Lecture 9: Classical Statistical Inference

Ziyu Shao

School of Information Science and Technology
ShanghaiTech University

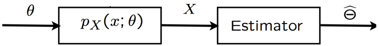
December 21, 2021

Outline

1. Inference Rule: Maximum Likelihood Estimation
2. Normal Distribution: New Perspective
3. Central Limit Theorem
4. Confidence Interval

Classical vs. Bayesian

- Inference using the Bayes rule:
  unknown Θ and observation X are both random variables
  - Find $p_{\Theta|X}$
- Classical statistics: unknown constant θ



  - also for vectors X and θ: $p_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_m)$
  - $p_X(x;\theta)$ are NOT conditional probabilities; θ is NOT random
  - mathematically: many models, one for each possible value of θ

Outline

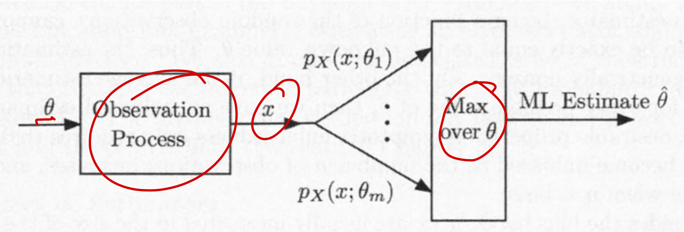
Maximum Likelihood Estimation (MLE)

- Joint distribution of the vector of observations $X = (X_1, \ldots, X_n)$: PMF $P_X(x;\theta)$ (or PDF $f_X(x;\theta)$)
- θ: unknown (scalar or vector) parameter θ.
- We observe a particular value $x = (x_1, \ldots, x_n)$ of X, then a **maximum likelihood estimate** (MLE) is a value of the parameter that maximizes the numerical function $P_X(x_1, \ldots, x_n; \theta)$ (or $f_X(x_1, \ldots, x_n; \theta)$) over all θ:

$$\hat{\theta}_n = \arg\max_{\theta} P_X(x_1, \ldots, x_n; \theta)$$

$$\hat{\theta}_n = \arg\max_{\theta} f_X(x_1, \ldots, x_n; \theta)$$

Maximum Likelihood Estimation



MLE under Independent Case

- Observations X_i are independent, and we observe a particular value $x = (x_1, \ldots, x_n)$ of X.
- We define the log-likelihood function as follows:

$$\log[P_X(x_1, \ldots, x_n; \theta)] = \log \prod_{i=1}^{n} P_{X_i}(x_i; \theta) = \sum_{i=1}^{n} \log[P_{X_i}(x_i; \theta)]$$

$$\log[f_X(x_1, \ldots, x_n; \theta)] = \log \prod_{i=1}^{n} f_{X_i}(x_i; \theta) = \sum_{i=1}^{n} \log[f_{X_i}(x_i; \theta)]$$

MLE under Independent Case

- Thus a **maximum likelihood estimate** (MLE) under independent case is a value of the parameter that maximizes the numerical function $P_X(x_1, \ldots, x_n; \theta)$ (or $f_X(x_1, \ldots, x_n; \theta)$) over all θ:

$$\hat{\theta}_n = \arg\max_{\theta} \sum_{i=1}^{n} \log[P_{X_i}(x_i; \theta)]$$

$$\hat{\theta}_n = \arg\max_{\theta} \sum_{i=1}^{n} \log[f_{X_i}(x_i; \theta)]$$

Example: Revisit Biased Coin Problem

1°. n independent Bernoulli trials. $X_1, \dots X_n \sim \text{Bern}(p)$

p: unknown constant

2°. $x_1, \dots x_n$ real number. $x_i = 1$ or 0.

$$\underline{P_{\mathbf{X}}(x;p)} = \prod_{i=1}^{n} \underline{P_{X_i}(x_i;p)} = \prod_{i=1}^{n} \underline{p^{x_i}(1-p)^{1-x_i}} \qquad \begin{cases} p & x_i=1 \\ 1-p & x_i=0 \end{cases}$$

$$= p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i} = p^{S_n}(1-p)^{n-S_n}$$

$S_n = \sum_{i=1}^n x_i$: # of heads in n coin tosses

3°. $\log P_X(x;p) = \underline{S_n \log p + (n-S_n)\log(1-p)} = f(p)$

$$\hat{p}_{MLE} = \arg\max_p f(p) \qquad [\ f'(p)=0,\ f''(p) \le 0 \ \dots\ ]$$

$$\Rightarrow \hat{p}_{MLE} = \frac{1}{n} S_n = \underline{\frac{1}{n}(x_1 + \dots + x_n)}$$

Outline

Normal Distribution: MLE Perspective

1°. θ : real, unknown, ; n independent measurements.

$X_1, X_2, \ldots X_n$ r.v.s.

error $E_1, E_2, \ldots E_n$ r.v.s.

$E_i = X_i - \theta$; ($e_i = x_i - \theta$)

2°. Suppose E_i is i.i.d. with PDF $f_{E_i}(e_i) = f(e_i)$

Since $X_i = E_i + \theta$, PDF of X_i is also i.i.d.

with PDF $f_{X_i}(x_i)$

$\mathbf{X} = (X_1, \ldots X_n)$

$x = (x_1, \ldots x_n)$

$= f_{E_i}(e_i) = f(x_i - \theta)$

thus the likelihood function $L(\theta) = f_{\mathbf{X}}(x) = \prod_{i=1}^{n} f_{X_i}(x_i)$

$= f(x_1 - \theta) \cdots f(x_n - \theta)$

Normal Distribution: MLE Perspective

$3^\circ.\quad \theta = \hat{\theta} = \frac{1}{n}(x_1 + \dots + x_n) = \underset{\theta}{\arg\max}\, L(\theta)$

$L'(\theta)\Big|_{\theta=\hat{\theta}} = 0 \quad \Rightarrow \frac{\partial \log L(\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}} = 0$

$\Rightarrow \frac{\partial}{\partial\theta}\left[\sum_{i=1}^{n} \log f(x_i - \theta)\right]\Big|_{\theta=\hat{\theta}} = 0$

$\Rightarrow \sum_{i=1}^{n} \frac{f'(x_i - \theta)}{f(x_i - \theta)}\Big|_{\theta=\hat{\theta}} = 0 \qquad \vdots\ g(x) = \frac{f'(x)}{f(x)}.$

$\Rightarrow \sum_{i=1}^{n} g(x_i - \hat{\theta}) = 0 \qquad \left(\hat{\theta} = \frac{1}{n}(x_1 + \dots + x_n) = \bar{x}\right)$

$\Rightarrow \underline{\sum_{i=1}^{n} g(x_i - \bar{x}) = 0}$

$\Leftrightarrow n = 2;\ g(x_1 - \bar{x}) + g(x_2 - \bar{x}) = 0$

$x_1 - \bar{x} = x_1 - \frac{1}{2}(x_1 + x_2) = -\left(x_2 - \frac{x_1 + x_2}{2}\right) = -(x_2 - \bar{x})$

$\Rightarrow \underline{g(-x) = -g(x)}$

Normal Distribution: MLE Perspective

<2> Let $n = m+1$, $x_1 = x_2 = \cdots = x_m = -x$, $x_{m+1} = mx$, $\bar{x} = 0$

$$\Rightarrow \sum_{i=1}^{n} g(x_i - \bar{x}) = \underline{\sum_{i=1}^{n} g(x_i)} = m g(-x) + g(mx) = 0$$

$$\Rightarrow g(mx) = -m\underline{g(-x)} = m g(x) \quad \forall m$$

$$\Rightarrow g(x) = cx \quad \Rightarrow \frac{f'(x)}{f(x)} = cx$$

$$\Rightarrow \frac{df(x)}{f(x)} = cx\,dx \quad \Rightarrow \underline{f(x) = M \cdot e^{c'x^2}}$$

Since $f(x)$ is a valid PDF, $\int_{-\infty}^{\infty} f(x)dx = 1$

$$\Rightarrow f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \quad (\sim N(0, \sigma^2))$$

$$L = f(x_1 - \theta) \cdots f(x_n - \theta) = f(e_1) \cdots f(e_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\left(\underline{\sum_{i=1}^{n} e_i^2}\right)\right\}$$

$\underline{\text{Maximize } L}$ $\Leftrightarrow$ minimize $\underline{\sum_{i=1}^{n} e_i^2}$. $\underline{\text{Least square}}$

Normal Distribution: Information Theory Perspective

Given a continuous r.v. X . PDF $f(x)$.

Entropy $H(X) = -\int f(x) \log f(x) dx$

Optimization problem :

$$\max_f H(x)$$

$$\text{s.t.} \quad \int x f(x) dx = \mu \quad (E(X) = \mu)$$

$$\underline{\int (x-\mu)^2 f(x) dx = \sigma^2} \quad (Var(x) = \sigma^2)$$

$$\Rightarrow f^*(x) \sim N(\mu, \sigma^2)$$

Normal Distribution: Information Theory Perspective

$[\ x>0,\ \log x \le x-1\ ]$

Consider $f(x)$, $q(x)$ (two valid PDFs).

Kullback-Leibler divergence

1° $\int f(x)\log\frac{q(x)}{f(x)}dx \le \int f(x)\left[\frac{q(x)}{f(x)}-1\right]dx = \int[q(x)-f(x)]dx$

$= \int q(x)dx - \int f(x)dx = 1-1 = 0$

2°. $\int f(x)[\log q(x) - \log f(x)]\,dx \le 0$

$\Rightarrow H(x) = -\int f(x)\log f(x)dx \le -\int f(x)\log q(x)dx$

3°. $q(x) \sim N(\mu,\sigma^2)$; $\Rightarrow H(x) \le -\int f(x)\log\left(\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)dx$

$= \int f(x)\left[\frac{(x-\mu)^2}{2\sigma^2}+\log\sqrt{2\pi\sigma^2}\right]dx = \int f(x)\frac{(x-\mu)^2}{2\sigma^2}dx + \log\sqrt{2\pi\sigma^2}$

$= \frac{1}{2}+\log\sqrt{2\pi\sigma^2} = \frac{1}{2}(1+\log(2\pi\sigma^2))$

when $f^*(x) \sim N(\mu,\sigma^2)$, the inequality holds.

~~Normal Distribution~~ MLE: Information Theory Perspective

$E(X) \approx \frac{1}{N}(x_1 + \cdots + x_N)$

$Pr(x)$: real distribution

$P_a(X;\theta)$: approximated distribution.

N samples iid. $\sim Pr(x)$ $\qquad X = (x_1, \dots, x_N)$ real samples.

$$P_a(X;\theta) = \prod_{i=1}^{N} P_a(x_i;\theta) = L(\theta)$$

$$\theta^* = \arg\max_\theta L(\theta) = \arg\max_\theta \log L(\theta) = \arg\max_\theta \sum_{i=1}^{N} \log P_a(x_i;\theta)$$

$$= \arg\max_\theta \frac{1}{N}\sum_{i=1}^{N} \log P_a(x_i;\theta) \approx \arg\max_\theta \mathop{E}_{x\sim Pr(x)}[\log P_a(x;\theta)]$$

$$= \arg\max_\theta \int_x Pr(x)\log P_a(x;\theta)\,dx = \arg\min_\theta \underbrace{-\int_x Pr(x)\log P_a(x;\theta)\,dx}_{\text{cross-entropy}}$$

MLE $\Leftrightarrow$ minimize cross-entropy

Outline

Central Limit Theorem

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

$E(X_i) = \mu$

$Var(X_i) = \sigma^2$

Theorem

As $n \to \infty$,

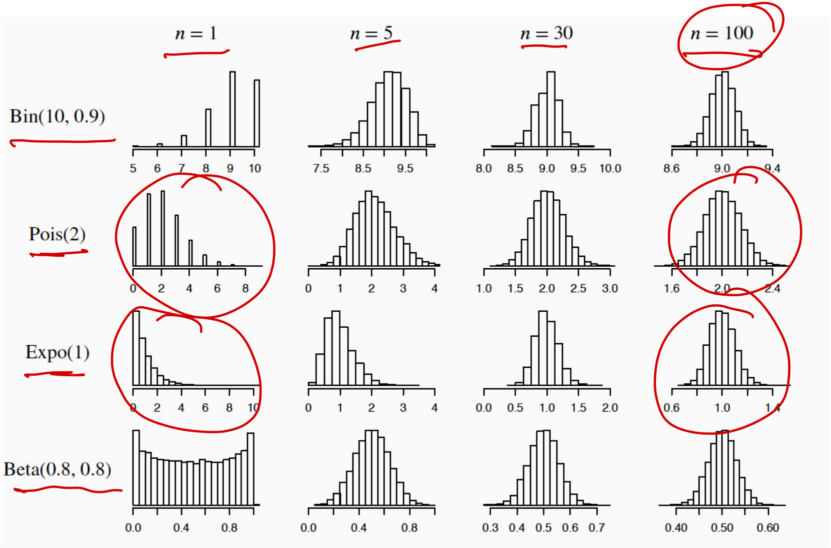
$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \to \mathcal{N}(0, 1) \text{ in distribution.}$$

In words, the CDF of the left-hand side approaches the CDF of the standard Normal distribution.

CLT Approximation

- For large n, the distribution of $\bar{X}_n$ is approximately $\mathcal{N}(\mu, \sigma^2/n)$.
- For large n, the distribution of $n\bar{X}_n = X_1 + \ldots + X_n$ is approximately $\mathcal{N}(n\mu, n\sigma^2)$.

CLT Approximation: Example



Poisson Convergence to Normal

Let $Y \sim Pois(n)$. We can consider Y to be a sum of n i.i.d. Pois(1) r.v.s. Therefore, for large n,

$$Y \sim \mathcal{N}(n, n)$$

Gamma Convergence to Normal

Let $Y \sim Gamma(n, \lambda)$. We can consider Y to be a sum of n i.i.d. $Expo(\lambda)$ r.v.s. Therefore, for large n,

$$Y \sim \mathcal{N}(\frac{n}{\lambda}, \frac{n}{\lambda^2}).$$

Binomial Convergence to Normal

Let $Y \sim Bin(n, p)$. We can consider Y to be a sum of n i.i.d. Bern(p) r.v.s. Therefore, for large n,

$$Y \sim \mathcal{N}(np, np(1-p)).$$

Continuity Correction: De Moivre-Laplace Approximation

$$P(Y = k) = P(k - \frac{1}{2} < Y < k + \frac{1}{2})$$
$$\approx \Phi(\frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}) - \Phi(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}).$$

- Poisson approximation: when n is large and p is small
- Normal approximation: when n is large and p is around $1/2$.

De Moivre-Laplace Approximation

$$P(k \leq Y \leq l) = P(k - \frac{1}{2} < Y < l + \frac{1}{2})$$
$$\approx \Phi(\frac{l + \frac{1}{2} - np}{\sqrt{np(1-p)}}) - \Phi(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}).$$

- Very good approximation when $n \leq 50$ and p is around $1/2$.

Example

Let $Y \sim Bin(n, p)$ with $n = 36$ and $p = 0.5$.

- An exact calculation: $P(Y \leq 21) = 0.8785$
- CLT approximation:
  $P(Y \leq 21) \approx \Phi(\frac{21-np}{\sqrt{np(1-p)}}) = \Phi(1) = 0.8413$
- DML approximation:
  $P(Y \leq 21) \approx \Phi(\frac{21.5-np}{\sqrt{np(1-p)}}) = \Phi(1.17) = 0.879$

History

- 1733: normal distribution was introduced by French mathematician Abraham DeMoivre
- Abraham DeMoivre (1667–1754): worked at betting shop, computing the probability of gambling bets in all types of games of chance. Also a close friend of Isaac Newton.
- 1809: rediscovered by German mathematician Karl Friedrich Gauss, and then people call it the Gaussian distribution.

History

- During the mid-to-late 19th century, most statisticians started to believe that the majority of data sets would have histograms conforming to the Gaussian bell-shaped form.
- Indeed, it came to be accepted that it was “normal” for any well-behaved data set to follow this curve.
- Following the lead of the British statistician Karl Pearson, we also call “normal distribution”.

Family of Normal Distribution

- Chi-Square Distribution: Found by Karl Pearson
- Student-t Distribution: Found by Student (William Gosset)
- F-distribution: Found by Ronald Fisher

Family of Normal Distribution

Given i.i.d. r.v.s $X_i \sim \mathcal{N}(0,1)$, $Y_j \sim \mathcal{N}(0,1)$, $i = 1, \ldots, n$, $j = 1, \ldots, m$. Then we have

- Chi-Square Distribution

$$\chi_n^2 = X_1^2 + \ldots + X_n^2$$

- Student-t Distribution

$$t = \frac{Y_1}{\sqrt{\frac{X_1^2 + \ldots + X_n^2}{n}}}$$

- F-distribution:

$$F = \frac{\frac{X_1^2 + \ldots + X_n^2}{n}}{\frac{Y_1^2 + \ldots + Y_m^2}{m}}$$

Chi-Square Distribution

Definition

Let $V = Z_1^2 + ... + Z_n^2$ where $Z_1, Z_2, ..., Z_n$ are i.i.d. $\mathcal{N}(0, 1)$. Then V is said to have the *Chi-Square distribution with n degrees of freedom*. We write this as $V \sim \chi_n^2$.

Chi-Square & Gamma

Theorem

The χ^2_n distribution is the Gamma$(\frac{n}{2}, \frac{1}{2})$ distribution.

Distribution of Sample Variance

For i.i.d. $X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$, the sample variance is the r.v.

$$S_n^2 = \frac{1}{n-1} \sum_{j=1}^{n} \left(X_j - \bar{X}_n\right)^2.$$

and we have

$$\frac{(n-1)\, S_n^2}{\sigma^2} \sim \mathcal{X}_{n-1}^2.$$

Student-t Distribution

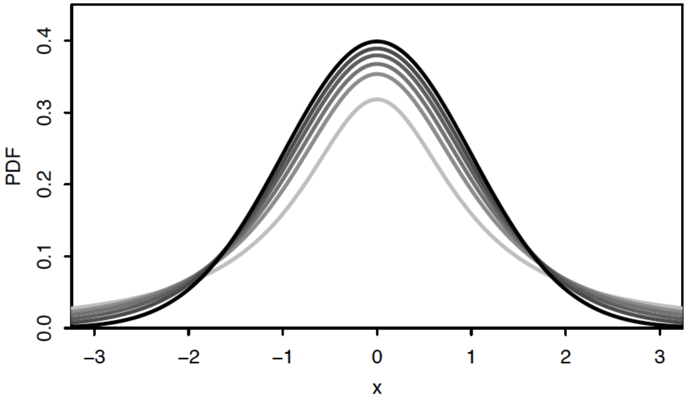
Definition

Let

$$T = \frac{Z}{\sqrt{V/n}},$$

where $Z \sim \mathcal{N}(0,1)$, $V \sim \mathcal{X}_n^2$, and Z is independent of V. Then T is said to have the *Student-t distribution with n degrees of freedom*. We write this as $T \sim t_n$. Often "Student-t distribution" is abbreviated to "t distribution".

PDF of Student-t Distribution



Properties of Student-t Distribution

Theorem

The Student-t distribution has the following properties.

1. *Symmetry: If $T \sim t_n$, then $-T \sim t_n$ as well.*
2. *Cauchy as special case: The t_1 distribution is the same as the Cauchy distribution.*
3. *Convergence to Normal: As $n \to \infty$, the t_n distribution approaches the standard Normal distribution.*

Sample Mean and Sample Variance

Theorem

For i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, the sample mean and sample variance are shown as follows

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

$$S_n^2 = \frac{1}{n-1} \sum_{j=1}^{n} \left(X_j - \bar{X}_n\right)^2.$$

The random variable

$$T = \frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n}}}$$

has a student t-distribution with $n-1$ degrees of freedom.

Outline

Confidence Intervals

Confidence Intervals

- A **confidence interval** for a scalar unknown parameter θ is an interval whose endpoints $\hat{\Theta}_n^-$ and $\hat{\Theta}_n^+$ bracket θ with a given high probability.
- $\hat{\Theta}_n^-$ and $\hat{\Theta}_n^+$ are random variables that depend on the observations $X_1, \ldots, X_n$.
- A $1 - \alpha$ confidence interval is one that satisfies

$$\mathbf{P}_\theta\big(\hat{\Theta}_n^- \leq \theta \leq \hat{\Theta}_n^+\big) \geq 1 - \alpha,$$

$\alpha = 0.05$

for all possible values of θ.

Example: I.I.D. Normal Random Variables

1^o. $X_i \sim \text{i.i.d. } N(\theta, v)$. θ : unknown constant

v : known constant.

Sample mean $\hat{\theta}_n = \frac{X_1+\dots+X_n}{n} \sim N(\theta, \frac{v}{n})$

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\frac{v}{n}}} \sim N(0,1)$$

2^o. θ : confidence interval. $1-\alpha = 0.95$ $\alpha = 0.05$.

$\Phi(x) = P(Z \le x)$

$Z \sim N(0,1)$

$\Phi(1.96) = P(Z \le 1.96) = 0.975 = 1 - \frac{\alpha}{2}$.

$$P_\theta\left(\frac{|\hat{\theta}_n - \theta|}{\sqrt{\frac{v}{n}}} \le 1.96\right) = P_\theta(|Z| \le 1.96)$$

$$= P_\theta(-1.96 \le Z \le 1.96) = 2P(Z \le 1.96) - 1$$

$$= P_\theta(Z \le 1.96) - P_\theta(Z < -1.96) = P_\theta(Z \le 1.96) - P_\theta(Z > 1.96)$$

Example: I.I.D. Normal Random Variables

$$P_\theta\left(\frac{|\hat{\theta}_n-\theta|}{\sqrt{\frac{v}{n}}}\le 1.96\right) = 2P_\theta(Z\le 1.96)-1$$

$$= 2\phi(1.96)-1 = 2\left(1-\frac{\alpha}{2}\right)-1$$

$$= 2-\alpha-1 = 1-\alpha = 0.95$$

3^0. $P_\theta\left(\underline{\hat{\theta}_n-1.96\sqrt{\frac{v}{n}}}\le\theta\le\underline{\hat{\theta}_n+1.96\sqrt{\frac{v}{n}}}\right)=0.95$

$\left[\hat{\theta}_n-1.96\sqrt{\frac{v}{n}},\ \hat{\theta}_n+1.96\sqrt{\frac{v}{n}}\right]$ is a

95% Confidence interval.

Reference

- Chapter 9 in Textbook **BT**
- Chapter 10 in Textbook **BH**