统计决策与贝叶斯估计

第一节：统计决策的基本概念

- 胡政发，肖海霞，应用数理统计与随机过程，电子工业出版社，2021年第一版
- 师义民，徐伟，秦超英，许勇，数理统计，科学出版社，2015年第四版

统计判决问题的三个要素

- 为了估计一个未知参数，需要给出一个合适的估计量，该估计量也称为该统计问题的解。一般地说，一个统计问题的解就是所谓的统计决策函数。
- 为了明确统计决策函数这一重要概念，需对构成一个统计决策问题的基本要素作一介绍。
- 这些要素是：样本空间和分布族、行动空间以及损失函数。

统计判决问题的三个要素

1. 样本空间和分布族

设总体X的分布函数为$F(x;\theta)$，θ是未知参数$\theta \in \Theta$， Θ称为参数空间。若$(X_1, X_2, \cdots, X_n)^T$为取自总体X的一个样本，则样本所有可能值组成的集合称为样本空间，记为$\mathfrak{R}$，由于X_i的分布函数为$F(x_i;\theta), i = 1,2,\cdots,n$，则$(X_1, X_2, \cdots, X_n)^T$的联合分布函数为

$$F(x_1, \cdots, x_n;\theta) = \prod_{i=1}^{n} F(x_i;\theta), \theta \in \Theta$$

若记$F^* = \{\prod_{i=1}^{n} F(x_i;\theta), \theta \in \Theta\}$，则称$F^*$为样本$(X_1, X_2, \cdots, X_n)^T$的概率分布族，简称分布族。

统计判决问题的三个要素

➢例3. 1　设总体X服从两点分布$B(1,p),p$为未知参数，$0<p<1$，$(X_1,X_2,\cdots,X_n)^T$是取自总体X的样本，则样本空间是集合$\Re=\{(x_1,x_2,\cdots,x_n):x_i=0,1,i=1,2,\cdots,n\}$

它含有2^n个元素，样本$(X_1,X_2,\cdots,X_n)^T$的分布族为

$$F^*=\{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i},x_i=0,1,i=1,2,\cdots,n,0<p<1\}$$

统计判决问题的三个要素

2. 行动空间（或称判决空间）

对于一个统计问题，如参数θ的点估计，区间估计及其他统计问题，我们常常要给予适当的回答。对参数θ的点估计，一个具体的估计值就是一个回答。在统计决策中，每一个具体的回答称为一个决策，**一个统计问题中可能选择的全部决策组成的集合称为决策空间，记为$\mathcal{R}$**。一个决策空间$\mathcal{R}$至少应含有两个决策，假如$\mathcal{R}$中只含有一个决策，那人们就无需选择，从而也形成不了一个统计决策问题。

◘统计判决问题的三个要素

➢例如，要估计正态分布$N(\mu,\sigma^2)$中的参数μ，$\mu\in\Theta=(-\infty,+\infty)$。因为$\mu$在$(-\infty,+\infty)$ 中取值，所以每一个实数都可用来估计μ，故每一个实数都代表一个决策，决策空间为$\mathfrak{R}=(-\infty,+\infty)$ 。

◆值得注意的是，在$\mathfrak{R}$中具体选取哪个决策与抽取的样本和所采用的统计方法有关。

➢例3. 2 某厂打算根据各年度市场的销售量来决定下年度应该扩大生产还是缩减生产，或者维持原状，这样的决策空间$\mathcal{R}$为

$\mathcal{R}$={扩大生产，缩减生产，维持原状}。

3. 损失函数

- 统计决策的一个基本观点和假定是，每采取一个决策，必然有一定的后果（经济的或其他的），决策不同，后果各异。对于每个具体的统计决策问题，一般有多种优劣不同的决策可采用。
- 例如，要估计正态分布$N(\mu, 0.02^2)$中的参数μ，假设μ的真值为3，那么采用3.5这个决策显然比10这个决策好的多。如果要作μ的区间估计，则显然$[2, 4]$这个决策比$[-5, 10]$这个决策好。

3. 损失函数

统计决策理论的一个基本思想是把上面所谈的优劣性，以数量的形式表现出来，其方法是引入一个依赖于参数值$\theta \in \Theta$和决策$d \in \Re$的二元实值非负函数$L(\theta, d) \geq 0$，称之为损失函数，它表示当参数真值为θ而采取决策d时所造成的损失，决策越正确，损失就越小。

由于在统计问题中人们总是利用样本对总体进行推断，所以误差是不可避免的，因而总会带来损失，这就是损失函数定义为非负函数的原因。

3. 损失函数

➢例3. 3 设总体X服从正态分布$N(\theta,1)$，θ为未知参数，参数空间$\Theta=(-\infty,+\infty)$，决策空间自然地取为$\Re=(-\infty,+\infty)$，一个可供考虑的损失函数是

$$L(\theta,d)=(\theta-d)^2$$

当$d=\theta$，即估计正确时损失为0，估计d与实际值θ的距离$|d-\theta|$越大，损失也越大。

如果要求未知参数θ的区间估计，损失函数可取为：

$$L(\theta,d)=d_2-d_1,\theta\in\Theta,d=[d_1,d_2]\in\Re$$

- 其中$\mathfrak{R}=\{[d_1,d_2]:-\infty<d_1<d_2<\infty\}$，这个损失函数表示以区间估计的长度来度量采用决策$d=[d_1,d_2]$所带来的损失，也可以取损失函数为

$$L(\theta,d)=1-I_{[d_1,d_2]}(\theta),\theta\in\Theta,d=[d_1,d_2]\in\mathfrak{R}$$

其中$I_{[d_1,d_2]}(\theta)$是集合$[d_1,d_2]$的示性函数，即

$$I_{[d_1,d_2]}(\theta)=\begin{cases}0, & \text{当}\theta\notin[d_1,d_2]\\ 1, & \text{当}\theta\in[d_1,d_2]\end{cases}$$

这个损失函数表示当决策d正确（即区间$[d_1,d_2]$覆盖未知参数的实际值）时损失为0，反之损失为1.

- 对于不同的统计问题，可以选取不同的损失函数，常见的损失函数有以下几种。

（1）线性损失函数

$$L(\theta,d)=\begin{cases}k_0(\theta-d), d\le\theta\\ k_1(d-\theta), d>\theta\end{cases}$$

其中k_0和k_1是两个常数，它们的选择常反映行动d低于参数θ和高于参数θ的相对重要性，当

$k_0=k_1=1$时就得到：

绝对值损失函数

$$L(\theta,d)=|\theta-d|$$

（2）平方损失函数

$$L(\theta, d) = (\theta - d)^2$$

（3）凸损失函数

$$L(\theta, d) = \lambda(\theta) W(|\theta - d|)$$

其中$\lambda(\theta) > 0$是θ的已知函数，且有限，$W(t)$是t>0上的单调非降函数且$W(0) = 0$。

（4）多元二次损失函数，当$\boldsymbol{\theta}$和d均为多维向量时，可取如下二次型作为损失函数

$$L(\boldsymbol{\theta}, \boldsymbol{d}) = (\boldsymbol{d} - \boldsymbol{\theta})^T \boldsymbol{A} (\boldsymbol{d} - \boldsymbol{\theta})$$

其中$\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \cdots, \theta_p)^{\boldsymbol{T}}$，d= $(\boldsymbol{d_1}, \cdots, d_p)^{\boldsymbol{T}}$ ， $\boldsymbol{A}$ 为$p \times p$阶正定矩阵，p为大于1的某个自然数，当A为对角阵即$\boldsymbol{A} = \boldsymbol{diag}(\boldsymbol{w_1}, \cdots, \boldsymbol{w_p})$时，则p元这个损失函数为

$$L(\boldsymbol{\theta}, \boldsymbol{d}) = \sum_{i=1}^{p} w_i (d_i - \theta_i)^2$$

其中$w_i (i = 1,2, \cdots, p)$可看作各参数重要性的加权。

- 将统计决策方法用于实际问题时，如何选择损失函数是一个关键问题，也是一个难点。一般来说，选取的损失函数应与实际问题相符合，同时也要在数学上便于处理。上面提到的二次损失（又称为平方损失）函数是参数点估计中常用的一种损失函数。

统计决策函数及其风险函数

统计决策函数

给定了样本空间g和概率分布族F^*，决策空间$\Re$及损失函数$L(\theta,d)$这三个要素后，统计决策问题就确定了，此后，我们的任务就是在$\Re$中选取一个好的决策d，所谓好是指有较小的损失。对样本空间g中每一点$x=(x_1,\cdots,x_n)^T$，可在决策空间中寻找一点$d(x)$与其对应，这样一个对应关系可看作定义在样本空间g上而取值于决策空间$\Re$内的函数$d(x)$。

- 定义3.1 定义在样本空间$\mathcal{g}$上，取值于决策空间$\mathfrak{R}$内的函数，称为统计决策函数$d(x)$，简称决策函数。
- 形象地说，决策函数$d(x)$就是一个“行动方案”。当有了样本x后，按既定的方案采取行动（决策）$d(x)$，在不致误解的情况下，也称$d(x) = d(X_1, \cdots, X_n)$为决策函数，此时表示当样本值为$x = (x_1, \cdots, x_n)^T$时采取决策$d(x) = d(x_1, \cdots, x_n)$，因此，决策函数本质上是一个统计量.

➢例如，设总体X服从正态分布$N(\mu,\sigma^2)$，σ^2已知，$(X_1,X_2,\cdots,X_n)^T$为取自X的样本，求参数μ的点估计。此时可用$d(x)=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$来估计$\mu$，$d(x)=\bar{x}$就是一个决策函数。

◘如果要求区间估计，那么$d(x)=(\bar{x}-u_{\frac{\sigma}{2}}\frac{\sigma}{\sqrt{n}},\bar{x}+u_{\frac{\sigma}{2}}\frac{\sigma}{\sqrt{n}})$就是一个决策函数

◘ 风险函数

◘给定一个决策函数$d(x)$之后，所采取的决策完全取决于样本X，从而损失必然与X有关，也就是说决策函数与损失函数$L(\theta,d)$都是样本X的函数，因此都是随机变量。

当样本X取不同的值x时，对应的决策$d(x)$可能不同，由此带来的损失$L(\theta, d(x))$也不相同，这样就不能运用基于样本所采取的决策而带来的损失$L(\theta, d(x))$来衡量决策的好坏，而应该从整体上来评价。为了比较决策函数的优劣，一个常用的数量指标是平均损失，即所谓的风险函数。

□定义3. 2 设样本空间g和概率分布族F^*，决策空间为$\mathfrak{R}$，损失函数$L(\theta, d)$，$d(X)$为决策函数，则由下式确定的θ函数$R(\theta, d)$称为决策函数$d(X)$的风险函数

$$R(\theta, d) = E_\theta\left[L\left(\theta, d(X)\right)\right] = E_\theta[L(\theta, d(X_1, \cdots, X_n))]$$

- $R(\theta,d)$表示当参数为θ时，采用决策（行动）d所遭受的平均损失，其中E_θ表示当参数为θ时，对样本的函数$L(\theta,d(X))$求数学期望，显然风险越小，即损失越小决策函数就越好。但是，对于给定的决策函数$d(X)$，风险函数仍是θ的函数，所以，两个决策函数风险大小的比较涉及两个函数的比较，情况比较复杂，因此就产生了种种优良性准则，下面仅介绍两种。

定义3.3 设$d_1(X)$和$d_2(X)$是统计决策问题中的两个决策函数，若其风险函数满足不等式

$$R(\theta, d_1) \leq R(\theta, d_2), \qquad \forall \theta \epsilon \Theta$$

且存在一些θ使上述严格不等式$R(\theta, d_1) < R(\theta, d_2)$成立，则称决策函数$d_1(X)$一致优于$d_2(X)$。

假如下列关系式成立

$$R(\theta, d_1) = R(\theta, d_2), \qquad \forall \theta \epsilon \Theta$$

则称决策函数$d_1(X)$与$d_2(X)$等价。

□定义3.4　设$D=\{d(X)\}$是一切定义在样本空间上取值于决策空间$\Re$上的决策函数的全体，若存在一个决策函数$d^*(X)(d^*(X)\epsilon D)$，使对任一个$d(X)\in D$，都有$R(\theta,d^*)\leq R(\theta,d),\quad \forall\theta\epsilon\Theta$

则称$d^*(X)$为（该决策函数类D的）一致最小风险决策函数，或称为一致最优决策函数。

□上述两个定义都是对某个给定的损失函数而言的，当损失函数改变了，相应的结论也可能随之而变。

□定义3.4的结论还是对某个决策函数类而言的。当决策函数类改变了，一致最优性可能就不具备了。

➢例3.4 设总体X服从正态分布$N(\mu,1)$，$\mu\in(-\infty,+\infty)$，$X=(X_1,X_2,\cdots,X_n)^T$为取自X的样本，欲估计未知参数μ，选取损失函数为

$$L(\mu,d)=(d-\mu)^2$$

则对μ的任一估计$d(X)$，风险函数为

$$R(\mu,d)=E_\mu[L(\mu,d)]=E_\mu(d-\mu)^2$$

若进一步要求$d(X)$是无偏估计，即$E_\mu[d(X)]=\mu$，则风险函数是

$$R(\mu,d)=E_\mu(d-Ed)^2=D_\mu(d(X))$$

即风险函数为估计量$d(X)$的方差。

若取$d(X)=\overline{X}$，则$R(\mu,d)=D\overline{X}=\frac{1}{n}$.

若取$d(X)=X_1$，则$R(\mu,d)=DX_1=1$.

显然，当n>1时，后者的风险比前者大，即$\overline{X}$优于X_1。

➢例3.5 设x_1和x_2是从下列分布获得的两个观察值

$$P\{X=\theta-1\}=P\{X=\theta+1\}=0.5,\quad \theta\in\Theta=R$$

现研究θ的估计问题，为此取决策空间$\Re=R$，取损失函数为

$$L(\theta,d)=1-I(d)$$

其中$I(d)$为示性函数，当$\mathrm{d}=\theta$时它为1，否则为0. 我们知道，从样本空间$\mathcal{g}=\{(x_1,x_2)\}$到决策空间$\Re$上的决策函数有许多。

现考察其中三个。

（1）$d_1(x_1, x_2) = (x_1 + x_2)/2$，其风险函数为

$$R(\theta, d_1) = 1 - P\{d_1 = \theta\} = 1 - P\{x_1 \neq x_2\} = 0.5, \ \forall\theta\epsilon\Theta$$

（2）$d_2(x_1, x_2) = x_1 - 1$，其风险函数为

$$R(\theta, d_2) = 1 - P\{d_2 = \theta\} = 1 - P\{x_1 = \theta + 1\} = 0.5, \ \forall\theta\epsilon\Theta$$

（3）$d_3(x_1, x_2) = \begin{cases} (x_1 + x_2)/2, x_1 \neq x_2 \\ x_1 - 1, x_1 = x_2 \end{cases}$其风险函数为

$$R(\theta, d_3) = 1 - P\{d_3 = \theta\} = 1 - P\{x_1 \neq x_2 \text{或} x_1 = \theta + 1\}$$

$$= 0.25, \ \forall\theta\epsilon\Theta$$

- 假如只限于考察这三个决策函数组成的类$D = \{d_1, d_2, d_3\}$，那么d_3是决策函数类中一致最优决策函数，当决策函数类扩大或损失函数改变时，d_3的最优性可能会消失。

统计决策与贝叶斯估计

第二节：统计决策中的常用分布族

- 胡政发，肖海霞，应用数理统计与随机过程，电子工业出版社，2021年第一版
- 师义民，徐伟，秦超英，许勇，数理统计，科学出版社，2015年第四版

在第1章中，我们介绍过一些分布族，它们是χ^2分布族$\{\chi_n^2: n \geq 1\}$，t分布族$\{t(n): n \geq 1\}$，F分布族$\{F(n_1, n_2): n_1 \geq 1, n_2 \geq 1\}$。

在统计决策中还经常会遇到Gamma分布族、贝塔分布族等。

Gamma分布族

定义3.5 若随机变量X的密度函数为

$$f(x;\alpha,\beta)=\begin{cases}\dfrac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, & x>0,\\ 0, & x\leq 0,\end{cases}$$

则称X服从Gamma分布，记作$X\sim\Gamma(\alpha,\beta)$，$\alpha>0$，$\beta>0$为参数

$$\Gamma(\alpha)=\int_0^{+\infty}x^{\alpha-1}e^{-x}\,dx$$

$\Gamma(\alpha+1)=\alpha\Gamma(\alpha),\Gamma(1)=\Gamma(0)=1,\Gamma(n+1)=n!,\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$。

Gamma分布族记作$\{\Gamma(\alpha,\beta)：\ \alpha>0,\ \beta>0\}$。

Gamma分布具有下列性质：

□性质1 若$X\sim\Gamma(\alpha,\beta)$，则

$$E(X^k)=\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)\beta^k}=\frac{(\alpha+k-1)(\alpha+k-2)\cdots\alpha}{\beta^k}$$

它的数学期望与方差分别为$EX=\frac{\alpha}{\beta}$，$D(X)=\frac{\alpha}{\beta^2}$

- **性质2** 若$X\sim\Gamma(\alpha,\beta)$，则X的特征函数为 $g(t)=\left(1-\frac{it}{\beta}\right)^{-\alpha}$

- **性质3** 若$X_j\sim\Gamma(\alpha_j,\beta),j=1,2,\cdots,n$,且诸$X_j$间相互独立，则

$$\sum_{j=1}^{n}X_j\sim\Gamma(\sum_{j=1}^{n}\alpha_j,\beta)$$

这个性质称为Gamma分布的可加性。

- **性质4** 若$X_1,X_2,\cdots,X_n$相互独立，同服从指数分布$e(\beta)$（即$\Gamma(1,\beta)$分布），则

$$\sum_{i=1}^{n}X_i\sim\Gamma(n,\beta)$$

性质5 若$X\sim\Gamma(\alpha,1)$，则$Y=\frac{X}{\beta}\sim\Gamma(\alpha,\beta)$

在Γ分布中，令$\alpha=\frac{n}{2}$，$\beta=\frac{1}{2}$则得到自由度为n的$\chi^2(n)$，即$\Gamma\left(\frac{n}{2},\frac{1}{2}\right)=\chi^2(n)$.

贝塔分布族

定义3. 6 若随机变量X的密度函数为

$$f(x;a,b)=\begin{cases}\dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1},0<x<1,\\ 0,\text{其他}\end{cases}$$

则称X服从β分布，记作$Be(a,b)$其中a>0, b>0是参数，β分布族记作$\{Be(a,b):\ a>0,\ b>0\}$。

◘关于β分布族我们作如下讨论

（1）β变量X的k阶矩为

$$E(X^k)=\frac{a(a+1)\cdots(a+k-1)}{(a+b)(a+b+1)\cdots(a+b+k-1)}=\frac{\Gamma(a+k)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+k)}$$

它的数学期望与方差分别为：

$$EX=\frac{a}{a+b},\quad DX=\frac{ab}{(a+b)^2(a+b+1)}$$

（2）设随机变量X与Y独立，$X\sim\Gamma(a,1)$，$Y\sim\Gamma(b,1)$，则$Z=\frac{X}{X+Y}\sim Be(a,b)$。

（3）若$X\sim\chi^2(n_1)$，$Y\sim\chi^2(n_2)$且相互独立，则$Z=\frac{X}{X+Y}\sim Be(\frac{n_1}{2},\frac{n_2}{2})$。

统计决策与贝叶斯估计

第三节：贝叶斯估计

- 胡政发，肖海霞，应用数理统计与随机过程，电子工业出版社，2021年第一版
- 师义民，徐伟，秦超英，许勇，数理统计，科学出版社，2015年第四版

- 在一个统计问题中，可供选择的决策函数往往很多，自然希望寻找使风险最小的决策函数，然而在这种意义下的最优决策函数往往是不存在的。这是因为风险函数$R(\theta,d)$是既依赖于参数θ又依赖于决策函数d的二元函数，它往往会使得在某些θ处决策函数d_1的风险函数值较小，而在另一些θ处决策函数d_2的风险函数值较小。要解决这个问题，就要建立一个整体指标比较准则。贝叶斯方法通过引进先验分布把两个风险函数的点点比较转化为用一个整体指标的比较来代替，从而可以决定优劣。

先验分布与后验分布

在第2章讨论参数估计问题时，我们都是把待估参数θ视为参数空间Θ中的一个未知常数（或常数向量），在估计时仅利用样本所提供的关于总体的信息，而没有利用关于θ其他信息。然而在许多实际问题中，往往在抽样前便对参数θ有所了解，这种在抽样前对未知参数θ所了解的信息，称为先验信息。

- 例3.6　某学生通过物理实验确定当地的重力加速度，测得如下数据（m/s^2）：

$$9.80，9.79, 9.78, 6.81, 6.80$$

问如何估计当地的重力加速度？

- 如果用样本均值$\bar{x} = 8.596$来估计，你一定会认为这个结果很差，这是因为在未做实验之前你对重力加速度已有了一个先验的认识，比如你已经知道它大致在9.80左右，误差最大不大超过0.1.因此，参数的先验信息对于正确估计参数往往是有益的。

- 要利用参数θ的先验信息，通常是将θ看作在参数空间Θ中取值的随机变量。在实际中这种作法可以有两种理解：一是从某一范围考察，参数确是随机的，如用p表示某工厂每日的废品率，尽管从某一天看，p确是一个未知常数，但从数天或更长一段时间看，每天的p会有一定变化，一般来说p的变化范围呈现一定的分布规律，我们可以利用这种分布规律来作为某日废品率估计的先验信息；另一种理解是参数可能确是某一常数，但人们无法知道或无法准确知道它，只可能通过它的观测值去认识它，像例3. 6中的当地重力加速度.

- 这时，我们不妨把它看成一个随机变量，认为它所服从的分布可以通过它的先验知识获得。例如，可以认为当地的重力加速度服从正态分布$N(9.80, 0.1^2)$。这一观点在实际中是很有用处的。它将使我们能够充分地利用参数的先验信息对参数作出更准确的估计。
- 贝叶斯估计方法就是把未知参数θ视为一个已知分布$\pi(\theta)$的随机变量，从而将先验信息数学形式化并加以利用的一种方法，通常称$\pi(\theta)$为先验分布。先验分布$\pi(\theta)$与其他分布一样也有离散型和连续型之分，这要视θ是离散型随机变量还是连续型随机变量而定。

- 设总体X的分布密度为$p(x,\theta)$，$\theta \in \Theta$，θ的先验分布为$\pi(\theta)$，由于θ为随机变量并假定已知θ的先验分布，所以总体X的分布密度$p(x,\theta)$应看作给定θ时X的条件分布密度，于是总体X的分布密度$p(x,\theta)$需改用$p(x|\theta)$来表示。
- 设$\mathrm{X} = (X_1, X_2, \cdots, X_n)^T$为取自总体X的一个样本，当给定样本值$\mathrm{x} = (x_1, x_2, \cdots, x_n)^T$时，样本$\mathrm{X} = (X_1, X_2, \cdots, X_n)^T$的联合密度为

$$q(x_1, x_2, \cdots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

或表示为

$$q(x\,|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

由此，样本X和θ的联合概率分布为

$$f(x,\theta) = q(x\,|\theta)\pi(\theta)$$

由乘法公式知

$$f(x,\theta) = \pi(\theta)q(x\,|\theta) = m(x)h(\theta|x)$$

于是有

$$h(\theta|x) = \frac{\pi(\theta)q(x\,|\theta)}{m(x)}$$

称$h(\theta|x)$为给定样本X=x时，θ的后验分布，它是给定样本后θ的条件分布。其中m(x)是（X,θ）关于样本X的边缘分布。

- 如果θ是连续型随机变量，则

$$m(x)=\int_{\Theta} q(x|\theta)\pi(\theta)\,d\theta$$

- 如果θ是离散型随机变量，则

$$m(x)=\sum_{\Theta} q(x|\theta)\pi(\theta)$$

- 贝叶斯估计方法认为后验分布集中体现了样本和先验分布两者所提供的关于总体信息的总和，因而估计应建立在后验分布的基础上来进行。

共轭先验分布

前面提到后验分布在贝叶斯统计中起着重要作用，然而，在某些场合后验分布的计算较为复杂，为了简便地计算参数θ_1的后验分布，我们引入共轭先验分布的概念。

定义3.7 设总体X的分布密度为$p(x|\theta)$，F^*为θ的一个分布族，$\pi(\theta)$为θ的任意一个先验分布，$\pi(\theta) \in F^*$，若对样本的任意观察值x，θ的后验分布$h(\theta|x)$仍在分布族F^*内，则称F^*是关于分布密度$p(x|\theta)$的共轭先验分布族，或简称为共轭族。

- 应当注意，共轭先验分布是对某分布中的参数而言的，如正态均值、正态方差、泊松均值等。离开指定的参数及所在的分布去谈共轭先验分布是没有意义的。
- 引入共轭分布族后，使得数学运算较为简便，因为当θ的先验分布为共轭分布时，其后验分布也属于同一类型，这一点使得在共轭先验分布下，贝叶斯估计问题易于处理。
- 在实际中，如何获得参数θ的共轭先验分布，是我们关心的一个重要问题。为此我们引入后验分布核的概念，随后介绍两种计算共轭先验分布的求法。

- 当给定样本的分布（或称为似然函数）$q(x|\theta)$和先验分布$\pi(\theta)$后，由贝叶斯公式知θ的后验分布为

$$h(\theta|x) = \pi(\theta)q(x|\theta)/m(x)$$

其中$m(x)$为样本$\mathrm{X} = (X_1, X_2, \cdots, X_n)^T$的边缘分布。由于$m(x)$不依赖于$\theta$，在计算$\theta$的后验分布中仅起到一个正则化因子的作用，若把$m(x)$省略，可将贝叶斯公式改写为如下等价形式

$$h(\theta|x) \propto \pi(\theta)q(x|\theta)$$

- 其中符号“$\propto$”表示两步仅差一个不依赖于θ的常数因子。（3.9）式的右端虽不是正常的密度函数，但它是后验分布$h(\theta|x)$的主要部分，称为$h(\theta|x)$的核。
- 共轭先验分布可以用下述方法获得：首先求出似然函数$q(x|\theta)$，根据$q(x|\theta)$中所含的θ因式情况，选取与似然函数（θ的函数）具有相同核的分布作为先验分布，这个分布往往就是共轭先验分布。

例3.8 设$(X_1, X_2, \cdots, X_n)^T$是来自正态分布$N(\theta, \sigma^2)$的一个样本，其中$\theta$已知，现要寻求方差$\sigma^2$的共轭先验分布。由于该样本的似然函数为

$$q(x|\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\}$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\}$$

设X服从Γ分布$\Gamma(\alpha, \lambda)$，其中$\alpha > 0$为形状参数，$\lambda > 0$为尺度参数，其密度函数为

$$p(x|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, x > 0$$

通过概率运算可以求得$Y = X^{-1}$的密度函数为

$$p(y|\alpha,\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}\left(\frac{1}{y}\right)^{\alpha-1} e^{-\lambda/y}, y > 0$$

该分布称为逆Γ分布IΓ(α,λ)。假如取此逆Γ分布为σ^2的先验分布，其中参数α与λ已知，则其密度函数为

$$\pi(\sigma^2) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1} e^{-\lambda/\sigma^2}, \quad \sigma^2 > 0$$

于是后验分布为：

$$h(\sigma^2|x) \propto \pi(\sigma^2)q(x|\sigma^2)$$

$$\propto \sum_{i=1}^{n}(x_i-\theta)^2\left(\frac{1}{\sigma^2}\right)^{\alpha+\frac{n}{2}+1} \exp\{-\frac{1}{\sigma^2}\left[\lambda+\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right]\}$$

例3.9 设总体X服从二项分布$B(N,\theta)$，$(X_1,X_2,\cdots,X_n)^T$为取自X的样本，其似然函数为

$$q(x|\theta)=\prod_{i=1}^{n}C_N^{x_i}\theta^{x_i}(1-\theta)^{N-x_i}\propto\theta^{\sum_{i=1}^{n}x_i}(1-\theta)^{nN-\sum_{i=1}^{n}x_i},$$

$$x_i=0,1,\cdots,N,$$

$q(x|\theta)$中所含θ的因式为贝塔分布的核，从而设θ的先验分布为贝塔分布$Be(\alpha,\beta)$，其核为$\theta^{\alpha-1}(1-\theta)^{\beta-1}$，其中$\alpha,\beta$已知。于是可写出$\theta$的后验分布密度为

$$h(\theta|x)=\frac{\Gamma(\alpha+\beta+nN)\theta^{\alpha+\sum_{i=1}^{n}x_i-1}(1-\theta)^{\beta+nN-\sum_{i=1}^{n}x_i-1}}{\Gamma(\alpha+\sum_{i=1}^{n}x_i-1)\Gamma(\beta+nN-\sum_{i=1}^{n}x_i-1)},0<\theta<1$$

- 当参数θ存在适当的充分统计量时，一般可用下面的方法构造共轭先验分布族。

- 设总体X的分布密度为$p(x|\theta)$，$(X_1, X_2, \cdots, X_n)^T$为取自X的样本，$T = T(X) = T(X_1, \cdots, X_n)$是参数$\theta$的充分统计量，则由因子分解定理有

$$\prod_{i=1}^{n} p(x_i|\theta) = g_n(T|\theta)h(x_1, \cdots, x_n)$$

其中$\mathrm{h}(x_1, \cdots, x_n)$与$\theta$无关，$T = T(X_1, \cdots, X_n)$。

定理3.1 设$f(\theta)$为任一固定的函数，满足条件

（1）$f(\theta) \geq 0, \theta \in \Theta$

（2）$0 < \int_{\Theta} g_n(\mathrm{T}|\theta)\, f(\theta) d\theta < \infty$

则
$$D_J = \left\{ \frac{g_n(T|\theta) f(\theta)}{\int_{\Theta} g_n(\mathrm{T}|\theta)\, f(\theta) d\theta} : n = 1,2,\cdots \right\}$$

是共轭先验分布族。

例3.10 设总体X服从两点分布$B(1,\theta)$，其分布为$p(x|\theta) = \theta^x(1-\theta)^{1-x}$，x=0, 1，$(X_1, X_2, \cdots, X_n)^T$为取自总体X的一个样本，则似然函数为

$$q(x|\theta)=\prod_{i=1}^{n}p(x_i|\theta)=\theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}}=g_n(\mathrm{T}|\theta)\cdot 1$$

其中$T=n\bar{x}$，$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，$g_n(T|\theta)=\theta^T(1-\theta)^{n-T}$所以$T=n\bar{X}$是充分统计量，取$f(\theta)=1$，则

$$D_J=\left\{\frac{\theta^T(1-\theta)^{n-T}}{\int_0^1\theta^T(1-\theta)^{n-T}\,d\theta}:n=1,2,\cdots,\mathrm{T}=0,1,2,\cdots,n\right\}$$

是共轭先验分布族。

容易看出D_J是贝塔分布族的子族。可以证明贝塔分布族的全体$\{Be(a,b):a>0,b>0\}$仍是共轭先验分布族，其中$Be(a,b)$的密度

为$p(\theta)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1},0<\theta<1$

我们将常用的共轭先验分布列于表3.1

总体分布	参数	共轭先验分布
二项分布	成功概率	贝塔分布$Be(\alpha,\beta)$
泊松分布	均值	Γ分布$\Gamma(\alpha,\lambda)$
指数分布	均值的倒数	Γ分布$\Gamma(\alpha,\lambda)$
正态分布（方差已知）	均值	正态分布$N(\mu,\tau^2)$
正态分布（均值已知）	方差	逆Γ分布$I\Gamma(\alpha,\lambda)$

贝叶斯风险

将参数θ视为Θ上具有先验分布$\pi(\theta)$的随机变量后，风险函数$R(\theta,d)$可写为

$$R(\theta,d)=E_\theta[L(\theta,d(X))]=\int_C L(\theta,d(x))q(x|\theta)\,dx$$

它是θ的函数，仍是随机变量，关于θ再求期望，得

$$R_B(d)=E[R(\theta,d)]=\int_\Theta R(\theta,d)\pi(\theta)\,d\theta$$

$R_B(d)$称为决策函数d在给定先验分布$\pi(\theta)$下的贝叶斯风险，简称d的贝叶斯风险。

- 当总体X和θ都是连续型随机变量时，上式可写为

$$R_B(d) = \int_{\Theta} R(\theta, d)\pi(\theta)\, d\theta = \int_{\Theta}\int_{C} L(\theta, d(x))q(x|\theta)\pi(\theta)\, dxd\theta$$

$$= \int_{\Theta}\int_{C} L(\theta, d(x))m(x)h(\theta|x)\, dxd\theta = \int_{C} m(x)\{\int_{\Theta} L(\theta, d(x))\, h(\theta|x)d\theta\}\, dx$$

- 当总体X和都是离散型随机变量时，有

$$R_B(d) = \sum_{C} m(x)\,\{L\big(\theta, d(x)\big)h(\theta|x)\}$$

- 由上式可见，贝叶斯风险可看作是随机损失函数$L(\theta, d(X))$求两次期望而得到的，即第一次先对θ的后验分布求期望，第二次关于样本X的边缘分布求期望。此时，由于$R_B(d)$已不依赖于参数θ而仅依赖于决策函数d(X)，因此，以贝叶斯风险的大小作为衡量决策函数优劣的标准是合理的。

- 贝叶斯估计

- 定义3.8 设总体X的分布函数$F(x, \theta)$中参数θ为随机变量，$\pi(\theta)$为θ的先验分布。若在决策函数类D中存在一个决策函数$d^*(X)$，使得对决策函数类D中任一决策函数d(X)，均有

$$R_B(d^*) = inf R_B(d), \forall d \in D$$

则称$d^*(X)$为参数θ的贝叶斯估计量。

- 由定义可见，贝叶斯估计量$d^*(X)$就是使贝叶斯风险$R_B(d)$达到最小的决策函数。应该注意，贝叶斯估计量是依赖于先验分布$\pi(\theta)$的，即对于不同的$\pi(\theta)$，θ的贝叶斯估计量是不同的，在常用损失函数下，贝叶斯估计有如下几个结论。
- 定理3.2 设θ的先验分布为$\pi(\theta)$和损失函数为$L(\theta, d) = (\theta - d)^2$，则$\theta$的贝叶斯估计是 $d^*(x) = E(\theta|X = x) = \int_{\Theta} \theta h(\theta|x)\, d\theta$

其中$h(\theta|x)$为参数θ的后验密度。

定理3.3 设θ的先验分布为$\pi(\theta)$，取损失函数为加权平方损失函数

$$L(\theta, d) = \lambda(\theta)(d-\theta)^2$$

则θ的贝叶斯估计为

$$d^*(x) = \frac{E(\lambda(\theta)\theta|x)}{E(\lambda(\theta)|x)}$$

定义3.9 设d=d(x)为决策函数类D中任一个决策函数，损失函数为$L(\theta, d(x))$，则$L(\theta, d(x))$对后验分布$h(\theta|x)$的数学期望称为**后验风险**，记为

$$R(d|x) = E[L(\theta, d(x))|x] = \begin{cases} \int_{\Theta} L(\theta, d(x))h(\theta|x)d\theta\,, \text{当}\theta\text{为连续型变量} \\ \sum_{i} L(\theta_i, d(x))h(\theta_i|x)\,, \text{当}\theta\text{为离散型变量} \end{cases}$$

假如在D中存在这样一个决策函数$d^*(x)$，使得

$$R(d^*|x) = \inf_d R(d|x), \forall d \in D$$

则称$d^*(x)$为该统计决策问题在后验风险准则下的最优决策函数，或称为贝叶斯（后验型）决策函数，在估计问题中，它又称为贝叶斯（后验型）估计。

□定理3.4 设参数θ为随机向量，$\theta=(\theta_1,\cdots,\theta_p)^T$，对给定的先验分布$\pi(\theta)$和二次损失函数

$$L(\theta,d)=(d-\theta)^TQ(d-\theta)$$

其中Q为正定矩阵，则θ的贝叶斯估计为后验分布$h(\theta|x)$的均值向量，即

$$d^*(x)=E(\theta|x)=\begin{bmatrix}E(\theta_1|x)\\ \vdots \\ E(\theta_p|x)\end{bmatrix}$$

□这个结论表明，在正定二次损失下，θ的贝叶斯估计不受正定矩阵Q的选取干扰，这一特性常被称为θ的贝叶斯估计关于Q是稳健的。

定理3.5 对给定的统计决策问题（包括先验分布给定的情形）和决策函数类D，当贝叶斯风险满足如下条件

$$\inf_d R_B(d) < \infty, \forall d \in D$$

则贝叶斯决策函数$d^*(x)$与贝叶斯后验型决策函数$d^{**}(x)$是等价的。即使后验风险最小的决策函数 $d^{**}(x)$也使贝叶斯风险最小。反之使贝叶斯风险最小的决策函数$d^*(x)$同时也使后验风险最小。

定理3.6 设θ的先验分布为$\pi(\theta)$，损失函数为绝对值损失

$$L(\theta, d) = |d - \theta|$$

则θ的贝叶斯估计$d^*(x)$为后验分布$h(\theta|x)$的中位数。

□定理3.7 线性损失函数

$$L(\theta,d)=\begin{cases}k_0(\theta-d), d\le\theta\\ k_1(d-\theta), d>\theta\end{cases}$$

下，θ的贝叶斯估计$d^*(x)$为后验分布$h(\theta|x)$的$\frac{k_1}{k_0+k_1}$上侧分位数。

例3.11 设总体X服从两点分布$B(1,p)$，其中参数p未知而p在[0,1]上服从均匀分布，$(X_1,X_2,\cdots,X_n)^T$是来自X的样本。假定损失函数是二次损失函数$L(p,d)=(p-d)^2$，求参数p的贝叶斯估计及贝叶斯风险。

解 由定理3.2知，当损失函数为二次损失函数时，欲求p的贝叶斯估计需先求p的后验分布$h(p|x)=q(x|p)\pi(p)/m(x)$。

由于给定p,X的条件概率是$q(x|p)=p^x(1-p)^{1-x}$，其中x=0,1.所以$(X_1,X_2,\cdots,X_n)^T$的条件概率是

$$q(x|p)=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}=p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i}$$

而p的先验概率密度为$\pi(p)=1, p\in[0,1]$，所以$(X_1,X_2,\cdots,X_n)^T$与p的联合密度为 $f(x,p)=p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$

$(X_1,X_2,\cdots,X_n)^T$的边缘分布是

$$m(x)=\int_0^1 p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}\,dp=\beta\left(\sum_{i=1}^n x_i+1, n+1-\sum_{i=1}^n x_i\right)$$

$$=\left(\sum_{i=1}^n x_i\right)!\left(n-\sum_{i=1}^n x_i\right)!/(n+1)!$$

最后两个等号成立是根据$\beta(p,q)=\int_0^1 x^{p-1}(1-x)^{q-1}dx$和$\beta(p,q)=\Gamma(p)\Gamma(q)/\Gamma(p+q)$,$\Gamma(n+1)=n!$而得。所以p的后验分布为

$$h(p|x)=\frac{f(x,p)}{m(x)}=\frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}{\left[\left(\sum_{i=1}^n x_i\right)!\left(n-\sum_{i=1}^n x_i\right)!\right]/(n+1)!}$$

$$=\frac{(n+1)!}{\left(\sum_{i=1}^n x_i\right)!\left(n-\sum_{i=1}^n x_i\right)!}p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$$

因此p的贝叶斯估计是

$$\hat{p} = \int_0^1 p h(p|x) dp$$

$$= \int_0^1 \frac{(n+1)!}{\left(\sum_{i=1}^n x_i\right)!\left(n - \sum_{i=1}^n x_i\right)!} p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i} dp$$

$$= \frac{(n+1)!}{\left(\sum_{i=1}^n x_i\right)!\left(n - \sum_{i=1}^n x_i\right)!} \cdot \frac{\left[\left(\sum_{i=1}^n x_i + 1\right)!\left(n - \sum_{i=1}^n x_i\right)!\right]}{(n+2)!} = \frac{\sum x_i + 1}{n+2}$$

这个估计的贝叶斯风险为

$$R_B(\hat{p}) = \int_0^1 E[L(p,d)|p]\pi(p)dp = \int_0^1 E(\hat{p}-p)^2 dp$$

$$= \int_0^1 E[\frac{\sum_{i=1}^n X_i + 1}{n+2} - p]^2 dp$$

$$= \frac{1}{(n+2)^2} \int_0^1 E\left[\sum_{i=1}^n X_i + 1 - (n+2)p\right]^2 dp$$

而$E\left[\sum_{i=1}^n X_i + 1 - (n+2)p\right]^2 = E[Y - np + 1 - 2p]^2$，其中服从二项分布B(n,p)，再把上式平方展开并分别求期望得

$$E\left[\sum_{i=1}^{n} X_i + 1 - (n+2)p\right]^2 = np(1-p) + (1-2p)^2$$

所以

$$R_B(\hat{p}) = \frac{1}{(n+2)^2}\int_0^1 [np(1-p) + (1-2p)^2]dp$$

$$= \frac{1}{(n+2)^2}\int_0^1 [(4-n)p^2 + (n-4)p + 1]dp$$

$$= \frac{1}{(n+2)^2}\left(\frac{4-n}{3} + \frac{n-4}{2} + 1\right) = \frac{1}{6(n+2)}$$

附带说明一点，对于p的最大似然估计 $\hat{p}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$，可求出其贝叶斯风险为$1/6n$。

例3.12 假设总体X服从正态分布$N(\mu,1)$，其中参数μ是未知的，假定μ服从正态分布$N(0,1)$，并假设$(X_1,X_2,\cdots,X_n)^T$是来自该总体的样本。对于给定的损失函数$L(\mu,d)=(\mu-d)^2$，试求μ的贝叶斯估计量。

解 给定μ，$(X_1,X_2,\cdots,X_n)^T$的条件分布密度为

$$q(x_1,x_2,\cdots,x_n|\mu)=\frac{1}{(\sqrt{2\pi})^n}\exp\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\}$$

$(X_1,X_2,\cdots,X_n)^T$与μ的联合密度是

$$f(x,\mu)=\frac{1}{(2\pi)^{\frac{n+1}{2}}}\exp\{-\frac{1}{2}[\sum_{i=1}^{n}x_i^2+(n+1)\mu^2-2\mu n\bar{x}]\}$$

$(X_1, X_2, \cdots, X_n)^T$的边缘分布密度为

$$m(x) = \int_{-\infty}^{+\infty} f(x,\mu)\, d\mu = \int_{-\infty}^{+\infty} \frac{1}{(2\pi)^{\frac{n+1}{2}}} \exp\{-\frac{1}{2}[\sum_{i=1}^{n} x_i^2 + (n+1)\mu^2 - 2\mu n\bar{x}]\} d\mu$$

$$= \frac{1}{(2\pi)^{\frac{n+1}{2}}} \exp\{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\} \int_{-\infty}^{+\infty} \exp\{-\frac{1}{2}[(n+1)\mu^2 - 2\mu n\bar{x}]\} d\mu$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\{-\frac{1}{2}[\sum_{i=1}^{n} x_i^2 - \frac{n^2}{n+1} x^2]\} \left(\frac{1}{n+1}\right)^{\frac{1}{2}}$$

于是μ的后验分布密度是

$$\mathrm{h}(\mu|x)=\frac{f(x,\mu)}{m(x)}=\left(\frac{n+1}{2\pi}\right)^{\frac{1}{2}}\exp\{-\frac{n+1}{2}\left(\mu-\frac{n\bar{x}}{n+1}\right)^{2}\}$$

于是μ的贝叶斯估计为

$$\hat{\mu}=\int_{-\infty}^{+\infty}\mu\,h(\mu|x)d\mu$$

$$=\frac{\sqrt{n+1}}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\mu\exp\{-\frac{n+1}{2}\left(\mu-\frac{n\bar{x}}{n+1}\right)^{2}\}d\mu=\frac{n\bar{x}}{n+1}$$

$$=\frac{1}{n+1}\sum_{i=1}^{n}x_i$$

若X服从$N(\mu,1)$，μ服从$N(0,k^2)$，$L(\mu,d)=(\mu-d)^2$的贝叶斯估计为 $\hat{\mu}_k=\frac{k^2}{1+nk^2}\sum_{i=1}^{n}x_i$ 贝叶斯风险为 $B(\hat{\mu}_k)=\frac{k^2}{1+nk^2}$。

- 由上所述可知，构造贝叶斯估计量主要取决两点：参数的先验分布和损失函数。在满足一定的条件下，可以证明贝叶斯估计量具有一致性，渐近正态性和渐近有效性。

- 例3.13 设$X=(X_1,X_2,\cdots,X_n)^T$是来自均匀分布$U(0,\theta)$的一个样本，又设θ的先验分布为pareto分布，其分布函数与密度函数分别为

$$F(\theta)=1-\left(\frac{\theta_0}{\theta}\right)^{\alpha},\theta\geq\theta_0,\pi(\theta)=\frac{\alpha\theta_0^{\alpha}}{\theta^{\alpha+1}},\theta\geq\theta_0$$

其中$0<\alpha<1$和$\theta_0>0$为已知。该分布记作$Pa(\alpha,\theta_0)$。θ的数学期望$\mathrm{E}(\theta)=\alpha\theta_0/(\alpha-1)$。在上述假设下，样本X与$\theta$的联合分布为

$$f(x,\theta)=\frac{\alpha\theta_0^{\alpha}}{\theta^{\alpha+n+1}},0<x_i<\theta,i=1,2,\cdots,n,0<\theta_0<\theta$$

设$\theta_1=\max(x_1,x_2,\cdots,x_n,\theta_0)$，则样本X的边缘分布为

$$m(x)=\int_{\theta_1}^{\infty}\frac{\alpha\theta_0^{\alpha}}{\theta^{\alpha+n+1}}d\theta=\frac{\alpha\theta_0^{\alpha}}{(\alpha+n)\theta_1^{\alpha+n}},0<x_i<\theta_1$$

由此可得θ的后验密度函数

$$h(\theta|x)=\frac{f(x,\theta)}{m(x)}=\frac{(\alpha+n)\theta_1^{\alpha+n}}{\theta^{\alpha+n+1}},\ \theta>\theta_1$$

这仍是pareto分布$Pa(\alpha+n,\theta_1)$。

在绝对值损失下，θ的贝叶斯估计 $\hat{\theta}_B$是后验分布的中位数，即$\hat{\theta}_B$是下列方程的解，

$$F(\theta|x)=\int_{\Theta} h(\theta|x)d\theta=\int_{\theta_1}^{\hat{\theta}_B}\frac{(\alpha+n)\theta_1^{\alpha+n}}{\theta^{\alpha+n+1}}d\theta=1-\left(\frac{\theta_1}{\hat{\theta}_B}\right)^{\alpha+n}=\frac{1}{2}$$

解之可得$\hat{\theta}_B=\frac{1}{2^{\alpha+n}}\theta_1$。

若取平方损失函数，则θ的贝叶斯估计$\hat{\theta}_{B_1}$是后验均值，即

$$\hat{\theta}_{B_1} = \frac{\alpha + n}{\alpha + n - 1} \max(x_1, \cdots, x_n, \theta_0)$$

- 定理3.8 设参数θ的先验分布为$\pi(\theta)$，$g(\theta)$为θ的连续函数，则在平方损失函数下，g(θ)的贝叶斯估计为$d(x) = E[g(\theta)|x]$。

□例3.14设$X=(X_1,X_2,\cdots,X_n)^T$是来自Γ分布$\Gamma(r,\theta)$的一个样本，

其中r已知。其期望$EX=\frac{r}{\theta}$与θ^{-1}成正比。通常人们对θ^{-1}有兴趣，现求θ^{-1}的贝叶斯估计。为此取Γ分布$\Gamma(\alpha,\beta)$作为θ的先验分布。容易获得θ的后验分布。

$$h(\theta|x)\propto\theta^{\alpha+n-1}e^{-\theta\left(\sum_{i=1}^{n}x_i+\beta\right)},\theta>0$$

若取如下平方损失函数

$$L(\theta,d)=(d-\frac{1}{\theta})^2$$

则θ^{-1}的贝叶斯估计为

$$\hat{\theta}_B = E(\theta^{-1}|x) = \frac{\left(\sum_{i=1}^{n} x_i + \beta\right)^{\alpha+nr}}{\Gamma(\alpha+nr)} \int_0^{+\infty} \frac{1}{\theta} \theta^{\alpha+nr-1} e^{-\theta\left(\sum_{i=1}^{n} x_i + \beta\right)} d\theta$$

$$= \left(\sum_{i=1}^{n} x_i + \beta\right)/(\alpha + nr - 1)$$

若取如下损失函数:$L(\theta, d) = \theta^2(d - \frac{1}{\theta})^2$,这时$\theta^{-1}$的贝叶斯估计为

$$\hat{\theta}_B = \frac{E(\theta^2\theta^{-1}|x)}{E(\theta^2|x)} = \frac{\int_0^{+\infty} \theta^{\alpha+nr} e^{-\theta\left(\sum_{i=1}^{n} x_i + \beta\right)} d\theta}{\int_0^{+\infty} \theta^{\alpha+nr-1} e^{-\theta\left(\sum_{i=1}^{n} x_i + \beta\right)} d\theta} = \frac{\sum_{i=1}^{n} x_i + \beta}{\alpha + nr + 1}$$

$$= \frac{\alpha + nr - 1}{\alpha + nr + 1} \hat{\theta}_B^{-1}$$

贝叶斯估计的误差

- 设$\hat{\theta}$是θ的一个贝叶斯估计，在给定样本后，$\hat{\theta}$是一个数，在综合各种信息后θ是按后验分布$h(\theta|x)$取值，所以评定一个贝叶斯估计的误差的最好而又简便的方法是用θ对$\hat{\theta}$的后验均方差或其平方根来度量，具体定义如下：

- 定义3.10 设参数θ的后验分布为$h(\theta|x)$，θ的贝叶斯估计为$\hat{\theta}$，则$(\hat{\theta}-\theta)^2$的后验期望

$$MSE(\hat{\theta}|x)=E_{\theta|x}(\hat{\theta}-\theta)^2$$

称为$\hat{\theta}$的后验均方差，而其平方根$MSE(\hat{\theta}|x)^{\frac{1}{2}}$称为$\hat{\theta}$的后验标准误差，其中符号$E_{\theta|x}$表示对条件分布$h(\theta|x)$求期望。估计量$\hat{\theta}$的后验均方差越小，贝叶斯估计的误差就有越小。当$\hat{\theta}$为$\theta$的后验期望$\hat{\theta}_E = E(\theta|x)$时，有 $MSE(\hat{\theta}_E|x) = E_{\theta|x}(\hat{\theta}_E - \theta)^2 = Var(\theta|x)$

称为后验方差，其平方根$[Var(\theta|x)\ ]^{\frac{1}{2}}$称为后验标准差。

后验均方差与后验方差有如下关系：

$$MSE(\hat{\theta}|x) = E_{\theta|x}(\hat{\theta} - \theta)^2 = E_{\theta|x}[(\hat{\theta} - \hat{\theta}_E) + (\hat{\theta}_E - \theta)]^2$$

$$= E_{\theta|x}(\hat{\theta}_E - \hat{\theta})^2 + Var(\theta|x) = (\hat{\theta}_E - \hat{\theta})^2 + Var(\theta|x)$$

这表明，当$\hat{\theta}$为后验均值$\hat{\theta}_E = E(\theta|x)$时，可使后验均方差达到最小，所以在实际中常常取后验均值作为θ的贝叶斯估计值。

- 从这个定义还可以看出，后验方差及后验均方差只依赖于样本X，不依赖于θ，故当样本给定后，它们都是确定的实数，立即可以应用。
- 在经典统计中，估计量的方差常常还依赖于被估参数θ，使用时常用估计$\hat{\theta}$去代替θ，获得其近似方差才可应用。另外在计算上，后验方差的计算在本质上不会比后验均值的计算复杂许多。

因为它们都用同一个后验分布计算。而在经典统计中，估计量的方差计算有时还要涉及抽样分布（估计量的分布）。我们知道，寻求抽样分布在经典统计学中常常是一个困难的数学问题，然而，在贝叶斯估计中从不涉及寻求抽样分布问题，这是因为贝叶斯估计对未出现的样本不加考虑之故。

- 值得注意，在贝叶斯估计中不用无偏性来评价一个估计量的好坏。

这是因为在无偏估计的定义中，$E\hat{\theta}(X)=\theta$，其中$X=(X_1,X_2,\cdots,X_n)^T$为样本。这里，数学期望是对样本空间中所有可能样本X而求的，但在实际中绝大多数样本尚未出现过，

甚至重复数百次也不会出现的样本也要在评价评估量中占一席之地，这是不合理的。另一方面，在实际使用中不少估计量只使用一次或数次，所以贝叶斯学派认为，评价一个估计量的好坏只能依据在试验中所收集到的观察值，不应该使用尚未观察到的数据。这一观点被贝叶斯学派称为“条件观点”。据此，估计的无偏性在贝叶斯估计中不予考虑。

区间估计

前面曾经提到，后验分布在贝叶斯统计中占有重要地位，当求得参数θ的后验分布$h(\theta|x)$以后，我们可以计算θ落在某区间[a,b]内的后验概率$P\{a \le \theta \le b|x\}$，当$\theta$为连续型变量，且其后验概率为$1-\alpha(0<\alpha<1)$时，我们有等式

$$P\{a \le \theta \le b|x\} = 1-\alpha$$

反之若给定概率$1-\alpha$，要找一个区间[a,b]，使上式成立，这样求得的区间称为θ的贝叶斯区间估计。

又称为贝叶斯置信区间。

□当θ为离散型随机变量时，对给定的概率$1-\alpha$，满足上式的区间不一定存在，这时只要略微放大上式左端概率，才能找到a与b，使得

$$P\{a \le \theta \le b|x\} = 1-\alpha$$

这样的区间[a,b]，也称为θ的贝叶斯区间估计。下面给出参数θ的贝叶斯区间估计的一般定义。

□定义3.11 设参数θ的后验分布为$h(\theta|x)$，对给定的样本$X=(X_1,X_2,\cdots,X_n)^T$和概率$1-\alpha(0<\alpha<1)$，若存在两个统计量$\hat{\theta}_L=\hat{\theta}_L(X)$和$\hat{\theta}_U=\hat{\theta}_U(X)$，使得

$$P\{\hat{\theta}_L \le \theta \le \hat{\theta}_U|x\} \ge 1-\alpha$$

则称区间$[\hat{\theta}_L, \hat{\theta}_U]$为参数$\theta$的置信度为$1-\alpha$的贝叶斯置信区间，或简称$\theta$的$1-\alpha$置信区间。而满足下式 的$\hat{\theta}_L$称为$\theta$的$1-\alpha$（单侧）置信下限：

$$P\{\theta \geq \hat{\theta}_L | x\} \geq 1-\alpha$$

满足下式的$\hat{\theta}_U$称为的$1-\alpha$（单侧）置信上限：

$$P\{\theta \leq \hat{\theta}_U | x\} \geq 1-\alpha$$

- 在经典统计学中寻求参数θ的置信区间有时是困难的，因为首先要设法构造一个函数（含有待估参数的随机变量），且使该函数的概率分布为已知，分布中不含任何未知参数，这是一项

技术性很强的工作，不熟悉“抽样分布”的人是很难完成的，但寻求参数θ的贝叶斯置信区间只要利用θ的后验分布，而不需要再去寻求另外的分布，二者相比，贝叶斯置信区间的寻求要简单得多。

□例3.15 设$X=(X_1,X_2,\cdots,X_n)^T$是来自正态总体$N(\theta,\sigma^2)$的一个样本，其中σ^2已知。取θ的先验分布为正态分布$N(\mu,\tau^2)$，则θ的密度函数为，

$$\pi(\theta)=\frac{1}{\sqrt{2\pi}\tau}\exp\{-\frac{1}{2\tau^2}(\theta-\mu)^2\},-\infty<\theta<+\infty$$

其中μ与τ^2为已知常数，由此可求得样本X与θ的联合密度函数为

$$f(x,\theta)=k_1\exp\{-\frac{1}{2}[\frac{1}{\sigma^2}\left(n\theta^2-2n\theta\bar{x}+\sum_{i=1}^{n}x_i^2\right)+\frac{1}{\tau^2}(\theta^2-2\mu\theta+\mu^2)\}$$

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

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

则有

$$f(x,\theta)=k_1\exp\{-\frac{1}{2}[A\theta^2-2\theta B+C]\}=k_2\exp\{-\frac{1}{2A^{-1}}(\theta-B/A)^2\}$$

其中$k_2 = k_1 \exp\{-\frac{1}{2}(\mathrm{C} - \frac{B^2}{A})\}$。由此容易算得样本X的边缘分布为

$$m(x) = \int_{-\infty}^{+\infty} f(x,\theta)\, d\theta = k_2 \left(\frac{2\pi}{A}\right)^{\frac{1}{2}}$$

因而θ的后验分布为

$$\mathrm{h}(\theta|x) = \frac{f(x,\theta)}{m(x)} = \left(\frac{A}{2\pi}\right)^{\frac{1}{2}} \exp\{-\frac{(\theta - B/A)^2}{2/A}\}$$

- 这正好是正态总体 $N(\mu_1, \sigma_1^2)$的密度函数。其中，

$$\mu_1 = \frac{B}{A} = \frac{\bar{x}\sigma_0^{-2}}{\sigma_0^{-2} + \tau^{-2}}, \sigma_1^2 = \frac{\sigma_0^2 \tau^2}{\sigma_0^2 + \tau^2}$$

据此可知$\frac{\theta-\mu_1}{\sigma_1}$服从标准正态分布$N(0,1)$，于是可得

$$P\left\{\left|\frac{\theta-\mu_1}{\sigma_1}\right| \leq u_{\frac{\alpha}{2}}\right\}=1-\alpha$$

即 $$P\left\{\mu_1-\sigma_1 u_{\frac{\alpha}{2}} \leq \theta \leq \mu_1+\sigma_1 u_{\frac{\alpha}{2}}\right\}=1-\alpha$$

其中$u_{\frac{\alpha}{2}}$为标准正态分布的上侧分位数$\frac{\alpha}{2}$。故可得θ的$1-\alpha$贝叶斯置信区间为 $$\left[\mu_1-\sigma_1 u_{\frac{\alpha}{2}}, \mu_1+\sigma_1 u_{\frac{\alpha}{2}}\right]$$

□例3.16 对某个儿童做智力测验，设测验结果$X\sim N(\theta,100)$，其中θ在心理学中定义为儿童的智商，根据多次测验，可设θ服从正态分布$N(100,225)$，应用例3.15的结论，当n=1时，可得在给定X=x条件下，该儿童智商θ的后验分布服从正态分布$N(\mu_1,\sigma_1^2)$，其中

$$\mu_1=\frac{100\cdot 100+225x}{100+225}=\frac{400+9x}{13}$$

$$\sigma_1^2=\frac{100\cdot 225}{100+225}=\frac{900}{13}=69.23=8.32^2$$

若该儿童在一次智商测验中得x=115，则可得其智商θ的后验分布为$N(110.38,8.32^2)$，于是有

$$p\{-u_{\alpha/2} \leq \frac{\theta - 110.38}{8.32} \leq u_{\alpha/2}\} = 1 - \alpha$$

其中$u_{\frac{\alpha}{2}}$为标准正态分布的上侧分位数。当给定$\alpha = 0.05$时，查正态分布数值表求得$u_{\frac{\alpha}{2}} = 1.96$，故有

$$p\{110.38 - 1.96 * 8.32 \leq \theta \leq 110.38 + 1.96 * 8.32\}$$

$$= \mathrm{P}\{94.07 \leq \theta \leq 126.69\} = 1 - \alpha = 0.95$$

- 在这个例子中，若不利用先验信息，仅利用当前抽样信息，则也可运用经典方法求出θ的置信区间。由于X服从正态分布$N(\theta, 100)$和 $\bar{x} = x = 115$，可求得的0.95置信区间为

$$[\bar{x} - \mu_{\alpha/2}\sigma, \bar{x} + \mu_{\alpha/2}\sigma] = [115 - 1.96 \cdot 10, 115 + 1.96 \cdot 10]$$
$$= [95.4, 134.6]$$

- 我们发现在上述问题中，置信度相同（均为0.95）但两个区间长度不同，贝叶斯置信区间的长度短一些（区间长度短时，估计的误差小），这是由于使用了先验分布之故。

统计决策与贝叶斯估计

第四节：minimax估计

- 胡政发，肖海霞，应用数理统计与随机过程，电子工业出版社，2021年第一版
- 师义民，徐伟，秦超英，许勇，数理统计，科学出版社，2015年第四版

- 风险函数提供了一个衡量决策函数好坏的尺度，我们自然希望选取一个决策函数，使得它的风险尽可能的小。
- 定义3.12 给定一个统计决策问题，设D^*是由全体决策函数组成的类，如果存在一个决策函数$d^* = d^*(x_1, \cdots, x_n), d^* \in D^*$，使得对$D^*$中任意一个决策函数$\mathrm{d}(x_1, \cdots, x_n)$，总有

$$\max_{\Theta} R(\theta, d^*) \leq \max_{\Theta} R(\theta, d), \qquad \forall d \in D^*$$

则称d^*为这个统计决策问题的最小最大决策函数（在这里我们假定R关于θ的最大值能达到，如果最大值达不到，可以理解为上确界）。

- 由定义可见，我们是以最大风险的大小作为衡量决策函数好坏的准则。因此，使最大风险达到最小的决策函数是考虑到最不利的情况，要求最不利的情况尽可能地好，也就是人们常说的从最坏处着想，争取最好的结果。它是一种出于稳妥的考虑，也是一种偏于保守的考虑。
- 如果我们讨论的问题是一个估计问题，则称满足式（3.12）的决策函数$d^*(X_1,\cdots,X_n)$为θ的最小最大估计量。

- 寻求最小最大决策函数的一般步骤为：
- 1.对D^*中每个决策函数$\mathrm{d}(x_1,\cdots,x_n)$，求出其风险函数在$\Theta$上的最大风险值$\max\limits_{\Theta} R(\theta,d)$；
- 2.在所有最大风险值中选取相对最小值，此值对应的决策函数便是最小最大决策函数。
- 例3.17 地质学家要根据某地区的地层结构来判断该地是否蕴藏石油。地层结构总是0,1两种状态之一，记该地无油为θ_0，该地有油为θ_1，已知它们的分布规律如表3.2所示（其中x表示地层结构的状态，θ表示石油的状态），它表示如果该地区蕴藏石油，

那么地层结构呈现状态0的概率为0.3，呈现状态1的概率为0.7，如果该地区不蕴藏石油，那么地层结构呈现状态0的概率为0.6，呈现状态1的概率为0.4，土地所有者希望根据地质学家对地层结构的分析来决定自己投资钻探石油，还是出卖土地所有权或者在该地区开辟旅游点，分别记这三种决策为$\alpha_1, \alpha_2, \alpha_3$，于是决策空间$\Re = \{\alpha_1, \alpha_2, \alpha_3\}$。土地所有者权衡利弊之后取损失函数$L(\theta, \alpha)$如表3.3所示。

表3.2 地层结构分布规律表

θ / x	0	1
θ_0（无油）	0.6	0.4
θ_1（有油）	0.3	0.7

表3.3 损失函数$L(\theta,\alpha)$取值表

θ \ a	α_1	α_2	α_3
θ_0（无油）	12	1	6
θ_1（有油）	0	10	5

假如我们仅取一个观察X_1（样本大小为1），如果土地所有者打算采用决策函数

$$d_4(x_1)=\begin{cases}\alpha_1, x_1=1\\ \alpha_2, x_1=0\end{cases}$$

那么风险函数$R(\theta,d_4)$在$\theta=\theta_0$处的值为

$$R(\theta_0, d_4) = L(\theta_0, \alpha_1)P_{\theta_0}\{X_1 = 1\} + L(\theta_0, \alpha_2)P_{\theta_0}\{X_1 = 0\}$$

$$= 12 \cdot 0.4 + 1 \cdot 0.6 = 5.4$$

在$\theta = \theta_1$处的值为

$$R(\theta_1, d_4) = L(\theta_1, \alpha_1)P_{\theta_1}\{X_1 = 1\} + L(\theta_1, \alpha_2)P_{\theta_1}\{X_1 = 0\}$$

$$= 0 \cdot 0.7 + 10 \cdot 0.3 = 3$$

如果土地所有者打算采用决策函数

$$d_7(X_1) = \begin{cases} \alpha_1, x_1 = 1 \\ \alpha_3, x_1 = 0 \end{cases}$$

那么风险函数$R(\theta, d_7)$在$\theta = \theta_0$处的值为

$$R(\theta_0, d_7) = L(\theta_0, \alpha_1)P_{\theta_0}\{X_1 = 1\} + L(\theta_0, \alpha_3)P_{\theta_0}\{X_1 = 0\}$$

$$= 0 \cdot 0.7 + 5 \cdot 0.3 = 1.5$$

在本例中，可供土地所有者选择的决策函数共有9个，将它们列于表3.4.

表3.4 决策函数表

$\boldsymbol{x}_1$	$\boldsymbol{d}_1(\boldsymbol{x}_1)$	$\boldsymbol{d}_2(\boldsymbol{x}_1)$	$\boldsymbol{d}_3(\boldsymbol{x}_1)$	$\boldsymbol{d}_4(\boldsymbol{x}_1)$	$\boldsymbol{d}_5(\boldsymbol{x}_1)$	$\boldsymbol{d}_6(\boldsymbol{x}_1)$	$\boldsymbol{d}_7(\boldsymbol{x}_1)$	$\boldsymbol{d}_8(\boldsymbol{x}_1)$	$\boldsymbol{d}_9(\boldsymbol{x}_1)$
0	α_1	α_1	α_1	α_2	α_2	α_2	α_3	α_3	α_3
1	α_1	α_2	α_3	α_1	α_2	α_3	α_1	α_2	α_3

我们在上面已经计算出决策函数$\boldsymbol{d}_4(\boldsymbol{x}_1)$与$\boldsymbol{d}_7(\boldsymbol{x}_1)$的风险函数，现把这9个决策函数的风险函数及其最大值列成表3.5.

表3.5 风险函数及最大值表

$\boldsymbol{d}_i(\boldsymbol{x}_1)$	$\boldsymbol{d}_1$	$\boldsymbol{d}_2$	$\boldsymbol{d}_3$	$\boldsymbol{d}_4$	$\boldsymbol{d}_5$	$\boldsymbol{d}_6$	$\boldsymbol{d}_7$	$\boldsymbol{d}_8$	$\boldsymbol{d}_9$
$R(\theta_0, d_i)$	12	7.6	9.6	5.4	1	3	8.4	4	6
$R(\theta_1, d_i)$	0	7	3.5	3	10	6.5	1.5	8.5	5
$\max_{\Theta} R(\theta, d_i$	12	7.6	9.6	5.4	10	6.5	8.4	8.5	6

- 如果土地所有者希望使得承担可能产生的最大风险尽量小，那么应当采用决策函数$\boldsymbol{d}_4(\boldsymbol{x}_1)$。由定义知$\boldsymbol{d}_4(\boldsymbol{x}_1)$是这个统计决策问题的minimax决策函数。

- 下面介绍如何借用贝叶斯方法来求最小最大决策函数。当然，使用贝叶斯方法必须预先引进未知参数θ的先验分布。但是，这里仅仅是借用这个先验分布以得到minimax决策函数而已。
- 定理3.9 给定一个统计决策问题，如果存在某个先验分布，使得在这个先验分布下的贝叶斯决策函数$d_B(x_1,\cdots,x_n)$的风险函数是一个常数，那么$d_B(x_1,\cdots,x_n)$必定是这个统计决策问题的一个minamax决策函数。

- 对于上述定理，若给定的统计决策问题为参数的点估计，且定理的条件满足，则相应的决策函数$d_B(x_1,\cdots,x_n)$必为参数的minimax估计量。

- 例3.18 在例3.11中若取参数p的先验分布为贝塔分布$Be(\frac{\sqrt{n}}{2},\frac{\sqrt{n}}{2})$，则在平方损失函数下,p的贝叶斯估计

$$\hat{p}=\frac{2\sqrt{n}(\overline{X})+1}{2(\sqrt{n}+1)}$$ 为p的minimax估计。

- 解 因p的先验分布为贝塔分布$Be(\frac{\sqrt{n}}{2},\frac{\sqrt{n}}{2})$，其密度函数

$$\pi(p)=\begin{cases}\dfrac{\Gamma(\sqrt{n})}{\Gamma(\frac{\sqrt{n}}{2})\Gamma(\frac{\sqrt{n}}{2})}p^{\frac{\sqrt{n}}{2}-1}(1-p)^{\frac{\sqrt{n}}{2}-1}, 0<p<1, \\ 0, \quad \text{其余}\end{cases}$$

所以p的后验分布密度为

$$h(p|x)=h(p|x_1,\cdots,x_n)=\frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}A(p)}{\int_0^1 p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}A(p)\,dp}$$

$$=\frac{\Gamma(n+\sqrt{n})}{\Gamma(\sum_{i=1}^n x_i+\frac{\sqrt{n}}{2})\Gamma(n-\sum_{i=1}^n x_i+\frac{\sqrt{n}}{2})}p^{\sum_{i=1}^n x_i+\frac{\sqrt{n}}{2}-1}(1-p)^{n-\sum_{i=1}^n x_i+\frac{\sqrt{n}}{2}-1}$$

其中$0 < p < 1$, $A(p) = \frac{\Gamma(\sqrt{n})p^{\frac{\sqrt{n}}{2}-1}}{\Gamma(\frac{\sqrt{n}}{2})\Gamma(\frac{\sqrt{n}}{2})}(1-p)^{\frac{\sqrt{n}}{2}-1}$ 由p的后验密度

知，p的后验分布为$Be(\sum_{i=1}^{n} x_i + \frac{\sqrt{n}}{2}, n - \sum_{i=1}^{n} x_i + \frac{\sqrt{n}}{2})$，通过计算

可求得p的贝叶斯估计为

$$\hat{p} = E(p|x) = \int_0^1 ph(p|x)dp = \frac{2\sqrt{n}(\bar{X}) + 1}{2(\sqrt{n}+1)}$$

$\hat{p}$的风险函数为

$$R(p,\hat{p}) = E\left[\frac{2\sqrt{nX}+1}{2(\sqrt{n}+1)} - p\right]^2$$

$$= D[\frac{2\sqrt{nX}+1}{2(\sqrt{n}+1)}] + \left\{E[\frac{2\sqrt{nX}+1}{2(\sqrt{n}+1)}] - p\right\}^2$$

$$= \frac{4n}{4(\sqrt{n}+1)^2}\frac{p(1-p)}{n} + \left[\frac{2\sqrt{n}p+1}{2(\sqrt{n}+1)} - p\right]^2 = \frac{1}{4(\sqrt{n}+1)^2}$$

由上式知$\hat{p}$的风险函数是与p无关的常数$\frac{1}{4(\sqrt{n}+1)^2}$，从而由定理3.8

知$\hat{p} = \frac{2\sqrt{n}(\overline{X})+1}{2(\sqrt{n}+1)}$为p的minimax估计。

- 定理3. 10 给定一个贝叶斯决策问题，设$\{\pi_k(\theta): k \geq 1\}$为参数空间$\Theta$上的一个先验分布列，$\{d_k: k \geq 1\}$和$\{R_B(d_k): k \geq 1\}$分别为相应的贝叶斯估计列和贝叶斯风险列。若$d_0$是$\theta$的一个估计，且它的风险函数$R(\theta, d_0)$满足

$$\max_{\theta \in \Theta} R(\theta, d_0) \leq \lim_{k \to \infty} R_B(d_k)$$

则d_0为θ的minimax估计。

定理3.11 给定一个贝叶斯决策问题，若θ的一个估计d_0的风险函数$R(\theta, d_0)$在Θ上为常数ρ，且存在一个先验分布列$\{\pi_k(\theta): k \geq 1\}$，使得相应的贝叶斯估计$d_k$的贝叶斯风险满足

$$\lim_{k\to\infty} R_B(d_k) = \rho$$

则d_0为θ的minimax估计.

例3.19 设$X=(X_1,X_2,\cdots,X_n)^T$是来自正态总体$N(\theta,1)$的一个样本，取参数θ的先验分布为正态分布$N(0,\tau^2)$，其中τ已知，损失函数取为如下的0-1损失

$$L(\theta,d)=\begin{cases}1,|d-\theta|>\varepsilon,\varepsilon>0\\0,|d-\theta|\leq\varepsilon,\varepsilon>0\end{cases}$$

在上述损失下，可求得θ的贝叶斯估计为

$$d_\tau(X)=\overline{X}_n\left(1+\frac{1}{n\tau^2}\right)^{-1},\text{其中}\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$$

现利用定理3.11证明样本均值$\bar{X}_n$是θ的最小最大估计。

□证明　首先求出θ的贝叶斯估计。由例3. 15知θ的后验分布

$h(\theta|x)$为正态分布 $N(\sum_{i=1}^{n} x_i\,(n+\tau^{-2})^{-1}, (n+\tau^{-2})^{-1})$

对任一个决策函数$d = d(X) \in \Re$，其后验风险为

$$R(d|x) = \int_{-\infty}^{+\infty} L(\theta, d)h(\theta|x)d\theta = P_\theta\{|d-\theta| > \varepsilon\}$$

$$= 1 - P_\theta\{|d-\theta| \le \varepsilon\}$$

- 要使上述后验风险最小，就要使上式中概率$P_\theta\{|d-\theta| \le \varepsilon\}$最大，由于后验分布$h(\theta|x)$为正态分布，所以，要在定长区间（长度为$2\varepsilon$）上的概率为最大，d(X)只能取后验分布的均值，即在此场合下，θ的贝叶斯估计为

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

- 应用定理3.11的关键在于选取先验分布列，现选用正态分布列$\{N(0,\tau^2): \tau_1 < \tau_2 < \cdots < \tau_i < \cdots \to \infty\}$作为先验分布列，相应的贝叶斯估计列为$\{d_{r_i}(X), i = 1,2,\cdots\}$。为了计算$d_\tau(X)$的贝叶斯风险，需要先作一些计算。由于$d_\tau(X)$仍服从正态分布，其数学

和方差分别为

$$E(d_\tau)=\theta\left(1+\frac{1}{n\tau^2}\right)^{-1},Var(d_\tau)=\frac{1}{n}\left(1+\frac{1}{n\tau^2}\right)^{-2}$$

而d_τ的风险函数为

$$R(\theta,d_\tau)=P_\theta\{|d_\tau-\theta|\geq\varepsilon\}=1-P_\theta\{\theta-\varepsilon<d_\tau<\theta+\varepsilon\}$$

$$=2-\Phi(\sqrt{n}[\varepsilon(1+\frac{1}{n\tau^2})+\frac{\theta}{n\tau^2}])-\Phi(\sqrt{n}\left[\varepsilon\left(1+\frac{1}{n\tau^2}\right)-\frac{\theta}{n\tau^2}\right])$$

并且有

$$\lim_{k\to\infty}R(\theta,d_\tau)=2-2\Phi(\sqrt{n}\varepsilon)$$

对序列$\tau_1<\tau_2<\cdots<\tau_i<\cdots\to\infty$，有$R(\theta,d_\tau)<2$，于是利用勒贝格收敛定理知

$$\lim_{i\to\infty} R_B(d_{\tau_i}) = \lim_{i\to\infty} E_\theta[R(\theta, d_{\tau_i})] = E_\theta \lim_{i\to\infty} R(\theta, d_{\tau_i}) = E_\theta[2 - 2\Phi(\sqrt{n}\varepsilon)]$$
$$= 2[1 - \Phi(\sqrt{n}\varepsilon)] = \rho$$

其中ρ是不依赖于θ的常数，从而知定理3. 11的条件全部满足，故

知在0−1损失函数下，样本均值 $\overline{X}_n$ 是θ的最小最大估计。

统计决策与贝叶斯估计

第五节：经验贝叶斯估计

- 胡政发，肖海霞，应用数理统计与随机过程，电子工业出版社，2021年第一版
- 师义民，徐伟，秦超英，许勇，数理统计，科学出版社，2015年第四版

- 贝叶斯方法的最大困难是要求未知参数具有确定的先验分布，在某些场合下，即使这个要求是合理的，未知参数的先验分布也无法确定，于是，人们往往对它做一种人为的规定。当先验分布的指定与实际情况不符合时，所得的贝叶斯解会受到较大的影响。因而在对先验分布无法基本确定时，贝叶斯方法的适用性和优越性就值得怀疑了。
- 经验贝叶斯方法的提出可以很好地解决这个问题。

- 经验贝叶斯方法是通过历史资料以决定先验分布的方法，这个方法使用的前提是未知参数的确是通常意义下的随机变量，其分布在历史资料的积累过程中一直保持不变。
- 经验贝叶斯方法是H. Robbins在1955年提出的，他提出这种方法的思想受到国际统计学界的高度重视，统计学元老J.Neyman甚至称它为统计判决的两大突破之一。

- 经验贝叶斯方法分为两类：第一类是非参数EB方法，第二类是参数EB方法。两类方法的不同之处在于，非参数EB方法不指明具体的先验分布，利用数据来估计有关分布。而参数EB方法则指明先验分布族，并且每一级先验分布中都含有未知参数需要确定，参数EB方法利用观测数据来估计在某一级中的参数。

- 非参数经验贝叶斯估计

- 设参数 $\theta \in \Theta$，Θ 为参数空间，$G(\theta)$ 为先验分布， A为决策空间，X为样本空间， $L(\theta, d)$为损失函数。对于随机变量 $X \in \mathcal{X}$，当给定 θ时，$\mathbf{X}$ 的概率密度为 $p(x|\theta)$。

- 设 $d(x)$ 为决策函数，则风险函数为

$$R(\theta, d) = E_\theta[L(\theta, d)] = E_\theta[L(\theta, d(X_1, X_2, \cdots, X_n))]$$

$$R_G(d) = E[L(\theta, d)] = \int_\Theta R(\theta, d)\, dG(\theta)$$

称为 $d(x)$ 的相对先验分布 $G(\theta)$ 的贝叶斯风险。记使贝叶斯风险达到最小的贝叶斯决策为 $d_G(x)$

- 在实际问题中，先验分布 $G(\theta)$ 往往是未知的，因而无法得到 $d_G(x)$。假定我们关心的决策问题独立地出现，在第 i 次碰到这个问题时，样本为 X_i，参数真值为 θ_i。我们假定 θ 有一定

的先验分布 $G(\theta)$，但是只知道 $G(\theta)$ 属于某个分布族 F^*，而 $\theta_1,\theta_2,\cdots,\theta_n$ 可以看成从分布 $G(\theta)$ 中抽出的具有独立同分布的样本，但是并不知道 $\theta_1,\theta_2,\cdots,\theta_n$ 的确定值，只知道 $X_1,X_2,\cdots,X_n$ 独立同分布，分布密度为

$$m_G(x)=\int_{\Theta} p(x\mid\theta)\mathbf{d}G(\theta)$$

由于 $m_G(x)$ 与先验分布 $G(\theta)$ 有关，从而样本 $X_1,X_2,\cdots,X_n$ 中就包含了 $G(\theta)$ 的信息，n 越大，所包含的信息应当越多。

- 若再一次面对上述统计决策问题时，得到样本X，真参数为θ，在求贝叶斯解时，可以参考历史资料$X_1, X_2, \cdots, X_n$获得的关于$G(\theta)$的信息，以选定一个决策函数d，这个d将与$X_1, X_2, \cdots, X_n$有关

- 因而可以记为$d_n(X|X_1, X_2, \cdots, X_n)$，我们希望它的贝叶斯风险接近真正的贝叶斯解$d_G(X)$的贝叶斯风险$R_G(d_G)$，并且当$n \to \infty$时，趋向于$R_G(d_G)$。设$x_1, x_2, \cdots, x_n, x$分别表示$X_1, X_2, \cdots, X_n$的观测值，下面给出$d_n(X|X_1, X_2, \cdots, X_n)$的计算方法。

- 首先固定 $x_1, x_2, \cdots, x_n$ ，这时 $d_n(X|X_1, X_2, \cdots, X_n)$ 只与 x 有关，若以 E^* 表示关于 (X, θ) 的联合分布求期望，则决策函数 d_n 的贝叶斯风险为

$$R_G(d_n \mid x_1, \cdots, x_n) = E^*[L(\theta, d_n(X \mid x_1, \cdots, x_n))]$$
$$= \int_{\Theta}\int_{\mathbf{X}} L(\theta, d_n(X \mid x_1, \cdots, x_n)) p(x \mid \theta) dx dG(\theta)$$

- 上式中涉及 $n+1$ 个变量 $X_1, X_2, \cdots, X_n$ 与 X，而 $X_1, X_2, \cdots, X_n$ 也是随机变量，故还要对它们求一次期望，这个期望称为 d_n 的全面贝叶斯风险，记为 $R_G^*(d_n)$

$$R_G^*(d_n) = \int_X R_G(d_n \mid x_1, \cdots, x_n) m_G(x_1, \cdots, x_n) \mathbf{d}x_1 \cdots \mathbf{d}x_n$$

□定义3.13. 任何同时依赖于历史样本$X_1, X_2, \cdots, X_n$和当前样本 X 的决策函数$d_n(X|X_1, X_2, \cdots, X_n)$ 称为经验贝叶斯决策函数。

□设F^* 为先验分布族，参数 θ 的先验分布为 $G(\theta)$，若对任何$G(\theta) \in F^*$，有

$$\lim_{n\to\infty} R_G^*(d_n) = R_G(d_G)$$

则称 d_n 为渐近最优的经验贝叶斯决策函数。当 d_n 为 θ 的估计时，称 d_n 为 θ 的渐近最优经验贝叶斯估计。

- 应当注意：在经验贝叶斯决策函数中，历史样本 $X_1,\cdots,X_n$ 与当前样本 X 的作用是不同的，$X_1,\cdots,X_n$ 的作用在于由之获得关于先验分布 $G(\theta)$ 的信息，以帮助选定一个尽可能接近真正的贝叶斯 $d_G(X)$ 的决策函数 $d(X|X_1,X_2,\cdots,X_n)$，而推断当前参数真值的任务落在当前样本 X 的头上。
- 关于经验贝叶斯估计渐近最优性的讨论是比较复杂的工作，可以证明，某些参数的渐近最优经验贝叶斯估计是不存在的。
- 例3.20. 设随机变量服从泊松分布，分布律为

$$p(x|\theta)=\theta^x e^{-\theta}/x!\quad (x=0,1,2,\cdots,\theta>0)$$

设参数θ 的先验分布为 $G(\theta)$ ，则 X 的边缘分布为

$$m_G(x)=\int_0^{\infty}\frac{e^{-\theta}\theta^x}{x!}dG(\theta)$$

- 对于先验分布$G(\theta)$，在平方损失下，可求得θ的贝叶斯估计为

$$d_G(x)=E(\theta|x)=\frac{\int_0^{\infty}\theta p(x|\theta)dG(\theta)}{\int_0^{\infty}p(x|\theta)dG(\theta)}=(x+1)\frac{m_G(x+1)}{m_G(x)}$$

- 若先验分布未知，但有了历史样本$X_1,X_2,\cdots,X_n$，它们是从分布$m_G(x)$中抽出的相互独立的同分布样本，故由它们可对$m_G(x)$进行估计，为了避免上式分母为0，取这个估计为

$$\widehat{m_G}(x_1,\cdots,x_n,x)=\frac{1}{n+1}\{(x_1\cdots x_n\text{中等于}x\text{的个数})+1\}$$

以 $\hat{m}_G(x)$ 代替上式中的 $m_G(x)$ ，以 $\hat{m}_G(x_1,\cdots,x_n,x+1)$ 代替$m_G(x+1)$

可得 θ 的经验贝叶斯估计量为

$$d_G(X \mid X_1,\cdots,X_n)=(\mathbf{X}+1)\frac{\hat{m}_G(X+1)}{\hat{m}_G(X)}$$

- 例3.21 设随机变量的分布密度为

$$P(x|\theta)=(2\pi)^{-1}e^{-(x-\theta)^2/2}$$

- θ 的先验分布为$G(\theta)$，$\Theta=(a,b)$。在平方损失下θ的贝叶斯估计为

$$d_G(x)=x-\frac{m'_G(x)}{m_G(x)}$$

- 其中 $m'_G(x)$ 为 $m_G(x)$ 的导数，$m_G(x)$ 为 X 的边缘密度。可以采用非参数密度估计方法来获得 $m_G(x)$ 和 $m'_G(x)$ 的估计 $\hat{m}_G(x)$ 和 $\hat{m}'_G(x)$。于是得到 θ 的经验贝叶斯估计

$$d_G(X|X_1,\cdots,X_n) = X - \frac{\hat{m}'_G(X)}{\hat{m}_G(X)}$$

参数经验贝叶斯估计

- 设随机变量 X 的分布密度为 $p(x|\theta)$，参数 θ 的先验分为 $G(\theta|\beta)$，β 是未知参数。若 $X_1 \cdots X_n$ 为历史样本，X 为当前样本，则 X 的边缘分布为

$$m_G(x|\beta) = \int_{\Theta} p(x|\theta) dG(\theta|\beta)$$

当 β 的估计存在时，θ 的先验分布可取为 $G(\theta|\widehat{\beta})$。当损失函数给定时，可以求出 θ 的经验贝叶斯估计

- 例3.22（续例3.21）在例3.21中，若设 θ 的先验分布族为 $F^* = \{N(0,\sigma^2), \sigma > 0\}$，则 X 在 θ 的先验分布之下的边缘分布为 $N(0,\sigma^2+1)$。由于 σ^2 未知，可求得 σ^2 的最大似然估计为

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 1$$

- 设当前样本为 X，取 θ 的先验分布为 $N(0,\hat{\sigma}_n^2)$，则在平方损失下的经验贝叶斯估计为

$$d_n(X|X_1\cdots X_n) = \frac{\widehat{\sigma}_n^2}{1+\widehat{\sigma}_n^2}X = X\left[\sum_{i=1}^{n} X_i^2 - n\right] \Big/ \sum_{i=1}^{n} X_i^2$$

其贝叶斯风险为

$$R_G(d_n|X_1\cdots X_n)=\frac{\hat{\sigma}_n^2}{1+\hat{\sigma}_n^2}$$

因而得到 d_n 的全面贝叶斯风险为

$$R_G^*(d_n)=E\left[\frac{\hat{\sigma}_n^2}{1+\hat{\sigma}_n^2}\right]$$

由大数定律可知，当 $n\to\infty$时， $\hat{\sigma}_n^2\to(1+\sigma^2)-1=\sigma^2$

于是，由控制收敛定理得

$$\lim_{n\to\infty}R_G^*(d_n)=\frac{\sigma^2}{1+\sigma^2}$$

当 σ^2 已知时，θ 的先验分布为 $N(0,\sigma^2)$，上式右端为 θ 的贝叶斯估计的贝叶斯风险， $d_n(X|X_1\cdots X_n)$也是相对于先验分布族 $F^*=\{N(0,\sigma^2),\sigma>0\}$ 的渐近最优经验贝叶斯估计。