Lecture 11 Statistics Estimation

§1 Statistical inference（统计推断）

1、适用情境

已知大量 samples 的数据，但不知道这些数据所符合的确切的 model.

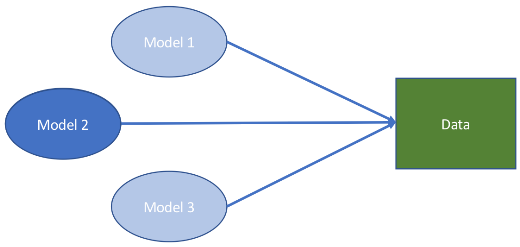
2、目标

基于这些数据，推断 true model，即选出一个最符合数据的概率模型.

The general setup

Example: We have a data set, and we want to **infer** the properties of the underlying distribution from this data set.

- A variety of different probability models must at least be explored to see which model best "fits" the data.



3、类型

1° estimate a parameter（已知数据大致服从某种分布，通过数据求该分布的系数）

2° hypothesis testing（检测某一假设正确的可能性）

§2 Estimate a parameter

1、The hat notation

用来 estimate 某系数 θ 的 point estimator（(点)估计值）通常被表示为 $\hat{\theta}$

2、Definition

给出 samples：$X_1, X_2, \cdots\cdots X_n$

推断 true model 的未知系数 θ

3、估计值 $\hat{\theta}$

属于一种 statistic（统计量）

1° 是一个关于样本 $(X_1, X_2, \cdots\cdots, X_n)$ 的函数

2° $\hat{\theta}$ 从样本中提取出了有效信息.

4、Best $\hat{\theta}$ 的选择

1° 从有限的系数中选出最佳的一个

Example 1

Data

Model 1: Head occurs with Probability 0.5

Model 2: Head occurs with probability 0.3

Model 3: Head occurs with probability 0.7

How likely?

$P(\{X_1=0, X_2=1, X_3=1, X_4=1\} | P=0.5\}$

Likelihood of generating the sample from Model 1: 0.5*0.5*0.5*0.5= 0.0625

$P(\{X_1=0, X_2=1, X_3=1, X_4=1\} | P=0.3\}$

Likelihood of generating the sample from Model 2: 0.3*0.3*0.3*0.7= 0.0189

$P(\{X_1=0, X_2=1, X_3=1, X_4=1\} | P=0.7\}$

Likelihood of generating the sample from Model 3: 0.7*0.7*0.7*0.3= 0.1029

highest

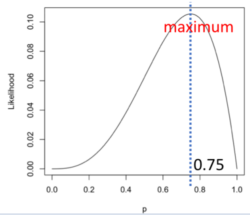
2° 直接根据样本数据推断最佳系数

Example 1

Data

Model:
Head occurs with Probability p

How likely?
Likelihood of generating the sample from Model:
$p*p*p*(1-p)$



Under the model **p=0.75**, the data appears with the maximum probability/likelihood.

Sample proportion $\frac{\Sigma_i x_i}{n}$

要取出最佳的$\hat{\theta}$，就要求 $\max P(\theta=\hat{\theta}|\{x_1,x_2,\cdots\cdots,x_n\})$

$P(\theta=\hat{\theta}|\{x_1,x_2,\cdots\cdots,x_n\})$（在已知样本数据的情况下，真实系数$\theta$=估计值$\hat{\theta}$的概率）

$$= \frac{P(\theta=\hat{\theta}\ \&\ \{x_1,x_2,\cdots\cdots,x_n\})}{P(\{x_1,x_2,\cdots\cdots,x_n\})}$$

$$= \frac{P(\{x_1,x_2,\cdots\cdots,x_n\}|\theta=\hat{\theta})\cdot P(\theta=\hat{\theta})}{P(\{x_1,x_2,\cdots\cdots,x_n\})}$$

（在已知真实系数θ=估计值$\hat{\theta}$的情况下，样本数据为现有数据的概率）

因为在知道最佳的$\hat{\theta}$之前，对任意估计值$\hat{\theta}$，$P(\theta=\hat{\theta})$与$P(\{x_1,x_2,\cdots\cdots,x_n\})$都是相等的，因此要选出最佳的$\hat{\theta}$，就要求$P(\{x_1,x_2,\cdots\cdots,x_n\}|\theta=\hat{\theta})$最大.

5. Formal definition

1° 对任意θ，产生现有样本数据的概率为 likehood

$L(\theta)=P(x_1,x_2,\cdots\cdots,x_n|\theta)$

2° 要求 best θ，我们选择$\hat{\theta}$使得 $L(\theta)$在$\theta=\hat{\theta}$处取最大值

3° 定义 $l(\theta)=\log L(\theta)$ （log 表示 natural logarithm）

被称为 log-likehood

4° Maximum likehood estimate (MLE)（最大似然估计）

选择$\hat{\theta}$使 $l(\theta)$在$\theta=\hat{\theta}$处最大

6. Likehood function

已知 a model with an unknown parameter θ

已知 samples：$x_1, x_2, \cdots\cdots, x_n$.

1° 对于 discrete model：

Likehood：$L(\theta)=\prod_i P(x_i|\theta)$

Log-Likehood：$l(\theta)=\sum_i \log(P(x_i|\theta))$

2° 对于 Continuous model：

Likehood：$L(\theta)=\prod_i f(x_i|\theta)$ （即使对于某一x_i，x_i发生的概率为0）

Log-Likehood：$l(\theta)=\sum_i \log(f(x_i|\theta))$

f为PDF，P为PMF

5、案例分析

案例一：

- 一种药能否救活病人是一个 Bernoulli RV

  Yes $(X=1)$ w.p. p　　　No $(X=0)$ w.p. $1-p$

  p 未知
- 在 N 个样本中，有 M 个成功了
- 推测 p 的值
- 对任意 p，产生现有样本数据的 likehood 为

  $L(p)=p^M\cdot(1-p)^{N-M}$ （无需加上 $\binom{N}{M}$，对结果无影响）
- 要求 $L(p)$ 的最大值，不妨先对两边取对数

  $l(p)=\log L(p)$

  $\quad = M\cdot\log p+(N-M)\cdot\log(1-p)$
- 仅需找到 p 满足 $l'(p)=0$ 即可

  $l'(p)=\frac{M}{p}-\frac{N-M}{1-p}=0$

  $\Rightarrow\ p=\frac{M}{N}$

* 这也解释了为什么我们用 $\frac{\sum_i x_i}{n}$ 去推算 p

案例二：

- Drug 1: $p=0.8$

  Drug 2: $p=0.9$

  Samples: $M=81$ successes from $N=100$ experiments.
- Which drug is being experimented
- $p\in\{0.8,\ 0.9\}$

  $l=\log L=M\cdot\log p+(N-M)\log(1-p)$
- Drug 1: -48.6539

  Drug 2: -52.2833
- Conclusion: drug 1 is more likely to be experimented

案例三：

- 灯泡的寿命服从指数分布，pdf 为 $\lambda e^{-\lambda x}$
- 给出样本 $x_1, x_2, \cdots\cdots, x_n$
- $L(\lambda)=\lambda e^{-\lambda x_1}\cdot\lambda e^{-\lambda x_2}\cdots\cdots$

  $\quad =\lambda^n e^{-\lambda\sum_i x_i}$

  $l(\lambda)=\log L(\lambda)=n\log\lambda-\lambda\sum_i x_i$

- $l'(\lambda) = \frac{n}{\lambda} - \sum_i x_i = 0$

  $\hat{\lambda} = \frac{n}{\sum_i x_i} = \frac{1}{\bar{x}}$

案例四：

- 若数据服从正态分布 $N(\mu, 1)$
- 给出样本 $x_1, x_2, \cdots\cdots, x_n$
- $L(\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot e^{-\frac{1}{2}\sum_i (x_i-\mu)^2}$

  $l(\mu) = \log L(\mu) = -\frac{1}{2}\sum_i (x_i-\mu)^2 - \frac{n}{2}\log 2\pi$
- $l'(\mu) = \sum_i (x_i - \mu) = 0$

  $\hat{\mu} = \bar{x}$

6. 特殊案例：对平均分布进行 MLE

1° Indicator function（指示函数）

① 记作 $\mathbb{1}\{x_i \in [a,b]\}$ 或 $\mathbb{1}_A(x)$（A为指定区间）

② 表示若 x_i 落在 $[a,b]$ / A 内，函数值为 1
若 x_i 不落在 $[a,b]$ / A 内，函数值为 0

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

2° 令 X 为区间 $[0,\theta]$ 上一个均匀分布的变量，

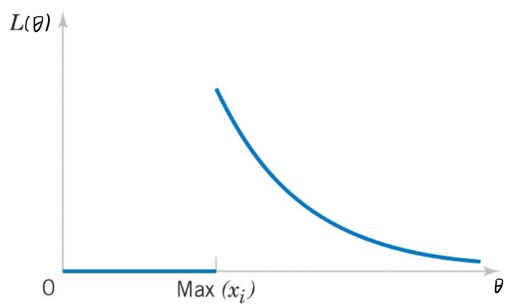
$$f(x) = \begin{cases} \frac{1}{\theta}, & \text{for } 0 \le x \le \theta \\ 0, & \text{otherwise} \end{cases}$$

$$= \frac{1}{\theta} \cdot \mathbb{1}\{x \in [0,\theta]\}$$

则一个大小为 n 的 random sample 的 likehood function 为：

$$L(\theta) = \prod_{i=1}^{n} f(x_i) = \frac{1}{\theta^n} \cdot \prod_{i=1}^{n} \mathbb{1}\{x_i \in [0,\theta]\} = \begin{cases} \frac{1}{\theta^n}, & \text{if } \theta \ge \max\{x_1, x_2, \cdots, x_n\} \\ 0, & \text{if } \theta < \max\{x_1, x_2, \cdots, x_n\} \end{cases}$$

因此，$\hat{\theta} = \max\{x_1, x_2, \cdots\cdots, x_n\}$



例: Baseball Team

- The weight for a baseball team players are
  {150, 143, 132, 160, 175, 190, 123, 154}

- Assume their weights are **uniformly** distributed over an interval [a, b]

- What are good estimators for a? for b?

This example will show that the **MLE could be complicated to solve**, e.g., the equation $l'(\theta) = 0$ may be difficult to solve, or it may **not** always to possible to use calculus methods **directly** to find the maximum of L(θ).

The density function on X is:

$$f(a,b) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

The like hood is:

$$L(a,b) = \begin{cases} \frac{1}{(b-a)^n} & a \le X_i \le b \\ 0 & \text{otherwise} \end{cases}$$

Thus, L(a,b) attains its maximum when b-a gets its minimum.

$$X_{(0)} = \min\{x_1, \cdots\cdots, x_n\}$$

$$X_{(n)} = \max\{x_1, \cdots\cdots, x_n\}$$

we have

$$L(a,b) = \frac{1}{(b-a)^n} \le \frac{1}{(X_{(n)} - X_{(0)})^n}$$

Clearly, L(a,b) attains its maximum when $a = X_{(0)}$ and $b = X_{(n)}$