Lecture 14

§1 Review

1. Statistics 的定义

Definition

- Given samples: $X_1, X_2 \ldots, X_n$
- Goal: estimate an unknown parameter θ of the true model, (we use samples).

The hat notation

The point estimator used to estimate a parameter θ is usually denoted as $\hat{\theta}$.

2. MLE

Formal Definition

- Given a model with an unknown parameter θ
- Given samples: $X_1,\ X_2 \ldots,\ X_n$
- The probability that the models generates the samples is called **likelihood**. $L(\theta) = P(X_1,\ X_2 \ldots,\ X_n \mid \theta)$
- To determine the best θ, we choose $\hat{\theta}$ such that $L(\theta)$ (or equivalently, the **log-likelihood** $l(\theta) = \log L(\theta)$)is maximize at $\theta = \hat{\theta}$. **Maximum likelihood estimate (MLE)**

$l'(\theta) = 0$

Likelihood function

- Given a model with an unknown parameter θ
- Given samples: $X_1,\ X_2 \ldots,\ X_n$

Continuous model:

- Likelihood: $L(\theta) = \Pi_i\ f(X_i \mid \theta)$
- Log-Likelihood: $l(\theta) = \sum_i \log(f(X_i \mid \theta))$

f: the model's probability density function (PDF)

Discrete model:

- Likelihood: $L(\theta) = \Pi_i P(X_i \mid \theta)$
- Log-Likelihood: $l(\theta) = \sum_i \log(P(X_i \mid \theta))$

P: the model's probability mass function (PMF)

例:

Example continued

- For drug experiments,
  - **N** experiments
  - **M** successes

$N_1 + N_2 = N$

$M_1 + M_2 = M$

$\frac{M_1}{N_1}$ $\frac{M_2}{N_2}$

- Possible models: Bernoulli distribution with probability p.
- Given one model, the probability of generating such samples:

$$L(p) = p^M (1-p)^{N-M}$$

- Maximize the probability L(p)

Example continued

- Maximize the probability

$$L(p) = p^{M}(1-p)^{N-M}$$

We often maximize the **logarithm** of the probability (usually **easier** to maximize) of generating such samples

$$l(p)=\log L(p) = M * \log p + (N-M) * \log (1-p)$$

Example continued

We often maximize the **logarithm of** the probability (usually easier to maximize) of generating such samples

$$l(p)=\log L(p) = M * \log p + (N-M) * \log (1-p)$$

- For continuous function l(p)
  - Find the point $l'(p) = 0$ and $l''(p)<0$
  - Before we discuss optimization, all exercises with continuous cases just require you to show the point $l'(p) = 0$.

- $l'(p)= \frac{M}{p} - \frac{N-M}{1-p} = 0$  $\hat{p} = M/N$ (sample cure rate)

例

Example continued

Suppose we have two coins with head probability being p_1 and p_2=0.3, respectively.
N_1 experiments with M_1 heads using coin 1 ;
N_2 experiments with M_2 heads using coin 2.

$\hat{p}_1 = \frac{M_1}{N_1}$ $\hat{p}_2 = \frac{M_2}{N_2}$

Given p, the probability of generating such samples (likelihood function):

$$L(p) = p^{M_1}(1-p)^{N_1-M_1}(\underbrace{p+0.3}_{p_2})^{M_2}(\underbrace{0.7-p}_{1-p_2})^{N_2-M_2}$$

Loglikelihood:

$$l(p) = M_1 \log p + (N_1 - M_1)\log(1-p) + M_2 \log(p+0.3) + (N_2 - M_2)\log(0.7-p)$$

$$l'(p) = \frac{M_1}{p} + \frac{N_1 - M_1}{1-p} + \frac{M_2}{p+0.3} + \frac{N_2 - M_2}{0.7-p}$$  $\hat{p}$

例

Example: uniform

Let *X* be a Uniform random variable on the interval [*0, θ*]

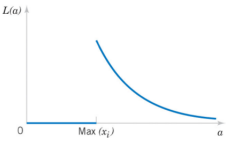
$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, & \text{for } 0 \le x \le \theta, \\ 0, & \text{otherwise,} \end{cases} = \frac{1}{\theta}\mathbf{1}_{\{0\le x\le\theta\}}$$

indicator function $\mathbf{1}_A(x)$

$$\mathbf{1}_A(x) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{otherwise} \end{cases}$$

The likelihood function of a random sample of size *n* is:

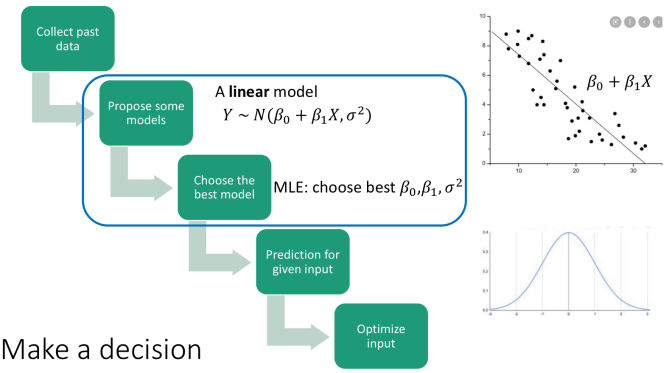
$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \frac{1}{\theta^n}\prod_{i=1}^{n}\mathbf{1}_{\{0\le x_i\le\theta\}} = \begin{cases} \frac{1}{\theta^n}, & \text{if } \theta \ge max\{X_1, X_2, \dots, X_n\} \\ 0, & \text{if } \theta < max\{X_1, X_2, \dots, X_n\} \end{cases}$$



$$\hat{\theta} = max\{X_1, X_2, \dots, X_n\}$$

Calculus methods don't work here because $L(\theta)$ is maximized at the **discontinuity**.
Clearly, θ cannot be smaller than max(x_i), thus the MLE is $max\{X_1, X_2, \dots, X_n\}$.

3. linear regression



- PDF for normal $\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$.
- Samples: $(X_1, Y_1), \ldots, (X_N, Y_N)$
- For the model with $\beta_0, \beta_1, \sigma^2$, the likelihood is

$$\frac{1}{(\sqrt{2\pi})^n \sigma^n}\exp\left[-\frac{1}{2}\frac{\Sigma_i(Y_i-\beta_1X_i-\beta_0)^2}{\sigma^2}\right]$$

- Given σ^2, to maximize the likelihood, we only need to minimize

$$\Sigma_i(Y_i-\beta_1X_i-\beta_0)^2$$

- Taking derivative over β_0 and β_1, we have

$$\Sigma_i(Y_i-\beta_1X_i-\beta_0)X_i=0$$
$$\Sigma_i(Y_i-\beta_1X_i-\beta_0)=0$$

➡ MLE: $\widehat{\beta_1}=\frac{\Sigma_i(X_i-\bar{X})(Y_i-\bar{Y})}{\Sigma_i(X_i-\bar{X})^2}$, $\widehat{\beta_0}=\bar{Y}-\widehat{\beta_1}\bar{X}$

A **linear** model $Y \sim N(\beta_0+\beta_1X, \sigma^2)$

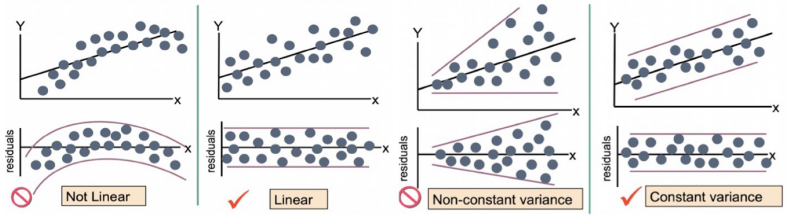
Residual Analysis: check assumptions

Linear regression assumes that…

1. The relationship between X and Y is **linear**
2. The variance of $Y-\beta_0-\beta_1X$ at every value of X is the **same** (homogeneity)

Residual: $e_i = Y_i - \widehat{\beta_0} - \widehat{\beta_1}X_i$

MLE: $\widehat{\beta_1}=\frac{\Sigma_i(X_i-\bar{X})(Y_i-\bar{Y})}{\Sigma_i(X_i-\bar{X})^2}$, $\widehat{\beta_0}=\bar{Y}-\widehat{\beta_1}\bar{X}$



4. Interval estimation

Recap: interval estimation

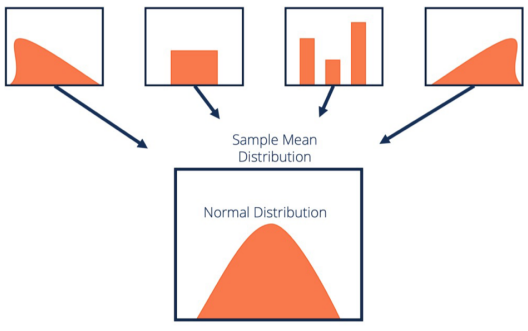
Whether a drug can cure a disease: $\hat{p} = \frac{\Sigma_i X_i}{n}$

- Drug 1: $\hat{p}_1 = 90\%$. 10 experiments.
- Drug 2: $\hat{p}_2 = 80\%$. 10000 experiments.

Which drug do you think is more effective?

With more data, we believe the estimator is closer to the true parameter.

Main technique: Central limit theorem



Sample mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots X_n}{n}$$

$X_1, \ldots, X_n$ can be non-normal

Mean: μ; Variance: σ^2

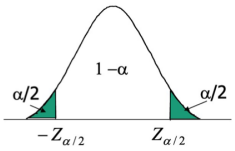
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

No matter what the true distribution is, the **sample mean** will be very close to the **normal distribution**, as long as the sample size is **large**.

Interval Estimation for mean

- We have data $X_1, X_2, \ldots, X_n$ that are sampled from some distribution
- Their mean is $\boldsymbol{\mu}$, which we want to estimate
- We can easily give a point estimate: $\bar{X}$ (sample mean)
- How to get an interval estimate??
  - Use Central Limit Theorem!

Interval Estimation - example



$$Z = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$$

$$\Rightarrow \quad P(-z_{\alpha/2} \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \leq z_{\alpha/2}) = 1-\alpha$$

$$\Rightarrow \quad P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

With probability 1 - α, μ is within the interval $[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$

You can also write as $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Some observations

1-α Confidence interval:

$$[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

We typically call $1-\alpha$ as the **confidence level**.

The length of the confidence interval is affected by several factors

- As the sample size n **increases**, the length of CI **decreases**
- As the variance σ^2 **increases**, the length of CI **increases**
- As the confidence level increases (α **decreases**), the length of CI **increases**.

Interval Estimation for proportion (mean)

Mean: p
Variance: $p(1-p)$
$\hat{p} = \bar{X}$ (sample cure rate)

$$Z = \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}} \sim N(0,1)$$

Recall:
$Z = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$
Interval is $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

$$\Rightarrow \quad P(-z_{\alpha/2} \leq \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}} \leq z_{\alpha/2}) = 1-\alpha$$

$$\Rightarrow \quad P(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}) = 1-\alpha$$

With probability 1 - α, use the approximate interval

$$[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}] \qquad \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

§2 Exercises

例: # Additional Exercise 1:

- Suppose there are **100** white or black balls in the bag, 90 of them are one color and 10 of them are another color. I randomly pick 1 ball out of the bag and find out that it's white. So, by the maximum likelihood estimate, which is the majority color in the bag (what color are the 90 balls)?

case 1: 90 white 10 black $L(\text{case 1}) = \frac{90}{100}$ X_1 = white

case 2: 90 black 10 white $L(\text{case 2}) = \frac{10}{100}$

100 balls

90 in one color
10 in another

Draw one sample randomly

例: # Additional Exercise 2:

- Suppose we ask people to select their favorite ice cream brand: Haagen-Dazs (type 1), Wall's (type 2), Ben & Jerry's (type 3).
- We can view the large population contains 3 different types of individuals, and let $\theta_1, \theta_2, \theta_3$ denote the proportion of individuals of type each type. Clearly, $0 \le \theta_i \le 1$ and $\theta_1 + \theta_2 + \theta_3 = 1$. $\theta_1, \theta_2, \theta_3$ unknown.
- Suppose we ask 100 students in CUHK(SZ), and 45 choose HD, 30 choose Walls, and 25 choose BJ. Find the MLE for $\theta_1, \theta_2, \theta_3$.

HD W BJ

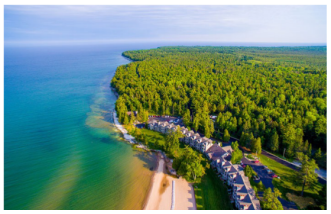
$L(\theta_1, \theta_2, \theta_3) = \theta_1^{45} \cdot \theta_2^{30} \cdot \theta_3^{25}$ $\ell(\theta_1, \theta_2, \theta_3) = 45\log\theta_1 + 30\log\theta_2 + 25\log\theta_3$

$\begin{cases} \frac{45}{\theta_1} - \frac{25}{1-\theta_1-\theta_2} = 0 \\ \frac{30}{\theta_2} - \frac{25}{1-\theta_1-\theta_2} = 0 \end{cases}$ $\hat{\theta}_1 = 0.45$ $\hat{\theta}_2 = 0.3$ $\hat{\theta}_3 = 0.25$

例: # Additional Exercise 3:

- A geologist studied the composition of rocks in the shoreline area of Lake Michigan. He randomly selected 100 samples from the area, each containing 10 stones, and recorded the number of limestone stones in each sample. The geologist's data are as follows:

number of limestone in a sample	0	1	2	3	4	5	6	7	8	9	10
number of samples	0	1	6	7	23	26	21	12	3	1	0



Additional Exercise 3:

number of limestone in a sample	0	1	2	3	4	5	6	7	8	9	10
number of samples	0	1	6	7	23	26	21	12	3	1	0

X_i = # limestone in sample i.

- Assume that these 100 observations are independent of each other. Find the maximum likelihood estimate (MLE) of the proportion p of limestone in the stones in this area.

$L(p) = \prod_{i=1}^{100}\left[\binom{10}{x_i} p^{x_i}(1-p)^{10-x_i}\right]$ 10×100 M

$\hat{p} = \frac{M}{1000}$

$= \left[\prod_{i=1}^{100}\binom{10}{x_i}\right] \cdot p^{\sum x_i} \cdot (1-p)^{1000-\sum x_i}$

$\ell(p) = \log\left[\prod_{i=1}^{100}\binom{10}{x_i}\right] + \sum x_i \cdot \log p - (1000 - \sum x_i)\log(1-p)$