4.3 点估计的优良性准则

从前节的例子中我们累累看到:同一个参数往往有不止一种看来都合理的估计法.因此,自然会提出其优劣比较的问题.

初一看觉得这个问题很容易回答:设$\hat{\theta}_1$和$\hat{\theta}_2$两个估计量都用于估计θ,则看哪一个的误差小,就哪一个为优.但是,由于θ本身未知,就不知道估计误差有多大,这还不是最主要的.主要问题在于:$\hat{\theta}_1$,$\hat{\theta}_2$之值都与样本有关.一般情况是:对某些样本,$\hat{\theta}_1$的误差小于$\hat{\theta}_2$的误差,而对另一些样本则反之.一个从整体上看不好的估计,在个别场合下可能表现很好.反之,一个很不错的估计,由于抽到了不易出现的样本,其表现也可以很差.如例1.2估计学生学习成绩(以其考分衡量)的问题,大家都会同意:如抽出100个学生,以其平均成绩作为估计值,比以抽出的第一个学生的成绩作为估计值要好.但也可以发生这种情况:所抽第一个学生的成绩很接近于全校总平均,而100个学生的平均成绩反而与这个总平均有较大差距.

由此可见,在考虑估计量的优劣时,必须从某种整体性能去衡量它,而不能看它在个别样本之下的表现如何.这里所谓"整体性能",有两种意义:一是指估计量的某种特性,具有这种特性就是好的,否则就是不好的.如下文要讲的"无偏性",即属于此类.二是指某种具体的数量性指标.两个估计量,指标小者为优.如下文讲到的"均方误差",即属于此类.应当注意的是:这种比较,归根到底,也还是相对性的.具有某种特性的估计是否一定就好?这在一定

程度上要看问题的具体情况，不是绝对的．下文在讲述无偏估计时还会涉及这一点，作为比较准则的数量性指标，也可以有很多种．很有可能：在甲指标之下 $\hat{\theta}_1$ 优于 $\hat{\theta}_2$，而在乙指标下则反之．

我们这样说，当然不是认为优良性准则和估计量的优劣比较毫无意义．相反，这些很有意义，且是参数估计这个分支学科研究的中心问题．我们是想提醒读者，不要把这些准则绝对化了．每种准则在某种情况下都有其局限性．

4.3.1 估计量的无偏性

设某统计总体的分布包含未知参数 $\theta_1,\cdots,\theta_k$，$X_1,\cdots,X_n$ 是从该总体中抽出的样本，要估计 $g(\theta_1,\cdots,\theta_k)$．$g$ 为一已知函数．设 $\hat{g}(X_1,\cdots,X_n)$ 是一个估计量．如果对任何可能的 $(\theta_1,\cdots,\theta_k)$ 都有

$$E_{\theta_1,\cdots,\theta_k}[\hat{g}(X_1,\cdots,X_n)] = g(\theta_1,\cdots,\theta_k) \tag{3.1}$$

则称 $\hat{g}$ 是 $g(\theta_1,\cdots,\theta_k)$ 的一个无偏估计量．记号 $E_{\theta_1,\cdots,\theta_k}$ 是指：求期望值时，是在各样本 $X_1,\cdots,X_n$ 的分布中的参数为 $\theta_1,\cdots,\theta_k$ 时去做的．比如，我说 X_1,X_2 是取自正态总体 $N(\theta,1)$ 的样本，让计算和 X_1+X_2 的期望值．这要看参数值 θ 等于多少：$\theta=1$ 时，期望值为 2；$\theta=2.5$ 时，期望值为 5．标出 E_θ，就明白显示是在哪个 θ 值之下去算期望值，也表示 θ 值可以流动．这在定义 3.1 式中尤其有意义．因为在参数估计问题中，我们并不知参数的真值，它能在一定范围内流动．如废品率 p，可在 $[0,1]$ 内流动．当比较两个估计量时，需要对种种可能的参数值去比较．故在 $E_{\theta_1,\cdots,\theta_k}$ 这个记号中强调指出 $(\theta_1,\cdots,\theta_k)$ 以及其可以流动，是重要的．在不致引起混淆时，我们也可以简写为 E．

估计量的无偏性有两个含义．第一个含义是没有系统性的偏差，不论你用什么样的估计量 $\hat{g}$ 去估计 g，总是时而（对某些样本）偏低，时而（对另一些样本）偏高．无偏性表示，把这些正负偏差在

概率上平均起来,其值为0.比如用一把秤去秤东西,误差来源有二:一是秤本身结构制作上的问题,使它在秤东西时,倾向于给出偏高或偏低之值,这属于系统误差.另一种是操作上和其他随机性原因,使秤出的结果有误差,这属于随机误差.在此,无偏性的要求相应于秤没有系统误差,但随机误差总是存在.因此,无偏估计不等于在任何时候都给出正确无误的估计.

另一个含义是由定义(3.1)结合大数定理(见第三章定理4.1)引伸出来的.设想每天把这个估计量$\hat{g}(X_1,\cdots,X_n)$用一次,第i天的样本记为$\hat{g}(X_1^{(i)},\cdots,X_n^{(i)})$,$i=1,2,\cdots,N,\cdots$.则按大数定理,当$N\to\infty$时,各次估计值的平均,即$\sum_{i=1}^{N}\hat{g}(X_1^{(i)},\cdots,X_n^{(i)})/N$,依概率收敛到被估计的值$g(\theta_1,\cdots,\theta_k)$.所以,若估计量有无偏性,则在大量次数使用取平均时,能以接近于100%的把握无限逼近被估计的量.如果没有无偏性,则无论使用多少次,其平均也会与真值保持一定距离——这距离就是系统误差.

由此可见,估计量的无偏性是一种优良的性质.但是,在一个具体的问题中,无偏性的实际价值如何,还必须结合这问题的具体情况去考察.如在秤东西那个例中,若你经常去这家商店买东西而该店用的秤是无系统误差的.这等于说,店里在秤上显示的重量,是你所买的东西的真实重量的无偏估计,则尽管在具体某一次购买中店里可能少给或多给了你一些,从长期平均看,无偏性保证了双方都不吃亏.在此,无偏性有很现实的意义.

现在设想另一种情况:工厂每周进原料一批.在投入使用前,由实验室对原料中某些成分含量的百分率p作一估计,根据估计值$\hat{p}$采取相应的工艺调整措施.无论$\hat{p}$比真正的p偏高或偏低,都会有损于产品质量.在此,即使$\hat{p}$是p的无偏估计,在长期使用中,估计的正负偏差的效应并不能抵消.这样$\hat{p}$的无偏性就不见得很有实用意义了.

例3.1 设$X_1,\cdots,X_n$是从某总体中抽出的样本,则样本均

值$\overline{X}$是总体分布均值θ的无偏估计.

这是因为,按定义,每个样本X_i的分布,与总体分布一样,因此其均值$E(X_i)$就是θ,而

$$E(\overline{X}) = \sum_{i=1}^{n} E(X_i)/n = n\theta/n = \theta$$

据此可知:在正态总体$N(\mu,\sigma^2)$中用$\overline{X}$估计μ,在指数分布总体中用$\overline{X}$估计$1/\lambda$,在二项分布总体中用$\overline{X}/N$估计p,以及在波哇松分布总体中用$\overline{X}$估计λ等,都是无偏估计.

例 3.2 由(1.1)式定义的样本方差S^2,是总体分布方差σ^2的无偏估计.

为证明这一点,以a记总体分布均值:$E(X_i)=a$. 也有$E(\overline{X})=a$,把$X_i-\overline{X}$写为$(X_i-a)-(\overline{X}-a)$,有

$$\sum_{i=1}^{n}(X_i-\overline{X})^2 = \sum_{i=1}^{n}[(X_i-a)-(\overline{X}-a)]^2$$

$$= \sum_{i=1}^{n}(X_i-a)^2 - 2(\overline{X}-a)\sum_{i=1}^{n}(X_i-a) + n(\overline{X}-a)^2$$

注意到$\sum_{i=1}^{n}(X_i-a) = n(\overline{X}-a)$,有

$$\sum_{i=1}^{n}(X_i-\overline{X})^2 = \sum_{i=1}^{n}(X_i-a)^2 - n(\overline{X}-a)^2$$

因$a=E(X_i)=E(\overline{X})$,有

$$E(X_i-a)^2 = \mathrm{Var}(X_i) = \sigma^2, i = 1,\cdots,n$$

$$E(\overline{X}-a)^2 = \mathrm{Var}(\overline{X}) = \sum_{i=1}^{n}\mathrm{Var}(X_i)/n^2 = n\sigma^2/n^2 = \sigma^2/n$$

于是得到

$$E(S^2) = \frac{1}{n-1}E\Big(\sum_{i=1}^{n}(X_i-\overline{X})^2\Big) = \frac{1}{n-1}(n\sigma^2 - n\cdot\sigma^2/n) = \sigma^2$$

这就说明了S^2是σ^2的无偏估计.

这就解释了为什么要在样本二阶中心矩$m_2 = \sum_{i=1}^{n}(X_i -$

$\overline{X})^2/n$ 的基础上，把分母 n 修正为 $n-1$ 以得到 S^2. 这与以前讲过的一点也相合：在第二章的附录 B 中我们曾讲到 $\sum_{i=1}^{n}(X_i-\overline{X})^2$ 的自由度为 $n-1$. 这正好是正确的除数，这件事不是一个巧合.

在这里我们还可以对“自由度”这个概念赋予另一种解释：一共有 n 个样本，有 n 个自由度. 用 S^2 估计方差 σ^2，自由度本应为 n. 但总体均值 a 也未知，用 $\overline{X}$ 去估计之，用掉了一个自由度，故只剩下 $n-1$ 个自由度.

如果总体均值 a 已知，则不用 S^2 而用 $\sum_{i=1}^{n}(X_i-a)^2/n$ 去估计总体方差 σ^2（在 a 未知时不能用）. 这是 σ^2 的无偏估计，分母为 n 不用改为 $n-1$. 因为此处 n 个自由度全保留下了（a 已知，不用估计，没有用去自由度）.

例 3.3 由上例易推知：用 S 去估计总体分布的标准差 σ（方差 σ^2 的正平方根），不是无偏估计. 事实上，据第三章(2.2)式及上例的结果，有

$$\sigma^2 = E(S^2) = \mathrm{Var}(S) + (ES)^2$$

由于方差总非负：$\mathrm{Var}(S)\geqslant 0$，有 $\sigma\geqslant E(S)$. 因而 $E(S)\leqslant\sigma$. 即如用 S 去估计 σ，总是系统地偏低. 在一些情况下，可以通过简单的调整达到无偏估计. 办法是把 S 乘上一个大于 1 的、与样本大小 n 有关的因子 c_n，得 c_nS. 适当选择 c_n 可以使 $E(c_nS)=c_nE(S)=\sigma$. 对正态分布总体 $N(\mu,\sigma^2)$ 而言，不难证明（习题 21）

$$c_n=\sqrt{\frac{n-1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)\Big/\Gamma\left(\frac{n}{2}\right) \tag{3.2}$$

由 $E(S)\leqslant\sigma$ 看出：在例 2.3 中给出的均匀分布 $R(\theta_1,\theta_2)$ 中 θ_1，θ_2 的估计量(2.2)，即使把 m_2 改成 S^2，也是有偏的（$\hat{\theta}_1$ 偏高，$\hat{\theta}_2$ 偏低）. 可以证明（习题 22）：能找到常数 c_n，使 $\overline{X}-c_nS$ 和 $\overline{X}+c_nS$ 分别是 θ_1,θ_2 的无偏估计，但 c_n 的具体数值不易定出来.

例 3.4 我们已经知道：矩估计不必是无偏的，极大似然估计也如此. 事实上，在例 2.7 中，我们已求出：正态总体 $N(\mu,\sigma^2)$ 的

方差 σ^2 的极大似然估计，就是样本二阶中心矩 m_2，而我们已知后者不是无偏的. 再看一个例子：例 2.9 中我们找出均匀分布 $R(0,\theta)$ 中 θ 的极大似然估计是 $\theta^* = \max(X_1,\cdots,X_n)$. 不用计算即知 θ^* 偏低. 因为，每个样本 X_i 都在 $(0,\theta)$ 内，故其最大值，即 θ^*，也在这个区间内. 下面通过计算 $E_\theta(\theta^*)$ 证明这一点，并找出调整因子 c_n，此例对下面还有用.

先算 θ^* 的分布函数 $G(x,\theta)$. 因为 $0<\theta^*<\theta$，有

$$G(x,\theta) = 0, \text{当 } x \leqslant 0; G(x,\theta) = 1, \text{当 } x \geqslant \theta$$

若 $0<x<\theta$，则为了事件 $\{\theta^* \leqslant x\}$ 发生，必须 $\{X_1 \leqslant x\},\cdots,\{X_n \leqslant x\}$ 这 n 个事件同时发生. 由于各样本独立，且都有均匀分布 $R(0,\theta)$，有 $P(X_i \leqslant x) = x/\theta$，因而

$$G(x,\theta) = (x/\theta)^n$$

对 x 求导数，得到 θ^* 的概率密度函数为

$$g(x,\theta) = nx^{n-1}/\theta^n, \text{当 } 0 < x < \theta; \text{此外为 } 0 \qquad (3.3)$$

由此得到

$$E_\theta(\theta^*) = \int_0^\theta xg(x,\theta)\mathrm{d}x = n\int_0^\theta x^n \mathrm{d}x/\theta^n = \frac{n}{n+1}\theta \qquad (3.4)$$

看出以 θ^* 估计 θ 系统偏低，且 $\frac{n+1}{n}\theta^*$ 为 θ 的无偏估计.

4.3.2 最小方差无偏估计

一个参数往往有不止一个无偏估计，从这些众多的无偏估计中，我们想挑出那个最优的. 这牵涉到两个问题：一是为优良性制定一个准则，二是在已定的准则之下，如何去找到最优者. 这涉及较深的理论问题，许多内容都超出本课程范围之外，这里我们只能作一个很初步的介绍.

1. 均方误差，设 $X_1,\cdots,X_n$ 是从某一带参数 θ 的总体中抽出的样本，要估计 θ. 若我们采用估计量 $\hat{\theta} = \hat{\theta}(X_1,\cdots,X_n)$，则其误差为 $\hat{\theta}(X_1,\cdots,X_n) - \theta$. 这误差随样本 $X_1,\cdots,X_n$ 的具体值而定，也是随机的，因而其本身无法取为优良性指标. 我们把它平方以消

除符号，得$(\hat{\theta}(X_1,\cdots,X_n)-\theta)^2$，然后取它的均值，即取

$$M_{\hat{\theta}}(\theta)=E_\theta[\hat{\theta}(X_1,\cdots,X_n)-\theta)]^2 \tag{3.5}$$

作为$\hat{\theta}$的误差大小从整体角度的一个衡量.这个量愈小，就表示$\hat{\theta}$的误差平均讲比较小，因而也就愈优.$M_{\hat{\theta}}(\theta)$就称为估计量θ的“均方误差”(误差平方的平均).不言而喻，均方误差小并不能保证$\hat{\theta}$在每次使用时一定给出小的误差.它有时也可以有较大的误差，但这种情况出现的机会较少.

用均方误差的观点就容易回答前面提到过的一个问题：用100个学生的平均成绩作为全校学生平均成绩的估计，比用抽出的第一个学生的成绩去估计好.事实上，这两个估计分别是$\overline{X}=(X_1+\cdots+X_{100})/100$和$X_1$.总体分布为正态$N(\mu,\sigma^2)$.$\overline{X}$和$X_1$的均方误差分别为

$$E(\overline{X}-\mu)^2=\sigma^2/100, E(X_1-\mu)^2=\sigma^2$$

故X_1的均方误差是$\overline{X}$的100倍.

均方误差并不是唯一可供选择的准则.例如，平均绝对误差$E_\theta|\hat{\theta}(X_1,\cdots,X_n)-\theta|$，以及其他许多别的准则，看来都很合理且在某些场合下还确有其优点，但是，由于平方这个函数在数学上最易处理，使这个准则成为一切准则中应用和研究得最多的.

按第三章(2.2)式，有

$$M_{\hat{\theta}}(\theta)=\mathrm{Var}_\theta(\hat{\theta})+[E_\theta(\hat{\theta})-\theta]^2 \tag{3.6}$$

即均方误差由两部分构成：一部分是$\mathrm{Var}_\theta(\hat{\theta})$，即$\hat{\theta}$的方差，表示$\hat{\theta}$自身变异的程度，另一部分中，$E_\theta(\hat{\theta})-\theta$表示$\hat{\theta}$这个估计量的系统偏差.如果$\hat{\theta}$为$\theta$的无偏估计，则第二项为0，而这时有

$$M_{\hat{\theta}}(\theta)=\mathrm{Var}_\theta(\hat{\theta}) \tag{3.7}$$

2. 最小方差无偏估计.从前面的讨论看到：若局限于无偏估计的范围，且采用均方误差的准则，则两个无偏估计$\hat{\theta}_1$和$\hat{\theta}_2$的比较，归结为其方差的比较：方差小者为优.

例 3.5 设 $X_1,\cdots,X_n$ 是从均匀分布总体 $R(0,\theta)$ 中抽出的样本. 在例 3.4 中已指出过 θ 的两个无偏估计: $\hat{\theta}_1=2\overline{X}$, $\hat{\theta}_2=\frac{n+1}{n}\max(X_1,\cdots,X_n)$. 有(参看第三章,例 2.5)

$$\mathrm{Var}_\theta(\hat{\theta}_1)=4\mathrm{Var}_\theta(\overline{X})=\frac{4}{n}\mathrm{Var}_\theta(X_1)=\frac{4}{n}\frac{1}{12}\theta^2=\frac{\theta^2}{3n}$$

为计算 $\hat{\theta}_2$ 的方差,仍以 θ^* 记 $\max(X_1,\cdots,X_n)$. 按 θ^* 的密度函数(3.3),得

$$E_\theta(\theta^*)=\frac{n}{n+1}\theta,E_\theta(\theta^{*2})=n\int_0^\theta x^{x+1}\mathrm{d}x/\theta^n=\frac{n}{n+2}\theta^2$$

因此

$$\mathrm{Var}_\theta(\theta^*)=E_\theta(\theta^{*2})-[E_\theta(\theta^*)]^2=\frac{n}{(n+1)^2(n+2)}\theta^2$$

而

$$\mathrm{Var}_\theta(\hat{\theta}_2^*)=\left(\frac{n+1}{n}\right)^2\mathrm{Var}_\theta(\theta^*)=\frac{1}{n(n+2)}\theta^2$$

当 $n>1$ 时,总有 $n(n+2)>3n$. 故除非 $n=1$, $\hat{\theta}_2$ 的方差总比 $\hat{\theta}_1$ 的方差为小,且这一点不论未知参数 θ 取什么值都对. 因此,在"方差小者为优"这个准则下, $\hat{\theta}_2$ 优于 $\hat{\theta}_1$,当 $n=1$ 时, $\hat{\theta}_1$ 与 $\hat{\theta}_2$ 重合.

如果 $\hat{\theta}$ 是 θ 的一个无偏估计,且它的方差对 θ 的任何可能取的值,都比任何其他的无偏估计的方差为小,或至多等于它,则在"方差愈小愈好"这个准则下, $\hat{\theta}$ 就是最好的,它称为 θ 的"最小方差无偏估计",简记为 MVU 估计*.

定义 3.1 设 $\hat{\theta}$ 为 $g(\theta)$ 之无偏估计. 若对 $g(\theta)$ 的任何一个无偏估计 $\hat{\theta}_1$ 都有

$$\mathrm{Var}_\theta(\hat{\theta})\leqslant\mathrm{Var}_\theta(\hat{\theta}_1)$$

---

* MVU 是"最小方差无偏"的英语 Minimum Variance Unbiased 的缩写.

对 θ 的任何可能取的值都成立,则称$\hat{\theta}$为$g(\theta)$的一个最小方差无偏估计(MVU 估计).

从例 3.5 知$\hat{\theta}_2$ 的方差小于$\hat{\theta}_1$ 的方差.但我们并不能由此就肯定$\hat{\theta}_2$ 就是 θ 的 MVU 估计,因为也可能还存在其他的无偏估计,其方差比$\hat{\theta}_2$ 的更小.那么,怎样去寻找 MVU 估计呢? 在数理统计学中给出了一些方法,我们只能简略地介绍其中的一个.这个方法的思想如下:先研究一下,在 $g(\theta)$的一切无偏估计中,方差最小能达到多少呢? 如果我们求出了这样一个方差的下界,则如某个估计$\hat{\theta}$的方差达到这个下界,那它必定就是 MVU 估计.

3. 求 MVU 估计的一种方法:克拉美-劳不等式.

我们只考虑单参数的情况.设总体的概率密度函数或概率函数 $f(x,\theta)$只包含一个参数,$X_1,\cdots,X_n$ 为从该总体中抽出的样本,要估计 $g(\theta)$.记

$$I(\theta)=\int\left[\left(\frac{\partial f(x,\theta)}{\partial\theta}\right)^2\Big/f(x,\theta)\right]\mathrm{d}x \tag{3.8}$$

这里积分的范围为 x 可取的范围.例如,对指数分布总体,$0<x<\infty$,对正态总体则 $-\infty<x<\infty$.如果总体分布是离散的,则(3.8)改为

$$I(\theta)=\sum_i\left(\frac{\partial f(a_i,\theta)}{\partial\theta}\right)^2\Big/f(a_i,\theta) \tag{3.9}$$

这里求和 $\sum_i$ 遍及总体的全部可能值$a_1,a_2,\cdots$.为确定计,我们下面就连续型的情况去讨论.对离散型的情况,只须作相应的修改,有如把(3.8)修改为(3.9).

克拉美-劳不等式:在一定的条件下,对 $g(\theta)$的任一无偏估计 $\hat{g}=\hat{g}(X_1,\cdots,X_n)$,有

$$\mathrm{Var}_\theta(\hat{g})\geqslant(g'(\theta))^2/(nI(\theta)) \tag{3.10}$$

n 是样本大小.

这个不等式给出了 $g(\theta)$的无偏估计的方差的一个下界,即

(3.10)式右边. 如果 $g(\theta)$的某个无偏估计其方差正好达到了(3.10)右端,则它就是 $g(\theta)$的 MVU 估计,这不等式的成立有一定的条件. 实际上,在其表述中,就包含了要求$\partial f(x,\theta)/\partial\theta$ 和 $g'(\theta)$存在的条件,其他的条件将在下文推导中看出.

记

$$S = S(X_1,\cdots,X_n,\theta) = \sum_{i=1}^{n}\partial\log f(X_i,\theta)/\partial\theta$$

$$= \sum_{i=1}^{n}\frac{\partial f(X_i,\theta)}{\partial\theta}\Big/f(X_i,\theta)$$

因为 $f(x,\theta)$为密度,有$\int f(x,\theta)\mathrm{d}x = 1$. 两边对 θ 求导,并假定(这就是条件之一)左边求导可搬到积分号内,有

$$\int\frac{\partial f(x,\theta)}{\partial\theta}\mathrm{d}x = 0$$

因此

$$E_\theta\left[\int\frac{\partial f(X_i,\theta)}{\partial\theta}\Big/f(X_i,\theta)\right]$$

$$= \int\left(\frac{\partial f(x,\theta)}{\partial\theta}\Big/f(x,\theta)\right)f(x,\theta)\mathrm{d}x$$

$$= \int\left(\frac{\partial f(x_i,\theta)}{\partial\theta}\right)\mathrm{d}x = 0 \qquad (3.11)$$

于是,由 $X_1,\cdots,X_n$ 的独立性,有

$$\mathrm{Var}_\theta(S) = \sum_{i=1}^{n}\mathrm{Var}_\theta\left(\frac{\partial f(X_i,\theta)}{\partial\theta}\Big/f(X_i,\theta)\right)$$

$$= \sum_{i=1}^{n}E_\theta\left[\frac{\partial f(X_i,\theta)}{\partial\theta}\Big/f(X_i,\theta)\right]^2$$

$$= n\int\left[\frac{\partial f(x,\theta)}{\partial\theta}\Big/f(x_i,\theta)\right]^2 f(x,\theta)\mathrm{d}x = nI(\theta)$$

按第三章定理 3.1 的 2°,有

$$[\mathrm{Cov}_\theta(\hat{g},S)]^2 \leqslant \mathrm{Var}_\theta(\hat{g})\mathrm{Var}_\theta(S) = nI(\theta)\mathrm{Var}_\theta(\hat{g}) \qquad (3.12)$$

由(3.11)有 $E_\theta(S)=0$. 按第三章(3.2)式,有

$$\begin{aligned}\mathrm{Cov}_\theta(\hat{g},S)&=E_\theta(\hat{g}S)\\&=\int\cdots\int\hat{g}(x_1,\cdots,x_n)\sum_{i=1}^{n}\left[\frac{\partial f(x_i,\theta)}{\partial\theta}\Big/f(x_i,\theta)\right]\\&\quad\cdot\prod_{i=1}^{n}f(x_i,\theta)\cdot\mathrm{d}x_1\cdots\mathrm{d}x_n\end{aligned}$$

由乘积的导数公式可知

$$\begin{aligned}&\sum_{i=1}^{n}\left[\frac{\partial f(x_i,\theta)}{\partial\theta}\Big/f(x_i,\theta)\right]\prod_{i=1}^{n}f(x_i,\theta)\\&=\frac{\partial f(x_1,\theta)\cdots f(x_n,\theta)}{\partial\theta}\end{aligned}$$

以此代入上式,并假定对 θ 求偏导数可移至积分号外面(这又是一个条件!),则得

$$\mathrm{Cov}_\theta(\hat{g},S)=\frac{\partial}{\partial\theta}\int\cdots\int\hat{g}(x_1,\cdots,x_n)f(x_1,\theta)\cdots f(x_n,\theta)\mathrm{d}x_1\cdots\mathrm{d}x_n$$

但上式右边的积分就是 $E_\theta(\hat{g})$,因 $\hat{g}$ 为 $g(\theta)$ 的无偏估计,这积分就是 $g(\theta)$. 故上式右边为 $g'(\theta)$,因而得到 $\mathrm{Cov}_\theta(\hat{g},S)=g'(\theta)$,以此代入(3.12),即得(3.10).

不等式(3.10)是瑞典统计学家 H. 克拉美和印度统计学家 C. R. 劳在 1945—1946 年各自独立得出的,故文献中一般称为克拉美—劳不等式. 这个不等式在数理统计学中有多方面的应用,此处求 MVU 估计是其中之一.

顺便提一下:(3.10)中 $I(\theta)$ 这个量的表达式(3.8),最初是英国统计学家 R.A. 费歇尔在 20 年代提出的,后人称之为"费歇尔信息量". 此量出现在(3.10)中,并非偶然的巧合. 从(3.10)我们可以对为什么把 $I(\theta)$ 称为"信息量"获得一点直观的理解:$I(\theta)$ 愈大,(3.10)式中的下界愈低,表示 $g(\theta)$ 的无偏估计更有可能达到较小的方差——即更有可能被估计得更准确一些. $g(\theta)$ 是通过样本去估计的,$g(\theta)$ 能估得更准,表示样本所含的信息量愈大. 一共有 n 个样本,如把总信息量说成是(3.10)右边的分母 $nI(\theta)$,则

一个样本正好占有信息量 $I(\theta)$，$I(\theta)$这个量在数理统计学中很重要，有多方面的应用，但大多超出本课程的范围.

不等式(3.10)并不直接给出找 MVU 估计的方法. 它的使用方式是：先要由直观或其他途径找出一个可能是最好的无偏估计，然后计算其方差，看是否达到了(3.10)式右端的界限，若达到了，就是 MVU 估计. 同时，还得仔细验证不等式推导过程中所有的条件是否全满足，这有时是不大容易的，在以下诸例中，我们都略去了这步验证.

例 3.6 设 $X_1,\cdots,X_n$ 为抽自正态总体 $N(\theta,\sigma^2)$的样本，σ^2 已知(因而只有一个参数 θ)，要估计 θ. 本例

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

因而

$$I(\theta)=(\sqrt{2\pi}\sigma)^{-1}\int_{-\infty}^{\infty}\frac{1}{\sigma^4}(x-\theta)^2\exp\left[-\frac{1}{2\sigma}(x-\theta)\right]^2\mathrm{d}x$$

$$=\frac{1}{\sigma^4}\sigma^2=\frac{1}{\sigma^2}$$

故按不等式(3.10)，θ 的无偏估计的方差，不能小于 σ^2/n. 而$\overline{X}$是 θ 的一个无偏估计，方差正好是 σ^2/n，故$\overline{X}$就是θ 的 MVU 估计.

虽然我们是在 σ^2 已知的条件下证得$\overline{X}$为 θ 的 MVU 估计，但不难推知，这个结论当 σ^2 未知时也对. 证明留给读者(习题 23).

例 3.7 指数分布的费歇尔信息量 $I(\lambda)$为

$$I(\lambda)=\int_0^{\infty}\left(\frac{1}{\lambda}-x\right)^2\lambda\mathrm{e}^{-\lambda x}\mathrm{d}x=\lambda^{-2}$$

故若要由大小为 n 的样本去估计总体均值 $g(\lambda)=1/\lambda$，则按(3.10)，$1/\lambda$ 的无偏估计的方差不能小于

$$[g'(\lambda)]^2/(nI(\lambda))=1/(n\lambda^2)$$

而样本均值$\overline{X}$是 $1/\lambda$ 的一无偏估计，方差正好为 $1/(n\lambda^2)$. 故$\overline{X}$是 $1/\lambda$ 的 MVU 估计.

例 3.8 回到例 3.6. 若均值 θ 已知而要估计方差，则不难证

明：$\sum_{i=1}^{n}(X_i-\theta)^2/n$ 是σ^2 的 MVU 估计，计算留给读者(在计算费歇尔信息量时，注意要把.σ^2 作为一个整体看.可以引进新参数 $\lambda=\sigma^2$再计算).

如果 θ,σ^2 都未知而要估计 σ^2，则可以证明：样本方差 S^2 为 σ^2 的 MVU 估计，但这个证明已超出本方法的范围之外.

例 3.9 为估计均匀分布 $R(0,\theta)$中的参数 θ，在例 3.5 中引进过两个无偏估计$\hat{\theta}_1=2\overline{X}$和$\hat{\theta}_2=\frac{n+1}{n}\max(X_1,\cdots,X_n)$，并证明了$\hat{\theta}_2$ 优于$\hat{\theta}_1$.事实上可以证明：$\hat{\theta}_2$ 就是 θ 的 MVU.但这个结论不能利用不等式(3.10)去证明.这是因为总体的密度函数并非 θ 的连续函数.它有一个间断点：$\theta=x$(注意：是把 $f(x,\theta)$中的 x 固定，作为 θ 的函数时的间断点)，故导数$\partial f(x,\theta)/\partial\theta$ 非处处存在.证明$\hat{\theta}_2$ 为 θ 的 MVU 估计要用另外的方法，此处不能讲了.

下面举一个离散型总体的例子.

例 3.10 总体分布为二项分布 $B(N,p)$，概率函数为

$$f(x,p)=\binom{N}{x}p^x(1-p)^{N-x},x=0,1,\cdots,N$$

由此算出费歇尔信息量(按(3.9)式)

$$I(p)=\frac{1}{p^2(1-p)^2}\sum_{x=0}^{N}(x-Np)^2\binom{N}{x}p^x(1-p)^{N-x}$$

右边这个和不是别的，正是总体方差，故这个和等于 $Np(1-p)$(第三章例 2.2).因此

$$I(p)=Np^{-1}(1-p)^{-1}$$

按(3.10)，p 的无偏估计(基于大小为 n 的样本)的方差，不能小于 $p(1-p)/(nN)$.现$\overline{X}/N$ 为p 之一无偏估计，其方差为

$$\begin{aligned}(\overline{X}\text{ 的方差})/N^2&=\text{总体方差}/(nN^2)=Np(1-p)/(nN)^2\\&=p(1-p)/(nN)\end{aligned}$$

因此，$\overline{X}/N$ 就是p 的 MVU 估计.

特别当 $N=1$ 时，得出："用频率估计概率"，是 MVU 估计.在

例 2.13 中，我们曾求出 p 的贝叶斯估计(2.14)，并指出过它与频率这个估计比，可能有某些优点. 这就看出："最小方差无偏"这个准则也不是绝对的.

例 3.11 仿例 3.10 可以证明：在波哇松分布 $P(\lambda)$ 的总体中估计 λ，$\overline{X}$ 是 MVU 估计. 证明留给读者.

4.3.3 估计量的相合性与渐近正态性

1. 相合性. 在第三章中我们曾证明大数定理. 这个定理说：若 $X_1, X_2, \cdots, X_n, \cdots$ 独立同分布，其公共均值为 θ. 记 $\overline{X}_n = \sum_{i=1}^{n} X_i/n$，则对任给 $\varepsilon>0$，有

$$\lim_{n\to\infty} P(|\overline{X}_n - \theta| \geqslant \varepsilon) = 0 \tag{3.13}$$

(在证明这个定理时假定了 X_i 的方差存在有限. 但我们曾指出：方差存在的条件并非必要).

现在我们可以从估计的观点对(3.13)作一个解释. 我们把 $X_1, X_2, \cdots, X_n$ 看作是从某一总体中抽出的样本. 抽样的目的是估计该总体的均值 θ. 概率 $P(|\overline{X}_n - \theta| \geqslant \varepsilon)$ 是："当样本大小为 n 时，样本均值 $\overline{X}_n$ 这个估计与真值 θ 的偏离达到 ε 这么大或更大"的可能性. (3.13)表明：随着 n 的增加，这种可能性愈来愈小以至趋于 0. 这就是说，只要样本大小 n 足够大，用样本均值去估计总体均值，其误差可以任意小. 在数理统计学上，就把 $\overline{X}_n$ 称为是 θ 的"相合估计". 字面的意思是：随着样本大小的增加，被估计的量与估计量逐渐"合"在一起了.

相合性的一般定义就是这个例子的引伸：

定义 3.2 设总体分布依赖于参数 $\theta_1, \cdots, \theta_k$，$g(\theta_1, \cdots, \theta_k)$ 是 $\theta_1, \cdots, \theta_k$ 之一给定函数. 设 $X_1, X_2, \cdots, X_n$ 为自该总体中抽出的样本，$T(X_1, \cdots, X_n)$ 是 $g(\theta_1, \cdots, \theta_k)$ 的一个估计量. 如果对任给 $\varepsilon>0$ 有

$$\lim_{n\to\infty} P_{\theta_1, \cdots, \theta_k}(|T(X_1, \cdots, X_n) - g(\theta_1, \cdots, \theta_k)| \geqslant \varepsilon) = 0 \tag{3.14}$$

而且这对$(\theta_1,\cdots,\theta_k)$一切可能取的值都成立，则称 $T(X_1,\cdots,X_n)$ 是 $g(\theta_1,\cdots,\theta_k)$的一个相合估计.

记号 $P_{\theta_1,\cdots,\theta_k}$的意义，表示概率是在参数值为$(\theta_1,\cdots,\theta_k)$时去计算的(参看前面关于记号 $E_{\theta_1,\cdots,\theta_k}$的说明).在讲述大数定理时我们曾引进过“依概率收敛”的术语.使用这个术语，相合性可简单地描述为:如果当样本大小无限增加时，估计量依概率收敛于被估计的值，则称该估计量是相合估计.

相合性是对一个估计量的最基本的要求.如果一个估计量没有相合性，那么，无论样本大小多大，我们也不可能把未知参数估计到任意预定的精度.这种估计量显然是不可取的.

如同样本均值的相合性那样，常见的矩估计量的相合性，都可以基于大数定理得到证明.我们再以用二阶中心矩 $m_2(n)=\sum_{i=1}^{n}(X_i-\overline{X}_n)^2/n$ 为例.以 a 和σ^2 分别记总体的均值和方差.注意到

$$\begin{aligned}\sum_{i=1}^{n}(X_i-a)^2&=\sum_{i=1}^{n}[(X_i-\overline{X}_n)+(\overline{X}_n-a)]^2\\&=\sum_{i=1}^{n}(X_i-\overline{X}_n)^2+n(\overline{X}_n-a)^2\end{aligned}$$

知

$$m_2(n)=\frac{1}{n}\sum_{i=1}^{n}(X_i-a)^2-(\overline{X}_n-a)^2$$

依大数定理，$\sum_{i=1}^{n}(X_i-a)^2/n$ 依概率收敛于$E(X_i-a)^2=\sigma^2$，而 $\overline{X}_n-a$ 依概率收敛于0.故 $m_2(n)$依概率收敛于σ^2，即它是总体方差 σ^2 的相合估计.因为样本方差与样本二阶中心矩只相差一个因子 $n/(n-1)$，而当 $n\to\infty$时这个因子趋于1，知样本方差也是总体方差的相合估计.这样可以证明:前面例子中的许多估计都有相合性.

极大似然估计在很一般的条件下也有相合性.其证明比较复

杂，不能在此讨论了.

2. 渐近正态性. 估计量是样本 $X_1, \cdots, X_n$ 的函数，其确切分布要用第二章 2.4 节的方法去求. 除了若干简单的情况以外，这常是难于实现的. 例如，样本均值可算是最简单的统计量，它的分布也不易求得.

可是，正如在中心极限定理中所显示的，当 n 很大时，和的分布渐近于正态分布. 理论上可以证明，这不只是和所独有的，许多形状复杂的统计量，当样本大小 $n \to \infty$ 时，其分布都渐近于正态分布. 这称为统计量的“渐近正态性”. 至于哪些统计量具有渐近正态性，其确切形式如何，这都是很深的理论问题，在我们这个课程的范围内无法细加介绍了.

估计量的相合性和渐近正态性称为估计量的大样本性质. 指的是：这种性质都是对样本大小 $n \to \infty$ 来谈的. 对一个固定的 n，相合性和渐近正态性都无意义. 与此相对，估计量的无偏性概念是对固定的样本大小来谈的，不需要样本大小趋于无穷. 这种性质称为“小样本性质”. 因此，大小样本性质之分不在于样本的具体大小如何，而在于样本大小趋于无穷与否.