§Bayes 点估计法

一、统计推断中可用的三种信息

E. L. Lehmann:

- 总体信息(population information): 总体分布或总体所属分布族所提供的信息. 如: 总体是正态分布, 总体是指数分布等.
- 样本信息(sample information): 样本提供给我们的信息.
- 先验信息(prior information): 先验信息来源于所考察的统计推断问题之前, 即, 在抽样之前就有的有关统计推断问题的信息. 常常是过去同类统计推断问题提供的信息.

例如: 某工厂考察某天所生产的产品的不合格率时, 过去抽检这种产品质量的资料(历史数据) 对我们估计这一天的不合格率是有好处的, 这些资料提供的信息就是先验信息.

经典统计学派: 在统计推断中只利用总体信息和样本信息

Bayes统计学派: 建议在统计推断中在利用总体信息和样本信息的同时,还利用先验信息.

Example

英国统计学家Savage, L. J. 曾考察了如下两个统计试验:

(1) 一位常饮牛奶加茶的妇女声称, 她能分辨出先倒进杯子里的是茶还是牛奶.对此做了十次试验, 她都正确地说出来.

(2) 一位音乐家声称, 他能从一页乐谱辨别出是海顿(Haydn) 还是莫扎特(Mozart) 的作品. 在十次这样的试验中, 他都分辨正确.

在这两个统计试验中, 假如认为被试验者是在猜测, 每次成功概率为0.5, 那么十次都猜中的概率为$2^{-10} = 0.0009766$. 这是一个小概率事件,是几乎不可能发生的. 因此不能认为是猜测,而是他们的经验帮了他们的忙.

$$Bayes \quad T.R.(1702? - 1761)$$

♣ 《An Essay toward Solving a Problem in the Doctrine of Chances》(论有关机遇问题的求解), 1763.

♣ 《女士品茶》全名《The Lady Tasting Tea——How Statistics Revolutionized Science in the Twentieth Century 》(女士品茶——20世纪统计学怎样变革了科学)

二、先验分布(prior distribution)和后验分布(posterior distribution)

如何将先验信息归纳到统计模型中?

设总体的密度函数(pdf)或分布列(pmf)为$p(x;\theta)$, $\theta \in \Theta$是未知参数.

经典统计学派认为: 参数θ虽然是未知的,但是固定的常数.

Bayes统计学派认为: 参数θ不是常数,而是变化的量. 例如: 某工厂每日生产的产品的次品率并不是固定的,而是逐日不同的(可能围绕某个平均值上下浮动). Bayes统计学派认为: 未知参数是变化的量, 它的波动情况可用一个概率分布来描述.

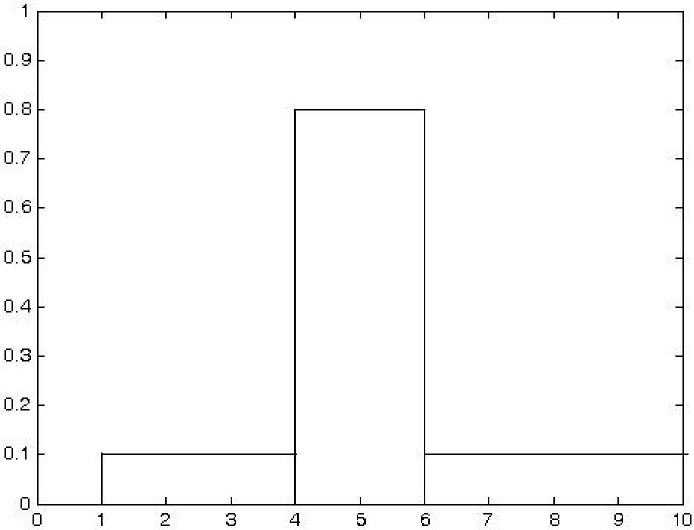
因此Bayes统计学派认为: 未知参数θ是个随机变量, 它有一个分布(pdf或pmf)$\pi(\theta)$, 称为先验分布.

先验分布$\pi(\theta)$可从先验信息中归纳出来. 常常是主观概率, 取决于试验者在试验前对θ的先验信息了解的程度和他对这些信息的信任程度.

Example

某地区煤的平均储存量θ在几百年内不会有多大变化，可以看作是一个常量，但对人们来说，它是未知的、不确定的量. 有位专家研究了有关资料，结合他的经验认为：该地区的平均储存量θ"大概有5亿吨左右".

如果把"左右"理解为4亿吨到6亿吨之内，把"大概"理解为80%的把握，还有20%的可能性在此区间之外. 这无形中用一个概率分布去描述未知量θ，而具有概率分布的量当然是随机变量.


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
9
10

而对总体的概率分布$p(x;\theta)$, Bayes统计学派认为: $p(x;\theta)$是作为随机变量的θ在取值给定后的一个条件概率分布, 因此记为$p(x|\theta)$.

Bayes统计的三个基本假设:

假设I: 总体X有一个概率分布(pdf或pmf)$p(x;\theta)$, 其中θ是参数, 不同的θ对应着不同的分布. 在Bayes统计中, $p(x;\theta)$是给定θ后的一个条件概率分布, 记为$p(x|\theta)$. $p(x|\theta)$提供的关于θ的信息就是总体信息.

假设II: 当给定θ后,从总体$p(x|\theta)$中随机抽取一个样本$\widetilde{X}=(X_1,X_2,\cdots,X_n)$, 该样本中含有的有关$\theta$的信息,就是样本信息.

当给定θ后, 样本的联合条件概率分布(pdf或pmf)为

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

这个条件概率分布综合了总体信息和样本信息.

假设III: 未知参数θ是一个随机变量, 它有一个概率分布(pdf或pmf)$\pi(\theta)$(称为先验分布). 先验分布是已知的. 先验分布提供的信息就是先验信息.

这样, 样本$\widetilde{X}=(X_1,X_2,\cdots,X_n)$和参数的$\theta$的联合概率分布是

$$p(\widetilde{x},\theta)=p(\widetilde{x}|\theta)\pi(\theta).$$

这个联合概率分布综合了总体信息、样本信息和先验信息.

给定样本$\widetilde{X}=\widetilde{x}$，由样本$\widetilde{X}$和参数的$\theta$的联合概率分布求得参数$\theta$的条件概率分布

$$\pi(\theta|\widetilde{x}) = \frac{p(\widetilde{x},\theta)}{p(\widetilde{x})} = \frac{p(\widetilde{x}|\theta)\pi(\theta)}{\int p(\widetilde{x}|\theta)\pi(\theta)d\theta} \text{——Bayes公式.}$$

$\pi(\theta|\widetilde{x})$是有了试验结果后,得到的参数$\theta$的条件分布，称为后验分布(posterior distribution). 而$p(\widetilde{x})$是样本的边际分布.

$$\widetilde{X}=\widetilde{x}$$
$$\vdots$$
$$\pi(\theta) \quad \longrightarrow \quad \pi(\theta|\widetilde{x})$$

Bayes方法对参数θ的统计推断是建立在后验分布$\pi(\theta|\tilde{x})$的基础上的.

在Bayes公式中，分子和分母上同乘上一个与θ无关的量$h(\widetilde{x})$不改变后验分布.
特别地如果$T=T(\widetilde{X})$是充分统计量，

$$p(\widetilde{x}|\theta)=p(\widetilde{x}|T=t)p_T(t|\theta)\propto p_T(t|\theta).$$

其中$t=T(\widetilde{x})$. 从而

$$\pi(\theta|\widetilde{x})=\frac{p_T(t|\theta)\pi(\theta)}{\int p_T(t|\theta)\pi(\theta)d\theta}=\pi(\theta|t).$$

用充分统计量代替样本所得的后验分布是一样的.

三、Bayes点估计

1. 后验分布$\pi(\theta|\tilde{x})$的众数作为θ的估计—-众数型Bayes估计
2. 后验分布$\pi(\theta|\tilde{x})$的中位数作为θ的估计—-中位数型Bayes估计
3. 后验分布$\pi(\theta|\tilde{x})$的期望作为θ的估计—期望型Bayes估计

$$\widehat{\theta}_B(\tilde{x}) = \mathsf{E}(\theta|\tilde{x}) = \int \theta \cdot \pi(\theta|\tilde{x}) d\theta.$$

$\widehat{\theta}_B(\widetilde{X})$就是期望型Bayes点估计量.

Example

(Binomial Bayes Estimation) 设某产品的废品率为θ，为估计θ，随机地抽取n件产品进行检查，发现$T=t$件废品$(0\le t\le n)$. 求θ的(期望型)Bayes点估计值，并与MLE相比较.

解: 从产品中随机取一件，检查其质量. 设总体为

$$X=\begin{cases}1, & \text{取到的产品为废品;}\\ 0, & \text{取到的产品不为废品.}\end{cases}$$

则总体X来自分布族$\{B(1,\theta),\theta\in[0,1]\}$. 抽取了容量为$n$的一组样本$\tilde{X}=\{X_1,\cdots,X_n\}$. 记样本中的废品数$T=T(\tilde{X})=\sum_{i=1}^n X_i$，由题意知，观测到$T=t$，由之前的知识得$\theta$的极大似然估计值为

$$\hat{\theta}_{MLE}=\frac{t}{n} \text{ ——经典统计学派的估计.}$$

且已知废品数T是θ的充分统计量，它的分布列为

$$p(t|\theta)=\mathsf{P}(T=t|\theta)=\binom{n}{t}\theta^t(1-\theta)^{n-t},\quad t=0,1,\cdots,n.$$

为求Bayes估计，先给定一个先验分布$\pi(\theta)$，然后求后验分布

$$\pi(\theta|t)=\frac{\binom{n}{t}\theta^t(1-\theta)^{n-t}\pi(\theta)}{\int_\Theta\binom{n}{t}\theta^t(1-\theta)^{n-t}\pi(\theta)d\theta}\propto_\theta \theta^t(1-\theta)^{n-t}\pi(\theta).$$

再求后验期望就得到Bayes估计

$$\widehat{\theta}_B=\int_\Theta\theta\pi(\theta|t)d\theta=\frac{\int_\Theta\theta\theta^t(1-\theta)^{n-t}\pi(\theta)d\theta}{\int_\Theta\theta^t(1-\theta)^{n-t}\pi(\theta)d\theta}.$$

先验分布的选取:

1. ”同等无知”

在没有先验信息的情况下，对未知参数θ的所有可能取值同等对待. 现θ取值范围为$[0,1]$，故在没有先验信息的情况下，取均匀分布$U[0,1]$作为先验分布:

$$\pi(\theta)=1; \quad 0\le\theta\le 1.$$

这时Bayes点估计值为

$$\widehat{\theta}_B=\frac{\int_0^1 \theta\theta^t(1-\theta)^{n-t}d\theta}{\int_0^1 \theta^t(1-\theta)^{n-t}d\theta}=\frac{t+1}{n+2}.$$

No.	n	t	$\widehat{\theta}_{MLE}=\frac{t}{n}$	$\widehat{\theta}_{B}=\frac{t+1}{n+2}$
1	5	5	1	0.867
2	20	20	1	0.955
3	5	0	0	0.143
4	20	0	0	0.045

事实上，此时θ的后验概率密度函数为

$$\begin{aligned}
\pi(\theta|t) =& \frac{\binom{n}{t}\theta^t(1-\theta)^{n-t}\cdot 1}{\int_0^1 \binom{n}{t}\theta^t(1-\theta)^{n-t}\cdot 1 d\theta} = \frac{\theta^t(1-\theta)^{n-t}}{\int_0^1 \theta^t(1-\theta)^{n-t} d\theta} \\
=& \frac{1}{B(t+1, n-t+1)}\theta^t(1-\theta)^{n-t} \\
=& \frac{\Gamma(n+2)}{\Gamma(t+1)\Gamma(n-t+1)}\theta^{(t+1)-1}(1-\theta)^{(n-t+1)-1}, \quad 0 \le \theta \le 1.
\end{aligned}$$

即θ的后验分布为$beta(t+1, n-t+1)$.

类似的例子: Laplace 在1786年研究了巴黎男婴诞生的比例θ是否大于0.5. 为此他收集了1745年到1770年在巴黎诞生的婴儿数据，其中男婴251527个,女婴241945 个. 记$T=\sum_{i=1}^n X_i$为样本中男婴的个数. 他选用$U(0,1)$作为θ的先验分布，计算了在观测到男婴数为251527的条件下，事件"$\theta\le 0.5$"的条件概率

$$\mathsf{P}(\theta\le 0.5|T=t)=\int_{-\infty}^{0.5}\pi(\theta|t)d\theta=\frac{\Gamma(n+2)}{\Gamma(t+1)\Gamma(n-t+1)}\int_0^{0.5}\theta^t(1-\theta)^{n-t}d\theta,$$

其中$n=251527+241945=493472$, $t=251527$. 当年Laplace 将被积函数$\theta^t(1-\theta)^{n-t}$在最大值$\frac{t}{n}$处展开,计算积分得到

$$\mathsf{P}(\theta\le 0.5|T=t)=1.15\times 10^{-42}.$$

由于这一个概率很小，Laplace以很大的把握断言:男婴的出生概率大于0.5, 这一结果在当时很有影响.

2. 共轭分布法 我们希望先验分布和后验分布是同类型的分布，即它们在同一个分布族中.

现在

$$\pi(\theta|t) \propto_\theta \theta^t(1-\theta)^{n-t}\pi(\theta).$$

为使得先验分布和后验分布有同类型的分布，可取

$$\pi(\theta) \propto_\theta \theta^{a-1}(1-\theta)^{b-1}.$$

因此取beta分布作为先验分布，beta分布族作为先验分布族.

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}; \quad 0 \le \theta \le 1.$$

这时

$$\pi(\theta|t) \propto_\theta \theta^{t+a-1}(1-\theta)^{n-t+b-1}.$$

即$\pi(\theta|t) \sim beta(t+a, n-t+b)$. 则期望型Bayes(点)估计为

$$\widehat{\theta}_B = \mathsf{E}[beta(t+a, n-t+b)] = \frac{t+a}{n+a+b}.$$

$\widehat{\theta}_B$可写成

$$\widehat{\theta}_B = \frac{n}{a+b+n} \cdot \widehat{\theta}_{MLE} + \frac{a+b}{a+b+n} \cdot \frac{a}{a+b}.$$

它是$\widehat{\theta}_{MLE}$和先验期望$a/(a+b)$的加权平均.

由此得到当n 较小时，先验信息在估计中占主要地位；当n较大时，试验信息在估计中占主要地位.

参数a, b不同,对应的先验分布也不同, 得到的Bayes估计也不同. a, b一般会根据试验者所掌握的先验信息来确定.

例如: 1. 如果抽样前知道θ的均值$\overline{\theta}$和方差s_θ^2, 则可由下列方程解出a, b:

$$\begin{cases} \frac{a}{a+b} = \overline{\theta}, \\ \frac{ab}{(a+b)^2(a+b+1)} = s_\theta^2. \end{cases}$$

2. 如果抽样前知道θ的分位数, 则也可解出a, b; 如

$$\begin{cases} \int_0^{\theta_{0.1}} \pi(\theta) d\theta = 0.1, \\ \int_0^{\theta_{0.5}} \pi(\theta) d\theta = 0.5. \end{cases}$$

3. 如果根据先验信息只能获得先验均值$\overline{\theta}$, 可令

$$\frac{a}{a+b}=\overline{\theta}.$$

但从一个方程不能唯一确定两个未知参数.
方差

$$\mathsf{Var}(\theta)=\frac{ab}{(a+b)^2(a+b+1)}=\frac{\overline{\theta}(1-\overline{\theta})}{a+b+1}$$

随$a+b$的增大而减少, 方差减少意味着分布向均值$\mathsf{E}(\theta)=\overline{\theta}$集中, 从而提高了$\mathsf{E}(\theta)=\overline{\theta}$的确信程度. 这样一来, 选择$a,b$的问题转化为决策人对$\mathsf{E}(\theta)=\overline{\theta}$的确信程度大小的问题. 如果对$\mathsf{E}(\theta)=\overline{\theta}$很确信, 那么$a+b$可选得大一些, 否则就选得小一些.

均值为0.4的Beta分布中参数与方差的关系

Beta分布	a	$a+b$	$\mathsf{E}(\theta)$	$\mathsf{Var}(\theta)$
beta(2,3)	2	5	0.4	0.0400
beta(4,6)	4	10	0.4	0.0218
beta(8,12)	8	20	0.4	0.0114
beta(10,15)	10	25	0.4	0.0092
beta(14,21)	14	35	0.4	0.0067

Example

设$\widetilde{X}=(X_1,\ldots,X_n)$是从均匀分布$U(0,\theta)$中抽取的简单随机样本，样本观测值为$\widetilde{x}=(x_1,\ldots,x_n)$, $n\geq 1$. 试求θ的(期望型)Bayes 点估计值.

解: $(X_1,\ldots,X_n)$的联合密度函数为$p(\widetilde{x}|\theta)=\frac{1}{\theta^n}I\{\theta>x_{(n)}\}\cdot I\{0<x_{(1)}\}$, 其中$x_{(n)}=\max\{x_1,\cdots,x_n\}$, $x_{(1)}=\min\{x_1,\cdots,x_n\}$.

设先验分布为$\pi(\theta)$, 则后验分布

$$\pi(\theta|\widetilde{x})=\frac{p(\widetilde{x}|\theta)\pi(\theta)}{\int p(\widetilde{x}|\theta)\pi(\theta)d\theta}\propto_\theta p(\widetilde{x}|\theta)\pi(\theta)\propto_\theta \frac{1}{\theta^n}I\{\theta>x_{(n)}\}\pi(\theta).$$

为使$\pi(\theta)$ 和$\pi(\theta|\widetilde{x})$有相同的形式, 取

$$\pi(\theta)\propto_\theta \frac{1}{\theta^{\alpha+1}}I\{\theta>\theta_0\}\quad(\alpha>0).$$

所以

$$\pi(\theta) = \frac{\alpha\theta_0^\alpha}{\theta^{\alpha+1}} I\{\theta > \theta_0\}.$$

这一分布称为帕累托(Pareto)分布，参数为$\alpha > 0$和$\theta_0 > 0$.

http://en.wikipedia.org/wiki/Pareto_distribution

当$\alpha > 1$时，其数学期望为

$$\int_{\theta_0}^{+\infty} \theta\pi(\theta)d\theta = \frac{\alpha}{\alpha-1}\theta_0.$$

此时后验分布为

$$\pi(\theta|\tilde{x}) \propto_\theta \frac{1}{\theta^{\alpha+n+1}} I\{\theta > \theta_1\}, \quad \theta_1 = \max\{x_{(n)}, \theta_0\}.$$

所以

$$\widehat{\theta}_B = \int \theta\pi(\theta|\tilde{x})d\theta = \frac{\alpha+n}{\alpha+n-1}\theta_1.$$

Example

彩电的寿命服从指数分布，它的密度函数为

$$p(t|\theta) = \frac{1}{\theta}e^{-t/\theta}, \quad t > 0.$$

其中，$\theta > 0$是彩电的平均寿命．现从一批彩电中随机抽取n台进行寿命试验．试验进行到第r台失效时为止$(1 \le r \le n)$．记录失效时间为$t_{(1)} \le t_{(2)} \le \cdots \le t_{(r)}$．另外，$n-r$台直到试验停止时还没失效．求$\theta$的期望型Bayes点估计值．

解: 设$(T_1, T_2, \cdots, T_n)$为抽取的n台彩电的寿命，$t_{(1)}, t_{(2)}, \cdots, t_{(r)}$ 是观察到的前r个次序统计量$T_{(1)}, T_{(2)}, \cdots, T_{(r)}$的观察值. $(T_{(1)}, T_{(2)}, \cdots, T_{(r)})$ 是截断样本，这个截断样本的联合概率密度函数为

$$\begin{aligned} p(t_{(1)}, \cdots, t_{(r)}|\theta) \propto_\theta & \prod_{i=1}^{r} (\theta^{-1} e^{-t_{(i)}/\theta}) \cdot [e^{-t_{(r)}/\theta}]^{n-r} \\ = & \theta^{-r} \exp\{-s/\theta\}, \quad 0 < t_{(1)} \le t_{(2)} \le \cdots \le t_{(r)}, \end{aligned}$$

其中$s = t_{(1)} + \cdots + t_{(r)} + (n-r)t_{(r)}$为总试验时间.

若设$\pi(\theta)$为先验分布，则后验分布为

$$\pi(\theta|t_{(1)},\cdots,t_{(r)}) \propto_\theta \theta^{-r}\exp\{-s/\theta\}\pi(\theta).$$

为使得$\pi(\theta)$和$\pi(\theta|t_{(1)},\cdots,t_{(r)})$有同样的形式，应取

$$\pi(\theta) \propto_\theta \theta^{-(a+1)}\exp\{-b/\theta\}.$$

即

$$\pi(\theta) = \frac{b^a}{\Gamma(a)}\theta^{-(a+1)}\exp\{-b/\theta\}, \quad \theta > 0.$$

这样的密度函数是倒gamma分布的密度函数，记为$\Gamma^{-1}(a,b)$.

易得

$$\mathsf{E}[\Gamma^{-1}(a,b)] = \frac{b}{a-1}.$$

现在后验密度函数为

$$\pi(\theta|t_{(1)},\cdots,t_{(r)}) \propto_\theta \theta^{-(a+r+1)}\exp\{-(b+s)/\theta\}, \quad \theta > 0.$$

即后验分布为$\Gamma^{-1}(a+r,b+s)$. 那么期望型Bayes点估计为

$$\widehat{\theta}_B = \mathsf{E}[\Gamma^{-1}(a+r,b+s)] = \frac{b+s}{a+r-1}.$$

根据已掌握的资料，我国已做了大量的彩电寿命试验，由这些试验数据可估算出彩电的平均寿命不低于30000小时，10%分位数$\theta_{0.1}$大约为11250小时. 这样解方程

$$\begin{cases} \frac{b}{a-1} = 30000 \\ \int_0^{\theta_{0.1}} \pi(\theta) d\theta = 0.1. \end{cases}$$

得

$$a = 1.956, \qquad b = 2868.$$

从而

$$\widehat{\theta}_B = \frac{2868 + s}{0.956 + r}.$$

若随机抽取了100台，进行了400小时试验，发现没有一台失效，则$s = 100 \times 400 = 40000$(小时).从而$\widehat{\theta}_B = 44841$(小时).

共轭先验分布

Definition

设θ是某分布的一个参数，$\pi(\theta)$是其先验分布. 假如由抽样信息算得的后验分布$\pi(\theta|\tilde{x})$与$\pi(\theta)$是属于同一类分布族，则称$\pi(\theta)$ 是θ 的共轭先验分布.

常用的共轭先验分布

总体分布	参数	共轭先验分布
二项分布	成功概率	beta分布
泊松分布	均值	gamma分布
指数分布	均值	倒gamma分布
指数分布	均值倒数	gamma分布
正态分布(方差已知)	均值	正态分布
正态分布(均值已知)	方差	倒gamma分布

Example

设$\widetilde{X}=(X_1,\ldots,X_n)$为从Poisson分布$P(\lambda)$中抽取的简单样本，$\lambda$的先验分布为Gamma分布$\Gamma(\alpha,\beta)$. 证明给定$\widetilde{X}=\widetilde{x}$时，$\lambda$的后验分布仍为Gamma分布.

证： λ的后验密度函数为

$$\pi(\lambda|\widetilde{x}) \propto_\lambda \lambda^{\sum_{i=1}^n x_i}\frac{1}{x_1!\cdots x_n!}e^{-n\lambda}\pi(\lambda) \propto_\lambda \lambda^{n\bar{x}}e^{-n\lambda}\pi(\lambda), \quad \lambda>0,$$

其中$\bar{x}=\frac{1}{n}\sum_{i=1}^n x_i$. λ的先验分布是

$$\pi(\lambda)=\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}, \quad \lambda>0.$$

所以

$$\pi(\lambda|\widetilde{x}) \propto_\lambda \lambda^{n\overline{x}+\alpha-1}e^{-(n+\beta)\lambda}, \quad \lambda > 0.$$

添上正则化因子得

$$\pi(\lambda|\widetilde{x}) = \frac{(n+\beta)^{n\overline{x}+\alpha}}{\Gamma(n\overline{x}+\alpha)}\lambda^{n\overline{x}+\alpha-1}e^{-(n+\beta),\lambda} \quad \lambda > 0.$$

即后验分布为Gamma分布$\Gamma(n\overline{x}+\alpha, n+\beta)$.

Example

设$X_1, X_2, \ldots, X_n$是取自正态分布$N(\theta, \sigma^2)$的样本,其样本观察值为$x_1, x_2, \ldots, x_n$ 其中σ^2已知. 取$N(\mu, \tau^2)$作为θ的先验分布. 求θ的期望型Bayes点估计值.

解: 样本的联合密度函数为

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

θ的先验分布为

$$\pi(\theta) = \frac{1}{\sqrt{2\pi}\tau} \exp\left\{-\frac{1}{2\tau^2}(\theta - \mu)^2\right\}.$$

样本$\widetilde{X}$和参数θ的联合密度函数为

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

若再记

$$\sigma_0^2=\frac{\sigma^2}{n},\quad A=\frac{1}{\sigma_0^2}+\frac{1}{\tau^2},\quad B=\frac{\overline{x}}{\sigma_0^2}+\frac{\mu}{\tau^2}.$$

则

$$p(\widetilde{x},\theta)\propto_\theta \exp\left\{-\frac{1}{2}\left[A\theta^2-2\theta B\right]\right\}\propto_\theta \exp\left\{-\frac{(\theta-B/A)^2}{2/A}\right\}.$$

所以θ的后验分布为

$$\pi(\theta|\widetilde{x}) = \frac{p(\widetilde{x},\theta)}{\int p(\widetilde{x},\theta)d\theta} = \frac{\exp\left\{-\frac{(\theta-B/A)^2}{2/A}\right\}}{\int \exp\left\{-\frac{(\theta-B/A)^2}{2/A}\right\}d\theta}$$
$$= \frac{1}{\sqrt{2\pi/A}}\exp\left\{-\frac{(\theta-B/A)^2}{2/A}\right\}.$$

即$\theta|_{\widetilde{x}} \sim N(\mu_1,\sigma_1^2)$, 其中

$$\mu_1 = \frac{B}{A} = \frac{\overline{x}\sigma_0^{-2}+\mu\tau^{-2}}{\sigma_0^{-2}+\tau^{-2}}, \quad \sigma_1^2 = 1/A = (\sigma_0^{-2}+\tau^{-2})^{-1}.$$

从而θ的期望型Bayes点估计值为

$$\widehat{\theta}_B = \mu_1 = \frac{\overline{x}\sigma_0^{-2} + \mu\tau^{-2}}{\sigma_0^{-2} + \tau^{-2}} = \overline{x}\frac{\sigma_0^{-2}}{\sigma_0^{-2} + \tau^{-2}} + \mu\frac{\tau^{-2}}{\sigma_0^{-2} + \tau^{-2}}.$$

其中$\sigma_0^2 = \sigma^2/n$为样本均值$\overline{X}$的方差, μ为先验均值, τ^2为先验方差.

Example

在儿童智商测验中，儿童的智商$X \sim N(\theta, 100)$，而$\theta \sim N(100, 225)$. 一儿童在一次智商测验中得$x = 115$分，求θ的Bayes估计.

$$\widehat{\theta}_B = \frac{x \times 100^{-1} + 100 \times 225^{-1}}{100^{-1} + 225^{-1}} = \frac{400 + 9x}{13} = 110.38.$$

Example

设有一枚面值为1元的人民币硬币，重复抛掷100次,出现了100次国徽. 请预测在第101次抛掷时再出现国徽的概率是多少?

此题的总体可取为

$$X = \begin{cases} 1, & \text{抛一次硬币出现国徽;} \\ 0, & \text{否则.} \end{cases}$$

则总体X来自分布族$\{B(1,\theta), \theta \in [0,1]\}$. 抽取了容量$n$为100的一组样本，记出现国徽的次数为$T = T(\widetilde{X}) = \sum_{i=1}^n X_i$，由题意知，观测到$T = 100$. 问:

$\mathsf{P}(X_{101} = 1 | T = 100) = ?$

概率为$\frac{1}{2}$？　概率为1?　概率为0?　概率为其它值?

解: 这个问题的一般提法是:在n次独立的Bernoulli试验中成功了$T=t$次,现要对未来的k次独立的Bernoulli试验中有可能发生的成功次数Z取各个可能取值的概率作出预测. 现设成功概率为θ, 则n次独立的Bernoulli 试验中成功次数T的pmf为

$$p(t|\theta)=\binom{n}{t}\theta^t(1-\theta)^{n-t},\quad t=0,1,\ldots,n.$$

取θ的先验分布为beta分布$beta(a,b)$, 当$t=0,1,\cdots,n$时, 则后验分布为beta分布$beta(t+a,n-t+b)$, 即密度函数为

$$\pi(\theta|t)=\frac{p(t|\theta)\pi(\theta)}{\int_0^1 p(t|\theta)\pi(\theta)d\theta}=\frac{1}{\beta(t+a,n-t+b)}\theta^{t+a-1}(1-\theta)^{n-t+b-1},\quad 0<\theta<1.$$

设Z表示"未来的k次独立Bernoulli试验中成功的次数",其pmf为

$$g(z|\theta) = \binom{k}{z}\theta^z(1-\theta)^{k-z}, \quad z = 0, 1, \ldots, k.$$

在给定$T = t$时, (Z, θ)的联合概率函数为

$$g(z, \theta|t) = g(z|t, \theta)\pi(\theta|t) = g(z|\theta)\pi(\theta|t).$$

所以在给定$T = t$时, Z的边际概率函数为

$$\begin{aligned}
p_{Z|T}(z|t) &= \int_0^1 g(z, \theta|t)d\theta = \int_0^1 g(z|\theta)\pi(\theta|t)d\theta \\
&= \binom{k}{z}\frac{1}{\beta(t+a, n-t+b)}\int_0^1 \theta^{z+t+a-1}(1-\theta)^{k-z+n-t+b-1}d\theta \\
&= \binom{k}{z}\frac{\beta(z+t+a, k-z+n-t+b)}{\beta(t+a, n-t+b)}.
\end{aligned}$$

对本问题$k=1, z=1$, 所得概率为

$$\widehat{p_{Z|T}(1|t)} = \frac{\beta(1+t+a, n-t+b)}{\beta(t+a, n-t+b)} = \frac{t+a}{n+a+b}.$$

此即为要求的Bayes点预测值. 若取先验分布为$U(0,1)$ (即$a=b=1$), 现在$n=100$, $t=100$, 则

$$\widehat{p_{Z|T}(1|100)} = \frac{101}{102}.$$

Bayes 预测

一般地，预测的典型问题是：设$X \sim p(x|\theta)$，在获取数据$T = t$后(其中T为θ的充分统计量，否则就用$\widetilde{X} = \tilde{x}$来做)，对具有概率密度函数(也可以为分布列)的$g(z|\theta)$ 的随机变量Z的未来观察值作出预测. 通常假设在θ给定的条件下，T和Z 独立(或$\widetilde{X}$和Z独立).

在Bayes框架下，设θ的先验密度为$\pi(\theta)$，那么在给定$T = t$时，θ的后验密度为$\pi(\theta|t)$. 从而在给定$T = t$时，(Z, θ)的联合密度为$g(z|t, \theta)\pi(\theta|t) = g(z|\theta)\pi(\theta|t)$. 因此，给定$T = t$后，$Z$ 的后验预测密度函数(posterior predictive density function)(也可为后验预测分布列) 为

$$\widehat{p_{Z|T}(z|t)} = \int_{\Theta} g(z|\theta)\pi(\theta|t)d\theta.$$

在经典统计框架下，由于T和Z独立，在给定$T=t$时，Z的密度函数(也可为分布列)仍然是

$$g(z;\theta).$$

由于θ是未知参数，需要根据$p(t;\theta)$和T的观察值t得到的估计$\widehat{\theta}$作为θ的替代，这样$g(z;\widehat{\theta})$可以作为$g(z;\theta)$的替代，把它当作Z的预测密度，但是$g(z;\widehat{\theta})$不是Z的真实密度函数，只是密度函数的替代.

先验分布的确定方法-Bayes推断的关键和难点

主观概率法

主观概率(subjective probability)是人们根据经验对事件发生机会的个人信念.

客观法–利用客观的先验信息确定先验分布.

- 非参数法, 利用先验信息采用非参数方法估计pdf或pmf $\pi(\theta)$, 得到$\hat{\pi}(\theta)$, 使得$\hat{\pi}(\theta)$ 和$\pi(\theta)$ 很接近, 则$\hat{\pi}(\theta)$即为选定的先验密度函数(分布列).
  直方图, 先验信息的经验分布函数, 密度函数估计,...

- 参数法:
  - 先确定θ的先验密度(或分布列)的形式$\pi(\theta|\gamma)$, 其中γ称为超参数(hyperparameter); (常使用共轭先验分布法)
  - 根据先验信息,对超参数γ作出估计,得到$\widehat{\gamma}$, 使得$\pi(\theta|\widehat{\gamma})$和$\pi(\theta|\gamma)$很接近, 则$\pi(\theta|\widehat{\gamma})$即为选定的先验密度函数(分布列).

如: 之前提到的彩电寿命估计问题.

无信息先验

Bayes分析的一个重要特点就是在统计推断时要利用先验信息. 但是实际中常常会出现这样的情况: 没有先验信息或者只有极少的先验信息可利用,但仍然要用Bayes方法.

此时所需要的是一种无信息先验(noninformative prior), 即对参数空间Θ中的任何一点θ没有偏爱的先验信息分布("同等无知").

均匀分布与广义先验分布

- 若$\Theta = \{\theta_1, \ldots, \theta_l\}$为有限集，无信息先验给每个元素以相同的概率，即$\mathsf{P}(\theta = \theta_i) = 1/l$, $i = 1, \ldots, l$.
- 若$\Theta = [a, b]$为有限区间，则取无信息先验为区间$[a, b]$上的均匀分布$U(a, b)$.
- 问题是若参数空间Θ无界，无信息先验如何选取？例如，总体分布为$N(\theta, 1)$，此时$\Theta = (-\infty, \infty)$. 若无信息先验，可取先验分布为$\pi(\theta) \equiv 1$，但$\pi(\theta)$不是通常的密度函数. 这就要引入广义先验分布的概念.

Definition

设样本$\widetilde{X} \sim p(\widetilde{x}|\theta)$, 若$\pi(\theta)$满足下列条件:

1. $\pi(\theta) \geq 0$ 且$\int_\Theta \pi(\theta)d\theta = \infty$,
2. 后验密度函数(或后验概率分布列)

$$\pi(\theta|\widetilde{x}) = \frac{p(\widetilde{x}|\theta)\pi(\theta)}{\int_\Theta p(\widetilde{x}|\theta)\pi(\theta)d\theta}$$

是正常的密度函数(或概率分布列),

则称$\pi(\theta)$为θ的广义先验密度(improper prior density)(或广义先验分布列).

位置参数的无信息先验

设总体$X \sim f(x-\theta)$, $-\infty < \theta < \infty$, θ 为位置参数(location parameter). 对X做平移变换$Y = X + c$, 同时对θ也作同样的平移变换$\eta = \theta + c$, 显然$Y \sim f(y-\eta)$, η为位置参数. 所以(X, θ)与(Y, η)的统计结构相同. 因此主张它们有相同的无信息先验是合理的. 这样θ与$\eta = \theta + c$有相同的分布, 从而θ的先验分布不依赖于原点的选择, 它在等长度区间内的先验概率应当一样. 所以可取

$$\pi(\theta) \equiv 1,$$

它是一个广义先验密度.

Example

设$\widetilde{X} = (X_1, \ldots, X_n)$为从总体$N(\theta, \sigma^2)$中抽取的简单随机样本， 其中$\sigma^2$已知. 样本观测值为$\widetilde{x} = (x_1, \ldots, x_n)$. 若$\theta$无任何先验信息可用, 求$\theta$的后验分布.

解 θ为位置参数, 取无信息先验$\pi(\theta) \equiv 1$. 记$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$, 并注意到$\theta$的充分统计量为$\overline{X}$, 且服从分布$N(\theta, \sigma^2/n)$(在参数取定为$\theta$的条件下), 其密度函数为

$$p(\overline{x}|\theta) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left\{-\frac{1}{2\sigma^2/n}(\overline{x}-\theta)^2\right\}.$$

所以

$$\pi(\theta|\widetilde{x}) = \pi(\theta|\overline{x}) \propto_\theta p(\overline{x}|\theta)\pi(\theta) \propto_\theta \exp\left\{-\frac{1}{2\sigma^2/n}(\theta-\overline{x})^2\right\}.$$

因此在观测到$\{\widetilde{X}=\widetilde{x}\}$的条件下，$\theta$的后验分布为$N(\overline{x},\sigma^2/n)$.
由此还可得θ的期望型Bayes点估计量为

$$\widehat{\theta}_B=E[\theta|\widetilde{X}]=E[\theta|\overline{X}]=\overline{X}.$$

事实上，利用后验分布————正态分布的单峰对称性，知θ的众数型，中位数型Bayes点估计量也是$\overline{X}$.

尺度参数的无信息先验

设总体$X \sim \frac{g(x/\sigma)}{\sigma}$, $\sigma > 0$ 为尺度参数(scale parameter). 对X作尺度变换$Y = cX$ $(c > 0)$, 同时对σ 也作同样的尺度变换$\eta = c\sigma$, 容易知道$Y \sim \frac{g(y/\eta)}{\eta}$, η为尺度参数, 所以(X, σ)与(Y, η)的统计结构相同. 因此主张它们有相同的无信息先验是合理的. 这样σ与$\eta = c\sigma$有相同的分布, 从而

$$\pi(\sigma) = \pi(\frac{\sigma}{c})\frac{1}{c}, \forall \sigma > 0, c > 0.$$

由此得$\pi(c) = \pi(1)/c$, 故可取σ的无信息先验为

$$\pi(\sigma) = \frac{1}{\sigma}, \ \sigma > 0,$$

它是一个广义先验密度.

Example

设总体X为指数分布，其密度函数为

$$p(x|\theta)=\theta^{-1}\exp\{-x/\theta\}, x>0,$$

其中$\theta>0$为尺度参数.令$\widetilde{X}=(X_1,\ldots,X_n)$为从上述分布中抽取的简单随机样本，样本观测值为$\widetilde{x}=(x_1,\ldots,x_n)$. 若$\theta$无任何先验信息可用，求$\theta$的后验分布.

解 θ为尺度参数，取无信息先验$\pi(\theta)=\frac{1}{\theta}, \theta>0$. 记$T=\sum_{i=1}^n X_i$,$t=\sum_{i=1}^n x_i$. 已知$T$为$\theta$的充分统计量，且服从$\Gamma(n,\frac{1}{\theta})$，则

$$\pi(\theta|\widetilde{x})=\pi(\theta|t)\propto_\theta p(t|\theta)\pi(\theta)\propto_\theta \theta^{-n}\exp\{-t/\theta\}\cdot\theta^{-1}.$$

所以

$$\pi(\theta|\tilde{x}) = \pi(\theta|t) = \frac{\theta^{-n-1}\exp\{-t/\theta\}}{\int_0^\infty \theta^{-n-1}\exp\{-t/\theta\}d\theta}$$
$$= \frac{t^n}{\Gamma(n)}\theta^{-(n+1)}\exp\{-t/\theta\}, \quad \theta > 0,$$

即后验分布为倒gamma分布，参数为n, t. 故θ的期望型Bayes点估计量为

$$\widehat{\theta}_B = \mathsf{E}[\theta|\widetilde{X}] = \mathsf{E}[\theta|T] = \frac{1}{n-1}T = \frac{1}{n-1}\sum_{i=1}^n X_i.$$

习题:

1.求θ的MLE和众数型Bayes点估计量;

2*. 求θ的中位数型Bayes点估计量(提示：寻找θ的分布与χ^2分布的关系).

Jeffrey's prior

设样本$\widetilde{X}=(X_1,\dots X_n)\sim p(\widetilde{x}|\boldsymbol{\theta})$, $\boldsymbol{\theta}=(\theta_1,\dots,\theta_p)$为$p$为参数向量. 在对$\boldsymbol{\theta}$ 无先验信息可用时, Jeffreys 用Fisher信息矩阵行列式的平方根作为$\boldsymbol{\theta}$的无信息先验, 这样的无信息先验称为Jeffrey先验(Jeffrey's Prior).其求解步骤如下:

- 写出样本的对数似然函数
$$l(\boldsymbol{\theta}|\widetilde{x})=\log p(\widetilde{x}|\boldsymbol{\theta}),$$
- 求样本的信息矩阵
$$\boldsymbol{I}(\boldsymbol{\theta})=(I_{ij}(\boldsymbol{\theta}))_{p\times p},\quad I_{ij}(\boldsymbol{\theta})=\mathsf{E}_{\widetilde{X}|\boldsymbol{\theta}}\left\{-\frac{\partial^2 l}{\partial\theta_i\partial\theta_j}\right\},$$
- $\boldsymbol{\theta}$的无信息先验的密度函数取为$\pi(\boldsymbol{\theta})\propto|\boldsymbol{I}(\boldsymbol{\theta})|^{1/2}$, 其中$|\boldsymbol{I}(\boldsymbol{\theta})|$为$I(\boldsymbol{\theta})$ 的行列式.

Example

设θ是Bernoulli试验中的成功概率，则在n次独立重复的Bernoulli试验中，成功次数$X \sim B(n,\theta)$. 求θ的Jeffery's prior.

解 X的分布为

$$P(X=x|\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x},\quad x=0,1,\ldots,n.$$

对数似然函数为

$$l(\theta)=\log\binom{n}{x}+x\log\theta+(n-x)\log(1-\theta),$$

$$\frac{\partial^2 l}{\partial\theta^2}=-\frac{x}{\theta^2}-\frac{n-x}{(1-\theta)^2}.$$

故有

$$I(\theta) = \mathsf{E}_{X|\theta}\left[\frac{X}{\theta^2} + \frac{n-X}{(1-\theta)^2}\right] = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)}.$$

因此取$\pi(\theta) \propto I(\theta)^{1/2} \propto \theta^{-1/2}(1-\theta)^{-1/2}$. 添上正则化因子的无信息密度为

$$\pi(\theta) = \frac{1}{\pi}\frac{1}{\sqrt{\theta(1-\theta)}}, \quad 0 < \theta < 1; \quad F(\theta) = \frac{2}{\pi}\arcsin(\sqrt{\theta}).$$

它是beta分布$Beta(\frac{1}{2}, \frac{1}{2})$.

Brown 运动在$[0,1]$上的最大零点服从这一分布; 大维随机矩阵理论中的"半圆律"与之密切相关.

Example

设$\widetilde{X}=(X_1,\ldots,X_n)$是从总体$N(\mu,\sigma^2)$中抽取的简单样本，记$\boldsymbol{\theta}=(\mu,\sigma)$，求$(\mu,\sigma)$的联合无信息先验.

(注意：不是(μ,σ^2))

解： 给定$\widetilde{X}=\widetilde{x}$，$\boldsymbol{\theta}$的似然函数为

$$L(\boldsymbol{\theta}|\widetilde{x})=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}.$$

对数似然函数为

$$l(\boldsymbol{\theta}|\widetilde{x})=-\frac{n}{2}\log 2\pi-n\log\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

求偏导得到

$$\frac{\partial l}{\partial \mu}=\sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^2},\quad \frac{\partial l}{\partial \sigma}=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2,$$

$$\frac{\partial^2 l}{\partial \mu^2}=-\frac{n}{\sigma^2},\quad \frac{\partial^2 l}{\partial \sigma^2}=\frac{n}{\sigma^2}-\frac{3}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2,$$

$$\frac{\partial^2 l}{\partial \sigma\partial \mu}=-2\sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^3}.$$

所以

$$I_{11}(\boldsymbol{\theta})=\frac{n}{\sigma^2},\quad I_{12}(\boldsymbol{\theta})=I_{21}(\boldsymbol{\theta})=0,$$

$$I_{22}(\boldsymbol{\theta})=-\frac{n}{\sigma^2}+\frac{3}{\sigma^4}\sum_{i=1}^{n}E(X_i-\mu)^2=\frac{2n}{\sigma^2}.$$

所以

$$\boldsymbol{I}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{pmatrix}, \quad |\boldsymbol{I}(\boldsymbol{\theta})|^{1/2} = \frac{\sqrt{2}n}{\sigma^2}.$$

所以(μ, σ)的Jeffreys先验为

$$\pi(\mu, \sigma) = \frac{1}{\sigma^2}.$$

- 当σ已知时, $I(\mu)=n/\sigma^2$不依赖于未知参数, 故取$\pi_1(\mu)\equiv 1$;
- 当μ已知时, $I(\sigma)=2n/\sigma^2$, 故取$\pi_2(\sigma)=1/\sigma$, $\sigma>0$. 这与前面讲过的尺度参数的无信息先验相同;
- 当μ和σ无信息先验独立时, $\pi(\mu,\sigma)=\pi_1(\mu)\pi_2(\sigma)=1/\sigma$, $\sigma>0$;
- 当μ和σ无信息先验不独立时, $\pi(\mu,\sigma)=1/\sigma^2$, $\sigma>0$.

选取先验分布的其它方法

Example

$$X_1, \ldots, X_n\big|_\theta \sim N(\theta, \sigma^2), \quad \sigma^2 \text{已知},$$
$$\theta \sim N(0, \tau^2).$$

为了求给定$\widetilde{X} = \widetilde{x}$时的后验分布，需要知道$\tau^2$的值，先验信息比较充分时,可以利用先验信息估计出$\tau^2$. 如果没有先验信息，就遇到了麻烦.

下面，从另一个角度出发.

$(\widetilde{X},\theta)$的联合密度为

$$
\begin{aligned}
p(\widetilde{x},\theta) =& p(\widetilde{x}|\theta)\pi(\theta|\tau^2) \\
=& \frac{1}{(2\pi\sigma^2)^{n/2}}\frac{1}{\sqrt{2\pi\tau^2}}\exp\left\{-\frac{\sum_{i=1}^n(x_i-\theta)^2}{2\sigma^2}-\frac{\theta^2}{2\tau^2}\right\}
\end{aligned}
$$

$\widetilde{X}$的边际密度为

$$m(\widetilde{x}|\tau^2) = \int p(\widetilde{x}|\theta)\pi(\theta|\tau^2)d\theta.$$

它是参数τ^2的函数，也可视为参数τ^2的似然函数$L(\tau^2|\widetilde{x})$，含了参数τ^2的信息，可以求得极大似然估计$\widehat{\tau^2}$.

这样我们选取先验分布为$N(0, \widehat{\tau^2})$.

这种方法叫做经验Bayes方法(Empirical Bayesian Method).

由于$\overline{X}$是充分统计量，所以

$$p(\widetilde{x}|\theta) \propto_\theta \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{n(\overline{x}-\theta)^2}{2\sigma^2}} \propto_\theta e^{-\frac{n(\overline{x}-\theta)^2}{2\sigma^2}}.$$

所以

$$m(\widetilde{x}|\tau^2) \propto_{\tau^2} \int e^{-\frac{n(\overline{x}-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{\theta^2}{2\tau^2}} d\theta$$
$$= \left(\frac{\sigma^2}{\sigma^2+n\tau^2}\right)^{1/2} e^{-\frac{1}{2}\frac{n\overline{x}^2}{\sigma^2+n\tau^2}}.$$

$\sigma^2+n\tau^2$的MLE为$\max\{\sigma^2, n\overline{x}^2\}$. 所以

$$\widehat{\tau^2}_{MLE} = \max\{\sigma^2/n, \overline{x}^2\} - \sigma^2/n.$$

θ 的Bayes估计为

$$\widehat{\theta}_B = \mathsf{E}[\theta|\widetilde{x}, \widehat{\tau^2}_{MLE}] = \left(1 - \frac{\sigma^2}{\widehat{\sigma^2+n\tau^2}}\right)\overline{x}$$
$$= \left(1 - \frac{\sigma^2}{\max\{\sigma^2, n\overline{x}^2\}}\right)\overline{x}.$$

Empirical Bayes

一般地,设样本$\widetilde{X}$的pdf(Or pmf)为$p(\widetilde{x}|\theta)$，先验分布为$\pi(\theta|\gamma)$．在Bayes意义下，$\widetilde{X}$的pdf(Or pmf)为

$$m(\widetilde{x}|\gamma) = \int p(\widetilde{x}|\theta)\pi(\theta|\gamma)d\theta.$$

可以求得γ的(似然估计) $\widehat{\gamma}$. 然后，以$\pi(\theta|\widehat{\gamma})$为$\theta$的先验分布进行Bayes统计分析.

Efron, B. (2009), Empirical Bayes estimates for larger-scale prediction problems, *Journal of the American Statistical Association*, **104**: 1025-1028.

Efron, B. (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation Testing, and Prediction.* Institute of Mathematical Statistics Monographs, Vol. I, Cambridge, Cambridge University Press.

Hierarchical Bayes

设 $\widetilde{X}\big|_\theta \sim p(\widetilde{x}|\theta), \qquad \theta\big|_\gamma \sim \pi(\theta|\gamma).$

为了确定θ 的先验分布，我们再把参数γ当作随机变量，选取一个分布$\psi(\gamma)$作为γ的先验分布，即

$$\begin{array}{c} \widetilde{X}\big|_\theta \sim p(\widetilde{x}|\theta), \\ \theta\big|_\gamma \sim \pi(\theta|\gamma), \\ \gamma \sim \psi(\gamma). \end{array} \qquad \left\{ \begin{array}{c} \widetilde{X}\big|_\theta \sim p(\widetilde{x}|\theta), \\ \theta\big|_\gamma \sim \pi(\theta), \\ \pi(\theta) = \int \pi(\theta|\gamma)\psi(\gamma)d\gamma. \end{array} \right.$$

这种方法叫做分层Bayes方法. 通常取$\pi(\theta|\gamma)$为对应于$p(\widetilde{x}|\theta)$的共轭先验分布，取$\psi(\gamma)$为对应于$\pi(\theta|\gamma)$ 的共轭先验分布或无先验分布.

The development of Hierarchical Bayes is another interesting story. For example,

- How to compute $\mathsf{E}[\theta|\widetilde{x}]$?—Markov chain Monte Carlo (MCMC) technique.
- How to look for a decomposition of the prior $\pi(\theta)$ of the form $\pi(\theta) = \int \pi(\theta|\gamma)\psi(\gamma)d\gamma$? —hidden Markov models, hidden mixtures, deconvolution.

加权平均均方误差与Bayes 估计

设θ为待估参数, $\widetilde{X}$为一个样本.

前面我们已证明: 在均方误差意义下, θ的最优估计量一般是不存在的, 即, 不存在这样的估计量$\delta(\widetilde{X})$使得

$$\mathsf{E}_\theta\{\theta-\delta(\widetilde{X})\}^2 \text{ 关于}\theta\text{一致地最小(对}\theta\in\Theta\text{)}.$$

而将寻找范围缩小到无偏估计类时, 这样的最优估计量在一定条件下存在且唯一.

下面我们换一个角度思考最优估计的问题. 我们不需要

$$R(\delta,\theta)=\mathsf{E}_\theta\{\theta-\delta(\widetilde{X})\}^2 \text{ 关于}\theta\text{一致地最小},$$

而考察使得均方误差$R(\delta,\theta)$关于θ的加权平均最小的问题:

$$\min_\delta \int R(\delta,\theta)\pi(\theta)d\theta,$$

其中$\pi(\theta)\geq 0$是某个给定的权函数.

我们不妨设

$$\int \pi(\theta)d\theta=1.$$

这样的$\pi(\theta)$可以看作一个概率密度函数.

若记样本的联合概率密度(pdf)或分布列(pmf)为$p(\tilde{x}|\theta)$, 则

$$\int R(\delta,\theta)\pi(\theta)d\theta = \int\int \{\theta-\delta(\tilde{x})\}^2 p(\tilde{x}|\theta)\pi(\theta)d\tilde{x}d\theta.$$

这恰好是将$\theta\sim\pi(\theta)$看作随机变量时, 在Bayes意义下, $(\widetilde{X},\theta)$的函数$\{\theta-\delta(\widetilde{X})\}^2$ 的期望, 即

$$\int R(\delta,\theta)\pi(\theta)d\theta = \mathsf{E}\{\theta-\delta(\widetilde{X})\}^2.$$

$\mathsf{E}\{\theta-\delta(\widetilde{X})\}^2$称为Bayes风险.

如果写$p(\tilde{x}|\theta)\pi(\theta) = \pi(\theta|\tilde{x})p(\tilde{x})$, 其中$\pi(\theta|\tilde{x})$ 是θ的后验分布, $p(\tilde{x})$是$\tilde{X}$的边际分布, 那么

$$\begin{aligned}\int R(\delta,\theta)\pi(\theta)d\theta =& \mathsf{E}\{\theta-\delta(\tilde{X})\}^2\\ =& \int\left[\int\{\theta-\delta(\tilde{x})\}^2\pi(\theta|\tilde{x})d\theta\right]p(\tilde{x})d\tilde{x}\\ =& \int \mathsf{E}\left[\{\theta-\delta(\tilde{X})\}^2\middle|\tilde{X}=\tilde{x}\right]p(\tilde{x})d\tilde{x}.\end{aligned}$$

$\mathsf{E}\left[\{\theta-\delta(\tilde{X})\}^2\middle|\tilde{X}=\tilde{x}\right]$称为后验期望损失 (posterior expected loss), 或者后验均方误差(posterior mean square error, PMSE).

为使得$\int R(\delta,\theta)\pi(\theta)d\theta$最小，只要

$$\min_{\delta(\widetilde{x})} \mathsf{E}\left[\{\theta-\delta(\widetilde{X})\}^2\big|\widetilde{X}=\widetilde{x}\right].$$

而

$$\arg\min_{\delta(\widetilde{x})} \mathsf{E}\left[\{\theta-\delta(\widetilde{X})\}^2\big|\widetilde{X}=\widetilde{x}\right]=\mathsf{E}[\theta|\widetilde{X}=\widetilde{x}]=\int\theta\pi(\theta|\widetilde{x})d\theta$$

即为后验期望，这就是期望型Bayes点估计.

所以

$$\mathsf{E}[\theta|\widetilde{X}=\widetilde{x}]=\arg\min_{\delta(\widetilde{x})}\int R(\delta,\theta)\pi(\theta)d\theta.$$

即

$$\mathsf{E}[\theta|\widetilde{X}]=\arg\min_{\delta(\widetilde{X})}\int \mathsf{E}_\theta\{\theta-\delta(\widetilde{X})\}^2\pi(\theta)d\theta.$$

结论 对给定的权函数$\pi(\theta)$，在加权均方误差(或称Bayes均方风险意义)下，期望型Bayes点估计是最优估计.

即

$$
\begin{aligned}
&\arg\min_{\delta(\tilde{x})} \int \mathsf{E}_\theta(\theta - \delta(\widetilde{X}))^2 \pi(\theta) d\theta \\
=&\arg\min_{\delta(\tilde{x})} \mathsf{E}\left[(\theta - \delta(\widetilde{X}))^2 \middle| \widetilde{X} = \tilde{x}\right] \\
=&\mathsf{E}[\theta|\tilde{x}] = \text{mean of } \ \pi(\theta|\tilde{x}).
\end{aligned}
$$

若以$|\theta-\delta(\widetilde{X})|$代替$\{\theta-\delta(\widetilde{X})\}^2$，类似可得

$$
\begin{aligned}
&\arg\min_{\delta(\widetilde{x})}\int \mathsf{E}_\theta|\theta-\delta(\widetilde{X})|\pi(\theta)d\theta \\
=&\arg\min_{\delta(\widetilde{x})}\mathsf{E}\left[|\theta-\delta(\widetilde{X})|\,\middle|\,\widetilde{X}=\widetilde{x}\right] \\
=&\text{median of }\ \pi(\theta|\widetilde{x}).
\end{aligned}
$$

推广：Loss Function Optimality

在$|\theta-\delta(\tilde{x})|$ 和$\{\theta-\delta(\tilde{x})\}^2$ 中，$\delta(\tilde{x})$是根据样本$\tilde{x}$采取的决策，称为决策函数．而$|\theta-\delta(\tilde{x})|$ 和$\{\theta-\delta(\tilde{x})\}^2$是这种决策导致的损失，叫做损失函数(loss function)，前者叫做绝对值损失，后者叫做平方损失．

一般地，设$\delta=\delta(\tilde{x})$是根据样本$\tilde{x}$采取的决策行动，参数为θ，由此造成的损失是$L(\delta,\theta)$. 称

$$R(\delta,\theta)=E[L(\delta(\tilde{X}),\theta)|\theta]$$

为δ的风险函数(risk function). 而称平均损失

$$R_\pi(\delta)=\int_\Theta R(\delta,\theta)\pi(\theta)d\theta$$

为δ在先验分布π下的Bayes风险(Bayesian risk).

称

$$R(\delta|\tilde{x})=E_{\theta|\tilde{x}}[L(\delta,\theta)]=\int_\Theta L(\delta,\theta)\pi(\theta|\tilde{x})d\theta$$

为决策函数δ的后验风险(posterior risk)，其中$\pi(\theta|\tilde{x})$为θ的后验分布(pdf或pmf).

显然

$$
\begin{array}{lcl}
R(\delta) & = & \displaystyle\int_{\Theta}\left[\int_{\mathscr{X}} L(\delta,\theta)p(\widetilde{x}|\theta)d\widetilde{x}\right]\pi(\theta)d\theta \\
\| & & \\
E[R(\delta|\widetilde{X})] & = & \displaystyle\int_{\mathscr{X}}\left[\int_{\Theta} L(\delta,\theta)\pi(\theta|\widetilde{x})d\theta\right]p(\widetilde{x})d\widetilde{x}.
\end{array}
$$

后验风险最小原则

Theorem

设先验分布为π. 如果δ_π是在后验风险意义下的最优决策，即

$$R(\delta_\pi|\tilde{x}) = \inf_{\delta} R(\delta|\tilde{x}),$$

则δ_π是$Bayes$风险意义下的最优决策,即

$$R(\delta_\pi) \le R(\delta), \quad \forall \delta.$$

也就是说

$$\arg\inf_{\delta} R(\delta) = \arg\inf_{\delta} R(\delta|\tilde{x}).$$

加权平方损失下的Bayes估计

Theorem

设$\delta = \delta(\tilde{x})$为一决策函数，则在加权平均损失$L(\delta, \theta) = w(\theta)(\delta - \theta)^2$下，$\theta$的$Bayes$估计值为

$$\widehat{\theta}_B = \frac{E[\theta w(\theta)|\tilde{x}]}{E[w(\theta)|\tilde{x}]},$$

其中$w(\theta)$在参数空间上恒为正.

证明： 后验风险为

$$\begin{aligned} R(\delta|\tilde{x}) =& \mathsf{E}[w(\theta)(\theta-\delta)^2|\tilde{x}] \\ =& \mathsf{E}[\theta^2 w(\theta)|\tilde{x}] - 2\delta \mathsf{E}[\theta w(\theta)|\tilde{x}] + \delta^2 \mathsf{E}[w(\theta)|\tilde{x}]. \end{aligned}$$

显然，当

$$\delta = \frac{\mathsf{E}[\theta w(\theta)|\tilde{x}]}{\mathsf{E}[w(\theta)|\tilde{x}]}$$

时，后验风险$R(\delta|\tilde{x})$达到最小. 故

$$\widehat{\theta}_B = \frac{\mathsf{E}[\theta w(\theta)|\tilde{x}]}{\mathsf{E}[w(\theta)|\tilde{x}]}.$$