AUA CS 108, Statistics, Fall 2019

Lecture 30

Michael Poghosyan
YSU, AUA
michael@ysu.am, mpoghosyan@aua.am

1 Nov 2019

Contents

Last Lecture ReCap

- Nothing new $\ddot\smile$

Some Notes about MLE

Note: The MLE does not always exist;

Some Notes about MLE

Note: The MLE does not always exist;

Note: The MLE is not always unique;

Some Notes about MLE

Note: The MLE does not always exist;

Note: The MLE is not always unique;

Note: Sometimes it is necessary to find it numerically, say, using the NR Method

Some Notes about MLE

Note: The MLE does not always exist;

Note: The MLE is not always unique;

Note: Sometimes it is necessary to find it numerically, say, using the NR Method

Note: It is remarkable, that ML Estimators, in general (if they exist, of course $\ddot\smile$), possess some nice properties. These properties make MLE one of the widely used methods of Estimation.

Properties of the MLE

It can be proven that, under some regularity conditions on the Parametric Family $\mathcal{F}_\theta$,

- The MLE $\hat{\theta}_n^{MLE}$ is Consistent, i.e.,

Properties of the MLE

It can be proven that, under some regularity conditions on the Parametric Family $\mathcal{F}_\theta$,

- The MLE $\hat{\theta}_n^{MLE}$ is Consistent, i.e.,

$$\hat{\theta}_n^{MLE} \xrightarrow{\mathbb{P}} \theta$$

Properties of the MLE

It can be proven that, under some regularity conditions on the Parametric Family $\mathcal{F}_\theta$,

- The MLE $\hat{\theta}_n^{MLE}$ is Consistent, i.e.,

$$\hat{\theta}_n^{MLE} \xrightarrow{\mathbb{P}} \theta$$

- The MLE $\hat{\theta}_n^{MLE}$ is Asymptotically Normal and Efficient:

$$\frac{\hat{\theta}_n^{MLE} - \theta}{\sqrt{\frac{1}{n \cdot \mathcal{I}(\theta)}}} \xrightarrow{D} \mathcal{N}(0, 1)$$

Properties of the MLE

It can be proven that, under some regularity conditions on the Parametric Family $\mathcal{F}_\theta$,

- The MLE $\hat{\theta}_n^{MLE}$ is Consistent, i.e.,

$$\hat{\theta}_n^{MLE} \xrightarrow{\mathbb{P}} \theta$$

- The MLE $\hat{\theta}_n^{MLE}$ is Asymptotically Normal and Efficient:

$$\frac{\hat{\theta}_n^{MLE} - \theta}{\sqrt{\frac{1}{n \cdot \mathcal{I}(\theta)}}} \xrightarrow{D} \mathcal{N}(0, 1)$$

  or, as people usually use:

$$\sqrt{n} \cdot \left(\hat{\theta}_n^{MLE} - \theta\right) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{\mathcal{I}(\theta)}\right),$$

Properties of the MLE

It can be proven that, under some regularity conditions on the Parametric Family $\mathcal{F}_\theta$,

- The MLE $\hat{\theta}_n^{MLE}$ is Consistent, i.e.,

$$\hat{\theta}_n^{MLE} \xrightarrow{\mathbb{P}} \theta$$

- The MLE $\hat{\theta}_n^{MLE}$ is Asymptotically Normal and Efficient:

$$\frac{\hat{\theta}_n^{MLE} - \theta}{\sqrt{\frac{1}{n \cdot \mathcal{I}(\theta)}}} \xrightarrow{D} \mathcal{N}(0, 1)$$

or, as people usually use:

$$\sqrt{n} \cdot \left(\hat{\theta}_n^{MLE} - \theta\right) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{\mathcal{I}(\theta)}\right),$$

or, put in other way,

$$\hat{\theta}_n^{MLE} \overset{D}{\approx} \mathcal{N}\left(\theta, \frac{1}{n \cdot \mathcal{I}(\theta)}\right)$$

Properties of the MLE, Cont'd

Let us give an interpretation to this last property:

$$\hat{\theta}_n^{MLE} \overset{D}{\approx} \mathcal{N}\left(\theta, \frac{1}{n \cdot \mathcal{I}(\theta)}\right)$$

Properties of the MLE, Cont'd

Let us give an interpretation to this last property:

$$\hat{\theta}_n^{MLE} \stackrel{D}{\approx} \mathcal{N}\left(\theta, \frac{1}{n \cdot \mathcal{I}(\theta)}\right)$$

We can interpret this as:

- $\hat{\theta}_n^{MLE}$ is approximately (Asymptotically) Normal;

Properties of the MLE, Cont'd

Let us give an interpretation to this last property:

$$\hat{\theta}_n^{MLE} \overset{D}{\approx} \mathcal{N}\left(\theta, \frac{1}{n \cdot \mathcal{I}(\theta)}\right)$$

We can interpret this as:

- $\hat{\theta}_n^{MLE}$ is approximately (Asymptotically) Normal;
- The Mean of $\hat{\theta}_n^{MLE}$ is approximately θ, so it is Asymptotically Unbiased;

Properties of the MLE, Cont'd

Let us give an interpretation to this last property:

$$\hat{\theta}_n^{MLE} \overset{D}{\approx} \mathcal{N}\left(\theta, \frac{1}{n \cdot \mathcal{I}(\theta)}\right)$$

We can interpret this as:

- $\hat{\theta}_n^{MLE}$ is approximately (Asymptotically) Normal;
- The Mean of $\hat{\theta}_n^{MLE}$ is approximately θ, so it is Asymptotically Unbiased;
- The Variance of $\hat{\theta}_n^{MLE}$ is approximately $\dfrac{1}{n \cdot \mathcal{I}(\theta)}$. And we know that, by CR LB, this is the possible smallest Variance we can have with Unbiased Estimators.

Properties of the MLE, Cont'd

Let us give an interpretation to this last property:

$$\hat{\theta}_n^{MLE} \overset{D}{\approx} \mathcal{N}\left(\theta, \frac{1}{n \cdot \mathcal{I}(\theta)}\right)$$

We can interpret this as:

- $\hat{\theta}_n^{MLE}$ is approximately (Asymptotically) Normal;
- The Mean of $\hat{\theta}_n^{MLE}$ is approximately θ, so it is Asymptotically Unbiased;
- The Variance of $\hat{\theta}_n^{MLE}$ is approximately $\dfrac{1}{n \cdot \mathcal{I}(\theta)}$. And we know that, by CR LB, this is the possible smallest Variance we can have with Unbiased Estimators. So we interpret this as **MLE is Asymptotically Efficient**.

Properties of the MLE, Cont'd

Let us give an interpretation to this last property:

$$\hat{\theta}_n^{MLE} \overset{D}{\approx} \mathcal{N}\left(\theta, \frac{1}{n \cdot \mathcal{I}(\theta)}\right)$$

We can interpret this as:

- $\hat{\theta}_n^{MLE}$ is approximately (Asymptotically) Normal;
- The Mean of $\hat{\theta}_n^{MLE}$ is approximately θ, so it is Asymptotically Unbiased;
- The Variance of $\hat{\theta}_n^{MLE}$ is approximately $\dfrac{1}{n \cdot \mathcal{I}(\theta)}$. And we know that, by CR LB, this is the possible smallest Variance we can have with Unbiased Estimators. So we interpret this as **MLE is Asymptotically Efficient**.

So, MLE is **Consistent** and **Asymptotocally Efficient**.

Properties of the MLE, Cont'd

Let us give an interpretation to this last property:

$$\hat{\theta}_n^{MLE} \stackrel{D}{\approx} \mathcal{N}\left(\theta, \frac{1}{n \cdot \mathcal{I}(\theta)}\right)$$

We can interpret this as:

- $\hat{\theta}_n^{MLE}$ is approximately (Asymptotically) Normal;
- The Mean of $\hat{\theta}_n^{MLE}$ is approximately θ, so it is Asymptotically Unbiased;
- The Variance of $\hat{\theta}_n^{MLE}$ is approximately $\dfrac{1}{n \cdot \mathcal{I}(\theta)}$. And we know that, by CR LB, this is the possible smallest Variance we can have with Unbiased Estimators. So we interpret this as **MLE is Asymptotically Efficient**.

So, MLE is **Consistent** and **Asymptotocally Efficient**. And this is why, for large Sample Size n, MLE is the Top 1 Choice, is (almost) unbeatable.

Properties of the MLE, Cont'd

- Also,

$$\frac{\hat{\theta}_n^{MLE} - \theta}{\sqrt{\frac{1}{n \cdot \mathcal{I}\left(\hat{\theta}_n^{MLE}\right)}}} \xrightarrow{D} \mathcal{N}\left(0, 1\right)$$

Properties of the MLE, Cont'd

- Also,

$$\frac{\hat{\theta}_n^{MLE} - \theta}{\sqrt{\frac{1}{n \cdot \mathcal{I}\left(\hat{\theta}_n^{MLE}\right)}}} \xrightarrow{D} \mathcal{N}(0, 1)$$

or, in more non-rigorous terms,

$$\hat{\theta}_n^{MLE} \overset{D}{\approx} \mathcal{N}\left(\theta, \frac{1}{n \cdot \mathcal{I}\left(\hat{\theta}_n^{MLE}\right)}\right)$$

Properties of the MLE, Cont'd

- Also,

$$\frac{\hat{\theta}_n^{MLE} - \theta}{\sqrt{\frac{1}{n \cdot \mathcal{I}\left(\hat{\theta}_n^{MLE}\right)}}} \xrightarrow{D} \mathcal{N}(0, 1)$$

  or, in more non-rigorous terms,

$$\hat{\theta}_n^{MLE} \overset{D}{\approx} \mathcal{N}\left(\theta, \frac{1}{n \cdot \mathcal{I}\left(\hat{\theta}_n^{MLE}\right)}\right)$$

Note: This is almost the above Property,

$$\hat{\theta}_n^{MLE} \overset{D}{\approx} \mathcal{N}\left(\theta, \frac{1}{n \cdot \mathcal{I}(\theta)}\right),$$

but, instead of $\mathcal{I}(\theta)$ we have $\mathcal{I}\left(\hat{\theta}_n^{MLE}\right)$.

Properties of the MLE, Cont'd

- Also,

$$\frac{\hat{\theta}_n^{MLE} - \theta}{\sqrt{\frac{1}{n \cdot \mathcal{I}\left(\hat{\theta}_n^{MLE}\right)}}} \xrightarrow{D} \mathcal{N}(0, 1)$$

  or, in more non-rigorous terms,

$$\hat{\theta}_n^{MLE} \overset{D}{\approx} \mathcal{N}\left(\theta, \frac{1}{n \cdot \mathcal{I}\left(\hat{\theta}_n^{MLE}\right)}\right)$$

Note: This is almost the above Property,

$$\hat{\theta}_n^{MLE} \overset{D}{\approx} \mathcal{N}\left(\theta, \frac{1}{n \cdot \mathcal{I}(\theta)}\right),$$

but, instead of $\mathcal{I}(\theta)$ we have $\mathcal{I}\left(\hat{\theta}_n^{MLE}\right)$.

Note: We will use this later, to construct an (approximate) Confidence Interval for θ and for testing Hypotheses about θ.

Properties of the MLE, Cont'd

- If $\hat{\theta}$ is the MLE for θ, then for any function g, the MLE of $g(\theta)$ is $g(\hat{\theta})$, i.e.,
$$\widehat{g(\theta)}^{MLE} = g\left(\hat{\theta}^{MLE}\right).$$

Properties of the MLE, Cont'd

- If $\hat{\theta}$ is the MLE for θ, then for any function g, the MLE of $g(\theta)$ is $g(\hat{\theta})$, i.e.,

$$\widehat{g(\theta)}^{MLE} = g\left(\hat{\theta}^{MLE}\right).$$

This is sometimes called the **invariance Principle of MLE**.

Properties of the MLE, Cont'd

- If $\hat{\theta}$ is the MLE for θ, then for any function g, the MLE of $g(\theta)$ is $g(\hat{\theta})$, i.e.,
$$\widehat{g(\theta)}^{MLE} = g\left(\hat{\theta}^{MLE}\right).$$

This is sometimes called the **invariance Principle of MLE**.

Example Find the MLE for σ in $\mathcal{N}(\mu, \sigma^2)$ Model.

Solution: OTB

Other Methods to construct Point Estimators/Estimates

There are other important methods to construct Estimators: e.g.

- Bayesian Estimation: Maximum APosteriori (MAP) Estimators
- Bayesian Estimation: Bayes Estimators;
- OLS
- etc

Confidence Intervals

Prelude

So far we have studied the Parameteric Point Estimation Problem: we were assuming that our data comes from a Parametric Family of Distribution, and our aim was to find a best guess for that Parameter.

Prelude

So far we have studied the Parameteric Point Estimation Problem: we were assuming that our data comes from a Parametric Family of Distribution, and our aim was to find a best guess for that Parameter.

Assume θ is our Parameter to be Estimated, and $\hat{\theta}$ is a good Estimator for θ.

Prelude

So far we have studied the Parameteric Point Estimation Problem: we were assuming that our data comes from a Parametric Family of Distribution, and our aim was to find a best guess for that Parameter.

Assume θ is our Parameter to be Estimated, and $\hat{\theta}$ is a good Estimator for θ. Assume $\hat{\theta}$ is a Continuous r.v.

Prelude

So far we have studied the Parameteric Point Estimation Problem: we were assuming that our data comes from a Parametric Family of Distribution, and our aim was to find a best guess for that Parameter.

Assume θ is our Parameter to be Estimated, and $\hat{\theta}$ is a good Estimator for θ. Assume $\hat{\theta}$ is a Continuous r.v. If the True value of our Parameter is θ^*, then

$$\mathbb{P}(\hat{\theta} = \theta^*) = 0,$$

i.e., we will (almost) **never** be correct in our guess.

Prelude

So far we have studied the Parameteric Point Estimation Problem: we were assuming that our data comes from a Parametric Family of Distribution, and our aim was to find a best guess for that Parameter.

Assume θ is our Parameter to be Estimated, and $\hat{\theta}$ is a good Estimator for θ. Assume $\hat{\theta}$ is a Continuous r.v. If the True value of our Parameter is θ^*, then

$$\mathbb{P}(\hat{\theta} = \theta^*) = 0,$$

i.e., we will (almost) **never** be correct in our guess. Sad news!

Prelude No. 2

But the good news is that even when we cannot exactly find the True value of our Parameter using $\hat{\theta}$, if $\hat{\theta}$ possesses some good properties, we believe that the Estimate obtained is a good approximation/Estimate for θ^*.

[1] Recall the $\widehat{SE}$, the Estimated Standard Error reporting story.

Prelude No. 2

But the good news is that even when we cannot exactly find the True value of our Parameter using $\hat{\theta}$, if $\hat{\theta}$ possesses some good properties, we believe that the Estimate obtained is a good approximation/Estimate for θ^*.

In fact, we think as

$$\theta^* = \hat{\theta} \pm \mathit{error},$$

where *error* is small[1],

[1] Recall the $\widehat{SE}$, the Estimated Standard Error reporting story.

Prelude No. 2

But the good news is that even when we cannot exactly find the True value of our Parameter using $\hat{\theta}$, if $\hat{\theta}$ possesses some good properties, we believe that the Estimate obtained is a good approximation/Estimate for θ^*.

In fact, we think as

$$\theta^* = \hat{\theta} \pm \mathit{error},$$

where *error* is small[1], or, rather, as

$$\theta^* \in (\hat{\theta} - \mathit{error}, \hat{\theta} + \mathit{error}).$$

[1] Recall the $\widehat{SE}$, the Estimated Standard Error reporting story.

Prelude No. 2

But the good news is that even when we cannot exactly find the True value of our Parameter using $\hat{\theta}$, if $\hat{\theta}$ possesses some good properties, we believe that the Estimate obtained is a good approximation/Estimate for θ^*.

In fact, we think as

$$\theta^* = \hat{\theta} \pm \mathit{error},$$

where *error* is small[1], or, rather, as

$$\theta^* \in (\hat{\theta} - \mathit{error}, \hat{\theta} + \mathit{error}).$$

And you were asking nice questions about how small is the *error* or how much sure are we in our Estimate (for the Unknown Parameter), and how large n needs to be to have a good estimate.

[1] Recall the $\widehat{SE}$, the Estimated Standard Error reporting story.

Prelude No. 2

But the good news is that even when we cannot exactly find the True value of our Parameter using $\hat{\theta}$, if $\hat{\theta}$ possesses some good properties, we believe that the Estimate obtained is a good approximation/Estimate for θ^*.

In fact, we think as

$$\theta^* = \hat{\theta} \pm \mathit{error},$$

where *error* is small[1], or, rather, as

$$\theta^* \in (\hat{\theta} - \mathit{error}, \hat{\theta} + \mathit{error}).$$

And you were asking nice questions about how small is the *error* or how much sure are we in our Estimate (for the Unknown Parameter), and how large n needs to be to have a good estimate.

Here we want to develop the theory of Confidence Intervals, which will contain answers to these questions.

---

[1] Recall the $\widehat{SE}$, the Estimated Standard Error reporting story.

CI Problem Setting

Assume we have a Random Sample from a Parametric Model $\mathcal{F}_\theta$:

$$X_1, X_2, ..., X_n \sim \mathcal{F}_\theta, \qquad \theta \in \Theta.$$

CI Problem Setting

Assume we have a Random Sample from a Parametric Model $\mathcal{F}_\theta$:

$$X_1, X_2, ..., X_n \sim \mathcal{F}_\theta, \qquad \theta \in \Theta.$$

Our aim is to find an interval containing the Unknown Parameter value.

CI Problem Setting

Assume we have a Random Sample from a Parametric Model $\mathcal{F}_\theta$:

$$X_1, X_2, ..., X_n \sim \mathcal{F}_\theta, \qquad \theta \in \Theta.$$

Our aim is to find an interval containing the Unknown Parameter value.

More rigorously, we want to have an interval

- which contains the True Parameter value with high Probability;

CI Problem Setting

Assume we have a Random Sample from a Parametric Model $\mathcal{F}_\theta$:

$$X_1, X_2, ..., X_n \sim \mathcal{F}_\theta, \qquad \theta \in \Theta.$$

Our aim is to find an interval containing the Unknown Parameter value.

More rigorously, we want to have an interval

- which contains the True Parameter value with high Probability;
- which has the possible smallest length.

CI Problem Setting

Assume we have a Random Sample from a Parametric Model $\mathcal{F}_\theta$:

$$X_1, X_2, ..., X_n \sim \mathcal{F}_\theta, \qquad \theta \in \Theta.$$

Our aim is to find an interval containing the Unknown Parameter value.

More rigorously, we want to have an interval

- which contains the True Parameter value with high Probability;
- which has the possible smallest length.

Let us state this in Mathematical terms. We will consider here only 1D case, i.e., we will assume $\theta \in \Theta \subset \mathbb{R}$.

Random Intervals and CI

We start by **Random Intervals:**

Random Intervals and CI

We start by **Random Intervals:**

Assume X and Y are two r.v. on the same probability space and $X \leq Y$ a.s., that is, $\mathbb{P}(X > Y) = 0$.

Random Intervals and CI

We start by **Random Intervals:**

Assume X and Y are two r.v. on the same probability space and $X \leq Y$ a.s., that is, $\mathbb{P}(X > Y) = 0$. We will call intervals of the form $[X, Y]$ or (X, Y) or $(X, Y]$ or $[X, Y)$ or $(X, +\infty)$,... , **random intervals**.

Random Intervals and CI

We start by **Random Intervals:**

Assume X and Y are two r.v. on the same probability space and $X \leq Y$ a.s., that is, $\mathbb{P}(X > Y) = 0$. We will call intervals of the form $[X, Y]$ or (X, Y) or $(X, Y]$ or $[X, Y)$ or $(X, +\infty)$,... , **random intervals**.

Example: Assume $X \sim Pois(2.3)$. Then

$$[X, X + 2)$$

is a Random Interval.

Random Intervals and CI

We start by **Random Intervals:**

Assume X and Y are two r.v. on the same probability space and $X \leq Y$ a.s., that is, $\mathbb{P}(X > Y) = 0$. We will call intervals of the form $[X, Y]$ or (X, Y) or $(X, Y]$ or $[X, Y)$ or $(X, +\infty)$,... , **random intervals**.

Example: Assume $X \sim Pois(2.3)$. Then

$$[X, X + 2)$$

is a Random Interval.

Example: Let $X_1, X_2, ..., X_n$ are IID r.v.s. Then

$$(\overline{X} - 0.1,\ \overline{X} + 0.1)$$

is a Random Interval.