第十章：Statistical Inference

一. Overview

Important concepts: Population → 分布, $\mathcal{F}$

Sample → 样本, random vector $\vec{X} = (X_1, \cdots, X_n)$, n为sample size

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

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

Statistic → 统计, sample $\vec{X}$的函数

Goal: From sample to infer property of population

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

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

例: 尝试使用正态分布拟合:

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

(Bayesian & Frequentist) ↱ take θ as a constant

贝叶斯方法与频率方法的区别在于: 如何对待 θ ?

↳ take θ as r.v. with prior distribution.

Core task of statistical Inference:

- Point Estimation
- Interval Estimation
- Hypothesis Testing

二. Point Estimation: Frequentist Perspective

$$\xrightarrow{\theta} \boxed{P_X(x;\theta)} \xrightarrow{X} \boxed{\text{estimator}} \xrightarrow{\hat{\theta}}$$

可见关键在于如何design estimator, 以: "$\hat{\theta}-\theta$"误差小!

No.

Date

Pipeline：生成 $p(x;\theta) \longrightarrow$ 采样 $\vec{X}=(x_1,\cdots,x_n)$
$\longrightarrow \hat{\theta}=g(\vec{X}) \longrightarrow$ 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 θ 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 θ：

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

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

$$=\log\prod_{i=1}^{n} p(x_i;\theta)=\sum_{i=1}^{n}\log[p(x_i;\theta)]$$

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

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

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

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

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

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

No.

Date

三. Point Estimation: Bayesian Statistical Inference

Pipeline: prior p_Θ → observation process → x → posterior calculation → $p_{\Theta|X}(\cdot|X=x)$

Conditional $p_{X|\Theta}$ → observation process

最后给出的是 $p_{\Theta|X}(\cdot|x)$ or. $f_{\Theta|X}(\cdot|x)$.

然后拟合出 θ 的手段:

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

② MAP (Maximum A Posteriori Probability)

$$\hat{\theta} = \arg\max_{\theta} p_{\Theta|X}(\theta|x)$$

or: $$\hat{\theta} = \arg\max_{\theta} p_\Theta(\theta)\, p_{X|\Theta}(x|\theta)$$

而 prior p_Θ 选什么好呢? 常见的含 Beta & Gamma 分布, 见下:

四: Beta & Gamma 分布

Beta: Def: 参数有 a.b, a,b>0, PDF 为:

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

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, λ, $a, \lambda>0$, PDF 为:

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

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

No.

Date

Theorem: $X_1, \cdots, X_n$ i.i.d. $\sim$ Expo(λ). 则:

$$X_1 + \cdots + X_n \sim Gamma(n, \lambda)$$

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

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

Proof (Gamma矩母为它↑):

$$E(e^{tY}) = \int_0^{+\infty} e^{ty} \frac{1}{\Gamma(n)} (\lambda y)^n e^{-\lambda y} \frac{dy}{y}$$

$$= \frac{\lambda^n}{(\lambda - t)^n} \int_0^{+\infty} \underbrace{\frac{1}{\Gamma(n)} e^{-(\lambda - t)y} [(\lambda - t)y]^n \frac{dy}{y}}_{\text{Gamma}(n,\ \lambda - t)\text{的PDF}} = \frac{\lambda^n}{(\lambda - t)^n}$$

Gamma与 Beta之间也有联系:

Independent Gamma r.v.s X & Y, λ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)