4.2 矩估计、极大似然估计和贝叶斯估计

4.2.1 参数的点估计问题

设有一个统计总体,以 $f(x,\theta_1,\cdots,\theta_k)$记其概率密度函数(若总体分布为连续型的),或其概率函数(若总体分布为离散型的),以后,为避免每次重复交代这两种情况,我们约定称 $f(x,\theta_1,\cdots,\theta_k)$为“总体分布”,其具体含义视其为连续型或离散型而定.这分布包含 k 个未知参数$\theta_1,\cdots,\theta_k$.例如对正态总体 $N(\mu,\sigma^2)$,有 $\theta_1=\mu,\theta_2=\sigma^2$,而

$$f(x,\theta_1,\theta_2) = \left(\sqrt{2\pi\theta_2}\right)^{-1}\exp\left(-\frac{1}{2\theta_2}(x-\theta_1)^2\right), -\infty < x < \infty$$

若总体有二项分布 $B(n,p)$,则 $\theta_1=p$,而

$$f(x,\theta_1) = \binom{n}{x}\theta_1^x(1-\theta_1)^{n-x}, x = 0,1,\cdots,n$$

当 $k=1$,即只有一个参数时,就用 θ 代替 θ_1.

参数估计问题的一般提法是:设有了从总体中抽出的样本 $X_1,\cdots,X_n$(在 4.1 节 4.1.3 段中已说明过,当不作特殊申明时,样本就是指独立随机样本,即 $X_1,\cdots,X_n$ 独立同分布,其公共分布就是总体分布),要依据这些样本去对参数 $\theta_1,\cdots,\theta_k$ 的未知值作出估计.当然,我们也可以只要求估计 $\theta_1,\cdots,\theta_k$ 中的一部分,或估计它们的某个已知函数 $g(\theta_1,\cdots,\theta_k)$.例如,为要估计 θ_1,我们需要构造出适当的统计量 $\hat{\theta}_1=\hat{\theta}_1(X_1,\cdots,X_n)$.每当有了样本 $X_1,\cdots,X_n$,就代入函数 $\hat{\theta}_1(X_1,\cdots,X_n)$ 算出一个值,用来作为 θ_1 的估计值.为着这样的特定目的而构造的统计量 $\hat{\theta}_1$,叫做(θ_1 的)估计量.由于未知参数 θ_1 是数轴上的一个点,用 $\hat{\theta}_1$ 去估计 θ_1,等于用一个点去估计另一个点,所以这样的估计叫做点估计,以别于将在 4.4 节讨论的区间估计.

在本节中我们要讨论几种常用的点估计方法,这些方法大多是基于某种直观上的考虑.同一个参数往往可以用若干个看来都合理的方法去估计.因此有一个判断优劣的问题,这就要为估计量的优劣制定准则,进而研究在某种准则下寻找最优估计量的问题.这就是参数估计这个数理统计学分支的重要内容.这些概念将在以后作更具体的解释.

4.2.2 矩估计法

矩估计法是 K. 皮尔逊在上世纪末到本世纪初的一系列文章中引进的.这个方法的思想很简单:设总体分布为 $f(x,\theta_1,\cdots,\theta_k)$,则它的矩(原点矩和中心矩都可以,此处以原点矩为例)

$$\alpha_m = \int_{-\infty}^{\infty} x^m f(x, \theta_1, \cdots, \theta_k) \mathrm{d}x$$

$$\left(或 \sum_i x_i^m f(x_i, \theta_i, \cdots, \theta_k)\right)$$

依赖于 $\theta_1, \cdots, \theta_k$. 另一方面,至少在样本大小 n 较大时,α_m 又应接近于样本原点矩 a_m. 于是

$$\alpha_m = \alpha_m(\theta_1, \cdots, \theta_k) \approx a_m = \sum_{i=1}^{n} X_i^m / n$$

取 $m=1, \cdots, k$,并让上面的近似式改成等式,就得到一个方程组:

$$\alpha_m(\theta_1, \cdots, \theta_k) = a_m, m = 1, \cdots, k \tag{2.1}$$

解此方程组,得其根 $\hat{\theta}_i = \hat{\theta}_i(X_1, \cdots, X_n), i = 1, \cdots, k$. 就以 $\hat{\theta}_i$ 作为 θ_i 的估计. $i=1, \cdots, k$. 如果要估计的是 $\theta_1, \cdots, \theta_k$ 的某函数 $g(\theta_1, \cdots, \theta_k)$,则用 $\hat{g} = \hat{g}(X_1, \cdots, X_n) = g(\hat{\theta}_1, \cdots, \hat{\theta}_k)$ 去估计它. 这样定出的估计量就叫做矩估计.

我们来举几个例子说明这个方法.

例 2.1 设 $X_1, \cdots, X_n$ 是从正态总体 $N(\mu, \sigma^2)$ 中抽出的样本,要估计 μ 和 σ^2. μ 是总体的一阶原点矩,按矩估计,用样本一阶原点矩即样本均值 $\overline{X}$ 去估计之. σ^2 是总体方差,即总体二阶中心矩,可用样本二阶中心矩 m_2 去估计. 一般,在估计方差时常用样本方差 s^2 而不用 m_2,即对矩估计作了一定的修正. 这种修正的理由将在下节中指出.

如果要估计的是标准差 σ,则由 $\sigma = \sqrt{\sigma^2}$,按矩估计法,它可以用 $\sqrt{m_2}$ 去估计,一般用 $\sqrt{s^2} = s$ 去估计,或者还作点修正(见下节). 又当 $\mu \neq 0$ 时(特别在 $\mu > 0$ 时,在有些问题中 μ 虽未知,但事先可知 $\mu > 0$. 如例 1.2,μ 是该校大学生的平均成绩,它必须大于 0),σ / μ 称为总体的变异系数——变异系数是以均值为单位去衡量的总体的标准差. 在有些问题中,反映变异程度的标准差意义如何,要看总体均值 μ 而定. 比如一大群人收入的标准差为 50 元. 若其平均工资只有 70 元,则这个变异程度可算很大了. 但若平均

工资为 850 元,则这变异程度不算大.所以,变异系数 σ/μ 不过是一定意义下的"相对误差".按矩法,为估计 σ/μ,可用$\sqrt{m_2}/\overline{X}$,一般用 $s/\overline{X}$.

例 2.2 设 $X_1,\cdots,X_n$ 是从指数分布总体中抽出的样本,要估计参数 λ 的倒数 $1/\lambda$.前已指出:$1/\lambda$ 就是总体分布的均值,故按矩法,就用$\overline{X}$去估计之.如要估计的是参数 λ 本身,就用 $1/\overline{X}$.

另一方面,如在第三章例 2.5 中指出的,指数分布的方差为 $1/\lambda^2$,即 $1/\lambda=\sqrt{\text{总体二阶中心矩}}$.按矩法,$1/\lambda$ 也可以用$\sqrt{m_2}$(或 s)去估计.这个估计与$\overline{X}$哪个更好?这就是需要研究的问题,见下节.

例 2.3 设 $X_1,\cdots,X_n$ 是从区间$[\theta_1,\theta_2]$上均匀分布的总体中抽出的样本,要估计 θ_1,θ_2.

前已指出(见第三章例 1.3 和例 2.5).这总体分布的均值、方差分别为$(\theta_1+\theta_2)/2$ 和$(\theta_2-\theta_1)^2/12$.因此按矩法,建立方程

$$\overline{X}=(\theta_1+\theta_2)/2,m_2=(\theta_2-\theta_1)^2/12$$

得出 θ_1,θ_2 的解$\hat{\theta}_1,\hat{\theta}_2$ 分别为

$$\hat{\theta}_1=\overline{X}-\sqrt{3m_2},\hat{\theta}_2=\overline{X}+\sqrt{3m_2} \tag{2.2}$$

也可以用 s 代替$\sqrt{m_2}$.

例 2.4 在第三章(2.8),(2.9)式中曾定义了分布的偏度系数 $\beta_1=\frac{\mu_3}{\mu_2^{3/2}}$及峰度系数 $\beta_2=\frac{\mu_4}{\mu_2^2}$(或 β_2-3),并阐述了它的意义.根据矩法,这些量可分别用$\frac{m_3}{m_2^{3/2}}$和$\frac{m_4}{m_2^2}$去估计之.

本例与前几例不同之处在于:它并不要求总体分布有特定的参数形式,如正态分布,指数分布之类.总体分布为任何分布都可以,只要其三阶(对 β_1)或四阶(对 β_2)矩存在就行.凡是被估计的对象能直接用矩表达出来时,都属于这种情况,其中最重要的例子是均值方差.只要总体分布的均值方差存在,则总可以用样本均值$\overline{X}$或样本方差 S^2 去估计之,而不论其分布有如何的形式.不过,在

总体分布已知有某种参数形式时，总体的均值方差也可以有比$\overline{X}$或S^2更好的估计（见后面有关的例子）.

例 2.5 设总体有二项分布 $B(N,p)$，$X_1,\cdots,X_n$ 为从该总体中抽出的样本.要估计 p，矩估计为$\overline{X}/N$.

例 2.6 设总体有波哇松分布 $P(\lambda)$，$X_1,\cdots,X_n$ 为从该总体中抽出的样本，要估计 λ.

由于 λ 是总体分布的均值，按矩估计法，用样本均值$\overline{X}$去估计之；另一方面，λ 也是总体分布的方差，故按矩法，也可以用 m_2 或 S^2 去估计.这又有一个优劣的问题.对本例及例 2.2 来说，在合理的准则下，都可以证明用样本均值$\overline{X}$为优.在一般情况下通常总是采取这样的原则：能用低阶矩处理的就不用高阶矩.

4.2.3 极大似然估计法

设总体有分布 $f(X;\theta_1,\cdots,\theta_k)$，$X_1,\cdots,X_n$ 为自这总体中抽出的样本，则样本$(X_1,\cdots,X_n)$的分布（即其概率密度函数或概率函数）为

$$f(X_1;\theta_1,\cdots,\theta_k)f(X_2;\theta_1,\cdots,\theta_k)\cdots f(X_n;\theta_1,\cdots,\theta_k)$$

记之为 $L(X_1,\cdots,X_n;\theta_1,\cdots,\theta_k)$.

固定 $\theta_1,\cdots,\theta_k$ 而看作是$X_1,\cdots,X_n$ 的函数时，L 是一个概率密度函数或概率函数，可以这样理解：若 $L(Y_1,\cdots,Y_n;\theta_1,\cdots,\theta_k)>L(X_1,\cdots,X_n;\theta_1,\cdots,\theta_k)$，则在观察时出现$(Y_1,\cdots,Y_n)$这个点的可能性，要比出现$(X_1,\cdots,X_n)$这个点的可能性大.把这件事反过来说，可以这样想：当已观察到 $X_1,\cdots,X_n$ 时，若 $L(X_1,\cdots,X_n;\theta'_1,\cdots,\theta'_k)>L(X_1,\cdots,X_n;\theta''_1,\cdots,\theta''_k)$，则被估计的参数$(\theta_1,\cdots,\theta_k)$是$(\theta'_1,\cdots,\theta'_k)$的可能性，要比它是$(\theta''_1,\cdots,\theta''_k)$的可能性大.

当$X_1,\cdots,X_n$ 固定而把L 看作$\theta_1,\cdots,\theta_k$ 的函数时，它称为“似然函数”.这名称的意义，可根据上述分析得到理解：这函数对不同的$(\theta_1,\cdots,\theta_k)$的取值，反映了在观察结果$(X_1,\cdots,X_n)$已知的条件下，$(\theta_1,\cdots,\theta_k)$的各种值的“似然程度”.注意这里有些像贝叶

斯公式中的推理(见第一章(3.18)式):把观察值 $X_1,\cdots,X_n$ 看成结果而参数值($\theta_1,\cdots,\theta_k$)看成是导致这结果的原因.现已有了结果,要反过来推算各种原因的概率.这里参数 $\theta_1,\cdots,\theta_k$ 有一定的值(虽然未知),并非事件或随机变量,无概率可言,于是就改用"似然"这个词.

从上述分析就自然地导致如下的方法:应该用似然程度最大的那个点($\theta_1^*,\cdots,\theta_k^*$),即满足条件

$$
\begin{aligned}
&L(X_1,\cdots,X_n;\theta_1^*,\cdots,\theta_k^*)\\
&=\max_{\theta_1,\cdots,\theta_k} L(X_1,\cdots,X_n;\theta_1,\cdots,\theta_k)
\end{aligned}
\tag{2.3}
$$

的($\theta_1^*,\cdots,\theta_k^*$)去作为($\theta_1,\cdots,\theta_k$)的估计值,因为在已得样本 $X_1,\cdots,X_n$ 条件下,这个"看来最像"是真参数值.这个估计($\theta_1^*,\cdots,\theta_k^*$)就叫做($\theta_1,\cdots,\theta_k$)的"极大似然估计".如果要估计的是 $g(\theta_1,\cdots,\theta_k)$,则 $g(\theta_1^*,\cdots,\theta_k^*)$是它的极大似然估计.

因为

$$
\log L=\sum_{i=1}^{n}\log f(X_i;\theta_1,\cdots,\theta_k) \tag{2.4}
$$

且为使 L 达到最大,只须使 $\log L$ 达到最大,故在 f 对 $\theta_1,\cdots,\theta_k$ 存在连续的偏导数时,可建立方程组(称为似然方程组):

$$
\frac{\partial \log L}{\partial \theta_i}=0,\ i=1,\cdots,k \tag{2.5}
$$

如果这方程组有唯一的解,又能验证它是一个极大值点,则它必是使 L 达到最大之点,即极大似然估计.在几个常见的重要例子中这一点不难验证.可是,在较复杂的场合,方程组(2.5)可以有不止一组解,求出这些解很费计算,且不易判定那一个使 L 达到最大.

有时,函数 f 并不对 $\theta_1,\cdots,\theta_k$ 可导,甚至 f 本身也不连续,这时方程组(2.5)就无法用,必须回到原始的定义 2.3.

现举一些例子来说明求极大似然估计的过程.

例 2.7 设 $X_1,\cdots,X_n$ 是从正态总体 $N(\mu,\sigma^2)$中抽出的样本,则似然函数为

$$L = \prod_{i=1}^{n}\left[\left(\sqrt{2\pi\sigma^2}\right)^{-1}\exp\left(-\frac{1}{2\sigma^2}(X_i-\mu)^2\right)\right] \qquad (2.6)$$

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$$

求方程组(2.5)(把 σ^2 作为一个整体看):

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) = 0$$

$$\frac{\partial \log L}{\partial(\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i-\mu)^2 = 0$$

由第一式得出 μ 的解为

$$\mu^* = \sum_{i=1}^{n} X_i/n = \overline{X}$$

以此代入第二式的 μ,得到 σ^2 的解为

$$\sigma^{*2} = \sum_{i=1}^{n}(X_i-\overline{X})^2/n = m_2$$

我们看到:μ 与 σ^2 的极大似然估计 μ^* 和 σ^{*2},与其矩估计完全一样.在本例中,容易肯定(μ^*,σ^{*2})确是使似然函数 L 达得最大值之点.因为,似然方程组只有唯一的根(μ^*,σ^{*2}),而这个点不可能是 L 的极小值点.因为,由 L 的表达式(2.6)可知,当$|\mu|\to\infty$ 或 $\sigma^2\to 0$ 时,L 趋向于 0,而 L 在每个点处都大于 0.以下几个例子都可以按照这个方式去验证,我们就不一一重复了.

例 2.8 设 $X_1,\cdots,X_n$ 是从指数分布总体中抽出的样本,求参数 λ 的极大似然估计.

有

$$L = \prod_{i=1}^{n}(\lambda e^{-\lambda x_i})$$

故

$$\log L = n\log\lambda - \lambda\sum_{i=1}^{n} X_i$$

解方程

$$\frac{\partial \log L}{\partial \lambda}=\frac{n}{\lambda}-\sum_{i=1}^{n}X_i=0$$

得 λ 的极大似然估计为

$$\lambda^*=n\Big/\sum_{i=1}^{n}X_i=1/\overline{X}$$

仍与其矩估计一样.但是在这里,极大似然估计只有一个,而如在例 2.2 中所指出的,λ 的矩估计依使用不同阶的矩,可以有几个.

例 2.9 设 $X_1,\cdots,X_n$ 是从均匀分布 $R(0,\theta)$ 的总体中抽出的样本,求 θ 的极大似然估计.

X_i 的密度函数为 $1/\theta$,当 $0<X_i<\theta$,此外为 0.故似然函数 L 为

$$L=\begin{cases}\theta^{-n}, & \text{当}\ 0<X_i<\theta, i=1,\cdots,n\\ 0, & \text{其他情况}\end{cases}$$

对固定的 $X_1,\cdots,X_n$,此函数为 θ 的间断函数,故无法使用似然方程(2.5).但此例不难直接用最初的定义 2.3 去解决:为使 L 达到最大,θ 必须尽量小,但又不能太小以致 L 为 0.这界线就在 $\theta^*=\max(X_1,\cdots,X_n)$处:当 $\theta\geqslant\theta^*$ 时,L 大于 0 且为 θ^{-n}.当 $\theta<\theta^*$ 时,L 为 0.故唯一使 L 达到最大的 θ 值,即 θ 的极大似然估计,为 θ^*.

如果用矩估计法,则因总体分布的均值为 $\theta/2$,θ 的矩估计为 $\hat{\theta}=2\overline{X}$.这两个估计的优劣比较将在后面讨论.

例 2.10 再考虑例 2.5,有

$$L=\prod_{i=1}^{n}\left[\binom{N}{X_i}p^{X_i}(1-p)^{N-X_i}\right]$$

$$\log L=\sum_{i=1}^{n}\log\binom{N}{X_i}+\sum_{i=1}^{n}X_i\log p+\sum_{i=1}^{n}(N-X_i)\log(1-p)$$

作方程

$$\frac{\partial \log L}{\partial p}=\frac{1}{p}\sum_{i=1}^{n}X_i-\left(nN-\sum_{i=1}^{n}X_i\right)\frac{1}{1-p}=0$$

此方程之解,即 p 的极大似然估计,为 $p^*=\overline{X}/N$,与矩估计相

同.

例 2.11 考虑例 2.6. 容易证明: λ 的极大似然估计 $\lambda^* = \overline{X}$, 与矩估计相同.

在我们所举的这些例子中(这些例子都是在应用上最常见的), 矩估计与极大似然估计在多数情况下一致. 这更多地是一种巧合, 并非一般情形. 有意思的是: 在这些例子中这两种估计方法结果一致, 说明这些估计是良好的. 这一点当然还需要一定的理论证明.

也有这样的情况, 用这两个估计方法都行不通或不易实行. 下面是一个例子.

例 2.12 设总体分布有密度函数

$$f(x,\theta) = \frac{1}{\pi[1 + (x - \theta)^2]}, \quad -\infty < x < \infty \qquad (2.7)$$

这分布包含一个参数 θ, θ 可取任何实数值. 这分布叫柯西分布, 其密度作为 x 的函数, 关于 θ 点对称. 故 θ 是这个分布的中位数(见第三章 3.1.4).

现设 $X_1, \cdots, X_n$ 为自这总体中抽出的样本, 要估计 θ. 由于

$$\int_{-\infty}^{\infty} |x| f(x,\theta) \mathrm{d}x = \infty$$

柯西分布的一阶矩也不存在, 更不用说更高阶的矩了. 因此, 矩估计无法使用. 若用极大似然法, 则将得出方程

$$\sum_{i=1}^{n} \frac{X_i - \theta}{1 + (X_i - \theta)^2} = 0$$

这方程有许多根且求根不容易. 因此, 对本例而言, 极大似然法也不是理想的方法.

为估计参数 θ, 有一个较简单易行但看来合理的方法可用. 这个方法是基于 θ 是总体分布的中位数这个事实. 既如此, 我们就要设法在样本 $X_1, \cdots, X_n$ 中找一种对应于中位数的东西. 这个思想其实在矩估计法中就已用过, 因为总体矩在样本中的对应物就是样本矩.

现在把 $X_1,\cdots,X_n$ 按由小到大排成一列:

$$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)} \tag{2.8}$$

它们称为次序统计量.既然中位数是“居中”的意思,我们就在样本中找居中者:

$$\hat{m} = \begin{cases} X_{((n+1)/2)}, \text{当 } n \text{ 为奇数时} \\ (X_{(n/2)} + X_{(n/2+1)})/2, \text{当 } n \text{ 为偶数时} \end{cases} \tag{2.9}$$

当 n 为奇数时,有一个居中者为 $X_{((n+1)/2)}$;若 n 为偶然,就没有一个居中者,就把两个最居中者取平均.这样定义的 $\hat{m}$ 叫作“样本中位数”.我们就拿 $\hat{m}$ 作为 θ 的估计.

就正态总体 $N(\mu,\sigma^2)$ 而言,μ 也是总体的中位数,故 μ 也可以用样本中位数去估计.从这些例子中,我们看出一点:统计推断问题的解,往往可以从许多看来都合理的途径去考虑,并无一成不变的方法,不同解固然有优劣之分,但这种优劣也是相对于一定的准则而言.并无绝对的价值.下述情况也并非不常见:估计甲在某一准则下优于乙,而乙又在另一准则下优于甲.

极大似然估计法的思想,始于高斯的误差理论,到 1912 年由 R. A. 费歇尔在一篇论文中把它作为一个一般的估计方法提出来.自 20 年代以来,费歇尔自己及许多统计学家对这一估计法进行了大量的研究.总的结论是:在各种估计方法中,相对说它一般更为优良,但在个别情况下也给出很不理想的结果.与矩估计法不同,极大似然估计法要求分布有参数的形式.比方说,如对总体分布毫无所知而要估计其均值方差,极大似然法就无能为力.

4.2.4 贝叶斯法

贝叶斯学派是数理统计学中的一大学派.在这一段中,我们简略地介绍一下这个学派处理统计问题的基本思想.

拿我们目前讨论的点估计问题来说,无论你用矩估计也好,用极大似然估计或其他方法也好,在我们心目中,未知参数 θ 就简单地是一个未知数,在抽取样本之前,我们对 θ 没有任何了解,所

有的信息全来自样本.

贝叶斯学派则不然,它的出发点是:在进行抽样之前,我们已对 θ 有一定的知识,叫做先验知识.这里"先验"的意思并非先验论,而只是表示这种知识是"在试验之先"就有了的,也有人把它叫做验前知识,即"在试验之前"的意思.

贝叶斯学派进一步要求:这种先验知识必须用 θ 的某种概率分布表达出来,这概率分布就叫做 θ 的"先验分布"或"验前分布".这个分布总结了我们在试验之前对未知参数 θ 的知识.

举一个例子.设某工厂每日生产一大批某种产品,我们想要估计当日的废品率 θ.该厂在以前已生产过很多批产品,如果过去的检验有记录在,则它确实提供了关于废品率 θ 的一种有用信息,据此可以画出 θ 的密度曲线,如图 4.1(a),(b).

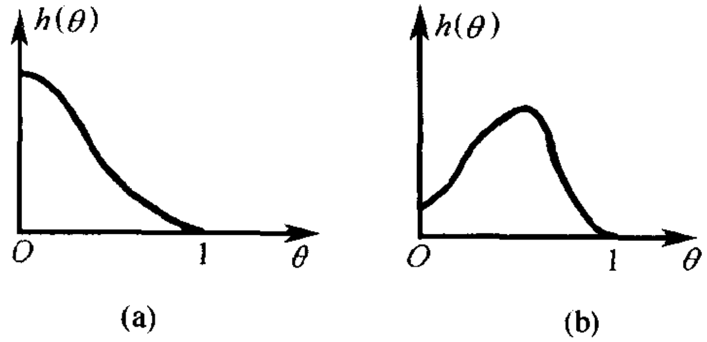

图 4.1

图中 $h(\theta)$表示 θ 的密度函数,$0\leqslant\theta\leqslant1$.(a)表示一个较好的情况:$h(\theta)$在 $\theta=0$ 附近很大而当 θ 增加时,下降很快.这表示该厂以往的废品率通常都很低.(b)则表示一个不大好的情况:比较大的废品率出现的比率相当高.容易理解:这种关于 θ 的历史知识(即先验知识),在当前估计废品率 θ 时,应适当地加以使用而不应弃之不顾.这种思想与我们日常处事的习惯符合:当我们面临一个问题时,除考虑当前的情况外,往往还要注意以往的先例和经验.

问题就来了:如果这个工厂以往没有记录,或甚至是一个新开工的工厂,该怎么办?怎样去获得上文所指的先验密度 $h(\theta)$?贝

叶斯统计的一个基本要求是:你必须设法去定出这样一个 $h(\theta)$,甚至出于你自己的主观认识*也可以,这要成为问题中一个必备的要素.正是在这一点上,贝叶斯统计遭到不少的反对和批评,而一个初接触这个问题的人,也容易这样想:"这怎么行?我没有根据怎么能凭主观想像去定出一个先验密度 $h(\theta)$".关于这一点,贝叶斯学派的信奉者有自己的一套说法,这问题非三言两语能说清楚.本书作者有一篇通俗形式的文章(见《数理统计与应用概率》1990 年第四期,p.389—400),其中对这个问题及有关问题作了仔细说明,有兴趣的读者可以参考.

现在我们转到下一个问题:已定下了先验密度之后,怎样去得出参数 θ 的估计.

设总体有概率密度 $f(X,\theta)$(或概率函数,若总体分布为离散的),从这总体抽样本 $X_1,\cdots,X_n$,则这样本的密度为 $f(X_1,\theta)\cdots f(X_n,\theta)$.它可视为在给定 θ 值时$(X_1,\cdots,X_n)$的密度,根据第二章(3.5)式及该式下的一段说明,$(\theta,X_1,\cdots,X_n)$的联合密度为

$$h(\theta)f(X_1,\theta)\cdots f(X_n,\theta)$$

由此,算出$(X_1,\cdots,X_n)$的边缘密度为

$$p(X_1,\cdots,X_n)=\int h(\theta)f(X_1,\theta)\cdots f(X_n,\theta)\mathrm{d}\theta \tag{2.10}$$

积分的范围,要看参数 θ 的范围而定.如上例 θ 为废品率,则 $0\leqslant\theta\leqslant1$.若 θ 为指数分布中的参数 λ,则 $0<\theta<\infty$,等等.由(2.10),再根据第二章的公式(3.4),得到在给定 $X_1,\cdots,X_n$ 的条件下,θ 的条件密度为

$$h(\theta|X_1,\cdots,X_n)=h(\theta)f(X_1,\theta)\cdots f(X_n,\theta)/p(X_1,\cdots,X_n) \tag{2.11}$$

照贝叶斯学派的观点,这个条件密度代表了我们现在(即在取得样本 $X_1,\cdots,X_n$ 后)对 θ 的知识,它综合了 θ 的先验信息(以 $h(\theta)$反映)与由样本带来的信息.通常把(2.11)称为 θ 的"后验(或验后)

---

* 就是说,这里允许使用主观概率,见第一章 1.1 节.

密度”,因为他是在做了试验以后才取得的.

如果把上述过程和我们在第一章中讲过的贝叶斯公式相比,就可以理解:现在我们所做的,可以说不过是把贝叶斯公式加以“连续化”而已,看下表中的比较.

	问　题	先验知识	当前知识	后验(现在)知识
贝叶斯公式	事件 $B_1,\cdots,B_n$ 中那一个发生了?	$P(B_1)$, $\cdots,P(B_n)$	事件 A 发生了	$P(B_1\mid A),\cdots$, $P(B_n\mid A)$
此处的问题	$\theta=?$	$h(\theta)$	样本 $X_1,\cdots,X_n$	后验密度(2.11)

由这里我们就理解到:为什么一个看来不起眼的贝叶斯公式会有如此大的影响.这一点我们在第一章中已有所论述了.

贝叶斯学派的下一个重要观点是:在得出后验分布(2.11)后,对参数 θ 的任何统计推断,都只能基于这个后验分布.至于具体如何去使用它,可以结合某种准则一起去进行,统计学家也有一定的自由度.拿此处讨论的点估计问题来说,一个常用的方法是:取后验分布(2.11)的均值作为 θ 的估计.

还有一点需要说明一下:按上文,$h(\theta)$必须是一个密度函数,即必须满足 $h(\theta)\geqslant 0$, $\int h(\theta)\mathrm{d}\theta=1$ 这两个条件.但在有些情况下,$h(\theta)\geqslant 0$,但 $\int h(\theta)\mathrm{d}\theta$ 不为 1 甚至为∞,不过积分(2.10)仍有限,这时,由(2.11)定义的 $h(\theta\mid X_1,\cdots,X_n)$作为 θ 的函数,仍满足密度函数的条件.这就是说,即使这样的 $h(\theta)$取为先验密度也无妨.当然,由于 $\int h(\theta)\mathrm{d}\theta$ 不为 1,它已失去了密度函数的通常的概率意义.这样的 $h(\theta)$通常称为“广义先验密度”.

例 2.13　作 n 次独立试验,每次观察某事件 A 是否发生,A 在每次试验中发生的概率为 p,要依据试验结果去估计 p.

这问题我们以往就“用频率估计概率”的方法去处理(这也是它的矩估计与极大似然估计).这方法不用 p 的先验知识.现在我

们用贝叶斯统计的观点来处理这个问题.

引进 $X_i=1$ 或 0,视第 i 次试验时 A 发生与否而定,$i=1,\cdots,n$. $P(X_i=1)=p$,$P(X_i=0)=1-p$. 因此$(X_1,\cdots,X_n)$的概率函数为 $p^x(1-p)^{n-x}$,$X=\sum_{i=1}^{n}X_i$. 取 p 的先验密度 $h(p)$,则 p 的后验密度为

$$h(p|X_1,\cdots,X_n)$$
$$=h(p)p^x(1-p)^{n-x}\Big/\int_0^1 h(p)p^x(1-P)^{n-x}\mathrm{d}p,0\leqslant p\leqslant 1$$

此分布的均值为

$$\tilde{p}=\tilde{p}(X_1,\cdots,X_n)=\int_0^1 ph(P|X_1,\cdots,X_n)\mathrm{d}p$$
$$=\int_0^1 h(p)p^{x+1}(1-p)^{n-x}\mathrm{d}p\Big/\int_0^1 h(p)p^x(1-p)^{n-x}\mathrm{d}p \tag{2.12}$$

$\tilde{p}$就是 p 在先验分布 $h(p)$之下的贝叶斯估计.

如何选择 $h(p)$? 贝叶斯本人曾提出"同等无知"的原则,即事先认为 p 取[0,1]内一切值都有同等可能,就是说取[0,1]内均匀分布 $R(0,1)$作为 p 的先验分布. 这时 $h(p)=1$ 当$0\leqslant p\leqslant 1$,而(2.12)中的两个积分都可以用 β 函数表出(见第二章(4.22)式). 由此得

$$\tilde{p}=\beta(X+2,n-X+1)/\beta(X+1,n-X+1) \tag{2.13}$$

根据 β 函数与 Γ 函数的关系式(4.25),以及当 k 为自然数时 $\Gamma(k)=(k-1)!$,由(2.13)不难得到

$$\tilde{p}=(X+1)/(n+2) \tag{2.14}$$

这个估计与频率 X/n 有些差别,当 n 很大时不显著,而在 n 很小时颇为显著. 从一个角度看,当 n 相当小时,用贝叶斯估计(2.14)比用 X/n 更合理. 因为当 n 很小时,试验结果可能出现 $X=0$ 或 $X=n$ 的情况. 这时,依 X/n 应把 p 估计为 0 或 1,这就太极端了(我们不能仅根据在少数几次试验中 A 会不出现或全出现,就判

定它为不可能或必然). 若按(2.14),则在这两种情况下分别给出估计值$1/(n+2)$和$(n+1)/(n+2)$. 这就留有一定的余地.

这个"同等无知"的原则,又称贝叶斯原则,被广泛用到一些其他的情况. 不过随着所估计的参数的范围和性质的不同,该原则的具体表现形式也不同. 例如,为估计正态分布$N(\mu,\sigma^2)$中的μ,同等无知原则给出一个广义先验密度$h(\mu)\equiv 1$. 若估计σ,则应取$h(\sigma)=\sigma^{-1}(\sigma>0)$. 若估计指数分布中的$\lambda$,则取$h(\lambda)=\lambda^{-1}(\lambda>0)$. 这些都是广义先验密度. 其所以这样做的理由,不能在此处细谈了.

这个原则也受到一些批评,其中最有力的批评是其不确定性. 理由是:拿本例的p来说,若对p同等无知,则对p^2(或$p^3,p^4,\cdots$等)也应是同等无知,因而也可以把p^2的密度函数取为$R(0,1)$的密度. 这时不难算出p的密度将为$h(p)=2p$(当$0\leqslant p\leqslant 1$,其外为0),与本例所给不一致. 另外,不言而喻,同等无知的原则是一个在确实没有什么信息时,不得已而采用的办法. 在实际问题中,有时是存在更确实的信息的,如本段开始讲到的那个估计废品率的情况. 又如,估计一个基本上均匀的铜板在投掷时出现正面的概率p. 我们有理由事先肯定p离1/2不远. 这时,可考虑取一个适当的数$\varepsilon>0$,而把p的先验分布取为$[1/2-\varepsilon,1/2+\varepsilon]$内的均匀分布. 这肯定比用同等无知的原则效果要好,尤其是在试验次数n不大时.

例 2.14 设$X_1,\cdots,X_n$是自正态总体$N(\theta,1)$中抽出的样本. 为估计θ,给出θ的先验分布为正态分布$N(\mu,\sigma^2)$(μ,σ^2当然都已知). 求θ的贝叶斯估计. 在本例中有

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

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

故按公式(2.11)知,θ的后验密度为

$$h(\theta|X_1,\cdots,X_n)=\exp\left[-\frac{1}{2\sigma^2}(\theta-\mu)^2-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2\right]\Big/ I \tag{2.15}$$

其中 I 是一个与 θ 无关而只与 $\mu,\sigma,X_1,\cdots,X_n$ 有关的数. 简单的代数计算表明

$$-\frac{1}{2\sigma^2}(\theta-\mu)^2-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2=-\frac{1}{2\eta^2}(\theta-t)^2+J \tag{2.16}$$

其中

$$t=(n\overline{X}+\mu/\sigma^2)/(n+1/\sigma^2) \tag{2.17}$$

$$\eta^2=1/(n+1/\sigma^2) \tag{2.18}$$

而 J 与 θ 无关. 以(2.16)代入(2.15),得

$$h(\theta|X_1,\cdots,X_n)=I_1\exp\left[-\frac{1}{2\eta^2}(\theta-t)^2\right]$$

这里 $I_1=Ie^J$ 与 θ 无关. I_1 不必直接算,因为,$h(\theta|X_1,\cdots,X_n)$作为 θ 的函数是一个概率密度函数,它必须满足条件

$$\int_{-\infty}^{\infty}h(\theta|X_1,\cdots,X_n)\mathrm{d}\theta=1$$

这就决定了 $I_1=(\sqrt{2\pi}\eta)^{-1}$. 因此,$\theta$ 的后验分布就是正态分布 $N(t,\eta^2)$,其均值 t 就是 θ 的贝叶斯估计$\tilde{\theta}$:

$$\tilde{\theta}=t=\frac{n}{n+1/\sigma^2}\overline{X}+\frac{1/\sigma^2}{n+1/\sigma^2}\mu \tag{2.19}$$

把$\tilde{\theta}$写成(2.19)的形状很有意思. 设想两个极端情况:一个是只有样本信息而毫无先验信息,这就是我们以前讨论的情况,这时用样本均值$\overline{X}$去估计θ. 另一个是只有先验信息 $N(\mu,\sigma^2)$而没有样本. 这时,我们只好用先验分布的均值 μ 作为 θ 的估计. 由(2.19)式看出:当两种信息都存在时,θ 的估计为二者的折衷. 它是上述两个极端情况下的估计$\overline{X}$和μ 的加权平均,权之比为 $n:1/\sigma^2$. 这个比值很合理:n 为样本数目,n 愈大,样本信息愈多,$\overline{X}$的权就该更大. 对 μ 而言,其重要性则要看 σ^2 的大小. σ^2 愈大,表示先验信息愈不肯定(θ 在μ 周围的散布很大). 反之,σ^2 很小时,仅根据先验信息,已有很大把握肯定 θ 在μ 附近不远处. 因此,μ 的权应与σ^2 成反比. 公式(2.19)恰好体现了上述分析.

目前在国际统计界及应用统计工作者中,贝叶斯学派已有很大影响.其原因在于它确实有一些别的方法所不具备的优点.这些在今后我们还将看到.在我国,贝叶斯方法也开始受到重视并得到一些应用.对把数理统计学方法作为一种工具的应用工作者来说,对这个学派的方法有必要有一定的了解.