

# 第十章: Statistical Inference

## - Overview

Important concepts: Population  $\rightarrow$  分布,  $F$

Sample  $\rightarrow$  样本, random vector  $\vec{X} = (X_1, \dots, X_n)$ ,  $n$  sample size

Random Sample  $\rightarrow \{X_i\}$  are i.i.d r.v,  $X_i \sim F$

Data  $\rightarrow$  数据, 真实的向量  $\vec{x} = (x_1, \dots, x_n)$ , i.e.,  $\vec{X}$  值

Statistic  $\rightarrow$  统计, sample  $\vec{X}$  的函数

Goal: From sample to infer property of population

$\hookrightarrow$  给出一个参数模型:  $F = \{p(x; \theta) : \theta \in R\}$

已知 random sample from model,  $\vec{X} = (X_1, \dots, X_n)$ ,  
如何去参数化地统计推理呢?

例: 尝试使用正态分布拟合:

$$\{p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \mu \in R, \sigma > 0\}$$

(Bayesian & Frequentist)  $\hookrightarrow$  take  $\theta$  as a constant

贝叶斯方法与频率方法的区别在于: 如何对待  $\theta$ ?

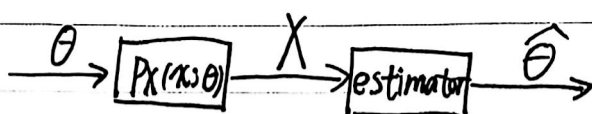
$\hookrightarrow$  take  $\theta$  as r.v. with prior distribution.

Core task of statistical Inference:

• Point Estimation • Interval Estimation

• Hypothesis Testing

## 二. Point Estimation: Frequentist Perspective



可见关键在于如何 design

estimator, 以: " $\hat{\theta} - \theta$ " 误差小!

KOKUYO



Pipeline: 生成  $p(x_i; \theta) \rightarrow$  采样  $\vec{x} = (x_1, \dots, x_n)$   
 $\rightarrow \hat{\theta} = g(\vec{x}) \rightarrow$  预测  $\theta = g(\vec{x})$ ,  $\vec{x}$  为观测数据, i.e.,  $\vec{x} = \vec{x}$

Def: Likelihood, 一个关于  $\theta$  的函数, 输出值是见到  $\vec{x}$  观测数据的概率, i.e.,  $p(\vec{x}; \theta)$

则: Maximum Likelihood Estimate (MLE) 是使  $p(\vec{x}; \theta)$  最大的参数  $\theta$ :

$$\hat{\theta} = \operatorname{argmax}_{\theta} p(\vec{x}; \theta)$$

若  $\{x_i\}$  are i.i.d, 则考虑:  $\log [p(\vec{x}; \theta)]$   
 $= \log \prod_{i=1}^n p(x_i; \theta) = \sum_{i=1}^n \log [p(x_i; \theta)]$

$$\text{则 } \hat{\theta} = \operatorname{argmax}_{\theta} \sum_{i=1}^n \log [p(x_i; \theta)]$$

例: 一枚硬币,  $p$  概率头朝上。在采的数据中, 有  $n$  次投掷, 令  $X$  代表  $n$  次投掷中头朝上次数, 发现  $X = k$ 。  
 用 MLE 求  $\hat{p}$ , i.e., 近似  $p$ 。

Solution: 生成  $p(x_i; \theta)$ :  $p_{x_i} = p^{x_i} (1-p)^{1-x_i}$

因为:  $x_i = 1$  代表 Head,  $p_{x_i} = p$ ;  $x_i = 0$ , Tail,  $p_{x_i} = 1-p$ 。

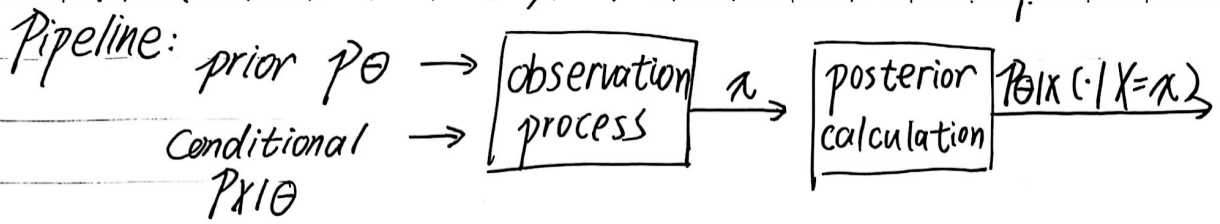
$$\text{则 } P_{\vec{X}}(\vec{x}; p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^k (1-p)^{n-k}, \quad \boxed{X=k}$$

$$\text{令 } g(p) = \log [p^k (1-p)^{n-k}] = k \log p + (n-k) \log (1-p)$$

$$g'(p) = (k \log p + (n-k) \log (1-p))' = \frac{k}{p} - \frac{n-k}{1-p} = 0 \quad \therefore \hat{p} = \frac{k}{n}$$



### 三. Point Estimation: Bayesian Statistical Inference



最后给出的是  $P_{\theta|X}(\cdot|x)$  or  $J_{\theta|X}(\cdot|x)$ .

然后拟合出  $\theta$  的手段:

①  $\hat{\theta} = E(\theta|X=x)$  (Posterior Mean)

② MAP (Maximum A Posteriori Probability)

$$\hat{\theta} = \operatorname{argmax}_{\theta} P_{\theta|X}(\theta|x)$$

or:  $\hat{\theta} = \operatorname{argmax}_{\theta} P_{\theta}(\theta) p_{X|\theta}(x|\theta)$

而 prior  $p_\theta$  选什么好呢? 常见的有 Beta & Gamma 分布, 见下:

#### 四: Beta & Gamma 分布

Beta: Def: 参数有  $a, b$ ,  $a, b > 0$ , PDF 为:

$$f(x) = \frac{1}{\beta(a, b)} x^{a-1} (1-x)^{b-1}, \quad x \in (0, 1), \quad X \sim \text{Beta}(a, b)$$

而  $\beta(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx$  (用于归一化)

Gamma: Def: 参数有  $a, \lambda$ ,  $a, \lambda > 0$ , PDF 为:

$$f(y) = \frac{1}{\Gamma(a)} (\lambda y)^a e^{-\lambda y} \frac{1}{y}, \quad y > 0, \quad Y \sim \text{Gamma}(a, \lambda)$$

而:  $\Gamma(a) = \int_0^{\infty} x^a e^{-x} \frac{dx}{x}$





Theorem:  $X_1, \dots, X_n$  i.i.d,  $\sim \text{Expo}(\lambda)$ . 则:

$$X_1 + \dots + X_n \sim \text{Gamma}(n, \lambda)$$

☆:  $X_i$  矩母为:  $\frac{\lambda}{\lambda-t}$ , 则  $X_1 + \dots + X_n = \text{Gamma}(n, \lambda)$  矩母为:

$$M_n(t) = \left(\frac{\lambda}{\lambda-t}\right)^n$$

Proof (Gamma 矩母为它  $\uparrow$ ):

$$\begin{aligned} E(e^{tx}) &= \int_0^{+\infty} e^{ty} \frac{1}{\Gamma(n)} (ly)^{n-1} e^{-ly} \frac{dy}{y} \\ &= \frac{\lambda^n}{(\lambda-t)^n} \int_0^{+\infty} \frac{1}{\Gamma(n)} e^{-(\lambda-ty)y} \frac{dy}{y} = \frac{\lambda^n}{(\lambda-t)^n} \end{aligned}$$

Gamma( $n, \lambda-t$ ) 的 PDF

Gamma 与 Beta 之间也有联系:

Independent Gamma r.v.s  $X$  &  $Y$ ,  $\lambda$  相同; 则  $X+Y$  有相同分布, 而  $\frac{X}{X+Y}$  也是一个 Beta 分布

## 五. Conjugate Prior: A Weapon of Bayesian

☆: 先验分布与似然模型共轭, if 先验与后验分布都在一个 distribution family 中

而 Gamma 与 Beta 正好有共轭性: We say that Beta is the conjugate prior of the Binomial. (If we have a Beta Prior on  $p$  and data are conditionally Binomial given  $p$ )

