (\mu_Q - \mu_P)^2 > 0$$

does not decrease to 0, even when $n \to \infty$.

Interpretation of the Bias-Variance Decomposition

This example shows the importance of the data distribution Q.

- If possible, we should collect data $X_1, \ldots, X_n$ generated from the same distribution P as the target random variable X, i.e., $Q = P$.

- In this case, the bias becomes 0: $(\mu_Q - \mu_P)^2 = 0$, and the MSE is

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = \frac{\sigma_P^2}{n},$$

where $\sigma_P^2 = \mathbb{V}[X]$ is the variance of $X \sim P$.

- Thus, the MSE decreases as the sample size n increases.

Interpretation of the Bias-Variance Decomposition

- On the other hand, the variance term

$$\mathbb{E}[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2] = \frac{\sigma_Q^2}{n}$$

depends only on the data $X_1, \ldots, X_n$, and not on the target μ_P.

- Therefore, whatever the data distribution Q is, the variance term converges to 0 as $n \to \infty$:

$$\lim_{n\to\infty} \mathbb{E}[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2] = \lim_{n\to\infty} \frac{\sigma_Q^2}{n} = 0.$$

Interpretation of the Bias-Variance Decomposition

- Note that in the derivation of the variance term, we used

$$\mathbb{V}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{V}[\frac{1}{n}X_i].$$

- This follows from the independence between $X_1, \ldots, X_n$. (see pp.21-22)

- Therefore, if the independence between $X_1, \ldots, X_n$ does not hold, the variance may not decrease to 0 (we'll see an example later).

Interpretation of the Bias-Variance Decomposition

- For example, recall the example where $X \sim P$ represents the income of a randomly picked-up French person.

- Assume that data $X_1, \ldots, X_n \sim Q$ $(i.i.d.)$ are the incomes of randomly picked-up French persons in French Riviera.

- Then we would have

$$\mu_Q := \mathbb{E}_{Y \sim Q}[Y] > \mathbb{E}_{X \sim P}[X] =: \mu_P$$

i.e., the average income of French Riviera people μ_Q is higher than the average income of the whole population μ_P.

Interpretation of the Bias-Variance Decomposition

- Thus, the empirical average of the data

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

has a non-zero bias:

$$\mathbb{E}[\hat{\mu}] - \mu_P = \mu_Q - \mu_P \neq 0.$$

- Therefore, the MSE of the empirical average

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = (\mu_Q - \mu_P)^2 + \frac{\sigma_Q^2}{n}$$

does not decrease to 0, even when n is very large.

- Thus, we should make sure that data $X_1, \ldots, X_n$ are randomly picked-up from the whole French population. (i.e., $Q = P$).

Mean Estimation in the Multivariate Case

- Let $X \sim P$ be a random vector in $\mathbb{R}^d$. Define

$$\mu_P := \mathbb{E}_{X \sim P}[X] \in \mathbb{R}^d$$

- Let $X_1, \ldots, X_n \sim Q$ ($i.i.d.$) be random vectors in $\mathbb{R}^d$, and let $Y \sim Q$. Define

$$\mu_Q := \mathbb{E}_{Y \sim Q}[Y] \in \mathbb{R}^d, \quad \sigma_Q^2 := \mathbb{E}_{Y \sim Q}[\|Y - \mu_Q\|^2] \geq 0.$$

Theorem. Assume that

$$\|\mu_P\| < \infty, \quad \|\mu_Q\| < \infty, \quad \sigma_Q^2 < \infty.$$

Then, the empirical average estimator $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ satisfies

$$\begin{aligned} \mathbb{E}[\|\hat{\mu} - \mu_P\|^2] &= \mathbb{E}[\|\hat{\mu} - \mathbb{E}[\hat{\mu}]\|^2] + \|\mathbb{E}[\hat{\mu}] - \mu_P\|^2 \\ &= \frac{\sigma_Q^2}{n} + \|\mu_Q - \mu_P\|^2. \end{aligned}$$

Exercise. Prove this. (The first identity is the bias-variance decomposition)

How Large should the Sample Size be?

- In the mean estimation problem, when $X_1, \ldots, X_n \sim P$ i.i.d., the MSE is given by

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = \frac{\sigma_P^2}{n}, \quad \sigma_P^2 := \mathbb{V}[X].$$

for the empirical average estimate $\hat{\mu} := \frac{1}{n}\sum_{i=1}^n X_i$.

- Assume that one wants to make the MSE small in that sense that

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] \leq \varepsilon^2,$$

for some $\varepsilon > 0$. Then the sample size n should satisfy

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = \frac{\sigma_P^2}{n} \leq \varepsilon^2$$

or equivalently

$$n \geq \frac{\sigma_P^2}{\varepsilon^2}.$$

How Large should the Sample Size be?

For instance, consider the example of estimating the average income.

- Assume $\mu_P = 2,000$ EUR/month (mean) and $\sigma_P = 500$ (standard deviation).

Then, the sample size n should satisfy

$$n \geq \frac{500^2}{\varepsilon^2}.$$

For instance,

- to achieve the precision of $\varepsilon = 10$, we need $n \geq 2500$.

- to achieve the precision of $\varepsilon = 1$, we need $n \geq 250,000$.

Outline

Consistency

- Let $\theta^* \in \Theta \subset \mathbb{R}$ be an estimand (i.e., the quantity of interest).

- Let $X_1, \ldots, X_n$ be random variables such that $X_i \in \mathcal{X}$, and define the data as

$$D_n := (X_1, \ldots, X_n) \in \mathcal{X}^n$$

- Let $F_n : \mathcal{X}^n \to \mathbb{R}$ be an estimator, and let $\hat{\theta}_n := F_n(D_n)$ be an estimate.

Definition. We call F_n a consistent estimator of θ^*, if the estimate $\hat{\theta}_n$ converges to θ^* as $n \to \infty$ in an appropriate sense, e.g.,

$$\mathbb{E}[(\hat{\theta}_n - \theta^*)^2] \to 0 \quad \text{as } n \to \infty.$$

- The consistency means that, as we have more data $X_1, \ldots, X_n$, the estimate $\hat{\theta}_n$ becomes more accurate (in estimating θ^*).

- Consistency is one of the most important concepts in statistics.

Unbiasedness

Definition. We call F_n an unbiased estimator of θ^*, if the bias is zero for every $n \in \mathbb{N}$, i.e.,

$$\mathbb{E}[F_n(D_n)] - \theta^* = \mathbb{E}[\hat{\theta}_n] - \theta^* = 0, \quad \forall n \in \mathbb{N}.$$

- If this is not satisfied, we call F_n a biased estimator of θ^*.

Unbiasedness

For instance, consider the mean estimation problem.

- If the data $X_1, \ldots, X_n$ are i.i.d. with $X \sim P$, then the empirical average $\hat{\mu} := \frac{1}{n}\sum_{i=1}^{n} X_i$ satisfies

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

- So, in this case, the empirical average $\hat{\mu}$ is an unbiased estimator of the mean μ_P.

- If $X_1, \ldots, X_n$ are i.i.d. with $Y \sim Q$, and if $\mu_Q \neq \mu_P$, then

$$\mathbb{E}[\hat{\mu}] = \mu_Q \neq \mu_P.$$

- So, in this case, the empirical average $\hat{\mu}$ is a biased estimator of the mean μ_P.

Unbiasedness

- If F_n is an unbiased estimator, then the MSE is given by

$$\mathbb{E}[(\hat{\theta}_n - \theta^*)^2] = \mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2] = \mathbb{V}[\hat{\theta}_n]$$

i.e., the MSE is equal to the variance of the estimate $\hat{\theta}_n$.

Some important consequences of unbiasedness:

- If the variance $\mathbb{V}[\hat{\theta}_n]$ decreases to 0 as $n \to \infty$, then $\hat{\theta}_n$ converges to θ^*; thus F_n becomes a consistent estimator.

- If we can estimate the variance $\mathbb{V}[\hat{\theta}_n]$, then we can estimate the amount of error (MSE):

  - In other words, we can estimate how far the estimate $\hat{\theta}_n$ is from the target θ^*.
  - Thus, an estimate of the variance $\mathbb{V}[\hat{\theta}_n]$ can be used for constructing a confidence interval for θ^* (not covered in the course).

Unbiasedness and Consistency

Note that

- the unbiasedness does not imply the consistency;
  - An unbiased estimator can be inconsistent.
- the consistency does not require the unbiasedness;
  - A biased estimator can be consistent (we'll see this later).

Example of an Unbiased Estimator that is not Consistent

Consider the mean estimation problem.

- Let $X \sim P$, and assume that $X_1, \dots, X_n \sim P$ $(i.i.d.)$.
- Define an estimator F_n by

$$\hat{\mu} := F_n(X_1, \dots, X_n) := X_1.$$

i.e., we only use X_1, and discard $X_2, \dots, X_n$.

- Then, this estimator is unbiased: In fact,

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}[X_1] = \mathbb{E}[X] = \mu_P.$$

Example of an Unbiased Estimator that is not Consistent

- However, the variance of the estimate $\hat{\mu}$ is a constant:

$$\mathbb{V}[\hat{\mu}] = \mathbb{E}[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2] = \mathbb{E}[(X_1 - \mu_P)^2] = \mathbb{E}[(X - \mu_P)^2] = \sigma_P^2.$$

- Thus, the MSE of this estimator is

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = \mathbb{E}[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2] = \sigma_P^2.$$

- Thus, the MSE does not decrease to 0, even if $n \to \infty$, i.e., the estimator is not consistent.

This example demonstrates that the unbiasedness does not imply consistency.

- For consistency, we need to make sure that the variance of the estimate decreases to 0 as $n \to \infty$.

Constructing an Unbiased Estimator by Weighting

Consider again the mean estimation problem.

- Let $X \sim P$, and assume $X_1, \ldots, X_n \sim Q$ $(i.i.d.)$.

- Assume that the data distribution Q is different from the target P.

- We show here that we can still construct an unbiased estimator of the mean

$$\mu_P = \mathbb{E}_{X \sim P}[X]$$

from the data $X_1, \ldots, X_n \sim Q$ $(i.i.d.)$.

Constructing an Unbiased Estimator by Weighting

- To this end, assume that distributions P and Q have density functions p and q, respectively.

- Define a weight function by

$$w(x) := \frac{p(x)}{q(x)}, \quad x \in \mathbb{R}$$

- Assume that this weight function is well-defined and bounded:

$$\max_{x \in \mathbb{R}} w(x) =: C < \infty.$$

- Note that this requires $p(x)/q(x) < C$, and thus

$$p(x) < Cq(x) \quad \text{for all } x \in \mathbb{R}.$$

- Thus, if the target density has a positive value $p(x) > 0$, then the data density should also have a positive value $q(x) > 0$.

Constructing an Unbiased Estimator by Weighting

- We assume for simplicity that this weight function $w(x) = p(x)/q(x)$ is known.

  - Otherwise we need to estimate it from data.

- Define an estimator F_n of the mean μ_P as:

$$\hat{\mu} := F_n(X_1, \ldots, X_n) := \frac{1}{n} \sum_{i=1}^{n} w(X_i) X_i.$$

- This is an unbiased estimator of the mean μ_P of P: This can be shown as follows.

Constructing an Unbiased Estimator by Weighting

- Recall that $X_1, \ldots, X_n$ are i.i.d. with $Y \sim Q$. Therefore,

$$\begin{aligned}
\mathbb{E}[\hat{\mu}] &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} w(X_i)X_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[w(X_i)X_i] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[w(Y)Y] \\
&= \mathbb{E}[w(Y)Y] = \int x\ w(x)dQ(x) = \int x\ \frac{p(x)}{q(x)}q(x)dx \\
&= \int x\ p(x)dx = \int x\ dP(x) = \mu_P.
\end{aligned}$$

- Thus, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} w(X_i)X_i$ is an unbiased estimator of μ_P.

Constructing an Unbiased Estimator by Weighting

- On the other hand, the variance of the estimator is

$$\begin{aligned}\mathbb{V}[\hat{\mu}] &= \mathbb{V}[\frac{1}{n}\sum_{i=1}^{n} w(X_i)X_i] = \sum_{i=1}^{n} \mathbb{V}[\frac{1}{n} w(X_i)X_i] \\ &= \sum_{i=1}^{n} \frac{1}{n^2}\mathbb{V}[w(X_i)X_i] = \sum_{i=1}^{n} \frac{1}{n^2}\mathbb{V}[w(Y)Y] \\ &= \frac{1}{n}\mathbb{V}[w(Y)Y].\end{aligned}$$

- This can be upper-bounded as

$$\begin{aligned}\frac{1}{n}\mathbb{V}[w(Y)Y] &= \frac{1}{n}\left(\mathbb{E}[(w(Y)Y)^2] - (\mathbb{E}[w(Y)Y])^2\right) \\ &\leq \frac{1}{n}(\mathbb{E}[C^2Y^2] + \mu_P^2) = \frac{1}{n}(C^2\mathbb{E}[Y^2] - \mu_P^2) \\ &= \frac{1}{n}(C^2(\sigma_Q^2 + \mu_Q^2) - \mu_P^2).\end{aligned}$$

Constructing an Unbiased Estimator by Weighting

- To summarize, the MSE of the estimator is

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = \frac{1}{n}\mathbb{V}[w(Y)Y] \leq \frac{1}{n}(C^2(\sigma_Q^2 + \mu_Q^2) - \mu_P^2).$$

- Therefore, the MSE decreases to 0 as $n \to \infty$:

  - i.e., the estimator $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} w(X_i)X_i$ is consistent in estimating μ_P.

- The weight function $w(x)$ is called the importance weight of a point x.

- The way of constructing an estimator by weighting each sample point X_i by $w(X_i)$ is called importance weighting.

Constructing an Unbiased Estimator by Weighting

- Importance weighting is a widely used technique, examples including:

  - Domain shift adaptation in machine learning.
  - Estimation of treatment effects in causal inference.
  - Monte Carlo for efficient simulations.

- If you are interested in the first, you can for instance look at [Sugiyama and Kawanabe, 2012].

Outline

Variance of Unbiased Estimators may be Large

- We demonstrate here that sometimes biased estimators may be "better" than unbiased estimators.

- The key is an approach called shrinkage or regularization, which is ubiquitous in statistics and machine learning.

Variance of Unbiased Estimators may be Large

- We have seen the bias-variance decomposition of the MSE:

$$\mathbb{E}[\|\hat{\theta}_n - \theta^*\|^2] = \underbrace{\mathbb{E}[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2]}_{\textit{Variance}} + \| \underbrace{\mathbb{E}[\hat{\theta}_n] - \theta^*}_{\textit{Bias}} \|^2$$

- The MSE decomposes into the bias and variance.

- For an unbiased estimator (i.e.,the bias is zero), the MSE is equal to the variance:

$$\mathbb{E}[\|\hat{\theta}_n - \theta^*\|^2] = \underbrace{\mathbb{E}[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2]}_{\textit{Variance}}.$$

- This variance may be large if, e.g.,

- the sample size n is small
- the dimensionality of $\hat{\theta}_n$ is large (in multivariate cases).

- In such a situation, a biased estimator with a lower variance may have a smaller MSE than the unbiased estimator.

Variance Reduction in Mean Estimation

- To describe this, consider the mean estimation problem.

- Let $X \sim P$, and $X_1, \ldots, X_n \sim P$ $(i.i.d.)$.

- We saw that the empirical average

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

is an unbiased estimator of the mean of

$$\mu_P := \mathbb{E}_{X \sim P}[X],$$

and the MSE is given by

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = \frac{\mathbb{V}[X]}{n}.$$

- We'll show that there are biased estimators that have smaller MSE than the empirical average.

Empirical Average as a Least-Squares Solution

- We first show that the empirical average $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the solution to the following optimization problem

$$\hat{\mu} = \arg\min_{\alpha\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}(\alpha - X_i)^2.$$

- i.e., we consider a least-squares problem (fitting a constant α to the data $X_1, \ldots, X_n$).

- To solve this, set the the derivative of the objective function with respect to α to be zero:

$$\frac{d}{d\alpha}\left(\frac{1}{n}\sum_{i=1}^{n}(\alpha - X_i)^2\right) = \frac{1}{n}\sum_{i=1}^{n} 2(\alpha - X_i) = 2\alpha - \frac{2}{n}\sum_{i=1}^{n} X_i = 0.$$

- Thus, the α that minimizes the objective function is

$$\alpha = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

i.e., the empirical average.

Regularized Least Squares and Shrinkage Estimator

- We then consider a modified optimization problem, adding a regularization term:

$$\hat{\mu}_\lambda := \arg\min_{\alpha \in \mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}(\alpha - X_i)^2 + \lambda\alpha^2,$$

where $\lambda \geq 0$ is a regularization constant.

- The solution is given by setting the derivative of the objective function to be 0:

$$\frac{d}{d\alpha}\left(\frac{1}{n}\sum_{i=1}^{n}(\alpha - X_i)^2 + \lambda\alpha^2\right) = \frac{1}{n}\sum_{i=1}^{n} 2(\alpha - X_i) + 2\lambda\alpha$$

$$= 2\alpha - \frac{2}{n}\sum_{i=1}^{n} X_i + 2\lambda\alpha = 2\alpha(1+\lambda) - \frac{2}{n}\sum_{i=1}^{n} X_i = 0.$$

- Thus, the solution is given by

$$\alpha = \frac{1}{(1+\lambda)}\frac{1}{n}\sum_{i=1}^{n} X_i =: \hat{\mu}_\lambda$$

Regularized Least Squares and Shrinkage Estimator

$$\hat{\mu}_\lambda = \frac{1}{(1+\lambda)} \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- Large λ shrinks the solution $\hat{\mu}_\lambda$ towards 0.

  - In this sense, this is called a shrinkage estimator.

- $\lambda = 0$ recovers the empirical average $\hat{\mu}_0 = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Mean Square Error of the Shrinkage Estimator

- The expectation of $\hat{\mu}_\lambda$ is

$$\mathbb{E}[\hat{\mu}_\lambda] = \mathbb{E}[\frac{1}{(1+\lambda)}\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{(1+\lambda)}\mu_P.$$

- Thus, the (squared) bias of $\hat{\mu}_\lambda$ is

$$(\mathbb{E}[\hat{\mu}_\lambda] - \mu_P)^2 = (\frac{1}{(1+\lambda)}\mu_P - \mu_P)^2 = \frac{\lambda^2\mu_P^2}{(1+\lambda)^2}.$$

- Thus, the bias increases as λ increases.

- On the other hand, the variance of $\hat{\mu}_\lambda$ is

$$\mathbb{V}[\hat{\mu}_\lambda] = \mathbb{V}[\frac{1}{1+\lambda}\frac{1}{n}\sum_{i=1}^{n} X_i]$$

$$= \frac{1}{(1+\lambda)^2}\mathbb{V}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{(1+\lambda)^2}\frac{\mathbb{V}[X]}{n}.$$

- Thus, the variance decreases as λ increases.

Mean Square Error of the Shrinkage Estimator

- Thus, the MSE of $\hat{\mu}_\lambda$ is

$$\mathbb{E}[(\hat{\mu}_\lambda - \mu_P)^2] = \mathbb{V}[\hat{\mu}_\lambda] + (\mathbb{E}[\hat{\mu}_\lambda] - \mu_P)^2$$
$$= \frac{1}{(1+\lambda)^2}\frac{\mathbb{V}[X]}{n} + \frac{\lambda^2\mu_P^2}{(1+\lambda)^2}$$

- Let's draw some observations. Assume $\mu_P \neq 0$.

Mean Square Error of the Shrinkage Estimator

- By an easy calculation, the MSE of $\hat{\mu}_\lambda = \frac{1}{(1+\lambda)} \frac{1}{n} \sum_{i=1}^{n} X_i$ can be shown to be smaller than that of the empirical average $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$:

$$\mathbb{E}[(\hat{\mu}_\lambda - \mu_P)^2] < \mathbb{E}[(\hat{\mu} - \mu_P)^2]$$

if $\lambda > 0$ is chosen so that

$$\frac{\lambda}{2+\lambda} \leq \frac{\mathbb{V}[X]}{n\ \mu_P^2}.$$

Some interpretations:

- When $\mathbb{V}[X]/n$ is large (e.g., when n is small), a large λ can be taken (and more shrinkage).

- When the mean μ_P^2 is small, a large λ can be taken (and more shrinkage).

Exercise: Perform numerical experiments to confirm that the shrinkage estimator can have a smaller MSE.

Mean Square Error of the Shrinkage Estimator

- For a right choice of $\lambda > 0$, we need to know $\mathbb{V}[X]$ and μ_P.

  - Therefore this estimator is not practically useful.

- However, under some assumptions (e.g., P is a Gaussian), there is a way of choosing λ without the knowledge of $\mathbb{V}[X]$ and μ_P.

  - This resulting estimator is called the James-Stein estimator; see [Efron and Hastie, 2016, Section7] [Berger, 1985, Section 5.4].

Regularization for Variance Reduction

- Anyway, this example illustrates that artificially introducing a bias is often useful to reduce the variance.

- In this spirit, regularization has been widely used in many statistical methods: e.g.,

  - L_2 and L_1 regularization in regression and classification (supervised learning)
  - Early stopping in optimization algorithms for machine learning algorithms.

- In supervised learning problems, a good regularization constant can be chosen by, e.g., cross validation

  - See e.g. the MALIS and ASI courses.

Summary of the Lecture

- We introduced several important concepts in statistical estimation.

- When constructing statistical estimators, always pay attention to

  - what is your quantity of interest (in the population).
  - how your data were generated.
  - whether your estimator is biased or unbiased.
  - how much your estimate would have variance.

Berger, J. O. (1985).
Statistical Decision Theory and Bayesian Analysis.
Springer Science & Business Media.

Efron, B. and Hastie, T. (2016).
Computer Age Statistical Inference.
Cambridge University Press.

Sugiyama, M. and Kawanabe, M. (2012).
Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation.
MIT Press, Cambridge, MA, USA.