



# 点估计

苏勤亮

中山大学计算机学院

超算中心503M

[suqliang@mail.sysu.edu.cn](mailto:suqliang@mail.sysu.edu.cn)

- 参数估计及评价准则
- 矩估计
- 最大似然估计
- 最小方差无偏估计
- Cramer-Rao不等式
- 贝叶斯估计

# 参数估计问题

**问题描述：** 设 $X_1, \dots, X_n$ 是从总体 $F_\theta$ 中抽取的简单随机样本， $F_\theta$ 的分布形式已知，但参数 $\theta$ 未知. 参数估计就是要利用样本 $X_1, \dots, X_n$ 对未知**参数 $\theta$** 或其**函数 $g(\theta)$** 做出估计

- **例如：** 设 $X_1, \dots, X_n$ 是从正态分布族 $\{N(\mu, \sigma^2), -\infty < \mu < +\infty, \sigma^2 > 0\}$ 抽出的简单随机样本
  - 可以根据样本去估计分布参数 $\theta = (\mu, \sigma^2)$ 的值
  - 若对 $\mu, \sigma^2$ 没有兴趣，而是对它的变异系数 $\sigma/\mu$ 感兴趣，也可以根据样本去估计该分布的参数的函数 $g(\theta) = \sigma/\mu$

## ■ 参数估计方法的分类

- 1) 点估计：用一个具体数值去估计未知参数 $\theta$
- 2) 区间估计：用一个区间去估计未知参数 $\theta$ 的范围
- 3) 贝叶斯估计：用一个分布去估计未知参数 $\theta$ 的概率分布

## ■ 点估计的基本概念

- 在正态分布族的例子中，抽样样本的观察值 $x_1, x_2, \dots, x_n$ 一旦确定，参数 $\mu$ 的估计值为

$$\hat{\mu} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

- 估计量（统计量）

$$\hat{\mu} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

- 用来估计未知参数的统计量  $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$  称为参数  $\theta$  的点估计量（point estimator）；而样本观察值点函数  $\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$  称为参数的估计值
- 可以看出，估计量  $\hat{\theta}(X_1, X_2, \dots, X_n)$  也具有二重性，当抽样样本值不知道时，它是随机变量；但当观察到样本的值后，变成一个固定值

# 估计量评价准则——无偏性

- **无偏性：** 设 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 是参数 $\theta$ 的一个估计量。若 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 的期望等于真实值 $\theta$ ，即：

$$E[\hat{\theta}] = \theta, \quad \forall \theta \in \Theta$$

则称 $\hat{\theta}$ 为 $\theta$ 的无偏估计 (unbiased estimator)

**例：** 设 $X_1, X_2, \dots, X_n$ 是取自均值为 $\mu$ ，方差为 $\sigma^2$ 总体的一个样本。

试说明样本均值估计量 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ，样本方差估计量 $S_n^2 =$

$\frac{1}{n} (X - \bar{X})^2$ 是否为无偏估计量

$$\begin{aligned}
 E[\bar{X}] &= E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] \\
 &= \frac{E[X_1] + E[X_2] + \cdots + E[X_n]}{n} \\
 &= \frac{n\mu}{n} = \mu
 \end{aligned}$$

➤ 因此，估计量 $\bar{X}$ 是无偏估计量

- **无偏性含义：** 如果无偏估计器多次重复使用，那么每次估计结果的平均值会随着使用次数的增加而趋近于真实值
- $|E[\hat{\theta}] - \theta|$ 表示估计量的系统性偏差

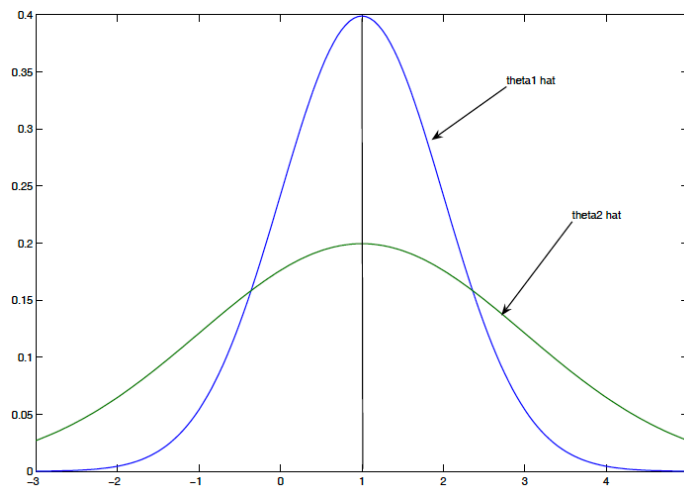
$$\begin{aligned}
E[S_n^2] &= E\left[\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right] = \frac{1}{n}E\left[\sum_{i=1}^n ((X_i - \mu) - (\bar{X} - \mu))^2\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^n (X_i - \mu)^2 - 2\left(\sum_{i=1}^n (X_i - \mu)\right)(\bar{X} - \mu) + n(\bar{X} - \mu)^2\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^n E[(X_i - \mu)^2] - E[(\bar{X} - \mu)^2] \\
&= \sigma^2 - \frac{1}{n}\sigma^2 \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

估计量  $S_n^2$  不是方差  $\sigma^2$  的无偏估计量



# 估计量评价准则——有效性

- 对于下图所示的两个无偏估计量  $\hat{\theta}_1$  和  $\hat{\theta}_2$ ，你会选择用那个呢？



- 定义：** 设  $\hat{\theta}_1(X_1, \dots, X_n)$  和  $\hat{\theta}_2(X_1, \dots, X_n)$  是参数  $\theta$  的两个无偏估计量，若

$$Var_{\theta}(\hat{\theta}_1) \leq Var_{\theta}(\hat{\theta}_2), \quad \forall \theta \in \Theta$$

且至少存在一个  $\theta \in \Theta$  使得不等式成立，则称  $\hat{\theta}_1$  比  $\hat{\theta}_2$  有效

- 根据方差的定义

$$\text{Var}_{\theta}(\hat{\theta}) \triangleq E \left[ (\hat{\theta} - E[\hat{\theta}])^2 \right],$$

当估计量是无偏估计量时，有 $E[\hat{\theta}] = \theta$ ，此时可以知道

$$\text{Var}_{\theta}(\hat{\theta}) = E \left[ (\hat{\theta} - \theta)^2 \right]$$

- 因此，当估计量是无偏估计量时， $\text{Var}_{\theta}(\hat{\theta})$ 反映了估计值 $\hat{\theta}$ 与真实值 $\theta$ 间偏差的平方的平均值
  - 但当估计量有偏时， $\text{Var}_{\theta}(\hat{\theta}) \neq E \left[ (\hat{\theta} - \theta)^2 \right]$ ，其不再能反映平均偏差的大小

# 估计量评价准则——均方误差

- 当 $\hat{\theta}$ 是有偏估计量时，可以直接使用均方误差来反映估计值与真实值间的平均偏差，即：

$$MSE_{\theta}(\hat{\theta}) \triangleq E[(\hat{\theta} - \theta)^2]$$

MSE stands for mean squared error（均方误差）

- 均方误差 $MSE_{\theta}(\hat{\theta})$ 与方差 $Var_{\theta}(\hat{\theta})$ 有何关系？

$$\begin{aligned} MSE_{\theta}(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\ &= E\left[\left((\hat{\theta} - E[\hat{\theta}]) - (\theta - E[\hat{\theta}])\right)^2\right] \\ &= E\left[(\hat{\theta} - E[\hat{\theta}])^2 - 2(\hat{\theta} - E[\hat{\theta}])(\theta - E[\hat{\theta}]) + (\theta - E[\hat{\theta}])^2\right] \\ &= E[(\hat{\theta} - E[\hat{\theta}])^2] + (\theta - E[\hat{\theta}])^2 = Var_{\theta}(\hat{\theta}) + (\theta - E[\hat{\theta}])^2 \end{aligned}$$

# 估计量评价准则——相合性 (Consistency)

- 定义：设  $X_i \sim \{f(x; \theta), \theta \in \Theta\}$ ，对每个自然数  $n$ ， $\hat{\theta}_n = \hat{\theta}(X_1, X_2, \dots, X_n)$  是  $\theta$  的一个估计量，若对任意  $\epsilon > 0$ ，都有

$$P(|\hat{\theta}_n - \theta| \geq \epsilon) \rightarrow 0, \quad n \rightarrow +\infty, \quad \forall \theta \in \Theta$$

或

$$\hat{\theta}_n \xrightarrow{a.s.} \theta, \quad n \rightarrow +\infty, \quad \forall \theta \in \Theta$$

则称  $\hat{\theta}_n$  是  $\theta$  的相合估计量

- 该性质考虑样本数  $n \rightarrow +\infty$  的情形，是大样本性质

# 估计量评价准则——渐近无偏性

- 定义：设对每个自然数 $n$ ， $\hat{\theta}_n = \hat{\theta}(X_1, X_2, \dots, X_n)$ 是 $\theta$ 的一个估计量，若偏差 $Bias_{\theta}(\hat{\theta}_n) \triangleq E[\hat{\theta}_n] - \theta$ 满足

$$Bias_{\theta}(\hat{\theta}_n) \rightarrow 0, \quad n \rightarrow +\infty,$$

则称 $\hat{\theta}_n$ 具有渐近无偏性

- 该性质考虑样本数 $n \rightarrow +\infty$ 的情形，也是一个大样本性质

# 提纲

---

- 参数估计及评价准则
- 矩估计
- 最大似然估计
- 最小方差无偏估计
- Cramer-Rao不等式
- 贝叶斯估计

# 总体的 $k$ 阶矩

- 设随机变量 $X \sim \{f(x; \theta), \theta \in \Theta\}$ , 其中 $\Theta$ 是参数空间, 称

$$\mu_k = E[X^k] = \int_{-\infty}^{+\infty} x^k f(x; \theta) dx$$

为总体的 $k$ 阶原点矩, 特别的 $\mu_1 = E[X]$ 为总体均值; 称

$$\alpha_k = E[(X - E[X])^k] = \int_{-\infty}^{+\infty} (x - E[X])^k f(x; \theta) dx$$

为总体的 $k$ 阶中心矩, 特别的 $\alpha_2 = Var(X)$ 为总体方差

# 样本矩

- 设 $X_1, X_2, \dots, X_n$ 是从总体 $F$ 中抽取的简单随机样本。总体的 $k$ 阶原点矩和中心矩可通过样本来估计

$$m_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

$$a_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$$

- 特别的，可以看出 $m_1 = \bar{X}$ ,  $a_2 = S_n^2$



# 矩估计量

- 设总体分布含有 $k$ 个参数 $\theta_1, \dots, \theta_k$ ，那么它的前 $k$ 阶原点矩（或中心矩）一般可以表示成这 $k$ 个参数的函数：

$$\mu_i = h_i(\theta_1, \theta_2, \dots, \theta_k), \quad i = 1, 2, \dots, s_1$$

$$\alpha_j = \tilde{h}_j(\theta_1, \theta_2, \dots, \theta_k), \quad j = 1, 2, \dots, s_2$$

- 如果能解出 $\theta_j$ ，将其表示成 $\mu_1, \dots, \mu_{s_1}$ 和 $\alpha_1, \dots, \alpha_{s_2}$ 的函数：

$$\theta_j = g_j(\mu_1, \mu_2, \dots, \mu_{s_1}, \alpha_1, \dots, \alpha_{s_2}), \quad j = 1, 2, \dots, k$$

然后将样本矩 $m_i, a_i$ 代替总体矩 $\mu_i, \alpha_i$ ，得到参数 $\theta_j$ 的矩估计

$$\hat{\theta}_j = g_j(m_1, m_2, \dots, m_{s_1}, a_1, \dots, a_{s_2}), \quad j = 1, 2, \dots, k$$

例：设 $X_1, X_2, \dots, X_n$ 是来自均匀分布 $X \sim U(a, b)$ 总体的一个样本. 求参数 $a$ 与 $b$ 的矩法估计.

解：由均匀分布的性质可知

$$\mu_1 = E[X] = \frac{a+b}{2}, \quad \sigma^2 = \text{Var}(X) = \frac{(b-a)^2}{12}$$

解方程组得

$$a = \mu_1 - \sqrt{3\sigma^2}, \quad b = \mu_1 + \sqrt{3\sigma^2}$$

用 $\bar{X}$ 代替 $\mu_1$ ,  $S_n^2$ 代替 $\sigma^2$ , 得 $a, b$ 的矩估计量为

$$\hat{a} = \bar{X} - \sqrt{3}S_n, \quad \hat{b} = \bar{X} + \sqrt{3}S_n$$

其中 $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$

例：设 $X_1, X_2, \dots, X_n$ 是来自二项分布 $X \sim \text{binomial}(k, p)$ 总体的一个样本. 求参数 $k$ 与 $p$ 的矩法估计.

解：由二项分布的性质可知

$$\mu_1 = E[X] = kp, \quad \sigma^2 = \text{Var}(X) = kp(1 - p)$$

解方程组得

$$p = \frac{(\mu_1 - \sigma^2)}{\mu_1}, \quad k = \frac{\mu_1^2}{\mu_1 - \sigma^2}$$

用 $\bar{X}$ 代替 $\mu_1$ ,  $S_n^2$ 代替 $\sigma^2$ , 得 $p, k$ 的矩估计量为

$$\hat{p} = \frac{(\bar{X} - S_n^2)}{\bar{X}}, \quad \hat{k} = \frac{\bar{X}^2}{\bar{X} - S_n^2}$$

例：设 $X_1, X_2, \dots, X_n$ 是来自指数分布 $X \sim E(\lambda)$ 总体的一个样本。  
求参数 $\lambda$ 的矩法估计。

解：由于 $X \sim E(\lambda)$ ，可以知道 $E[X] = \frac{1}{\lambda}$

因此，矩估计量可表示为

$$\hat{\lambda} = \frac{1}{\bar{X}} \quad \text{该估计量是否无偏?}$$

由于指数分布是Gamma分布的一个特例，即： $E(\lambda) = \Gamma(1, \lambda)$ 。

因此，由Gamma分布性质可知 $\sum_{i=1}^n X_i = \text{Gamma}(n, \lambda)$ ，从而

$$E[\hat{\lambda}] = E\left[\frac{n}{\sum_{i=1}^n X_i}\right] = \frac{n}{n-1} \lambda \quad \text{有偏}$$

➤ 怎样获得无偏估计量？  $\hat{\lambda} = \frac{n-1}{n} \frac{1}{\bar{X}}$

# 矩估计的相合性

- 根据大数定律，设 $X_1, X_2, \dots, X_n$ 是一系列独立同分布的随机变量序列，若其数学期望 $\mu$ 有限，则对任意给定的 $\epsilon > 0$ ，有

$$\frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow{a.s.} E[X^k] \text{ as } n \rightarrow +\infty$$

即：样本原点（中心） $k$ 阶矩会依概率收敛于 $k$ 阶总体原点（中心）矩

**定理：** 设 $\mathbf{X} = (X_1, X_2, \dots, X_n)$ 为从总体 $F$ 中抽取的简单随机样本，待估函数 $g(\theta) = G(\alpha_1, \dots, \alpha_k, \mu_1, \dots, \mu_s)$ ，其矩估计量为 $\hat{g} = G(a_1, \dots, a_k, m_1, \dots, m_s)$ ，且 $G$ 为连续函数，则 $\hat{g} = G(a_1, \dots, a_k, m_1, \dots, m_s)$ 为 $g(\theta)$ 的相合估计

**证：** 根据大数定律，有 $a_i \xrightarrow{a.s.} \alpha_i$ 和 $m_j \xrightarrow{a.s.} \mu_j$ ，因此，

$$G(a_1, \dots, a_k, m_1, \dots, m_s) \xrightarrow{a.s.} G(\alpha_1, \dots, \alpha_k, \mu_1, \dots, \mu_s)$$

# 矩估计的优缺点

---

- 优点

- 简单
- 满足相合性

- 缺点

- 在一般场合，矩估计量不唯一
- 在小样本场合，无突出性质（如：无偏、有效性）
- 没有充分利用已知参数分布族提供的信息

# 提纲

---

- 参数估计及评价准则
- 矩估计
- 最大似然估计
- 最小方差无偏估计
- Cramer-Rao不等式
- 贝叶斯估计



# 似然函数的定义

- **定义：** 设 $f(\mathbf{x}; \theta)$ 为样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)$ 的概率函数. 当 $\mathbf{x}$ 固定时, 把 $f(\mathbf{x}; \theta)$ 看成是 $\theta$ 的函数, 称其为似然函数, 记为

$$L(\theta, \mathbf{x}) = f(\mathbf{x}, \theta), \quad \theta \in \Theta$$

称 $\log L(\theta, \mathbf{x})$ 为**对数似然函数**, 记为 $\ell(\theta, \mathbf{x})$

- $f(\mathbf{x}, \theta)$ 分别被看作是 $\mathbf{x}$ 和 $\theta$ 的函数时表示的意义是不同的
  - 关于 $\mathbf{x}$ 的函数:  $f(\mathbf{x}, \theta)$ 是一个概率密度函数
  - 关于 $\theta$ 的函数:  $f(\mathbf{x}, \theta)$ 是一个关于 $\theta$ 的普通函数

**例：** 设罐子里有许多黑球和红球. 假定已知它们的比例是1:3，但不知道是黑球多还是红球多，即：抽出一个黑球的概率是1/4或3/4. 如果有放回的从罐子中抽 $n$ 个球，要根据抽样数据说明抽到黑球的概率是1/4，还是3/4. 记 $X_i$ 为第 $i$ 次抽球的结果， $X_i = 1$ 表示黑球、 $X_i = 0$ 表示红球、抽出黑球的概率为 $\theta$ 。记 $X = \sum_{i=1}^n X_i$ ，其服从二项分布 $X \sim b(n; \theta)$

**解：** 考虑 $n = 3$ 的情形. 这时如果 $x \leq 1$ ，**直觉上**，我们更倾向认为 $\theta = 1/4$ ；如果 $x \geq 2$ ，则我们会更倾向认为 $\theta = 3/4$

当 $X$ 给定时，似然函数

$$L(\theta, x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

当 $x = 0, 1, 2, 3$ 时, 可以计算出似然函数在 $\theta = 1/4$ 或 $3/4$ 时的取值分别为

$x$	0	1	2	3
$L(\theta_1, x)$	27/64	27/64	9/64	1/64
$L(\theta_2, x)$	1/64	9/64	27/64	27/64

$$\theta_1 = \frac{1}{4}, \theta_2 = \frac{3}{4}$$

可以看出,

$$\text{当 } x \leq 1 \text{ 时, } L(\theta_1, x) > L(\theta_2, x)$$

$$\text{当 } x \geq 2 \text{ 时, } L(\theta_1, x) < L(\theta_2, x)$$

根据似然函数值的大小, 当 $x \leq 1$ 时, 应该选择 $\theta = 1/4$ ; 而当 $x \geq 2$ 时, 应该选择 $\theta = 3/4$

这与我们前面的直觉选择一致

# 极大似然原理

---

- 若样本  $\mathbf{X} = (X_1, \dots, X_n)$  的分布族为  $\mathcal{F} = \{F_\theta, \theta \in \Theta\}$ 
  - 当样本  $\mathbf{x}$  给定时, 若  $\theta^*$  使似然函数  $L(\theta^*, \mathbf{x})$  为似然函数集合  $\{L(\theta, \mathbf{x}), \forall \theta \in \Theta\}$  中的最大者
  - 即:  $\theta^*$  的“似然性”比参数空间  $\Theta$  中任何其它参数值的“似然性”更大, 则取“似然性”最大的  $\theta^*$  作为  $\theta$  的估计值
  - 这一方法得到的参数  $\theta$  的估计, 称为“极大似然估计”

定义： 设  $X = (X_1, \dots, X_n)$  是从参数分布族  $\mathcal{F} = \{F_\theta, \theta \in \Theta\}$  中抽取的简单随机样本,  $L(\theta, \mathbf{x})$  是似然函数, 若存在统计量  $\theta^*(\mathbf{X}) = \theta^*(X_1, \dots, X_n)$ , 满足条件

$$L(\hat{\theta}^*, \mathbf{x}) = \sup_{\theta \in \Theta} L(\theta, \mathbf{x})$$

或

$$\hat{\theta}^* = \arg \max_{\theta \in \Theta} L(\theta, \mathbf{x})$$

则称  $\hat{\theta}^*(\mathbf{X})$  为  $\theta$  的极大似然估计

- 极大似然估计的求法

- 求解似然方程组

$$\frac{\partial \ell(\theta, \mathbf{x})}{\partial \theta_i} = 0, \quad i = 1, 2, \dots, k$$

其中 $\ell(\theta, \mathbf{x}) = \log L(\theta, \mathbf{x})$ 是对数似然函数，与 $L(\theta, \mathbf{x})$ 有相同的极值点

在有些情形下，该方程不一定有解，或者求出的 $\theta^* \notin \Theta$

- 通过分析的方法找到让 $\ell(\theta, \mathbf{x})$ 的值最大的 $\theta$ ，如函数单调性等性质

例：设 $X_1, X_2, \dots, X_n$ 是取自正态总体 $N(\mu, \sigma^2)$ 的一个样本，求未知参数 $\mu$ 和 $\sigma^2$ 的极大似然估计

解： 似然函数 
$$L(\mathbf{x}, \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\}$$

对数似然函数 
$$\ell(\mathbf{x}, \theta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

其中 $-\infty < \mu < +\infty, \sigma^2 > 0$

似然方程

$$\frac{\partial \ell(\mathbf{x}, \theta)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0$$

$$\frac{\partial \ell(\mathbf{x}, \theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0$$

求解上述似然方程，得到

$$\hat{\mu}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

$$\hat{\sigma}^2(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = s_n^2$$

可以验证 $\bar{x}, s_n^2$ 确实最大化对数似然函数

最大似然函数估计量

$$\hat{\mu}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$$

$$\hat{\sigma}^2(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = S_n^2$$

注意：这里 $S_n^2$ 除以 $n$



例：设 $X_1, X_2, \dots, X_n$ 是从两点分布族 $\{b(1, p): 0 < p < 1\}$ 中抽取的简单样本，求 $p$ 和 $g(p) = p(1 - p)$ 的极大似然估计

解：

$$L(\mathbf{x}, \theta) = p^{\sum_{i=1}^n x_i} (1 - p)^{n - \sum_{i=1}^n x_i}$$

$$\Rightarrow \ell(\mathbf{x}, \theta) = \left( \sum_{i=1}^n x_i \right) \log p + \left( n - \sum_{i=1}^n x_i \right) \log(1 - p)$$

$$\frac{\partial \ell(\mathbf{x}, \theta)}{\partial p} = \frac{1}{p} \left( \sum_{i=1}^n x_i \right) - \frac{1}{1 - p} \left( n - \sum_{i=1}^n x_i \right) = 0$$

$$\Rightarrow \hat{p} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \quad \Rightarrow \hat{g}^* = \hat{p}^*(1 - \hat{p}^*) = \bar{X}(1 - \bar{X})$$

例：设 $X_1, X_2, \dots, X_n$ 是从Poisson分布族 $\{P(\lambda): \lambda > 0\}$ 中抽取的简单样本，求 $\lambda$ 和 $g(\lambda) = e^{-\lambda}$ 的最大似然估计

解：

$$L(\mathbf{x}, \theta) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

$$\Rightarrow \ell(\mathbf{x}, \theta) = \left( \sum_{i=1}^n x_i \right) \log \lambda - n\lambda - \sum_{i=1}^n \log x_i!$$

$$\Rightarrow \frac{\partial \ell(\mathbf{x}, \theta)}{\partial \lambda} = \frac{1}{\lambda} \sum_{i=1}^n x_i - n$$

$$\Rightarrow \hat{\lambda}^* = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}, \quad \hat{g}^* = e^{-\bar{X}}$$

例：设 $X_1, X_2, \dots, X_n$ 是从均匀分布族 $\{U(0, \theta): \theta > 0\}$ 中抽取的简单样本，

1) 求 $\theta$ 的MLE

2) 说明 $\hat{\theta}^*$ 是否为 $\theta$ 的无偏估计. 若不然，作适当修正获得 $\theta$ 的无偏估计 $\hat{\theta}_1^*$

3) 试将 $\hat{\theta}_1^*$ 与 $\theta$ 的矩估计 $\hat{\theta}_1$ 进行比较，看哪一个有效？

4) 证明 $\theta$ 的极大似然估计 $\hat{\theta}^*$ 是 $\theta$ 的相合估计

解： 1) 似然函数

$$L(\mathbf{x}, \theta) = \begin{cases} \frac{1}{\theta^n}, & 0 < x_1, \dots, x_n < \theta \\ 0, & \text{其它} \end{cases}$$

这时，不能通过求解似然方程的方式来获得最优解 $\theta^*$ . 为使 $L(\mathbf{x}, \theta)$ 达到最大值，由上式可看出

$$\hat{\theta}^* = \max(X_1, X_2, \dots, X_n) = X_{(n)}$$

2) 为说明 $\hat{\theta}^*$ 是否无偏, 需要求得 $E[X_{(n)}]$ . 为此, 需要知道随机变量 $T = X_{(n)}$ 的概率分布. 根据次序统计量的分布结果, 可知

$$g(t, \theta) = \begin{cases} \frac{nt^{n-1}}{\theta^n}, & 0 < t < \theta \\ 0, & \text{其它} \end{cases}$$

因此, 可以知道

$$E[\hat{\theta}^*] = E[T] = \frac{n}{n+1}\theta$$

为获得无偏估计 $\hat{\theta}_1^*$ , 可对MLE进行如下修正

$$\hat{\theta}_1^* = \frac{n+1}{n}\hat{\theta}^* = \frac{n+1}{n}X_{(n)}$$

3) 均匀分布  $X \sim U(0, \theta)$  的期望  $E[X] = \frac{\theta}{2}$ . 因此,  $\theta$  的矩估计是

$$\hat{\theta}_1 = \frac{2}{n} \sum_{i=1}^n X_i = 2\bar{X}$$

$$\Rightarrow \text{Var}(\hat{\theta}_1) = \frac{4}{n^2} \times n \times \frac{\theta^2}{12} = \frac{\theta^2}{3n}$$

$$\begin{aligned} \text{Var}(\hat{\theta}_1^*) &= \frac{(n+1)^2}{n^2} \text{Var}(X_{(n)}) \\ &= \frac{(n+1)^2}{n^2} \frac{n}{(n+2)(n+1)^2} \theta^2 \\ &= \frac{1}{(n+2)n} \theta^2 \end{aligned}$$

可以看出, 当  $n \geq 2$  时,  $\hat{\theta}_1^*$  比  $\hat{\theta}_1$  更有效

4) 由 $T$ 的概率密度  $g(t, \theta)$ 可知, 对于任意 $\epsilon > 0$ , 有

$$\begin{aligned}P(|\hat{\theta}^* - \theta| \geq \epsilon) &= 1 - P(|\hat{\theta}^* - \theta| < \epsilon) \\&= 1 - P(\theta - \epsilon < T < \theta + \epsilon) \\&= 1 - P(\theta - \epsilon < T < \theta) \\&= 1 - \int_{\theta - \epsilon}^{\theta} \frac{nt^{n-1}}{\theta^n} dt \\&= 1 - \frac{1}{\theta^n} [\theta^n - (\theta - \epsilon)^n] \\&= \left(1 - \frac{\epsilon}{\theta}\right)^n\end{aligned}$$

因此,

$$\lim_{n \rightarrow +\infty} P(|\hat{\theta}^* - \theta| \geq \epsilon) = \lim_{n \rightarrow +\infty} \left(1 - \frac{\epsilon}{\theta}\right)^n = 0$$

因此, 可以看出 $\hat{\theta}^* = X_{(n)}$ 为 $\theta$ 的相合估计

# 极大似然估计的渐近正态性

**定理：** 设  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  为满足下面条件 (1) - (3) 的总体中抽取的简单随机样本，且设对数似然方程

$$\sum_{i=1}^n \frac{\partial \log f(x_i, \theta)}{\partial \theta} = 0$$

有唯一根  $\hat{\theta}^* = \hat{\theta}^*(X_1, \dots, X_n)$ ，则当  $n \rightarrow \infty$ ， $\hat{\theta}^*$  依概率收敛于真实值  $\theta$  且服从正态分布，即：

$$\hat{\theta}^* \sim N\left(\theta, \frac{1}{n \cdot I(\theta)}\right)$$

其中

$$I(\theta) = E \left[ \left( \frac{\partial \log p(X; \theta)}{\partial \theta} \right)^2 \right] \quad \text{Fisher 信息量}$$

条件(1)-(3):

(1) 对一切  $\theta \in \Theta$ ,  $f(x; \theta)$  对如下偏导数都存在

$$\frac{\partial \ln f(x; \theta)}{\partial \theta}, \quad \frac{\partial^2 \ln f(x; \theta)}{\partial^2 \theta}, \quad \frac{\partial^3 \ln f(x; \theta)}{\partial^3 \theta}$$

(2) 对一切  $\theta \in \Theta$ , 有

$$\left| \frac{\partial \ln f(x; \theta)}{\partial \theta} \right| < F_1(x), \quad \left| \frac{\partial^2 \ln f(x; \theta)}{\partial^2 \theta} \right| < F_2(x), \quad \left| \frac{\partial^3 \ln f(x; \theta)}{\partial^3 \theta} \right| < H(x)$$

其中  $F_1(x), F_2(x)$  在实轴上可积, 而  $H(x)$  满足  $\int_{-\infty}^{+\infty} H(x) f(x; \theta) dx < M$

(3) 对一切  $\theta \in \Theta$ , 有

$$-\infty < E \left[ \left( \frac{\partial \log p(X; \theta)}{\partial \theta} \right)^2 \right] < +\infty$$



例：设 $X_1, X_2, \dots, X_n$ 是从0-1分布 $b(1, \theta)$ 抽取的简单随机样本，求 $\theta$ 的最大似然估计及其渐近分布

解：

$$L(\mathbf{x}; \theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} \quad \text{其中 } x_i \in \{0, 1\}$$

$$\ell(\mathbf{x}; \theta) = \left( \sum_{i=1}^n x_i \right) \log \theta + \left( n - \sum_{i=1}^n x_i \right) \log(1 - \theta)$$

$$\hat{\theta}^* = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

$$\begin{aligned} I(\theta) &= E \left[ \left( \frac{\partial \log p(x; \theta)}{\partial \theta} \right)^2 \right] = E \left[ \left( \frac{x - \theta}{\theta(1 - \theta)} \right)^2 \right] \\ &= \frac{\theta(1 - \theta)}{(\theta(1 - \theta))^2} = \frac{1}{\theta(1 - \theta)} \end{aligned} \quad \longrightarrow \quad \hat{\theta}^* \sim N \left( \theta, \frac{\theta(1 - \theta)}{n} \right)$$

例：设 $X_1, X_2, \dots, X_n$ 是从正态分布 $N(\mu, \sigma_0^2)$ 抽取的简单随机样本， $\sigma_0^2$ 是已知值，求 $\mu$ 的fisher信息量，并给出其MLE的渐近分布

解：已知 $\mu$ 的MLE为

$$\hat{\mu}^* = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$$

$$I(\theta) = E \left[ \left( \frac{\partial \log p(x; \theta)}{\partial \theta} \right)^2 \right] = E \left[ \left( \frac{x - \mu}{\sigma_0^2} \right)^2 \right] = \frac{\sigma_0^2}{\sigma_0^4} = \frac{1}{\sigma_0^2}$$

因此，MLE  $\hat{\mu}^*$ 的渐近分布为 $\hat{\mu}^* \sim N\left(\mu, \frac{\sigma_0^2}{n}\right)$

通过正态分布的性质也可以得到 $\bar{X} \sim N\left(\mu, \frac{\sigma_0^2}{n}\right)$ 。可看出，这两种方法得到的结果一致

# 提纲

---

- 参数估计及评价准则
- 矩估计
- 最大似然估计
- 最小方差无偏估计
- Cramer-Rao不等式
- 贝叶斯估计

# 一致最小均方误差估计

- 设有一参数分布族  $F = \{F_\theta, \theta \in \Theta\}$ , 其中  $\Theta$  为参数空间.  $\mathbf{X} = (X_1, \dots, X_n)$  为从总体  $F_\theta$  中抽取的简单样本,  $\hat{\theta}(\mathbf{X}) = \hat{\theta}(X_1, \dots, X_n)$  为  $\theta$  的一个估计量, 如何评价  $\hat{\theta}(\mathbf{X})$  的优劣?
- 评价估计值优劣的一个合理指标是估计值与真实值间的距离

$$E \left[ (\hat{\theta}(\mathbf{X}) - \theta)^2 \right]$$

- 由于  $\hat{\theta}(\mathbf{X})$  是一个随机变量, 所以采用差距平方的期望来度量
- $E \left[ (\hat{\theta}(\mathbf{X}) - \theta)^2 \right]$  也称为均方误差 (Mean-Squared Error, MSE)

- 设有两个估计量 $\hat{\theta}_1(\mathbf{X})$ 和 $\hat{\theta}_2(\mathbf{X})$ ，若

$$E \left[ (\hat{\theta}_1(\mathbf{X}) - \theta)^2 \right] \leq E \left[ (\hat{\theta}_2(\mathbf{X}) - \theta)^2 \right]$$

对一切 $\theta \in \Theta$ 都成立，且不等号至少对某个 $\theta \in \Theta$ 成立，则称在MSE准则下 $\hat{\theta}_1(\mathbf{X})$ 优于 $\hat{\theta}_2(\mathbf{X})$

与有效性什么关系？

定义：若存在估计量 $\hat{\theta}^*(\mathbf{X})$ ，使得对 $\theta$ 的任一估计量 $\hat{\theta}(\mathbf{X})$ ，都有

$$E \left[ (\hat{\theta}^*(\mathbf{X}) - \theta)^2 \right] \leq E \left[ (\hat{\theta}(\mathbf{X}) - \theta)^2 \right]$$

对一切 $\theta \in \Theta$ 都成立，则称 $\hat{\theta}^*(\mathbf{X})$ 为 $\theta$ 的一致最小均方误差估计

- 均方误差与方差的关系

$$\begin{aligned}MSE_{\theta}(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\&= E\left[\left((\hat{\theta} - E[\hat{\theta}]) - (\theta - E[\hat{\theta}])\right)^2\right] \\&= E\left[(\hat{\theta} - E[\hat{\theta}])^2 - 2(\hat{\theta} - E[\hat{\theta}])(\theta - E[\hat{\theta}]) + (\theta - E[\hat{\theta}])^2\right] \\&= E[(\hat{\theta} - E[\hat{\theta}])^2] + (\theta - E[\hat{\theta}])^2 \\&= Var_{\theta}(\hat{\theta}) + (\theta - E[\hat{\theta}])^2\end{aligned}$$

➤ 当估计量是无偏的时 ( $E[\hat{\theta}] = \theta$ ), 有

$$MSE_{\theta}(\hat{\theta}) = Var_{\theta}(\hat{\theta})$$

例：设 $X_1, X_2, \dots, X_n$ 是从正态分布 $N(\mu, \sigma^2)$ 抽取的简单随机样本，

$S_n^{*2} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 和 $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ 都是 $\sigma^2$ 的估计.

前者是无偏估计. 计算它们的均方误差

解：由于 $\frac{(n-1)S_n^{*2}}{\sigma^2} \sim \chi^2(n-1)$ , 因此,

$$\begin{aligned} \text{Var}(S_n^{*2}) &= \frac{\sigma^4}{(n-1)^2} \cdot \text{Var}(\chi^2(n-1)) \\ &= \frac{\sigma^4}{(n-1)^2} \cdot 2(n-1) = \frac{2\sigma^4}{n-1} \end{aligned}$$

由于 $S_n^{*2}$ 无偏, 因此,

$$\text{MSE}(S_n^{*2}) = \text{Var}(S_n^{*2}) = \frac{2\sigma^4}{n-1}$$

另一方面，由于  $S_n^2 = \frac{n-1}{n} S_n^{*2}$ ，可以知道

$$\text{Var}(S_n^2) = \frac{(n-1)^2}{n^2} \cdot \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2}$$

$$E[S_n^2] = \frac{n-1}{n} \sigma^2$$

$$\text{MSE}(S_n^2) = \frac{2(n-1)\sigma^4}{n^2} + \left( \frac{n-1}{n} \sigma^2 - \sigma^2 \right)^2 = \frac{2n-1}{n^2} \sigma^4$$

容易看出

$$\text{MSE}(S_n^2) < \text{MSE}(S_n^{*2})$$

有偏估计量的MSE可能比无偏估计量的更好



- 最小均方误差估计量经常都不存在
  - 为寻找“最优”的估计量，我们将寻找的范围缩小一些
    - 原先要求从所有的估计量中寻找一个估计量 $\hat{\theta}^*$ ，使得它对所有的 $\theta \in \Theta$ 值都有MSE最小，
    - 现在要求从所有无偏的估计量中寻找一个估计量 $\hat{\theta}^*$ ，使得它对所有的 $\theta \in \Theta$ 值都有MSE最小，
- 在无偏的前提下，我们有

$$MSE_{\theta}(\hat{\theta}) = Var_{\theta}(\hat{\theta})$$

# 一致最小方差无偏估计定义

定义：设有一参数分布族  $F = \{F_\theta, \theta \in \Theta\}$ ，其中  $\Theta$  为参数空间.  $\mathbf{X} = (X_1, \dots, X_n)$  为自总体  $F_\theta$  中抽取的简单样本,  $\hat{\theta}^*(\mathbf{X})$  为  $\theta$  的一个无偏估计量，若对  $\theta$  的任一无偏估计量  $\hat{\theta}(\mathbf{X})$ ，都有

$$\text{Var}_\theta \left( \hat{\theta}^*(\mathbf{X}) \right) \leq \text{Var}_\theta \left( \hat{\theta}(\mathbf{X}) \right)$$

对一切  $\theta \in \Theta$  都成立，则称  $\hat{\theta}^*(\mathbf{X})$  为  $\theta$  的一致最小方差无偏估计 (Uniform minimum variance unbiased estimator, UMVUE)

问题：如何找到UMVUE？

定理 (Rao-Blackwell): 设  $T = T(\mathbf{X})$  是参数  $\theta \in \Theta$  的充分统计量,  $\varphi(\mathbf{X})$  是参数  $\theta$  的一个无偏估计, 那么

$$g(T) = E[\varphi(\mathbf{X})|T]$$

是  $\theta$  的无偏估计, 并且

$$\text{Var}_{\theta}(g(T)) \leq \text{Var}_{\theta}(\varphi(\mathbf{X})), \quad \forall \theta \in \Theta$$

其中等号当且仅当

$$P(\varphi(\mathbf{X}) = g(T)) = 1, \quad \forall \theta \in \Theta$$

■ 定理表明:

- 1) 改善现有无偏估计量  $\varphi(\mathbf{X})$  的方法——求其关于某充分统计量的条件期望  $E[\varphi(\mathbf{X})|T]$
- 2) UMVUE 一定是充分统计量的函数 Why?

证明:  $\mathbf{X}$ 和充分统计量 $T$ 共同构成的联合分布如下所示

$$p_{\theta}(\mathbf{x}, t) = p_{\theta}(\mathbf{x})I(t = T(\mathbf{x}))$$

因为 $T$ 是充分统计量, 在给定 $T$ 的条件下 $\mathbf{X}$ 的分布 $p(\mathbf{x}|T)$ 与参数 $\theta$ 无关. 因此, 可以知道

$$\underbrace{E[\varphi(\mathbf{X})|T = t]}_{g(T)} = \int \varphi(\mathbf{x})p_{\theta}(\mathbf{x}|T = t)d\mathbf{x}$$

与参数 $\theta$ 无关

$$\begin{aligned} E[g(T)] &= \int \int \varphi(\mathbf{x})p_{\theta}(\mathbf{x}|T = t)p_{\theta}(t)d\mathbf{x} dt \\ &= \int \varphi(\mathbf{x})p_{\theta}(\mathbf{x}) d\mathbf{x} = \theta \end{aligned}$$

因此,  $g(T)$ 也是 $\theta$ 的无偏估计量

$$\begin{aligned}
\text{Var}_\theta(\varphi(\mathbf{X})) &= E_X[(\varphi - \theta)^2] \\
&= E_{X,T}[(\varphi - E_X[\varphi|T] + E_X[\varphi|T] - \theta)^2] \\
&= E_{X,T}[(\varphi - E_X[\varphi|T])^2] + \underbrace{E_T[(E_X[\varphi|T] - \theta)^2]}_{\text{Var}_\theta(g(T))} \\
&\quad + 2E_{X,T}[(\varphi - E_X[\varphi|T])(E_X[\varphi|T] - \theta)]
\end{aligned}$$

另一方面，可以证明  $E_{X,T}[(\varphi - E_X[\varphi|T])(E_X[\varphi|T] - \theta)] = 0$

$$\begin{aligned}
E_{X,T}[(\varphi - E_X[\varphi|T])(E_X[\varphi|T] - \theta)] &= E_T \left[ E_{X|T}[(\varphi - E_X[\varphi|T])(E_X[\varphi|T] - \theta)] \right] \\
&= E_T[(E_X[\varphi|T] - E_X[\varphi|T])(E_X[\varphi|T] - \theta)] \\
&= 0
\end{aligned}$$

因此，可以得到

$$\begin{aligned} \text{Var}_\theta(\varphi(\mathbf{X})) &= E_{X,T}[(\varphi - E[\varphi|T])^2] + \text{Var}_\theta(g(T)) \\ &\geq \text{Var}_\theta(g(T)) \end{aligned}$$

等号成立当且仅当

$$E_{X,T}[(\varphi - E[\varphi|T])^2] = 0$$

即：

$$P(\varphi(\mathbf{X}) = g(T)) = 1$$

例：设  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  是从 0-1 分布族  $\{b(1, p): 0 < p < 1\}$  中抽取的简单随机样本. 显然,  $X_1$  是  $p$  的一个无偏估计量,  $T(\mathbf{X}) = \sum_{i=1}^n X_i$  是  $p$  的充分统计量, 试利用  $T = T(\mathbf{X})$  构造一个比  $X_1$  方差更小的无偏估计

解：根据上述定理, 构造一个无偏估计器

$$\begin{aligned} g(T = t) &= E[X_1 | T = t] \\ &= 1 \cdot P(X_1 = 1 | T = t) + 0 \cdot P(X_1 = 0 | T = t) \\ &= \frac{P(X_1 = 1, T = t)}{P(T = t)} = \frac{P(X_1 = 1, \sum_{i=2}^n X_i = t - 1)}{P(T = t)} \\ &= \frac{p \cdot \binom{n-1}{t-1} p^{t-1} (1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{t}{n} = \bar{x} \end{aligned}$$

可以验证  $g(T) = \bar{X}$  是  $p$  的无偏估计. 同时,  $g(T) = \bar{X}$  的方差为  $p(1-p)/n$ , 而  $X_1$  的方差为  $p(1-p)$ . 当  $n \geq 2$  时,  $\bar{X}$  的方差更小

- **问题：**通过上述定理可以获得一个方差更小的无偏估计量，但改进后的估计量是否为UMVUE呢？

**定理：** 设 $\hat{g}(\mathbf{X})$ 是 $\theta$ 的一个无偏估计且 $D_{\theta}(\hat{g}(\mathbf{X})) < \infty \quad \forall \theta \in \Theta$ 。

若对任何满足条件 “ $E_{\theta}[\ell(\mathbf{X})] = 0, \forall \theta \in \Theta$ ” 的统计量 $\ell(\mathbf{X})$ 都有

$$\text{Cov}_{\theta}(\hat{g}(\mathbf{X}), \ell(\mathbf{X})) = 0, \quad \forall \theta \in \Theta$$

成立，那么 $\hat{g}(\mathbf{X})$ 是 $\theta$ 的UMVUE

➤  $\text{Cov}_{\theta}(\hat{g}(\mathbf{X}), \ell(\mathbf{X})) = 0$ 等价于  $E_{\theta}[\hat{g}(\mathbf{X}) \cdot \ell(\mathbf{X})] = 0$

➤ “ $E_{\theta}[\ell(\mathbf{X})] = 0, \forall \theta \in \Theta$ ” 可理解为 “ $\ell(\mathbf{X})$ 在所有 $\forall \theta \in \Theta$ 上都是0的无偏估计”。因此  $\ell(\mathbf{X})$ 表示所有零无偏估计的集合（ $\forall \theta \in \Theta$ ）



- 定理直白理解

- 如果一个无偏估计去  $\hat{g}(\mathbf{X})$  与所有的零无偏估计器的协方差都为0，那么，可以断定  $\hat{g}(\mathbf{X})$  是UMVUE
- 零无偏估计器  $\ell(\mathbf{X})$  是要求对所有的  $\theta \in \Theta$ ，其期望都等于0，即：  $E_{\theta}[\ell(\mathbf{X})] = 0, \forall \theta \in \Theta$

- 在实际使用的时候，可以先找到零无偏估计器的集合  $\{\ell(\mathbf{X})\}$ ，然后判断  $\hat{g}(\mathbf{X})$  与该集合中的每一个元素的协方差是否为0

**证明：** 设 $\hat{g}_1(\mathbf{X})$ 为 $\theta$ 的任一无偏估计。记 $\ell(\mathbf{X}) = \hat{g}_1(\mathbf{X}) - \hat{g}(\mathbf{X})$ ，可以看出 $E[\ell(\mathbf{X})] = 0, \forall \theta \in \Theta$ ，因而 $\ell(\mathbf{X})$ 是零的无偏估计。因此，对于任一无偏估计器 $\hat{g}_1(\mathbf{X})$ ，都可将其表示成 $\hat{g}_1(\mathbf{X}) = \hat{g}(\mathbf{X}) + \ell(\mathbf{X})$ 。因此，我们有

$$\begin{aligned} D_{\theta}(\hat{g}_1(\mathbf{X})) &= D_{\theta}(\hat{g}(\mathbf{X}) + \ell(\mathbf{X})) \\ &= D_{\theta}(\hat{g}(\mathbf{X})) + D_{\theta}(\ell(\mathbf{X})) + 2Cov_{\theta}(\hat{g}(\mathbf{X}), \ell(\mathbf{X})) \end{aligned}$$

由于对任一 $\ell(\mathbf{X})$ ，结论 $Cov_{\theta}(\hat{g}(\mathbf{X}), \ell(\mathbf{X})) = 0, \forall \theta \in \Theta$ 恒成立。因此，

$$\begin{aligned} D_{\theta}(\hat{g}_1(\mathbf{X})) &= D_{\theta}(\hat{g}(\mathbf{X})) + D_{\theta}(\ell(\mathbf{X})) \\ &\geq D_{\theta}(\hat{g}(\mathbf{X})), \quad \forall \theta \in \Theta \end{aligned}$$

推论： 设  $T = T(\mathbf{X})$  是一个充分统计量，  $\delta(T)$  是  $\theta$  的一个无偏估计，  
 $D_\theta(\delta(T)) < \infty \quad \forall \theta \in \Theta$ 。若对任何满足条件 “ $E_\theta[\ell(T)] = 0,$   
 $\forall \theta \in \Theta$ ” 的统计量  $\ell(T)$  都有

$$\text{Cov}_\theta(\delta(T), \ell(T)) = 0, \quad \forall \theta \in \Theta$$

成立，那么  $g(T)$  是  $\theta$  的UMVUE

➤  $\text{Cov}_\theta(\delta(T), \ell(T)) = 0$  等价于  $E_\theta[\delta(T) \cdot \ell(T)] = 0$

例：证明上个例子中得到的估计量  $g(T) = \frac{T}{n}$  ( $T = \sum_{i=1}^n X_i$ ) 是UMVUE

解：由前例已知  $T = \sum_{i=1}^n X_i$  是充分统计量， $g(T) = \frac{T}{n}$  是参数  $p$  的无偏估计。显然， $g(T)$  的方差有限。因此，为证明  $g(T)$  是UMVUE，只需要证明

$$E_{\theta}[g(T) \cdot \ell(T)] = 0, \quad \forall \theta \in \Theta$$

其中  $\ell(T)$  为任一零无偏估计量

由于  $T = \sum_{i=1}^n X_i$ ，可知  $T \sim b(n, p)$ 。因此，若记  $a_i = \ell(i)$ ，由  $\ell(T)$  为任一零无偏估计量可知  $a_i$  需满足如下条件

$$\begin{aligned} \sum_{i=0}^n a_i \binom{n}{i} p^i (1-p)^{n-i} &= 0, \quad \forall 0 < p < 1 \\ \Leftrightarrow \sum_{i=0}^n a_i \binom{n}{i} \theta^i &= 0, \quad \forall 0 < \theta < +\infty \quad \theta = p/(1-p) \end{aligned}$$

由于  $\sum_{i=0}^n a_i \binom{n}{i} \theta^i$  是关于  $\theta$  的多项式, 当系数  $a_i \neq 0$  时,  $\sum_{i=0}^n a_i \binom{n}{i} \theta^i = 0$  最多只有  $n$  个实数根. 为了使得  $\sum_{i=0}^n a_i \binom{n}{i} \theta^i = 0, \forall 0 < \theta < +\infty$  成立, 必然要求

$$a_i = 0, \quad i = 0, 1, \dots, n$$

由于  $a_i = 0$ , 可以容易得到

$$\begin{aligned} E_{\theta}[g(T) \cdot \ell(T)] &= \sum_{t=0}^n g(t) \cdot a_t \binom{n}{t} p^t (1-p)^{n-t} \\ &= 0 \end{aligned}$$

因此,  $g(T) = \bar{X}$  是UMVUE

例：设  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  为从均匀分布  $U(0, \theta)$  中抽取的简单样本，求  $\theta$  的UMVUE

解：由前面例子已经知道  $T = X_{(n)}$  是参数  $\theta$  的充分统计量，而且已经证明  $g(T) = \frac{n+1}{n}T$  是  $\theta$  的无偏估计

容易看出  $g(T)$  的方差有限. 因此，为证明  $g(T)$  是UMVUE，只需要证明

$$E_{\theta}[g(T) \cdot \ell(T)] = 0, \quad \forall \theta \in \Theta$$

其中  $\ell(T)$  为任一零无偏估计量  $\forall \theta \in \Theta$ . 若记  $\delta(t) = \ell(t)$ ，由  $\ell(T)$  为任一零无偏估计量可知  $\delta(t)$  需满足如下条件

$$\begin{aligned} E_{\theta}[\ell(T)] &= \int_0^{\theta} \delta(t) \cdot \frac{nt^{n-1}}{\theta^n} dt = 0, \quad \forall \theta > 0 \\ \Rightarrow \int_0^{\theta} \delta(t) \cdot t^{n-1} dt &= 0, \quad \forall \theta > 0 \end{aligned}$$

将上式两边对 $\theta$ 求导得

$$\delta(\theta) \cdot \theta^{n-1} = 0, \quad \forall \theta > 0$$

如果上式成立，那么一定有

$$\delta(\theta) \equiv 0, \quad \forall \theta > 0$$

由于 $\delta(\theta) \equiv 0, \forall \theta > 0$ ，容易看出

$$E_{\theta}[g(T) \cdot \ell(T)] = 0$$

因此， $g(T) = \frac{n+1}{n}T$ 是UMVUE

定理 (Lehmann-Scheffe): 设 $T(\mathbf{X})$ 为一个充分完全统计量, 若 $g(T)$ 为 $\theta$ 的一个无偏估计, 则 $g(T)$ 是 $\theta$ 的唯一的UMVUE

证明: 唯一性: 假设 $g_1(T(\mathbf{X}))$ 为 $\theta$ 的任一无偏估计. 令 $\delta(T(\mathbf{X})) = g(T) - g_1(T)$ , 则容易看出

$$E[\delta(T)] = E[g(T) - g_1(T)] = 0$$

由于 $T(\mathbf{X})$ 是完全统计量, 可以知道如果 $E[g(T) - g_1(T)] = 0$ , 那么

$$g(T) = g_1(T)$$

因此, 以统计量 $T(\mathbf{X})$ 为输入的无偏估计量就只有一个.



**一致最小方差性：** 假设 $\varphi(\mathbf{X})$ 为 $\theta$ 的任一无偏估计. 令 $h(T(\mathbf{X})) = E[\varphi(\mathbf{X})|T]$ , 由于 $T(\mathbf{X})$ 为充分统计量, 故知 $h(T(\mathbf{X}))$ 与 $\theta$ 无关, 是统计量. 而且, 由Rao-Blackwell定理可知

$$E[h(T(\mathbf{X}))] = \theta$$

$$\text{Var}(h(T(\mathbf{X}))) \leq \text{Var}(\varphi(\mathbf{X}))$$

由于 $E[h(T(\mathbf{X}))] = \theta$ , 结合前面证得的唯一性可知 $g(T(\mathbf{X})) = h(T(\mathbf{X}))$

由于 $\text{Var}(h(T(\mathbf{X}))) \leq \text{Var}(\varphi(\mathbf{X}))$ , 因此可以得到

$$\text{Var}(g(T(\mathbf{X}))) \leq \text{Var}(\varphi(\mathbf{X})), \quad \forall \theta \in \Theta$$

所以,  $g(T(\mathbf{X}))$ 的方差最小.

- 由该定理的证明过程可知：

若 $T(\mathbf{X})$ 为一个充分完全统计量， $\varphi(\mathbf{X})$ 是 $\theta$ 的一个无偏估计量，那么 $\theta$ 唯一的UMVUE估计量就是

$$E[\varphi(\mathbf{X})|T]$$

推论： 设样本 $\mathbf{X} = (X_1, X_2, \dots, X_n)$ 的分布为指数族

$$f(\mathbf{x}, \boldsymbol{\theta}) = C(\boldsymbol{\theta}) \exp \left\{ \sum_{j=1}^k \theta_j T_j(\mathbf{x}) \right\} h(\mathbf{x}), \quad \boldsymbol{\theta} \in \Theta^*$$

令 $T(\mathbf{X}) = (T_1(\mathbf{X}), \dots, T_k(\mathbf{X}))$ ，若自然参数空间 $\Theta^*$ 作为 $R_k$ 的子集有内点，且 $h(T(\mathbf{X}))$ 为 $\boldsymbol{\theta}$ 的无偏估计，那么 $h(T(\mathbf{X}))$ 为 $\boldsymbol{\theta}$ 的唯一的UMVUE

证明： 由指数族的性质可知 $T(\mathbf{X})$ 为充分完全统计量。根据条件 $h(T(\mathbf{X}))$ 为 $\boldsymbol{\theta}$ 的无偏估计，由Lehmann-Scheffe定理可知， $h(T(\mathbf{X}))$ 为 $\boldsymbol{\theta}$ 唯一的UMVUE

例：设 $\mathbf{X} = (X_1, X_2, \dots, X_n)$ 是从0-1分布族 $\{b(1, p): 0 < p < 1\}$ 中抽取的简单随机样本. 证明 $p$ 的无偏估计 $h(T) = \frac{T}{n} = \bar{X}$ 为 $p$ 的UMVUE

证明：由因子分解定理可知 $T = T(\mathbf{X}) = \sum_{i=1}^n X_i$ 为两点分布 $b(1, p)$ 中参数 $p$ 的充分统计量. 第二章已经证明 $T$ 是完全统计量，故 $T$ 是完全充分统计量

另外，容易看出

$$E[h(T)] = p$$

因此，根据L-S定理可知 $h(T) = \bar{X}$ 为 $p$ 唯一的UMVUE

例：设  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  是从指数分布  $Exp(\lambda)$  中抽取的简单随机样本.  
求参数  $\lambda$  的UMVUE

解：由因子分解定理可知  $T = T(\mathbf{X}) = \sum_{i=1}^n X_i$  为指数分布中参数  $\lambda$  的充分统计量. 第二章已经证明  $T$  是完全统计量，故  $T$  是完全充分统计量  
另外，由Gamma分布的性质可知

$$T(\mathbf{X}) \sim \Gamma(n, \lambda)$$

可以知道

$$E\left[\frac{1}{T}\right] = \int_0^{+\infty} \frac{1}{t} \cdot \frac{\lambda^n}{\Gamma(n)} t^{n-1} e^{-\lambda t} dt = \frac{\lambda}{n-1}$$

因此，

$$h(T) = \frac{n-1}{T}$$

是  $\lambda$  的无偏估计. 由L-S定理可知， $h(T)$  是  $\lambda$  的UMVUE

例：设 $\mathbf{X} = (X_1, X_2, \dots, X_n)$ 是从正态分布 $N(a, \sigma^2)$ 中抽取的简单随机样本，记 $\theta = (a, \sigma^2)$ 。求 $a$ 和 $\sigma^2$ 的UMVUE

解：第二章已经表明 $T = (T_1, T_2)$ 为充分完全统计量，其中 $T_1 = T_1(\mathbf{X}) = \bar{X}$ ， $T_2 = T_2(\mathbf{X}) = \sum_{i=1}^n (X_i - \bar{X})^2$

由于 $h_1(T) = \bar{X}$ 和 $h_2(T) = T_2/(n-1)$ 分别为 $a$ 和 $\sigma^2$ 的无偏估计，它们又是充分完全统计量的函数，故由L-S定理可知它们分别是 $a$ 和 $\sigma^2$ 的UMVUE

结论： $\bar{X}$ 和 $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 是整体分布参数均值和方差的最优无偏估计器

例：设 $\mathbf{X} = (X_1, X_2, \dots, X_n)$ 是从均匀分布 $U(0, \theta)$ 中抽取的简单随机样本，求参数 $\theta$ 的UMVUE

解：前面已经证明 $T = T(\mathbf{X}) = \max(X_1, X_2, \dots, X_n) = X_{(n)}$ 为充分完全统计量

前面例子也已经表明

$$h(T) = (n + 1)T/n$$

为参数 $\theta$ 的无偏估计，故由L-S定理可知 $h(T)$ 为参数 $\theta$ 的UMVUE

# 提纲

---

- 参数估计及评价准则
- 矩估计
- 最大似然估计
- 最小方差无偏估计
- Cramer-Rao不等式
- 贝叶斯估计



- 这一方法的思想如下：设 $\mathcal{U}_\theta$ 是 $\theta$ 的一切无偏估计构成的类。 $\mathcal{U}_\theta$ 中的估计量的方差有一个下界，这个下界称为Cramer-Rao下界（简称C-R下界）
- 因此，若 $\theta$ 的一个无偏估计 $\hat{\theta}$ 的方差达到此下界，则 $\hat{\theta}$ 就是 $\theta$ 的一个UMVUE
- 但在一些场合，虽然 $\theta$ 的UMVUE  $\hat{\theta}$ 存在，但其方差大于C-R 下界。在这情况下，用C-R不等式就无法判定 $\theta$ 的UMVUE存在。因此，这一方法的适用范围不广
- C-R 不等式除了用于判断 $\theta$ 的UMVUE之外，它在数理统计理论上还有其它的用处，如估计量的效率和有效估计的概念以及Fisher信息量都与之有关。

# C-R正则条件

■ 若单参数概率函数族  $\mathcal{F} = \{f(x, \theta), \theta \in \Theta\}$  满足:

- (1) 参数空间  $\Theta$  是直线上的某个开区间;
- (2)  $\forall x \in \mathcal{X}$  及  $\theta \in \Theta$ ,  $f(x, \theta) > 0$ , 即分布族具有共同支撑集;
- (3)  $\forall x \in \mathcal{X}$  及  $\theta \in \Theta$ ,  $\frac{\partial f(x, \theta)}{\partial \theta}$  存在;
- (4) 概率函数  $f(x, \theta)$  的积分与微分运算可交换, 即

$$\frac{\partial}{\partial \theta} \int f(x, \theta) dx = \int \frac{\partial}{\partial \theta} f(x, \theta) dx,$$

若 r.v.  $X$  分布为离散型, 上式为无穷级数和微分运算可交换;

- (5) 下列数学期望存在, 且  $0 < I(\theta) = E_{\theta} \left[ \frac{\partial \log f(X, \theta)}{\partial \theta} \right]^2 < \infty$ ,

则称该分布族为C-R正则分布族. 条件(1)-(5)称为C-R正则条件, 其中 $I(\theta)$ 称为该分布的Fisher信息量

# C-R不等式

**定理：** 设 $\mathcal{F} = \{f(x, \theta), \theta \in \Theta\}$  是C-R正则分布族,  $\theta$ 是 $\Theta$ 上的可微函数. 设 $X = (X_1, \dots, X_n)$ 是由总体 $F$ 中抽取的*i.i.d.*样本,  $g(\mathbf{X})$ 是 $\theta$ 的任一无偏估计, 则有

$$\text{Var}(g(\mathbf{X})) \geq \frac{1}{nI(\theta)}, \quad \forall \theta \in \Theta \quad \text{C-R不等式}$$

## ■ 两点说明：

- 验证样本分布族是否满足C-R正则条件十分麻烦. 但幸运的是对指数族这些正则条件皆成立
- 若 $\text{Var}(g(\mathbf{X}))$ 达不到C-R下界, 并不能得出结论说 $\theta$ 的UMVUE不存在. 存在这样的例子,  $\theta$ 的UMVUE存在, 但其方差大于C-R下界

证：由于 $X_1, \dots, X_n$ 为i. i. d. 样本，因此 $f(\mathbf{x}, \theta) = \prod_{i=1}^n f(x_i, \theta)$ . 记

$$S(\mathbf{x}, \theta) = \frac{\partial \log f(\mathbf{x}, \theta)}{\partial \theta} = \sum_{i=1}^n \frac{\partial \log f(x_i, \theta)}{\partial \theta}$$

由正则条件(3)和(4)可知

$$\begin{aligned} E_{\theta}\{S(\mathbf{X}, \theta)\} &= \sum_{i=1}^n E_{\theta}\left\{\frac{\partial \log f(X_i, \theta)}{\partial \theta}\right\} = \sum_{i=1}^n \int \frac{1}{f(x_i, \theta)} \frac{\partial f(x_i, \theta)}{\partial \theta} \cdot f(x_i, \theta) dx_i \\ &= \sum_{i=1}^n \int \frac{\partial f(x_i, \theta)}{\partial \theta} dx_i = \sum_{i=1}^n \frac{\partial}{\partial \theta} \int f(x_i, \theta) dx = 0. \end{aligned}$$

由正则条件(3)和(4)可知

$$\begin{aligned} \text{Cov}_{\theta}(\hat{g}(\mathbf{X}), S(\mathbf{X}, \theta)) &= E_{\theta}\{\hat{g}(\mathbf{X}) \cdot S(\mathbf{X}, \theta)\} \\ &= \int \cdots \int \hat{g}(\mathbf{x}) \left[ \frac{1}{f(\mathbf{x}, \theta)} \frac{\partial f(\mathbf{x}, \theta)}{\partial \theta} \right] f(\mathbf{x}, \theta) d\mathbf{x} \\ &= \int \cdots \int \hat{g}(\mathbf{x}) \frac{\partial f(\mathbf{x}, \theta)}{\partial \theta} d\mathbf{x} \\ &= \frac{\partial}{\partial \theta} \int \cdots \int g(\mathbf{x}) f(\mathbf{x}, \theta) d\mathbf{x} = \frac{\partial \theta}{\partial \theta} = 1 \end{aligned}$$

另一方面，我们有

$$\text{Var}(S(\mathbf{x}, \theta)) = \sum_{i=1}^n \text{Var}\left(\frac{\partial \log f(\mathbf{x}, \theta)}{\partial \theta}\right) = \sum_{i=1}^n E\left[\left(\frac{\partial \log f(\mathbf{x}, \theta)}{\partial \theta}\right)^2\right] = nI(\theta)$$

由 Cauchy-Schwartz 不等式可知

$$[\text{Cov}(g(\mathbf{X}), S(\mathbf{X}, \theta))]^2 \leq \text{Var}(g(\mathbf{X}))\text{Var}(S(\mathbf{X}, \theta))$$

将  $\text{Cov}(g(\mathbf{X}), S(\mathbf{X}, \theta)) = 1$  和  $\text{Var}(S(\mathbf{X}, \theta)) = nI(\theta)$  代入上述不等式，  
可得

$$\text{Var}(g(\mathbf{X})) \geq \frac{1}{nI(\theta)}$$

例：设 $\mathbf{X} = (X_1, \dots, X_n)$ 为从两点分布 $b(1, p)$ 中抽取的简单样本，

证明样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 为 $p$ 的UMVUE

解：设随机变量 $X \sim b(1, p)$ ，则其概率分布为 $f(x, p) = p^x(1-p)^{1-x}$ ，其中 $x \in \{0, 1\}$ ,  $0 < p < 1$ 。由于两点分布族是指数族，因此C-R正则条件成立。Fisher信息量为

$$\begin{aligned} I(p) &= E_p \left[ \left( \frac{\partial \log f(X, p)}{\partial p} \right)^2 \right] = E_p \left[ \left( \frac{X - p}{p(1-p)} \right)^2 \right] \\ &= \frac{\text{Var}(X)}{p^2(1-p)^2} = \frac{1}{p(1-p)} \end{aligned}$$

因此，C-R下界为 $p(1-p)/n$

另一方面，已知 $\bar{X}$ 为 $p$ 的无偏估计，其方差为 $p(1-p)/n$ ，达到C-R下界，故 $\bar{X}$ 为 $p$ 的UMVUE

例：设  $\mathbf{X} = (X_1, \dots, X_n)$  为从Poisson分布  $P(\lambda)$  中抽取的简单样本，

用C-R不等式验证样本均值  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  为  $\lambda$  的UMVUE

解：设随机变量  $X \sim P(\lambda)$ ，则其概率分布为  $f(x, \lambda) = e^{-\lambda} \lambda^x / x!$ ，其中  $x = 0, 1, 2, \dots, \lambda > 0$ 。由于Poisson分布族是指数族，因此C-R正则条件成立。Fisher信息量为

$$I(\lambda) = E_{\lambda} \left[ \frac{\partial \log f(X, \lambda)}{\partial \lambda} \right]^2 = E_{\lambda} \left[ \frac{X - \lambda}{\lambda} \right]^2 = \frac{\text{Var}(X)}{\lambda^2} = \frac{1}{\lambda}$$

因此，C-R下界为  $\lambda/n$

另一方面，已知  $\bar{X}$  为  $\lambda$  的无偏估计，其方差为  $\lambda/n$ ，达到C-R下界，故  $\bar{X}$  为  $\lambda$  的UMVUE

**例：** 设 $\mathbf{X} = (X_1, \cdots, X_n)$ 为从 $N(a, \sigma^2)$ 中抽取的简单随机样本，其中 $\sigma^2$ 已知，用C-R不等式验证样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 为 $a$ 的UMVUE

**解：** 由于正态分布族是指数族，因此C-R正则条件成立. 正态分布 $N(a, \sigma^2)$ 的密度函数为

$$f(x, a) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2} (x - a)^2\right\}$$

Fisher信息量为

$$I(a) = E_a \left[ \left( \frac{\partial \log f(X, a)}{\partial a} \right)^2 \right] = E_a \left[ \frac{(X - a)^2}{\sigma^4} \right] = \frac{\text{Var}(X)}{\sigma^4} = \frac{1}{\sigma^2}$$

因此，C-R下界为 $\sigma^2/n$

另一方面，已知 $\bar{X}$ 为 $a$ 的无偏估计，其方差为 $\sigma^2/n$ ，达到C-R下界，故 $\bar{X}$ 为 $a$ 的UMVUE



# 估计量的效率及有效估计

定理：设 $g(\mathbf{X})$ 为 $\theta$ 的无偏估计，比值

$$e_n(\theta) = \frac{1/(nI(\theta))}{\text{Var}(g(\mathbf{X}))}$$

称为无偏估计 $g(\mathbf{X})$ 的效率

- 显然， $0 < e_n(\theta) \leq 1$ . 当 $e_n(\theta) = 1$ 时，称 $g(\mathbf{X})$ 是 $\theta$ 的有效估计. 若 $\lim_{n \rightarrow +\infty} e_n(\theta) = 1$ ，则称 $g(\mathbf{X})$ 为 $\theta$ 的渐近有效估计
- 这一概念的不足之处：从定义看，有效估计一定是UMVUE，但很多UMVUE不是有效估计. 这是因为C-R下界偏小，在很多场合UMVUE的方差达不到C-R下界

例：设 $\mathbf{X} = (X_1, \dots, X_n)$ 为从 $N(a, \sigma^2)$ 中抽取的简单随机样本，

1) 当 $a$ 未知时，证明样本方差 $S^2$ 不是 $\sigma^2$ 的有效估计，但是渐近有效估计

2) 当 $a$ 已知时，求 $\sigma^2$ 的有效估计

解：1) 当 $a$ 未知时， $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)$ 的方差为 $\frac{2\sigma^4}{n-1}$ ，达不到C-R下界 $2\sigma^4/n$ ，因此，估计的效率为

$$e_n(\sigma^2) = \frac{2\sigma^4/n}{2\sigma^4/(n-1)} = \frac{n-1}{n}$$

因此，它不是 $\sigma^2$ 的有效估计，但是渐近有效估计

2) 当 $a$ 已知时， $S^2 = \sum_{i=1}^n (X_i - a)^2 / n$ 的方差为 $\frac{2\sigma^4}{n}$ ，达到C-R下界 $2\sigma^4/n$ ，因此，这时 $S^2 = \sum_{i=1}^n (X_i - a)^2 / n$ 是有效估计

# 提纲

---

- 参数估计及评价准则
- 矩估计
- 最大似然估计
- 最小方差无偏估计
- Cramer-Rao不等式
- 贝叶斯估计

# 统计推断中可用的三种信息

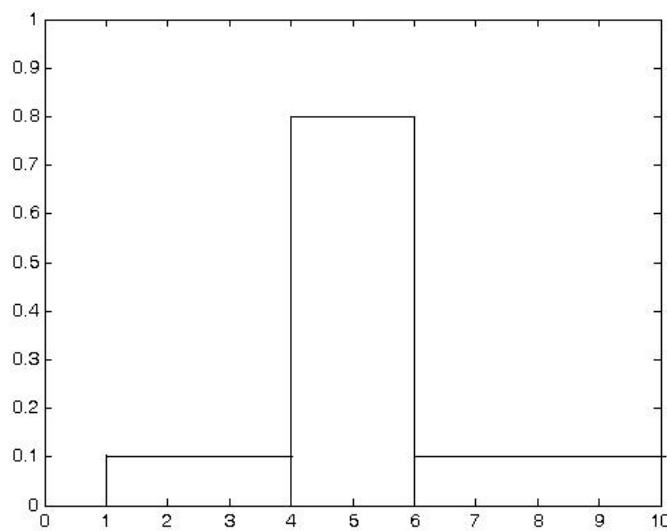
---

- 1) 总体信息 (population information): 总体分布或总体所属分布族所提供的信息. 如: 总体是正态分布, 总体是指数分布等
- 2) 样本信息 (sample information): 样本提供我们的信息
- 3) 先验信息 (prior information): 先验信息来源千所考察的统计推断问题之外, 在抽样之前就有的有关统计推断问题的信息. 常常是过去同类统计推断问题提供的信息

**例：**某工厂考察某天所生产的产品的不合格率时，过去抽检这种产品质量的资料(历史数据)对我们估计这一天的不合格率是有好处的，这些资料提供的信息就是先验信息.

- 某工程师根据自己多年积累的经验对正在设计的某种彩电的平均寿命所给出的估计

**例：**某地区煤的储存量  $\theta$  在几百年内不会有多大变化，可以看作是一个常量，但对人们来说，它是未知的、不确定的量。有位专家研究了有关资料，结合他的经验认为：该地区的储存量  $\theta$  “大概有5亿吨左右”。如果把“5亿吨左右”理解为4亿吨到6亿吨之内，把“大概”理解为80%的把握，还有20%的可能性在此区间之外。这无形中就使用一个概率分布去描述未知量  $\theta$ ，而具有概率分布的量当然是随机变量



- 经典（频率）统计学派：在统计推断中只利用总体信息和样本信息
- Bayes统计学派：建议在统计推断中在利用总体信息和样本信息的同时, 还利用先验信息

# 先验分布 (Prior Distribution)

---

- 如何表示先验信息，并让其统计模型中得到体现？

记总体的密度函数或分布函数为 $p(x; \theta)$

- 频率学派认为：参数 $\theta$ 虽然未知，但是一个**固定常数**
- Bayes学派认为：参数 $\theta$ 不是常数，而是一个**随机变量**

**例如：**某工厂每日生产的产品的次品率并不是固定的，而是在某一个范围内逐日变化的。Bayes学派认为：未知参数是变化的量，它的波动情况可用一个概率分布来描述



- Bayes学派认为未知参数 $\theta$ 是个随机变量，因此它会服从某一个概率分布，该分布即称为先验分布，常记为 $\pi(\theta)$ 
  - 先验分布 $\pi(\theta)$ 可从先验信息中归纳出来
  - 先验分布常常是主观概率, 取决于试验者在试验前对 $\theta$ 的先验信息了解的程度和他对这些信息的信任程度.

# Bayes统计的基本思想

## ■ Bayes统计中的联合分布

- 在 $\theta$ 给定后，总体服从条件分布 $p(x|\theta)$
- 简单随机样本 $X_1, X_2, \dots, X_n$ 在 $\theta$ 给定的条件下相互独立，即：  
在 $\theta$ 给定的条件下，它们的联合分布为

$$p(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n p(x_i | \theta)$$

(即：  $X_1, X_2, \dots, X_n$  条件独立)

该模型综合了总体信息和样本信息，但没有包括先验信息


- 综合先验信息 $\pi(\theta)$ 后, 得到的联合分布为

$$p(x_1, x_2, \dots, x_n, \theta) = p(\mathbf{x}, \theta) = \prod_{i=1}^n p(x_i | \theta) \cdot p(\theta)$$

该联合分布 综合了总体信息、样本信息、先验信息

- Bayes统计中对未知参数 $\theta$ 的推断

- 没有任何观察值的情况下,  $\theta$ 的分布就是先验分布 $\pi(\theta)$
- 在观察到 $X_1, X_2, \dots, X_n$ 的值后,  $\theta$ 的分布是

后验分布   $p(\theta | \mathbf{x}) = \frac{p(\mathbf{x} | \theta) \pi(\theta)}{\int_{\theta} p(\mathbf{x} | \theta) \pi(\theta) d\theta}$

例：设 $Y_1, \dots, Y_n \sim b(1, \theta)$ , 因此,  $X = \sum_{i=1}^n Y_i \sim b(n, \theta)$ . 试根据观察到的 $X$ , 求参数 $\theta$ 的Bayes估计 (假设无任何关于 $\theta$ 的先验信息)

解：  $\theta$ 给定的条件下,  $X$ 的分布

$$f(X = x|\theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

事先没有任何 $\theta$ 的信息, 可以认为

$$\pi(\theta) = \begin{cases} 1, & 0 \leq \theta \leq 1 \\ 0, & \text{其它} \end{cases}$$

$\theta$ 的后验分布等于

$$f(\theta|x) = \frac{\binom{n}{x} \theta^x (1 - \theta)^{n-x}}{\int_0^1 \binom{n}{x} \theta^x (1 - \theta)^{n-x} d\theta} = \frac{\theta^x (1 - \theta)^{n-x}}{\int_0^1 \theta^x (1 - \theta)^{n-x} d\theta}$$

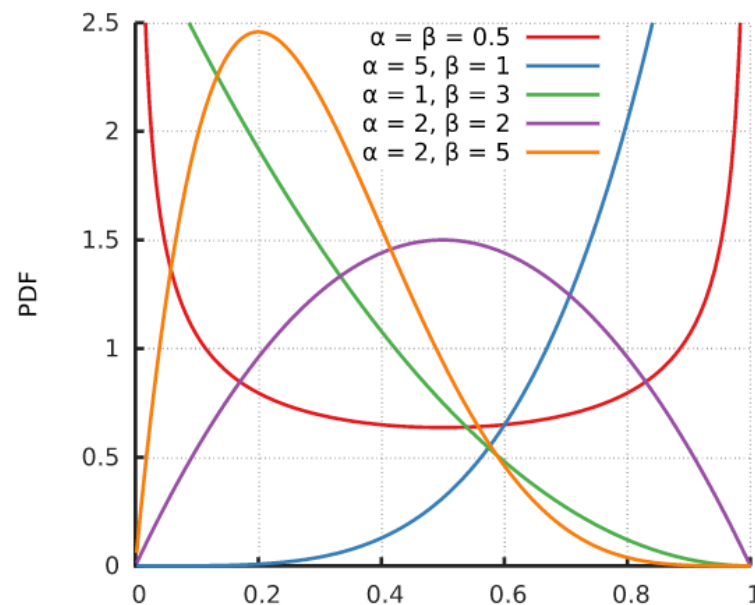
已知

$$\int_0^1 \theta^x (1 - \theta)^{n-x} d\theta = \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}$$

因此,  $\theta$  的后验分布为

$$f(\theta|x) = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} \theta^{(x+1)-1} (1-\theta)^{(n-x+1)-1}$$

这是参数为  $x+1$  和  $n-x+1$  的贝塔分布  $Be(x+1, n-x+1)$ , 其中  $Be(1, 1) = U(0, 1)$



例：上述例子中，如果事先对参数 $\theta$ 并不是毫无信息，而是知道 $\pi(\theta) = Be(a, b)$ . 试求 $\theta$ 的Bayes估计.

解：由 $f(x|\theta)$ 和 $\pi(\theta)$ 可知，

$$\begin{aligned} f(\theta|x) &= \frac{\binom{n}{x} \theta^x (1-\theta)^{n-x} \pi(\theta)}{\int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} \pi(\theta) d\theta} \\ &= \frac{\theta^x (1-\theta)^{n-x} Be(a, b)}{\int_0^1 \theta^x (1-\theta)^{n-x} Be(a, b) d\theta} \\ &= \frac{\theta^{x+a-1} (1-\theta)^{n-x+b-1}}{\int_0^1 \theta^{x+a-1} (1-\theta)^{n-x+b-1} d\theta} \\ &= Be(x+a, n-x+b) \end{aligned}$$

例：设 $X_1, X_2, \dots, X_n \sim N(\theta, \sigma^2)$ ,  $\sigma^2$ 已知而 $\theta$ 未知. 令 $\theta$ 的先验分布 $\pi(\theta)$ 是 $N(\mu, \tau^2)$ , 其中 $\mu$ 和 $\tau^2$ 已知, 求 $\theta$ 的后验分布 $\pi(\theta|x)$

解：样本的联合条件分布为

$$f(\mathbf{x}|\theta) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2\right\}$$

$\theta$ 的先验分布

$$\pi(\theta) = \frac{1}{\sqrt{2\pi}\tau} \exp\left\{-\frac{1}{2\tau^2} (\theta - \mu)^2\right\}$$

由此可写成 $\mathbf{x}$ 和 $\theta$ 的联合分布

$$f(\mathbf{x}, \theta) = k_1 \exp\left\{-\frac{n\theta^2 - 2n\theta\bar{x} + \sum_{i=1}^n x_i^2}{2\sigma^2} - \frac{\theta^2 - 2\mu\theta + \mu^2}{2\tau^2}\right\}$$

$$\text{其中 } k_1 = \frac{1}{(\sqrt{2\pi}\sigma)^n \sqrt{2\pi}\tau}, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

记

$$A = \frac{n}{\sigma^2} + \frac{1}{\tau^2} \quad B = \frac{n\bar{x}}{\sigma^2} + \frac{\mu}{\tau^2} \quad C = \frac{\sum_{i=1}^n x_i^2}{\sigma^2} + \frac{\mu^2}{\tau^2}$$

联合分布可表示成

$$\begin{aligned} f(\mathbf{x}, \theta) &= k_1 \exp \left\{ -\frac{A\theta^2 - 2B\theta + C}{2} \right\} \\ &= k_1 \exp \left\{ -\frac{(\theta - B/A)^2}{2/A} - \frac{1}{2} \left( C - \frac{B^2}{A} \right) \right\} \\ &= k_1 \exp \left\{ -\frac{1}{2} \left( C - \frac{B^2}{A} \right) \right\} \cdot \exp \left\{ -\frac{(\theta - B/A)^2}{2/A} \right\} \end{aligned}$$

后验分布

$$f(\theta|\mathbf{x}) = \frac{f(\mathbf{x}, \theta)}{m(\mathbf{x})} = \frac{k_1 \exp \left\{ -\frac{1}{2} \left( C - \frac{B^2}{A} \right) \right\}}{m(\mathbf{x})} \cdot \exp \left\{ -\frac{(\theta - B/A)^2}{2/A} \right\}$$



$$f(\theta|\mathbf{x}) = k_2 \cdot \exp \left\{ -\frac{(\theta - B/A)^2}{2/A} \right\}$$

$$\text{其中 } k_2 = \frac{k_1 \exp \left\{ -\frac{1}{2} \left( C - \frac{B^2}{A} \right) \right\}}{m(\mathbf{x})}$$

由  $f(\theta|\mathbf{x})$  的形式可知，其一定是如下分布

$$f(\theta|\mathbf{x}) = N \left( \frac{B}{A}, \frac{1}{A} \right)$$

$k_2$  一定等于高斯分布的归一化系数，即：  $k_2 = \frac{1}{\sqrt{2\pi \cdot 1/A}}$

将  $A$  和  $B$  的值带入可得

$$f(\theta|\mathbf{x}) = N \left( \frac{n\bar{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}, \frac{1}{n/\sigma^2 + 1/\tau^2} \right)$$

# 利用分布的核简化后验分布计算

- 给定条件分布 $f(x|\theta)$ , 先验分布 $\pi(\theta)$ ,  $\theta$ 后验一般表达式为

$$f(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}$$

- $f(\theta|x)$ 完全由 $f(x|\theta)\pi(\theta)$ 决定, 与 $m(x)$ 无关.  $m(x)$ 一定等于“归一化”系数, 即:  $m(x) = \int f(x|\theta)\pi(\theta)d\theta$
- 事实上, 如果 $f(x|\theta)\pi(\theta)$ 可拆解成 $f(x|\theta)\pi(\theta) = h(x)g(\theta)$ , 那么后验分布一定是由 $g(\theta)$ 完全决定

$$f(\theta|x) = \frac{1}{C} g(\theta)$$

其中 $C = \int g(\theta)d\theta$ 且  $C = \frac{m(x)}{h(x)}$ , 其中 $m(x) = \int f(x|\theta)\pi(\theta)d\theta$

- 因此，在计算后验分布的时候，只需要写出分布的核（只包含所感兴趣参数 $\theta$ 的项），其它与 $\theta$ 无关项可以不用写出

➤ 这些与 $\theta$ 无关的项综合在一起后一定是起到“归一化”的作用

- 这样，我们在表示后验分布的时候，可以写成如下形式

$$f(\theta|x) \propto f(x|\theta)\pi(\theta)$$

其中  $\propto$  表示两边在“不考虑常数缩放”的前提下相等

- $f(x|\theta)\pi(\theta)$ 并不是  $\theta$  的概率密度，但却唯一确定了一个概率密度（相差一个归一化因子）。因此， $f(x|\theta)\pi(\theta)$ 也称为后验分布 $f(\theta|x)$ 的核

例：设 $X_1, X_2, \dots, X_n \sim b(1, \theta)$ , 其中成功概率 $\theta$ 的先验分布为贝塔分布 $Be(a, b)$ , 求 $\theta$ 的后验分布

解：条件分布 $f(x|\theta)$ 和先验分布 $\pi(\theta)$ 的核

$$f(\mathbf{x}|\theta) = C_n^T \theta^T (1 - \theta)^{n-T}, \quad T = x_1 + \dots + x_n$$

$$\pi(\theta) \propto \theta^{a-1} (1 - \theta)^{b-1}, \quad 0 < \theta < 1$$

因此，后验分布的核为

$$f(\theta|\mathbf{x}) \propto \theta^{T+a-1} (1 - \theta)^{n+b-T-1}$$

可以看出，该核是贝塔分布 $Be(T + a, n + b - T)$ 的核。因此，可以直接写成后验分布为贝塔分布 $Be(T + a, n + b - T)$

例：设  $X \sim N(\theta, \sigma^2)$ ,  $\sigma^2$  已知而  $\theta$  未知. 令  $\theta$  的先验分布  $\pi(\theta)$  是  $N(\mu, \tau^2)$ , 其中  $\mu$  和  $\tau^2$  已知, 求  $\theta$  的后验分布  $\pi(\theta|x)$

解：后验分布的核

$$\begin{aligned}\pi(\theta|x) &\propto f(x|\theta)\pi(\theta) \propto \exp\left\{-\frac{1}{2}\left[\frac{(x-\theta)^2}{\sigma^2} + \frac{(\theta-\mu)^2}{\tau^2}\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}[A\theta^2 - 2B\theta + C]\right\}\end{aligned}$$

$$\text{其中 } A = \frac{1}{\sigma^2} + \frac{1}{\tau^2}, \quad B = \frac{x}{\sigma^2} + \frac{\mu}{\tau^2}, \quad C = \frac{x^2}{\sigma^2} + \frac{\mu^2}{\tau^2}$$

$$\begin{aligned}\pi(\theta|x) &\propto \exp\left\{-\frac{A}{2}\left(\theta - \frac{B}{A}\right)^2 - \frac{1}{2}\left(C - \frac{B^2}{A}\right)\right\} \\ &= \exp\left\{-\frac{A}{2}\left(\theta - \frac{B}{A}\right)^2\right\} \cdot \exp\left\{-\frac{1}{2}\left(C - \frac{B^2}{A}\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2\eta_x^2}(\theta - \mu_x)^2\right\}\end{aligned}$$

其中

$$\eta_x^2 = \frac{1}{A} = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}$$

$$\mu_x = \frac{B}{A} = \frac{\sigma^2 \mu + \tau^2 x}{\sigma^2 + \tau^2}$$

因此，可以知道

$$\pi(\theta|x) = N(\mu_x, \eta_x^2)$$

# 共轭先验分布

---

- 如何确定先验分布是贝叶斯统计中的关键一步，它会影响最后的贝叶斯统计推断结果
- 如何确定先验分布的形式，主要要考虑两个因素
  - 先验分布符合先验信息（经验和历史资料）
  - 使用简单，即：后验分布容易获得

在前面几个例子中，有没有发现后验分布和先验分布具有什么特点？

- 在前面贝塔和正态分布的例子中，我们看到后验分布的形式会与先验分布的形式保持一致

➤ 条件分布 $b(n, \theta)$ 、先验 $Be(a, b)$ 、后验 $Be(x + a, n - x + b)$

➤ 条件分布 $N(\theta, \sigma^2)$ 、先验 $N(\mu, \tau^2)$ 、后验 $N\left(\frac{n\bar{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}, \frac{1}{n/\sigma^2 + 1/\tau^2}\right)$

这是否是偶然？

这种能让后验分布保持与先验分布类型一致的先验分布叫作**共轭先验分布**

- 从上述两个例子可以看出，**不同的条件分布对应不同的共轭先验分布**



- 事实上，共轭先验分布不仅跟条件分布有关，还与条件分布中的具体参数有关。同一个条件分布中的不同参数对应的共轭先验分布不相同
  - 例如：对于条件分布 $N(\theta, \sigma^2)$ ，参数 $\theta$ 的共轭先验分布是高斯分布，而参数 $\sigma^2$ 的共轭先验分布是倒Gamma分布

总体条件分布	参数	共轭先验分布
二项分布	成功概率	贝塔分布
泊松分布	均值	伽马分布
指数分布	均值倒数	伽马分布
正态分布（方差已知）	均值	正态分布
正态分布（均值已知）	方差	倒伽马分布

# 贝叶斯估计

- 后验分配 $\pi(\theta|\mathbf{x})$ 综合了总体分布 $p(\mathbf{x}|\theta)$ 、样本 $\mathbf{x}$ 和先验 $\pi(\theta)$ 中有关 $\theta$ 的信息
- 如果要寻求参数 $\theta$ 的点估计 $\hat{\theta}$ ，只需从后验分布 $\pi(\theta|\mathbf{x})$ 合理提取信息即可
- 如何确定点估计 $\hat{\theta}$ ？
- 一种常用方法是希望找到一个 $\hat{\theta}$ ，使 $\theta$ 到该值的MSE最小，即：

$$\begin{aligned}MSE(\hat{\theta}|\mathbf{x}) &= \int (\theta - \hat{\theta})^2 \pi(\theta|\mathbf{x}) d\theta \\ &= E_{\theta|\mathbf{x}} [(\theta - \hat{\theta})^2]\end{aligned}$$

$$\begin{aligned}MSE(\hat{\theta}|\mathbf{x}) &= \int (\theta - \hat{\theta})^2 \pi(\theta|\mathbf{x}) d\theta \\&= E_{\theta|\mathbf{x}}[\theta^2] - 2E_{\theta|\mathbf{x}}[\theta]\hat{\theta} + \hat{\theta}^2\end{aligned}$$

- $\hat{\theta}$ 等于什么值时,  $MSE(\hat{\theta}|\mathbf{x})$ 最小?

$$\hat{\theta} = E_{\theta|\mathbf{x}}[\theta] = \int \theta \pi(\theta|\mathbf{x}) d\theta$$

- $\hat{\theta} = E_{\theta|\mathbf{x}}[\theta]$ 常称为 $\theta$ 的贝叶斯估计, 记为 $\hat{\theta}_B$

例：设 $X_1, X_2, \dots, X_n \sim N(\theta, \sigma^2)$ ， $\sigma^2$ 已知而 $\theta$ 未知。令 $\theta$ 的先验分布 $\pi(\theta)$ 是 $N(\mu, \tau^2)$ ，其中 $\mu$ 和 $\tau^2$ 已知。前面例题已经知道 $\theta$ 的后验分布为

$$\pi(\theta|\mathbf{x}) = N\left(\frac{n\bar{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}, \frac{1}{n/\sigma^2 + 1/\tau^2}\right)$$

求 $\theta$ 的贝叶斯估计

解： $\theta$ 的贝叶斯估计为

$$\hat{\theta} = E_{\theta|\mathbf{x}}[\theta] = \frac{n\bar{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}$$

# 贝叶斯估计中的充分统计量

- 在经典统计中，判断一个统计量 $T(\mathbf{x})$ 是否为充分统计量的充要条件是因子分解定理，即：若样本 $X_1, \dots, X_n$ 的分布 $p(\mathbf{x}|\theta)$ 可以分解为如下形式

$$p(\mathbf{x}|\theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$$

- 在贝叶斯统计中，判断一个统计量是否为充分统计量也有一个充要条件，即：若 $\theta$ 的后验分布 $\pi(\theta|\mathbf{x})$ 可以表示为 $\theta$ 和某个统计量 $T(\mathbf{x})$ 的函数

$$\pi(\theta|\mathbf{x}) = \pi(\theta|T(\mathbf{x}))$$

则 $T(\mathbf{x})$ 为 $\theta$ 的充分统计量

**例：** 设  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  是来自正态总体  $N(\mu, 1)$  的一个样本， $\mu$  的先验分布取为共轭先验  $N(0, \tau^2)$ ，其中  $\tau^2$  已知。在经典统计学中，已经知道样本均值  $\bar{x}$  是  $\mu$  的充分统计量。现要验证，在贝叶斯统计中， $\bar{x}$  仍是  $\mu$  的充分统计量。

**解：** 已知  $x_i \sim N(\mu, 1)$  和  $\mu \sim N(0, \tau^2)$ ，利用它们的核可知

$$\begin{aligned}\pi(\mu|\mathbf{x}) &\propto p(\mathbf{x}|\mu)\pi(\mu) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2\right\} \cdot \exp\left\{-\frac{\mu^2}{2\tau^2}\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\mu^2\left(n + \frac{1}{\tau^2}\right) - 2n\bar{x}\mu\right]\right\} \\ &\propto \exp\left\{-\frac{n + \tau^{-2}}{2}\left(\mu - \frac{n\bar{x}}{n + \tau^{-2}}\right)^2\right\}\end{aligned}$$

可以看出  $\pi(\mu|\mathbf{x}) = N\left(\frac{n\bar{x}}{n+\tau^{-2}}, \frac{1}{n+\tau^{-2}}\right)$ ，只与样本均值  $\bar{x}$  有关，因此在贝叶斯统计中， $\bar{x}$  也是  $\mu$  的充分统计量

# 实际应用例子

- 数据模型

$$y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$$

其中 $\mathbf{x}_i$ 表示模型的第 $i$ 个输入， $y_i$ 对应的观察值，噪声 $\epsilon_i$ 服从均值为0、方差为 $\sigma_0^2$ 的高斯分布 $N(\epsilon; 0, \sigma_0^2)$

- 给定 $\mathbf{x}_i$ 的情况下，输出 $y_i$ 服从什么分布？

$$p(y_i | \mathbf{x}_i; \mathbf{w}) = \mathcal{N}(y_i; \mathbf{w} \mathbf{x}_i, \sigma_0^2)$$

其中 $\mathbf{w}$ 表示模型参数

问题一：给定一组数据 $\{\mathbf{x}_i, y_i\}_{i=1}^n$ ，试求模型参数 $\mathbf{w}$ 的最大似然估计

概率模型： 
$$p(\mathbf{y}|\mathbf{X}; \mathbf{w}) = \prod_{i=1}^n p(y_i|\mathbf{x}_i; \mathbf{w})$$

其中 $\mathbf{y} = [y_1, y_2, \dots, y_n]^T$ ， $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$

对数似然： 
$$\begin{aligned} L(\mathbf{w}) &= \log p(\mathbf{y}|\mathbf{X}; \mathbf{w}) = \sum_{i=1}^n \log p(y_i|\mathbf{x}_i; \mathbf{w}) \\ &= -\frac{n}{2} \log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2} \sum_{i=1}^n (y_i - \mathbf{w}^T \mathbf{x}_i)^2 \end{aligned}$$

计算 $L(\mathbf{w})$ 的导数，得到

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} = -\frac{1}{\sigma_0^2} \sum_{i=1}^n \mathbf{x}_i (y_i - \mathbf{x}_i^T \mathbf{w}) = -\frac{1}{\sigma_0^2} (\mathbf{X}\mathbf{y} - \mathbf{X}\mathbf{X}^T \mathbf{w})$$



将 $L(\mathbf{w})$ 设置为0，求解得到

$$\mathbf{w} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y}$$

问题二：给定一组数据 $\{\mathbf{x}_i, y_i\}_{i=1}^n$ ，假定模型参数 $\mathbf{w}$ 的先验分布为

$p(\mathbf{w}) = N(\mathbf{w}; \mathbf{0}, \sigma_w^2 \mathbf{I})$ ，试求模型参数 $\mathbf{w}$ 的贝叶斯估计

概率模型：  $p(y_i, \mathbf{w} | \mathbf{x}_i) = \mathcal{N}(y_i; \mathbf{w}\mathbf{x}_i, \sigma_0^2)N(\mathbf{w}; \mathbf{0}, \sigma_w^2 \mathbf{I})$

$$= \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(y_i - \mathbf{w}^T \mathbf{x}_i)^2}{2\sigma_0^2}} \frac{1}{(2\pi\sigma_w^2)^{\frac{m}{2}}} e^{-\frac{\mathbf{w}^T \mathbf{w}}{2\sigma_w^2}}$$

其中 $m$ 表示 $\mathbf{w}$ 的维度

$$\begin{aligned}
p(\mathbf{w}|\mathbf{x}_i, y_i) &\propto \exp \left\{ -\frac{(y_i - \mathbf{w}^T \mathbf{x}_i)^2}{2\sigma_0^2} - \frac{\mathbf{w}^T \mathbf{w}}{2\sigma_w^2} \right\} \\
&\propto \exp \left\{ -\frac{\mathbf{w}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{w} - 2y_i \mathbf{x}_i^T \mathbf{w}}{2\sigma_0^2} - \frac{\mathbf{w}^T \mathbf{w}}{2\sigma_w^2} \right\} \\
&\propto \exp \left\{ -\frac{1}{2} \left[ \mathbf{w}^T \left( \frac{1}{\sigma_0^2} \mathbf{x}_i \mathbf{x}_i^T + \frac{1}{\sigma_w^2} \mathbf{I} \right) \mathbf{w} - \frac{2y_i \mathbf{x}_i^T}{\sigma_0^2} \mathbf{w} \right] \right\} \\
&\propto \exp \left\{ -\frac{1}{2} \left[ \left( \mathbf{w} - \frac{1}{\sigma_0^2} \mathbf{P}_i^{-1} \mathbf{x}_i y_i \right)^T \mathbf{P} \left( \mathbf{w} - \frac{1}{\sigma_0^2} \mathbf{P}_i^{-1} \mathbf{x}_i y_i \right) \right] \right\}
\end{aligned}$$

其中  $\mathbf{P}_i = \frac{1}{\sigma_0^2} \mathbf{x}_i \mathbf{x}_i^T + \frac{1}{\sigma_w^2} \mathbf{I}$

$$p(\mathbf{w}|\mathbf{x}_i, y_i) \propto N \left( \mathbf{w}; \frac{1}{\sigma_0^2} \mathbf{P}_i^{-1} \mathbf{x}_i y_i, \mathbf{P}_i^{-1} \right)$$

- 前面只给出了一个观察点下的贝叶斯估计，多个观察值共同决定 $\mathbf{w}$ 的贝叶斯估计是什么呢？

概率模型： 
$$p(\mathbf{y}, \mathbf{w} | \mathbf{X}) = \prod_{i=1}^n \mathcal{N}(y_i; \mathbf{w}^T \mathbf{x}_i, \sigma_0^2) \times N(\mathbf{w}; \mathbf{0}, \sigma_w^2 \mathbf{I})$$
$$= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(y_i - \mathbf{w}^T \mathbf{x}_i)^2}{2\sigma_0^2}} \times \frac{1}{(2\pi\sigma_w^2)^{\frac{m}{2}}} e^{-\frac{\mathbf{w}^T \mathbf{w}}{2\sigma_w^2}}$$

$$p(\mathbf{w} | \mathbf{X}, \mathbf{y}) \propto \exp \left\{ -\frac{\sum_{i=1}^n (y_i - \mathbf{w}^T \mathbf{x}_i)^2}{2\sigma_0^2} - \frac{\mathbf{w}^T \mathbf{w}}{2\sigma_w^2} \right\}$$

$$\propto \exp \left\{ -\frac{\mathbf{w}^T \mathbf{X} \mathbf{X}^T \mathbf{w} - 2\mathbf{y}^T \mathbf{X}^T \mathbf{w}}{2\sigma_0^2} - \frac{\mathbf{w}^T \mathbf{w}}{2\sigma_w^2} \right\}$$

$$\begin{aligned}
&\propto \exp \left\{ -\frac{\mathbf{w}^T \mathbf{X} \mathbf{X}^T \mathbf{w} - 2 \mathbf{y}^T \mathbf{X}^T \mathbf{w}}{2\sigma_0^2} - \frac{\mathbf{w}^T \mathbf{w}}{2\sigma_w^2} \right\} \\
&\propto \exp \left\{ -\frac{1}{2} \left[ \mathbf{w}^T \left( \frac{1}{\sigma_0^2} \mathbf{X} \mathbf{X}^T + \frac{1}{\sigma_w^2} \mathbf{I} \right) \mathbf{w} - \frac{2}{\sigma_0^2} \mathbf{y}^T \mathbf{X}^T \mathbf{w} \right] \right\} \\
&\propto \exp \left\{ -\frac{1}{2} \left[ \left( \mathbf{w} - \frac{1}{\sigma_0^2} \mathbf{P}^{-1} \mathbf{X} \mathbf{y} \right)^T \mathbf{P} \left( \mathbf{w} - \frac{1}{\sigma_0^2} \mathbf{P}^{-1} \mathbf{X} \mathbf{y} \right) \right] \right\}
\end{aligned}$$

其中  $\mathbf{P} = \frac{1}{\sigma_0^2} \mathbf{X} \mathbf{X}^T + \frac{1}{\sigma_w^2} \mathbf{I}$

$$p(\mathbf{w} | \mathbf{X}, \mathbf{y}) \propto N \left( \mathbf{w}; \frac{1}{\sigma_0^2} \mathbf{P}^{-1} \mathbf{X} \mathbf{y}, \mathbf{P}^{-1} \right)$$

- 能看出贝叶斯估计与最大似然估计间的联系吗？

# 本章作业

---

- 2.1节: 2, 3, 6, 8
- 2.2节: 1, 2, 4, 10, 14
- 2.3节: 2, 5, 7, 9, 12, 17
- 2.4节: 1, 6
- 2.5节: 1, 6, 7