《概率论与数理统计》习题

第十八讲　频率学派常见点估计方法

1. 设总体分布列/密度函数如下，$x_1, x_2, \cdots, x_n$ 是样本，试求未知参数的矩估计：

   (a) $P(X=x)=(x-1)\theta^2(1-\theta)^{x-2}, x=2,3,\cdots, 0<\theta<1$.

   (b) $p(x;\theta)=(\theta+1)x^\theta, 0<x<1, \theta>0$.

   解：

   (1)

$$
\begin{aligned}
E(X) &= \sum_{x=2}^{+\infty} x(x-1)\theta^2(1-\theta)^{x-2} \\
&= \theta^2 \sum_{x=2}^{+\infty} x(x-1)(1-\theta)^{x-2} \\
&= \theta^2 \cdot \frac{2}{\theta^3} \\
&= \frac{2}{\theta}.
\end{aligned}
$$

   即 $\theta=\frac{2}{E(X)}$，所以 θ 的矩估计为 $\hat{\theta}=\frac{2}{\bar{x}}$.

   (2)

$$
\begin{aligned}
E(X) &= \int_0^1 x(\theta+1)x^\theta dx \\
&= \frac{\theta+1}{\theta+2}.
\end{aligned}
$$

   即 $\theta=\frac{1-2E(X)}{E(X)-1}$，所以 θ 的矩估计为 $\hat{\theta}=\frac{1-2\bar{x}}{\bar{x}-1}$.

2. 设总体为 $N(\mu,1)$，现对该总体观测 n 次，发现有 k 次观测值为正，使用频率替换方法求 μ 的估计。

   解：

   观测值为正的频率是 $f=\frac{k}{n}$，接下来将计算观测值为正的概率：

   因为总体 $X\sim N(\mu,1)$，所以，

$P(X>0)=1-P(X<0)=1-P(X-\mu<-\mu)=1-\Phi(-\mu)=\Phi(\mu)$.

利用频率替换概率的想法，有 $\Phi(\hat{\mu})=\frac{k}{n}$，

所以，μ 的矩估计为：$\hat{\mu}=\Phi^{-1}(\frac{k}{n})$.

3. 设总体的概率密度函数如下，$x_1,x_2,\cdots,x_n$ 是样本，试求未知参数的最大似然估计：

   (a) $p(x;\theta)=\frac{1}{2\theta}e^{-|x|/\theta},\theta>0$;

   (b) $p(x;\theta_1,\theta_2)=\frac{1}{\theta_2-\theta_1},\theta_1<x<\theta_2$;

解：

(1) 似然函数为：

$$L(\theta)=\left(\frac{1}{2\theta}\right)^n e^{-\frac{\sum_{i=1}^n|x_i|}{\theta}}.$$

对数似然函数为：

$$lnL(\theta)=-nln2\theta-\frac{\sum_{i=1}^n|x_i|}{\theta}.$$

关于 θ 求导：

$$\frac{\partial lnL(\theta)}{\partial\theta}=\frac{-n}{\theta}+\frac{\sum_{i=1}^n|x_i|}{\theta^2}=0.$$

$$\hat{\theta}=\frac{\sum_{i=1}^n|x_i|}{n}.$$

并且，

$$\frac{\partial^2 lnL(\theta)}{\partial\theta^2}|_{\hat{\theta}}=-\frac{n^3}{(\sum_{i=1}^n|x_i|)^2}<0.$$

所以，$\hat{\theta}$ 是 θ 的最大似然估计。

(2) 似然函数为：

$$L(\theta)=\frac{1}{(\theta_2-\theta_1)^n}I_{\theta_1<x_{(1)}<x_{(n)}<\theta_2}.$$

由于，要使 $L(\theta)$ 尽量大，首先，示性函数的取值应为 1，即 $\theta_1<x_{(1)}<x_{(n)}<\theta_2$；其次，要使 $\theta_2-\theta_1$ 要尽量小，所以，θ_1 的最大似然估计为 $x_{(1)}$，θ_2 的最大似然估计为 $x_{(n)}$。

4. 设 $x_1, x_2, \cdots, x_n$ 是来自密度函数为 $p(x;\theta) = e^{-(x-\theta)}, x > \theta$ 的总体的样本。

(a) 求 θ 的最大似然估计 $\hat{\theta}_1$,它是否是相合估计？是否是无偏估计？

(b) 求 θ 的矩估计 $\hat{\theta}_2$，它是否是相合估计？它是否是无偏估计？

解:

(1) 似然函数为:

$$L(\theta) = \prod_{i=1}^{n} \left(e^{-(x_i-\theta)} I_{\{x_i>\theta\}}\right) = e^{-\sum_{i=1}^{n} x_i + n\theta} I_{\{x_1>\theta\}}.$$

可以看出，$L(\theta)$ 在示性函数为 1 的条件下是 θ 的增函数，因此 θ 的最大似然估计为 $\hat{\theta}_1 = x_{(1)}$.

$x_{(1)}$ 的密度函数为 $f(x) = ne^{-n(x-\theta)}$，$x > \theta$，所以，

$$E(\hat{\theta}_1) = \int_{\theta}^{+\infty} xne^{-n(x-\theta)} dx = \frac{1}{n} + \theta.$$

因此，$\hat{\theta}_1$ 不是 θ 的无偏估计，却是渐进无偏估计。

当 $n \to +\infty$ 时，$E(\hat{\theta}_1) \to \theta$，接下来计算 $Var(\hat{\theta}_1)$.

$$E({\hat{\theta}_1}^2) = \int_{\theta}^{+\infty} x^2 ne^{-n(x-\theta)} dx = \frac{2}{n^2} + \frac{2}{n}\theta + \theta^2.$$

$$Var(\hat{\theta}_1) = E({\hat{\theta}_1}^2) - E^2(\hat{\theta}_1) = \frac{1}{n^2} \to 0.$$

说明 $\hat{\theta}_1$ 是 θ 的相合估计。

(2) 矩估计:

$$E(X) = \int_{\theta}^{+\infty} xe^{-(x-\theta)} dx = \theta + 1.$$

所以，θ 的矩估计为:

$$\hat{\theta}_2 = \bar{x} - 1.$$

接下来计算 $Var(\hat{\theta}_2)$.

$$E(X^2) = \int_{\theta}^{+\infty} x^2 e^{-(x-\theta)} dx = \theta^2 + 2\theta + 2.$$

$$Var(X) = E(X^2) - E^2(X) = 1.$$

所以，

$$E(\hat{\theta_2}) = E(\bar{x}) - 1 = \theta.$$

$$Var(\hat{\theta_2}) = \frac{1}{n} Var(X) = \frac{1}{n} \to 0 (n \to +\infty).$$

因此，$\hat{\theta_2}$ 是 θ 的无偏估计和相合估计。

5. 众所周知，双胞胎可分为同卵双胞胎与异卵双胞胎。在一项针对双胞胎的研究中，研究者关心的是一对双胞胎是同卵双胞胎的概率，记为 p，而且研究者也关心一个孩子是男孩的概率，记为 q. 在这项研究中，研究者招募到了 n 对双胞胎（包括龙凤胎），其中 n_1 是两个男孩的双胞胎，n_2 是两个女孩的双胞胎，$n_3 = n - (n_1 + n_2)$ 是龙凤胎（一个男孩一个女孩）。另外，研究者虽然知道不同性别的双胞胎一定不是同卵双胞胎，但并不知道其中哪些相同性别的双胞胎为同卵双胞胎。

   (a) 采用 EM 算法，写出 $\theta = (p, q)$ 的最大似然估计的形式。（提示：构造合适的潜变量，定义似然函数，写出 EM 算法中的 E 步和 M 步）

   (b) (编程题) 在 $n = 1000, n_1 = 432, n_2 = 232, n_3 = 336$ 时，编程计算出 θ 的最大似然估计。

   解：

   (1) 令 MM 代表双胞胎均是男孩子，FF 代表双胞胎均是女孩子，MF 代表龙凤胎，其中，龙凤胎一定是异卵双胞胎，但是性别相同的双胞胎可能为同卵或异卵双胞胎。

   因此，我们可得：$P(MM) = pq + (1-p)q^2$，$P(FF) = p(1-q) + (1-p)(1-q)^2$，$P(MF) = 2(1-p)(1-q)q$.

   令 $\mathbf{x} = (n_1, n_2, n_3)$ 为可观测变量，设 $\mathbf{z} = (n_{11}, n_{12}, n_{21}, n_{22}, n_3)$，其中 n_{11} 是双胞胎均为男孩并来自同卵双胞胎的人数，$n_{12} = n_1 - n_{11}$ 是双胞胎均为男孩并来自异卵双胞胎的人数，同理，n_{21} 是双胞胎均为女孩并来自同卵双胞胎的人数，$n_{22} = n_2 - n_{21}$ 是双胞胎均为女孩并来自异卵双胞胎的人数。

   关于参数 $\theta = (p, q)$ 的似然函数：

$$\begin{aligned} L(\theta, \mathbf{z}) &\propto (pq)^{n_{11}}((1-p)q^2)^{n_{12}}(p(1-q))^{n_{21}}((1-p)(1-q)^2)^{n_{22}}(2(1-p)(1-q)q)^{n_3} \\ &\propto p^{n_{11}+n_{21}}(1-p)^{n_{12}+n_{22}+n_3}q^{n_{11}+2n_{12}+n_3}(1-q)^{n_{21}+2n_{22}+n_3}. \end{aligned}$$

对数似然函数为：

$$
\begin{aligned}
lnL(\theta,\mathbf{z}) \propto\ & (n_{11}+n_{21})lnp+(n_{12}+n_{22}+n_3)ln(1-p) \\
& +(n_{11}+2n_{12}+n_3)lnq+(n_{21}+2n_{22}+n_3)ln(1-q).
\end{aligned}
$$

已知 $\mathbf{x}=(n_1,n_2,n_3)$ 和 $\theta=(p,q)$，可得：

$$
\begin{aligned}
n_{11} &\sim b(n_1, \frac{pq}{pq+(1-p)q^2}). \\
n_{21} &\sim b(n_2, \frac{p(1-q)}{p(1-q)+(1-p)(1-q)^2}).
\end{aligned}
$$

E 步：

$$
\begin{aligned}
Q(\theta|\mathbf{x},\theta^{(k)}) &= E_{\mathbf{z}}lnL(\theta,\mathbf{z}) \\
&\propto (n_{11}^{(k)}+n_{21}^{(k)})lnp+(n_{12}^{(k)}+n_{22}^{(k)}+n_3)ln(1-p) \\
&\quad +(n_{11}^{(k)}+2n_{12}^{(k)}+n_3)lnq+(n_{21}^{(k)}+2n_{22}^{(k)}+n_3)ln(1-q).
\end{aligned}
$$

其中，

$$
\begin{aligned}
n_{11}^{(k)} &= E(n_{11}|\mathbf{x},\theta^{(k)}) = n_1\frac{p^{(k)}q^{(k)}}{p^{(k)}q^{(k)}+(1-p^{(k)})q^{(k)^2}}. \\
n_{12}^{(k)} &= n_1-n_{11}^{(k)} = n_1\frac{(1-p^{(k)})q^{(k)^2}}{p^{(k)}q^{(k)}+(1-p^{(k)})q^{(k)^2}}. \\
n_{21}^{(k)} &= E(n_{21}|\mathbf{x},\theta^{(k)}) = n_2\frac{p^{(k)}(1-q^{(k)})}{p^{(k)}(1-q^{(k)})+(1-p^{(k)})(1-q^{(k)})^2}. \\
n_{22}^{(k)} &= n_2-n_{21}^{(k)} = n_2\frac{(1-p^{(k)})(1-q^{(k)})^2}{p^{(k)}(1-q^{(k)})+(1-p^{(k)})(1-q^{(k)})^2}.
\end{aligned}
$$

M 步：

$$
\begin{aligned}
p^{(k+1)} &= \frac{n_{11}^{(k)}+n_{21}^{(k)}}{n_{11}^{(k)}+n_{12}^{(k)}+n_{21}^{(k)}+n_{22}^{(k)}+n_3} = \frac{n_{11}^{(k)}+n_{21}^{(k)}}{n}. \\
q^{(k+1)} &= \frac{n_{11}^{(k)}+2n_{12}^{(k)}+n_3}{n_{11}^{(k)}+2n_{12}^{(k)}+n_3+n_{21}^{(k)}+2n_{22}^{(k)}+n_3} = \frac{n_{11}^{(k)}+2n_{12}^{(k)}+n_3}{n+n_{12}^{(k)}+n_{22}^{(k)}+n_3}.
\end{aligned}
$$

(2) 当 n= 1000，n_1= 432，n_2= 232，n_3= 336 时，p=0.3，q=0.6.
代码仅供参考

```
In [1]: import random

In [2]: n1 = 432
        n2 = 232
        n3 = 336
        n=1000

In [3]: def _n11(p,q):
            return n1*p*q/(p*q+(1-p)*q*q)

In [4]: def _n12(p,q):
            return n1*(1-p)*q*q/(p*q+(1-p)*q*q)

In [5]: def _n21(p,q):
            return n2*p*(1-q)/(p*(1-q)+(1-p)*(1-q)*(1-q))

In [6]: def _n22(p,q):
            return n2*(1-p)*(1-q)*(1-q)/(p*(1-q)+(1-p)*(1-q)*(1-q))

In [9]: p=random.random()
        q=random.random()
        #设置随机初值
        new_p = p

        new_q = q

        while True: #迭代

            p = new_p

            q = new_q

            n11 = _n11(p,q)
            n12 = _n12(p,q)

            n21 = _n21(p,q)
            n22 = _n22(p,q)

            new_p = (n11+n21)/n
            new_q = (n11+2*n12+n3)/(n+n12+n22+n3)

            if abs(new_p-p)<0.000001 and abs(new_q-q)<0.000001:#是否收敛
                p = new_p
                q = new_q
                break

In [10]: print(p,q)

0.2999972714118231 0.6000000808603186
```