数理统计笔记

作者：肖程哲

时间：November 17, 2022

苟日新，日日新，又日新

目录

第 3 章 点估计

考试重点

- 矩估计和极大似然估计
- 点估计的其他评价方法
- UMVUE
- (弱) 相合和强相合
- Fisher 信息量和有效估计
- 贝叶斯估计

定义 3.1 (点估计)

用于估计参数 $\gamma = g(\theta)$ 的统计量 $T(\mathbf{X})$ 称为**估计量** (estimator), 记为 $\hat{\gamma} = T(X)$. 估计式是随机变量, 若给定统计模型, 则其分布由参数 θ 决定, 将观测数值代入估计式后得到的值称为**估计值**.

将数据代入估计式得到的估计值并不是真实的参数值，若使用新的样本，很可能会得到不同的估计值。

3.1 点估计的评价方法

3.1.1 无偏性

定义 3.2 (无偏估计)

估计式的**偏差** (standard error) 定义为其抽样分布的均值与实际参数的偏差, 即:

$$E(\hat{\theta}) - \theta$$

若其为零, 则称此估计为参数 θ 的**无偏估计**

例题 3.1 方差的无偏估计

$$\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=0}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}[\sum_{i=0}^{n} X_i^2 - n\bar{X}^2] \\
&= \frac{1}{n-1}[\sum_{i=0}^{n} X_i^2 - \frac{1}{n}(\sum_{i=0}^{n} X_i)^2] = \frac{1}{n-1}[\sum_{i=0}^{n} X_i^2 - \frac{1}{n}(\sum_{i=0}^{n} X_i^2 + \sum_{i\neq j}^{n} X_i X_j)] \\
&= \frac{1}{n}\sum_{i=0}^{n} X_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j}^{n} X_i X_j
\end{aligned}$$

$$E(s^2) = \frac{1}{n}\sum_{i=0}^{n} E(X_i^2) - \frac{1}{n(n-1)}\sum_{i\neq j}^{n} E(X_i X_j) = \mu_2 - m_1^2 = \sigma^2$$

定义 3.3 (渐进无偏估计)

若样本量趋于无穷时，估计式的偏差为零，即:

$$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$$

则称此估计为参数 θ 的**渐近无偏估计**

例题 3.2 样本方差的渐进无偏性 对于样本方差 $s_n^2 = \frac{1}{n}\sum_{i=0}^{n}(X_i - \bar{X})^2$,其期望为 $E(s_n^2) = \frac{n-1}{n}\sigma^2$。当样本量趋于无穷时，有 $E(s^{*2}) \to \sigma^2$。故 s^{*2} 为 σ^2 的渐近无偏估计。

无偏性不具有不变性，即若 $\hat{\theta}$ 是 θ 的无偏估计，一般而言，其函数 $g(\hat{\theta})$ 不是 $g(\theta)$ 的无偏估计，除非 $g(\theta)$ 是 θ 的线性函数。譬如，s^2 是 σ^2 的无偏估计，但 s 不是 σ 的无偏估计。

例题 3.3 设总体为 $N(\mu, \sigma^2), x_1, \cdots, x_n$ 是样本，考察 s 是否是 σ 的无偏估计。$Y = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$，其密度函数为

$$p(y) = \frac{1}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})} y^{\frac{n-1}{2}-1} \mathrm{e}^{-\frac{y}{2}}, \quad y > 0$$

从而

$$\begin{aligned} E(Y^{1/2}) &= \int_0^{+\infty} y^{1/2} p(y) \mathrm{d}y = \frac{1}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})} \int_0^{\infty} y^{\frac{n}{2}-1} \mathrm{e}^{-\frac{y}{2}}\, \mathrm{d}y \\ &= \frac{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})} = \sqrt{2}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \end{aligned}$$

由此，我们有

$$E(s) = \frac{\sigma}{\sqrt{n-1}} E(Y^{1/2}) = \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \cdot \sigma$$

这说明 s 不是 σ 的无偏估计。

当参数存在无偏估计时，我们称该参数是**可估**的，否则称它是不可估的。

3.1.2 有效性

参数的无偏估计可以有很多，如何在无偏估计中进行选择？直观的想法是希望该估计围绕参数真值的波动越小越好，波动大小可以用方差来衡量，因此人们常用无偏估计的方差的大小作为度量无偏估计优劣的标准，这就是有效性.

定义 3.4 (有效性)

设 $\hat{\theta}_1, \hat{\theta}_2$ 是 θ 的两个无偏估计，如果对任意的 $\theta \in \Theta$ 有

$$\mathrm{Var}(\hat{\theta}_1) \le \mathrm{Var}(\hat{\theta}_2),$$

且至少有一个 $\theta \in \Theta$ 使得上述不等号严格成立，则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ **有效**。

3.1.3 均方误差

无偏性是估计的一个优良性质，对无偏估计我们还可以通过其方差进行有效性比较. 然而不能由此认为：有偏估计一定是不好的估计.

在有些场合，有偏估计比无偏估计更优，这就涉及如何对有偏估计进行评价. 一般而言，在样本量一定时，评价一个点估计的好坏使用的度最指标总是点估计值与参数真值 8 的距离的函数，最常用的函数是距离的平方. 由于具有随机性，可以对该函数求期望，

定义 3.5

估计式与参数的平方的期望称为**均方误差**

$$\mathrm{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

♣

注意到

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) &= [E(\hat{\theta} - E\hat{\theta}) + (E\hat{\theta} - \theta)^2] \\ &= E(\hat{\theta} - E\hat{\theta})^2 + (E\hat{\theta} - \theta)^2 + 2E[(\hat{\theta} - E\hat{\theta})(E\hat{\theta} - \theta)] \\ &= \mathrm{Var}(\hat{\theta}) + (E\hat{\theta} - \theta)^2\end{aligned}$$

因此，均方误差由点估计的方差与偏差的平方两部分组成。如果 $\hat{\theta}$ 是 θ 的无偏估计，则 $\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta})$，此时用均方误差评价点估计与用方差是一样的；当 $\hat{\theta}$ 不是 θ 的无偏估计时，就要看其均方误差 $\mathrm{MSE}(\hat{\theta})$，即不仅要看其方差大小，还要看其偏差大小。

定义 3.6 (一致最小均方误差估计)

在一个估计类中，若其中的一个估计式 $\hat{\theta}$ 对估计类中任意一个估计式 $\tilde{\theta}$，在参数空间 Θ 上满足

$$\mathrm{MSE}_\theta(\hat{\theta}) \le \mathrm{MSE}_\theta(\tilde{\theta}), \quad \forall \theta \in \Theta$$

则称此估计式为该估计类中 θ 的**一致最小均方误差估计**。

♣

例题 3.4 对均匀总体 $U(0,\theta)$，考虑 θ 的形如 $\hat{\theta}_\alpha = \alpha \cdot x_{(n)}$ 的估计，其均方误差为

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}_\alpha) &= \mathrm{Var}(\alpha \cdot x_{(n)}) + (\alpha E x_{(n)} - \theta)^2 \\ &= \alpha^2 \mathrm{Var}(x_{(n)}) + \left(\alpha \frac{n}{n+1}\theta - \theta\right)^2 \\ &= \alpha^2 \frac{n}{(n+1)^2(n+2)} + \left(\frac{n \cdot \alpha}{n+1} - 1\right)^2 \theta^2.\end{aligned}$$

用求导的方法不难求出当 $\alpha_0 = (n+2)/(n+1)$ 时上述均方误差达到最小，此时 $\mathrm{MSE}\left(\frac{n+2}{n+1}x_{(n)}\right) = \frac{\theta^2}{(n+1)^2}$。$\hat{\theta}_0 = \frac{n+2}{n+1}x_{(n)}$ 虽是 θ 的有偏估计，但对于无偏估计 $\hat{\theta} = \frac{n+1}{n}x_{(n)}$ 而言，其均方误差 $\mathrm{MSE}(\hat{\theta}_0) = \frac{\theta^2}{(n+1)^2} < \frac{\theta^2}{n(n+2)} = \mathrm{MSE}(\hat{\theta})$。所以在均方误差的标准下，有偏估计 $\hat{\theta}_0$ 优于无偏估计 $\hat{\theta}$。并且 $\hat{\theta}_0$ 为形如 $\alpha \cdot x_{(n)}$ 的估计类中 θ 的一致最小均方误差估计。

若不对估计加以限制（即考虑所有可能的估计），则一致最小均方误差估计是不存在的，从而没有意义。事实上，若 $\hat{\theta}$ 是 θ 的所有估计中的一致最小均方误差估计，取定任一个 $\theta_0 \in$

$Theta$，令 $\tilde{\theta} = \theta_0$，则 $\mathrm{MSE}_{\theta_0}(\tilde{\theta}) = 0$，于是要求 $\hat{\theta}_0$ 也有 $\mathrm{MSE}_{\theta_0}(\hat{\theta}) = 0$，由 θ_0 的任意性，这意味着 MSE_θ 处处为 0，这显然是做不到的。

3.1.4 相合性

定义 3.7 (相合)

若当样本容量 $n \to \infty$ 时, 有

$$\hat{\gamma} \xrightarrow{\mathbb{P}} \gamma, \quad \forall \theta \in \Theta.$$

则称 $\hat{\gamma}$ 是**相合的/一致的** (consistent), 如果上式中 $\xrightarrow{\mathbb{P}}$ 可以增强为 $\xrightarrow{\text{a.s.}}$, 则称 $\hat{\gamma}$ 是**强相合的**.

注 由于参数是常数，所以依概率收敛与依分布收敛等价。

若把依赖于样本量 n 的估计量 $\hat{\theta}_n$. 看作一个随机变量序列, 相合性就是 $\hat{\theta}_n$. 依概率收敛于 θ，所以证明估计的相合性可应用依概率收敛的性质及各种大数定律.

例题 3.5 设 $x_1, \cdots, x_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本，则由辛钦大数定律及依概率收敛的性质知:

- $\bar{x}$ 是 μ 的相合估计;
- s^{*2} 是 σ^2 的相合估计;
- s^2 也是 σ^2 的相合估计.

由此可见参数的相合估计不止一个.

定理 3.1 (相合性定理)

设 $\hat{\theta}_n = \hat{\theta}_n(x_1, \cdots, x_n)$ 是 θ 的一个估计量, 若

$$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta, \quad \lim_{n\to\infty} \mathrm{Var}(\hat{\theta}_n) = 0$$

则 $\hat{\theta}_n$ 是 θ 的相合估计，

证明 由切比雪夫不等式有

$$P(|\hat{\theta}_n - E(\hat{\theta}_n)| \geq \varepsilon/2) \leq \frac{4}{\varepsilon^2}\mathrm{Var}(\hat{\theta}_n), \quad \forall \varepsilon > 0$$

另一方面，由 $\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$ 可知

$$\exists N, \forall n > N, \quad \text{s.t.}\ |E(\hat{\theta}_n) - \theta| < \varepsilon/2$$

注意到如果 $|\hat{\theta}_n - E(\hat{\theta}_n)| < \varepsilon/2$，就有

$$|\hat{\theta}_n - \theta| \leq |\hat{\theta}_n - E(\hat{\theta}_n)| + |E(\hat{\theta}_n) - \theta| < \varepsilon$$

故

$$\{|\hat{\theta}_n - E(\hat{\theta}_n)| < \varepsilon/2\} \subset |\hat{\theta}_n - \theta| < \varepsilon$$

等价地

$$\{|\hat{\theta}_n - E(\hat{\theta}_n)| < \varepsilon/2\} \supset \{|\hat{\theta}_n - \theta| < \varepsilon\}$$

由此即有

$$P(|\hat{\theta}_n-\theta|>\varepsilon)\le P(|\hat{\theta}_n-E(\hat{\theta}_n)|\ge\varepsilon/2)\le\frac{4}{\varepsilon^2}\operatorname{Var}(\hat{\theta}_n)\to 0$$

定理 3.2 (相合估计的连续映射)

若 $\hat{\theta}_{n1},\cdots,\hat{\theta}_{nk}$ 分别是 $\theta_1,\cdots,\theta_k$ 的相合估计，$\eta=g(\theta_1,\cdots,\theta_k)$ 是 $\theta_1,\cdots,\theta_k$ 的连续函数，则 $\hat{\eta}_n=g(\hat{\theta}_{n1},\cdots,\hat{\theta}_{nk})$ 是 η 的相合估计。

证明 由函数 g 的连续性，对 $\forall\varepsilon>0$，存在 $\exists\delta>0$，当满足 $|\hat{\theta}_j-\theta_j|<\delta,j=1,\cdots,k$ 时，有

$$|g(\hat{\theta}_1,\cdots,\hat{\theta}_k)-g(\theta_1,\cdots,\theta_k)|<\varepsilon$$

即

$$\bigcap_{j=1}^{k}\{|\hat{\theta}_{nj}-\theta_j|<\delta\}\subset\{|\hat{\eta}_n-\eta|<\varepsilon\}$$

又由 $\hat{\theta}_{n1},\cdots,\hat{\theta}_{nk}$ 的相合性得

$$\forall v>0,\exists N\in N_+,\forall n>N,\quad \text{s.t. } P(|\hat{\theta}_{nj}-\theta_j|\ge\delta)<v/k,\quad j=1,\cdots,k$$

从而有

$$\begin{aligned}
P\left(|\hat{\eta}_n-\eta|<\varepsilon\right)&>P\left(\bigcap_{i=1}^{k}\{|\hat{\theta}_{nj}-\theta_j|<\delta\}\right)\\
&=1-P\left(\bigcup_{j=1}^{k}\{|\hat{\theta}_{nj}-\theta_j|\ge\delta\}\right)\\
&\ge 1-\sum_{j=1}^{k}P(|\hat{\theta}_{nj}-\theta_j|\ge\delta)\\
&>1-k\cdot v/k=1-v
\end{aligned}$$

由 v 的任意性，定理得证。

3.1.5 渐近正态性

定义 3.8 (渐近正态性)

对于参数 $\tau(\theta)$ 的相合估计 T_n，若存在趋于 0 的非负常数序列 $\sigma_n(\theta)$，使得

$$\frac{T_n-\tau(\theta)}{\sigma_n(\theta)}\xrightarrow{d}N(0,1)$$

则称 T_n 是**渐近正态**的，服从渐近正态分布 $N(\tau(\theta),\sigma_n(\theta)^2)$，记为 $T_n\sim AN(\tau(\theta),\sigma_n(\theta)^2)$，$\sigma_n(\theta)^2$ 为 T_n 的**渐近方差**或 T_n 的极限分布的方差。

注 $\sigma_n(\theta)$ 表示着估计量 T_n 依概率收敛于 $\tau(\theta)$ 的速度。由中央极限定理可看出，通常为 $\frac{1}{\sqrt{n}}$ 量级。

例题 3.6 设总体为泊松分布 $P(\lambda)$，以样本均值为其参数估计，即

$$\hat{\lambda}_n = \bar{x_n}$$

由中心极限定理，$\frac{\hat{\lambda}_n-\lambda}{\sqrt{\lambda/n}}$ 依分布收敛于 $N(0,1)$，因此，$\hat{\lambda}_n$ 是渐近正态的，

3.2 点估计的方法

3.2.1 矩估计

定义 3.9 (总体矩与样本矩)

若 $X_1,\cdots,X_n \overset{\text{i.i.d.}}{\sim} P_\theta$，记 k 阶**总体矩**为:

$$\mu_k(\theta) = \mathbb{E}_\theta[X_i^k], \quad k \in \mathbb{N}$$

记 k 阶**样本矩**为：

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^n X_i^k, \quad k \in \mathbb{N}$$

♣

命题 3.1 (矩估计的无偏性)

k 阶样本矩是关于总体分布 k 阶矩的无偏估计。

♠

证明

$$E(\hat{\mu}_k) = \frac{1}{n}\sum_{i=1}^n E(X_i^k) = \mu_k$$

矩方法的步骤：

1. 将低阶矩通过参数表达，一般阶数与待估参数个数相同 $\mu = \mathbf{f}(\theta)$；
2. 找出上一步骤的反函数，通过矩表达参数 $\theta = \mathbf{f}^{-1}(\mu)$；
3. 将样本矩代入，得到参数的估计式 $\hat{\theta} = f^{-1}(\hat{\mu})$
4. 将参数的矩估计式代入函数 $g(\theta)$，得到参数的函数 $\eta = g(\theta)$ 的矩估计。

注 对于参数的函数的估计，矩估计和极大似然估计都是将对参数的估计代入。但极大似然估计的做法是根据其不变性；而矩估计是源于定义，不保证与函数的矩估计相同。

例题 3.7 泊松分布的矩估计 设 $X_1,\cdots,X_n \overset{\text{i.i.d.}}{\sim} P(\lambda)$，则其一阶矩 $\mu_1 = \lambda$，所以 $\lambda = \mu_1$，其矩估计为：

$$\hat{\lambda} = \hat{\mu_1} = \overline{X}$$

例题 3.8 设 $X_1,\ldots,X_n$ 是独立同分布的连续型随机变量, 具有 p.d.f.

$$f_{\lambda,a}(x) = \lambda \mathrm{e}^{-\lambda(x-a)}\mathbb{1}_{[x>a]}, \quad x \in \mathbb{R},$$

其中 $\lambda > 0$ 和 $a \in \mathbb{R}$ 未知 (注: 相应的统计模型是带有位置 (location) 参数和速率 (rate) 参数的指数分布). 易见 $X_i \overset{d}{=} a + Y/\lambda$, 其中 $Y \sim \text{Exponential}(1)$. 利用 $\mathbb{E}[Y] = 1$ 和 $\mathbb{E}[Y^2] = 2$, 可以得

到

$$\mu_1(\lambda, a) = a + 1/\lambda, \quad \& \quad \mu_2(\lambda, a) = a^2 + 2a/\lambda + 2/\lambda^2.$$

方程

$$\mu_k(\hat{\lambda}^{\text{MoM}}, \hat{a}^{\text{MoM}}) = \hat{\mu}_k, \quad k = 1, 2$$

的解

$$\hat{\lambda}^{\text{MoM}} = 1\Big/\sqrt{\hat{\mu}_2 - \hat{\mu}_1^2}, \quad \& \quad \hat{a}^{\text{MoM}} = \hat{\mu}_1 - \sqrt{\hat{\mu}_2 - \hat{\mu}_1^2}$$

即为 (λ, a) 的一种矩估计.

3.2.2 极大然似估计

极大然似估计的基本思想：对于参数空间中的每一个参数，计算在此参数下，观测数据的发生概率，选取最大概率对应的参数。

定义 3.10 (然似函数)

设随机变量 $X_1, \cdots, X_n$ 的联合密度函数（或质量函数）为 $f(x_1, \cdots, x_n|\theta)$。对于某一组观测数据 $x_1^*, \cdots, x_n^*$，其**然似函数** (likelihood function) 为：

$$\mathfrak{L}(\theta) = f(x_1^*, \cdots, x_n^*|\theta)$$

对数然似函数 (log likelihood function) 为 $l(\theta) = \log \mathfrak{L}(\theta)$

♣

注 联合密度质量函数代表概率，但联合密度函数不是，代表概率所占比例。然似函数是关于参数 θ 的函数，不是概率，对整个参数空间的积分未必等于一。

定义 3.11 (极大然似估计)

可使然似函数取最大值的参数被称为**极大然似估计**，即

$$\hat{\theta} = \max_{\theta \in \Theta} \mathfrak{L}(\theta)$$

♣

注 由于对数函数单调递增，最大化然似函数等价于最大化对数然似函数。

若样本来源变量独立同分布，即 $f(x_1, \cdots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta)$，则其然似函数和对数然似函数分布可写为：

$$\mathfrak{l}(\theta) = \prod_{i=1}^n f(x_i^*|\theta), \quad l(\theta) = \sum_{i=1}^n \log f(x_i^*|\theta)$$

命题 3.2 (极大然似估计的不变性)

设 $\hat{\theta}$ 是参数 θ 的极大然似估计，那么对于参数 θ 的任意函数 $\gamma = g(\theta)$，其极大然似估计 $\hat{\gamma} = g(\hat{\theta})$

♠

证明

3.2.3 贝叶斯估计

贝叶斯估计的基本观点是：任一未知量 θ 都可看作随机变量，可用概率分布 $\pi(\theta)$ 描述，称为**先验分布**；在获得样本之后，总体分布、样本与先验分布通过贝叶斯公式结合起来得到关于未知量 θ 的后验分布 $\pi(\theta|\mathbf{X})$。任何关于 θ 的统计推断都基于 θ 的后验分布进行。利用后验分布 $\pi(\theta|\mathbf{X})$ 估计 θ 有三种常用的方法：

- 使用后验分布的密度函数最大值点作为 θ 的点估计的最大后验估计;
- 使用后验分布的中位数作为 θ 的点估计的后验中位数估计;
- 使用后验分布的均值作为 θ 的点估计的后验期望估计.

用得最多的是后验期望估计, 它一般也简称为贝叶斯估计, 记为的 $\hat{\theta}_B$。

获取后验分布的步骤：

1. 根据总体信息与样本信息确定模型 $p(\mathbf{X}|\theta)$，根据先验信息确定参数 θ 的先验分布 $\pi(\theta)$
2. 综合总体信息、样本信息和先验信息得到样本 $\mathbf{X}$ 和参数 θ 的联合分布
$$h(\mathbf{X},\theta)=\mathbf{p}(\mathbf{X}|\theta)\pi(\theta)$$
3. 为求后验分布 $\pi(\theta|\mathbf{X})=\frac{\mathbf{h}(\mathbf{X},\theta)}{\mathbf{m}(\mathbf{X})}$，先求 X 的边际概率函数:
$$m(\mathbf{X})=\int_\theta \mathbf{h}(\mathbf{X},\theta)\mathrm{d}\theta=\int_\theta \mathbf{p}(\mathbf{X}|\theta)\pi(\theta)\mathrm{d}\theta$$
4. 得到后验分布
$$\pi(\theta|\mathbf{X})=\frac{\mathbf{h}(\mathbf{X},\theta)}{\mathbf{m}(\mathbf{X})}=\frac{\mathbf{p}(\mathbf{X}|\theta)\pi(\theta)}{\int_\theta \mathbf{p}(\mathbf{X}|\theta)\pi(\theta)\mathrm{d}\theta}$$

定义 3.12 (共扼先验分布)

设 θ 是总体参数, $\pi(\theta)$ 是其先验分布, 若对任意的样本观测值得到的后验分布 $\pi(\theta|\mathbf{X})$ 与 $\pi(\theta)$ 属于同一个分布族, 则称该分布族是 θ 的**共扼先验分布（族）**。 ♣

例题 3.9 设某事件 A 在一次试验中发生的概率为 θ，为估计 θ，对试验进行了 n 次独立观测，其中事件 A 发生了 X 次，显然 $X|\theta\sim b(n,\theta)$，即
$$P(X=x|\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x},\quad x=0,1,\cdots,n$$
假若我们在试验前对事件 A 没有什么了解，从对其发生的概率也没有任何信息。在这种场合，贝叶斯本人建议采用“同等无知”的原则，使用区间 $(0,1)$ 上的均匀分布 $U(0,1)$ 作为 θ 的先验分布。贝叶斯的这个建议被后人称为贝叶斯假设。

解 先写出 X 和 θ 的联合分布
$$h(x,\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x},\quad x=0,1,\cdots,n,\quad 0<\theta<1$$
然后求 X 的边际分布
$$m(x)=\binom{n}{x}\int_0^1\theta^x(1-\theta)^{n-x}\mathrm{d}\theta=\binom{n}{x}\frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}$$

最后求出 θ 的后验分布

$$\pi(\theta|x)=\frac{h(x,\theta)}{m(x)}=\frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\theta^{(x+1)-1}(1-\theta)^{(n-x+1)-1},\quad 0<\theta<1$$

即 $\theta|x\sim Be(x+1,n-x+1)$，故贝塔分布是伯努利试验中成功概率的共栭先验分布。其后验期望估计为

$$\hat{\theta}_{\mathrm{B}}=E(\theta|x)=\frac{x+1}{n+2}$$

例题 3.10 设 $x_1,\cdots,x_n$ 是来自 $N(\mu,\sigma_0^2)$ 的一个样本，其中 σ_0^2 已知，μ 未知。假设 μ 的先验分布亦为正态分布 $N(\theta,\tau^2)$，其中先验均值 θ 和先验方差 τ^2 均已知，试求 μ 的贝叶斯估计。

解 样本 $\mathbf{X}$ 的分布和 μ 的先验分布分别为

$$p(\mathbf{X}|\mu)=(2\pi\sigma_0^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

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

由此可以写出 $\mathbf{X}$ 与 μ 的联合分布为

$$h(\mathbf{X},\mu)=\mathbf{k_1}\cdot\exp\left\{-\frac{\mathbf{1}}{\mathbf{2}}\left[\frac{\mathbf{n}\mu^2-\mathbf{2n}\mu\bar{\mathbf{x}}+\sum_{\mathbf{i=1}}^{\mathbf{n}}\mathbf{x_i^2}}{\sigma_\mathbf{0}^\mathbf{2}}+\frac{\mu^\mathbf{2}-\mathbf{2}\theta\mu+\theta^\mathbf{2}}{\tau^\mathbf{2}}\right]\right\}$$

其中 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,k_1=(2\pi)^{-(n+1)/2}\tau^{-1}\sigma_0^{-n}$。若记

$$A=\frac{n}{\sigma_0^2}+\frac{1}{\tau^2},\quad B=\frac{n\bar{x}}{\sigma_0^2}+\frac{\theta}{\tau^2},\quad C=\frac{\sum_{i=1}^{n}x_i^2}{\sigma_0^2}+\frac{\theta^2}{\tau^2}$$

则有

$$\begin{aligned}h(\mathbf{X},\mu)&=k_1\exp\left\{-\frac{1}{2}[A\mu^2-2B\mu+C]\right\}\\&=k_1\exp\left\{-\frac{(\mu-B/A)^2}{2/A}-\frac{1}{2}(C-B^2/A)\right\}\end{aligned}$$

样本的边际密度函数为

$$m(\mathbf{X})=\int_{-\infty}^{+\infty}\mathbf{h}(\mathbf{X},\mu)\mathrm{d}\mu=\mathbf{k_1}\exp\left\{-\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{C}-\mathbf{B}^\mathbf{2}/\mathbf{A})\right\}(\mathbf{2}\pi/\mathbf{A})^{\mathbf{1/2}}$$

应用贝叶斯公式即可得到后验分布

$$\pi(\mu|\mathbf{X})=\frac{\mathbf{h}(\mathbf{X},\mu)}{\mathbf{m}(\mathbf{X})}=(\mathbf{2}\pi/\mathbf{A})^{-\mathbf{1/2}}\exp\left\{-\frac{\mathbf{1}}{\mathbf{2/A}}(\mu-\mathbf{B/A})^\mathbf{2}\right\}$$

这说明在样本给定后，μ 的后验分布为 $N(B/A,1/A)$, 即

$$\mu|\mathbf{X}\sim N\left(\frac{n\bar{x}\tau^2+\theta\sigma_0^2}{n\tau^2+\sigma_0^2},\frac{\sigma_0^2\tau^2}{n\tau^2+\sigma_0^2}\right)$$

故正态分布是正态总体的共栭先验分布。后验均值即为其贝叶斯估计:

$$\hat{\mu}=\frac{n\tau^2}{n\tau^2+\sigma_0^2}\bar{x}+\frac{\sigma_0^2}{n\tau^2+\sigma_0^2}\theta$$

它是样本均值 $\bar{x}$ 与先验均值 θ 的加权平均. 当总体方差 σ_0^2 较小或样本量 n 较大时, 样本均值 $\bar{x}$ 的权重较大; 当先验方差 τ^2 较小时, 先验均值 θ 的权重较大, 这一综合很符合人们的经验。

总体	参数	共轭分布族
$U(\theta-a,\theta+b)$	θ	均匀分布
$B(n,p)$	p	Beta 分布
$Poi(\lambda)$	λ	$Ga(\alpha,\beta)\rightarrow Ga(\alpha+n\bar{x},\beta+n)$
$N(\mu,\sigma_0^2)$	μ	$N(\theta,\tau^2)\rightarrow N\left(\frac{n\bar{x}\tau^2+\theta\sigma_0^2}{n\tau^2+\sigma_0^2},\frac{\sigma_0^2\tau^2}{n\tau^2+\sigma_0^2}\right)$
$N(\mu,\sigma^2)$	σ^2	$IGa(\alpha,\lambda)\rightarrow IGa(\alpha+\frac{n}{2},\beta+\frac{n}{2}\overline{(x-\mu)^2})$

表 3.1: 常见共轭分布族

3.2.4 最小二乘估计

3.3 最小方差无偏估计

由于一致最小均方误差估计（定义3.6）一般都不存在，为方便起见，需对估计加一些合理性前置要求，前述无偏性就是一个最常见的合理性要求。

3.3.1 最小方差无偏估计

定义 3.13 (一致最小方差无偏估计)

设 $\hat{\theta}$ 是 θ 的一个无偏估计，如果对任意一个 θ 的无偏估计 $\tilde{\theta}$，在参数空间 Θ 上都有

$$\mathrm{Var}_\theta(\hat{\theta})\leq\mathrm{Var}_\theta(\tilde{\theta})$$

则称 $\hat{\theta}$ 是 θ 的**一致最小方差无偏估计** (uniformly minimum variance unbiased estimation, UMVUE)。

定理 3.3 (UMVUE 的充要条件)

设 $\mathbf{X}=(\mathbf{x}_1,\Pi,\mathbf{x}_n)$ 是来自某总体的一个样本，$\hat{\theta}=\hat{\theta}(\mathbf{X})$ 是 θ 的一个无偏估计，$\mathrm{Var}(\hat{\theta})<+\infty$. 如果对任意一个满足 $E(\varphi(\mathbf{X}))=\mathbf{0}$ 的 $\varphi(\mathbf{X})$，都有

$$\mathrm{Cov}_\theta(\hat{\theta},\varphi)=0,\quad\forall\theta\in\Theta$$

则 $\hat{\theta}$ 是 θ 的 UMVUE。

证明 **充分性**：对 θ 的任意一个无偏估计 $\tilde{\theta}$

$$\mathrm{Var}(\tilde{\theta})=E(\tilde{\theta}-\theta)^2=E[(\tilde{\theta}-\hat{\theta})+(\hat{\theta}-\theta)]^2$$

代入 $\varphi(\mathbf{X})=\tilde{\theta}-\hat{\theta}$ 得：

$$\mathrm{Var}(\tilde{\theta})=E(\varphi^2)+\mathrm{Var}(\hat{\theta})+2\,\mathrm{Cov}(\varphi,\hat{\theta})\geq\mathrm{Var}(\hat{\theta})$$

必要性：若存在 $\theta_0 \in \Theta$，使得 $\mathrm{Cov}_{\theta_0}(\hat{\theta}, \phi(\mathbf{X})) = \mathbf{a} \neq \mathbf{0}$。则代入 $\varphi(\mathbf{X}) = \mathbf{b}(\tilde{\theta} - \hat{\theta})$ 得：

$$\mathrm{Var}(\tilde{\theta}) = b^2 E(\varphi^2) + \mathrm{Var}(\hat{\theta}) + 2b\,\mathrm{Cov}(\varphi, \hat{\theta}) = \mathrm{Var}(\hat{\theta}) + b^2 E(\varphi^2) + 2ab$$

再令 $b = -\frac{a}{E(\varphi^2)} \neq 0$，则可得 $b^2 E(\varphi^2) + 2ab < 0$，即 $\mathrm{Var}(\tilde{\theta}) < \mathrm{Var}(\hat{\theta})$，与假设不符。

例题 3.11 设 $x_1, \cdots, x_n$ 是来自指数分布 $Exp(1/\theta)$ 的样本，对于充分统计量 $T = x_1 + \cdots + x_n$，由于 $ET = n\theta$，所以 $\overline{x} = T/n$ 是 θ 的无偏估计。设 $\varphi = \varphi(x_1, \cdots, x_n)$ 是 θ 的任一无偏估计, 则

$$E\varphi(\mathbf{X}) = \int_0^{+\infty} \cdots \int_0^{+\infty} \varphi(x_1, \cdots, x_n) \cdot \prod_{i=1}^{n} \left\{\frac{1}{\theta} \cdot \mathrm{e}^{-x_i/\theta}\right\} \mathrm{d}x_1 \cdots \mathrm{d}x_n = 0$$

即

$$\int_0^{+\infty} \cdots \int_0^{+\infty} \varphi(x_1, \cdots, x_n) \cdot \exp\left\{-\frac{n\overline{x}}{\theta}\right\} \mathrm{d}x_1 \cdots \mathrm{d}x_n = 0$$

两端对 θ 求导, 得

$$\int_0^{+\infty} \cdots \int_0^{+\infty} \varphi(x_1, \cdots, x_n) \frac{n\overline{x}}{\theta^2} \cdot \exp\left\{-\frac{n\overline{x}}{\theta}\right\} dx_1 \cdots \mathrm{d}x_n = 0$$

这说明 $E(\overline{x} \cdot \varphi) = 0$, 从而

$$\mathrm{Cov}(\overline{x}, \varphi) = E(\overline{x} \cdot \varphi) - E(\overline{x}) \cdot E(\varphi) = 0$$

即 $\overline{x}$ 是 θ 的 UMVUE.

3.3.2 充分性原则

引理 3.1 (Rao–Blackwell 定理)

设 X 和 Y 是两个随机变量，$EX = \mu, \mathrm{Var}(X) > 0$。用条件期望构造一个新的随机变量 $\varphi(Y) = E(X|Y)$，则有

$$E\varphi(Y) = \mu, \mathrm{Var}\left(\varphi(Y)\right) \leq \mathrm{Var}(X)$$

其中等号成立的充分必要条件是 X 和 $\varphi(Y)$ 几乎处处相等.

证明 由重期望公式易得

$$E\varphi(Y) = E[E(X|Y)] = EX = \mu$$

将 $\mathrm{Var}(X)$ 写成如下的形式：

$$\begin{aligned}\mathrm{Var}(X) &= E[(X - \varphi(Y)) + (\varphi(Y) - \mu)]^2 \\ &= E(X - \varphi(Y))^2 + \mathrm{Var}(\varphi(Y) - \mu) + 2E[(X - \varphi(Y))(\varphi(Y) - \mu)]\end{aligned}$$

由重期望公式

$$\begin{aligned}E[(X - \varphi(Y))(\varphi(Y) - \mu)] &= E\{E[(X - \varphi(Y))(\varphi(Y) - \mu)|Y]\} \\ &= E\{[\varphi(Y) - \mu][E(X|Y) - \varphi(Y)]\} = 0\end{aligned}$$

由此即有

$$\mathrm{Var}(X) = E(X - \varphi(Y))^2 + \mathrm{Var}(\varphi(Y)) \leq \mathrm{Var}(\varphi(Y))$$

等号成立的充要条件为

$$E[X-\varphi(Y)]^2=0$$

即 X 和 $\varphi(Y)$ 几乎处处相等。

定理 3.4 (充分性原则)

设，$T=T(x_1,\cdots,x_n)$ 是 θ 的充分统计量，则对 θ 的任一无偏估计 $\hat{\theta}$，令 $\tilde{\theta}=E(\hat{\theta}|T)$，则 $\tilde{\theta}$ 也是 θ 的无偏估计，且

$$\mathrm{Var}(\tilde{\theta})\le\mathrm{Var}(\hat{\theta})$$

证明 由于 T 是充分统计量，故而 $\tilde{\theta}=E(\hat{\theta}|T)$ 与 θ 无关，因此它也是一个估计（统计量）。再将引理3.1取 $X=\hat{\theta},Y=T$ 即可。

例题 3.12 设 $x_1,\cdots,x_n\overset{\text{i.i.d.}}{\sim}b(1,p)$，则 $\bar{x}$（或 $T=n\bar{x}$）是 p 的充分统计量. 为估计 $\theta=p^2$，可令

$$\hat{\theta}_1=\begin{cases}1, & x_1=1,x_2=1;\\ 0, & \text{其他}.\end{cases}$$

由于

$$E(\hat{\theta}_1)=P(x_1=1,x_2=1)=p\cdot p=\theta$$

所以，$\hat{\theta}_1$ 是 θ 的无偏估计，这个估计并不好，它只使用了两个观测值，下面我们用 Rao–Blackwel 定理对之加以改进：求 $\hat{\theta}_1$ 关于充分统计量 $T=\sum_{i=1}^n x_i$ 的条件期望

$$\begin{aligned}\hat{\theta}&=E(\hat{\theta}_1|T=t)\\&=P(\hat{\theta}_1=1|T=t)\\&=\frac{P(X_1=1,X_2=1,T=t)}{P(T=t)}\\&=\frac{P(X_1=1,X_2=1,\sum_{i=3}^n X_i=t-2)}{P(T=t)}\\&=\frac{p\cdot p\cdot\binom{n-2}{t-2}p^{t-2}(1-p)^{n-t}}{\binom{n}{t}p^t(1-p)^{n-t}}\\&=\binom{n-2}{t-2}\Big/\binom{n}{t}=\frac{t(t-1)}{n(n-1)},\end{aligned}$$

3.3.3 费希尔信息量

定义 3.14

若总体分布 $p(x;\theta),\theta\in\Theta$ 满足

1. 参数空间 Θ 是直线上的一个开区间;
2. 支撑 $S=\{x:p(x;\theta)>0\}$ 与 Θ 无关;
3. 导数 $\frac{\partial}{\partial\theta}p(x;\theta)$ 对一切 $\theta\in\Theta$ 都存在;

4. 对 $p(x;\theta)$, 积分与微分运算可交换次序, 即

$$\frac{\partial}{\partial\theta}\int_{-\infty}^{+\infty} p(x;\theta)\mathrm{d}x = \int_{-\infty}^{+\infty}\frac{\partial}{\partial\theta}p(x;\theta)\mathrm{d}x$$

5. 期望 $E\left[\frac{\partial}{\partial\theta}\ln p(x;\theta)\right]^2$ 存在,

则称

$$I(\theta) = E\left[\frac{\partial}{\partial\theta}\ln p(x;\theta)\right]^2$$

为总体分布的**费希尔信息量** (Fisher)。

注 "$I(\theta)$ 越大"代表总体分布中包含未知参数 θ 的信息越多。

命题 3.3

若总体 $p(x;\theta)$ 的费希尔信息量存在，且存在二阶导数 $\frac{\partial^2}{\partial\theta^2}p(x;\theta)\forall\theta\in\Theta$，则

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln p(x;\theta)\right]$$

证明 由于

$$\int p(x;\theta)\mathrm{d}x = 1$$

两边同时对 θ 求偏导得：

$$\frac{\partial}{\partial\theta}\int p(x;\theta)\mathrm{d}x = \int\frac{\partial}{\partial\theta}p(x;\theta)\mathrm{d}x = \int\frac{\partial}{\partial\theta}\ln p(x;\theta)p(x;\theta)\mathrm{d}x = 0$$

再同时对 θ 求偏导得：

$$\int\left[\frac{\partial^2}{\partial\theta^2}\ln p(x;\theta)\right]p(x;\theta)\mathrm{d}x + \int\left[\frac{\partial}{\partial\theta}\ln p(x;\theta)\right]^2 p(x;\theta)\mathrm{d}x = 0$$

即

$$E\left[\frac{\partial^2}{\partial\theta^2}\ln p(x;\theta)\right] + I(\theta) = 0$$

命题 3.4

对于总体分布 $p(x;\theta)$，若其参数 θ 的费希尔信息量存在且为 $I(\theta)$，则对于参数的函数 $\tau(\theta)$ 的费希尔信息量为

$$I(\tau(\theta)) = E\left[\frac{\partial}{\partial\tau(\theta)}\ln p(x;\theta)\right]^2 = \frac{I(\theta)}{[\tau'(\theta)]^2}$$

证明

$$\begin{aligned}I(\tau(\theta)) &= E\left[\frac{\partial}{\partial\tau(\theta)}\ln p(x;\theta)\right]^2 = E\left[\frac{\partial\theta}{\partial\tau(\theta)}\frac{\partial}{\partial\theta}\ln p(x;\theta)\right]^2\\ &= [\frac{\partial\theta}{\partial\tau(\theta)}]^2 E\left[\frac{\partial}{\partial\theta}\ln p(x;\theta)\right]^2 = \frac{I(\theta)}{[\tau'(\theta)]^2}\end{aligned}$$

3.3.4 Cramer-Rao 不等式

定理 3.5 (Cramer-Rao 不等式)

设总体分布 $p(x;\theta)$ 存在费希尔统计量。若对 $g(\theta)$ 的任一个无偏估计T，

$$g(\theta)=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}T\left(x_1,\cdots,x_n\right)\prod_{i=1}^{n}p\left(x_i;\theta\right)\mathrm{d}x_1\cdots\mathrm{d}x_n,\quad\forall\theta\in\Theta$$

的微商 $g'(\theta)=\frac{\mathrm{d}g(\theta)}{\mathrm{d}\theta}$ 存在，并且可在积分号下进行，即

$$\begin{aligned}g'(\theta)&=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}T(x_1,\cdots,x_n)\frac{\partial}{\partial\theta}\left(\prod_{i=1}^{n}p(x_i;\theta)\right)\mathrm{d}x_1\cdots\mathrm{d}x_n\\&=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}T(x_1,\cdots,x_n)\left[\frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n}p(x_i;\theta)\right]\prod_{i=1}^{n}p(x_i;\theta)\mathrm{d}x_1\cdots\mathrm{d}x_n\end{aligned}$$

对离散总体, 则将上述积分改为求和符号后, 等式仍然成立. 则有

$$\operatorname{Var}(T)\geq\frac{1}{nI(g(\theta))}$$

特别地，对 θ 的无偏估计 $\hat{\theta}$，有 $\operatorname{Var}(\hat{\theta})\geq(nI(\theta))^{-1}$。

证明 以连续总体为例：由于

$$\int p(x_i;\theta)\mathrm{d}x_i=1$$

两边同时对 θ 求偏导，由于积分与微分可交换次序，于是有：

$$\begin{aligned}\frac{\partial}{\partial\theta}\int p(x_i;\theta)\mathrm{d}x_i&=\int\frac{\partial}{\partial\theta}p(x_i;\theta)\mathrm{d}x_i=\int\frac{\partial}{\partial\theta}\ln p(x_i;\theta)p(x_i;\theta)\mathrm{d}x_i\\&=E\left[\frac{\partial}{\partial\theta}\ln p=(x_i;\theta=)\right]=0\end{aligned}$$

记

$$Z=\frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n}p(x_i;\theta)=\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln p(x_i;\theta)$$

则

$$E(Z)=\sum_{i=1}^{n}E\left[\frac{\partial}{\partial\theta}\ln p(x_i;\theta)\right]=0$$

从而

$$\begin{aligned}\operatorname{Var}(Z)&=\sum_{i=1}^{n}\operatorname{Var}\left(\frac{\partial}{\partial\theta}\ln p\left(x_i;\theta\right)\right)\\&=\sum_{i=1}^{n}E\left[\frac{\partial}{\partial\theta}\ln p\left(x_i;\theta\right)\right]^2=nI(\theta)\end{aligned}$$

又由

$$g'(\theta)=E(T\cdot Z)=E((T-g(\theta))\cdot Z)$$

据施瓦茨不等式, 有

$$\left[g'(\theta)\right]^2 \leq E\left[(T-g(\theta))^2\right] \cdot E\left(Z^2\right) = \operatorname{Var}(T)\operatorname{Var}(Z)$$

关于离散总体可类似证明。

定义 3.15 (CR 下界)

称 $\frac{1}{nI(g(\theta))}$ 为 $g(\theta)$ 的无偏估计的方差的 **C-R 下界**, 简称 $g(\theta)$ 的 CR 下界。若估计式的方差达到 C-R 下界，则称其为参数的**有效估计**，即 UMVUE。♣

错题记录

1. （茆 6.1.14）设 $x_1,\cdots,x_n$ 是来自二点分布 $b(1,p)$ 的一个样本，证明 $\frac{1}{p}$ 的无偏估计不存在。
2. （茆 6.2.3(2)）设总体分布列为 $P(X=k)=(k-1)\theta^2(1-\theta)^{k-2}, k=2,3,\cdots,0<\theta<1$ 如下,$x_1,x_2,\cdots,x_n$，试求未知参数的矩估计。
3. （茆 6.4.3）
4. （茆 6.4.4）
5. （茆 6.4.5）
6. （茆 6.4.6）
7. （茆 6.4.9）
8. （茆 6.4.11）
9. （茆 6.4.12）
10. （茆 6.4.13）
11. （茆 6.4.14）
12. （茆 6.4.16-20）
13. （茆 6.5.12-16）