24 数理统计期中

$\mathscr{NULIOUS}$

一.（20 分）设从总体

X	0	1	2
P	$(1-\theta)/3$	$1/3$	$(1+\theta)/3$

（其中 $-1<\theta<1$ 为未知参数）中抽取的一个简单样本 $X_1,\ldots,X_n$

（1）求 θ 的充分统计量，其是否为完全统计量？

（2）求 θ 的矩估计 $\tilde{\theta}$ 和最大似然估计 $\hat{\theta}$ ，是否为无偏估计？

解：（1）设 n_0,n_1,n_2 分别为 $\{x_n\}$ 中为取值为 1,2,3 的个数，则

$$n=n_0+n_1+n_2$$

样本联合密度为

$$f\left(x_1\cdots,x_n;\theta\right)=\left(\frac{1-\theta}{3}\right)^{n_0}\left(\frac{1}{3}\right)^{n_1}\left(\frac{1+\theta}{3}\right)^{n_2}$$

注意这是指数族，改写为

$$f\left(x_1\cdots,x_n;\theta\right)=\left(\frac{1}{3}\right)^{n_1}e^{n_0\ln\frac{1-\theta}{3}+n_2\ln\frac{1+\theta}{3}}$$

即 $(\ln\frac{1-\theta}{3},\ln\frac{1+\theta}{3})$ 作为自然参数，显然其在 $\mathbb{R}^2$ 中有内点

则 $T(\boldsymbol{X})=(n_0,n_2)$ 是充分完全统计量

（2）矩估计：

$$\mathbb{E}[X]=1+\frac{2\theta}{3}$$

令

$$\hat{\theta}_M=\frac{3}{2}(\overline{X}-1)$$

即可，显然无偏。

MLE：

$$\ln f\left(x_1,\cdots,x_n;\theta\right)=n_1\ln\frac{1}{3}+n_0\ln\frac{1-\theta}{3}+n_2\ln\frac{1+\theta}{3}$$

$$\frac{\partial\ln f}{\partial\theta}=\frac{-n_0}{1-\theta}+\frac{n_2}{1+\theta}=0$$

$$\hat{\theta}_{MLE}=\frac{n_2-n_0}{n_2+n_0}$$

用 $n=1\quad\theta=0.5$ 验证知非无偏。

二.（20 分）一个移动通讯公司随机抽取了其 900 个包月客户，计算得知他们一个月平均使用时间是 220 分钟，样本标准差是 90 分钟．假设使用时间服从正态分布．

（1）求包月客户平均使用时间和标准差的 95% 置信区间，并解释所得区间的含义．
（2）如果要求客户平均使用时间的 95% 置信区间的长度不超过 5 分钟，应至少抽取多少个客户？该公司的抽样规模是否满足要求？
解：（1）μ,σ^2 均未知
利用

$$\frac{\sqrt{n}(\overline{X}-\mu)}{S}\to t_{n-1}$$

估计 μ ，进而有置信区间

$$\left[\overline{X}-t_{n-1}\left(\frac{\alpha}{2}\right)\frac{S}{\sqrt{n}},\ \overline{X}+t_{n-1}\left(\frac{\alpha}{2}\right)\frac{S}{\sqrt{n}}\right]$$

代值有

$$[214.12, 225.885]$$

这里因为 $t_n\to N(0,1)$ ，用 $u_{\frac{\alpha}{2}}$ 代替 $t_{n-1}\left(\frac{\alpha}{2}\right)$
再利用

$$\frac{(n-1)S^2}{\sigma^2}\to\chi^2_{(n-1)}$$

σ^2 有置信区间

$$\left[\frac{(n-1)S^2}{\chi^2_{(n-1)}\left(\frac{\alpha}{2}\right)},\frac{(n-1)S^2}{\chi^2_{(n-1)}\left(1-\frac{\alpha}{2}\right)}\right]$$

开方有 σ 的置信区间

$$\left[\sqrt{\frac{(n-1)S^2}{\chi^2_{(n-1)}\left(\frac{\alpha}{2}\right)}},\sqrt{\frac{(n-1)S^2}{\chi^2_{(n-1)}\left(1-\frac{\alpha}{2}\right)}}\right]$$

代值为

$$[88.15, 92.02]$$

区间的含义：使用这个区间充分大次数后，落在置信区间的频率接近于置信系数
（2）即需要

$$2\times u_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\leq 5$$

这里同样因为 $t_n\to N(0,1)$ ，用 $u_{\frac{\alpha}{2}}$ 代替 $t_{n-1}\left(\frac{\alpha}{2}\right)$ 解得

$$n\geq 4979$$

不满足要求

三．（20 分）下表统计了某铁路局 122 个扳道员五年内由于操作失误引起的严重事故情况，其中 r 表示一扳道员某五年内引起严重事故的次数，s 表示扳道员人数．假设扳道员由于操作失误在五年内所引起的严重事故的次数服从 Poisson 分布．求

r	0	1	2	3	4	5	≥ 6
s	44	42	21	9	4	2	0

（1）一个扳道员在五年内未引起严重事故的概率 p 的最小方差无偏估计 $\hat{p}_1$ 和最大似然估计 $\hat{p}_2$ ．
（2）p 的一个（渐近）95% 水平的置信上界．

解：（1）下面记 Poisson 分布的参数为 λ

MLE：样本联合密度为

$$f\left(x_1,\cdots,x_n;\lambda\right)=\frac{e^{-n\lambda}\lambda^{x_1+\cdots+x_n}}{x_1!\cdots x_n!}$$

进而

$$\ln f\left(x_1,\cdots,x_n;\lambda\right)\propto\left(\sum_{i=1}^{n}x_i\right)\ln\lambda-n\lambda$$

则令

$$\frac{\partial\ln f}{\partial\lambda}=\frac{\sum\limits_{i=1}^{n}x_i}{\lambda}-n=0$$

解得

$$\hat{\lambda}_{MLE}=\frac{\sum\limits_{i=1}^{n}x_i}{n}$$

由于 MLE 的不变性有

$$\hat{p_2}=e^{\hat{-\lambda}_{MLE}}=e^{-\frac{\sum\limits_{i=1}^{n}x_i}{n}}=0.325$$

UMVUE：先找一个无偏估计，又由于 Poisson 分布是指数族，充分完全统计量是明显的，即 $T(\boldsymbol{X})=\sum\limits_{i=1}^{n}X_i$，对无偏估计取充分完全统计量的条件期望即得 UMVUE

选取

$$\mathbb{I}_{\{X_1=0\}}$$

作为无偏估计，因为 $\mathbb{E}\left[\mathbb{I}_{\{X_1=0\}}\right]=P(X_1=0)$

下一步取条件期望

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}_{\{X_1=0\}}\mid T(\boldsymbol{X})=t\right]&=P\left(X_1=0\mid T(\boldsymbol{X})=t\right)\\&=\frac{P\left(X_1=0,\sum\limits_{i=2}^{n}X_i=t\right)}{P(T(\boldsymbol{X})=t)}\end{aligned}$$

利用 Poisson 分布的可加性，有

$$\sum_{i=2}^{n}X_i\sim\mathrm{P}((n-1)\lambda)\quad\sum_{i=1}^{n}X_i\sim\mathrm{P}(n\lambda)$$

则

$$\mathbb{E}\left[\mathbb{I}_{\{X_1=0\}}\mid T(\boldsymbol{X})=t\right]=\left(\frac{n-1}{n}\right)^t$$

即 UMVUE 为

$$\hat{p_1}=\left(\frac{n-1}{n}\right)^{\sum\limits_{i=1}^{n}X_i}$$

代值为

$$\hat{p_1}=0.325$$

（2）解一：利用 MLE 的渐进正态性

$$\sqrt{n}\left(\hat{\lambda}_{MLE}-\lambda\right) \to N\left(0, \frac{1}{I(\lambda)}\right)$$

其中 $I(\lambda)$ 为总体（单个样本）的 Fisher 信息量，为

$$I(\lambda) = -\mathbb{E}\left[\frac{\partial^2 \ln f(x,\lambda)}{\partial \lambda^2}\right] = \frac{1}{\lambda}$$

这里

$$f(x,\lambda) = P(X=x,\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$$

得到

$$\sqrt{n}\left(\hat{\lambda}_{MLE}-\lambda\right) \to N\left(0, \lambda\right)$$

法一：因为 $e^{-\lambda}$ 单调递减，因此只需要求出 λ 的置信下界即可

$$\frac{\sqrt{n}\left(\hat{\lambda}_{MLE}-\lambda\right)}{\sqrt{\lambda}} \le u_\alpha$$

处理一：直接解一元二次方程，较复杂
处理二：利用 MLE 做二次近似
用 $\sqrt{\overline{X}}$ 代替 $\sqrt{\lambda}$

$$\frac{\sqrt{n}\left(\hat{\lambda}_{MLE}-\lambda\right)}{\sqrt{\overline{X}}} \le u_\alpha$$

得到

$$\lambda \ge \overline{X} - \frac{\sqrt{\overline{X}}u_\alpha}{\sqrt{n}}$$

则 e^{λ} 的置信上界为

$$e^{-\overline{X}+\frac{\sqrt{\overline{X}}u_\alpha}{\sqrt{n}}}$$

法二：使用 Δ 方法，取 $g(x) = e^{-x}$，则

$$\sqrt{n}\left(e^{\hat{-}\lambda}{}_{MLE} - e^{-\lambda}\right) \to N\left(0, e^{-2\lambda}\lambda\right)$$

即

$$\frac{\sqrt{n}\left(e^{\hat{-}\lambda}{}_{MLE} - e^{-\lambda}\right)}{\sqrt{\lambda e^{-2\lambda}}} \to N(0,1)$$

仍类似处理二，利用 MLE 做二次近似

$$\frac{\sqrt{n}\left(e^{\hat{-}\lambda}{}_{MLE} - e^{-\lambda}\right)}{\sqrt{\overline{X}e^{-2\overline{X}}}} \ge u_\alpha$$

解得置信上界为

$$e^{-\overline{X}} - \frac{\sqrt{\overline{X}e^{-2\overline{X}}}u_\alpha}{\sqrt{n}}$$

解二：利用 CLT

$$\sqrt{n}\left(\overline{X}-\lambda\right) \to N(0,\lambda)$$

余下解法同解一

解三：把 p 看作成功概率，则问题变为求两点分布的参数 p 的置信上界

$$Y_i = \mathbb{I}_{\{X_1=0\}} \sim \mathrm{B}(1,p)$$

则有

$$\frac{\sqrt{n}(\overline{Y}-p)}{\sqrt{\overline{Y}(1-\overline{Y})}} \xrightarrow{D} N(0,1)$$

直接可以得到 p 的置信上界

四.（15 分）设 $X_1,\ldots,X_n$ 为来自正态总体 $N(1,\sigma^2)$ 一组简单样本，$\sigma^2>0$ 为参数．试

（1）求 σ^2 的最小方差无偏估计 $\hat{\sigma}^2$ ，其是否达到 Cramer-Rao 下界？

（2）给出一个比最小方差无偏估计 $\hat{\sigma}^2$ 在均方误差准则下更优的估计．

解：（1）$N(1,\sigma^2)$ 是指数族，样本联合密度为

$$f\left(x_1,\cdots,x_n;\sigma^2\right)=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\cdot e^{-\frac{\sum\limits_{i=1}^n(x_i-1)^2}{2\sigma^2}}$$

显然自然参数空间有内点，进而

$$T(\boldsymbol{X})=\sum_{i=1}^n\left(X_i-1\right)^2$$

为 σ^2 的充分完全统计量

注意当 $\mu=1$ 已知的时候，σ^2 有无偏估计

$$\frac{\sum\limits_{i=1}^n\left(X_i-\mu\right)^2}{n}=\frac{\sum\limits_{i=1}^n\left(X_i-1\right)^2}{n}$$

它正是充分完全统计量的函数，因此 UMVUE 就是

$$\hat{\sigma^2}=\frac{\sum\limits_{i=1}^n\left(X_i-1\right)^2}{n}$$

对 C-R 下界,注意 $N(1,\sigma^2)$ 是单参数指数族,且 UMVUE 为充分完全统计量的线性函数,那么 UMVUE 可以达到 C-R 下界

下面进行数值验证：n 个样本的 Fisher 信息量为

$$\begin{aligned}I\left(\sigma^2\right)&=-\mathbb{E}\left[\frac{\partial^2}{\partial\left(\sigma^2\right)^2}\ln f\left(x_1,\cdots,x_n;\sigma^2\right)\right]\\&=\frac{n}{2\sigma^4}\end{aligned}$$

因此 C-R 下界为

$$\frac{1}{I\left(\sigma^2\right)}=\frac{2\sigma^4}{n}$$

另外一方面

$$\operatorname{Var}\left(\hat{\sigma^2}\right)=\operatorname{Var}\left(\frac{T(\boldsymbol{X})}{n}\right)=\frac{1}{n^2}\operatorname{Var}(T(\boldsymbol{X}))$$

注意

$$\frac{\sum\limits_{i=1}^{n}\left(X_i-\mu\right)^2}{\sigma^2}\sim\chi^2(n)$$

即

$$\operatorname{Var}\left(\frac{T(\boldsymbol{X})}{\sigma^2}\right)=2n$$

也就是

$$\operatorname{Var}\left(\hat{\sigma^2}\right)=\frac{2\sigma^4}{n}$$

达到 C-R 下界

（2）这时我们不要求无偏性

对估计 $\hat{\theta}$，均方误差

$$\begin{aligned}\operatorname{MSE}(\hat{\theta})&=\mathbb{E}\left[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\right]+(\mathbb{E}[\hat{\theta}]-\theta)^2\\&=\operatorname{Var}(\hat{\theta})+(\mathbb{E}[\hat{\theta}]-\theta)^2\end{aligned}$$

考虑

$$\tilde{\sigma^2}=c\hat{\sigma^2}$$

则

$$\begin{aligned}\operatorname{MSE}\left(\widetilde{\sigma}^2\right)&=\operatorname{Var}\left(c\hat{\sigma^2}\right)+\left(c\sigma^2-\sigma^2\right)^2\\&=c^2\cdot\frac{2\sigma^2}{4}+\left(c\sigma^2-\sigma^2\right)^2\end{aligned}$$

对 c 求导数并令其为 0

$$\frac{4c\sigma^4}{n}+2(c-1)\sigma^4=0$$

得到

$$c=\frac{n}{n+2}$$

即

$$\tilde{\sigma^2}=\frac{1}{n+2}\sum_{i=1}^{n}\left(X_i-1\right)^2$$

且满足 $\operatorname{MSE}\left(\tilde{\sigma^2}\right)<\operatorname{MSE}\left(\hat{\sigma^2}\right)$

五．（25 分）设 $X_1,\ldots,X_n$ 为来自如下指数总体的简单样本，总体密度函数为

$$f(x,a)=e^{-(x-a)}I(x\geq a),-\infty<a<1$$

其中 a 为未知参数．试

（1）求 a 的最大似然估计，并讨论其相合性和极限分布．

（2）证明 $T=X_{(1)}$ 为 a 的充分统计量但不是完全统计量．

（3）求 a 的最小方差无偏估计．

解：（1）样本联合密度为

$$f\left(x_1,\cdots x_n;a\right)=e^{-\sum_{i=1}^{n}x_i}\cdot e^{na}\cdot\mathbb{I}_{\left\{x_{(1)}\geq a\right\}}$$

由单调性

$$\hat{a}_{MLE} = X_{(1)}$$

$X_{(1)}$ 有密度

$$f(x) = ne^{-n(x-a)}, x \geq a$$

有分布函数

$$F(x) = 1 - e^{-n(x-a)}$$

进一步

$$P\left(|\hat{a}_{MLE} - a| \geqslant \varepsilon\right) = P(X_{(1)} \geqslant a + \varepsilon) = e^{-n\varepsilon} \to 0$$

这就说明了弱收敛

RK：进一步的

$$\sum_{i=1}^{\infty} P\left(|\hat{a}_{MLE} - a| \geqslant \varepsilon\right) < \infty$$

由 B-C 引理可知强收敛

极限分布：

$$X_i - a \sim \mathrm{Exp}(1) \triangleq Y_i$$

而

$$Y_{(1)} \sim \mathrm{Exp}(n)$$

则

$$n(X_{(1)} - a) \sim \mathrm{Exp}(1)$$

（2）由因子分解定理知 $T(\boldsymbol{X}) = X_{(1)}$ 充分

下面证明它不是完全的，即存在 $\phi(T)$ 使得 $\mathbb{E}[\phi(T)] = 0$ 但是 $\phi(T)$ 不恒为 0

条件

$$\begin{aligned}
\mathbb{E}[\phi(T)] &= \int_a^{+\infty} \phi(t) \cdot ne^{-n(t-a)} dt \\
&= \int_a^1 \phi(t) \cdot ne^{-n(t-a)} dt + \int_1^{+\infty} \phi(t) \cdot ne^{-n(t-a)} dt \\
&= 0
\end{aligned}$$

求导有

$$\phi(t) = 0 \quad \forall t < 1$$

下面构造 $t \geq 1$ 的部分，把积分分段成有限和无限的两段，这两段都不能为 0，且两段的积分之和为 0，去掉常数部分，只需要满足

$$\int_1^c \phi(t) \cdot e^{-nt} dt = -\int_c^{+\infty} \phi(t) \cdot e^{-nt} dt$$

不妨就设 $c = 2$ 且 $\phi(t) = 1, t \geq 2$，则

$$\int_2^{+\infty} \phi(t) \cdot e^{-nt} dt = \frac{e^{-2n}}{n}$$

另一方面

$$\int_1^2 e^{-nt}dt = \frac{-e^{-2n}+e^{-n}}{n}$$

那么只需要令

$$\phi(t) = \begin{cases} 0 & a \le t < 1 \\ 1 - e^{nt} \cdot \frac{e^{-n}}{n} & 1 \le t \le 2 \\ 1 & t > 2 \end{cases}$$

就构造出了这个反例，进而说明 $T(\boldsymbol{X}) = X_{(1)}$ 为 a 的充分统计量但不是完全统计量

RK：构造的 $\phi(T)$ 不应该与未知参数有关

另外地，对于包含无穷长区间的分布都可以类似地使用这个办法，将有限部分和无限部分分成两段再构造反例（这么做是为了让任何无限长的区间内 $\phi(t)$ 不为 0，否则 $\phi(t)$ 仍然有可能以概率 1 地为 0，如平均地从 $\mathbb{R}$ 中取得区间 $(0,1)$ 中的实数的概率为 0）

（3）对于充分但不完全的统计量，用零无偏法，注意参数取值范围 $a < 1$

设 $\mathbb{E}[\delta(T)] = 0$，且 $h(T)$ 为所求的 UMVUE，则有

$$\mathbb{E}[\delta(T)] = \int_a^{+\infty} \delta(t) \cdot ne^{-n(t-a)}dt = 0$$

也就是

$$\int_a^{+\infty} \delta(t) \cdot e^{-nt}dt = 0$$

即

$$\int_a^1 \delta(t) \cdot e^{-nt}dt + \int_1^{+\infty} \delta(t) \cdot e^{-nt}dt = 0$$

求导有

$$\delta(t) = 0 \quad \forall t \le 1$$

另一方面 $\mathbb{E}\left[\delta(T)h(T)\right] = 0$，即

$$\int_a^1 \delta(t)h(t)f(t)dt + \int_1^{+\infty} \delta(t)h(t)f(t)dt = 0$$

也就是

$$\int_1^{+\infty} \delta(t)h(t)f(t)dt = 0$$

为了满足这个要求，待定 $h(T) = c, T > 1$

又需要无偏性，即

$$\mathbb{E}[h(T)] = a$$

待定

$$h(t) = \begin{cases} bt + d & a \le t \le 1 \\ c & t > 1 \end{cases}$$

对积分逐项计算得到

$$b = 1 \quad d = -\frac{1}{n} \quad c = -1$$

最后

$$h(t)=\begin{cases}t-\frac{1}{n} & a\le t\le 1\\ -1 & t>1\end{cases}$$

附表：上分位数

$$u_{0.025}=1.960,\quad u_{0.05}=1.645,\quad \chi^2_{899}(0.025)=984,\quad \chi^2_{899}(0.975)=817.8$$