ch7 点估计

7.1 引言

一些定义

定义

- 随机变量$X_1,\ldots,X_n$的函数称为统计量（Statistic）
- 统计量T的概率分布称为T的样本分布

定义7.1.1

定义 7.1.1 样本的任何一个函数 $W(X_1,\cdots,X_n)$称为一个点估计量（point-estimator），即任何一个统计量就是一个点估计量.

注意

1. 估计量与估计值的区别在于：估计量是样本的一个函数，估计值是一个估计量的实现值，它是从样本抽取之后的实际观测值得到的。
2. 在记号上，估计量是随机变量$X_1,\ldots,X_n$的一个函数，而估计值是样本观测值$x_1,\ldots,x_n$的函数

7.2 求估计量的方法

7.2.1 矩法

原理

设 X_1，…，X_n 是来自以 $f(x|\theta_1,\cdots,\theta_k)$为其概率密度函数或概率质量函数的总体的样本. 矩法估计量是这样得到的：令前 k 阶的样本矩与相应的前 k 阶总体矩相等，这样就得到一个联立方程组，求解之，就得到矩估计量. 更清楚地说，我们定义

$$
\begin{aligned}
m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i^1, \quad \mu_1' = \mathrm{E}X^1, \\
m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \mu_2' = \mathrm{E}X^2, \\
&\vdots \\
m_k &= \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad \mu_k' = \mathrm{E}X^k.
\end{aligned} \tag{7.2.1}
$$

在典型的情况下，总体矩 μ_j'是参数 θ_1，…，θ_k 的一个函数，可以记作 $\mu_j'(\theta_1,\cdots,\theta_k)$. 于是$(\theta_1,\cdots,\theta_k)$的矩法估计量$(\tilde{\theta}_1,\cdots,\tilde{\theta}_k)$就可以通过求解下面的关于$(\theta_1,\cdots,\theta_k)$的方程组

$$
\begin{aligned}
m_1 &= \mu_1'(\theta_1,\cdots,\theta_k), \\
m_2 &= \mu_2'(\theta_1,\cdots,\theta_k), \\
&\vdots \\
m_k &= \mu_k'(\theta_1,\cdots,\theta_k)
\end{aligned} \tag{7.2.2}
$$

得到.

例

- 设$X_1,\ldots,X_n$是iid的$N(\theta,\sigma^2)$的样本，求θ,σ^2的矩估计?

例

- 设$X_1,\ldots,X_n$是iid的$binomial(k,p)$的样本，求k,p的矩估计?

注意1

这些估计公认地不是总体参数的最佳估计量. 尤其是 k 和 p 有可能取到负的估计值而显然它们必须是正的. （这是估计量的值域与被估计参数的值域不一致的一个实例.）然而公道地讲，我们注意这个矩法的负估计值只发生在样本均值小于样本方差的情况，而这种情况表明数据的变异程度大. 在这个例子中，矩法至少还是给了

我们一套 k 和 p 的候选估计量. 虽然直观有可能给我们一个关于参数 p 的候选估计量，但是要提出关于 k 的一个估计量却是相当困难的. ‖

注意2

矩法对获得统计量分布的近似是非常有用的. 这个技术有时被称为“矩匹配”，它是一种通过匹配分布的矩而给出的近似方法. 从理论上讲，任何一个统计量分布的矩都可以与任何一个分布的矩相匹配，但在实际中，最好采用与之相似的分布.

7.2.2 极大似然法

似然函数的定义

到目前为止，极大似然估计法是最为流行的求估计量的技术. 设 X_1，…，X_n 是来自以 $f(x|\theta_1,\cdots,\theta_k)$为其概率密度函数或概率质量函数的总体的 iid 样本，回忆似然函数的定义

$$(7.2.3)\qquad L(\theta \mid \boldsymbol{x}) = L(\theta_1,\cdots,\theta_k \mid x_1,\cdots,x_n) = \prod_{i=1}^{n} f(x_i \mid \theta_1,\cdots,\theta_k)$$

定义7.2.4

定义 7.2.4 对每一个固定的样本点 $\boldsymbol{x}$，令 $\hat{\theta}(\boldsymbol{x})$是参数 θ 的一个取值，它使得 $L(\theta|\boldsymbol{x})$作为 θ 的函数在该处达到最大值. 那么，基于样本 $\boldsymbol{X}$ 的极大似然估计量 (maximum likelihood estimator 缩写为 MLE) 就是 $\hat{\theta}(\boldsymbol{X})$.

注意，从这个定义本身的构造，就表明 MLE 的值域与参数值域相符合. 我们在谈到这个估计量的实现值的时候，也用缩写 MLE 代表极大似然估计值.

注意

一般求函数最大值的问题中，会有两个固有缺陷，它们在极大似然估计中也是存在的. 第一个问题就是怎样实际求出全局最大值并且验证它确实为最大. 很多情况之下这个问题就归结成一个简单的微分练习，但有的时候，即使对于普通的总体密度，也会产生困难. 第二个问题是数值敏感性. 就是说，对于数据的微小改变，估计值有多么敏感？（严格讲来，作为一个涉及极大化过程的问题，它是统计问题而更是一个数学问题，因为 MLE 就是通过极大化过程求得的. 然而它是我们必须应对的问题.）遗憾的是，有时样本一个微小的变化会使得 MLE 产生巨大的改变，导致对它的使用受到怀疑. 下面，我们先考虑求 MLE 的问题.

原理

- 怎么找 MLE？

如果似然函数是可微的（对于 θ_i），那么 MLE 的可能值就是满足

$$\frac{\partial}{\partial \theta_i} L(\theta|\boldsymbol{x})=0, \quad i=1,\cdots k \tag{7.2.4}$$

的解 $(\theta_1, \cdots, \theta_k)$. 注意，方程（7.2.4）的解仅是 MLE 的可能的选择，这是因为一阶导数为 0 只是成为极大值点的必要而非充分条件. 另外，一阶导数的零点只处于函数定义域内部的极值点上. 如果极值点出现在定义域的边界上，一阶导数未必是 0. 因此，我们必须另外对边界进行核查以发现极值点.

一阶导数为 0 的点有可能是局部或全局的极小点，极大点，也可能是拐点. 我们的工作是求全局最大值点.

另外一种找 MLE 的方法是不用微分法而直接极大化. 这种方法通常在代数上更简单，特别是当求导引起麻烦的情况更是这样，不过有的时候它是较难施行的，这是因为没有一定之规可以遵从. 一般的技术是给似然函数找出一个全局的上界，然后确定一个唯一的点，该点达到了这个上界.

例7.2.5

例 7.2.5（正态 似然） 设 $X_1, \cdots, X_n$ 是 iid $n(\theta,1)$的，用 $L(\theta|\boldsymbol{x})$记它的似然函数，则

$$L(\theta \mid \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} e^{-(1/2)(x_i-\theta)^2} = \frac{1}{(2\pi)^{n/2}} e^{(-1/2)\sum_{i=1}^{n}(x_i-\theta)^2}.$$

化简方程$(d/d\theta)L(\theta|\boldsymbol{x})=0$ 得到

$$\sum_{i=1}^{n}(x_i-\theta)=0$$

它有解 $\hat{\theta}=\bar{x}$. 因此 $\bar{x}$ 是 MLE 的一个可能的选择. 为了验证 $\bar{x}$ 事实上就是似然函数的一个全局最大值点，我们可以这样做：首先注意到 $\bar{x}$ 是 $\sum(x_i-\theta)=0$ 的唯一解，从而 $\bar{x}$ 就是一阶导数的唯一零点. 第二，验证

$$\frac{d^2}{d\theta^2} L(\theta|\boldsymbol{x})\big|_{\theta=\bar{x}}<0$$

这样，$\bar{x}$ 就是唯一的内部极值点而且是一个极大值点. 最后验证 $\bar{x}$ 是一个全局最大值点，这就需要我们核查两个边界，$\pm\infty$. 通过取极限易于得出似然函数在 $\pm\infty$处是 0. 因此 $\bar{x}$ 是一个全局最大值点从而 $\bar{x}$ 是 MLE.（事实上我们也可以些许聪明一点而不必去核查$\pm\infty$. 因为我们已得出 $\bar{x}$ 是唯一的内部极值点而且是极大值点，那么边界点$\pm\infty$就不会是最大值点. 如果它们是的话，就必有内部极小值点，这与唯一性矛盾.）‖

例7.2.6（例7.2.5续)

例 7.2.6（例 7.2.5 续） 回忆（见定理 5.2.4）对于任一个数 a，有

$$\sum_{i=1}^{n}(x_i-a)^2 \geqslant \sum_{i=1}^{n}(x_i-\overline{x})^2$$

等号成立当且仅当 $a=\overline{x}$. 这蕴涵对于任何 θ，

$$e^{-(1/2)\Sigma(x_i-\theta)^2} \leqslant e^{-(1/2)\Sigma(x_i-\overline{x})^2}$$

等号成立当且仅当 $\theta=\overline{x}$. 因此 $\overline{x}$ 是 MLE. ‖

例

- 设$X_1,\ldots,X_n$是iid的$Bernoulli(p)$的，求p的`MLE`？

注意

极大似然估计量的一个有用的性质是所谓不变性（不要与第 6 章论及的那种不变性混淆）. 假定一个分布以一个参数 θ 作为指标，而人们的兴趣在于找到 θ 的某个函数（记为 $\tau(\theta)$）的一个估计量. 非正式地讲，MLE 的不变性说的是如果 $\hat{\theta}$ 是 θ 的 MLE，则 $\tau(\hat{\theta})$ 就是 $\tau(\theta)$ 的 MLE. 例如，如果 θ 是正态分布的均值，那么 sin（θ）的 MLE 就是 $\sin(\overline{X})$. 这里按照 Zehna（1966）的方法来讲述不变性，也可以参阅 Pal and Berry（1992）关于 MLE 的不变性的讲述方法.

定理7.2.10

定理 7.2.10（极大似然估计的不变性） 若 $\hat{\theta}$ 是 θ 的 MLE，则对于 θ 的任何函数 $\tau(\theta)$，$\tau(\hat{\theta})$ 是 $\tau(\theta)$ 的 MLE.

例

- 求解下列问题的MLE

1

▷ Example: $X_1,\cdots,X_n \overset{iid}{\sim} N(\mu,\sigma^2)$. Known σ and Unknown σ

2

▷ Example: $X_{ij},\ i=1,\cdots,s;\ j=1,\cdots,n$ independently distributed as normal distribution with mean μ_i and variance σ^2. Find the mle of μ_i and σ^2.

3

Example (Uniform Distribution). Let $X_1, \ldots, X_n$ be iid with the uniform $(0, \theta)$ density, i.e.,

$$f(x;\theta)=\begin{cases}1/\theta, & 0<x\le\theta\\ 0, & elsewhere\end{cases}$$

Find the $\hat{\theta}_{MLE}$.

7.2.3 Bayes估计量

Bayes方法

在经典方法中，参数 θ 被认为是一个未知、但固定的量. 从以 θ 为指标的总体中抽取一组随机样本 X_1，…，X_n，基于样本的观测值来获得关于 θ 的知识. 在 Bayes 方法中，θ 被考虑成一个其变化可被一个概率分布描述的量，该分布叫做先验分布（prior distribution）. 这是一个主观的分布，建立在试验者的信念上，而且见到抽样数据之前就已经用公式制定好了（因而名为先验分布）. 然后从以 θ 为指标的总体中抽取一组样本，先验分布通过样本信息得到校正. 这个被校正的先验分布叫做后验分布（posterior distribution）. 这个校正工作是通过 Bayes 法则完成的（见第 1 章），因而称为 Bayes 统计.

Bayes估计

如果我们把先验分布记为 $\pi(\theta)$ 而把样本分布记为 $f(\boldsymbol{x}|\theta)$，那么后验分布是给定样本 $\boldsymbol{x}$ 的条件下 θ 的条件分布，就是

(7.2.7) $\quad \pi(\theta|\boldsymbol{x})=f(\boldsymbol{x}|\theta)\pi(\theta)/m(\boldsymbol{x}),\quad (f(\boldsymbol{x}|\theta)\pi(\theta)=f(\boldsymbol{x},\theta))$

这里 $m(\boldsymbol{x})$ 是 $\boldsymbol{X}$ 的边缘分布，由下式得出，

(7.2.8) $$m(\boldsymbol{x})=\int f(\boldsymbol{x}\mid\theta)\pi(\theta)\mathrm{d}\theta$$

注意这个后验分布是一个条件分布，其条件建立在观测样本上. 现在用这个后验分布来作出关于 θ 的推断，而 θ 仍被考虑为一个随机的量. 例如，后验分布的均值就可以被用作 θ 的点估计.

关于记号的注记：当处理一个参数 θ 的分布的时候，我们将打破以往使用大写字母表示随机变量，用小写字母表示自变量的记号习惯. 于是我们可以这样表述：此随机变量 θ 具有分布 $\pi(\theta)$. 这已经成为常见的用法而决不会因此而产生混淆.

例7.2.14

例 7.2.14（二项分布的 Bayes 估计） 设 X_1，…，X_n 是 iid Bernoulli（p）的，则 $Y=\sum X_i$ 是二项分布 binomial（n，p）的. 我们假定 p 的先验分布是贝塔

分布，$p\sim\text{beta}(\alpha,\beta)$. Y 和 p 的联合分布是

$$f(y,p)=\left[\binom{n}{y}p^y(1-p)^{n-y}\right]\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}\right] \quad \begin{pmatrix}\text{条件密度}\times\text{边缘密度}\\ f(y|p)\times\pi(p)\end{pmatrix}$$

$$=\binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{y+\alpha-1}(1-p)^{n-y+\beta-1}$$

Y 的概率密度函数是

$$(7.2.9)\qquad f(y)=\int_0^1 f(y,p)\mathrm{d}p=\binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}$$

这个分布称为贝塔-二项分布（见习题 4.34 和例 4.4.6）. 给定 y 的条件下 p 的分布，即后验分布是

$$f(p|y)=\frac{f(y,p)}{f(y)}=\frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}p^{y+\alpha-1}(1-p)^{n-y+\beta-1}$$

这是 beta$(y+\alpha,n-y+\beta)$分布.（记住这里 p 是变动的而 y 被当作固定的.） p 的一个自然的估计就是这个后验分布的均值，作为 p 的 Bayes 估计量，如下式给出，

$$\hat{p}_B=\frac{y+\alpha}{\alpha+\beta+n}. \qquad \|$$

定义7.2.15

定义 7.2.15 设 $\mathcal{F}$ 是概率密度函数或概率质量函数 $f(x|\theta)$ 的类（以 θ 为指标）. 称一个先验分布类 $\prod$ 为 $\mathcal{F}$ 的一个共轭族（conjugate family），如果对所有的 $f\in\mathcal{F}$，所有的 $\prod$ 中的先验分布和所有的 $x\in X$，其后验分布仍在 $\prod$ 中.

贝塔分布族是二项分布族的共轭族. 这样，如果我们由一个贝塔分布当作先验分布开始，那么我们将以一个贝塔分布的后验分布结束. 先验分布的校正表现为其参数的校正. 在数学上这样是非常方便的，它通常使得计算相当容易. 至于一个共轭族对一个特定的问题是否为一个合理的选择，则是一个留给试验者考虑的问题.

7.2.4 EM算法——了解

我们即将看到的最后一种求估计量的方法在其途径上有其固有的特别之处. 它是专门为寻找 MLE 而设计的. 与其详细讲解这个求解 MLE 的过程，不如在这里详述这样一种算法，它用以保证收敛到 MLE. 这种算法叫做 EM（期望-最大化，Expectation-Maximization）算法. 它基于这样的想法，把一个难于处理的似然函数最大化问题用一个易于最大化的序列取代，而其极限是原始问题的解. 这种算法特别适用于“缺失数据（missing data）”问题，因为存在缺失数据的情况，有时致使计算麻烦. 不过我们将看到，填充这些“缺失数据”常会使计算变得更加光滑.（我们还将看到“缺失数据”有不同的解释——例如见习题 7.30.）

在使用 EM 算法时我们考虑两个不同的似然问题. 我们的兴趣是解“不完全数据（incomplete-data）”问题，而我们实际解的是“完全数据（complete-data）问题”. 届时根据情况我们再决定由哪个问题开始.

7.3 估计量的评价方法

7.3.1 均方误差

定义7.3.1

我们首先研究有限样本时对一个估计量质量的度量，从下面的均方误差开始.

定义 7.3.1 参数 θ 的估计量 W 的均方误差（mean squared error，简记为 MSE）是由 $\mathrm{E}_\theta(W-\theta)^2$ 定义的关于 θ 的函数.

注意，MSE 度量的是估计量 W 与参数 θ 之差的平方的平均值，它是对于一个点估计性质的颇为合理的度量. 一般讲，绝对值距离 $|W-\theta|$ 的任何一个增函数都可以取作一个估计量优度的度量（平均绝对误差 $\mathrm{E}_\theta(|W-\theta|)$ 就是一个合理的选择），但是 MSE 至少有两个优点超过其他的距离度量：第一，它易于解析处理，第二，它有这样一个解释

(7.3.1) $$\mathrm{E}_\theta(W-\theta)^2=\mathrm{Var}_\theta W+(\mathrm{E}_\theta W-\theta)^2=\mathrm{Var}_\theta W+(\mathrm{Bias}_\theta W)^2$$

这里我们所讲一个估计量的偏倚 $\mathrm{Bias}_\theta W$ 是如下定义的：

定义7.3.2

定义 7.3.2 参数 θ 的点估计量 W 的偏倚（bias）是指的 W 的期望值与 θ 之差；即 $\mathrm{Bias}_\theta W=\mathrm{E}_\theta W-\theta$. 一个估计量如果它的偏倚（关于 θ）恒等于 0，则称为无偏的（unbiased），它满足 $\mathrm{E}_\theta W=\theta$ 对所有 θ 成立.

这样，MSE 由两部分组成，其一度量该估计量的变异性（精度）而其二度量它的偏倚（准确度）. 一个估计量具有好的 MSE 性质就在方差与偏倚两项上综合的小. 为求得一个有良好 MSE 性质的估计量我们需要寻找方差与偏倚两者都得到控制的估计量. 显然无偏估计量对控制偏倚再好不过.

对一个无偏估计量，我们有

$$\mathrm{E}_\theta(W-\theta)^2=\mathrm{Var}_\theta W$$

因此，如果一个估计量是无偏的，它的 MSE 就是它的方差.

例

- 设$X_1,\ldots,X_n$是iid的$N(\mu,\sigma^2)$的，证明统计量$\bar{X},S^2$都是无偏估计量

例 7.3.3（正态 MSE） 设 X_1，…，X_n 是 iid $\mathrm{n}(\mu,\sigma^2)$的. 则统计量 $\overline{X}$ 和 S^2 都是无偏估计量. 因为

$$E\overline{X}=\mu, ES^2=\sigma^2, \text{对所有的}\ \mu\ \text{和}\ \sigma^2\ \text{成立}$$

（这一结论不需要正态也是对的，见定理 5.2.6.）这两个估计量的 MSE 是

$$\mathrm{E}(\overline{X}-\mu)^2=\mathrm{Var}\overline{X}=\frac{\sigma^2}{n}$$

$$\mathrm{E}(S^2-\sigma^2)^2=\mathrm{Var}S^2=\frac{2\sigma^4}{n-1}$$

即使去掉正态假定，$\overline{X}$ 的 MSE 仍保持等于 σ^2/n. 但是如果去掉正态假定，S^2 的 MSE 就不再保持等于上边式子了（见习题 5.8）. ‖

例7.3.4（续）

例 7.3.4（例 7.3.3 续） σ^2的估计量的一个选择是极大似然估计量 $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2$. 直接计算得到

$$E\hat{\sigma}^2 = E\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\sigma^2$$

所以 $\hat{\sigma}^2$ 是 σ^2 的一个有偏的估计量. $\hat{\sigma}^2$ 的方差也可以计算如下

$$\mathrm{Var}\hat{\sigma}^2 = \mathrm{Var}\left(\frac{n-1}{n}S^2\right) = \left(\frac{n-1}{n}\right)^2 \mathrm{Var}S^2 = \frac{2(n-1)\sigma^4}{n^2}$$

于是，它的 MSE 是

$$\mathrm{E}(\hat{\sigma}^2 - \sigma^2)^2 = \frac{2(n-1)\sigma^4}{n^2} + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4$$

这样我们有

$$\mathrm{E}(\hat{\sigma}^2 - \sigma^2)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4 < \left(\frac{2}{n-1}\right)\sigma^4 = \mathrm{E}(S^2 - \sigma^2)^2$$

这说明 $\hat{\sigma}^2$ 具有比 S^2 更小的 MSE. 这样用偏倚抵换方差，MSE 得到改善. ‖

注意

这里要及时指出，上面例子并非意指应当放弃把 S^2 作为 σ^2 的一个估计量. 上面的议论是为了表明，如果用 MSE 作为度量的话，在平均的意义下 $\hat{\sigma}^2$ 比 S^2 更靠近 σ^2. 但是 $\hat{\sigma}^2$ 是有偏的而且在平均意义下对 σ^2 估计偏低. 仅这一点就可能令我们对使用 $\hat{\sigma}^2$ 作为 σ^2 的估计量感到不安. 而且我们还可以说，MSE 作为位置参数的准则是合理的而作为尺度参数的准则就不是合理的了，因此上面的比较甚至就不应当去做.（这里有一个问题是：MSE 对于估计偏低和估计偏高是平等处罚的，这在位置参数情况是好的，但是在尺度参数的情况，0 乃是自然下界，所以估计问题不对称. MSE 用于这种情况就有宽恕低估的倾向.）最终结论是这个问题得不到绝对的回答，而对于一个特定情况为了选取一个好的估计量，我们更应为估计量搜集较多的信息.

一般，由于 MSE 是参数的函数，因此将不会有一个“最优的”估计量. 常常两个估计量的 MSE 将出现相互交叉，表明每个估计量仅在参数空间的一部分上是较优的（相对于另外一个）. 不过即使这样的部分信息有时也能够为我们在估计量中作选择提供指导原则.

7.3.2 最佳无偏估计量

介绍

正如上一节所表明的，基于 MSE 的考虑对估计量进行比较未必能产生一个明显的优胜者. 确实是没有一个“最佳 MSE”的估计量. 很多人感觉到这种恼人情况，认为与其对候选估计量进行 MSE 比较工作，还不如有一个“被建议的”为好.

之所以没有一个“最佳 MSE”估计量，是因为全部估计量的类是一个太大的类. （例如取估计量 $\hat{\theta}=17$，那么 $\hat{\theta}$ 在 $\theta=17$ 这一点上的 MSE 是无与伦比的，但在其他情况 $\hat{\theta}$ 却是一个很糟的估计量.）为易于寻找“最佳”估计量，一种解决方法是限制估计量的类. 限制估计量类的常用方法就是本节所考虑的，即仅在无偏估计量的范围内做考虑.

若 W_1 和 W_2 都是参数 θ 的无偏估计量，即 $\mathrm{E}_\theta W_1=\mathrm{E}_\theta W_2=\theta$，则它们的均方误差就等于它们的方差，所以我们应当选择方差比较小的那个估计量. 如果我们能找到一个具有一致最小方差的无偏估计量——最佳无偏估计量——那么我们的任务就完成了.

在开始这个进程之前我们指出，虽然我们要处理无偏估计量，但是这里以及下一节的结论实际上更为一般. 假定有 θ 的一个估计量 W^*，其期望为 $\mathrm{E}_\theta W^*=\tau(\theta)\neq\theta$，而我们感兴趣于研究 W^* 的价值. 考虑估计类

$$\mathcal{C}_\tau=\{W:\mathrm{E}_\theta(W)=\tau(\theta)\}$$

由于对任何 $W_1,W_2\in\mathcal{C}_\tau$，$\mathrm{Bias}_\theta W_1=\mathrm{Bias}_\theta W_2$，于是

$$\mathrm{E}_\theta(W_1-\theta)^2-\mathrm{E}_\theta(W_2-\theta)^2=\mathrm{Var}_\theta W_1-\mathrm{Var}_\theta W_2$$

这样，在类 C_τ 中对 MSE 的比较就可以仅基于对方差的比较. 因此，虽然我们是在用无偏估计量的术语讲话，而实际上是比较具有相同期望 $\tau(\theta)$ 的估计量.

本节的目标是研究求得“最佳”无偏估计量的方法，我们以下面方式定义之.

定义7.3.7

定义 7.3.7 估计量 W^* 称为 $\tau(\theta)$ 的最佳无偏估计量（best unbiased estimator）如果它满足 $\mathrm{E}_\theta W^*=\tau(\theta)$ 对所有 θ 成立，并且对任何一个其他的满足 $\mathrm{E}_\theta(W)=\tau(\theta)$ 的估计量 W，都有 $\mathrm{Var}_\theta W^*\leqslant\mathrm{Var}_\theta W$ 对所有 θ 成立. W^* 也称为 $\tau(\theta)$ 的一致最小方差无偏估计量（uniform minimum variance unbiased estimator，简记 UMVUE).

例

- 设$X_1,\ldots,X_n$是$iidPoisson(\lambda)$的，试证明样本均值$\bar{X}$和样本方差S^2均是λ的无偏估计量?

注意

- UMVUE可能不存在，如果存在，则是唯一的。

怎么找UMVUE

用CRLB

定理7.3.9

定理 7.3.9（Cramér-Rao 不等式） 设 $X_1, \cdots, X_n$ 是具有概率密度函数 $f(\boldsymbol{x}|\theta)$ 的样本，令 $W(\boldsymbol{X})=W(X_1,\cdots,X_n)$ 是任意的一个估计量，满足

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta W(\boldsymbol{X})=\int_{\mathcal{X}}\frac{\partial}{\partial\theta}[W(\boldsymbol{x})f(\boldsymbol{x}\mid\theta)]\mathrm{d}\boldsymbol{x} \tag{7.3.4}$$

和

$$\mathrm{Var}_\theta W(\boldsymbol{X})<\infty$$

则有

$$\mathrm{Var}_\theta(W(\boldsymbol{x}))\geqslant\frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta W(\boldsymbol{X})\right)^2}{\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)^2\right)} \tag{7.3.5}$$

证明 这个定理的证明手法简洁漂亮，它是 Cauchy-Schwarz 不等式的一次聪明的运用，或用统计的语言说，证明利用了这样的事实：对于任意两个随机变量 X 和 Y，有

$$[\mathrm{Cov}(X,Y)]^2\leqslant(\mathrm{Var}X)(\mathrm{Var}Y) \tag{7.3.6}$$

重新安排一下式（7.3.6），我们就可以得到 X 方差的一个下界，

$$\mathrm{Var}X\geqslant\frac{[\mathrm{Cov}(X,Y)]^2}{\mathrm{Var}Y}$$

这个定理证明的聪明之处从 X 和 Y 的选择开始. 把 X 选为估计量 $W(\boldsymbol{X})$ 而 Y 选为 $\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)$，然后应用 Cauchy-Schwarz 不等式.

首先注意有

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta W(\boldsymbol{X})&=\int_{\mathcal{X}}W(\boldsymbol{x})\left[\frac{\partial}{\partial\theta}f(\boldsymbol{x}\mid\theta)\right]\mathrm{d}x\\&=\mathrm{E}_\theta\left[W(\boldsymbol{X})\frac{\frac{\partial}{\partial\theta}f(\boldsymbol{X}\mid\theta)}{f(\boldsymbol{X}\mid\theta)}\right]\quad\text{（前式乘以 } f(\boldsymbol{X}\mid\theta)/f(\boldsymbol{X}\mid\theta)\text{）}\\&=\mathrm{E}_\theta\left[W(\boldsymbol{X})\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}\mid\theta)\right]\quad\text{（对数的性质）}\end{aligned} \tag{7.3.7}$$

这暗示我们应考虑 $W(\boldsymbol{X})$ 与 $\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)$ 之间的协方差，为了使它化为一个协方差，需要减去两期望值的乘积，于是我们来计算 $\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)$. 假如我们在式（7.3.7）中特别地取 $W(\boldsymbol{x})=1$，就得到

$$\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)=\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta[1]=0 \tag{7.3.8}$$

因此 $\mathrm{Cov}_\theta\left(W(\boldsymbol{X}),\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)$ 就等于乘积的期望，于是由式（7.3.7）和式（7.3.8），就得到

$$\mathrm{Cov}_\theta\left(W(\boldsymbol{X}),\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)=\mathrm{E}_\theta\left(W(\boldsymbol{X})\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)=\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta W(\boldsymbol{X}) \tag{7.3.9}$$

同样，因为 $\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)=0$，我们有

$$\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)=\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)^2\right) \tag{7.3.10}$$

对式（7.3.9）和式（7.3.10）一并使用 Cauchy-Schwarz 不等式，我们就得到

$$\mathrm{Var}_\theta(W(\boldsymbol{x}))\geqslant\frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta W(\boldsymbol{X})\right)^2}{\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)^2\right)}$$

定理证毕. ■

推论7.3.10

推论 7.3.10（Cramér-Rao 不等式，iid 情况） 如果定理 7.3.9 的假定满足，而且附加假定 X_1，…，X_n 是 iid 的，具有概率密度函数 $f(x|\theta)$，则

$$\mathrm{Var}_\theta(W(\boldsymbol{X}))\geqslant\frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta W(\boldsymbol{X})\right)^2}{n\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right)}$$

证明 我们只需证明

$$\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)^2\right)=n\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right)$$

因为 X_1，…，X_n 相互独立，因此

$$\begin{aligned}\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}\mid\theta)\right)^2\right)&=\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(X_i\mid\theta)\right)^2\right)\\&=\mathrm{E}_\theta\left(\left(\sum_{i=1}^n\frac{\partial}{\partial\theta}\log f(X_i\mid\theta)\right)^2\right)\qquad\text{（对数的性质）}\\&=\sum_{i=1}^n\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X_i\mid\theta)\right)^2\right)+\qquad\text{（平方展开）}\\&\quad\sum_{i\neq j}^n\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i\mid\theta)\,\frac{\partial}{\partial\theta}\log f(X_j\mid\theta)\right)\end{aligned}\tag{7.3.11}$$

对于 $i\neq j$，我们有

$$\begin{aligned}&\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right)\\=&\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right)\qquad\text{（独立性）}\\=&0\qquad\text{（由式(7.3.8)）}\end{aligned}$$

因此，式（7.3.11）中第二个和式是 0，而第一项是

$$\sum_{i=1}^n\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X_i\mid\theta)\right)^2\right)=n\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X\mid\theta)\right)^2\right)\qquad\text{（同分布）}$$

这样就最后证得了推论. □

Fisher信息量

数量 $\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right)^2\right)$ 叫做样本的信息数（Information number），或 Fisher 信息量（Fisher information）. 这个术语反映这样一个事实，信息量为最佳无偏估计量在 θ 处的方差给出了一个界. 当信息量增大，我们就掌握关于 θ 更多的信息，从而就有一个较小的对于最佳无偏估计方差的界.

事实上，Cramér-Rao 不等式这个术语也可以换成信息不等式（Information Inequality），而信息不等式具有比这里更一般的存在形式. 这种更一般形式与这里的一个关键差别在于有关候选估计量的全部假定都被去掉而换之以基础密度函数上的假定. 在这种形式中，信息不等式对于比较估计量性能变得非常有用. 其中的细节参阅 Lehmann and Casella（1998，2.6 节）.

引理7.3.11

对于任何可微函数 $\tau(\theta)$，我们现在对于任何满足式（7.3.4）且 $\mathrm{E}_\theta W=\tau(\theta)$ 的估计量 W 的方差有一个下界. 这个界仅依赖于 $\tau(\theta)$ 和 $f(x|\theta)$ 并且是方差的一致下界. 任何一个估计量 W 满足 $\mathrm{E}_\theta W=\tau(\theta)$ 而且达到了这个下界，它就是 $\tau(\theta)$ 的一个最佳无偏估计量.

在观看几个例子之前，我们给出一个计算结果，它有助于以上定理的应用，证明留作习题 7.39.

引理 7.3.11 若 $f(x|\theta)$ 满足

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X\mid\theta)\right)=\int\frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log f(x\mid\theta)\right)f(x\mid\theta)\right]\mathrm{d}x$$

（对一个指数族为真），则

$$\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right)=-\mathrm{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right)$$

例

- 设$X_1,\ldots,X_n$是$iid Poisson(\lambda)$的，试证明样本均值$\bar{X}$是λ的最佳无偏估计量?

缺点

记住 Cramér-Rao 定理的关键假定非常重要，这个假定就是可以在积分号下微分，当然它多少有点限制了定理. 如我们已经看到的，指数族密度是满足这个假定的，但一般情况下，这个假定需要经过核对，否则将会出现像下面例子中的矛盾.

例7.3.13

例 7.3.13（均匀分布尺度的无偏估计量） 设 X_1，…，X_n 是 iid 均匀分布的，概率密度函数为 $f(x|\theta)=1/\theta$，$0<x<\theta$. 因为 $\frac{\partial}{\partial\theta}\log f\ (x\mid\theta)=-1/\theta$，所以有

$$\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right)=\frac{1}{\theta^2}$$

Cramér-Rao 定理似乎表明假如 W 是 θ 的任何一个无偏估计量，则有

$$\mathrm{Var}_\theta W\geqslant\frac{\theta^2}{n}$$

而我们现在要找出一个具有更小方差的无偏估计量. 作为首先的猜测，考虑充分统计量 $Y=\max(X_1,\cdots,X_n)$，即最大顺序统计量. Y 的概率密度函数是 $f_Y(y|\theta)=ny^{n-1}/\theta^n$，$0<y<\theta$，于是

$$\mathrm{E}_\theta Y=\int_0^\theta\frac{ny^n}{\theta^n}\,\mathrm{d}y=\frac{n}{n+1}\theta$$

这表明 $\frac{n+1}{n}Y$ 是 θ 的一个无偏估计量. 我们进一步计算

$$\begin{aligned}\mathrm{Var}_\theta\left(\frac{n+1}{n}Y\right)&=\left(\frac{n+1}{n}\right)^2\mathrm{Var}_\theta Y\\&=\left(\frac{n+1}{n}\right)^2\left[\mathrm{E}_\theta Y^2-\left(\frac{n}{n+1}\theta\right)^2\right]\\&=\left(\frac{n+1}{n}\right)^2\left[\frac{n}{n+2}\theta^2-\left(\frac{n}{n+1}\theta\right)^2\right]\\&=\frac{1}{n(n+2)}\theta^2\end{aligned}$$

它一致地小于 θ^2/n. 这表明 Cramér-Rao 定理不适用于这里的概率密度函数. 为了看明这一点，我们可以使用莱布尼茨法则（见 2.4 节）来计算

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\theta}\int_0^\theta h(x)f(x\mid\theta)\mathrm{d}x &= \frac{\mathrm{d}}{\mathrm{d}\theta}\int_0^\theta h(x)\,\frac{1}{\theta}\mathrm{d}x\\ &= \frac{h(\theta)}{\theta}+\int_0^\theta h(x)\,\frac{\partial}{\partial\theta}\Big(\frac{1}{\theta}\Big)\mathrm{d}x\\ &\neq \int_0^\theta h(x)\,\frac{\partial}{\partial\theta}f(x\mid\theta)\mathrm{d}x\end{aligned}$$

除非 $h(\theta)/\theta=0$ 对所有 θ 成立. 因此，Cramér-Rao 定理不适用. 一般地，如果概率密度函数不等于 0 的范围（即支撑集合）依赖于参数，该定理将不适用. ‖

缺点

用这种方法求最佳无偏估计量的一个缺点就是，即使可以使用 Cramér-Rao 定理，也不能保证下界是可达的. 这就是说 Cramér-Rao 下界有可能严格小于任何无偏估计量的方差. 事实上，即便对于常常受到青睐的单参数指数族 $f(x|\theta)$，我们所能讲的最多也不过是存在一个参数 $\tau(\theta)$，它有达到 Cramér-Rao 下界的无偏估计量. 而在其他的典型情况下，对于其他参数，这个下界可能是达不到的. 这些情况之所以引起关注是因为如果我们不能找到一个估计量达到下界，我们就必须确定是不存在达到下界的估计量还是我们必须考查更多的估计量.

例7.3.14

例 7.3.14（正态方差界） 设 $X_1,\cdots,X_n$ 是 iid $\mathrm{n}(\mu,\sigma^2)$的，考虑对 σ^2的估计，这里 μ 未知. 正态的概率密度函数是满足 Cramér-Rao 定理与引理 7.3.11 的假定的，所以我们有

$$\frac{\partial^2}{\partial(\sigma^2)^2}\log\Big(\frac{1}{(2\pi\sigma^2)^{1/2}}\mathrm{e}^{-(1/2)(x-\mu)^2/\sigma^2}\Big)=\frac{1}{2\sigma^4}-\frac{(x-\mu)^2}{\sigma^6}$$

和

$$\begin{aligned}-E\Big(\frac{\partial^2}{\partial(\sigma^2)^2}\log f(X|\mu,\sigma^2)\,\Big|\,\mu,\sigma^2\Big)&=-E\Big(\frac{1}{2\sigma^4}-\frac{(X-\mu)^2}{\sigma^6}\,\Big|\,\mu,\sigma^2\Big)\\ &=\frac{1}{2\sigma^4}\end{aligned}$$

于是，任何一个关于 σ^2的无偏估计量 W 必须满足

$$\mathrm{Var}(W|\mu,\sigma^2)\geqslant\frac{2\sigma^4}{n}$$

在例 7.3.3 中我们看到

$$\mathrm{Var}(S^2|\mu,\sigma^2)=\frac{2\sigma^4}{n-1}$$

所以 S^2 未达到 Cramér-Rao 下界. ‖

推论7.3.15

推论 7.3.15（达到下界） 设 $X_1,\cdots,X_n$ 是 iid 的，具有概率密度函数 $f(x|\theta)$，其 $f(x|\theta)$满足 Cramér-Rao 定理的条件. 令 $L(\theta\mid\boldsymbol{x})=\prod_{i=1}^n f(x_i\mid\theta)$ 表示似然函数. 如果 $W(\boldsymbol{X})=W(X_1,\cdots,X_n)$是 $\tau(\theta)$的任意一个无偏估计量，则 $W(\boldsymbol{X})$ 达

到 Cramér-Rao 下界当且仅当

$$a(\theta)[W(\boldsymbol{x})-\tau(\theta)]=\frac{\partial}{\partial\theta}\log L(\theta|\boldsymbol{x}) \tag{7.3.12}$$

对某一函数 $a(\theta)$ 成立.

证明 Cramér-Rao 不等式，根据式（7.3.6），它能写成

$$\left[\mathrm{Cov}_\theta\left(W(\boldsymbol{X}),\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(X_i\mid\theta)\right)\right]^2\leqslant\mathrm{Var}_\theta W(\boldsymbol{X})\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(X_i\mid\theta)\right)$$

回忆有 $\mathrm{E}_\theta W=\tau(\theta)$，$\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(X_i\mid\theta)\right)=0$，然后运用定理 4.5.7 的结论，我们就得到等号成立当且仅当 $W(\boldsymbol{x})-\tau(\theta)$ 和 $\frac{\partial}{\partial\theta}\prod_{i=1}^n f(x_i\mid\theta)$ 成比例. 而这就正是式（7.3.12）所表示的. ■

问题

本节所展开论述的理论仍然留下一些未回答的问题. 第一，如果 $f(x|\theta)$ 不满足 Cramér-Rao 定理的假定，我们能做什么？（在例 7.3.13 中，我们仍然不知道 $\frac{n+1}{n}Y$ 是否为一个最佳无偏估计量.）第二，如果下界不能被允许的估计量达到，像例 7.3.14 那样，将如何呢？在该例中我们仍然不知道 S^2 是否为一个最佳无偏估计量.

回答这些问题的一种途径是寻找适用范围更广，产生更大的下界的方法. 在这个题目上已经有了很多研究，也许最为著名的就是 Chapman-Robbins（1951）下界. Stuart，Ort and Arnold（1999，17 章）对这个题目有一个很好的处理. 这里我们不采用这个方法，而是从另外一个观点继续对最佳无偏估计量的研究，这就是利用充分性的概念.

用充分性、无偏性和Rao-Blackwell

定理7.3.17

定理 7.3.17（Rao-Blackwell） 设 W 是 $\tau(\theta)$ 的任意一个无偏估计量，而 T 是关于 θ 的一个充分统计量. 定义 $\phi(T)=\mathrm{E}(W|T)$. 则 $\mathrm{E}_\theta\phi(T)=\tau(\theta)$ 而且 $\mathrm{Var}_\theta\phi(T)\leqslant\mathrm{Var}_\theta W$ 对所有 θ 成立；即是说 $\phi(T)$ 是 $\tau(\theta)$ 的一个一致较优的无偏估计量.

证明 由式（7.3.13），我们有

$$\tau(\theta)=\mathrm{E}_\theta W=\mathrm{E}_\theta[\mathrm{E}(W|T)]=\mathrm{E}_\theta\phi(T)$$

所以 $\phi(T)$ 对 $\tau(\theta)$ 是无偏的. 而且

$$\begin{aligned}\mathrm{Var}_\theta W&=\mathrm{Var}_\theta[\mathrm{E}(W|T)]+\mathrm{E}_\theta[\mathrm{Var}(W|T)]\\&=\mathrm{Var}_\theta\phi(T)+\mathrm{E}_\theta[\mathrm{Var}(W|T)]\qquad(\mathrm{Var}(W|T)\geqslant0)\\&\geqslant\mathrm{Var}_\theta\phi(T)\end{aligned}$$

因此 $\phi(T)$ 一致地优于 W，现在只剩下证明 $\phi(T)$ 的确是一个估计量. 即，我们必须证明 $\phi(T)=\mathrm{E}(W|T)$ 仅是样本的函数，且特别地，它独立于 θ. 而根据充分性的定义以及 W 仅是样本的函数这一事实，就可以推出 $W|T$ 的分布独立于 θ. 所以 $\phi(T)$ 是 $\tau(\theta)$ 的一个一致较优的无偏估计量. ■

例7.3.18

例 7.3.18（给定以一个非充分的统计量为条件） 设 X_1，X_2 是 iid n$(\theta,1)$的. 统计量 $\overline{X}=\frac{1}{2}(X_1+X_2)$具有

$$\mathrm{E}_\theta\overline{X}=\theta \quad 和 \quad \mathrm{Var}_\theta\overline{X}=\frac{1}{2}$$

考虑以 X_1 为条件，它不是充分的. 设 $\phi(X_1)=\mathrm{E}_\theta(\overline{X}|X_1)$. 由式 (7.3.13)，则有 $\mathrm{E}_\theta\phi(X_1)=\theta$ 和 $\mathrm{Var}_\theta\phi(X_1)\leqslant\mathrm{Var}_\theta\overline{X}$，所以 $\phi(X_1)$优于 $\overline{X}$. 但是，

$$\begin{aligned}\phi(X_1)&=\mathrm{E}_\theta(\overline{X}|X_1)\\&=\frac{1}{2}\mathrm{E}_\theta(X_1|X_1)+\frac{1}{2}\mathrm{E}_\theta(X_2|X_1)\\&=\frac{1}{2}X_1+\frac{1}{2}\theta\end{aligned}$$

上边第三个等号是根据独立性，有 $\mathrm{E}_\theta(X_2|X_1)=\mathrm{E}_\theta X_2$. 所以 $\phi(X_1)$不是估计量. ‖

我们现在知道了，为求 $\tau(\theta)$的一个最佳无偏估计量，我们只需考虑基于一个充分统计量的估计量. 现在出现的问题是，假如我们有 $\mathrm{E}_\theta\phi=\tau(\theta)$而且 ϕ 基于一个充分统计量，即 $\mathrm{E}(\phi|T)=\phi$，我们怎么能知道 ϕ 是最佳无偏的？当然，如果 ϕ 达到 Cramér-Rao 下界，则它就是最佳无偏的，但如果它未达到，那我们又获得了什么呢？例如 ϕ^* 是 $\tau(\theta)$的一个无偏估计量，怎样比较 $\mathrm{E}(\phi^*|T)$与 ϕ？下一个定理通过证明一个最佳无偏估计量是唯一的而部分地回答了这个问题.

定理7.3.19

定理 7.3.19 如果 W 是 $\tau(\theta)$的一个最佳无偏估计量，则 W 是唯一的.

证明 假如 W'是另一个最佳无偏估计量，考虑估计量 $W^*=\frac{1}{2}(W+W')$. 注意到 $\mathrm{E}_\theta W^*=\tau(\theta)$并且

$$\begin{aligned}\mathrm{Var}_\theta W^*&=\mathrm{Var}_\theta\left(\frac{1}{2}W+\frac{1}{2}W'\right)\\&=\frac{1}{4}\mathrm{Var}_\theta W+\frac{1}{4}\mathrm{Var}_\theta W'+\frac{1}{2}\mathrm{Cov}_\theta(W,W') \qquad (习题\ 4.44)\\&\leqslant\frac{1}{4}\mathrm{Var}_\theta W+\frac{1}{4}\mathrm{Var}_\theta W'+\frac{1}{2}[(\mathrm{Var}_\theta W)(\mathrm{Var}_\theta W')]^{1/2}\\&\qquad (\text{Cauchy-Schwarz 不等式})\\&=\mathrm{Var}_\theta W \qquad (\mathrm{Var}_\theta W=\mathrm{Var}_\theta W')\end{aligned} \tag{7.3.14}$$

但如果以上不等式是严格的，则与 W 的最佳无偏性矛盾，所以上边式子必须对所有 θ 都是等式. 因为上边的不等式是 Cauchy-Schwarz 不等式的一个运用，所以只有在 $W'=a(\theta)W+b(\theta)$时才有等号成立. 现在使用协方差的性质，我们有

$$\begin{aligned}\mathrm{Cov}_\theta(W,W')&=\mathrm{Cov}_\theta[W,a(\theta)W+b(\theta)]\\&=\mathrm{Cov}_\theta[W,a(\theta)W]\\&=a(\theta)\mathrm{Var}_\theta W\end{aligned}$$

但是由于在式 (7.3.14) 中等号成立，从而 $\mathrm{Cov}_\theta(W,W')=\mathrm{Var}_\theta W$. 所以 $a(\theta)=1$，而且由于 $\mathrm{E}_\theta W'=\tau(\theta)$，因此一定有 $b(\theta)=0$，于是 $W=W'$，这就证明了 W 是唯一的. ■

定理7.3.20

定理 7.3.20 如果 $\mathrm{E}_\theta W=\tau(\theta)$，$W$ 是 $\tau(\theta)$ 的最佳无偏估计量当且仅当 W 与 0 的所有无偏估计量不相关.

证明 假如 W 是最佳无偏的，根据上面的讨论 W 必须满足 $\mathrm{Cov}_\theta(W,U)=0$ 对所有 θ 及任意满足 $\mathrm{E}_\theta U=0$ 的 U 都成立，因此必要性得以确立.

假定我们现在有一个无偏估计量 W，它与 0 的所有无偏估计量不相关. 设 W' 是任意一个满足 $\mathrm{E}_\theta W'=\mathrm{E}_\theta W=\tau(\theta)$ 的估计量. 我们要证明 W 优于 W'. 写成

$$W'=W+(W'-W)$$

然后计算

$$\begin{aligned}\mathrm{Var}_\theta W' &= \mathrm{Var}_\theta W+\mathrm{Var}_\theta(W'-W)+2\mathrm{Cov}_\theta(W,W'-W)\\ &= \mathrm{Var}_\theta W+\mathrm{Var}_\theta(W'-W)\end{aligned} \tag{7.3.15}$$

最后一个等式成立是因为 $W'-W$ 是 0 的一个无偏估计量以及根据假定，它和 W 不相关. 因为 $\mathrm{Var}_\theta(W'-W)\geqslant 0$，式 (7.3.15) 蕴涵 $\mathrm{Var}_\theta W'\geqslant \mathrm{Var}_\theta W$. 由于 W' 是任意的，所以由此得出 W 是 $\tau(\theta)$ 的最佳无偏估计量. ■

注意，0 的一个无偏估计量无异于随机噪声 (random noise)，就是说在 0 的一个估计量里没有信息. (有理由认定，使用 0 而不是随机噪声作为估计 0 的方法是最为明智的.) 因此，如果一个估计量能够通过加上随机噪声而被改善，这个估计量可能是有缺陷的. (或许，我们可以质询评价估计量的准则，但在这里的情况下，正是从该准则得到上述疑问的.) 这种直觉在定理 7.3.20 中被正式化了.

注意

注意，0 的一个无偏估计量无异于随机噪声 (random noise)，就是说在 0 的一个估计量里没有信息. (有理由认定，使用 0 而不是随机噪声作为估计 0 的方法是最为明智的.) 因此，如果一个估计量能够通过加上随机噪声而被改善，这个估计量可能是有缺陷的. (或许，我们可以质询评价估计量的准则，但在这里的情况下，正是从该准则得到上述疑问的.) 这种直觉在定理 7.3.20 中被正式化了.

虽然我们现在有了对最佳无偏估计量的一个有趣的刻画，但是它在应用上还是受限制的. 验证一个估计量与 0 的所有无偏估计量不相关常常是一件困难的工作，这是因为通常难以描述出 0 的所有无偏估计量. 然而，对于确定一个无偏估计量不是最佳无偏的，它有时却是有用的.

为了回答关于最佳无偏估计量的问题，需要的是关于 0 的所有无偏估计量的某一特征描述. 给出这样一个特征描述，我们就能够看出我们的最佳无偏估计量的候选者是不是最佳的.

描述 0 的无偏估计量的特征不是一件容易的工作，需要对当前的概率密度函数 (或概率质量函数) 附加条件. 注意在本节中直至现在我们还没有对概率密度函数附加条件 (例如像在 Cramér-Rao 下界中需要的那样)，而为这个一般性付出的代价就是验证最佳无偏估计量存在性时的困难.

如果一个概率密度函数或者概率质量函数族 $f(x|\theta)$ 具有这样性质，它没有 0 的无偏估计量 (0 本身除外)，那么我们的寻找工作就可结束了，因为任何统计量 W 都满足 $\mathrm{Cov}_\theta(W,0)=0$. 回忆在定义 6.2.21 中定义了完全性所具有的性质，它就保证了这样一种情况.

例

- 对于$X_1,\ldots,X_n$是$iidU(0,\theta)$的，令$Y=max\{X_1,\ldots,X_n\}$，证明$\frac{n+1}{n}Y$是θ的完全充分统计量，也是一个无偏估计量，也是最佳无偏估计量

注意

注意，这里重要的地方就是充分统计量的分布族的完全性，而原始族的完全性却是无关紧要的. 这是根据 Rao-Blackwell 定理推出的，它说明我们可以把注意力集中在充分统计量的函数上，于是所有期望都将相对于它的分布取得.

我们把完全性与最佳无偏性的关系概括总结为以下定理.

定理7.3.23

定理 7.3.23 设 T 是一个参数 θ 的完全充分统计量而 $\phi(T)$ 是任意的一个仅基于 T 的估计量. 则 $\phi(T)$ 是其期望值的唯一最佳无偏估计量.

我们用上述理论的一个有趣且有用的应用来结束这一节. 在很多情形下，没有明显的候选者当作 $\tau(\theta)$ 的无偏估计量，更不必说最佳无偏估计量的候选者了. 然而在完全性的面前，本节的理论告诉我们如果能找到任意的一个无偏估计量，我们就能找到那个最佳无偏估计量. 若 T 是参数 θ 的一个完全充分统计量，$h(X_1,\cdots,X_n)$ 是 $\tau(\theta)$ 的任意一个无偏估计量，则 $\phi(T)=\mathrm{E}(h(X_1,\cdots,X_n)\mid T)$ 是 $\tau(\theta)$ 的最佳无偏估计量（见习题 7.56）.

7.3.4 损失函数最优性

原理介绍

前面我们对点估计量的评价是基于它们的均方误差的. 均方误差是所谓损失函数（loss function）的特例. 通过损失函数研究估计量的性能与最优性是判决理论的一个分支.

当数据 $\boldsymbol{X}=\boldsymbol{x}$ 被观测到之后，这里 $X\sim f(x\mid\theta)$，$\theta\in\Theta$，就做出一个关于 θ 的判决. 容许判决的集合是行为空间（action space），记为 A. 在点估计问题中 A 常常等于 Θ，即参数空间，但在其他问题上（像假设检验—参见 8.3.5 节）这点会有改变.

损失函数在点估计问题里反映了这样的事实，如果一个行为 a 靠近 θ，则 a 是合理的且遭受小的损失. 如果 a 远离 θ，则遭受大的损失. 损失函数是一个非负函数，一般它随 a 与 θ 的距离增加而增加. 如果 θ 是实值的，两个常用损失函数是

绝对误差损失（absolute error loss），$L(\theta,a)=|a-\theta|$

和

平方误差损失（squared error loss），$L(\theta,a)=(a-\theta)^2$

这两个损失函数都随着 θ 与 a 的距离增加而增加，最小值是 $L(\theta,\theta)=0$. 就是说如果行为正确，损失最小. 平方误差损失对大的偏差给予相对更多的惩罚，而绝对误差损失给予小偏差相对更多的惩罚. 平方损失有一个变种，它对高估比低估给予更多的惩罚，如下所示

$$L(\theta,a)=\begin{cases}(a-\theta)^2 & \text{若 } a<\theta\\ 10(a-\theta)^2 & \text{若 } a\geqslant\theta\end{cases}$$

另一种损失函数是相对平方损失

$$L(\theta,a)=\frac{(a-\theta)^2}{|\theta|+1}$$

它在 θ 接近于 0 时对误差的惩罚要比 $|\theta|$ 较大时的惩罚大. 注意，基于绝对误差的损失也可以有类似的变种. 一般，试验者必须考虑到对不同 θ 值的估计的误差不同带来的后果并且指定一种能反应这种后果的损失函数.

在一个损失函数或判决理论分析中，一个估计量的质量被它的风险函数（risk function）量化；即，对 θ 的估计量 $\delta(\boldsymbol{x})$，其风险函数是 θ 的一个函数，定义为

(7.3.16) $$R(\theta,\delta)=\mathrm{E}_\theta L(\theta,\delta(\boldsymbol{X}))$$

在一给定 θ 处，风险函数就是假如使用估计量 $\delta(\boldsymbol{X})$ 的话，将遭受的平均损失.

因为 θ 的真值是未知的，我们愿意用一个对所有 θ 值都有小 $R(\theta,\delta)$ 值的估计量. 这将意味着不管 θ 的真值如何，该估计量将有小的期望损失. 如果要比较两个不同估计量 δ_1 与 δ_2 的质量，就可以通过比较它们的风险函数 $R(\theta,\delta_1)$ 与 $R(\theta,\delta_2)$ 来进行. 如果 $R(\theta,\delta_1)<R(\theta,\delta_2)$ 对所有 θ 成立，则 δ_1 是我们优先选用的统计量，因为 δ_1 对所有的 θ 表现得都更佳. 更具代表性的情况是两个风险函数交叉的情况. 这时判断哪个估计量更好可能就不这样鲜明了.

一个估计量 δ 的风险函数是期望损失，即如式（7.3.16）所定义. 对于平方误差损失，它的风险函数是个熟悉的数量，就是在 7.3.1 节使用的均方误差（MSE）. 那里一个估计量的 MSE 定义为 $\mathrm{MSE}(\theta)=\mathrm{E}_\theta(\delta(\boldsymbol{X})-\theta)^2$，它就是 $\mathrm{E}_\theta L(\theta,\delta(\boldsymbol{X}))=R(\theta,\delta)$，其中 $L(\theta,a)=(a-\theta)^2$. 如本章前边所得出，对于平方损失函数，

(7.3.17) $$R(\theta,\delta)=\mathrm{Var}_\theta\,\delta(\boldsymbol{X})+(\mathrm{E}_\theta\,\delta(\boldsymbol{X})-\theta)^2=\mathrm{Var}_\theta\,\delta(\boldsymbol{X})+(\mathrm{Bias}_\theta\,\delta(\boldsymbol{X}))^2$$

平方损失的风险函数清楚地指明一个好的估计量应当同时具有小的方差与小的偏倚. 判决理论分析要裁决的是一个估计量在同时成功最小化这两个量上有多好.

如果像我们在 7.3.2 节中所做的那样，把所容许的估计量的集合 $\mathcal{D}$ 限制到无偏估计类上，这将是一种非典型的判决分析. 这时最小化风险刚好就是最小化方差. 判决分析要比之更广泛，方差与偏倚都要在风险中，并且将被同时考虑. 如果一个估计量具有小的、但可能非零的偏倚同时结合以一个小的方差，则可能被判为好估计量.

总结

- Data: $\mathbf{X}$
- Model(Distribution): $f(\mathbf{x}|\boldsymbol{\theta})$, $\boldsymbol{\theta}\in\Theta$
- Action space: $\mathcal{A}$
  Point estimation: $\mathcal{A}=\Theta$
  Testing: $\mathcal{A}$=$\{$Reject H_0, Accept $H_0\}$
- Loss function: $L(\boldsymbol{\theta},a)$
- Decision rule: $\delta(\mathbf{x})$: Sample space $\to\mathcal{A}$
- Risk function: Expected loss

$$R(\boldsymbol{\theta},\delta)=E\left[L(\boldsymbol{\theta},\delta(\mathbf{X}))\right]=\int L(\boldsymbol{\theta},\delta(\mathbf{x}))f(\mathbf{x}|\boldsymbol{\theta})dx$$

- Goal: Find $\delta(\mathbf{x})$ that has small risk somehow.

Definition

A real valued function $L(\boldsymbol{\theta}, a)$ satisfying

1. $L(\boldsymbol{\theta}, a) \geq 0$ for all θ, a
2. $L(\boldsymbol{\theta}, a) = 0$ for $a = \theta$

is called a *loss function* of the action a.

Definition

Let $\delta(\mathbf{X})$ be an estimator of a parametric function $\tau(\boldsymbol{\theta})$. Then

$$R(\boldsymbol{\theta}, \delta) = E\left[L(\boldsymbol{\theta}, \delta(\mathbf{X}))\right]$$

is called the *risk function* of $\delta(\mathbf{X})$ in estimating $\tau(\boldsymbol{\theta})$.

1. An estimator $\delta_1(\mathbf{X})$ is said to be at least *as good as* another estimator $\delta_2(\mathbf{X})$ if

   $$R(\theta, \delta_1(\mathbf{X})) \leq R(\theta, \delta_2(\mathbf{X}))$$

   for all $\theta \in \Theta$.
2. An estimator $\delta_1(\mathbf{X})$ is *better than* $\delta_2(\mathbf{X})$ if

   $$R(\theta, \delta_1(\mathbf{X})) \leq R(\theta, \delta_2(\mathbf{X}))$$

   for all $\theta \in \Theta$ and

   $$R(\theta, \delta_1(\mathbf{X})) < R(\theta, \delta_2(\mathbf{X}))$$

   for at least one $\theta \in \Theta$.

例子

1. Squared error loss

   $$L(\theta, \delta(\mathbf{X})) = (\delta(\mathbf{X}) - \theta)^2,\ R(\boldsymbol{\theta}, \delta) = E\left[(\delta(\mathbf{X}) - \theta)^2\right]$$

2. Absolute error loss

   $$L(\theta, \delta(\mathbf{X})) = |\delta(\mathbf{X}) - \theta|,\ R(\boldsymbol{\theta}, \delta) = E\left[|\delta(\mathbf{X}) - \theta|\right]$$

3. Stein's loss

   $$L(\theta, \delta(\mathbf{X})) = \frac{\delta(\mathbf{X})}{\theta} - 1 - \ln\left(\frac{\delta(\mathbf{X})}{\theta}\right)$$

minmax估计量

An estimator $\delta(\mathbf{X})$ is called a *minimax estimator* if

$$\max_{\theta\in\Theta} R(\theta,\delta(\mathbf{X})) \le \max_{\theta\in\Theta} R(\theta,\tilde{\delta}(\mathbf{X}))$$

for all other estimator $\tilde{\delta}(\mathbf{X})$.

bayes方法处理损失函数最优化

我们也可以使用 Bayes 方法处理损失函数最优化的问题，此处我们假定有一个先验分布 $\pi(\theta)$. 在 Bayes 分析中，我们要利用这个先验分布来计算一个平均风险

$$\int_{\Theta} R(\theta,\delta)\pi(\theta)\mathrm{d}\theta$$

此即为 Bayes 风险（Bayes risk）. 我们可以用这个平均风险函数来评估一个估计量在一个给定的损失函数之下的表现. 进一步，我们还可尝试去求那个具有最小的 Bayes 风险值的估计量. 这样的估计量叫做关于先验分布 π 的 Bayes 法则（Bayes rule），常记作 δ^{π}.

求关于一个给定先验 π 的 Bayes 判决法则看起来可能像是一个吓人的任务，但实际上是相当机械的，就像下面所示的那样.（下面给出的求 Bayes 法则的方法在更为一般的情况仍然适用，参阅 Brown and Purves 1973.）

设 $\boldsymbol{X}\sim f(\boldsymbol{x}|\theta)$，$\theta\sim\pi$，一个判决法则 δ 的 Bayes 风险可以写为

$$\int_{\Theta} R(\theta,\delta)\pi(\theta)\mathrm{d}\theta = \int_{\Theta}\left(\int_{\chi} L(\theta,\delta(\boldsymbol{X}))f(\boldsymbol{x}\mid\theta)\mathrm{d}\boldsymbol{x}\right)\pi(\theta)\mathrm{d}\theta$$

由于 $f(\boldsymbol{x}|\theta)\pi(\theta)=\pi(\theta|\boldsymbol{x})m(\boldsymbol{x})$，这里 $\pi(\theta|\boldsymbol{x})$ 是 θ 的后验分布而 $m(\boldsymbol{x})$ 是 $\boldsymbol{X}$ 的边缘分布，则我们可以把 Bayes 风险写成

$$\int_{\Theta} R(\theta,\delta)\pi(\theta)\mathrm{d}\theta = \int_{\chi}\left[\int_{\Theta} L(\theta,\delta(\boldsymbol{X}))\pi(\theta\mid\boldsymbol{x})\mathrm{d}\theta\right]m(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

方括号的值是损失函数关于后验分布的期望，叫做后验期望损失（posterior expected loss）. 它仅是 $\boldsymbol{x}$ 的函数而非 θ 的函数. 这样对每个 $\boldsymbol{x}$，如果我们选择行为 $\delta(\boldsymbol{x})$ 去极小化后验期望损失，也就极小化了 Bayes 风险.

注意我们现在有了一个构造 Bayes 法则的既定做法. 对一个给定的观测 $\boldsymbol{x}$，Bayes 法则应该极小化后验期望损失. 这一点与我们以前各节中已讲的任何一个传统方法都不同. 例如，考虑前面讨论的求最佳无偏估计量的方法. 为了应用定理 7.3.23，首先我们需要找出一个完全充分统计量 T. 然后我们需要找出一个函数 $\phi(T)$，它是参数的一个无偏估计量. Rao-Blackwell 定理，即定理 7.3.17 在我们知道参数的某个无偏估计量时可能会有帮助. 但是如果我们不能设想出某个无偏估计量的时候，该方法并未告诉我们如何去构造.

即使后验期望损失的极小化不能解析地做出，也能用数值方法计算出积分并实施极小化. 事实上，观测完 $\boldsymbol{X}=\boldsymbol{x}$，我们需要做的极小化是仅针对这个特别 $\boldsymbol{x}$ 的. 不过对某些问题我们是可以用显式描述其 Bayes 法则的.

7.4 习题

1

7.1 对具有概率质量函数 $f(x|\theta)$（这里 $\theta\in\{1,2,3\}$）的离散型随机变量 X 进行一次观测. 求 θ 的 MLE.

x	$f(x\|1)$	$f(x\|2)$	$f(x\|3)$
0	$\frac{1}{3}$	$\frac{1}{4}$	0
1	$\frac{1}{3}$	$\frac{1}{4}$	0
2	0	$\frac{1}{4}$	$\frac{1}{4}$
3	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{2}$
4	$\frac{1}{6}$	0	$\frac{1}{4}$

2

7.6 设 $X_1,\cdots,X_n$ 是来自概率密度函数为下式的一组随机样本

$$f(x|\theta)=\theta x^{-2},0<\theta\leqslant x<\infty$$

(a) 关于 θ 的充分统计量是什么?

(b) 求 θ 的 MLE.

(c) 求 θ 的矩估计量.

3

7.7 设 $X_1,\cdots,X_n$ 是 iid 的，具有两种概率密度函数. 如果 $\theta=0$，则

$$f(x|\theta)=\begin{cases}1 & \text{当 } 0<x<1\\ 0 & \text{其他}\end{cases}$$

而如果 $\theta=1$，则

$$f(x|\theta)=\begin{cases}1/(2\sqrt{x}) & \text{当 } 0<x<1\\ 0 & \text{其他}\end{cases}$$

求 θ 的 MLE.

4

7.8 X 是来自一正态总体 $n(0,\sigma^2)$ 的一次观测.

(a) 求 σ^2 的一个无偏估计量.

(b) 求 σ 的 MLE.

(c) 讨论 σ 的矩估计量的求法.

5

7.9 设 X_1，…，X_n 是 iid 的，具有概率密度函数

$$f(x|\theta)=\frac{1}{\theta},\quad 0\leqslant x\leqslant\theta,\quad \theta>0$$

用矩法和极大似然法估计 θ. 计算两种估计量的均值与方差. 哪一个应该被优先选用，为什么？

6

7.10 设独立随机变量 X_1，…，X_n 具有共同的分布

$$P(X_i\leqslant x|\alpha,\beta)=\begin{cases}0 & \text{当 } x<0\\ (x/\beta)^{\alpha} & \text{当 } 0\leqslant x\leqslant\beta\\ 1 & \text{当 } x>\beta\end{cases}$$

其中参数 α，β 为正.

(a) 求一个关于（α，β）的二维充分统计量.

(b) 求 α，β 的极大似然估计.

(c) 在篱雀的巢中找到的杜鹃蛋的长度（单位：mm）可以用这个分布建模. 根据数据 22.0，23.9，20.9，23.8，25.0，24.0，21.7，23.8，22.8，23.1，23.1，23.5，23.0，23.0，求 α 和 β 的 MLE.

7

7.24 设 X_1，…，X_n 是 iid Poisson(λ)的，并设 λ 服从伽玛分布 gamma(α,β)，即 Poisson 分布的共轭族.

(a) 求 λ 的后验分布.

(b) 计算后验均值和后验方差.

8

7.37 设 X_1，…，X_n 是来自概率密度函数为下式的一组随机样本：

$$f(x|\theta)=\frac{1}{2\theta},\quad -\theta<x<\theta,\quad \theta>0$$

若存在的话则求出 θ 的一个最佳无偏估计量.

9

7.44 设 X_1，…，X_n 是 iid n(θ,1)的. 证明：θ^2 的最佳无偏估计量是 $\overline{X}^2-(1/n)$. 计算它的方差（利用 3.6 节的 Stein 恒等式）并且证明它大于 Cramér-Rao 下界.

7.48 设 X_1，…，X_n 是 iid Bernoulli（p）的.

（a）证明 p 的 MLE 的方差达到 Cramér-Rao 下界.

（b）对 $n \geqslant 4$，证明：乘积 $X_1X_2X_3X_4$ 是 p^4 的一个无偏估计量，并且利用此事实求 p^4 的最佳无偏估计量.