COMP 680
Statistics for Computing and Data Science
Week 4: Maximun Likelihood Estimate

Su Chen, Assistant Teaching Professor,
Rice D2K Lab

Outline

A Motivating Example

- One-dimensional parametric estimation:
  - let $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim}$ Bernoulli(p) observations. The problem is to estimate the parameter p.
  - if data is 6 Heads out of 10 flips, what would be the "best guess" of p, the probability of Head?

A Motivating Example

- One-dimensional parametric estimation:
  - let $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim}$ Bernoulli(p) observations. The problem is to estimate the parameter p.
  - if data is 6 Heads out of 10 flips, what would be the "best guess" of p, the probability of Head?
- Intuition of Maximum Likelihood Estimate(MLE)
  - what would be the value of p that most likely to generate the data I have observed?

The Likelihood Function

- The likelihood function is a function of the unknown parameter θ, but also depends on the observed data $X_1, X_2, \cdots X_n$.
  - notation: $L(\theta|X_1 = x_1, X_2 = x_2, \cdots X_n = x_n) = L_n(\theta|X)$
  - intuition: the "probability" of observing the data as a function of θ.

The Likelihood Function

- The likelihood function is a function of the unknown parameter θ, but also depends on the observed data $X_1, X_2, \cdots X_n$.
  - notation: $L(\theta|X_1 = x_1, X_2 = x_2, \cdots X_n = x_n) = L_n(\theta|X)$
  - intuition: the "probability" of observing the data as a function of θ.
- $L_n(p|X) =?$
  - $L_n(p|X)$ is a function of p defined in $p \in [0, 1]$
  - how does this function look like?
  - what does it tell us about the "best guess" of p?

MLE

- assume $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(x;\theta)$

$$L_n(\theta|X) = \begin{cases} \prod_{i=1}^n f_X(x_i;\theta) & \text{if } X \text{ is continuous} \\ \prod_{i=1}^n \mathbb{P}(X = x_i;\theta) & \text{if } X \text{ is discrete} \end{cases}$$

MLE

- assume $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(x;\theta)$

$$L_n(\theta|X) = \begin{cases} \prod_{i=1}^n f_X(x_i;\theta) & \text{if } X \text{ is continuous} \\ \prod_{i=1}^n \mathbb{P}(X = x_i;\theta) & \text{if } X \text{ is discrete} \end{cases}$$

- $\hat{\theta}^{MLE} = \arg\max_{\theta \in \Theta} L_n(\theta|X)$
  - we need to know the explicit format of the likelihood function in order to maximize it!

MLE

- assume $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(x;\theta)$

$$L_n(\theta|X) = \begin{cases} \prod_{i=1}^n f_X(x_i;\theta) & \text{if } X \text{ is continuous} \\ \prod_{i=1}^n \mathbb{P}(X = x_i;\theta) & \text{if } X \text{ is discrete} \end{cases}$$

- $\hat{\theta}^{MLE} = \arg\max_{\theta \in \mathbf{\Theta}} L_n(\theta|X)$
  - we need to know the explicit format of the likelihood function in order to maximize it!
- Notice both parameter θ and data X can be multi-dimensional
  - for illustration, we will keep it simple

Examples

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathsf{N}(x; \mu, \sigma = 1)$
  - $L(\mu|\cdot) = ?$

Examples

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathsf{N}(x; \mu, \sigma = 1)$
  - $L(\mu|\cdot) = ?$
  - what if σ is also unknown? $L(\mu, \sigma|\cdot) = ?$

Examples

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(x; \mu, \sigma = 1)$
  - $L(\mu|\cdot) = ?$
  - what if σ is also unknown? $L(\mu, \sigma|\cdot) = ?$
- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Poission}(x; \theta)$
  - $L(\theta|\cdot) = ?$

Examples

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(x; \mu, \sigma = 1)$
  - $L(\mu|\cdot) = ?$
  - what if σ is also unknown? $L(\mu, \sigma|\cdot) = ?$
- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Poission}(x; \theta)$
  - $L(\theta|\cdot) = ?$
- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}[0, \theta]$
  - $L(\theta|\cdot) = ?$

MLE in Practice

- Once you write down the likelihood function:
  - $\hat{\theta}^{MLE} = \arg\max_{\theta \in \mathbf{\Theta}} L_n(\theta|X)$
  - to get $\hat{\theta}^{MLE}$ is to solve an optimization problem

MLE in Practice

- Once you write down the likelihood function:
  - $\hat{\theta}^{MLE} = \arg\max_{\theta \in \mathbf{\Theta}} L_n(\theta|X)$
  - to get $\hat{\theta}^{MLE}$ is to solve an optimization problem
- Some tricks
  - ignore any "constant" that does not depend on θ
  - take the log transformation before maximize

MLE in Practice

- Once you write down the likelihood function:
  - $\hat{\theta}^{MLE} = \arg\max_{\theta \in \mathbf{\Theta}} L_n(\theta|X)$
  - to get $\hat{\theta}^{MLE}$ is to solve an optimization problem
- Some tricks
  - ignore any "constant" that does not depend on θ
  - take the log transformation before maximize
- Need numerical (convex) optimization
  - gradient decent
  - Newton's method

Outline

Point Estimate

- Reminder: a point estimate $\hat{\theta}$ is a function of data $X_1, X_2, \cdots X_n$ therefore by definition it is a statistic with a sampling distribution.

Point Estimate

- Reminder: a point estimate $\hat{\theta}$ is a function of data $X_1, X_2, \cdots X_n$ therefore by definition it is a statistic with a sampling distribution.
- A point estimate $\hat{\theta}$ is:

Point Estimate

- Reminder: a point estimate $\hat{\theta}$ is a function of data $X_1, X_2, \cdots X_n$ therefore by definition it is a statistic with a sampling distribution.
- A point estimate $\hat{\theta}$ is:
  - unbiased if $\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta = 0$

Point Estimate

- Reminder: a point estimate $\hat{\theta}$ is a function of data $X_1, X_2, \cdots X_n$ therefore by definition it is a statistic with a sampling distribution.
- A point estimate $\hat{\theta}$ is:
  - unbiased if $\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta = 0$
  - consistent if $\hat{\theta} \xrightarrow{\mathbb{P}} \theta$ as $n \to \infty$

Point Estimate

- Reminder: a point estimate $\hat{\theta}$ is a function of data $X_1, X_2, \cdots X_n$ therefore by definition it is a statistic with a sampling distribution.
- A point estimate $\hat{\theta}$ is:
  - unbiased if $\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta = 0$
  - consistent if $\hat{\theta} \xrightarrow{\mathbb{P}} \theta$ as $n \to \infty$
- Intuition and examples:
  - unbiased but not consistent:
  - consistent but not unbiased:

The Mean Squared Error

- For a point estimate $\hat{\theta}$, what can we say about its accuracy?

The Mean Squared Error

- For a point estimate $\hat{\theta}$, what can we say about its accuracy?
- Accuracy does not depend on bias **alone**, but on both:
  - $\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$
  - variance and standard error: $\text{se}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$

The Mean Squared Error

- For a point estimate $\hat{\theta}$, what can we say about its accuracy?
- Accuracy does not depend on bias **alone**, but on both:
  - $\text{bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$
  - variance and standard error: $\text{se}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$
- Thus the quality of a point estimate is often assessed by the Mean Squared Error (MSE):

$$\text{MSE} = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

Bias and Variance Decomposition

- Not surprisingly, the MSE can be written as

$$\text{MSE} = \text{bias}^2(\hat{\theta}) + \text{Var}(\hat{\theta})$$

Bias and Variance Decomposition

- Not surprisingly, the MSE can be written as

$$\text{MSE} = \text{bias}^2(\hat{\theta}) + \text{Var}(\hat{\theta})$$

- Proof:

Bias and Variance Decomposition

- Not surprisingly, the MSE can be written as

$$\text{MSE} = \text{bias}^2(\hat{\theta}) + \text{Var}(\hat{\theta})$$

- Proof:
- A criteria to compare two point estimates:
  - In general, there is the bias and variance trade-off

Asymptotic Normality

- A point estimate is Asymptotically Normal if

$$\frac{\hat{\theta} - \theta}{\mathsf{se}(\hat{\theta})} \xrightarrow{\mathbb{D}} \mathsf{N}(0, 1)$$

Asymptotic Normality

- A point estimate is Asymptotically Normal if

$$\frac{\hat{\theta} - \theta}{\text{se}(\hat{\theta})} \xrightarrow{\mathbb{D}} \text{N}(0, 1)$$

- Connection to Central Limit Theorem:
  - many $\hat{\theta}$ involves sample mean

Asymptotic Normality

- A point estimate is Asymptotically Normal if

$$\frac{\hat{\theta} - \theta}{\text{se}(\hat{\theta})} \xrightarrow{\mathbb{D}} \text{N}(0, 1)$$

- Connection to Central Limit Theorem:
  - many $\hat{\theta}$ involves sample mean
- Asymptotic Normality of MLE
  - in general true, under some mild regularity conditions

Outline

Definition

- A $1-\alpha$ confidence interval (CI) for a parameter θ is an interval $C = [a, b]$ where both a and b are functions of the data, such that:

$$\mathbb{P}(\theta \in C) \geq 1 - \alpha$$

Definition

- A $1-\alpha$ confidence interval (CI) for a parameter θ is an interval $C = [a, b]$ where both a and b are functions of the data, such that:

$$\mathbb{P}(\theta \in C) \geq 1 - \alpha$$

- Please notice in the above statement:
  - the interval C is random
  - θ is fixed (unknown parameter)

Definition

- A $1-\alpha$ confidence interval (CI) for a parameter θ is an interval $C = [a, b]$ where both a and b are functions of the data, such that:

$$\mathbb{P}(\theta \in C) \geq 1 - \alpha$$

- Please notice in the above statement:
  - the interval C is random
  - θ is fixed (unknown parameter)
- $\alpha = 0.05(0.1) \rightarrow$ 95% (90%) Confidence Interval

Interpretation

- Interpretation is the key!!!

Interpretation

- Interpretation is the key!!!
- It is not a probability statement about θ!
  - why?

Interpretation

- Interpretation is the key!!!
- It is not a probability statement about θ!
  - why?
- Textbook interpretation:
  - if repeatedly draw data from they population many times and construct a CI each time
  - about $1 - \alpha$ of the times, the CI will contain true θ value

Interpretation

- Interpretation is the key!!!
- It is not a probability statement about θ!
  - why?
- Textbook interpretation:
  - if repeatedly draw data from they population many times and construct a CI each time
  - about $1 - \alpha$ of the times, the CI will contain true θ value
- Real life interpretation???
  - interval estimate of parameter θ
  - what is wrong with point estimate?
  - quantify uncertainty

Recall CLT

- Central Limit Theorem: "asymptotic normality of sample mean "

Recall CLT

- Central Limit Theorem: "asymptotic normality of sample mean "
- **IF** $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(\cdot)$ with mean μ and variance σ^2, then:

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathbb{D}} N(0,1)$$

Recall CLT

- Central Limit Theorem: "asymptotic normality of sample mean "
- **IF** $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} f_X(\cdot)$ with mean μ and variance σ^2, then:

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathbb{D}} N(0,1)$$

- Construct CI for sample mean:
  - relies on "knowing" the sampling distribution

CI for Sample Mean

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$

CI for Sample Mean

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$
  - **IF** σ is known, then we have Normal CI for μ :

$$\left[\bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right] \approx \left[\bar{X}_n - 2\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 2\frac{\sigma}{\sqrt{n}}\right]$$

CI for Sample Mean

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$
  - **IF** σ is known, then we have Normal CI for μ :

$$\left[\bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right] \approx \left[\bar{X}_n - 2\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 2\frac{\sigma}{\sqrt{n}}\right]$$

  - **IF** σ is unknown and s^2 is sample variance, then we have t-CI for μ :

$$\left[\bar{X}_n + \frac{s}{\sqrt{n}} t_{\alpha/2}, \bar{X}_n + \frac{s}{\sqrt{n}} t_{1-\alpha/2}\right]$$

CI for Sample Mean

- $X_1, X_2, \cdots X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$
  - **IF** σ is known, then we have Normal CI for μ :

$$\left[\bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right] \approx \left[\bar{X}_n - 2\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 2\frac{\sigma}{\sqrt{n}}\right]$$

  - **IF** σ is unknown and s^2 is sample variance, then we have t-CI for μ :

$$\left[\bar{X}_n + \frac{s}{\sqrt{n}} t_{\alpha/2}, \bar{X}_n + \frac{s}{\sqrt{n}} t_{1-\alpha/2}\right]$$

- If population distribution is not normal:
  - asymptotic Normal CI for population mean.

The Delta Method

- If $\bar{X}_n$ has a limiting Normal distribution
- If g is differentiable and $g'(\mu) \neq 0$

The Delta Method

- If $\bar{X}_n$ has a limiting Normal distribution
- If g is differentiable and $g'(\mu) \neq 0$

$$\frac{\sqrt{n}(g(\bar{X}_n) - g(\mu))}{|g'(\mu)|\sigma} \xrightarrow{\mathbb{D}} N(0,1)$$

The Delta Method

- If $\bar{X}_n$ has a limiting Normal distribution
- If g is differentiable and $g'(\mu) \neq 0$

$$\frac{\sqrt{n}(g(\bar{X}_n) - g(\mu))}{|g'(\mu)|\sigma} \xrightarrow{\mathbb{D}} N(0,1)$$

- In other words,

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right) \text{ implies } g(\bar{X}_n) \approx N\left(g(\mu), g'(\mu)^2 \frac{\sigma^2}{n}\right)$$

The Delta Method

- If $\bar{X}_n$ has a limiting Normal distribution
- If g is differentiable and $g'(\mu) \neq 0$

$$\frac{\sqrt{n}(g(\bar{X}_n) - g(\mu))}{|g'(\mu)|\sigma} \xrightarrow{\mathbb{D}} N(0,1)$$

- In other words,

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right) \text{ implies } g(\bar{X}_n) \approx N\left(g(\mu), g'(\mu)^2 \frac{\sigma^2}{n}\right)$$

- Intuition: propagation of uncertainty