Introduction to Data Science With Probability and Statistics
Lecture 21: Unbiased Estimators and Mean Squared Error

CSCI 3022 - Summer 2020
Sourav Chakraborty
Dept. of Computer Science
University of Colorado Boulder

Estimators

- One of the tasks is to use a dataset to estimate a quantity of interest.
- We should be able to deal with a situation where the dataset is modeled as one of the parameters of the model distribution or as a certain function of the parameters.

ESTIMATE. An *estimate* is a value t that only depends on the dataset $x_1, x_2, \ldots, x_n$, i.e., t is some function of the dataset only:

$$t = h(x_1, x_2, \ldots, x_n).$$

Eg: $\bar{x} = \frac{\sum x_i}{n}$

Estimators

One can often think of several estimates for the parameter of interest. This raises questions like:-

- When is one estimate better than another?
- Does there exist a best possible estimate?

We can never say which of the two estimate values 'e_1' and 'e_2' computed from a dataset is closer to the 'true' parameter. This is because:-

- ➔ The measurements and the corresponding estimates are subject to randomness.
- ➔ One of the things we can say for each of them is how likely it is that they are within a given distance from the 'true' parameter.

Note that estimators are special cases of 'sample statistics'.

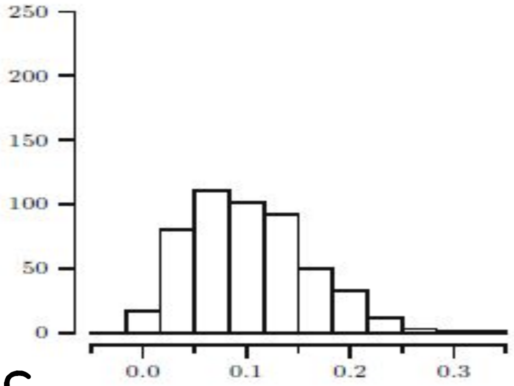
Estimators

ESTIMATOR. Let $t = h(x_1, x_2, \ldots, x_n)$ be an estimate based on the dataset $x_1, x_2, \ldots, x_n$. Then t is a realization of the random variable

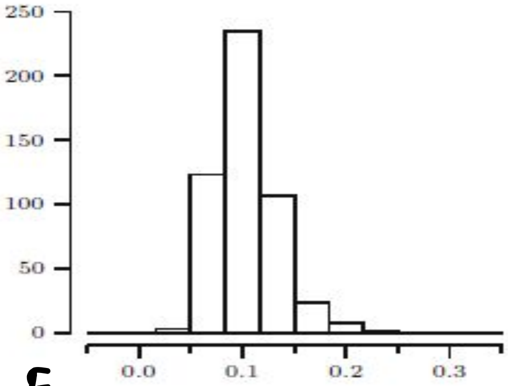
$$T = h(X_1, X_2, \ldots, X_n).$$

The random variable T is called an *estimator*.

Eg: $\bar{X}_n = \frac{\sum X_i}{n}$



E_1



E_2

sampling distributions of estimators

The sampling distribution and unbiasedness

THE SAMPLING DISTRIBUTION. Let $T = h(X_1, X_2, \ldots, X_n)$ be an estimator based on a random sample $X_1, X_2, \ldots, X_n$. The probability distribution of T is called the *sampling distribution* of T.

DEFINITION. An estimator T is called an *unbiased* estimator for the parameter θ, if

$$\mathrm{E}[T] = \theta$$

irrespective of the value of θ. The difference $\mathrm{E}[T] - \theta$ is called the *bias* of T; if this difference is nonzero, then T is called *biased*.

Unbiased estimators for expectation and variance

UNBIASED ESTIMATORS FOR EXPECTATION AND VARIANCE. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with finite expectation μ and finite variance σ^2. Then

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is an *unbiased estimator for* μ and

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

is an *unbiased estimator for* σ^2.

Unbiased estimators for expectation and variance

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots X_n}{n}$$

$$E[\bar{X}_n] = E\left[\frac{X_1 + \cdots + X_n}{n}\right]$$

$$= \frac{1}{n} E[X_1 + \cdots + X_n]$$

$$= \frac{1}{n}\left(E[X_1] + \cdots + E[X_n]\right)$$

$$= \frac{1}{n} \cdot n\mu = \mu$$

$$\therefore \boxed{E[\bar{X}_n] = \mu}$$

$$E[S_n^2] = \frac{1}{n-1} \sum_{i=1}^{n} E\left[(X_i - \bar{X}_n)^2\right]$$

$$E\left[\sum \left(X_i^2 - 2\bar{X}_n X_i + \bar{X}^2\right)\right]$$

$$= E\left[\sum X_i^2 - 2\bar{X}_n \underbrace{\sum X_i} + n\bar{X}_n^2\right]$$

$$= E\left[\sum X_i^2 - 2\bar{X}_n \cdot \underbrace{n\bar{X}_n} + n\bar{X}_n^2\right]$$

$$= E\left[\sum X_i^2 - n\bar{X}_n^2\right] = \sum\left(E[X_i^2] - E[n\bar{X}_n^2]\right)$$

Now;

$$E[X_i^2] = Var(X_i) + \mu^2$$

Unbiased estimators for expectation and variance

$$\therefore \sum\left(E[x_i^2] - n(E[\bar{x}^2])\right) = \sum\left[(\sigma^2+\mu^2) - n.\left(\frac{\sigma^2}{n}+\mu^2\right)\right]$$

$$= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2$$

$$= \sigma^2(n-1)$$

- This explains why we divide by $n - 1$ in the formula for S_n^2; only in this case S_n^2 is an unbiased estimator for the "true" variance σ^2.
- If we would divide by n instead of $n-1$, we would obtain an estimator with negative bias; it would systematically produce too-small estimates for σ^2.

So, $$E[S_n^2] = \frac{1}{n-1} E\left[\sum (x_i - \bar{x}_n)^2\right]$$

$$= \frac{1}{\cancel{n-1}} \cdot \sigma^2 \cancel{(n-1)} = \boxed{\sigma^2}$$

Unbiased estimators for expectation and variance

Consider the following estimator for σ^2:

$$V_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Compute the bias $\mathrm{E}[V_n^2] - \sigma^2$ for this estimator, where you can keep computations simple by realizing that $V_n^2 = (n-1)S_n^2/n$

Unbiased estimators for expectation and variance

Consider the following estimator for σ^2:

$$V_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Compute the bias $\mathrm{E}[V_n^2] - \sigma^2$ for this estimator, where you can keep computations simple by realizing that $V_n^2 = (n-1)S_n^2/n$

$$E[V_n^2] = E\left[\frac{n-1}{n} S_n^2\right] = \frac{n-1}{n} \cdot E[S_n^2] = \frac{n-1}{n}\sigma^2$$

Bias: $E[V_n^2] - \sigma^2 = -\frac{\sigma^2}{n} < 0$

Unbiased estimators

General Fact: Unbiasedness does not always carry over, i.e., if T is an unbiased estimator for a parameter θ, then $g(T)$ does not have to be an unbiased estimator for $g(\theta)$.

Exception:

if $g(T) = aT + b$; & if T is unbiased for θ

$\because \quad E[aT+b] = aE[T]+b = a\theta + b$

$\therefore \quad aT+b$ is unbiased for $\underline{\underline{a\theta + b}}$.

Unbiased estimators

Suppose our dataset is a realization of a random sample $X_1, X_2, \ldots, X_n$ from a uniform distribution on the interval *[−θ, θ]*, where *θ* is unknown.

a. Show that $T = \frac{3}{n}(X_1^2 + X_2^2 + \cdots + X_n^2)$ is an unbiased estimator for θ^2.

Unbiased estimators

Suppose our dataset is a realization of a random sample $X_1, X_2, \ldots, X_n$ from a uniform distribution on the interval $[-\theta, \theta]$, where θ is unknown.

a. Show that $T = \frac{3}{n}(X_1^2 + X_2^2 + \cdots + X_n^2)$ is an unbiased estimator for θ^2.

We have to show $E[T] = \theta^2$;

So; $E[T] = \frac{3}{n}\left(E[X_1^2] + E[X_2^2] + \cdots E[X_n^2]\right)$

$= \frac{3}{n}\left(\frac{\theta^2}{3} + \cdots + \frac{\theta^2}{3}\right)$

$= \frac{3}{n} \cdot n \cdot \frac{\theta^2}{3} = \boxed{\theta^2}$

$$E[X_i^2] = \int_{-\theta}^{\theta} x^2 \cdot \frac{1}{2\theta} \cdot dx$$

$$= \frac{1}{2\theta}\left[\frac{x^3}{3}\right]_{-\theta}^{\theta}$$

$$= \frac{1}{2\theta} \cdot \left[\frac{\theta^3}{3} + \frac{\theta^3}{3}\right] = \frac{\theta^2}{3}$$

Unbiased estimators

Suppose our dataset is a realization of a random sample $X_1, X_2, \ldots, X_n$ from a uniform distribution on the interval $[-\theta, \theta]$, where θ is unknown.

b. Is $\sqrt{T}$ also an unbiased estimator for θ? If not, argue whether it has positive or negative bias.

Skip

Unbiased estimators

Suppose our dataset is a realization of a random sample $X_1, X_2, \ldots, X_n$ from a uniform distribution on the interval $[-\theta, \theta]$, where θ is unknown.

b. Is $\sqrt{T}$ also an unbiased estimator for θ? If not, argue whether it has positive or negative bias.

Skip

Unbiased estimators

Suppose the random variables $X_1, X_2, \ldots, X_n$ have the same expectation μ.

a. Is $S = \frac{1}{2}X_1 + \frac{1}{3}X_2 + \frac{1}{6}X_3$ an unbiased estimator for μ?

Unbiased estimators

Suppose the random variables $X_1, X_2, \ldots, X_n$ have the same expectation μ.

a. Is $S = \frac{1}{2}X_1 + \frac{1}{3}X_2 + \frac{1}{6}X_3$ an unbiased estimator for μ? Yes.

$$E[S] = \frac{1}{2}E[X_1] + \frac{1}{3}E[X_2] + \frac{1}{6}E[X_3]$$

$$= \left(\frac{1}{2} + \frac{1}{3} + \frac{1}{6}\right)\mu = \underline{\underline{\mu}}$$

Unbiased estimators

Suppose the random variables $X_1, X_2, \ldots, X_n$ have the same expectation μ.

b. Under what conditions on constants $a_1, a_2, \ldots, a_n$ is

$$T = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

an unbiased estimator for μ?

Unbiased estimators

Suppose the random variables $X_1, X_2, \ldots, X_n$ have the same expectation μ.

b. Under what conditions on constants $a_1, a_2, \ldots, a_n$ is

$$T = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

an unbiased estimator for μ?

$$E[T] = E[a_1X_1 + a_2X_2 + \cdots + a_nX_n]$$

$$= a_1E[X_1] + a_2E[X_2] + \cdots + a_nE[X_n]$$

$$= a_1\mu + a_2\mu + \cdots + a_n\mu \quad \text{—— } 1$$

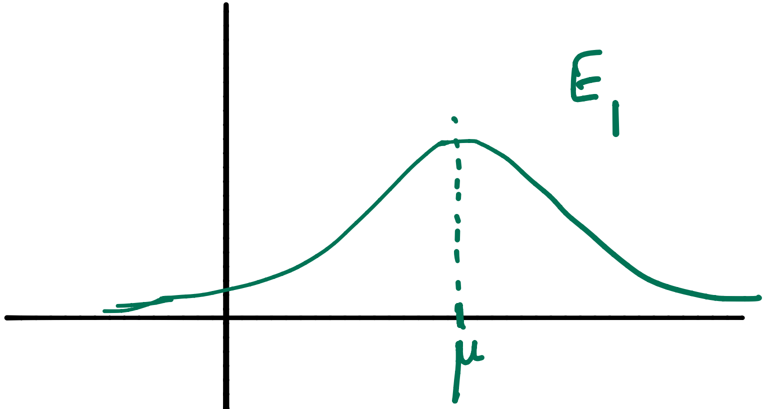
$$\therefore \; a_1 + a_2 + \cdots + a_n = 1 \text{ to make } 1 = \mu.$$

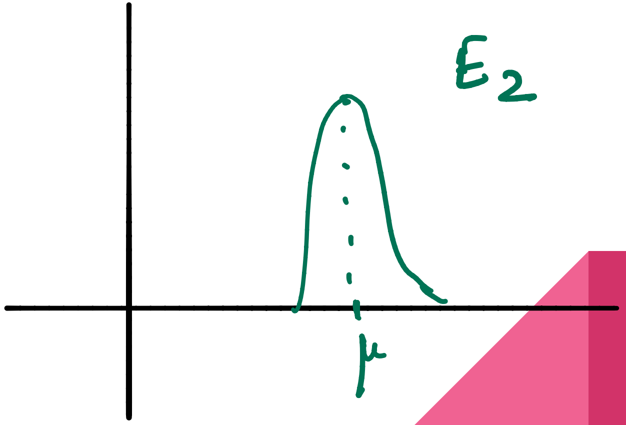
Efficiency and mean squared error

- If several unbiased estimators for the same parameter of interest exist, we need a criterion for comparison of these estimators.
- A natural criterion is some measure of spread of the estimators around the parameter of interest.
- For unbiased estimators we will use variance.
- For arbitrary estimators we introduce the notion of mean squared error (MSE), which combines variance and bias.

Variance of an estimator

EFFICIENCY. Let T_1 and T_2 be two unbiased estimators for the same parameter θ. Then estimator T_2 is called *more efficient* than estimator T_1 if $\mathrm{Var}(T_2) < \mathrm{Var}(T_1)$, irrespective of the value of θ.





Mean squared error

Definition. Let T be an estimator for a parameter θ. The *mean squared error* of T is the number $\mathrm{MSE}(T) = \mathrm{E}\left[(T-\theta)^2\right]$.

$$\begin{aligned}
\mathrm{MSE}(T) &= \mathrm{E}\left[(T-\theta)^2\right] \\
&= \mathrm{E}\left[(T-\mathrm{E}[T]+\mathrm{E}[T]-\theta)^2\right] \\
&= \mathrm{E}\left[(T-\mathrm{E}[T])^2\right] + 2\mathrm{E}[T-\mathrm{E}[T]]\,(\mathrm{E}[T]-\theta) + (\mathrm{E}[T]-\theta)^2 \\
&= \mathrm{Var}(T) + (\mathrm{E}[T]-\theta)^2.
\end{aligned}$$

Mean squared error

Given are two estimators S and T for a parameter θ. Furthermore it is known that $Var(S) = 40$ and $Var(T) = 4$.

a. Suppose that we know that $E[S] = \theta$ and $E[T] = \theta + 3$. Which estimator would you prefer, and why?

Mean squared error

Given are two estimators *S* and *T* for a parameter *θ*. Furthermore it is known that *Var(S) = 40* and *Var(T) = 4*.

a. Suppose that we know that *E[S] = θ* and *E[T] = θ + 3*. Which estimator would you prefer, and why?

T

$$MSE(S) = Var(S) + (E[S] - \theta)^2 = 40$$

$$MSE(T) = Var(T) + (E[T] - \theta)^2 = 4 + 9 = \boxed{13}$$

Mean squared error

Given are two estimators S and T for a parameter θ. Furthermore it is known that $Var(S) = 40$ and $Var(T) = 4$.

b. Suppose that we know that $E[S] = \theta$ *and* $E[T] = \theta + a$ for some positive number a. For each a, which estimator would you prefer, and why?

Mean squared error

Given are two estimators *S* and *T* for a parameter *θ*. Furthermore it is known that *Var(S) = 40* and *Var(T) = 4*.

b. Suppose that we know that *E[S] = θ and E[T] = θ + a* for some positive number *a*. For each *a*, which estimator would you prefer, and why?

- $MSE(S) = 40$.
- $MSE(T) = 4 + a^2$

$\therefore \quad 4 + a^2 < 40$

$\therefore \quad a < 6$ we will prefer T

else S.

Mean squared error

Suppose we have a random sample $X_1, \ldots, X_n$ from an $Exp(\lambda)$ distribution. Suppose we want to estimate the mean $1/\lambda$. Given an estimator:

$$T_1 = \bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

is an unbiased estimator of $1/\lambda$. Let M_n be the minimum of $X_1, X_2, \ldots, X_n$. M_n has an $Exp(n\lambda)$ distribution. Given that:

$$T_2 = nM_n$$

is another unbiased estimator for $1/\lambda$. Which of the estimators T_1 and T_2 would you choose for estimating the mean $1/\lambda$? Substantiate your answer.

Imp: $Var(x_i) = \frac{1}{\lambda^2}$ if $x_i \sim Exp(\lambda)$

Now;

$$Var(T_1) = Var\left(\frac{x_1}{n} + \frac{X_2}{n} + \cdots + \frac{x_n}{n}\right) = \frac{1}{n^2}\left(\sum Var(x_i)\right)$$

$$= \frac{1}{n^2}\left(\frac{1}{\lambda^2} + \frac{1}{\lambda^2} + \cdots \frac{1}{\lambda^2}\right) = \frac{1}{n^2} \cdot \frac{n}{\lambda^2} = \boxed{\frac{1}{n\lambda^2}}$$

$$\frac{1}{n\lambda^2} < \frac{1}{\lambda^2}$$

$Var(M_n) = \frac{1}{\lambda^2 n^2}$; $M_n \sim Exp(n\lambda)$

$$\therefore Var(T_2) = Var(nM_n) = n^2 \cdot \frac{1}{n^2 \cdot \lambda^2} = \boxed{\frac{1}{\lambda^2}}$$

Next: Linear Regression