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).

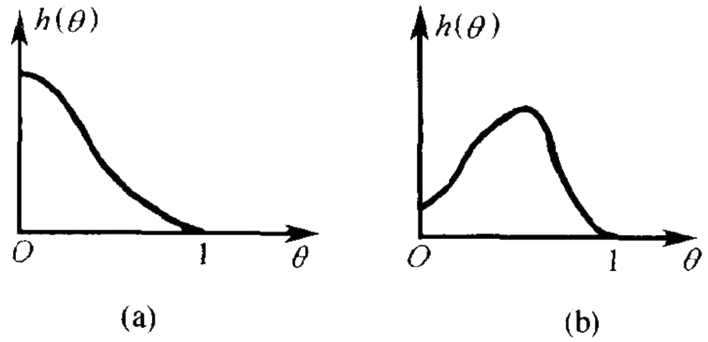

图 4.1

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

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

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

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

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	$P(B_1\|A),\cdots$, $P(B_n\|A)$
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 $h(\theta|X_1,\cdots,X_n)$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 λ,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, θ 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.