第二节 最小二乘信度信度

一、符号

设风险子集的一个投保人的风险水平可以通过一个参数θ来描述，但是θ的取值随投保人的不同而不同。这样，通过不同取值θ，我们可以区分不同投保人的风险水平的差异。θ具有如下特点：

θ是存在的，但θ是不可观察的，而且我们永远不知道θ的真实值。

θ随投保人变化，可以看作随机变量Θ的观察值，在风险子集内存在一个关于Θ的概率分布$\pi(\theta)$

假定$\pi(\theta)$是已知的，但每个投保人的具体风险参数θ的值都是未知的。

记号：

设投保人的风险参数为Θ，则

$F_{X|\Theta}(x|\theta)=F(x;\theta)$ 为投保人的损失分布函数

$\mu(\Theta)=E(X|\Theta)$ 为投保人的期望损失，称为风险保费。

$\sigma^2(\Theta)=\mathrm{var}(X|\Theta)=E[(X-\mu(\Theta))^2|\Theta]$ 为投保人的方差。

设 $\pi(\Theta)$ 为 Θ 的分布密度。则任意选取一个投保人的损失 X 的分布为

$$F_X(x)=\int F_{X\Theta}(x|\theta)d\pi(\theta)$$

投保人的平均损失为

$$\mu=E[X]=E[\mu(\Theta)]$$

投保人的方差为

$$\mathrm{var}(X) = E(\mathrm{var}(X \mid \Theta)) + \mathrm{var}(E(X \mid \Theta)) = v + a$$

其中

- $v = E[\sigma^2(\Theta)]$ 为组内方差的均值（average variance within subgroups），也被称为 Expected Value of the Process Variance (EPV)。Process Variance 是个体风险实际经验与其预期值的差异，是衡量个人风险损失经历变异性的指标。Process Variance 的预期值（EPV）是整个风险类别的过程差异的平均值。$a = \mathrm{var}[\mu(\Theta)]$ 为各组均值的方差(variance between subgroup averages), 也被称为 Variance of the Hypothetical Means (VHM)。VHM 衡量的是分类的同质性。如果 VHM 的值比较小，说明各组的差异性不大，具有同质性。反之，则说明具有异质性。
- μ称为无条件均值，可视为不知投保人任何信息时，对其风险的估计。

二、Bühlmann 的基本模型

- 基本思想：使用过去的数据 $\vec{X}$ 来估计 $\mu_{n+1}(\theta)$ 时，将估计量的形式限制在形如 $\alpha_0+\Sigma_{i=1}^n\alpha_i X_i$ 的线性估计类上，其中参数 $\alpha_0,\alpha_1,\cdots,\alpha_n$ 满足使平方误差

$$Q=E\left\{[\mu_{n+1}(\Theta)-(\alpha_0+\Sigma_{i=1}^n\alpha_i X_i)]^2\right\} \tag{1}$$

最小。

- 模型求解

为最小化 Q，我们采用求微分的方法。首先对 α_0 微分

$$\frac{\partial Q}{\partial\alpha_0}=E\{-2(\mu_{n+1}(\Theta)-\alpha_0-\Sigma_{i=1}^n\alpha_i X_i)\} \tag{2}$$

假设 $\hat{\alpha}_0,\hat{\alpha}_1,\cdots,\hat{\alpha}_n$ 使得 Q 达到最小，则 $\partial Q/\partial\hat{\alpha}_0=0$。因此

$$E(\mu_{n+1}(\Theta)) = E(X_{n+1}) = \hat{\alpha}_0 + \Sigma_{i=1}^{n}\hat{\alpha}_i E(X_i) \qquad （3）$$

可以看出，$\hat{\alpha}_0 + \Sigma_{i=1}^{n}\hat{\alpha}_i E(X_i)$ 是 X_{n+1} 的无偏估计，因此（3）式也称为是**无偏方程**。

对 $i = 1, 2, \cdots, n$，同样有

$$\frac{\partial Q}{\partial \hat{\alpha}_i} = E\{2\left[\mu_{n+1}(\Theta) - \hat{\alpha}_0 - \Sigma_{j=1}^{n}\hat{\alpha}_j X_j\right](-X_i)\} = 0$$

化简得

$$E[\mu_{n+1}(\Theta)X_i] = \hat{\alpha}_0 E(X_i) + \Sigma_{j=1}^{n}\hat{\alpha}_j E(X_i X_j). \qquad （4）$$

又由于

$$
\begin{aligned}
E[\mu_{n+1}(\Theta)X_i] &= E\{E[X_i\mu_{n+1}(\Theta)\,|\,\Theta]\} \\
&= E\{\mu_{n+1}(\Theta)E[X_i\,|\,\Theta]\} \\
&= E\{E[X_{n+1}\,|\,\Theta]E[X_i\,|\,\Theta]\} \\
&= E\{E[X_{n+1}X_i\,|\,\Theta]\} \\
&= E(X_{n+1}X_i)
\end{aligned}
\tag{5}
$$

注意第二个等式利用了条件期望的性质：设 X 和 Y 是任意两个随机变量， $g(\cdot)$ 是一个函数，则

$$E(Xg(Y)\,|\,Y) = g(Y)E(X\,|\,Y),$$

第四个等式是由于给定 Θ ， X_{n+1}, X_i 是相互独立的。

因此

$$E(X_iX_{n+1}) = \hat{\alpha}_0 E(X_i) + \sum_{j=1}^{n} \hat{\alpha}_j E(X_iX_j) \tag{6}$$

将（3）式两端同时乘以$E(X_i)$得

$$E(X_i)E(X_{n+1}) = \hat{\alpha}_0 E(X_i) + \sum_{j=1}^{n} \hat{\alpha}_j E(X_i)E(X_j) \qquad (7)$$

将（6）减去（7）得

$$\operatorname{cov}(X_i, X_{n+1}) = \sum_{j=1}^{n} \hat{\alpha}_j \operatorname{cov}(X_i, X_j), \quad i = 1, 2, \cdots, n \qquad (8)$$

由上面的 n 个方程和无偏方程（3），我们可解出$\hat{\alpha}_0, \hat{\alpha}_1, \cdots, \hat{\alpha}_n$。这时$\mu_{n+1}(\theta)$的估计值

$$\hat{\alpha}_0 + \sum_{j=1}^{n} \hat{\alpha}_j X_j$$

称为X_{n+1}的信度估计

下面两条定理将说明，这种 Bühlmann 线性估计不但是对 $E(X_{n+1}|\Theta)$ 的线性估计中均方误差最小估计，也是 X_{n+1} 和贝叶斯信度估计值的最小误差线性估计。

定理 1 Bühlmann 线性估计在所有用 $\vec{X}$ 对 X_{n+1} 的线性估计中均方误差最小。

定理 2 Bühlmann 线性估计在所有用 $\vec{X}$ 对贝叶斯信度估计值的线性估计中均方误差最小的。

例 1：如果 $E(X_j)=\mu, \mathrm{var}(X_j)=\sigma^2$，且 $\mathrm{cov}(X_i, X_j)=\rho\sigma^2$，$i \neq j$，其中 $-1<\rho<1$，求 X_{n+1} 的信度估计 $\hat{\alpha}_0+\sum_{j=1}^{n}\hat{\alpha}_j X_j$。

解：由方程（3）得到

$$\mu=\hat{\alpha}_0+\mu\sum_{j=1}^{n}\hat{\alpha}_j \tag{9}$$

或

$$\sum_{j=1}^{n}\hat{\alpha}_j=1-\hat{\alpha}_0/\mu$$

于是（8）的 n 个方程

$$\operatorname{cov}(X_i, X_{n+1}) = \sum_{j=1}^{n} \widehat{\alpha}_j \operatorname{cov}(X_i, X_j), \quad i = 1, 2, \cdots, n$$

变为

$$\rho = \sum_{\substack{j=1 \\ j \neq i}}^{n} \widehat{\alpha}_j \rho + \widehat{\alpha}_i$$

或者
$$\rho = \sum_{j=1}^{n} \widehat{\alpha}_j \rho + \widehat{\alpha}_i (1-\rho), \quad i = 1, \cdots, n$$

于是解得

$$\widehat{\alpha}_i = \frac{\rho(1-\sum_{j=1}^{n} \widehat{\alpha}_j)}{1-\rho} = \frac{\rho \widehat{\alpha}_0}{\mu(1-\rho)}$$

将 i 从 1 到 n 相加得到，

$$\sum_{i=1}^{n}\widehat{\alpha}_i=\sum_{j=1}^{n}\widehat{\alpha}_j=\frac{n\rho\widehat{\alpha}_0}{\mu(1-\rho)}$$

结合方程（9），得到

$$1-\widehat{\alpha}_0/\mu=\frac{n\rho\widehat{\alpha}_0}{\mu(1-\rho)}$$

解出 $\widehat{\alpha}_0$ 得到，

$$\widehat{\alpha}_0=\frac{(1-\rho)\mu}{1-\rho+n\rho}$$

从而得到，

$$\widehat{\alpha}_j=\frac{\rho\widehat{\alpha}_0}{\mu(1-\rho)}=\frac{\rho}{1-\rho+n\rho}$$

因此， X_{n+1} 的信度估计为

$$\widehat{\alpha}_0+\sum_{j=1}^{n}\widehat{\alpha}_j X_j=\frac{(1-\rho)\mu}{1-\rho+n\rho}+\sum_{j=1}^{n}\frac{\rho X_j}{1-\rho+n\rho}$$
$$=(1-z)\mu+z\bar{X}$$

其中 $z=n\rho/(1-\rho+n\rho)$ 为信度因子和 $\bar{X}=n^{-1}\Sigma_{j=1}^{n}X_j$ 为经验均值。

例 2：已知如下信息

First Observation	Unconditional Probability	Bayesian Estimate of Second Observation
1	1/3	1.50
2	1/3	1.50
3	1/3	3.00

Calculate the Bühlmann credibility estimate of the second observation, given that the first observation is 1.

解： 贝叶斯信度估计

$$E(X_{n+1} \mid \vec{X} = \vec{x}) == z\bar{x} + (1-z)\mu_{n+1},$$

在本例中，$\bar{x}$ 实际上是第一个观察值，Bühlmann credibility 应该是贝叶斯估计的最小二乘估计，因此 z 和 μ_{n+1} 应该使得下面式子最小化

$$\frac{1}{3}[Z+(1-Z)\mu-1.5]^2+\frac{1}{3}[2Z+(1-Z)\mu-1.5]^2+\frac{1}{3}[3Z+(1-Z)\mu-3]^2$$

将上式对μ和 Z 求导，并令其等于 0。并且注意到μ是贝叶斯估计的均值

$$\mu=\frac{1}{3}(1.5+1.5+3)=2.$$

于是，只需将μ代入，并对 Z 求导

$$2(-Z+0.5)(-1)+2(0.5)(0)+2(Z-1)(1)=0$$

$$4Z-3=0, Z=0.75.$$

因此，Bühlmann credibility 估计为

$$0.75(1)+0.25(2)=1.25.$$

二、Buhlmann 模型

1、假定：给定$\Theta=\theta$的条件下，$X_1,\cdots,X_n$是独立同分布的随机变量。

记号：

$$\mu(\theta)=E(X_j\mid\Theta=\theta)$$

$$v(\theta)=\mathrm{var}(X_j\mid\Theta=\theta)$$

定义

- $\mu=E(\mu(\Theta))=E[X]$

为组内方差的均值（average variance within subgroups(EPV)）

- $v=E[v(\Theta)]$ 为组内方差的均值（average variance within subgroups）,也被称为 Expected Value of the Process Variance (EPV)

- $a=\mathrm{var}[\mu(\Theta)]$ 为各组均值的方差(variance between subgroup averages)，也被称为 Variance of the Hypothetical Means (VHM)

2、信度估计

下面我们来推导 X_{n+1} 的信度估计。为确定 $\hat{\alpha}_0,\hat{\alpha}_1,\cdots,\hat{\alpha}_n$，我们需求解无偏方程（3）和（8）构成的方程组，即

$$E(\mu_{n+1}(\Theta))=E(X_{n+1})=\alpha_0+\Sigma_{i=1}^{n}\alpha_i E(X_i)$$

$$\mathrm{cov}(X_i,X_{n+1})=\sum_{j=1}^{n}\alpha_j\,\mathrm{cov}(X_i,X_j),\quad i=1,2,\cdots,n$$

由于

$$E(X_j)=E[E(X_j\mid\Theta)]=E(\mu(\Theta))=\mu$$

$$\begin{aligned}\operatorname{var}(X_j)&=E(\operatorname{var}(X_j\mid\Theta))+\operatorname{var}(E(X_j\mid\Theta))\\&=E[v(\Theta)]+\operatorname{var}(\mu(\Theta))\\&=v+a\\&=\text{VHM+EPV}\end{aligned}$$

当$i\neq j$时

$$\begin{aligned}\operatorname{cov}(X_i,X_j)&=E(X_iX_j)-E(X_i)E(X_j)\\&=E(E(X_iX_j\mid\Theta))-\mu^2\\&=E\{[E(X_i\mid\Theta)E(X_j\mid\Theta)]-\{E[\mu(\Theta)]\}^2\\&=E\{[\mu(\Theta)]^2\}-\{E[\mu(\Theta)]\}^2\\&=\operatorname{var}(\mu(\Theta))=a=VHM\end{aligned}$$

因此，Buhlmann 模型满足例 1 的条件，其中$\sigma^2=v+a$， $\rho=a/(v+a)$，

代入例 1 中 $\widehat{\alpha}_0,\widehat{\alpha}_1,\cdots,\widehat{\alpha}_n$ 的表示式

$$\widehat{\alpha}_0=\frac{(1-\rho)\mu}{1-\rho+n\rho}$$

$$\widehat{\alpha}_j=\frac{\rho}{1-\rho+n\rho}$$

化简得到

$$\widehat{\alpha}_0=\frac{(1-\rho)\mu}{1-\rho+n\rho}=\frac{(1-\frac{a}{(v+a)})\mu}{1-\frac{a}{(v+a)}+n(\frac{a}{(v+a)})}=\frac{v\mu}{v+na}$$

$$\widehat{\alpha}_j=\frac{\rho}{1-\rho+n\rho}=\frac{\frac{a}{(v+a)}}{1-\frac{a}{(v+a)}+n\frac{a}{(v+a)}}=\frac{a}{v+na}$$

即

$$\begin{cases} \widehat{\alpha}_0 = \dfrac{v}{v+na}\mu \\ \widehat{\alpha}_j = \dfrac{a}{v+na}, j=1,2,\cdots,n \end{cases}$$

X_{n+1} 的信度估计为

$$\begin{aligned} \widehat{\alpha}_0 + \sum_{j=1}^{n} \widehat{\alpha}_j X_j &= \frac{v}{v+na}\mu + \sum_{j=1}^{n} \frac{a}{v+na} X_j \\ &= (1-z)\mu + z\bar{X} \end{aligned}$$

其中

$$z = \frac{na}{v+na} = \frac{n}{n+k} \tag{10}$$

$$k=\frac{v}{a}=\frac{E(\mathrm{var}(X_j\,|\,\Theta))}{\mathrm{var}(E(X_j\,|\,\Theta))}=\frac{EPV}{VHM}$$

z 称为 Buhlmann 信度因子。

3、信度因子的直观解释

1、当样本数$n\to\infty$时，$z\to 1$，权重主要集中在$\bar{X}$上。

2、k 的值越小，z 的值越大，说明经验数据的可信度大。反之亦然。

3、如果在个体风险的实际损失经验中预期存在大量变化（EPV 值比较大），则观察到的实际经验可能远远低于其预期值，并且对于估计预期值不是非常有用。在这种情况下，应该将较少的权重，即较低的可信度分配给个人经验。

4、VHM 衡量的是分类的同质性。如果 VHM 的值比较小，说明个体的差异性不

大，具有同质性。个体的经验数据与预期值的差异应该是由随机波动引起的，因此，应该将较少的权重，即较低的可信度分配给个体经验

例 2：设$\{X_j|\Theta; j=1,\cdots,n\}$是独立同分布的泊松随机变量，参数为$\Theta$，而$\Theta$又服从 gamma 分布，

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

求 X_{n+1} 的 Buhlmann 信度估计。

解：由于$\{X_j|\Theta; j=1,\cdots,n\}$服从泊松分布，

$$\mu(\theta)=E(X_j|\Theta=\theta)=\theta$$

$$v(\theta)=\mathrm{var}(X_j|\Theta=\theta)=\theta$$

因此，

$$\mu = E[\mu(\Theta)] = E(\Theta) = \alpha\beta$$

$$v = EPV = E[v(\Theta)] = E(\Theta) = \alpha\beta$$

以及

$$a = VHM = \mathrm{var}[\mu(\Theta)] = \mathrm{var}(\Theta) = \alpha\beta^2$$

于是，

$$k = \frac{v}{a} = \frac{EPV}{VHM} = \frac{\alpha\beta}{\alpha\beta^2} = \frac{1}{\beta}$$

$$z = \frac{n}{n+k} = \frac{n}{n+1/\beta} = \frac{n\beta}{n\beta+1}$$

Buhlmann 信度估计为

$$z\bar{X} + (1-z)\mu = \frac{n\beta}{n\beta+1}\bar{X} + \frac{1}{n\beta+1}\alpha\beta$$

课堂练习

1. You are given:

(i) The annual number of claims for an insured has probability function:

$$p(x)=\binom{3}{x}q^{x}(1-q)^{3-x}, \quad x=0,1,2,3$$

(ii) The prior density is

$$\pi(q)=2q, \quad 0<q<1.$$

A randomly chosen insured has zero claims in Year 1.

Using Bühlmann credibility, estimate the number of claims in Year 2 for the selected insured.

(A) 0.33 (B) 0.50 (C) 1.00 (D) 1.33 (E) 1.50

解： 设 X 表示理赔次数，则给定 q，X 服从二项分布，

$$E(X\mid q)=3q\text{，}\mathrm{var}(X\mid q)=3q(1-q)$$

$$\mu=E(3q)=\int_0^1 3q2qdq=2q^3\mid_0^1=2$$

$$EPV=E(3q(1-q))=\int_0^1 3q(1-2q)2qdq=2q^3-1.5q^4\mid_0^1=0.5$$

$$VHM=a=\mathrm{var}(3q)=E(9q^2)-\mu^2=\int_0^1 9q^2 2qdq-2^2$$

$$=4.5q^4\mid_0^1-4=4.5-4=0.5$$

$$k=v/a=0.5/0.5=1$$

因此， $z=\dfrac{n}{n+k}=\dfrac{1}{1+1}=0.5$，选 C。

4、Bulhmann 模型的非参数估计

假设第 i 个投保人的损失为

$$X_i = (X_{i1}, X_{i2}, ..., X_{in})', \quad i = 1, ..., r$$

给定$\Theta_i = \theta_i$， $X_{ij}, j = 1, ..., n_i$是相互独立的，且

$$\mu(\theta_i) = E(X_{ij} | \Theta = \theta_i)$$

$$v(\theta_i) = Var(X_{ij} | \Theta = \theta_i)$$

另外还假设两个不同的投保人之间损失是相互独立的，即如果$i \neq l$，则 X_{ij}和X_{lt}是相互独立，求 Buhlmann 信度估计中 $a, v,$ 和 μ 的无偏估计。

Risk	Time Period				Risk Parameter
	1	**2**	. . .	**N**	
1	X_{11}	X_{12}	. . .	X_{1N}	θ_1
2	X_{21}	X_{22}	. . .	X_{2N}	θ_2
:	:	:	. . .	:	:
R	X_{R1}	X_{R2}	. . .	X_{RN}	θ_R

	Time Period				Risk's Sample Mean $\overline{X}_i$	Risk's Sample Process Variance
Risk	**1**	**2**	**. . .**	***N***		
1	X_{11}	X_{12}	. . .	X_{1N}	$\overline{X}_1 = \left(\frac{1}{N}\right)\sum_{t=1}^{N} X_{1t}$	$\hat{\sigma}_1^2 = \left(\frac{1}{N-1}\right)\sum_{t=1}^{N}(X_{1t} - \overline{X}_1)^2$
2	X_{21}	X_{22}	. . .	X_{2N}	$\overline{X}_2 = \left(\frac{1}{N}\right)\sum_{t=1}^{N} X_{2t}$	$\hat{\sigma}_2^2 = \left(\frac{1}{N-1}\right)\sum_{t=1}^{N}(X_{2t} - \overline{X}_2)^2$
:	:	:	. . .	:	:	:
R	X_{R1}	X_{R2}	. . .	X_{RN}	$\overline{X}_R = \left(\frac{1}{N}\right)\sum_{t=1}^{N} X_{Rt}$	$\hat{\sigma}_R^2 = \left(\frac{1}{N-1}\right)\sum_{t=1}^{N}(X_{Rt} - \overline{X}_R)^2$

（1） μ 的无偏估计

考虑统计量

$$\bar{X} = r^{-1}\sum_{i=1}^{r}\bar{X}_i = (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n}X_{ij}$$

下面证明 $\bar{X}$ 是无偏估计

$$E(\bar{X}) = (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n}E(X_{ij}) = (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n}E(E(X_{ij}\,|\,\Theta_i))$$

$$= (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n}E(\mu(\Theta_i)) = (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n}\mu = \mu$$

（2）v 的无偏估计
考虑统计量

$$\hat{v}_i = \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2$$

由于给定$\Theta_i = \theta_i$， $X_{ij}, j = 1,...,n$是相互独立的， $v(\theta_i) = Var(X_{ij} \mid \Theta = \theta_i)$，
因此给定$\Theta_i = \theta_i$， $\hat{v}_i$是$v(\theta_i)$的无偏估计，且

$$E(\hat{v}_i) = E(E(\hat{v}_i \mid \Theta_i)) = E(v(\Theta_i)) = v$$

因此$\hat{v}_i$是v的无偏估计。于是统计量

$$\hat{v} = \frac{1}{r}\sum_{i=1}^{r}\hat{v}_i = \frac{1}{r(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2$$

是v(EPV)的无偏估计。

（3）a 的无偏估计

由于

$$E(\bar{X}_i|\Theta_i=\theta_i)=n^{-1}\sum_{j=1}^{n}E(X_{ij}|\Theta_i=\theta_i)=n^{-1}\sum_{j=1}^{n}\mu(\theta_i)=\mu(\theta_i)$$

因此

$$E(\bar{X}_i)=E(E(\bar{X}_i|\Theta_i))=E(\mu(\Theta_i))=\mu$$

$$\begin{aligned}Var(\bar{X}_i)&=Var(E(\bar{X}_i|\Theta_i))+E(Var(\bar{X}_i|\Theta_i))\\&=Var(\mu(\Theta_i))+E(v(\Theta_i)/n))=a+v/n\end{aligned}$$

因此， $\bar{X}_1,...,\bar{X}_r$ 相互独立，具有相同的均值 μ 和方差 $a+v/n$ 。从而统计量

$$(r-1)^{-1}\sum_{i=1}^{r}(\bar{X}_i-\bar{X})^2$$ ，其中 $\bar{X}=r^{-1}\sum_{i=1}^{r}\bar{X}_i$ 是样本均值

是 $a+v/n$ 的无偏估计。于是 a 的无偏估计等于

$$\hat{VHM} = \hat{a} = \frac{1}{r-1}\sum_{i=1}^{r}(\bar{X}_i - \bar{X})^2 - \frac{\hat{v}}{n}$$

$$= \frac{1}{r-1}\sum_{i=1}^{r}(\bar{X}_i - \bar{X})^2 - \frac{1}{rn(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2$$

$$\hat{k} = \frac{E\hat{P}V}{V\hat{H}M} = \frac{\hat{v}}{\hat{a}}, \qquad \hat{z} = \frac{n}{n+k}, \qquad \hat{\mu}(\theta_i) = \hat{z}\bar{X}_i + (1-\hat{z})\cdot\bar{X}$$

直观解释

EPV（$\hat{v}$） ：组内均方误差

VHM（$\hat{a}$）中的第一项：组间均方误差

当组间均方误差$\hat{v}$相对组内均方误差较小时，即$\hat{a}$小于$\hat{v}$，k 较大，$\hat{Z}$将接近于 0，$\bar{X}_i$信度较小。

计算出来的$\hat{a}$可能小于 0，这时令$\hat{a}=\hat{Z}=0$。这种情况等价于方差分析中 F 检验统计量小于 1，导致无法拒绝等均值假设。

例：假设有两个投保人，他们在 3 年内的理赔额分别为$\vec{x}_1=(3,5,7)^T$，$\vec{x}_2=(6,12,9)^T$，求他们各自的 Buhlmann 信度估计。

解：

$$\left.\begin{array}{l}\bar{X}_1=(3+5+7)/3=5\\ \bar{X}_2=(6+12+9)/3=9\end{array}\right. \Rightarrow \bar{X}=(5+9)/2=7$$

$$\hat{v}_1=[(3-5)^2+(5-5)^2+(7-5)^2]/2=4$$

$$\hat{v}_2=?$$

$$E\hat{P}V=\hat{v}=(4+9)/2=13/2$$

$$V\hat{H}M=\hat{a}=\frac{1}{r-1}\sum_{i=1}^{r}(\bar{X}_i-\bar{X})^2-\frac{\hat{v}}{n}$$

$$=[(5-7)^2+(9-7)^2]-\frac{1}{3}\times\frac{13}{2}=\frac{35}{6}$$

$$\hat{k}=\frac{E\hat{P}V}{V\hat{H}M}=\frac{\hat{v}}{\hat{a}}=\frac{39}{35},\ \hat{z}=\frac{3}{3+k}=\frac{35}{48}$$

投保人 1 的信度估计 $\hat{z}\bar{X}_1+(1-z)\hat{\mu}=\left(\frac{35}{48}\right)\times 5+\frac{13}{48}\times 7=\frac{133}{24}$

投保人 2 的信度估计 $\hat{z}\bar{X}_2+(1-z)\hat{\mu}=\left(\frac{35}{48}\right)\times 9+\frac{13}{48}\times 7=\frac{203}{24}$

Example Two risks were selected at random from a population. Risk 1 had 0 claims in year one, 3 claims in year two, and 0 claims in year three: (0,3,0). The claims by year for Risk 2 were (2,1,2). In this case, R = 2 and N = 3. 从群体中随机选择两种风险。其中风险1在第一年有0个索赔，在第二年有3个索赔，在第三年有0个索赔：（0,3,0）。 风险2的年度索赔为（2,1,2）。在这种情况下，r= 2且n = 3。

$\bar{x}_1 = (0+3+0)/3 = 1$, $\bar{x}_2 = (2+1+2)/3 = 5/3$, and $\bar{x} = [1+(5/3)]/2 = 4/3$

$$\hat{\sigma}_1^2 = [(0-1)^2 + (3-1)^2 + (0-1)^2]/(3-1) = 3$$

$$\hat{\sigma}_2^2 = [(2-5/3)^2 + (1-5/3)^2 + (2-5/3)^2]/(3-1) = 1/3$$

$$E\hat{P}V = (\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2 = [3+(1/3)]/2 = 5/3$$

$$V\hat{H}M = \left(\frac{1}{2-1}\right)\left\{[1-(4/3)]^2 + [(5/3)-(4/3)]^2\right\} - (5/3)/3 = -1/3$$

The sample means for the two risks are close relative to the sizes of the sample process variances. The calculated $\hat{VHM}$ is negative, so a value of zero will be assumed. The hypothetical means are indistinguishable. This implies a credibility factor $\hat{Z} = 0$.

38. An insurer has data on losses for four policyholders for 7 years. The loss from the ith policyholder for year j is X_{ij} .

You are given:

$$\sum_{i=1}^{4}\sum_{j=1}^{7}(X_{ij} - \bar{X}_i)^2 = 33.60, \quad \sum_{i=1}^{4}(\bar{X}_i - \bar{X})^2 = 3.30$$

Using nonparametric empirical Bayes estimation, calculate the Bühlmann credibility factor for an individual policyholder.

For this problem, $r = 4$ and $n = 7$. Then,

$$\hat{v} = \frac{33.60}{4(7-1)} = 1.4, \ \hat{a} = \frac{3.3}{4-1} - \frac{1.4}{7} = 0.9.$$

Then,

$$k = \frac{1.4}{0.9} = \frac{14}{9}, \ Z = \frac{7}{7+(14/9)} = \frac{63}{77} = 0.82.$$

5、Bühlmann model的半参数估计

例：某汽车险的一个被保险人的理赔次数服从 Poisson 分布，参数为θ，θ随被保险人变化而变化。得到去年的理赔记录如下：

理赔次数	被保险人数
0	1563
1	271
2	32

3	7
4	2
合计	1875

对每个被保险人，求理赔次数的信度估计。

解：r＝1875，n_i＝1，m_{ij}＝1，

对投保人 i，（$i=1,\ldots,1875$），$X_{i1}|\Theta_i=\theta_i$ 服从 Poisson 分布，均值为 θ_i，方差 θ_i，无条件均值为

$$\mu=E(E(X_i|\Theta_i))=E(\Theta_i)$$

V=EPV=E(EPV=E($Var(X_{i1}|\Theta_i))=E(\Theta_i)$)

从而推出，$\mu=v$，因此可以用 μ 的无偏估计来代替 v 的无偏估计

$$\bar{X} = \frac{1}{1875}\left(\sum_{i=1}^{1875} X_{i1}\right)$$
$$= \frac{0(1563)+1\times 271+2\times 32+3\times 7+4\times 2}{1875} = 0.194$$

所以 EPV=0.194

$$Var(X_{i1}) = Var(E(X_{i1}\,|\,\Theta_i)) + E(Var(X_{i1}\,|\,\Theta_i))$$
$$= Var(\mu(\Theta_i)) + E(v(\Theta_i)) = a+v = a+\mu$$

$a+v$ 的无偏估计是样本方差

$$\frac{\sum_{i=1}^{1875}(X_{i1}-\bar{X})^2}{1874} = \frac{1563(0-0.194)^2+271(1-0.194)^2+\ldots+2(4-0.194)^2}{1874}$$
$$= 0.226$$

因此

$a = 0.226 - 0.194 = 0.032$， $k = 0.194/0.032 = 6.06$

$$z = 1/(1+6.06) = 0.14$$

所以一个被保险人的理赔次数的信度估计为

$$0.14X_{i1} + 0.86 \times 0.194$$

197. You are given:

(i) During a 2-year period, 100 policies had the following claims experience:

Total Claims in Years 1 and 2	Number of Policies
0	50
1	30
2	15
3	4
4	1

(ii) The number of claims per year follows a Poisson distribution.

(iii) Each policyholder was insured for the entire 2-year period.

A randomly selected policyholder had one claim over the 2-year period.

Using semiparametric empirical Bayes estimation, calculate the Bühlmann estimate for the number of claims in Year 3 for the same policyholder.

$$\hat{v} = \bar{x} = \frac{30+30+12+4}{100} = 0.76.$$

$$\hat{a} = \frac{50(0-0.76)^2 + 30(1-0.76)^2 + 15(2-0.76)^2 + 4(3-0.76)^2 + 1(4-0.76)^2}{99} - 0.76$$

$$= 0.090909,$$

$$\hat{k} = \frac{0.76}{0.090909} = 8.36, \quad \hat{Z} = \frac{1}{1+8.36} = 0.10684,$$

$$P = 0.10684(1) + 0.89316(0.76) = 0.78564.$$

The above analysis was based on the distribution of total claims for two years. Thus 0.78564 is the expected number of claims for the next two years. For the next one year the expected number is 0.78564/2 = 0.39282.

三、Buhlmann－Straub 模型

假定：在给定$\Theta=\theta$条件下，$X_1,\cdots,X_n$是相互独立的，具有相同的条件均值

$$\mu(\theta)=E(X_j\mid\Theta) \tag{11}$$

但条件方差不相同

$$\mathrm{var}(X_j\mid\Theta=\theta)=v(\theta)/m_j \tag{12}$$

符号：我们定义

$$\mu=E(\mu(\Theta)),$$
$$v=E(v(\Theta)),$$
$$a=\mathrm{var}(\mu(\Theta))$$

背景：

m_j 可以看作

第 j 年的风险单位数

第 j 年的某风险子集内投保人的个数。

第 j 年的保费收入

X_j 可以看作

第 j 年的每风险单位的平均赔付

第 j 年的某风险子集内每个投保人的平均赔付

第 j 年的保费收入的赔付率

例如

某风险子集内有 10 个投保人在某保单组合中；1998 年每个风险子集内 12 个投保人，1999 年有 15 个投保人。它在这三年的总索赔次数为 18（1997），20（1998）和 27（1999）。假设 2000 年该风险子集共有 20 个投保人在保单组合中，使用 Bühlmann－Straub 模型求 2000 年该风险子集的总索赔次数的信度估计。

一般来说

- 如果估计索赔频率或纯保费，那么 m 就是风险单位数（exposure）。
- 如果估计索赔强度，那么 m 是索赔次数。

数学解释

若令

$$X_j = \frac{1}{m_j}\sum_{i=1}^{m_j} Y_i^{(j)}$$

其中$Y_i^{(j)}$表示第 j 年的第 i 个风险单位（投保人）的赔付，固定 j，$Y_1^{(j)}, Y_2^{(j)}, \cdots$ 是独立同分布的，$E(Y_i^{(j)} \mid \Theta = \theta) = \mu(\theta)$，$Var(Y_i^{(j)} \mid \Theta = \theta) = v(\theta)$，则

$$E(X_j \mid \Theta) = \mu(\theta)$$

$$\operatorname{var}(X_j \mid \Theta = \theta) = v(\theta)/m_j$$

Buhlmann－Straub 信度估计

下面求解方程组（3）和（8），即

$$E(\mu_{n+1}(\Theta)) = E(X_{n+1}) = \alpha_0 + \Sigma_{i=1}^{n}\alpha_i E(X_i) \tag{3}$$

$$\operatorname{cov}(X_i, X_{n+1}) = \sum_{j=1}^{n} \alpha_j \operatorname{cov}(X_i, X_j), \quad i = 1, 2, \cdots, n \tag{8}$$

因为$E(X_j) = \mu$，由公式（11），方程（3）变为

$$\mu = \widehat{\alpha}_0 + \sum_{j=1}^{n} \widehat{\alpha}_j \mu$$

于是推出

$$\sum_{j=1}^{n} \widehat{\alpha}_j = 1 - \widehat{\alpha}_0 / \mu \tag{13}$$

又由于

$$
\begin{aligned}
\operatorname{cov}(X_i, X_j) &= E(X_i X_j) - E(X_i)E(X_j) \\
&= E(E(X_i X_j \mid \Theta)) - \mu^2 \\
&= E\{[E(X_i \mid \Theta)E(X_j \mid \Theta)] - \{E[\mu(\Theta)]\}^2 \\
&= E\{[\mu(\Theta)]^2\} - \{E[\mu(\Theta)]\}^2 \\
&= \operatorname{var}(\mu(\Theta)) = a
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{var}(X_j) &= E(\operatorname{var}(X_j \mid \Theta)) + \operatorname{var}(E(X_j \mid \Theta)) \\
&= E(v(\Theta)/m_j) + \operatorname{var}(\mu(\Theta)) \\
&= v/m_j + a
\end{aligned}
\tag{14}
$$

对于 $i = 1, \cdots, n$，方程（8）变为

$$a=\sum_{\substack{j=1\\ j\neq i}}^{n}\widehat{\alpha}_j a+\widehat{\alpha}_i(a+v/m_i)=\sum_{j=1}^{n}\widehat{\alpha}_j a+v\widehat{\alpha}_i/m_i$$

利用（13），这个公式可改写为如下形式

$$\widehat{\alpha}_i=\frac{a}{v}m_i(1-\sum_{j=1}^{n}\widehat{\alpha}_j)=\frac{a}{v}\frac{\widehat{\alpha}_0}{\mu}m_i,\quad i=1,\cdots,n \tag{15}$$

于是利用公式（14）和（15）得到

$$1-\frac{\widehat{\alpha}_0}{\mu}=\sum_{j=1}^{n}\widehat{\alpha}_j=\sum_{i=1}^{n}\widehat{\alpha}_i=\frac{a}{v}\frac{\widehat{\alpha}_0}{\mu}\sum_{i=1}^{n}m_i=\frac{a}{v}\frac{\widehat{\alpha}_0}{\mu}m$$

其中$m=m_1+\cdots+m_n$。由此解得

$$\widehat{\alpha}_0 = \frac{\mu}{1+\dfrac{am}{v}} = \frac{v/a}{m+v/a}\mu$$

$$\widehat{\alpha}_j = \frac{a\widehat{\alpha}_0}{\mu v}\cdot m_j = \frac{m_j}{m+v/a}$$

X_{n+1} 的 Buhlmann－Straub 信度估计可以写为

$$\widehat{\alpha}_0 + \sum_{j=1}^{n}\widehat{\alpha}_j X_j = \frac{v/a}{m+v/a}\mu + \sum_{j=1}^{n}\frac{m_j}{m+v/a}X_j$$
$$= z\bar{X} + (1-z)\mu$$

其中

$$z = \frac{m}{m+k}, k = v/a$$

$\bar{X}=\sum_{j=1}^{m}\frac{m_j}{m}X_j$，表示 m 年内每风险单位（投保人）的损失均值。

例：假设投保人可分为若干个风险子集，每个风险子集具有相同的风险参数Θ。假设某风险子集的一个投保人在一年内的总索赔次数服从二项式分布，参数 n＝3，$q=\theta$，其中参数Θ是一个随机变量，分布密度为$\pi(\theta)=6\theta(1-\theta)$，$0<\theta<1$。随机记录一个风险子集，假设 1997 年该风险子集内有 10 个投保人在某保单组合中；1998 年每个风险子集内 12 个投保人，1999 年有 15 个投保人。它在这三年的总索赔次数为 18（1997），20（1998）和 27（1999）。假设 2000 年该风险子集共有 20 个投保人在保单组合中，使用 Bühlmann－Straub 模型求 2000 年该风险

子集的总索赔次数的信度估计。

解：设 W 表示一个投保人在一年内的总索赔次数，则

$$\mu(\theta)=E(W\mid\Theta=\theta)=3\theta$$

$$v(\theta)=\mathrm{var}(W\mid\Theta=\theta)=3\theta(1-\theta)$$

又由 Θ 的分布，可以计算出

$$\mu=E(X)=E[\mu(\Theta)]=E[3\Theta]=\int_0^1 3\theta\cdot6\theta(1-\theta)d\theta=\frac{3}{2}$$

$$v=E[v(\Theta)]=E[3\Theta(1-\Theta)]=\int_0^1 3\theta(1-\theta)\cdot6\theta(1-\theta)d\theta=\frac{3}{5}$$

$$
\begin{aligned}
a &= \mathrm{var}[\mu(\Theta)] = \mathrm{var}[3(\Theta)] \\
&= 9\,\mathrm{var}(\Theta) = 9[E(\Theta^2)] - [E(\Theta)]^2 \\
&= 9[\int_0^1 \theta^2 6\theta(1-\theta)d\theta - (\frac{1}{2})^2] = \frac{9}{20}
\end{aligned}
$$

在 1997 年，由于每个风险子集有 10 个投保人，则随机选取一个风险子集的人均总索赔次数为

$$
X_1 = \frac{W_{1,1} + \cdots + W_{1,10}}{10} = \frac{18}{10}, m_1 = 10
$$

相应的，在 1998 和 1999 年有

$$
X_2 = \frac{20}{12}, m_2 = 12 \quad X_3 = \frac{27}{15}, m_3 = 15
$$

应用 Buhlmann－Straub 模型， $k = v/a = 4/3$，信度因子为

$$z=\frac{m}{m+k}=\frac{10+12+15}{10+12+15+\frac{4}{3}}=0.965$$

$$\bar{X}=\sum_{i=1}^{n}\frac{m_i}{m}X_i=\frac{10}{37}\times\frac{18}{10}+\frac{12}{37}\times\frac{20}{12}+\frac{15}{37}\times\frac{27}{15}=1.757$$

因此，2000 年随机选取一个风险子集的每个投保人信度保费为

$$z\bar{X}+(1-z)\mu=0.965\times1.757+0.035\times1.5=1.748$$

该风险子集的信度保费为 20×1.748＝34.96。

课堂练习

1104－9 Members of three classes of insureds can have 0, 1 or 2 claims, with the following probabilities:

Class	Number of Claims		
	0	1	2
I	0.9	0.0	0.1
II	0.8	0.1	0.1
III	0.7	0.2	0.1

A class is chosen at random, and varying numbers of insureds from that class are observed over 2 years, as shown below:

Year	Number of Insureds	Number of Claims
1	20	7
2	30	10

Determine the Bühlmann-Straub credibility estimate of the number of claims in Year 3 for 35 insureds from the same class.

(A) 10.6 (B) 10.9 (C) 11.1 (D) 11.4 (E) 11.6

设 X 表示一个投保人理赔数，$\theta_1,\theta_2,\theta_3$分别表示三种类型被保险人的风险，则由题意知

$$\mu(\theta_1)=0.9\times 0+0.1\times 2=0.2$$

$$\mu(\theta_2)=0.8\times 0+0.1\times 1+0.1\times 2=0.3$$

$$\mu(\theta_3)=0.7\times 0+0.2\times 1+0.1\times 2=0.4$$

$$\mu=\frac{1}{3}(0.2+0.3+0.4)=0.3$$

$$v(\theta_1)=0.9\times 0+0.1\times 4-0.2^2=0.36$$

$$v(\theta_2)=0.8\times 0+0.1\times 1+0.1\times 4-0.3^2=0.41$$

$$v(\theta_3)=0.7\times 0+0.2\times 1+0.1\times 4-0.4^2=0.44$$

$$v=(1/3)(0.36+0.41+0.44)=0.403333$$

$$a = (1/3)(0.2^2 + 0.3^2 + 0.4^2) - 0.3^2 = 0.006667$$

$$k = 0.403333 / 0.006667 = 60.5$$

$$m = m_1 + m_2 = 50$$

$$z = \frac{m}{m+k} = \frac{50}{50+60.5} = 0.45249$$

第 3 年一个被保险人理赔数的 Buhlmann－Straub 信度估计为

$$0.4529(17/50) + 0.54751(0.3) = 0.31810$$

35 个被保险人理赔数的 Buhlmann－Straub 信度估计 35×0.31810=11.31

Buhlmann－Straub 模型的非参数估计

记号和假设

（1）损失

假设有 r 个投保人，

X_{ij} 表示第 i 个投保人，第 j 年的每风险单位的损失

$X_i = (X_{i1}, X_{i2}, ..., X_{in_i})'$，表示第 i 个投保人的经验损失，$n_i$ 表示第 i 个投保人的观测值的个数

$\mathbf{X} = (X_1, X_2, ..., X_r)$，表示全部投保人的损失，r 表示投保人的个数。

（2）风险参数

Θ_i 表示第 i 个投保人的风险参数，$\Theta_1, ..., \Theta_r$ 独立同分布，

$\pi(\theta_i)$，$i = 1, ..., r$ 表示风险参数 Θ_i 的分布

给定 Θ_i，$X_{ij}, j = 1, ..., n_i$ 是相互独立的，分布为 $f_{X_{ij}|\Theta}(x_i | \theta_i)$

（3）风险单位数

m_{ij} 表示第 i 个投保人(团体)第 j 年的风险单位数，一般约定为 1

$m_i = \sum_{j=1}^{n_i} m_{ij}$ ，第 i 个投保人(团体)的风险单位数总和

$m = \sum_{i=1}^{r} m_i$ ，所有投保人(团体)的风险单位数的总和

（4）平均损失

$\bar{X}_i = \frac{1}{m_i}\sum_{j=1}^{n_i} m_{ij} X_{ij}$ 表示第 i 个投保人(团体)的每风险单位的平均损失

$\bar{X} = \frac{1}{m}\sum_{i=1}^{r} m_i \bar{X}_i = \frac{1}{m}\sum_{i=1}^{r}\sum_{j=1}^{n_i} m_{ij} X_{ij}$ 每风险单位的平均损失

数据形式

Risk	Periods of Experience					
1	X_{11}	X_{12}	$\cdots$	X_{1N_1}		
	m_{11}	m_{12}	$\cdots$	m_{1N_1}		
2	X_{21}	X_{22}	$\cdots$	$\cdots$	$\cdots$	X_{2N_2}
	m_{21}	m_{22}	$\cdots$	$\cdots$	$\cdots$	m_{2N_2}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\vdots$
R		X_{R1}	X_{R2}	$\cdots$	$\cdots$	X_{RN_R}
		m_{R1}	m_{R2}	$\cdots$	$\cdots$	m_{RN_R}

统计量

Risk	Periods	Total Exposure	Sample Mean	Sample Estimate for σ^2
1	N_1	$m_1=\sum_{t=1}^{N_1} m_{1t}$	$\bar{X}_1=\left(\frac{1}{m_1}\right)\sum_{t=1}^{N_1} m_{1t}X_{1t}$	$\hat{\sigma}_1^2=\left(\frac{1}{N_1-1}\right)\sum_{t=1}^{N_1} m_{1t}(X_{1t}-\bar{X}_1)^2$
2	N_2	$m_2=\sum_{t=1}^{N_2} m_{2t}$	$\bar{X}_2=\left(\frac{1}{m_2}\right)\sum_{t=1}^{N_2} m_{2t}X_{2t}$	$\hat{\sigma}_2^2=\left(\frac{1}{N_2-1}\right)\sum_{t=1}^{N_2} m_{2t}(X_{2t}-\bar{X}_2)^2$
:	:	:	:	:
R	N_R	$m_R=\sum_{t=1}^{N_R} m_{Rt}$	$\bar{X}_R=\left(\frac{1}{m_R}\right)\sum_{t=1}^{N_R} m_{Rt}X_{Rt}$	$\hat{\sigma}_R^2=\left(\frac{1}{N_R-1}\right)\sum_{t=1}^{N_R} m_{Rt}(X_{Rt}-\bar{X}_R)^2$
Total		$m=\sum_{i=1}^{R} m_i$	$\bar{X}=\sum_{i=1}^{R} m_i\bar{X}_i/m$	$E\hat{P}V=\sum_{i=1}^{R}(N_i-1)\hat{\sigma}_i^2/\left(\sum_{i=1}^{R}(N_i-1)\right)$

假设

$$E(X_{ij}\,|\,\Theta_i=\theta_i)=\mu(\theta_i)$$

$$\mathrm{var}(X_{ij}\,|\,\Theta_i=\theta_i)=v(\theta_i)/m_{ij}\,,\quad j=1,\ldots,n_i$$

下面求 a，v 和 μ 的无偏估计。

（1）μ 的无偏估计

考虑统计量

$$\bar{X}=\frac{1}{m}\sum_{i=1}^{r}m_i\bar{X}_i=\frac{1}{m}\sum_{i=1}^{r}\sum_{j=1}^{n_i}m_{ij}X_{ij}$$

由于 $E(X_{ij})=E(E(X_{ij}\mid\Theta_i))=E(\mu(\Theta_i))=\mu$，因此

$$E(\bar{X}_i\mid\Theta_i)=\sum_{j=1}^{n_i}\frac{m_{ij}}{m_i}E(X_{ij}\mid\Theta_i)=\sum_{j=1}^{n_i}\frac{m_{ij}}{m_i}\mu(\Theta_i)=\mu(\Theta_i)$$

从而

$$E(\bar{X}_i)=E(E(\bar{X}_i\mid\Theta_i))=E(\mu(\Theta_i))=\mu$$

$$E(\bar{X})=\frac{1}{m}\sum_{i=1}^{r}m_iE(\bar{X}_i)=\frac{1}{m}\sum_{i=1}^{r}m_i\mu=\mu$$

所以 $\bar{X}$ 是 μ 的无偏估计。

（2）v的无偏估计

###自行阅读

考虑统计量

$$\hat{v}_i = \frac{\sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X}_i)^2}{n_i - 1}, \quad i = 1, \ldots, r$$

由于

$$
\begin{aligned}
&\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\bar{X}_i)^2 \\
&=\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\mu(\Theta_i)-(\bar{X}_i-\mu(\Theta_i)))^2 \\
&=\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\mu(\Theta_i))^2-2\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\mu(\Theta_i))(\bar{X}_i-\mu(\Theta_i))+\sum_{j=1}^{n_i} m_{ij}(\bar{X}_i-\mu(\Theta_i))^2 \\
&=\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\mu(\Theta_i))^2-m_i(\bar{X}_i-\mu(\Theta_i))^2
\end{aligned}
$$

因此

$$E\left[\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\bar{X}_i)^2 \mid \Theta_i\right]$$

$$=\sum_{j=1}^{n_i} m_{ij}E\left[(X_{ij}-\mu(\Theta_i))^2 \mid \Theta_i\right]-m_iE\left[(\bar{X}_i-\mu(\Theta_i))^2 \mid \Theta_i\right]$$

$$=\sum_{j=1}^{n_i} m_{ij}Var(X_{ij} \mid \Theta_i)-m_iVar(\bar{X}_i \mid \Theta_i)$$

$$=\sum_{j=1}^{n_i} m_{ij}v(\Theta_i)/m_{ij}-v(\Theta_i)=v(\Theta_i)(n_i-1) \tag{*}$$

因此

$$E(\hat{v}_i \mid \Theta_i)=v(\Theta_i)$$

$$E(\hat{v}_i)=E(E(\hat{v}_i \mid \Theta_i))=E(v(\Theta_i))=v$$

所以 $\hat{v}_i$ 是 v 的无偏估计。

令 $\hat{v}=\Sigma_{i=1}^{r}w_i\hat{v}_i$，其中 w_i 表示权重，$0<w_i<1, \Sigma_{i=1}^{n}w_i=1$，则 $\hat{v}$ 是 v 的无偏估计。这里我们选择 $w_i=(n_i-1)/\Sigma_{i=1}^{n}(n_i-1)$，则

#####

$$E\hat{P}V=\hat{v}=\frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i}m_{ij}(X_{ij}-\bar{X}_i)^2}{\sum_{i=1}^{r}(n_i-1)}$$

是 v 的无偏估计

（3）a 的无偏估计

##由于固定 i，$X_{i1},X_{i2},\ldots,X_{in_i}$ 是关于 Θ_i 是条件独立的，因此

$$Var(\bar{X}_i \mid \Theta_i) = Var(\frac{1}{m_i}\sum_{j=1}^{n_i} m_{ij} X_{ij} \mid \Theta_i)$$

$$= \sum_{j=1}^{n_i}\left(\frac{m_{ij}}{m_i}\right)^2 Var(X_{ij} \mid \Theta_i)$$

$$= \sum_{j=1}^{n_i}\left(\frac{m_{ij}}{m_i}\right)^2 \frac{v(\Theta_i)}{m_{ij}}$$

$$= \frac{v(\Theta_i)}{m_i^2}\sum_{j=1}^{n_i} m_{ij} = \frac{v(\Theta_i)}{m_i}$$

$$Var(\bar{X}_i) = Var(E(\bar{X}_i \mid \Theta_i)) + E(Var(\bar{X}_i \mid \Theta_i))$$

$$= Var(\mu(\Theta_i)) + E(v(\Theta_i / m_i))$$

$$= a + v / m_i$$

则 $\bar{X}_1, ..., \bar{X}_r$ 相互独立，具有相同的均值 μ 和方差 $a + v / m_i$ 。

令

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{r} m_i \bar{X}_i$$

利用(*)类似的方法，可以证明

$$E(\sum_{i=1}^{r} m_i (\bar{X}_i - \bar{X})^2) = a\left(m - m^{-1}\sum_{i=1}^{r} m_i^2 \right) + v(r-1)$$

###

a 的无偏统计量为

$$V\hat{H}M = \hat{a} = \left(m - m^{-1}\sum_{i=1}^{r} m_i^2 \right)^{-1} \left(\sum_{i=1}^{r} m_i (\bar{X}_i - \bar{X})^2 - \hat{v}(r-1) \right)$$

17. You are given the following commercial automobile policy experience:

	Company	Year 1	Year 2	Year 3
Losses Number of Automobiles	I	50,000 100	50,000 200	? ?
Losses Number of Automobiles	II	? ?	150,000 500	150,000 300
Losses Number of Automobiles	III	150,000 50	? ?	150,000 150

Determine the nonparametric empirical Bayes credibility factor, *Z*, for Company III.

解答

The subscripts denote the three companies.

$$x_{I1}=\frac{50,000}{100}=500,\quad x_{I2}=\frac{50,000}{200}=250,\quad x_{II1}=\frac{150,000}{500}=300$$

$$x_{II2}=\frac{150,000}{300}=500,\quad x_{III1}=\frac{150,000}{50}=3,000,\quad x_{III2}=\frac{150,000}{150}=1,000$$

$$\overline{x}_I=\frac{100,000}{300}=333.33,\quad \overline{x}_{II}=\frac{300,000}{800}=375,\quad \overline{x}_{III}=\frac{300,000}{200}=1,500,\quad \overline{x}=\frac{700,000}{1,300}=538.46$$

$$\hat{v} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X}_i)^2}{\sum_{i=1}^{r}(n_i - 1)}$$

$$\hat{v} = \frac{\begin{array}{c}100(500-333.33)^2 + 200(250-333.33)^2 + 500(300-375)^2 + 300(500-375)^2 \\ + 50(3,000-1,500)^2 + 150(1,000-1,500)^2\end{array}}{(2-1)+(2-1)+(2-1)}$$

$$= 53,888,888.89$$

$$\hat{a} = \left(m - m^{-1}\sum_{i=1}^{r} m_i^2\right)^{-1}\left(\sum_{i=1}^{r} m_i(\bar{X}_i - \bar{X})^2 - \hat{v}(r-1)\right),$$ $m_1=300, m_2=800, m_3=200$

$$\hat{a} = \frac{300(333.33-538.46)^2 + 800(375-538.46)^2 + 200(1,500-538.46)^2 - 53,888,888.89(3-1)}{1,300 - \frac{300^2 + 800^2 + 200^2}{1,300}}$$

$$= 157,035.60$$

$$k=\frac{53,888,888.89}{157,035.60}=343.1635,\ Z=\frac{200}{200+343.1635}=0.3682$$

几种特殊情况的讨论：

$\hat{a}<0$

$n_i=1$

$r=1$，

μ 已知

a，v 和 μ 的无偏估计以及更深入的讨论，请阅读 loss model

平衡调整

值得注意的是，虽然所有参数的估计值都是无偏的，但这并不一定就说明经验信度估计值也是无偏的。事实上我们可以求出所有被保险人的经验总损失 $TL=\sum_{i=1}^{r} m_i \bar{X}_i$ 。依照上述模型，如果在过去也按照信度估计计算总损失，平均经验损失等于信度估计经验损失，应该是

$$\bar{X}=\sum_{i=1}^{R}\frac{m_i}{m}[\hat{Z}_i\bar{X}_i+(1-\hat{Z}_i)\bar{X}]\ .$$

但现实情况并不相等。假设存在μ使得

$$TP=\sum_{i=1}^{r} m_i[\hat{Z}_i\bar{X}_i+(1-\hat{Z}_i)\hat{\mu}]$$

$$=\sum_{i=1}^{r} m_i\bar{X}_i-\sum_{i=1}^{r} m_i[(1-\hat{Z}_i)\bar{X}_i-(1-\hat{Z}_i)\hat{\mu}]$$

$$=TL-\sum_{i=1}^{r}\frac{m_i\hat{k}}{m_i+\hat{k}}(\bar{X}_i-\hat{\mu})$$

理想情况是TL等于TP。因此

$$\sum_{i=1}^{r}\frac{m_i\hat{k}}{m_i+\hat{k}}(\bar{X}_i-\hat{\mu})=0$$

即$\sum_{i=1}^{r}\hat{Z}_i(\bar{X}_i-\hat{\mu})=0$。因此

$$\hat{\mu} = \frac{\sum_{i=1}^{r} \hat{Z}_i \bar{X}_i}{\sum_{i=1}^{r} \hat{Z}_i}$$

μ 的估计是个体样本均值的信度加权平均，而非是直接的按风险量加权平均。它的好处在于估计先验均值时考虑到了每个个体样本的可信度，使得总保费等于总损失。

例假设两组被保险人在过去三年的经验数据和第四年的被保险人数如下所示：

被保险人		保单年度			
		1	2	3	4
1	总损失	-	11,000	18,000	-
	团体人数	-	50	80	70
2	总损失	20,000	25,000	24,000	-
	团体人数	100	120	125	100

求（1）对两组被保险人第四年应收取的信度保费。

（2）使用信度加权平均估计μ，两组被保险人各自应缴纳的第四年的信度保费。

（3）如果第一组保险人的第二年总损失数据不是18,000而是15,000，

重新计算上面问题。

解：1）根据数据，可以看出$r=2$，$n_1=2$，$n_2=3$。对于第一组被保险人团体，下标表示哪一年对运算没有影响，方便起见，我们下标采取的年份和表中对应，

$$m_{12}=50，\quad X_{12}=\frac{11000}{50}=220，\quad m_{13}=80，\quad X_{13}=\frac{18000}{80}=225，$$ 以及

$$m_1=m_{12}+m_{13}=130，\quad \bar{X}_1=\frac{11000+18000}{130}=223.08$$

对于第二组被保险人团体$m_{21}=100$，$X_{21}=\dfrac{20000}{100}=200$，$m_{22}=120$，

$$X_{22}=\frac{25000}{120}=\frac{625}{3}\quad,\qquad m_{23}=125\quad,\qquad X_{23}=\frac{24000}{125}=192\quad,$$

$$m_2=m_{21}+m_{22}+m_{23}=345\,,\quad m=m_1+m_2=475$$

以及 $\bar{X}_2=\dfrac{20000+25000+24000}{345}=200$

于是 $\hat{\mu}=\bar{X}=\dfrac{223.08\times130+200\times345}{130+345}=206.31$

$$\hat{\upsilon} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X}_i)^2}{\sum_{i=1}^{r}(n_i - 1)}$$

$$= \frac{\begin{array}{c} 50\times(220-223.08)^2 + 80\times(225-223.08)^2 + 100\times(200-200)^2 \\ +120\times(625/3-200)^2 + 125\times(192-200)^2 \end{array}}{1+2}$$

$$= 5700.855$$

$$\hat{a} = [\sum_{i=1}^{r} m_i (\bar{X}_i - \bar{X})^2 - \hat{\upsilon}(r-1)] / (m - \frac{\sum_{i=1}^{r} m_i^2}{m})$$

$$= \frac{130 \times (223.08 - 206.31)^2 + 345 \times (200 - 206.31)^2 - 5700.855 \times 1}{475 - (130^2 + 345^2) / 475}$$

$$= 236.15$$

因此 $\hat{k} = \frac{\hat{\upsilon}}{\hat{a}} = 24.15$， $\hat{Z}_1 = \frac{m_1}{m_1 + \hat{k}} = 0.84$， $\hat{Z}_2 = \frac{m_2}{m_2 + \hat{k}} = 0.93$

被保险人团体 1 和被保险人团体 2 的个体平均信度保费估计值分别为

$$\hat{Z}_1 \bar{X}_1 + (1 - \hat{Z}_1)\hat{\mu} = 0.84 \times 223.08 + (1 - 0.84) \times 206.31 = 220.45$$

$$\hat{Z}_2\bar{X}_2+(1-\hat{Z}_2)\hat{\mu}=0.93\times200+(1-0.93)\times206.31=200.41$$

于是，被保险人团体 1 和被保险人团体 2 下一年要交的总的信度保费估计分别为 $220.45\times70=15431.5$ 和 $200.41\times100=20041$

（2）从上题中我们知道 $\hat{Z}_1=0.84$， $\hat{Z}_2=0.93$,，得到

$$\hat{\mu}=\frac{\sum_{i=1}^{r}\hat{Z}_i\bar{X}_i}{\sum_{i=1}^{r}\hat{Z}_i}=\frac{0.84\times223.08+0.93\times200}{0.84+0.93}=210.95$$

于是被保险人团体 1 和被保险人团体 2 的个体平均信度保费估计值分别为

$$\hat{Z}_1\bar{X}_1+(1-\hat{Z}_1)\hat{\mu}=0.84\times 223.08+(1-0.84)\times 210.95=221.18$$

$$\hat{Z}_2\bar{X}_2+(1-\hat{Z}_2)\hat{\mu}=0.93\times 200+(1-0.93)\times 206.31=200.72$$

于是，被保险人团体 1 和被保险人团体 2 下一年要交的总的信度保费估计分别为 $221.18\times 70=15482.6$ 和 $200.72\times 100=20072$

（3）此时，$X_{13}=\dfrac{15000}{80}=187.5$，$\bar{X}_1=\dfrac{11000+15000}{130}=200$，其他不变，经计算可得 $\hat{\mu}=200$，$\hat{\upsilon}=16888.57$，

$$\hat{a}=[\sum_{i=1}^{r} m_i(\bar{X}_i-\bar{X})^2-\hat{\upsilon}(r-1)]/(m-\frac{\sum_{i=1}^{r} m_i^{2}}{m})$$

$$=\frac{130\times(200-200)^2+345\times(200-200)^2-16888.57\times 1}{475-(130^2+345^2)/475}$$

$$=-89.43$$

此时 $\hat{a}<0$，因此取 $\hat{Z}_1=\hat{Z}_2=0$，两个团体中的个体平均保费估计均等于 $\hat{\mu}=200$。

两个被保险人团体总的信度保费估计值分别为14000和200000。 半

Bühlmann-Straub model的半参数估计

假设 $f_{X_{ij}|\Theta}(x_{ij}|\theta_i)$ 的分布已知，但 $\pi(\theta_i)$ 的参数未知。根据 $f_{X_{ij}|\Theta}(x_{ij}|\theta_i)$ 的情况具体分析。与 Bühlmann 类似，主要用于**理赔频率**模型。

假设给定 $\Theta_i=\theta_i$，第 i 类保单组有 m_{it} 个风险单位，在第 t 年的保单组的总理赔次数 $m_{it}X_{ij}$ 服从 Poisson 分布，参数为 $m_{it}\theta_i$，则

$$m_{it}\theta_i = E_{X|\Theta}[m_{it}X_{it}|\ \theta_i] = Var_{X|\Theta}[m_{it}X_{it}|\ \theta_i] = m_{it}^2\ Var_{X|\Theta}[X_{it}|\ \theta_i].$$

除以 m_{it}

$$E_{X|\Theta}[X_{it}|\ \theta_i] = m_{it}Var_{X|\Theta}[X_{it}|\ \theta_i] = \theta_i.$$

$$E_{X|\Theta}[X_{it}|\ \theta_i] = \mu(\theta_i) \text{ and } m_{it}Var_{X|\Theta}[X_{it}|\ \theta_i] = \sigma^2(\theta_i).$$

根据泊松分布的性质

$$\mu(\theta_i) = \sigma^2(\theta_i).$$

因此

$$E_\Theta[\mu(\Theta_i)] = E_\Theta[\sigma^2(\Theta_i)] = EPV$$

因此，我们可以用 $\hat{\mu} = \bar{X}$ 去估计 EPV。

Example Carpentry contractors A and B had insurance policies covering pickup trucks. Over a four-year period the following was observed:

		Year			
Insured		***Y***	***Y*+1**	***Y*+2**	***Y*+3**
A	Number of Claims	3	2	2	0
	Insured Vehicles	2	2	2	1
B	Number of Claims	2	1	0	
	Insured Vehicles	4	3	2	

Solution The random variables X_{it} representing claim frequency and the corresponding exposures m_{it} are as follows:

		Year			
Insured		Y	Y+1	Y+2	Y+3
A	Claims per Exposure	$X_{11} = 3/2$	$X_{12} = 1$	$X_{13} = 1$	$X_{14} = 0$
	Exposure = Number of Vehicles	$m_{11} = 2$	$m_{12} = 2$	$m_{13} = 2$	$m_{14} = 1$
B	Claims per Exposure	$X_{21} = 1/2$	$X_{22} = 1/3$	$X_{23} = 0$	
	Exposure = Number of Vehicles	$m_{21} = 4$	$m_{22} = 3$	$m_{23} = 2$	

The claim frequency X_{it} is (Number of Claims)/(Insured Vehicles). The first table shows number of claims that are the values for $(m_{it}X_{it})$.

$m_A = 2 + 2 + 2 + 1 = 7$, $m_B = 4 + 3 + 2 = 9$, and $m = 7 + 9 = 16$.

$\bar{x}_A = (3+2+2+0)/7 = 1$, $\bar{x}_B = (2+1+0)/9 = 1/3$, and
$\bar{x} = [(7)(1)+(9)(1/3)]/16 = 5/8$

$$\hat{VHM} = \frac{7[1-(5/8)]^2 + 9[(1/3)-(5/8)]^2 - (2-1)(5/8)}{16 - \left(\frac{1}{16}\right)(7^2 + 9^2)} = .1429$$

$$\hat{K} = \frac{\hat{EPV}}{\hat{VHM}} = \frac{5/8}{.1429} = 4.3737$$

$$\hat{Z}_{\mathrm{A}} = \frac{m_{\mathrm{A}}}{m_{\mathrm{A}} + \hat{K}} = \frac{7}{7 + 4.3737} = .6155 \quad \rightarrow \quad \hat{\mu}_{\mathrm{A}} = .6155(1) + (1 - .6155)(5/8) = .8558$$

$$\hat{Z}_{\mathrm{B}} = \frac{m_{\mathrm{B}}}{m_{\mathrm{B}} + \hat{K}} = \frac{9}{9 + 4.3737} = .6730 \quad \rightarrow \quad \hat{\mu}_{\mathrm{B}} = .6730(1/3) + (1 - .6730)(5/8) = .4287 \; .$$

257. You are given:

(i) Over a three-year period, the following claim experience was observed for two insureds who own delivery vans:

		Year		
Insured		**1**	**2**	**3**
A	Number of Vehicles	2	2	1
	Number of Claims	1	1	0
B	Number of Vehicles	N/A	3	2
	Number of Claims	N/A	2	3

(ii) The number of claims for each insured each year follows a Poisson distribution.

Calculate the semiparametric empirical Bayes estimate of the claim frequency per vehicle for Insured A in Year 4.

Question #257
Key: C

The estimate of the overall mean, μ, is the sample mean, per vehicle, which is 7/10 = 0.7. With the Poisson assumption, this is also the estimate of v = EPV. The means for the two insureds are 2/5 = 0.4 and 5/5 = 1.0. The estimate of a is the usual non-parametric estimate,

$$\text{VHM} = \hat{a} = \frac{5(0.4-0.7)^2 + 5(1.0-0.7)^2 - (2-1)(0.7)}{10 - \dfrac{1}{10}(25+25)} = 0.04$$

Then, $k = 0.7/0.04 = 17.5$ and so $Z = 5/(5+17.5) = 2/9$. The estimate for insured A is $(2/9)(0.4) + (7/9)(0.7) = 0.6333$.

自行阅读

例：考虑某团体保险的个体理赔发生概率。设θ_i表示保单持有人（团体）i 的理赔发生概率。X_{ij}表示理赔发生，即

$$X_{ij}=\begin{cases}1 & 理赔发生\\ 0 & 理赔不发生\end{cases}$$

m_{ij}表示保单持有人（团体）i 的第 j 年的个体数，则$m_{ij}X_{ij}$表示保单持有人（团体）i 在第 j 年发生的理赔总数，建立信度模型。

解： 显然$m_{ij}X_{ij}$服从参数为m_{ij},θ_i的二项分布，因此

$$E(m_{ij}X_{ij}\,|\,\Theta_i)=m_{ij}\Theta_i$$

$$Var(m_{ij}X_{ij}\,|\,\Theta_i)=m_{ij}\Theta_i(1-\Theta_i)$$

因此

$$\mu(\Theta_i)=\Theta_i, v(\Theta_i)=\Theta_i(1-\Theta_i)$$

于是

$$\mu=E(\Theta_i),$$

$$v=\mu-E(\Theta_i)^2$$

$$a=Var(\Theta_i)=E(\Theta_i)^2-\mu^2=\mu-v-\mu^2$$