YSU ASDS, Statistics, Fall 2019

Lecture 17

Michael Poghosyan

19 Oct 2019

Contents

- Consistency
- Fisher Information
- Cramer-Rao Lower Bound (Cramer-Rao Inequality)
- MVUE

Last Lecture ReCap

- Give the Bias-Variance Decomposition

Last Lecture ReCap

- Give the Bias-Variance Decomposition
- What is the Standard Error?

Last Lecture ReCap

- Give the Bias-Variance Decomposition
- What is the Standard Error?
- What is the Estimated Standard Error?

Last Lecture ReCap

- Give the Bias-Variance Decomposition
- What is the Standard Error?
- What is the Estimated Standard Error?
- What is the MVUE?

Consistency

Another important and desirable property of an Estimator is the Consistency: this is about its behaviour when we increase the number of Datapoints:

Consistency

Another important and desirable property of an Estimator is the Consistency: this is about its behaviour when we increase the number of Datapoints:

Definition: A point estimator $\hat{\theta}_n$ of the parameter θ is called

- **consistent**, if $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ for any $\theta \in \Theta$;

Consistency

Another important and desirable property of an Estimator is the Consistency: this is about its behaviour when we increase the number of Datapoints:

Definition: A point estimator $\hat{\theta}_n$ of the parameter θ is called

- **consistent**, if $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ for any $\theta \in \Theta$;
- **strongly consistent**, if $\hat{\theta}_n \xrightarrow{a.s.} \theta$ for any $\theta \in \Theta$;

Consistency

Another important and desirable property of an Estimator is the Consistency: this is about its behaviour when we increase the number of Datapoints:

Definition: A point estimator $\hat{\theta}_n$ of the parameter θ is called

- **consistent**, if $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ for any $\theta \in \Theta$;
- **strongly consistent**, if $\hat{\theta}_n \xrightarrow{a.s.} \theta$ for any $\theta \in \Theta$;
- **weakly or Mean Square consistent**, if $\hat{\theta}_n \xrightarrow{q.m.} \theta$ for any $\theta \in \Theta$, i.e., if

$$MSE(\hat{\theta}_n, \theta) = \mathbb{E}_\theta((\hat{\theta}_n - \theta)^2) \to 0 \qquad \forall \theta \in \Theta.$$

Example

Example: Consider a Random Sample

$$X_1, X_2, ..., X_n \sim Bernoulli(p),$$

where p is our unknown Parameter to be estimated.

Example

Example: Consider a Random Sample

$$X_1, X_2, ..., X_n \sim Bernoulli(p),$$

where p is our unknown Parameter to be estimated.

We consider the following Estimator for p:

$$\hat{p} = \frac{X_1 + X_2 + ... + X_n}{n+1}.$$

Example

Example: Consider a Random Sample

$$X_1, X_2, ..., X_n \sim Bernoulli(p),$$

where p is our unknown Parameter to be estimated.

We consider the following Estimator for p:

$$\hat{p} = \frac{X_1 + X_2 + ... + X_n}{n+1}.$$

Then:

- $\hat{p}$ is a Biased Estimator for p;
- $\hat{p}$ is Consistent Estimator for p.

Proof: OTB

Some Properties

Proposition:

- If the estimator is strongly consistent, then it is consistent;

Some Properties

Proposition:

- If the estimator is strongly consistent, then it is consistent;
- If the estimator is weakly consistent, then it is consistent;

Some Properties

Proposition:

- If the estimator is strongly consistent, then it is consistent;
- If the estimator is weakly consistent, then it is consistent;
- If $\hat{\theta}_n$ is an *Unbiased Estimator* for θ and

$$Var(\hat{\theta}_n) \to 0, \qquad n \to \infty,$$

  then $\hat{\theta}_n$ is consistent;

Some Properties

Proposition:

- If the estimator is strongly consistent, then it is consistent;
- If the estimator is weakly consistent, then it is consistent;
- If $\hat{\theta}_n$ is an *Unbiased Estimator* for θ and

$$Var(\hat{\theta}_n) \to 0, \qquad n \to \infty,$$

  then $\hat{\theta}_n$ is consistent;
- If $\hat{\theta}_n$ is an *Asymptotically Unbiased Estimator* for θ and

$$Var(\hat{\theta}_n) \to 0, \qquad n \to \infty,$$

  then $\hat{\theta}_n$ is consistent.

Example

Example: Assume $X_1, X_2, ..., X_n, ...$ are IID from a Distribution with the Mean μ, Variance σ^2 and finite 4-th order Moment $\mathbb{E}(X_1^4)$.

Example

Example: Assume $X_1, X_2, ..., X_n, ...$ are IID from a Distribution with the Mean μ, Variance σ^2 and finite 4-th order Moment $\mathbb{E}(X_1^4)$. Let us denote this as

$$X_1, X_2, ..., X_n \sim \mathcal{F}_{\mu,\sigma^2}$$

and let our aim be to estimate σ^2 based on $X_1, X_2, ...$

Example

Example: Assume $X_1, X_2, ..., X_n, ...$ are IID from a Distribution with the Mean μ, Variance σ^2 and finite 4-th order Moment $\mathbb{E}(X_1^4)$. Let us denote this as

$$X_1, X_2, ..., X_n \sim \mathcal{F}_{\mu,\sigma^2}$$

and let our aim be to estimate σ^2 based on $X_1, X_2, ...$

Let's consider the following Estimator for σ^2, the Sample Variance with the denominator n:

$$\widehat{\sigma^2} = \frac{\displaystyle\sum_{k=1}^{n}(X_k - \overline{X}_n)^2}{n}$$

Example

Example: Assume $X_1, X_2, ..., X_n, ...$ are IID from a Distribution with the Mean μ, Variance σ^2 and finite 4-th order Moment $\mathbb{E}(X_1^4)$. Let us denote this as

$$X_1, X_2, ..., X_n \sim \mathcal{F}_{\mu,\sigma^2}$$

and let our aim be to estimate σ^2 based on $X_1, X_2, ...$.

Let's consider the following Estimator for σ^2, the Sample Variance with the denominator n:

$$\widehat{\sigma^2} = \frac{\sum\limits_{k=1}^{n}(X_k - \overline{X}_n)^2}{n}$$

Then

- $\widehat{\sigma^2}$ is Biased;

Example

Example: Assume $X_1, X_2, ..., X_n, ...$ are IID from a Distribution with the Mean μ, Variance σ^2 and finite 4-th order Moment $\mathbb{E}(X_1^4)$. Let us denote this as

$$X_1, X_2, ..., X_n \sim \mathcal{F}_{\mu,\sigma^2}$$

and let our aim be to estimate σ^2 based on $X_1, X_2, ...$

Let's consider the following Estimator for σ^2, the Sample Variance with the denominator n:

$$\widehat{\sigma^2} = \frac{\displaystyle\sum_{k=1}^{n} (X_k - \overline{X}_n)^2}{n}$$

Then

- $\widehat{\sigma^2}$ is Biased;
- $\widehat{\sigma^2}$ is Consistent.

Proof: OTB. Use the relation $\widehat{\sigma^2} = \frac{\sum_{k=1}^{n} (X_k)^2}{n} - \left(\frac{\sum_{k=1}^{n} X_k}{n}\right)^2$.

Choosing Good Estimators

We have talked about different properties of Estimators:

Choosing Good Estimators

We have talked about different properties of Estimators:

- Unbiasedness

Choosing Good Estimators

We have talked about different properties of Estimators:

- Unbiasedness
- Consistency

Choosing Good Estimators

We have talked about different properties of Estimators:

- Unbiasedness
- Consistency

Unbiasedness is good if we can do resamplings: to estimate our parameter, we can calculate estimates for many samples and average the obtained values - that average will be close to the real value of the parameter;

Choosing Good Estimators

We have talked about different properties of Estimators:

- Unbiasedness
- Consistency

Unbiasedness is good if we can do resamplings: to estimate our parameter, we can calculate estimates for many samples and average the obtained values - that average will be close to the real value of the parameter;

Consistency is good when we have a large sample: we can just calculate the estimate for that sample, and the obtained value will be close to the real value of the parameter with high probability.

Choosing Good Estimators

We have talked about different properties of Estimators:

- Unbiasedness
- Consistency

Unbiasedness is good if we can do resamplings: to estimate our parameter, we can calculate estimates for many samples and average the obtained values - that average will be close to the real value of the parameter;

Consistency is good when we have a large sample: we can just calculate the estimate for that sample, and the obtained value will be close to the real value of the parameter with high probability.

Of course, it is better to have dva v odnom: an Unbiased and Consistent Estimator $\ddot\smile$

Choosing Good Estimators

We have talked about different properties of Estimators:

- Unbiasedness
- Consistency

Unbiasedness is good if we can do resamplings: to estimate our parameter, we can calculate estimates for many samples and average the obtained values - that average will be close to the real value of the parameter;

Consistency is good when we have a large sample: we can just calculate the estimate for that sample, and the obtained value will be close to the real value of the parameter with high probability.

Of course, it is better to have dva v odnom: an Unbiased and Consistent Estimator $\ddot\smile$

And also, the universal measure for goodness is: *an Estimator is good if it has a small MSE*.

Question

Question: Is sampling 200 times with a sample size 10 the same as sampling once with a size 2000?

Question

Question: Is sampling 200 times with a sample size 10 the same as sampling once with a size 2000? And, in general, isn't it the same to sample many times with a small sample size, as to take a large dataset once?

Question

Question: Is sampling 200 times with a sample size 10 the same as sampling once with a size 2000? And, in general, isn't it the same to sample many times with a small sample size, as to take a large dataset once?

Answer: No, in general. This is because, say,

- we can do a lot of resamplings even when our dataset is not big enough, but one large sample will not be available

Question

Question: Is sampling 200 times with a sample size 10 the same as sampling once with a size 2000? And, in general, isn't it the same to sample many times with a small sample size, as to take a large dataset once?

Answer: No, in general. This is because, say,

- we can do a lot of resamplings even when our dataset is not big enough, but one large sample will not be available
- when taking a large sample, we will take each individual just once. But if we are doing resamplings, we can have the same individual in different samples.

MVUE

Let us go back to MVUEs.

MVUE

Let us go back to MVUEs. So we restrict our attention to only Unbiased Estimators, and we want to find among them the one with the minimal MSE or, which is the same, the one with minimal Variance.

MVUE

Let us go back to MVUEs. So we restrict our attention to only Unbiased Estimators, and we want to find among them the one with the minimal MSE or, which is the same, the one with minimal Variance.

To find the one with the minimal Variance, we can use the Cramer-Rao inequality. But before stating that inequality, we need the notion of the Fisher Information.

Fisher Information

Assume we have a parametric family of distributions $\mathcal{F}_\theta$, $\theta \in \Theta \subset \mathbb{R}$, and $f(x|\theta)$ is the PD(M)F of $\mathcal{F}_\theta$.

Fisher Information

Assume we have a parametric family of distributions $\mathcal{F}_\theta$, $\theta \in \Theta \subset \mathbb{R}$, and $f(x|\theta)$ is the PD(M)F of $\mathcal{F}_\theta$.

Definition: The following quantity is called **the Fisher Information** of the parametric family $\mathcal{F}_\theta$:

$$I(\theta) = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\ln f(X|\theta)\right) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right)^2\right],$$

where $X \sim \mathcal{F}_\theta$.

Example

Example: Calculate the Fisher Information for the *Bernoulli*(p) family

Solution: OTB

Example

Example: Calculate the Fisher Information for the $Bernoulli(p)$ family

Solution: OTB

Example: Calculate the Fisher Information for the $Exp(\lambda)$ family

Solution: OTB

Example

Example: Calculate the Fisher Information for the *Bernoulli*(p) family

Solution: OTB

Example: Calculate the Fisher Information for the *Exp*(λ) family

Solution: OTB

Example: Calculate the Fisher Information for the $\mathcal{N}(\mu, \sigma^2)$ family (separately for the Parameter μ and σ^2)

Solution: OTB

Fisher Information, cont'd

Another interpretation of the Fisher Information is the following.

Fisher Information, cont'd

Another interpretation of the Fisher Information is the following.

It is easy to see that (under the regularity conditions)

$$\mathbb{E}\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right) = 0.$$

Fisher Information, cont'd

Another interpretation of the Fisher Information is the following.

It is easy to see that (under the regularity conditions)

$$\mathbb{E}\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right) = 0.$$

Then,

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right)^2\right] = Var\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right).$$

Fisher Information, cont'd

Another interpretation of the Fisher Information is the following.

It is easy to see that (under the regularity conditions)

$$\mathbb{E}\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right) = 0.$$

Then,

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right)^2\right] = Var\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right).$$

So the Fisher Information is the Variance of the Score function

$$\frac{\partial}{\partial\theta}\ln f(X|\theta).$$