TDAB01 Probability and Statistics

Maryna Prus
IDA, Linköping University

Lecture 8: Maximum Likelihood Estimator, Confidence Intervals

Overview

- **Maximum likelihood method**
- **Sampling distribution**
- **Confidence intervals**

Point estimator (or estimator), from Lecture 7

- Observing data we may guess suitable family of distributions
- Problem: **unknown** parameters
- Examples:
  - Average income in Sweden $\rightarrow$ Population expectation μ - ?
  - Proportion of defect products $\rightarrow$ Probability p - ?
- Use data to determine values for unknown parameters
- **Point estimate** - **best guess** about parameter based on data
- Data from different points of view
  - Practically: observations, **values**, sample - $x_1, \dots, x_n$
  - Analytically: **random variables**, iid - $X_1, \dots, X_n$
- (Point) **Estimate** - one **value**; mean - $\bar{x}$
- (Point) **Estimator** - function, **random variable**; mean - $\bar{X}$

Maximum likelihood method

- $X_1, \ldots, X_n$ - i.i.d. random variables
- Distribution of $X_1, \ldots, X_n$ depends on *unknown* parameter $\theta \in \Theta$
- $x_1, \ldots, x_n$ - observed data
- "Good" estimation of θ - ?
- Idea:

  "Good" estimation of θ -

  value of θ that *maximizes likelihood of observed data*

ML estimation, Discrete case

- $X_1, \ldots, X_n$ - *discrete* random variables
- $P_{X_i}(x_i)$ - *probability function* (pmf) of X_i, depends on parameter $\theta \in \Theta$
- **Maximum Likelihood estimation** $\hat{\theta}$ of θ maximizes

  joint pmf of $X_1, \ldots, X_n$:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} P_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$$

- **Likelihood function**:

$$\begin{aligned} L(\theta) &= P_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \\ &= \prod_{i=1}^{n} P_{X_i}(x_i) \end{aligned}$$

ML estimation, Continuous case

- $X_1, \ldots, X_n$ - *continuous* random variables
- $f_{X_i}(x_i)$ - *density function* (pdf) of X_i, , depends on parameter $\theta \in \Theta$
- **Maximum Likelihood estimation** $\hat{\theta}$ of θ maximizes

  joint pdf of $X_1, \ldots, X_n$:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$$

- **Likelihood function**:

$$\begin{aligned} L(\theta) &= f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \\ &= \prod_{i=1}^{n} f_{X_i}(x_i) \end{aligned}$$

Interpretation

- $X_1, \ldots, X_n$ - *discrete* random variables

  ML estimation $\hat{\theta}$ of θ maximizes

$$P_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = P(X_1 = x_1,\ldots,X_n = x_n)$$

- $X_1, \ldots, X_n$ - *continuous* random variables

$$P(x_i - h < X_i < x_i + h) = \int_{x_i-h}^{x_i+h} f_{X_i}(t)dt$$

$$\int_{x_i-h}^{x_i+h} f_{X_i}(t)dt \approx 2\,h\,f_{X_i}(x_i), \quad h > 0, \text{ small}$$

  ML estimation $\hat{\theta}$ maximizes probability for $X_1, \ldots, X_n$

  to take values "very close to" $x_1, \ldots, x_n$

Maximizing Likelihood

1. $\theta \in \mathbb{R}$, $L(\theta)$ twice differentiable

    - Solve equation
    $$\frac{\partial L(\theta)}{\partial \theta} = 0$$
    - Solution $\hat{\theta}$ is local maximum if
    $$\frac{\partial^2 L(\theta)}{\partial \theta^2} < 0, \text{ at } \theta = \hat{\theta}$$
    - Check for local maximum $\hat{\theta}$ if it is global maximum
    - Usually it is easier to maximize **Log-Likelihood function**
    $$\ell(\theta) = \ln L(\theta)$$
    Same result as $ln(x)$ is *strictly increasing* function

Example: Poisson Distribution

- $X_i \sim Po(\lambda),\ i = 1, \ldots, n$

Pmf of X_i:

$$P_{X_i}(x) = \frac{\exp(-\lambda) \cdot \lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

Likelihood and Log-Likelihood functions:

$$L(\lambda) = \frac{\exp(-n\lambda) \cdot \lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \quad \& \quad \ell(\lambda) = -n\lambda + \sum_{i=1}^n x_i \ln\lambda - \ln \prod_{i=1}^n x_i!$$

Derivatives of Log-Likelihood function:

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = -n + \frac{\sum_{i=1}^n x_i}{\lambda} \quad \& \quad \frac{\partial^2 \ell(\lambda)}{\partial \lambda^2} = -\frac{\sum_{i=1}^n x_i}{\lambda^2} < 0$$

ML estimation of λ: $\hat{\lambda} = \bar{x}$

Sampling distribution

- Estimator $\hat{\theta}$ - function of $X_1, \ldots, X_n \quad \rightarrow \quad \hat{\theta}$ - **random variable**
- Distribution of $\hat{\theta}$ - **sampling distribution**
- Sampling distribution describes variation of $\hat{\theta}$ **over all samples** of size n
- **Bias** of $\hat{\theta}$: $Bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$
- **Standard error** of $\hat{\theta}$: $\sqrt{Var(\hat{\theta})}$
- **Mean Squared Error**:

$$MSE(\hat{\theta}) = \mathbb{E}[\left(\hat{\theta} - \theta\right)^2] = \text{Var}(\hat{\theta}) + \left[Bias(\hat{\theta})\right]^2$$

- $\hat{\theta}$ good estimator for θ if
  - $\hat{\theta}$ has correct expected value (unbiased): $\mathbb{E}(\hat{\theta}) = \theta$, or small bias
  - $\hat{\theta}$ has small standard error / small variance
  - $\hat{\theta}$ has small MSE

Sampling distribution

- Poisson data: $X_1, \ldots, X_n$ iid. with $X_i \sim Po(\lambda)$ (Example 9.7 in textbook)
- ML estimator for λ: $\bar{X}$
- $\hat{\lambda}$ unbiased: $\mathbb{E}(\hat{\lambda}) = \lambda$
- $Var(\hat{\lambda}) = \frac{\sigma^2}{n} = \frac{\lambda}{n}$, σ^2 - variance of X_i
- $Var(\hat{\lambda})$ depends on unknown parameter λ
- Solution: replace λ by $\bar{x}$ or σ^2 by s^2, $s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$
- Techniques for deriving sampling distribution of estimator $\hat{\theta}$:
  - $X_1, \ldots, X_n$ iid from $N(\mu, \sigma^2)$

    $$\rightarrow \quad \hat{\theta} = \bar{X} \sim N(\mu, \sigma^2/n) \ \textbf{exactly}$$
  - $X_1, \ldots, X_n$ iid with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, n large

    $$\rightarrow \quad \hat{\theta} = \bar{X} \sim N(\mu, \sigma^2/n) \ \textbf{approximately}$$
  - **Bootstrap method**

Bootstrap method:

- Create N **bootstrap samples**

$$\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$$

  of the same size as the original sample by sampling **with replacement**

- Calculate estimates

$$\hat{\theta}(\mathbf{x}^{(1)}), \ldots, \hat{\theta}(\mathbf{x}^{(N)})$$

  for each of these N samples

- Empirical distribution of

$$\hat{\theta}(\mathbf{x}^{(1)}), \ldots, \hat{\theta}(\mathbf{x}^{(N)})$$

  - histogram - is approximation of sampling distribution for $\hat{\theta}$

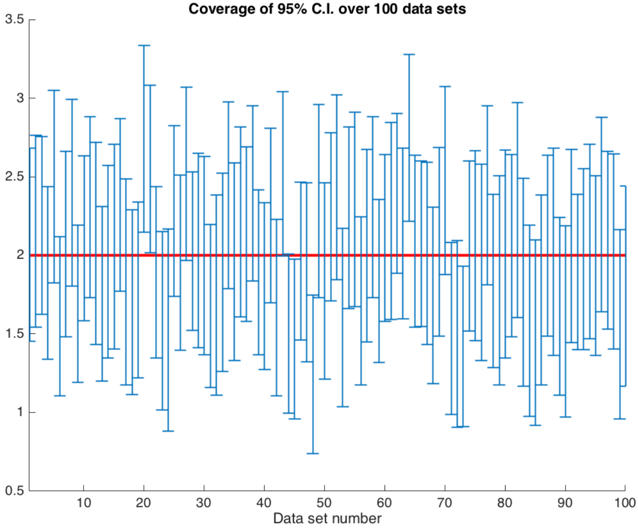
Confidence interval

- Point estimate - best guess for θ
  Confidence interval - describes uncertainty of θ
- **95% confidence interval** for θ - interval $[a, b]$ such that

$$\boldsymbol{P}\{a \leq \theta \leq b\} = 0.95$$

- **Important**: Parameter θ is fixed constant
  Interval is **random**, i.e. a and b - functions of sample
- **Interpretation**: 95% confidence interval $[a, b]$ **covers** parameter value θ, i.e. $\theta \in [a, b]$ in 95% of all possible samples
  If we count a and b from all samples, $[a, b]$ covers θ in 95% of cases
- 95% - **confidence level**
  Other commonly used confidence levels: 90% and 99%

Confidence interval

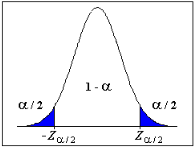


Confidence interval - general approach

- $\hat{\theta}$ - **normally distributed unbiased estimator** for θ
- Standardization

$$Z = \frac{\hat{\theta} - \theta}{\sigma(\hat{\theta})} \sim N(0,1)$$

- z_α - $(1-\alpha)$ quantile of $N(0,1)$ distribution



- Then

$$\boldsymbol{P}\Big(-z_{\alpha/2} \le \frac{\hat{\theta}-\theta}{\sigma(\hat{\theta})} \le z_{\alpha/2}\Big) = 1-\alpha$$

$$\Rightarrow \quad \boldsymbol{P}(\hat{\theta} - z_{\alpha/2}\cdot\sigma(\hat{\theta}) \le \theta \le \hat{\theta} + z_{\alpha/2}\cdot\sigma(\hat{\theta})) = 1-\alpha$$

- $(1-\alpha)\cdot 100\%$ confidence interval for θ:

$$[\hat{\theta} - z_{\alpha/2}\cdot\sigma(\hat{\theta}), \hat{\theta} + z_{\alpha/2}\cdot\sigma(\hat{\theta})]$$

- Example: $\alpha = 0.05$
- $z_{\alpha/2} = z_{0.025} = 1.96$ from Table A4 in textbook
- $[\hat{\theta} - 1.96\cdot\sigma(\hat{\theta}), \hat{\theta} + 1.96\cdot\sigma(\hat{\theta})]$ is 95% confidence interval for θ

Confidence interval for the population mean

- $X_1, \dots, X_n$ iid with $\mathbb{E}(X_i) = \mu$ and $Var(X_i) = \sigma^2$
- $\theta = \mu$ - **unknown**, σ - **known**
- $\hat{\theta} = \bar{X}$ - estimator for θ with

$$\mathbb{E}(\hat{\theta}) = \mathbb{E}(\bar{X}) = \mu \text{ and } \sigma(\hat{\theta}) = Std(\bar{X}) = \sigma/\sqrt{n}$$

- $X_1, \dots, X_n$ **normally distributed**

  $\rightarrow \quad [\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}]$ - exact $(1-\alpha) \cdot 100\%$ confidence interval for μ

- $X_1, \dots, X_n$ **not** normally distributed (any other distribution), n **large**

  $\rightarrow \quad \hat{\theta} = \bar{X}$ approximately normally distributed according to CLT

  $\rightarrow \quad [\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}]$ - **approximate** $(1-\alpha) \cdot 100\%$ confidence interval for μ

- Length of confidence interval: $2 \cdot z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$, decreasing with increasing n
- **Selection of sample size** n: Choose n to get given length
- Examples: Examples 9.13 and 9.15 in textbook

Confidence interval for population mean

- $X_1, \ldots, X_n$ iid, σ^2 **unknown**
- For **large** n replace σ by its estimator $s \quad \rightarrow \quad s(\hat{\theta}) = s/\sqrt{n}$

  $\rightarrow \quad [\bar{X} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}]$ - **approximate** $(1-\alpha) \cdot 100\%$ confidence interval for μ
- $X_1, \ldots, X_n$ **normally distributed**:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t\,(t-1)$$

  $\rightarrow$ **Exact** $(1-\alpha) \cdot 100\%$ confidence interval for μ:

$$[\bar{X} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}]$$

  $t_{\alpha/2}(n-1)$ - $(1-\alpha/2)$ quantile of t-**distribution** with $\nu = n-1$ degrees of freedom (Table A5 in textbook)
- Example: Example 9.19 in textbook
- **Small sample & non-normally distributed data** $\rightarrow$ bootstrap method

Confidence interval for proportion

- Some items from population have certain attribute

  Example: defect products
- p - probability for randomly selected item to have this attribute
- $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ where $X_i = 1$ if i-th sampled item has attribute and $X_i = 0$ otherwise
- $\hat{p}$ - estimator for p
- $\hat{p}$ is also mean $\quad\rightarrow\quad$ same approach as for population mean
- Then $X_1, \ldots, X_n \overset{iid}{\sim} Bernoulli(p)$ and $\mathbb{E}(X_i) = p$, $Var(X_i) = p(1-p)$
- In other words

$$\mathbb{E}(\hat{p}) = p \text{ and } Var(\hat{p}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

- $\sigma(\hat{p})$ depends on $p \quad\rightarrow\quad$ use $s(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$
- From CLT, for large n:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

  - approximate $(1-\alpha)100\%$ confidence interval for p

Thank you for your attention!