数理统计往年卷合集

NULIOUS

前言

本文档是数理统计（管院）往年卷合集，于 25 年夏天编写，整理了有试题的往年卷的答案，仅供参考，如文档有错误或者有其他问题请发送至我的邮箱3589851379@qq.com

数理统计期末往往考整本书，期中复习的时候可以参考往年期末卷前半本书的试题

我的个人主页：https://mulious.github.io/

如果试题或答案有误请先参考原卷，其中 24，25 期末均有答案：数理统计往年题原卷

预祝大家在考试前速通成功！

目录

Chapter 1

往年期中试卷

1.1 25 数理统计期中

一. 填空选择题（每空两分）

（1）设 $X_1, X_2, \ldots, X_n, X_{n+1}$ 为来自同一正态总体的一组简单随机样本，且记 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 及 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ 。若统计量 $c_n\left(X_{n+1} - \overline{X}\right)/S$ 服从 t 分布，则常数 $c_n =$_____ ,t 分布的自由度为_____, 且与 $\sum_{i=1}^{n+1} X_i$ 的相关系数为_____

答案：$(\pm)\sqrt{\frac{n}{n+1}}; n-1; 0$

首先对于正态分布，$\overline{X}$ 与 S^2 是独立的，这说明了 t 分布的分子分母的独立性

由 t 分布的定义：

$$t_{n-1} = \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}} \tag{1.1}$$

再有

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \tag{1.2}$$

$$X_{n+1} - \overline{X} \sim N\left(\mu, \sigma^2\right) - N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(0, \frac{n+1}{n}\sigma^2\right) \tag{1.3}$$

(3) 式来自两个变量的独立性

则

$$\sqrt{\frac{n}{n+1}}\left(X_{n+1} - \overline{X}\right)/S \sim t_{n-1} \tag{1.4}$$

特别的，正负号来自 t 分布是对称的（没有负号也没算错）

数理统计涉及独立性的几乎只有 $Basu$ 定理一个，猜测它们独立，即相关系数为 0

把 $(X_1, \cdots, X_{n+1})$ 视为样本，则由于指数族的性质，关于 λ 有充分完全统计量 $T = \sum_{i=1}^{n+1} X_i$

我们不考虑常数部分，则

$$\frac{(X_{n+1})-(\overline{X})}{S}=\frac{(X_{n+1}-\mu)-(\overline{X}-\mu)}{S} \tag{1.5}$$

是与 μ 无关的统计量（即辅助量），因此由 $Basu$ 定理它们独立，进而相关系数为 0

（2）设统计量 $\hat{\theta}$ 为总体参数 θ 的一个点估计，下列说法一般不成立的是_____

（A）若 $\hat{\theta}$ 为 θ 的矩估计，则 $\hat{\theta}^2$ 为 θ^2 的矩估计

（B）若 $\hat{\theta}$ 为 θ 的最大似然估计，则 $\hat{\theta}^2$ 为 θ^2 的最大似然估计

（C）若 $\hat{\theta}$ 为 θ 的无偏估计，则 $\hat{\theta}^2$ 为 θ^2 的无偏估计

（D）若 $\hat{\theta}$ 为 θ 的相合估计，则 $\hat{\theta}^2$ 为 θ^2 的相合估计

答案：C

一般一个随机变量的二阶矩不等于其一阶矩的平方，因此 C 错误

（3）如果极小充分统计量存在，那么充分完全统计量必是极小充分统计量，但是极小充分统计量不一定是完全的。这种说法_____

（A）正确

（B）错误

答案：A

（4）设 $X_1,\cdots,X_n$ 为来自于正态总体 $N(\mu,1)$ 的简单随机样本，若要求参数 μ 的置信系数为 95% 的置信区间长度不超过 1 ，则至少需要抽取的样本量 n 为 _____

（A）14

（B）16

（C）18

（D）20

答案：B

注意方差已知。则置信区间为 $\left[\overline{X}-\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}},\overline{X}+\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}\right]$, 带入数值 $\sigma=1$ 即可

（5）在给定一组样本值和先验下，采用后验期望作为感兴趣参数 θ 的估计，得到估计值 $\hat{\theta}=5$. 下述说法正确的是_____

（A）在重复抽取样本意义下 θ 的无偏估计值为 1.5

（B）$\hat{\theta}=1.5$ 是 θ 的有效估计

（C）估计值 1.5 是最小后验均方误差估计

（D）估计值 1.5 是 θ 的相合估计

答案：C

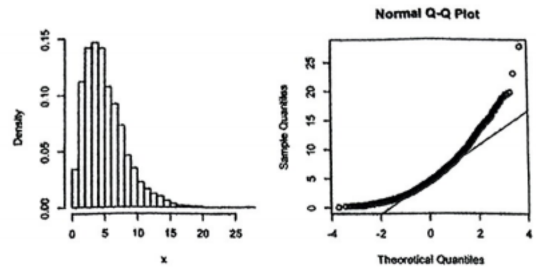
二。（16 分）随机调查了某保险公司 n 个独立的车险索赔额 $X_1,\ldots,X_n$（单位：千元），得到如下样本直方图和正态 Q-Q 图。据此回答

（1）该样本来自的总体分布有何特点？可以选择什么分布作为总体分布？给出理由.

（2）试选择合适的参数统计模型，并讨论参数的充分完全统计量.

解：

（1）总体分布为单峰且峰偏左（右偏分布），且正态 q-q 图在第一象限对角线 $y=x$ 的下方；我们可以选择 Γ 分布，卡方分布（也是一种 Γ 分布）等符合要求的分布



（2）以总体分布为 Γ 分布：$X \sim \Gamma(\alpha,\beta)$ 为例

样本 $(X_1,...,X_n)$ 有联合密度:

$$f(\vec{x};\alpha,\beta)=\frac{\beta^{n\alpha}}{\Gamma(\alpha)^n}\prod_{i=1}^{n}x_i^{\alpha-1}e^{-\beta\sum_{i=1}^{n}x_i} \tag{1.6}$$

把密度写成指数族的形式：

$$f(\vec{x};\alpha,\beta)=\frac{\beta^{n\alpha}}{\Gamma(\alpha)^n}e^{(\alpha-1)\sum_{i=1}^{n}\ln(x_i)-\beta\sum_{i=1}^{n}x_i} \tag{1.7}$$

自然参数分别为 $\alpha-1$，β，自然空间显然有内点，同时由因子分解定理，有充分完全统计量 $T=\left(\sum\limits_{i=1}^{n}\ln(X_i),\sum\limits_{i=1}^{n}X_i\right)$

特别的，若选择了卡方分布，则充分完全统计量为 $T=\sum\limits_{i=1}^{n}\ln(X_i)$

三.（20 分）设 $X_1,...,X_n$ 为来自均匀总体 $\mathrm{U}(\theta,\theta+1)$ 的简单样本，其中 $\theta\in R$ 为未知参数．试

（1）证明 $T=\left(X_{(1)},X_{(n)}\right)$ 为 θ 的极小充分统计量但不是完全统计量.

（2）求 θ 的最大似然估计，并讨论其相合性.

解：

（1）首先要利用因子分解定理证明 $(X_{(1)},X_{(n)})$ 是充分统计量

样本联合密度：

$$f(\vec{x};\theta)=\mathbb{I}_{\theta<X_{(1)}<X_{(n)}<\theta+1} \tag{1.8}$$

则 $h(\boldsymbol{X})=1$，$g(X_{(1)},X_{(n)};\theta)=\mathbb{I}_{\theta<X_{(1)}<X_{(n)}<\theta+1}$

由因子分解定理可以知道 $(X_{(1)},X_{(n)})$ 是充分统计量

下面再取相同总体中的 n 个样本 $(Y_1,...,Y_n)$，构造似然比：

$$\begin{aligned}\frac{f(\vec{x};\theta)}{f(\vec{y};\theta)}=C(\vec{x},\vec{y}) &\iff \frac{\mathbb{I}_{\theta<X_{(1)}<X_{(n)}<\theta+1}}{\mathbb{I}_{\theta<Y_{(1)}<Y_{(n)}<\theta+1}}=C(\vec{x},\vec{y})\\ &\iff (X_{(1)},X_{(n)})=(Y_{(1)},Y_{(n)})\end{aligned} \tag{1.9}$$

其中 $C(\vec{x},\vec{y})$ 表示仅与 $\vec{x}$，$\vec{y}$ 有关的常数，这就说明了 $(X_{(1)},X_{(n)})$ 是极小充分统计量

下面通过充分统计量来构造辅助量（与 θ 无关的统计量）来说明 $T=(X_{(1)},X_{(n)})$ 不是完全统计量

设 $Z_i = X_i - \theta \sim U(0,1)$，则

$$Z_{(n)} - Z_{(1)} = ((X_{(n)} - \theta) - (X_{(1)} - \theta)) \sim \beta(n-1,2)$$

与 θ 无关
上面的结论来自 $U(0,1)$ 的极差分布为 $\beta(n-1,2)$
取 a，b 使得

$$P(Z_{(n)} - Z_{(1)} > a) = P(Z_{(n)} - Z_{(1)} < b) > 0$$

再取

$$\varphi(x) = \begin{cases} 1, & x > a \\ 1, & x < b \\ 0, & \text{其他} \end{cases}$$

则 $\mathbb{E}[\varphi(T)] = 0$ 但是 $\varphi(T)$ 显然不处处为 0
则 T 不是完全统计量
RK：对于二元的充分统计量要说明其不是完全的，往往通过相减和相除构造辅助量，再取如 φ 这样的函数进行说明
（2）接下来求 θ 的最大似然估计
由式 (8) 可以看出

$$f(\vec{x};\theta) = \begin{cases} 1, & X_{(n)} - 1 < \theta < X_{(1)}, \\ 0, & \text{其他}. \end{cases} \tag{1.10}$$

则 θ 的最大似然估计 $\hat{\theta}_{MLE}$ 为 $(X_{(n)} - 1, X_{(1)})$ 中的任何值
下面利用 $Markov$ 不等式证明其弱相合性
只需说明 $tX_{(1)} + (1-t)(X_{(n)} - 1)$ 对于 θ 的相合性即可，其中 $0 < t < 1$

$$\begin{aligned} P\Big(\big|tX_{(1)} + (1-t)(X_{(n)} - 1) - \theta\big| \geq \epsilon\Big) &\leq \frac{\mathbb{E}\big[\big|tX_{(1)} + (1-t)(X_{(n)} - 1) - \theta\big|\big]}{\epsilon} \\ &\leq \frac{\mathbb{E}\big[\big|tX_{(1)} - t\theta\big|\big] + \mathbb{E}\big[\big|(1-t)(X_{(n)} - 1) - (1-t)\theta\big|\big]}{\epsilon} \end{aligned}$$

而

$$X_{(1)} - \theta \sim \beta(1,n), \tag{1.11}$$

$$X_{(n)} - \theta \sim \beta(n,1). \tag{1.12}$$

则 $\mathbb{E}[X_{(1)}] = \theta + \frac{1}{n+1}$，$\mathbb{E}[X_{(n)}] = \theta + \frac{n}{n+1}$
注意关系 $\theta < X_{(1)} < X_{(n)} < \theta + 1$
代入 $Markov$ 不等式后令 $n \to \infty$ 即证弱收敛

四.（25 分）某厂生产的产品分为三个质量等级 $(X = 1,2,3)$，各等级产品的分布如下

X	1	2	3
P	θ	2θ	$1-3\theta$

其中 $\theta \in (0,1/3)$ 未知．为了解该厂产品的质量分布情况，从该厂产品中随机有放回抽取 20 件产品检测后发现一等品有 5 件，二等品有 7 件，三等品有 8 件．试

（1）求 θ 的矩估计和最大似然估计量，是否都为无偏估计？给出估计值．

（2）求 θ 的最小方差无偏估计量，其方差是否达到了 Cramér-Rao 下界？

解：

（1）$\mathbb{E}[X] = 3-4\theta$

则 θ 的矩估计为 $\hat{\theta}_M = \frac{3-\overline{X}}{4}$，它自然是无偏的（因为就是拿期望算出来的）

记 $n_k = \sum_{i=1}^n \mathbb{I}_{X_i=k}$

则 $(X_1,...,X_n)$ 有联合密度

$$f(\vec{x};\theta) = \theta^{n_1}(2\theta)^{n_2}(1-3\theta)^{n_3}$$

则

$$\ln f(\vec{x};\theta) = n_1 \ln\theta + n_2\ln(2\theta) + n_3\ln(1-3\theta) \tag{1.13}$$

$$\frac{\partial \ln f\left(\vec{x};\theta\right)}{\partial\theta} = \frac{n_1}{\theta} + \frac{n_2}{\theta} - \frac{3n_3}{1-3\theta} = 0 \tag{1.14}$$

有

$$\hat{\theta}_{MLE} = \frac{n-n_3}{3n}$$

又 $n_3 \sim B(n, 1-3\theta)$

则

$$\mathbb{E}[\hat{\theta}_{MLE}] = \frac{1}{3} - \frac{\mathbb{E}[n_3]}{3n} = \frac{1}{3} - \frac{n(1-3\theta)}{3n} = \theta$$

即 $\hat{\theta}_{MLE}$ 为无偏估计

带入数值有 $\hat{\theta}_M = 0.2125$，$\hat{\theta}_{MLE} = 0.2$

（2）化为自然指数族的形式

$$f(\vec{x};\theta) = e^{n_2\ln 2}e^{n_1\ln\theta + n_2\ln\theta + n_3\ln(1-3\theta)} \tag{1.15}$$

$$= e^{n_2\ln 2}e^{\ln\theta}e^{n_3\ln(\frac{1-3\theta}{\theta})} \tag{1.16}$$

其中 $h(\boldsymbol{X}) = e^{n_2\ln 2}$ $\quad C(\theta) = e^{\ln\theta} = \theta$，自然参数为 $\ln(\frac{1-3\theta}{\theta})$

又 $\theta \in (0,\frac{1}{3})$, 则自然参数空间有内点，且由因子分解定理，$T = n_3$ 为 θ 的充分完全统计量

同时注意到 $\hat{\theta}_{MLE}$ 无偏且为充分完全统计量的函数，它也就是 θ 的 $UMVUE$

注意 $UMVUE$ 能达到 $C-R$ 下界 $\iff$ 分布为单参数指数族且 $UMVUE$ 为充分完全统计量 $T(\vec{x})$ 的线性函数

所以本题的 $UMVUE$ 能达到 $C-R$ 下界

下面额外给出数值验证：

$$\mathrm{Var}(\hat{\theta}_{\mathrm{MLE}}) = \frac{\mathrm{Var}(n_3)}{9n^2} = \frac{3n\theta(1-3\theta)}{9n^2} = \frac{\theta(1-3\theta)}{3n}, \tag{1.17}$$

$$\begin{aligned} n\text{个样本的 Fisher 信息量：} I(\theta) &= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(\vec{x};\theta)\right] \\ &= -\mathbb{E}[-\frac{n_1}{\theta^2} - \frac{n_2}{\theta^2} - \frac{9n_3}{(1-3\theta)^2}] \\ &= \frac{3n}{\theta(1-3\theta)} \end{aligned} \tag{1.18}$$

对于 θ 和 n 个样本的 $Fisher$ 信息量 $I(\theta)$，其 $C-R$ 下界为

$$\frac{1}{I(\theta)} = \frac{\theta(1-3\theta)}{3n}$$

这就说明了 $UMVUE$ 的方差能达到 $C-R$ 下界

RK：对于 n 个样本的 $Fisher$ 信息量 $I(\theta)$,$g(\theta)$ 的 $C-R$ 下界为 $\frac{[g'(\theta)]^2}{I(\theta)}$

对于单个样本（总体）的 $Fisher$ 信息量 $\widetilde{I(\theta)}$,$g(\theta)$ 的 $C-R$ 下界为 $\frac{[g'(\theta)]^2}{n\widetilde{I(\theta)}}$

这实际上是因为 $I(\theta)$ 是 $\widetilde{I(\theta)}$ 的 n 倍，另外对于 $Fisher$ 信息阵仍有相同的规律

五.（25 分）调查发现人们每天使用手机的时间（单位：分钟）服从正态分布 $N(\mu,\sigma^2)$，其中 $\mu \in R, \sigma^2 > 0$ 为未知参数。现随机调查了 25 个人每天使用手机时间，得到样本均值 $\overline{X} = 180$ 分钟，样本标准差 $S = 20$ 分钟。若取先验分布为 $\pi(\mu,\sigma^2) \propto \sigma^{-2}$. 试

（1）求 σ^2 的边际后验分布，并给出 σ^2 的后验期望估计值.

（2）求一个人每天平均使用手机时长 μ 的 95% 置信区间和可信区间，两者的解释有何不同？

解：

（1）先解释一下 $\pi(\mu,\sigma^2) \propto \sigma^{-2}$ 的含义，他表示先验分布与 μ 成常数倍的关系，但不意味着先验分布给出的**只有** σ 的信息，因此按照这个先验分布算出来的后验分布是 μ 与 σ 的联合分布，如果要求某一个参数的分布还需要对另一个参数进行积分

样本联合密度为：

$$f(\vec{x};\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}} \tag{1.19}$$

后验联合密度为：

$$\pi(\mu,\sigma^2|\vec{x}) \propto f(\vec{x};\mu,\sigma^2)\cdot\pi\left(\mu,\sigma^2\right) \propto (\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{\sum(x_i-\overline{X})^2 + n(\overline{X}-\mu)^2}{2\sigma^2}} \tag{1.20}$$

其中用到了

$$\sum_{i=1}^n (x_i-\mu)^2 = \sum_{i=1}^n (x_i-\overline{X})^2 + n(\overline{X}-\mu)^2$$

接下来对参数 μ 做积分来得到 σ^2 的后验密度

$$\begin{aligned} \pi(\sigma^2|\vec{x}) &= \int_{-\infty}^{+\infty} \pi(\mu,\sigma^2|\vec{x})d\mu \\ &= (\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{\sum_{i=1}^n (x_i-\overline{X})^2}{2\sigma^2}} \int_{-\infty}^{+\infty} e^{-\frac{n(\overline{X}-\mu)^2}{2\sigma^2}} d\mu \\ &= (\sigma^2)^{-\frac{n}{2}-\frac{1}{2}} e^{-\frac{\sum_{i=1}^n (x_i-\overline{X})^2}{2\sigma^2}} \end{aligned} \tag{1.21}$$

式 (26) 来自 $N(\overline{X},\frac{\sigma^2}{n})$ 的密度的积分为 1

即

$$\int_{-\infty}^{+\infty} e^{-\frac{n(\overline{X}-\mu)^2}{2\sigma^2}}d\mu=(\frac{2\pi\sigma^2}{n})^{\frac{1}{2}}$$

则

$$\sigma^2|\vec{x}\sim\Gamma^{-1}(-\frac{n}{2}+\frac{1}{2},\frac{\sum(x_i-\overline{X})^2}{2})$$

最后

$$\hat{\sigma^2}_E=\frac{\sum\limits_{i=1}^{n}(x_i-\overline{X})^2}{2(-\frac{n}{2}-\frac{1}{2})}=480$$

RK：在矩存在的条件下有

$$X\sim\Gamma(\alpha,\beta)\quad 有\forall n\in\mathbb{Z}\quad \mathbb{E}[X^n]=\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)\beta^n}$$

$$Y\sim\Gamma^{-1}(\alpha,\beta)\quad 有\forall n\in\mathbb{Z}\quad \mathbb{E}[Y^n]=\frac{\Gamma(\alpha-n)\beta^n}{\Gamma(\alpha)}$$

事实上，Γ 分布的 n 阶矩就是 Γ^{-1} 分布的 $-n$ 阶矩

（2）

先解释一下置信区间与可信区间的区别：

置信区间：经过多次重复实验，μ 落在置信区间的**频率**趋于 95%

可信区间：相当于把 μ 视为**随机变量**，μ 落在可信区间的**概率**为 95%

构造置信区间：

未知 μ，σ^2

取

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

那么置信区间为：

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

带入数值为 $[171.76,188.24]$

构造可信区间：

首先要求出 μ 的后验分布：

$$\begin{aligned}\pi(\mu|\vec{x})&=\int_0^{+\infty}\pi(\mu,\sigma^2|\vec{x})d\sigma^2\\&=\int_0^{+\infty}(\sigma^2)^{-\frac{n}{2}-1}e^{-\frac{\sum(x_i-\overline{X})^2+n(\overline{X}-\mu)^2}{2\sigma^2}}d\sigma^2\\&\overset{t=\frac{1}{\sigma^2}}{=}\int_0^{+\infty}(t)^{\frac{n}{2}-1}e^{-\frac{t(\sum(x_i-\overline{X})^2+n(\overline{X}-\mu)^2)}{2}}dt\end{aligned}\tag{1.22}$$

设 $\alpha=\frac{n}{2}$，$\beta=\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2+n(\overline{X}-\mu)^2}{2}$

有 $\Gamma(\alpha,\beta)$ 的密度的积分为 1

即

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

则

$$\begin{aligned}\pi(\mu|\vec{x}) &\propto \frac{\Gamma(\alpha)}{\beta^\alpha} \\ &= \frac{\Gamma(\frac{n}{2})}{(\frac{\sum_{i=1}^n (x_i-\overline{X})^2 - n(\overline{X}-\mu)^2}{2})^{\frac{n}{2}}}\end{aligned} \tag{1.23}$$

接下来的计算意义不大，因为考试时没有提供非标准 t 分布的密度

$$t_v(\mu,\sigma^2) \sim f(x) = \frac{\Gamma(\frac{v+1}{2})}{\sigma\sqrt{v\pi}\Gamma(\frac{v}{2})}(1+\frac{(x+\mu)^2}{v\sigma^2})^{-\frac{v+1}{2}}$$

带入数值可以得到

$$\mu|\vec{x} \sim t_{24}(180, 16)$$
$$\frac{\mu|\vec{x} - 180}{4} \sim t_{24}$$

取其上下 $\frac{\alpha}{2}$ 分位数，得到的可信区间也为 $[171.76, 188.24]$

附表：上分位数 $u_{0.025} = 1.960, u_{0.05} = 1.645, t_{24}(0.025) = 2.06, t_{24}(0.05) = 1.71$
伽马分布，逆伽马分布与 t 分布概率密度函数：

$$\begin{aligned}&Ga(\alpha,\beta): f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \alpha, \beta, x > 0 \\ &\operatorname{Inv} Ga(\alpha,\beta): f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{x}}, \alpha, \beta, x > 0 \\ &t_n: f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}, -\infty < x < \infty\end{aligned}$$

1.2 24 数理统计期中

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

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

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

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

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

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

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

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

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

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

对积分逐项计算得到

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

最后

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

附表：上分位数

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

1.3 23 数理统计期中残卷

1. 设从总体

X	0	1	2	3
$\mathbb{P}$	$3\theta/4$	$\theta/4$	2θ	$1-3\theta$

（其中 $0<\theta<1/3$ ）中抽取的一个简单样本 $X_1,\ldots,X_n$ ，试

（1）将样本分布表示为指数族自然形式，指出自然参数及自然参数空间。

（2）求 θ 的最大似然估计量 $\hat{\theta}$ ，其是否为一致最小方差无偏估计？

（3）证明 $\hat{\theta}$ 具有相合性和渐近正态性。

解：（1）样本联合密度为

$$f\left(x_1\cdots,x_n;\theta\right)=\left(\frac{3\theta}{4}\right)^{n_0}\left(\frac{\theta}{4}\right)^{n_1}\cdot(2\theta)^{n_2}\cdot(1-3\theta)^{n_3}$$

这里 n_i 为 n 个样本中取值为 i 的个数，把样本联合密度写成指数族的形式为

$$f\left(x_1,\cdots,x_n;\theta\right)=\left(\frac{3}{4}\right)^{n_0}\left(\frac{1}{4}\right)^{n_1}\cdot 2^{n_2}\cdot e^{(n-n_3)\ln\theta+n_3\ln(1-3\theta)}$$

其中

$$h(\boldsymbol{X})=\left(\frac{3}{4}\right)^{n_0}\left(\frac{1}{4}\right)^{n_1}2^{n_2}$$
$$C(\theta)=e^{n\ln\theta}$$

自然参数为 $\varphi=\ln\frac{1-3\theta}{\theta}$

又 $0<\theta<\frac{1}{3}$，则自然参数空间为 $\mathbb{R}$

由因子分解定理和自然参数空间有内点，充分完全统计量为 $T(\boldsymbol{X})=n_3$

（2）

$$\ln f\left(x_1,\cdots,x_n;\theta\right)\propto(n-n_3)\ln\theta+n_3\ln(1-3\theta)$$

令

$$\frac{\partial\ln f}{\partial\theta}=\frac{n-n_3}{\theta}-\frac{3n_3}{1-3\theta}=0$$

解得

$$\hat{\theta}_{MLE}=\frac{n-n_3}{3n}$$

又 $n_3\sim\mathrm{B}(n,1-3\theta)$

$$\mathbb{E}[n_3]=n(1-3\theta)\quad\mathbb{E}\left[\hat{\theta}_{MLE}\right]=\theta$$

也就是 MLE 为无偏估计量，且它是充分完全统计量的函数，则其为 UMVUE

（3）相合性：

强相合：注意

$$n_3\sim\mathrm{B}(n,1-3\theta)$$

为 n 个 $\mathrm{B}(1,1-3\theta)$ 的独立和

由 SLLN

$$\frac{n_3}{n}\xrightarrow{a.s.}1-3\theta$$

则

$$\frac{n-n_3}{3n} \xrightarrow{a.s.} \theta$$

即强相合

弱相合：

i. 直接由强相合得到

ii. 由 Chebyshev 不等式

$$P\left(\mid \hat{\theta}_{MLE} - \theta \mid \geq \epsilon\right) \leq \frac{\operatorname{Var}\left(\hat{\theta}_{MLE}\right)}{\epsilon^2}$$

另一方面

$$\begin{aligned}\operatorname{Var}(\hat{\theta}_{MLE}) &= \operatorname{Var}\left(\frac{n-n_3}{3n}\right) &= \operatorname{Var}\left(\frac{n_3}{3n}\right) \\ &= \frac{1}{9n^2}\operatorname{Var}\left(n_3\right) &= \frac{\theta(1-3\theta)}{3n}\end{aligned}$$

则

$$\lim_{n\to\infty} P(|\hat{\theta}_{MLE} - \theta| \geq \varepsilon) \leqslant \frac{\theta(1-3\theta)}{3n\varepsilon^2} \to 0$$

就说明了弱收敛

渐进正态性：

i.MLE 的渐近正态性

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

这里总体（单个样本）的 Fisher 信息量为

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

其中 $f(x,\theta)$ 为总体的密度

$$\mathbb{E}[\frac{\partial^2 \ln f(x,\theta)}{\partial \theta^2}] = \ln'' f(x=0,\theta)P(X=0) + \cdots + \ln'' f(x=3,\theta)P(X=3)$$

得到

$$I(\theta) = \frac{3}{\theta} + \frac{9}{1-3\theta}$$

则

$$\sqrt{n}\left(\hat{\theta}_{MLE} - \theta\right) \xrightarrow{D} N\left(0, \frac{\theta(1-3\theta)}{3}\right)$$

ii.CLT

$$\sqrt{n}\left(\frac{n_3}{n} - (1-3\theta)\right) \xrightarrow{D} N(0, (1-3\theta)3\theta)$$

则

$$\sqrt{n}\left(\hat{\theta}_{MLE} - \theta\right) \xrightarrow{D} N\left(0, \frac{\theta(1-3\theta)}{3}\right)$$

附表：$u_{0.025} = 1.960, u_{0.05} = 1.645, \Phi(1.36) = 0.9131, \Phi(1.4) = 0.9192.$

1.4 21 数理统计期中

一、（20 分）假设 $X_1, X_2, \cdots, X_n$ 为来自 Poisson 总体 $\mathrm{P}(\lambda)$ 的一组简单样本，试

（1）将抽样分布表示为指数族的自然形式，并给出自然参数空间；

（2）证明统计量 $T=\sum_{i=1}^n X_i$ 为充分完备统计量。

（3）证明 $g(\lambda)=e^{-\lambda}$ 不存在可以达到 C-R 不等式下界的无偏估计.

解：（1）样本联合密度

$$f\left(x_1, \cdots, x_n ; \lambda\right)=e^{-n \lambda} \frac{\lambda^{\sum\limits_{i=1}^n x_i}}{x_1! \cdots x_n!}$$

写成指数族的自然形式为

$$f\left(x_1, \cdots, x_n ; \lambda\right)=e^{-n \lambda} \cdot \frac{1}{x_1! \cdots x_n!} \cdot e^{\sum\limits_{i=1}^n x_i \ln \lambda}$$

这里

$$C(\theta)=e^{-n \lambda} \quad h(\boldsymbol{X})=\frac{1}{x_1! \cdots x_n!}$$

自然参数为 $\ln \lambda$，又 $\lambda \in(0,+\infty)$

则自然参数空间为

$$\Theta=\{\ln \lambda: \ln \lambda \in \mathbb{R}\}$$

（2）由因子分解定理和自然参数空间有内点，则 $T(\boldsymbol{X})=\sum\limits_{i=1}^n X_i$ 为充分完全统计量

（3）无偏估计能达到 C-R 下界 $\Longleftrightarrow$ 分布为单参数指数族且 UMVUE 为充分完全统计量 $T(\boldsymbol{X})$ 的线性函数

先求 UMVUE

解一：注意到

$$e^{-\lambda}=P\left(X_1=0\right)$$

再仿照 24 期中 3（1）可以得到 UMVUE

$$\widehat{e^{-\lambda}}=\hat{g}(T)=\left(\frac{n-1}{n}\right)^T$$

解二：首先

$$T(\boldsymbol{X}) \sim \mathrm{P}(n \lambda)$$

设 UMVUE 为 $\hat{g}(T)$，则

$$\mathbb{E}[\hat{g}(T)]=\sum_{k=0}^{\infty} \hat{g}(k) \frac{(n k)^k}{k!} \cdot e^{-n \lambda}=e^{-\lambda}$$

也就是

$$\sum_{k=0}^{\infty} \hat{g}(k) \frac{(n \lambda)^k}{k!}=e^{(n-1) \lambda}=\sum_{k=0}^{\infty} \frac{(n-1)^k \lambda^k}{k!}$$

就能得到

$$\hat{g}(T)=\left(\frac{n-1}{n}\right)^T$$

显然 UMVUE 不是充分完全统计量的线性函数，因此不能达到 C-R 下界
数值验证：
下面求总体（单个样本）的 Fisher 信息量

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

这里 X 为总体分布，且密度为

$$f(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$$

因此 C-R 下界为

$$\frac{\left(\frac{\partial\left(e^{-\lambda}\right)}{\partial\lambda}\right)^2}{nI(\lambda)} = \frac{\lambda e^{-2\lambda}}{n}$$

再计算 UMVUE 的方差

$$\begin{aligned}\operatorname{Var}(\hat{g}(T)) &= \mathbb{E}\left[\left(\frac{n-1}{n}\right)^{2T}\right] - (\mathbb{E}[\hat{g}(T)])^2 \\ &= \mathbb{E}\left[\left(\frac{n-1}{n}\right)^{2T}\right] - e^{-2\lambda}\end{aligned}$$

其中

$$\begin{aligned}E\left[\left(\frac{n-1}{n}\right)^{2T}\right] &= \sum_{k=0}^{\infty}\left(\frac{n-1}{n}\right)^{2k}\frac{(n\lambda)^k}{k!}e^{-n\lambda} \\ &= e^{-n\lambda}\sum_{k=0}^{\infty}\frac{\left(\frac{(n-1)^2\lambda}{n}\right)^k}{k!} \\ &= e^{-n\lambda}\cdot e^{\frac{(n-1)^2\lambda}{n}} \\ &= e^{-2\lambda+\frac{\lambda}{n}}\end{aligned}$$

最后

$$\operatorname{Var}(\hat{g}(T)) = e^{-2\lambda+\frac{\lambda}{n}} - e^{-2\lambda}$$

不能达到 C-R 下界

二.（30 分）设 $X_1,\dots,X_n$ 为来自 0-1 分布 $\mathrm{B}(1,p), 0<p<1$ 的一组简单样本，试
（1）求 $g(p)=(1-p)^2$ 的矩估计量和极大似然估计量，并说明是否为无偏估计。
（2）求 $g(p)$ 的 UMVUE，其方差是否达到 C-R 不等式的下界？
（3）证明 $g(p)$ 的极大似然估计量具有渐近正态性.
解：（1）由于 $\mathbb{E}[X]=p$，矩估计量为

$$\hat{g(p)}_M = (1-\overline{X})^2$$

又样本联合密度

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

求导有

$$\frac{\partial \ln f}{\partial p} = \frac{\sum_{i=1}^n x_i}{p} - \frac{n-\sum_{i=1}^n x_i}{1-p} = 0.$$

解得

$$\hat{p}_{MLE} = \overline{X}$$

再由 MLE 的不变性，有

$$g(\hat{p})_{MLE} = \left(1-\overline{X}\right)^2$$

无偏性：

$$\begin{aligned}
\mathbb{E}\left[(1-\overline{X})^2\right] &= \mathbb{E}\left[1-2\overline{X}+\overline{X}^2\right] \\
&= 1-2\mathbb{E}[\overline{X}] + \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^2\right] \\
&= 1-2p+p^2+\frac{p(1-p)}{n} \\
&\neq (1-p)^2
\end{aligned}$$

这就说明两个估计均不无偏

（2）注意二项分布为指数族，且自然参数空间有内点，易知 $T(\boldsymbol{X}) = \sum\limits_{i=1}^n X_i$ 为充分完全统计量

先找一个无偏估计

$$\mathbb{I}_{\{X_1=0, X_2=0\}}$$

则 UMVUE 为

$$\begin{aligned}
\hat{g}(p)_{UMVUE} &= \mathbb{E}\left[\mathbb{I}_{\{X_1=0,X_2=0\}} \mid \sum_{i=1}^n x_i = t\right] \\
&= P\left(X_1=0, X_2=0 \mid \sum_{i=1}^n X_i = t\right) \\
&= \frac{P\left(X_1=0, X_2=0, \sum_{i=3}^n X_i = t\right)}{P\left(\sum_{i=1}^n X_i = t\right)} \\
&= \frac{(n-t)(n-t-1)}{n(n-1)}
\end{aligned}$$

也就是

$$\hat{g}(p)_{UMVUE} = \frac{(n-T)(n-T-1)}{n(n-1)}$$

它不是 $T(\boldsymbol{X}) = \sum\limits_{i=1}^n X_i$ 的线性函数，因此不能达到 C-R 下界

数值验证：先求总体（单个样本）的 Fisher 信息量

$$\begin{aligned}
I\left(p\right) &= -\mathbb{E}\left[\frac{\partial^2}{\partial p^2}\ln f\left(x;p\right)\right] \\
&= \frac{1}{p(1-p)}
\end{aligned}$$

这里总体的密度为

$$f(x;p) = p^x(1-p)^{1-x}$$

因此 C-R 下界为

$$\frac{\frac{\partial g(p)}{\partial p}}{nI(p)} = \frac{4p(1-p)^3}{n}$$

又

$$\mathrm{Var}(\hat{g}(p)_{UMVUE}) = \mathrm{Var}\left(\frac{(n-T)(n-T-1)}{n(n-1)}\right) = \frac{\mathrm{Var}\left(n^2-n+(1-2n)T+T^2\right)}{n^2(n-1)^2}$$

其中

$$\begin{aligned}\mathrm{Var}\left(n^2-n+(1-2n)T+T^2\right) &= \mathrm{Var}\left((1-2n)T+T^2\right)\\ &= \mathbb{E}\left[\left[(1-2n)T+T^2\right]^2\right] - \left(\mathbb{E}\left[(1-2n)T+T^2\right]\right)^2\end{aligned}$$

这里由于 $T \sim \mathrm{B}(n,p)$

$$\begin{aligned}\mathbb{E}\left[\left[(1-2n)T+T^2\right]^2\right] &= \mathbb{E}\left[(1-2n)^2T^2+(2-4n)T^3+T^4\right]\\ &= \mathbb{E}\left[(1-2n)^2T^2\right]+\mathbb{E}\left[(2-4n)T^3\right]+\mathbb{E}\left[T^4\right]\\ &= (1-2n)^2np(1-p+np)+(2-4n)(1-3p+3np+(2-3n+n^2)p^2)\\ &+np(1-p)(1-6p+6p^2)+7np(1-p)(n-1)p+6n(n-1)(n-2)p^3+n^3p^4\end{aligned}$$

另一方面

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

这就说明了 UMVUE 不能达到 C-R 下界

（3）解一：直接利用 MLE 的渐进正态性

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

也就是

$$\sqrt{n}\left(\hat{p}_{MLE}-p\right) \xrightarrow{D} N\left(0, p(1-p)\right)$$

再使用 Δ 方法

$$\sqrt{n}\left(g(\hat{p})_{MLE}-g(p)\right) \xrightarrow{D} N\left(0, 4p(1-p)^3\right)$$

解二：$\overline{X}$ 有 CLT

$$\sqrt{n}\left(\overline{X}-p\right) \xrightarrow{D} N\left(0, p(1-p)\right)$$

再使用 Δ 方法

$$\sqrt{n}\left(g(\hat{p})_{MLE}-g(p)\right) \xrightarrow{D} N\left(0, 4p(1-p)^3\right)$$

三.（15 分）设 $X_1,\ldots,X_n$ 来自正态总体 $N(\mu,1)$ 的一组简单随机样本，试 $(0<\alpha<1)$

（1）证明样本平均值 $\overline{X}$ 与统计量 $X_1-\overline{X}$ 相互独立.

（2）求概率 $P\left(X_1 \le 0\right)$ 的 UMVUE 以及其置信水平为 $1-\alpha$ 的置信区间.

解：（1）注意方差已知，由 $N(\mu,1)$ 为指数族, 容易得到 $T(\boldsymbol{X})=\overline{X}$ 为充分完全统计量

设

$$Y_i = X_i - \mu \sim N(0,1)$$

则

$$X_1-\overline{X}=(X_1-\mu)-\frac{\sum\limits_{i=1}^{n}(X_i-\mu)}{n}=Y_1-\overline{Y}$$

分布与 μ 无关，是辅助量，由 Basu 定理知二者独立

（2）显然 $\mathbb{I}_{\{X_1\leq 0\}}$ 就是无偏估计，设要求的 UMVUE 为 $h(T)$

$$\begin{aligned}h(T)&=\mathbb{E}\left[\mathbb{I}_{\{X_1\leq 0\}}\mid T(X)\right]\\&=P\left(X_1\leqslant 0\mid T(X)\right)\\&=P\left(X_1-\overline{X}\leqslant-\overline{X}\mid\overline{X}\right)\\&\overset{\text{独立性}}{=}P\left(X_1-\overline{X}\leqslant-\overline{X}\right)\end{aligned}$$

注意 Y_1 与 $\overline{Y}$ 不独立，有

$$X_1-\overline{X}\sim Y_1-\overline{Y}\sim N\left(1,1-\frac{1}{n}\right)$$

则

$$h(T)=\Phi\left(-\frac{\overline{X}}{\sqrt{1-\frac{1}{n}}}\right)$$

这里 $\Phi(x)$ 为 $N(0,1)$ 的分布函数，又

$$\overline{X}\sim N\left(\mu,\frac{1}{n}\right)\qquad\sqrt{n}(\overline{X}-\mu)\sim N(0,1)$$

则

$$P(X_1\leq 0)=\Phi(-\mu)$$

即 $P(X_1\leq 0)$ 关于 μ 单调递减，只需要求 μ 的置信区间即可另外 μ 有置信系数为 $1-\alpha$ 的置信区间

$$\left[\overline{X}-\frac{u_{\frac{\alpha}{2}}}{\sqrt{n}},\overline{X}+\frac{u_{\frac{\alpha}{2}}}{\sqrt{n}}\right]$$

则 $P(X_1\leq 0)$ 有置信区间

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

四.（15 分）设 $X_1,X_2,\cdots,X_n$ 为来自如下分布的一组简单样本

X	0	1	2
P	$\frac{1}{2}\left[(1-\theta)^2+\theta^2\right]$	$2\theta(1-\theta)$	$\frac{1}{2}\left[(1-\theta)^2+\theta^2\right]$

其中 $0<\theta<1/2$ 为参数。试利用重参数化方法和极大似然估计的不变性

（1）求 θ 的极大似然估计量，并求其渐近分布.

（2）由此给出 θ 的一个（渐近）置信水平为 $1-\alpha$ 的置信区间 $(0<\alpha<1)$.

解：（1）重参数化

$$P(X=1)=P(X=3)\triangleq\mu=\frac{1}{2}\left[(1-\theta)^2+\theta^2\right]$$

则

$$P(X=2)=1-2\mu$$

同时令 n_0,n_1,n_2 为 n 个样本中取值为 0，1，2 的样本个数，有 $n_0+n_1+n_2=n$，则样本联合密度为

$$f\left(x_1,\cdots,x_n;\mu\right)=\mu^{n_0}(1-2\mu)^{n_1}\mu^{n_2}$$

进而

$$\ln f\left(x_1,\cdots,x_n;\mu\right)=n_0\ln\mu+n_1\ln(1-2\mu)+n_2\ln\mu$$
$$\frac{\partial\ln f}{\partial\mu}=\frac{n_0}{\mu}+\frac{n_2}{\mu}-\frac{2n_1}{1-2\mu}=0$$

得到

$$\hat{\mu}_{MLE}=\frac{n_0+n_2}{2n}$$

注意 $\theta<\frac{1}{2}$，反解 θ 有

$$\hat{\mu}_{MLE}=\frac{1}{2}\left[(1-\hat{\theta}_{MLE})^2+\hat{\theta}_{MLE}\right]$$

得到

$$\hat{\theta}_{MLE}=\frac{1-\sqrt{1-\frac{2n_1}{n}}}{2}$$

由 MLE 的渐进正态性

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

这里总体（单个样本）的 Fisher 信息量为

$$\begin{aligned}I(\theta)=&-\mathbb{E}\left[\ln'' f(x;\theta)\right].\\=&-\left[P(x=0)\cdot\ln'' f(x=0;\theta)+P(x=1)\cdot\ln'' f(x=1;\theta)\right.\\&\left.+P(x=2)\cdot\ln'' f\left(x=2;\theta\right)\right]\end{aligned}$$

对 $X=0$ 或 2

$$f(x;\theta)=\frac{(1-\theta)^2+\theta^2}{2}$$

进而

$$\frac{\partial^2}{\partial\theta^2}\ln f=\frac{4\left(2\theta^2-2\theta+1\right)-2(2\theta-1)(4\theta-2)}{\left(2\theta^2-2\theta+1\right)^2}$$

而对 $X=1$

$$\frac{\partial^2}{\partial\theta^2}\ln f(x=1;\theta)=-\frac{1}{\theta^2}-\frac{1}{(1-\theta)^2}$$

则

$$I(\theta)=\frac{2}{2\theta^2-2\theta+1}+\frac{1}{\theta(1-\theta)}$$

就有

$$\sqrt{n}\left(\hat{\theta}_{MLE}-\theta\right)\xrightarrow{D}N\left(0,\frac{1}{\frac{2}{2\theta^2-2\theta+1}+\frac{1}{\theta(1-\theta)}}\right)$$

（2）置信区间为

$$\left[\hat{\theta}_{MLE}-\frac{u_{\frac{\alpha}{2}}}{\sqrt{nI(\theta)}},\hat{\theta}_{MLE}+\frac{u_{\frac{\alpha}{2}}}{\sqrt{nI(\theta)}}\right]$$

五.（20 分）设 $X_1,\ldots,X_n, i.i.d \sim U(0,\theta)$ ，其中 $\theta>1$ 为未知参数．记 $X_{(n)}=\max\limits_{1\le i\le n} X_i$ ，试
（1）证明 $X_{(n)}$ 是充分但不完备的统计量.
（2）求 θ 的 UMVUE.
解：（1）样本联合密度为

$$f\left(x_1,\cdots,x_n;\theta\right)=\frac{1}{\theta^n}\mathbb{I}_{\left\{X_{(n)}<\theta\right\}}$$

取 $g(T(\boldsymbol{X});\theta)=\theta^{-n}\cdot I_{\left\{x_{(n)}<\theta\right\}}, h(\boldsymbol{X})=1$, 由因子分解定理知道 $T(\boldsymbol{X})=X_{(n)}$ 为充分统计量
注意 $\theta>1$，为了证明 $T(\boldsymbol{X})=X_{(n)}$ 不是 θ 的完全统计量．只需寻找 t 的某一实函数 $\varphi(t)$ ，满足 $\mathbb{E}_\theta[\varphi(T)]=0$ ，即

$$\int_0^\theta \varphi(t)t^{n-1}\,\mathrm{d}t=0,\quad \theta>1$$

但 $\mathbb{P}_\theta\left(\varphi\left(X_{(n)}\right)=0\right)<1$
注意到上式对 θ 求导后只能得到 $\varphi(t)=0, t>1$ ．因此我们只要构造合适的 $\varphi(t), t\le 1$ ，使得

$$\int_0^1 \varphi(t)t^{n-1}\,\mathrm{d}t=0$$

即可，例如

$$\varphi(t)=\begin{cases} t^{1-n} & t\le\frac{1}{2}\\ -t^{1-n} & \frac{1}{2}<t\le 1\\ 0 & 1<t\le\theta\end{cases}$$

（2）统计量只充分不完全，考虑零无偏法
$X_{(n)}$ 有密度

$$g(t)=n\theta^{-n}t^{n-1}\mathbb{I}_{(0,\theta)}(t)$$

设 $\mathbb{E}[\delta(t)]=0$，即

$$\int_0^\theta \delta(t)n\theta^{-n}t^{n-1}dt=0$$

也就是

$$\int_0^\theta \delta(t)t^{n-1}dt=0$$

进而

$$\int_0^1 \delta(t)t^{n-1}dt+\int_1^\theta \delta(t)t^{n-1}dt=0$$

求导有

$$\delta(t)\equiv 0\quad \forall t>1$$

设要求的 UMVUE 为 $h(T)$，它要满足以下条件

$$\mathbb{E}[h(T)\delta(T)]=0$$

$$\int_0^\theta \delta(t)h(t)n\theta^{-n}t^{n-1}dt=0$$

注意 $\delta(t) \equiv 0 \quad \forall t > 1$，上式也就说明

$$\int_0^1 \delta(t)h(t)n\theta^{-n}t^{n-1}dt = 0$$

又 $h(T)$ 无偏，即 $\mathbb{E}[h(T)] = \theta$

$$\theta = \int_0^\theta h(t)n\theta^{-n}t^{n-1}dt$$

待定 $h(T) = c, 0 \leq T < 1$，有

$$c\int_0^1 \frac{nt^{n-1}}{\theta^n}dt = c\theta^{-n}$$

想要

$$\int_1^\theta h(t)\frac{nt^{n-1}}{\theta^n}dt$$

含有 θ^{-n} 及 θ 项，待定剩下部分为

$$h(T) = \begin{cases} c & 0 \leq T < 1 \\ bT + d & T \geq 1 \end{cases}$$

代入积分的值则有

$$c = 1 \quad b = \frac{n+1}{n} \quad d = 0$$

最后

$$h(T) = \begin{cases} 1 & 0 \leq T < 1 \\ \frac{n+1}{n}T & T \geq 1 \end{cases}$$

Chapter 2

往年期末试卷

2.1 25 数理统计期末

一、填空及判断题（15 分）

1. 设 $X_1, \cdots, X_n, X_{n+1} \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$，若

$$\frac{\overline{X_n} - X_{n+1}}{\sqrt{b_n \sum_{i=1}^n X_i^2 - c_n(\overline{X_n})^2}} \sim t_{n-1}$$

其中 $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$，则 $b_n =$__________，$c_n =$__________。

答案：$b_n = \frac{n+1}{n(n-1)} \quad c_n = \frac{n+1}{n-1}$

先将分子归一化为 $N(0,1)$

$$\overline{X_n} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \overline{X_n} - X_{n+1} \sim N\left(0, \frac{(n+1)\sigma^2}{n}\right)$$

则

$$\frac{\overline{X_n} - X_{n+1}}{\sqrt{\frac{(n+1)\sigma^2}{n}}} \sim N(0,1)$$

再处理分母

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

其中

$$\begin{aligned}\sum_{i=1}^n \left(X_i - \overline{X_n}\right)^2 &= \sum_{i=1}^n X_i^2 - 2\overline{X_n}\sum_{i=1}^n X_i + n\overline{X_n}^2 \\ &= \sum_{i=1}^n X_i^2 - n\overline{X_n}^2\end{aligned}$$

则

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

也就是

$$\frac{\frac{\overline{X_n}-X_{n+1}}{\sqrt{\frac{(n+1)\sigma^2}{n}}}}{\sqrt{\frac{\sum\limits_{i=1}^{n}X_i^2-n\overline{X_n}^2}{(n-1)\sigma^2}}}\sim t_{n-1}$$

其中分子分母的独立性来自正态总体的 $\overline{X_n}$ 与 S^2 独立

化简得到

$$\frac{\overline{X_n}-X_{n+1}}{\sqrt{\frac{n+1}{n(n-1)}\sum_{i=1}^{n}X_i^2-\frac{n+1}{n-1}(\overline{X_n})^2}}\sim t_{n-1}$$

就有

$$b_n=\frac{n+1}{n(n-1)}\quad c_n=\frac{n+1}{n-1}$$

2. 设 $X\sim N(\mu,2)$，求 μ 的 95% 置信区间长度小于 1 所需的最小样本量__________。

答案：$n\geq 31$

已知方差，则置信区间为

$$\left[\overline{X}-\sqrt{\frac{2}{n}}u_{\alpha/2},\overline{X}+\sqrt{\frac{2}{n}}u_{\alpha/2}\right]$$

代入数值即得

3.×

4.×

5.×

6.✓;✓;×

7. 样本量小导致有些格子计数小于 5

二、（15 分）证明来自总体分布 $U(\theta,2\theta)$ 的 n 个样本 $(X_1,...,X_n)$ 的统计量 $(X_{(1)},X_{(n)})$ 是极小充分统计量但不是完全统计量。

解：均匀总体下的充分完全统计量.

通过因子分解定理说明充分统计量：根据题意有样本联合密度

$$f(\boldsymbol{x};\theta)=\frac{\mathbb{I}_{\{\theta<x_i<2\theta,i=1,2,\ldots,n\}}}{\theta^n}=\theta^{-n}\mathbb{I}_{\left\{x_{(n)}/2<\theta<x_{(1)}\right\}}.$$

因此由因子分解定理知 $T=\left(X_{(1)},X_{(n)}\right)$ 是 θ 的充分统计量.

通过定理 2.6.2 说明极小充分统计量：对任意 $\boldsymbol{x}$ 和 $\boldsymbol{y}$，

$$\frac{f(\boldsymbol{x};\theta)}{f(\boldsymbol{y};\theta)}=\frac{\mathbb{I}_{\left\{x_{(n)}/2<\theta<x_{(1)}\right\}}}{\mathbb{I}_{\left\{y_{(n)}/2<\theta<y_{(1)}\right\}}}$$

要使得上式与 θ 无关，当且仅当

$$\left(x_{(1)},x_{(n)}\right)=\left(y_{(1)},y_{(n)}\right)$$

因此由定理 2.6.2 知 $T=\left(X_{(1)},X_{(n)}\right)$ 是 θ 的极小充分统计量.

构造函数或利用辅助统计量说明不完全性：

法一: 注意到

$$\mathbb{E}_\theta\left[X_{(1)}\right]=\theta+\frac{1}{n+1}\theta=\frac{n+2}{n+1}\theta,\quad\mathbb{E}_\theta\left[X_{(n)}\right]=\theta+\frac{n}{n+1}\theta=\frac{2n+1}{n+1}\theta,$$

上式来自 n 个独立同分布的 U(0,1) 的次序统计量满足

$$U_{(1)} \sim \beta(1,n) \quad U_{(n)} \sim \beta(n,1)$$

取 $g\left(x_{(1)}, x_{(n)}\right) = (2n+1)x_{(1)} - (n+2)x_{(n)}$ ，则

$$\mathbb{E}_\theta\left[g\left(X_{(1)}, X_{(n)}\right)\right] = 0, \quad \forall\theta > 0,$$

但函数 g 并不几乎处处为零。因此 $\left(X_{(1)}, X_{(n)}\right)$ 不是 θ 的完全统计量.
法二: 注意到 $X_i/\theta \sim \mathrm{U}(1,2)$ ，构造辅助统计量

$$T = \frac{X_{(n)}}{X_{(1)}} = \frac{X_{(n)}/\theta}{X_{(1)}/\theta},$$

于是 T 的分布与 θ 无关，是一个辅助统计量．因此 $\left(X_{(1)}, X_{(n)}\right)$ 不是 θ 的完全统计量.

三、（15 分）设 X 的概率分布为：

X	1	2	3
$P(X)$	2θ	3θ	$1-5\theta$

取出 20 个样本，其中有 6 个取值为 1, 有 8 个取值为 2, 有 6 个取值为 3
（1）求 θ 的最大似然估计和矩估计，并判断该估计是否无偏；
（2）求 θ 的一致最小方差无偏估计 (UMVUE)，并比较其方差与 C-R 下界。
解：（1）矩估计：根据题意有

$$\mathbb{E}[X] = 1 \times 2\theta + 2 \times 3\theta + 3 \times (1-5\theta) = 3 - 7\theta.$$

因此反解得 θ 的矩估计量为

$$\hat{\theta}_M(\boldsymbol{X}) = \frac{3-\overline{X}}{7}$$

由 $\overline{x} = \frac{1\times6+2\times8+3\times6}{20} = 2$ 知 θ 的矩估计值为

$$\hat{\theta}_M(\boldsymbol{x}) = 1/7$$

而

$$\mathbb{E}_\theta\left[\hat{\theta}_M(\boldsymbol{X})\right] = \mathbb{E}_\theta\left(\frac{3-\overline{X}}{7}\right) = \frac{3-\mathbb{E}(\overline{X})}{7} = \theta.$$

所以矩估计 $\hat{\theta}_M(\boldsymbol{X})$ 是 θ 的无偏估计.
最大似然估计：记 $n_j = \sum_{i=1}^n \mathbb{I}_{(X_i=j)}, j=1,2,3$ 。由分布列知 θ 的似然函数为

$$L(\theta) = (2\theta)^{n_1}(3\theta)^{n_2}(1-5\theta)^{n_3} = 2^{n_1}3^{n_2}\theta^{n-n_3}(1-5\theta)^{n_3}.$$

由对数似然方程

$$\frac{\partial \log L(\theta)}{\partial\theta} = \frac{n-n_3}{\theta} - \frac{5n_3}{1-5\theta} = 0 \Rightarrow \theta = \frac{n-n_3}{5n}.$$

经微分检验其确为似然函数的最大值点，因此 θ 的最大似然估计为

$$\hat{\theta}_L(\boldsymbol{X}) = \frac{1}{5n}\sum_{i=1}^n \mathbb{I}_{(X_i \neq 3)}$$

由样本观测值知其估计值为

$$\hat{\theta}_L(\boldsymbol{x}) = 0.14$$

而

$$\mathbb{E}_\theta\left[\hat{\theta}_L(\boldsymbol{X})\right] = \mathbb{E}_\theta\left(\frac{1}{5n}\sum_{i=1}^n \mathbb{I}_{(X_i\neq 3)}\right) = \frac{1}{5n}\sum_{i=1}^n (5\theta) = \theta.$$

所以最大似然估计 $\hat{\theta}_L(\boldsymbol{X})$ 也是 θ 的无偏估计.

（2）一致最小方差无偏估计：样本联合密度函数为

$$f(\boldsymbol{x};\theta) = (2\theta)^{n_1}(3\theta)^{n_2}(1-5\theta)^{n_3} = 2^{n_1}3^{n_2}\theta^n \exp\left\{n_3 \log\frac{1-5\theta}{\theta}\right\}$$

自然参数空间 $\Theta^* = \{\eta : -\infty < \eta < \infty\}$ 有内点，于是 $T(\boldsymbol{X}) = n_3$ 是 θ 的充分完全统计量. 从而由 Lehmann-Scheffé 定理知，$\hat{\theta}_L(\boldsymbol{X})$ 作为基于 n_3 的无偏估计是 θ 的 UMVUE，其方差

$$\mathrm{Var}_\theta\left[\hat{\theta}_L(\boldsymbol{X})\right] = \frac{\mathrm{Var}_\theta\left(I\left(X_1\neq 3\right)\right)}{25n} = \frac{5\theta(1-5\theta)}{25n} = \frac{\theta(1-5\theta)}{5n}.$$

而 n 个样本的 Fisher 信息量

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 \log L(\theta)}{\partial\theta^2}\right] = \mathbb{E}_\theta\left[\frac{n-n_3}{\theta^2} - \frac{25n_3}{(1-5\theta)^2}\right] = \frac{5n\theta}{\theta^2} - \frac{25n(1-5\theta)}{(1-5\theta)^2} = \frac{5n}{\theta(1-5\theta)}.$$

因而 Cramér-Rao 下界为

$$1/I(\theta) = \mathrm{Var}_\theta\left[\hat{\theta}_L(\boldsymbol{X})\right]$$

即一致最小方差无偏估计 $\hat{\theta}_L(\boldsymbol{X})$ 的方差达到 θ 的无偏估计方差的下界.

四、（12 分）设 $(X_1,...,X_m)$ 来自某总体分布；$(Y_1,...,Y_n)$ 来自另外的总体分布；设 $\overline{X} = 70$，$\overline{Y} = 80$，样本容量 $n = m = 26$，样本方差 $S_1^2 = 10$，$S_2^2 = 12$。

（1）假设总体服从正态分布：判断两总体方差是否相同；判断两总体均值是否相同。

（2）若去掉正态性假设，检验两总体均值是否相同

解：（1）

i. 两正态总体方差齐性检验：假设检验问题为

$$H_0 : \sigma_1^2 = \sigma_2^2 \longleftrightarrow H_1 : \sigma_1^2 \neq \sigma_2^2.$$

取检验统计量为

$$F(\boldsymbol{X},\boldsymbol{Y}) = \frac{S_2^2}{S_1^2} \overset{H_0}{\sim} F_{n-1,m-1}$$

由 F 检验知水平 $\alpha = 0.2$ 的检验拒绝域为

$$D = \left\{(\boldsymbol{X},\boldsymbol{Y}) : F(\boldsymbol{X},\boldsymbol{Y}) > F_{25,25}(0.1)\ 或 F(\boldsymbol{X},\boldsymbol{Y}) < F_{25,25}(0.9) = \frac{1}{F_{25,25}(0.1)}\right\}.$$

现有

$$F(\boldsymbol{x},\boldsymbol{y}) = s_2^2/s_1^2 = 1.2 \in (1/1.68, 1.68)$$

所以不能拒绝原假设，可认为方差齐性.

ii. 两正态总体均值差的检验：假设检验问题为

$$H_0 : \mu_1 = \mu_2 \longleftrightarrow H_1 : \mu_1 \neq \mu_2.$$

由于我们认为方差齐性，所以可取检验统计量为

$$T(\boldsymbol{X},\boldsymbol{Y})=\frac{\overline{Y}-\overline{X}}{S_w\sqrt{1/m+1/n}}\overset{H_0}{\sim}t_{m+n-2}$$

其中 $(m+n-2)S_w^2=(m-1)S_1^2+(n-1)S_2^2$ ．由 t 检验知水平 $\alpha=0.05$ 的检验拒绝域为

$$D=\{(\boldsymbol{X},\boldsymbol{Y}):|T(\boldsymbol{X},\boldsymbol{Y})|>t_{50}(0.025)\}$$

现有

$$T(\boldsymbol{x},\boldsymbol{y})=\frac{80-70}{\sqrt{11}\times\sqrt{1/13}}=10.9>2.01$$

所以拒绝原假设，认为两班学生成绩均值有显著差异.

（2）两一般总体均值差的检验：假设检验问题仍为

$$H_0:\mu_1=\mu_2\longleftrightarrow H_1:\mu_1\neq\mu_2.$$

此时可用大样本检验，取检验统计量为

$$Z(\boldsymbol{X},\boldsymbol{Y})=\frac{\overline{Y}-\overline{X}}{\sqrt{S_1^2/m+S_2^2/n}}\xrightarrow{H_0}N(0,1),\quad \text{as } m,n\to\infty.$$

由 z 检验知水平 $\alpha=0.05$ 的检验拒绝域为

$$D=\{(\boldsymbol{X},\boldsymbol{Y}):|Z(\boldsymbol{X},\boldsymbol{Y})|>u_{0.025}\}.$$

现有

$$Z(\boldsymbol{x},\boldsymbol{y})=\frac{80-70}{\sqrt{10/26+12/26}}=10.9>1.96$$

所以拒绝原假设，认为两班学生成绩均值有显著差异.

五、（15 分）设有 n 个样本 $(X_1,...,X_n)$ 来自 Beta$(2,\beta)$ 分布。检验假设：

$$H_0:\beta=1\quad\leftrightarrow\quad H_1:\beta\neq 1$$

求似然比检验，Wald 检验和得分检验。

解：伽马总体下的三大检验.

似然比检验：由题意知似然函数

$$L(\beta)=\prod_{i=1}^n\beta^2x\exp\{-\beta x\}=\beta^{2n}\exp\left\{-\beta\sum_{i=1}^n x_i\right\}\prod_{i=1}^n x_i,\quad \beta>0$$

由对数似然方程

$$\frac{\partial\log L(\beta)}{\partial\beta}=\frac{2n}{\beta}-\sum_{i=1}^n x_i=0\Rightarrow\beta=\frac{2n}{\sum_{i=1}^n x_i}$$

经微分检验其确为似然函数的最大值点，因此 β 的最大似然估计为 $\hat\beta=2/\overline{X}$ ．于是似然比检验统计量

$$\Lambda(\boldsymbol{X})=\frac{\sup_{\beta>0}L(\beta)}{\sup_{\beta=1}L(\beta)}=\frac{L(\hat\beta)}{L(1)}=\frac{(2/\overline{X})^{2n}\exp\{-2n\}}{\exp\{-\overline{X}\}}=(2/\overline{X})^{2n}\exp\{\overline{X}-2n\}$$

由似然比检验统计量的极限分布知

$$2\log\Lambda(\boldsymbol{X}) = 4n\log\frac{2}{\overline{X}} + 2\overline{X} - 4n \xrightarrow{H_0} \chi_1^2, \quad \text{as } n\to\infty$$

因此水平为 α 的似然比检验的拒绝域为

$$D = \left\{\boldsymbol{X} : 4n\log\frac{2}{\overline{X}} + 2\overline{X} - 4n > \chi_1^2(\alpha)\right\}.$$

Wald 检验：由似然函数计算 n 个样本的 Fisher 信息量为

$$I_n(\beta) = -\mathbb{E}_\beta\left[\frac{\partial^2\log L(\beta)}{\partial\beta^2}\right] = \frac{2n}{\beta^2}$$

所以 Wald 检验统计量为

$$W(\boldsymbol{X}) = (\hat\beta - 1)^2 I_n(\hat\beta) = 2n(1-1/\hat\beta)^2 = 2n(1-\overline{X}/2)^2 \xrightarrow{H_0} \chi_1^2, \quad \text{as } n\to\infty.$$

因此水平为 α 的 Wald 检验的拒绝域为

$$D = \left\{\boldsymbol{X} : 2n(1-\overline{X}/2)^2 > \chi_1^2(\alpha)\right\}$$

得分检验：得分函数为

$$U_n(\beta) = \frac{\partial\log L(\beta)}{\partial\beta} = \frac{2n}{\beta} - \sum_{i=1}^n x_i.$$

所以得分检验统计量为

$$S(\boldsymbol{X}) = [U_n(1)]^2 I_n^{-1}(1) = \left(2n - \sum_{i=1}^n X_i\right)^2/(2n) = n(2-\overline{X})^2/2 \xrightarrow{H_0} \chi_1^2, \quad \text{as } n\to\infty.$$

因此水平为 α 的得分检验的拒绝域为

$$D = \left\{\boldsymbol{X} : n(2-\overline{X})^2/2 > \chi_1^2(\alpha)\right\}$$

该形式与 Wald 检验的形式是一致的。

六、（10 分）男女舒张压检测数据如下：共检测男性 16 人，其中舒张压 <60 的有 4 人，>90 的有 2 人；共检测女性 21 人，其中舒张压 <60 的有 5 人，>90 的有 2 人。

是否能认为男女舒张压分布相同？

解：未合并列的齐一性检验（酌情扣分）：首先根据题目描述写出列联表如下：

舒张压	<60	$60\sim 90$	>90	合计
男性人数	4	10	2	16
女性人数	5	14	2	21
合计	9	24	4	37

要检验的假设为 H_0 ：男女性舒张压分布没有显著差异．取检验统计量为

$$K(\boldsymbol{X}) = n\sum_{i=1}^r\sum_{j=1}^s\frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}} \xrightarrow{H_0} \chi^2_{(r-1)(s-1)}, \quad \text{as } n\to\infty.$$

因此水平 α 的检验拒绝域为

$$D=\left\{\boldsymbol{X}:K(\boldsymbol{X})>\chi^2_{(r-1)(s-1)}(\alpha)\right\}$$

由样本观测值计算检验统计量值时，可以考虑等价形式

$$K(\boldsymbol{x})=n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}-1\right)=0.1040<\chi^2_2(0.2)=3.22.$$

因此不拒绝原假设，认为男女性舒张压没有显著差异.
合并列的齐一性检验：由于有格子点计数过小，需要合并首尾两列得列联表如下：

舒张压	正常	过低或过高	合计
男性人数	10	6	16
女性人数	14	7	21
合计	24	13	37

要检验的假设为 H_0 ：男女性舒张压分布没有显著差异．取检验统计量为

$$K(\boldsymbol{X})=n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(n_{ij}-n_i\cdot n_{\cdot j}/n\right)^2}{n_i\cdot n_{\cdot j}}\xrightarrow{H_0}\chi^2_{(r-1)(s-1)},\qquad \text{as } n\to\infty.$$

因此水平 α 的检验拒绝域为

$$D=\left\{\boldsymbol{X}:K(\boldsymbol{X})>\chi^2_{(r-1)(s-1)}(\alpha)\right\}$$

由样本观测值计算检验统计量值时，可以考虑 2×2 列联表的等价形式

$$K(\boldsymbol{x})=\frac{n\left(n_{11}n_{22}-n_{12}n_{21}\right)^2}{n_{1\cdot}n_{\cdot 1}n_{2\cdot}n_{\cdot 2}}=0.0692<\chi^2_1(0.2)=1.64.$$

因此不拒绝原假设，认为男女性舒张压没有显著差异.

七、（18 分）n 个样本 $(X_1,...,X_n)$ 来自总体分布 $X\sim\Gamma\left(\alpha,\frac{1}{\theta}\right)$，其中 α 已知，先验密度为 $\pi(\theta)=\frac{1}{\theta}$ ，损失函数为

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

求贝叶斯解并证明其为 Minimax 解

解：后验分布的计算：样本 $\boldsymbol{X}=(X_1,\ldots,X_n)$ 的联合概率密度函数为

$$f(\boldsymbol{x}\mid\theta)=\prod_{i=1}^{n}\frac{1}{\theta^\alpha\Gamma(\alpha)}x_i^{\alpha-1}\exp\left\{-\frac{x_i}{\theta}\right\}=\frac{\theta^{-n\alpha}}{[\Gamma(\alpha)]^n}\exp\left\{-\frac{1}{\theta}\sum_{i=1}^{n}x_i\right\}\prod_{i=1}^{n}x_i^{\alpha-1}$$

因此在无信息先验 $\pi(\theta)=(1/\theta)I_{(0,\infty)}(\theta)$ 下，θ 的后验密度为

$$\pi(\theta\mid\boldsymbol{x})\propto f(\boldsymbol{x}\mid\theta)\pi(\theta)\propto\theta^{-n\alpha-1}\exp\left\{-\frac{1}{\theta}\sum_{i=1}^{n}x_i\right\},\quad \theta>0$$

添加归一化常数后可知 θ 的后验分布为

$$\theta\mid\boldsymbol{X}=\boldsymbol{x}\sim\Gamma^{-1}\left(n\alpha,\sum_{i=1}^{n}x_i\right)$$

加权平方损失下的 Bayes 估计：在加权平方损失下，θ 的 Bayes 估计为

$$\hat{\theta}_B(\boldsymbol{X}) = \frac{\mathbb{E}\left(\theta^{-1} \mid \boldsymbol{X}\right)}{\mathbb{E}\left(\theta^{-2} \mid \boldsymbol{X}\right)} = \frac{\frac{n\alpha}{n\overline{X}}}{\frac{n\alpha(n\alpha+1)}{(n\overline{X})^2}} = \frac{n\overline{X}}{n\alpha+1}$$

RK：在矩存在的条件下有

$$X \sim \Gamma(\alpha,\beta) \quad 有\forall n \in \mathbb{Z} \quad \mathbb{E}[X^n] = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)\beta^n}$$

$$Y \sim \Gamma^{-1}(\alpha,\beta) \quad 有\forall n \in \mathbb{Z} \quad \mathbb{E}[Y^n] = \frac{\Gamma(\alpha-n)\beta^n}{\Gamma(\alpha)}$$

事实上，Γ 分布的 n 阶矩就是 Γ^{-1} 分布的 $-n$ 阶矩
验证该 Bayes 估计为 Minimax 估计：Bayes 估计 $\hat{\theta}_B(\boldsymbol{X})$ 的风险函数，注意 $n\overline{X} = \sum_{i=1}^{n} X_i \sim \Gamma\left(n\alpha, \frac{1}{\theta}\right)$

$$\begin{aligned}
R\left(\hat{\theta}_B(\boldsymbol{X}), \theta\right) &= \mathbb{E}\left[\frac{(n\overline{X}/(n\alpha+1)-\theta)^2}{\theta^2}\right] \\
&= \frac{1}{\theta^2}\mathbb{E}\left(\frac{n\overline{X}+\theta}{n\alpha+1} - \theta - \frac{\theta}{n\alpha+1}\right)^2 \\
&= \frac{1}{\theta^2}\left[\operatorname{Var}\left(\frac{n\overline{X}+\theta}{n\alpha+1}\right) + \frac{\theta^2}{(n\alpha+1)^2}\right] \\
&= \frac{1}{n\alpha+1}
\end{aligned}$$

为常数，所以 θ 的 Bayes 估计 $\hat{\theta}_B(\boldsymbol{X}) = n\overline{X}/(n\alpha+1)$ 是 θ 的 Minimax 估计.

八、（20 分，附加题）将 Hardy-Weinberg 定律简化如下：n 个样本 $(X_1, ..., X_n)$ 来自总体分布 X
设 X 的概率分布为：

X	1	2	3
$P(X)$	p^2	$2p(1-p)$	$(1-p)^2$

对于检验问题：

$$H_0 : p \leq p_0 \quad \leftrightarrow \quad H_1 : p > p_0$$

求 UMPT。
解：先说明样本分布族是单参数指数族：样本联合概率质量函数为

$$\begin{aligned}
f(\boldsymbol{x}; p) &= \left(p^2\right)^{n_0} [2p(1-p)]^{n_1} \left[(1-p)^2\right]^{n_2} \\
&= 2^{n_1} p^{2n_0+n_1} (1-p)^{n_1+2n_2} \\
&= 2^{n_1}(1-p)^{2n} \exp\left\{(2n_0+n_1)\log\frac{p}{1-p}\right\}
\end{aligned}$$

这是单参数指数族，且 $Q(p) = \log\frac{p}{1-p}$ 为 p 的严格单调增函数，$T = T(\boldsymbol{X}) = 2n_0 + n_1$.
根据推论 5.4.2 给出检验的 UMPT：注意到原检验问题等价于

$$H_0 : \theta \leq \theta_0 \longleftrightarrow H_1 : \theta > \theta_0$$

其中 $\theta_0 = \log \frac{p_0}{1-p_0}$. 由推论 5.4.2 知，离散型要补上随机化常数，假设检验的 UMPT 可取为

$$\phi(T) = \begin{cases} 1, & T > c \\ r, & T = c \\ 0, & T < c \end{cases}$$

其中常数 c 和 r 满足条件

$$c = \arg\min_{c'} \left\{ \mathrm{P}_{p_0}(T > c) \leq \alpha \right\}, \quad r = \frac{\alpha - \mathrm{P}_{p_0}(T > c)}{\mathrm{P}_{p_0}(T = c)}.$$

确定检验函数中的常数 c 和 r ：记

$$p_t = \sum_{1 \leq i, j \leq n, i+j \leq n, 2i+j=t} \frac{n!}{i!j!(n-i-j)!} 2^i p_0^{2i+j} \left(1 - p_0\right)^{j+2(n-i-j)},$$

则

$$\mathrm{P}_{p_0}(T = c) = p_c, \quad \mathrm{P}_{p_0}(T > c) = \sum_{t=c+1}^{n} p_t.$$

2.2 24 数理统计期末

一.（20 分）单项选择填空题（每题 2 分）

1.设 $X_1,...,X_n$ 为来自均匀分布 $\mathrm{U}(-\theta,\theta)$ 的一组样本,θ 为未知参数,则下述量为统计量的是__________

（A）$\overline{X}-\theta$

（B）$\max_{1\leq i\leq n}(X_i-\theta)-\min_{1\leq i\leq n}(X_i-\theta)$

（C）$\max_{1\leq i\leq n}(X_i-\theta)$

（D）$\min_{1\leq i\leq n}(X_i-\theta)$

答案：B

统计量要与未知参数无关

2．设 $\hat{\theta}_n$ 为末知参数 θ 的一个估计量，如果 $\lim_{n\to\infty}\mathbb{E}\left[\hat{\theta}_n-\theta\right]=0$ ，则 $\hat{\theta}_n$ 为 θ 的__________

（A）无偏估计

（B）有效估计

（C）相合估计

（D）渐近正态估计

答案：C

3．假设样本 X 的密度为 $f_\theta(x)$ ，其中 θ 为参数，则下列表述不正确的是__________

（A）固定 x 时 $f_\theta(x)$ 为似然函数

（B）固定 θ 时 $f_\theta(x)$ 为似然函数

（C）固定 θ 时 $f_\theta(x)$ 为密度函数

（D）$f_\theta(x)$ 衡量了不同 θ 下观测到值 x 的可能性大小

答案：B

4．一个参数 θ 的 95% 区间估计为 $[0.1,0.3]$ ，则下列表述正确的是__________

（A）若该区间为置信区间，则表明 θ 位于该区间的概率是 0.95

（B）该区间的边际误为 0.2

（C）对假设 $H_0:\theta=0.2\leftrightarrow H_1:\theta\neq0.2$ ，会在 0.05 水平下拒绝原假设

（D）若该区间为贝叶斯可信区间，则表明 θ 位于该区间的概率是 0.95

答案：D

5．下列表述错误的是__________

（A）矩估计量一般不唯一

（B）无偏估计总是优于有偏估计

（C）相合性是一个估计量的基本性质

（D）最大似然估计可以不存在

答案：B

6．若 $\delta(X)$ 是一个损失下的 Bayes 法则，则下列表述正确的是__________

（A）$\delta(X)$ 的贝叶斯风险不超过 Minimax 风险

（B）$\delta(X)$ 不可能是一个 Minimax 法则

（C）$\delta(X)$ 是可容许的

（D）$\delta(X)$ 的风险为常数

答案：A

7．下述对一个显著性检验方法的描述错误的是__________

（A）原假设与对立假设地位不均等，原假设被保护起来

（B）p 值越显著表明原假设成立的依据越强烈

（C）在一个检验结果是不能拒绝零假设时，检验只可能会犯第二类错误

（D）双边假设的接受域等价于参数的置信区间

答案：B

8．设 $X_1,\ldots,X_n$ 为来自正态总体 $N(\mu,1)$ 的简单样本，考虑假设检验问题 $H_0:\mu=0\leftrightarrow H_1：\mu=0.5$ 。如果要求检验的第一类和第二类错误均不超过 $\alpha(0<\alpha<1)$ ，则样本量 n 应满足 $\underline{n\geq\lceil 16u_\alpha^2\rceil}$ （结果用分位数表示）.

解：两点假设的拒绝域形如 $R=\{\boldsymbol{X}:\overline{X}>c\}$ ．按要求

$$\begin{aligned}
&\alpha\geq \mathrm{P}\left(\boldsymbol{X}\in R\mid H_0\right)=\mathrm{P}_{\mu=0}(\overline{X}>c)=1-\Phi(\sqrt{n}c),\\
&\alpha\geq \mathrm{P}\left(\boldsymbol{X}\notin R\mid H_1\right)=\mathrm{P}_{\mu=0.5}(\overline{X}\leq c)=\Phi(\sqrt{n}(c-0.5)).
\end{aligned}$$

于是我们有

$$\left\{\begin{array}{l}\sqrt{n}c\geq u_\alpha,\\ \sqrt{n}(c-0.5)\leq -u_\alpha,\end{array}\right.\quad\Longrightarrow\quad 0.5\sqrt{n}\geq 2u_\alpha,\quad\Longrightarrow\quad n\geq\lceil 16u_\alpha^2\rceil.$$

9．设某种产品的质量等级可以划分为 " 优 " 、 " 合格 " 和 " 不合格 " ，为了判断生产此产品的三家工厂的产品是否有差异，使用拟合优度检验方法时的原假设为 <u>三家工厂生产的产品质量无差异</u>，渐近卡方分布的自由度为 <u>4</u>.

10. 设 $X_1,\ldots,X_n$ 为来自均匀分布 $\mathrm{U}(0,\theta),\theta>0$ 的一组简单样本，θ 的先验密度为 $\pi(\theta)=1/\left(2\theta^2\right),\theta\geq 1/2$ 。考虑假设检验问题 $H_0:\theta\leq 1\leftrightarrow H_1:\theta>1$ ，则其 Bayes 因子 BF_{01} 为 $\underline{\left[\left(x_{(n)}\vee 0.5\right)^{-n-1}-1\right]\vee 0}$

解：样本联合密度与先验分别为

$$f(\boldsymbol{x}\mid\theta)=\theta^{-n}\cdot\mathbb{I}_{\left\{0<x_{(n)}<\theta\right\}},\quad \pi(\theta)=0.5\theta^{-2}\cdot\mathbb{I}_{\{\theta\geq 0.5\}}$$

因此 θ 的后验密度为

$$\pi(\theta\mid\boldsymbol{x})\propto\theta^{-n-2}\cdot\mathbb{I}_{\left\{\theta>x_{(n)}\vee 0.5\right\}}$$

归一化后可得后验密度，进而求得后验分布函数为

$$\Pi(\theta\mid\boldsymbol{x})=\begin{cases}1-\left(\frac{\theta_*}{\theta}\right)^{n+1}, & \theta\geq\theta_*\\ 0, & \theta<\theta_*\end{cases}$$

其中 $\theta_*=x_{(n)}\vee 0.5$ ．于是当 $\theta_*<1$ ，即 $x_{(n)}<1$ 时，

$$\alpha_0=\mathrm{P}(\theta\leq 1\mid\boldsymbol{x})=\Pi(1\mid\boldsymbol{x})=1-\theta_*^{n+1},\quad \alpha_1=1-\alpha_0=\theta_*^{n+1}$$

当 $\theta_*\geq 1$ ，即 $x_{(n)}\geq 1$ 时，$\alpha_0=0,\alpha_1=1$ ．又因为 $\pi_0=\mathrm{P}(\theta\leq 1)=0.5,\pi_1=\mathrm{P}(\theta>1)=0.5$ ，所以贝叶斯因子为

$$\mathrm{BF}_{01}=\frac{\alpha_0/\alpha_1}{\pi_0/\pi_1}=\left\{\begin{array}{ll}\theta_*^{-n-1}-1, & \theta_*<1,\\ 0, & \theta_*\geq 1\end{array}\right.=\left[\left(x_{(n)}\vee 0.5\right)^{-n-1}-1\right]\vee 0$$

二.（20 分）设从总体

X	0	1	2
P	p_1	p_2	p_3

（其中 $0<p_1,p_2,p_3<1,p_1+p_2+p_3=1$ 为末知参数）中抽取的一个简单样本 $X_1,\ldots,X_n$ ，试
（1）求 p_1-p_2 的最大似然估计，并证明其为最小方差无偏估计。
（2）求检验问题 $H_0:p_1=p_2\leftrightarrow H_1:p_1\neq p_2$ 的一个（渐近）水平 α 检验。
解：（1）似然函数为

$$L\left(p_1,p_2;\boldsymbol{x}\right)=p_1^{n_0}p_2^{n_1}\left(1-p_1-p_2\right)^{n-n_0-n_1}$$

其中 $n_i=\sum_{j=1}^n\mathbb{I}_{\{X_j=i\}},i=0,1$ ．由对数似然方程

$$\begin{cases}\frac{\partial l(p_1,p_2;\boldsymbol{x})}{\partial p_1}=\frac{n_0}{p_1}-\frac{n-n_0-n_1}{1-p_1-p_2}=0\\ \frac{\partial l(p_1,p_2;\boldsymbol{x})}{\partial p_2}=\frac{n_1}{p_2}-\frac{n-n_0-n_1}{1-p_1-p_2}=0\end{cases}$$

解得 p_1,p_2 的最大似然估计分别为

$$\hat{p}_1=\frac{n_0}{n},\quad \hat{p}_2=\frac{n_1}{n}$$

进一步由最大似然估计的不变性可知 p_1-p_2 的最大似然估计为 $\hat{p}_1-\hat{p}_2=\left(n_0-n_1\right)/n$.
最小方差无偏估计：将样本联合密度函数写成指数族形式如下

$$f\left(\boldsymbol{x};p_1,p_2\right)=\exp\left\{n_0\log\frac{p_1}{1-p_1-p_2}+n_1\log\frac{p_2}{1-p_1-p_2}\right\}\cdot\left(1-p_1-p_2\right)^n$$

令

$$\eta_1=\log\frac{p_1}{1-p_1-p_2},\eta_2=\log\frac{p_2}{1-p_1-p_2}$$

于是自然参数空间

$$\Theta^*=\{(\eta_1,\eta_2):-\infty<\eta_1<\infty,-\infty<\eta_2<\infty\}$$

有内点，因此 $T=(n_0,n_1)$ 是 (p_1,p_2) 的充分完全统计量．又注意到 $\hat{p}_1$ 和 $\hat{p}_2$ 分别是 p_1 和 p_2 的无偏估计，因此由 Lehmann-Scheffé 定理知 p_1-p_2 的最大似然估计是最小方差无偏估计．
（2）法一：拟合优度检验，取检验统计量为

$$K(\boldsymbol{X})=\sum_{r=1}^3\frac{\left(n_{r-1}-n\hat{p}_r\right)^2}{n\hat{p}_r}\xrightarrow{H_0}\chi^2_{3-1-1},$$

其中 $\hat{p}_r$ 为 H_0 下的极大似然估计．注意到当 $p_1=p_2=p$ 时，样本的似然函数为

$$L(p;\boldsymbol{x})=p^{n_0+n_1}(1-2p)^{n_2}.$$

由对数似然方程可得 $\hat{p}=\left(n_0+n_1\right)/(2n)$ ．代入检验统计量表达式得

$$K(\boldsymbol{X})=\frac{\left(n_0-n_1\right)^2}{n_0+n_1}.$$

因此检验问题渐近水平 α 检验为

$$\phi(\boldsymbol{X})=\begin{cases}1, & \text{当}\left(n_0-n_1\right)^2>\left(n_0+n_1\right)\chi_1^2(\alpha),\\ 0, & \text{当}\left(n_0-n_1\right)^2\leq\left(n_0+n_1\right)\chi_1^2(\alpha).\end{cases}$$

法二：似然比检验，注意到似然比

$$\lambda(\boldsymbol{X})=\frac{\sup_{\theta\in\Theta}L(\theta)}{\sup_{\theta\in\Theta_0}L(\theta)}=\frac{\hat{p}_1^{n_0}\hat{p}_2^{n_1}\left(1-\hat{p}_1-\hat{p}_2\right)^{n_2}}{\hat{p}^{(n_0+n_1)}(1-2\hat{p})^{n_2}}$$

在大样本下，我们有 $2\log\lambda(\boldsymbol{X})\xrightarrow{H_0}\chi_1^2$. 代入检验统计量表达式得

$$2\log\lambda(\boldsymbol{X})=2\log\frac{\left(n_0/n\right)^{n_0}\cdot\left(n_1/n\right)^{n_1}}{\left(n_0+n_1\right)^{n_0+n_1}/(2n)^{n_0+n_1}}=2n_0\log\frac{2n_0}{n_0+n_1}+2n_1\log\frac{2n_1}{n_0+n_1}$$

因此检验问题渐近水平 α 检验为

$$\phi(\boldsymbol{X})=\begin{cases}1, & \text{当}2n_0\log\frac{2n_0}{n_0+n_1}+2n_1\log\frac{2n_1}{n_0+n_1}>\chi_1^2(\alpha),\\ 0, & \text{当}2n_0\log\frac{2n_0}{n_0+n_1}+2n_1\log\frac{2n_1}{n_0+n_1}\le\chi_1^2(\alpha).\end{cases}$$

法三：利用渐近正态检验，注意到

$$\hat{p}_1-\hat{p}_2=\frac{n_0-n_1}{n}=\frac{1}{n}\sum_{j=1}^{n}\left[\mathbb{I}_{\{X_j=0\}}-\mathbb{I}_{\{X_j=1\}}\right]$$

是独立随机变量之平均，于是由中心极限定理知

$$\frac{\sqrt{n}\left(\hat{p}_1-\hat{p}_2-\mathbb{E}\left[\mathbb{I}_{\{X_j=0\}}-\mathbb{I}_{\{X_j=1\}}\right]\right)}{\sqrt{\operatorname{Var}\left[\mathbb{I}_{\{X_j=0\}}-\mathbb{I}_{\{X_j=1\}}\right]}}\xrightarrow{D}N(0,1)$$

其中

$$\mathbb{E}\left[\mathbb{I}_{\{X_1=0\}}-\mathbb{I}_{\{X_1=1\}}\right]=p_1-p_2\quad\operatorname{Var}\left[\mathbb{I}_{\{X_1=0\}}-\mathbb{I}_{\{X_1=1\}}\right]=p_1+p_2-\left(p_1-p_2\right)^2$$

结合 Slutsky 定理，因此考虑取检验统计量为

$$U(\boldsymbol{X})=\frac{\sqrt{n}\left(\hat{p}_1-\hat{p}_2\right)}{\sqrt{\hat{p}_1+\hat{p}_2-\left(\hat{p}_1-\hat{p}_2\right)^2}}$$

在 H_0 下，我们有 $U(\boldsymbol{X})\xrightarrow{D}N(0,1)$. 于是检验问题渐近水平 α 检验为

$$\phi(\boldsymbol{X})=\begin{cases}1, & \text{当}|U(\boldsymbol{X})|>u_{\alpha/2},\\ 0, & \text{当}|U(\boldsymbol{X})|\le u_{\alpha/2}.\end{cases}$$

法四：利用 Wald 检验，记 $\boldsymbol{\theta}=\left(p_1,p_2\right)^T,\hat{\boldsymbol{\theta}}=\left(\hat{p}_1,\hat{p}_2\right)^T$ ，于是由中心极限定理有

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\xrightarrow{D}N\left(0,\boldsymbol{I}^{-1}(\boldsymbol{\theta})\right),$$

其中总体（单个样本）的 Fisher 信息阵为

$$\boldsymbol{I}(\theta)=\begin{pmatrix}p_1^{-1}+p_3^{-1} & p_3^{-1}\\ p_3^{-1} & p_2^{-1}+p_3^{-1}\end{pmatrix}$$

注意 $h(\boldsymbol{\theta})=p_1-p_2,\boldsymbol{B}=\partial h/\partial\boldsymbol{\theta}=(1,-1)$ ，因此取检验统计量为

$$W_n=nh(\hat{\boldsymbol{\theta}})\left[\boldsymbol{B}(\hat{\boldsymbol{\theta}})\boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{B}^T(\hat{\boldsymbol{\theta}})\right]^{-1}h(\hat{\boldsymbol{\theta}})=\frac{n\left(\hat{p}_1-\hat{p}_2\right)^2}{\hat{p}_1+\hat{p}_2-\left(\hat{p}_1-\hat{p}_2\right)^2}$$

在 H_0 下，我们有 $W_n \xrightarrow{D} \chi_1^2$. 于是检验问题渐近水平 α 检验为

$$\phi(\boldsymbol{X})=\begin{cases}1, & \text{当} W_n>\chi_1^2(\alpha),\\ 0, & \text{当} W_n\le \chi_1^2(\alpha).\end{cases}$$

三.（30 分）设 $X_1,\dots,X_n$ i.i.d.$\sim N(\mu,1)$ ，其中 μ 为参数. 对水平 α ，试

（1）求 $\mathrm{P}\left(X_1>0\right)$ 的最大似然估计，并求其渐近方差。

（2）证明检验问题 $H_0:\mu=\mu_0\leftrightarrow H_1:\mu\neq\mu_0$ 不存在 UMPT，其中 μ_0 为一已知数.

（3）若参数 μ 在 $\mu=\mu_0$ 上的先验概率为 0.6 ，在 $\mu\neq\mu_0$ 上的先验分布为 $N\left(\mu_0,4\right)$ ，损失函数取为 0-1 损失，求（2）中的假设检验问题的 Bayes 决策。

解：（1）似然函数

$$L(\mu;\boldsymbol{x})=(2\pi)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right\}=(2\pi)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2-\frac{n(\mu-\overline{x})^2}{2}\right\}$$

因此 μ 的最大似然估计为 $\hat{\mu}=\overline{X}$ 。

由最大似然估计的不变性可知，$p=\mathrm{P}\left(X_1>0\right)=\Phi(\mu)$ 的最大似然估计为 $\hat{p}=\Phi(\hat{\mu})=\Phi(\overline{X})$ 。

注意到 $\overline{X}\sim N(\mu,1/n)$ ，由 Delta 方法可知 $\hat{p}$ 的渐近方差为

$$[\phi(\mu)]^2\cdot\frac{1}{n}=\frac{\exp\left\{-\mu^2\right\}}{2n\pi}$$

（2）首先注意到正态分布族（方差已知，均值为未知参数）关于 $T=\overline{X}$ 是单调似然比族。因而对检验问题 $H_0:\mu=\mu_0\leftrightarrow H_1':\mu>\mu_0$ ，存在 UMPT 形如

$$\phi_1(\boldsymbol{x})=\begin{cases}1, & \text{当}\overline{x}>\mu_0+u_\alpha/\sqrt{n},\\ 0, & \text{其他}.\end{cases}$$

对检验问题 $H_0:\mu=\mu_0\leftrightarrow H_1'':\mu<\mu_0$ ，存在 UMPT 形如

$$\phi_2(\boldsymbol{x})=\begin{cases}1, & \text{当}\overline{x}<\mu_0-u_\alpha/\sqrt{n},\\ 0, & \text{其他}.\end{cases}$$

显然 ϕ_1 和 ϕ_2 都是检验问题 $H_0\leftrightarrow H_1$ 的水平 α 检验。

假设检验问题 $H_0\leftrightarrow H_1$ 的 UMPT 存在，令其为 ϕ_0 。对固定 $\mu_1>\mu_0$ 和 $\mu_2<\mu_0$ ，检验 ϕ_0 也是简单假设 $H_0:\mu=\mu_0\leftrightarrow K_1:\mu=\mu_1$ 和 $H_0:\mu=\mu_0\leftrightarrow K_2:\mu=\mu_2$ 的 UMPT. 因此由 Neyman-Pearson 引理可知 ϕ_0 有形式

$$\phi_0(\boldsymbol{x})=\begin{cases}1, & \text{当} f\left(\boldsymbol{x};\mu_1\right)>k_1 f\left(\boldsymbol{x};\mu_0\right),\\ 0, & \text{当} f\left(\boldsymbol{x};\mu_1\right)\le k_1 f\left(\boldsymbol{x};\mu_0\right),\end{cases}$$

$$\phi_0(\boldsymbol{x})=\begin{cases}1, & \text{当} f\left(\boldsymbol{x};\mu_2\right)>k_2 f\left(\boldsymbol{x};\mu_0\right),\\ 0, & \text{当} f\left(\boldsymbol{x};\mu_2\right)\le k_2 f\left(\boldsymbol{x};\mu_0\right).\end{cases}$$

法一：考虑 $\boldsymbol{x}\in\{\boldsymbol{x}:\phi_0(\boldsymbol{x})=1\}$ ，由单调似然比的性质可知

- 如果 $T(\boldsymbol{y}) > T(\boldsymbol{x})$，则由第一个检验形式知 $\phi_0(\boldsymbol{y}) = 1$。
- 如果 $T(\boldsymbol{y}) < T(\boldsymbol{x})$，则由第二个检验形式知 $\phi_0(\boldsymbol{y}) = 1$。

于是要么 $\phi_0(\boldsymbol{y}) = 1$ 对所有 $\boldsymbol{y}$ 成立，要么 $\phi_0(\boldsymbol{x}) \neq 1$ 对所有 $\boldsymbol{x}$ 成立．这时 ϕ_0 的功效比 ϕ_1 和 ϕ_2 在各自的检验问题 $H_0 \leftrightarrow K_1$ 和 $H_0 \leftrightarrow K_2$ 都要小，导出矛盾。
法二：由唯一性可知，在 $\mu_1 > \mu_0$ 上，$\phi_0 = \phi_1$ ，a.e.；在 $\mu_2 < \mu_0$ 上，$\phi_0 = \phi_2$ ，a.e. 由 ϕ_1 和 ϕ_2 的形式知这不可能成立。
（3）由题意知两个假设的后验概率分别为

$$\alpha_0 = \mathrm{P}\left(\mu = \mu_0 \mid \boldsymbol{x}\right) = \frac{\pi_0 f\left(\boldsymbol{x} \mid \mu_0\right)}{m(\boldsymbol{x})}, \quad \alpha_1 = \mathrm{P}\left(\mu \neq \mu_0 \mid \boldsymbol{x}\right) = \frac{\pi_1 m_1(\boldsymbol{x})}{m(\boldsymbol{x})},$$

或者直接注意到简单假设对复杂假设的贝叶斯因子有形式

$$\mathrm{BF}_{01}(\boldsymbol{x}) = \frac{f\left(\boldsymbol{x} \mid \mu_0\right)}{m_1(\boldsymbol{x})}, \Longrightarrow \frac{\alpha_0}{\alpha_1} = \frac{\pi_0 f\left(\boldsymbol{x} \mid \mu_0\right)}{\pi_1 m_1(\boldsymbol{x})},$$

其中

$$m(\boldsymbol{x}) = \pi_0 f\left(\boldsymbol{x} \mid \mu_0\right) + \pi_1 m_1(\boldsymbol{x}), m_1(\boldsymbol{x}) = \int_{\mu \neq \mu_0} f(\boldsymbol{x} \mid \mu) \pi(\mu) \mathrm{d}\mu$$

下面计算 $m_1(\boldsymbol{x})$ 如下

$$\begin{aligned}
m_1(\boldsymbol{x}) &= \int_{\mu \neq \mu_0} f(\boldsymbol{x} \mid \mu) \pi(\mu) \mathrm{d}\mu \\
&= \int_{\mu \neq \mu_0} (2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} \sum_{i=1}^{n} (x_i - \overline{x})^2 - \frac{n(\mu - \overline{x})^2}{2}\right\} \cdot (8\pi)^{-\frac{1}{2}} \exp\left\{-\frac{(\mu - \mu_0)^2}{8}\right\} \mathrm{d}\mu \\
&= \frac{(2\pi)^{-\frac{n+1}{2}}}{2} \exp\left\{-\frac{(n-1)s^2}{2}\right\} \int_{\mu \neq \mu_0} \exp\left\{-\frac{A\mu^2 - 2B\mu + C}{2}\right\} \mathrm{d}\mu \\
&= \frac{(2\pi)^{-\frac{n+1}{2}}}{2} \exp\left\{-\frac{(n-1)s^2}{2}\right\} \cdot (2\pi / A)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(C - \frac{B^2}{A}\right)\right\} \\
&= \frac{(2\pi)^{-\frac{n}{2}}}{\sqrt{4n+1}} \exp\left\{-\frac{(n-1)s^2}{2} - \frac{n\left(\overline{x} - \mu_0\right)^2}{2(4n+1)}\right\}
\end{aligned}$$

其中

$$A = n + \frac{1}{4}, B = n\overline{x} + \frac{\mu_0}{4}, C = n\overline{x}^2 + \frac{\mu_0^2}{4}$$

因此

$$\frac{\alpha_0}{\alpha_1} = \frac{0.6(2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{(n-1)s^2}{2} - \frac{n(\overline{x} - \mu_0)^2}{2}\right\}}{0.4 \frac{(2\pi)^{-\frac{n}{2}}}{\sqrt{4n+1}} \exp\left\{-\frac{(n-1)s^2}{2} - \frac{n(\overline{x} - \mu_0)^2}{2(4n+1)}\right\}} = \frac{3\sqrt{4n+1}}{2} \exp\left\{-\frac{2n^2\left(\overline{x} - \mu_0\right)^2}{4n+1}\right\}$$

在 0-1 损失下，该检验问题的贝叶斯决策为

$$\delta(\boldsymbol{x}) = \begin{cases} a_0, & \text{当} \left(\overline{x} - \mu_0\right)^2 \leq \frac{4n+1}{2n^2} \log \frac{3\sqrt{4n+1}}{2}, \\ a_1, & \text{其他}, \end{cases}$$

其中 a_0 表示接受假设 H_0, a_1 表示接受假设 H_1 .

四.（30 分）设 $X_1,\ldots,X_m$ i. i. d. $\sim \operatorname{Exp}(\lambda_1)$（期望是 $1/\lambda_1$ 的指数分布），$Y_1,\ldots,Y_n$ i. i. d. $\sim \operatorname{Exp}(\lambda_2)$，且样本 $X_1,\ldots,X_m$ 和 $Y_1,\ldots,Y_n$ 独立，其中 λ_1,λ_2 为正参数。记 $\overline{X}$ 和 $\overline{Y}$ 分别为两组样本的样本均值．试

（1）求 $\mathbb{E}\left[\left(X_1-Y_1\right)^2 \mid \overline{X},\overline{Y}\right]$ 。

（2）求 λ_1/λ_2 的置信系数为 $1-\alpha$ 的置信区间．

（3）求检验问题 $H_0:\lambda_1=c\lambda_2 \leftrightarrow H_1:\lambda_1\neq c\lambda_2$ 的水平 α 似然比检验。

解：（1）法一：由指数分布的性质知

$$\mathbb{E}\left[\left(X_1-Y_1\right)^2\right]=\mathbb{E}\left(X_1^2\right)-2\mathbb{E}\left(X_1Y_1\right)+\mathbb{E}\left(Y_1^2\right)=\frac{2}{\lambda_1^2}-\frac{2}{\lambda_1\lambda_2}+\frac{2}{\lambda_2^2}.$$

注意到 $(\overline{X},\overline{Y})$ 是 (λ_1,λ_2) 的充分完全统计量，由 UMVUE 的唯一性可知

$$\mathbb{E}\left[\left(X_1-Y_1\right)^2 \mid \overline{X},\overline{Y}\right]=\left(\frac{2}{\lambda_1^2}-\frac{2}{\lambda_1\lambda_2}+\frac{2}{\lambda_2^2}\right)_{\text{UMVUE}}$$

显然地，我们有

$$\mathbb{E}\left(\overline{X}^2\right)=\operatorname{Var}(\overline{X})+[\mathbb{E}(\overline{X})]^2=\frac{m+1}{m\lambda_1^2},\quad \mathbb{E}\left(\overline{Y}^2\right)=\operatorname{Var}(\overline{Y})+[\mathbb{E}(\overline{Y})]^2=\frac{n+1}{n\lambda_2^2}.$$

因此

$$\mathbb{E}\left[\left(X_1-Y_1\right)^2 \mid \overline{X},\overline{Y}\right]=\frac{2m\overline{X}^2}{m+1}-2\overline{XY}+\frac{2n\overline{Y}^2}{n+1}.$$

法二：由条件期望的线性性及独立性可知

$$\mathbb{E}\left[\left(X_1-Y_1\right)^2 \mid \overline{X},\overline{Y}\right]=\mathbb{E}\left(X_1^2 \mid \overline{X}\right)-2\mathbb{E}\left(X_1Y_1 \mid \overline{X},\overline{Y}\right)+\mathbb{E}\left(Y_1^2 \mid \overline{Y}\right).$$

注意到

$$X_1/(m\overline{X})\sim \operatorname{Be}(1,m-1), Y_1/(n\overline{Y})\sim \operatorname{Be}(1,n-1)$$

且 $(\overline{X},\overline{Y})$ 是 (λ_1,λ_2) 的充分完全统计量，由 Basu 定理可知

$$\begin{aligned}
&\mathbb{E}\left(X_1^2 \mid \overline{X}\right)=\overline{X}^2\cdot\mathbb{E}\left(\left.\frac{X_1^2}{\overline{X}^2}\right|\overline{X}\right)=\overline{X}^2\cdot\mathbb{E}\left(\frac{X_1^2}{\overline{X}^2}\right),\quad \mathbb{E}\left(Y_1^2 \mid \overline{Y}\right)=\overline{Y}^2\cdot\mathbb{E}\left(\frac{Y_1^2}{\overline{Y}^2}\right),\\
&\mathbb{E}\left(X_1Y_1 \mid \overline{X},\overline{Y}\right)=\overline{XY}\mathbb{E}\left(\left.\frac{X_1}{\overline{X}}\cdot\frac{Y_1}{\overline{Y}}\right|\overline{X},\overline{Y}\right)=\overline{XY}\mathbb{E}\left(\frac{X_1}{\overline{X}}\right)\cdot\mathbb{E}\left(\frac{Y_1}{\overline{Y}}\right).
\end{aligned}$$

由贝塔分布性质知

$$\mathbb{E}\left(\frac{X_1}{m\overline{X}}\right)=\frac{1}{m},\quad \mathbb{E}\left(\frac{X_1^2}{m^2\overline{X}^2}\right)=\frac{2}{m(m+1)},\quad \mathbb{E}\left(\frac{Y_1}{n\overline{Y}}\right)=\frac{1}{n},\quad \mathbb{E}\left(\frac{Y_1^2}{n^2\overline{Y}^2}\right)=\frac{2}{n(n+1)}.$$

于是代入可得

$$\mathbb{E}\left[\left(X_1-Y_1\right)^2 \mid \overline{X},\overline{Y}\right]=\frac{2m\overline{X}^2}{m+1}-2\overline{XY}+\frac{2n\overline{Y}^2}{n+1}$$

（2）注意 $2m\lambda_1\overline{X}\sim\chi_{2m}^2, 2n\lambda_2\overline{Y}\sim\chi_{2n}^2$ ，取枢轴变量为

$$\frac{\lambda_1\overline{X}}{\lambda_2\overline{Y}}\sim F_{2m,2n}$$

由 $\mathrm{P}\left(F_{2m,2n}(1-\alpha/2) \leq \frac{\lambda_1\overline{X}}{\lambda_2\overline{Y}} \leq F_{2m,2n}(\alpha/2)\right) = 1-\alpha$ ，反解得到 λ_1/λ_2 的置信系数为 $1-\alpha$ 的置信区间为

$$\left[F_{2m,2n}(1-\alpha/2)\cdot\frac{\overline{Y}}{\overline{X}}, F_{2m,2n}(\alpha/2)\cdot\frac{\overline{Y}}{\overline{X}}\right].$$

（3）似然函数

$$L\left(\lambda_1,\lambda_2;\boldsymbol{x},\boldsymbol{y}\right)=\lambda_1^m\exp\left\{-\lambda_1\sum_{i=1}^m x_i\right\}\cdot\lambda_2^n\exp\left\{-\lambda_2\sum_{j=1}^n y_j\right\}.$$

由似然比检验的思想，取检验统计量为

$$\Lambda(\boldsymbol{x},\boldsymbol{y})=\frac{\sup_{\lambda_1>0,\lambda_2>0}L\left(\lambda_1,\lambda_2;\boldsymbol{x},\boldsymbol{y}\right)}{\sup_{\lambda_1/\lambda_2=c}L\left(\lambda_1,\lambda_2;\boldsymbol{x},\boldsymbol{y}\right)}.$$

注意到全空间下 $\hat{\lambda}_1=1/\overline{X},\hat{\lambda}_2=1/\overline{Y}$ ．在原假设空间下，记 $\lambda_1=c\lambda,\lambda_2=\lambda$ ，考虑 λ 的似然函数为

$$L(\lambda;\boldsymbol{x},\boldsymbol{y})=c^m\lambda^{m+n}\exp\{-\lambda(cm\overline{x}+n\overline{y})\}$$

此时最大似然估计 $\hat{\lambda}=(m+n)/(cm\overline{x}+n\overline{y})$ ．于是似然比可化简为

$$\begin{aligned}\Lambda(\boldsymbol{x},\boldsymbol{y})=\frac{L\left(\hat{\lambda}_1,\hat{\lambda}_2;\boldsymbol{x},\boldsymbol{y}\right)}{L(c\hat{\lambda},\hat{\lambda};\boldsymbol{x},\boldsymbol{y})}&=\frac{(cm\overline{x}+n\overline{y})^{m+n}}{c^m(m+n)^{m+n}\overline{x}^m\cdot\overline{y}^n}\\&=\frac{1}{(m+n)^{m+n}}\left(m+n\frac{\overline{y}}{c\overline{x}}\right)^m\left(n+m\frac{c\overline{x}}{\overline{y}}\right)^n\\&\triangleq\frac{1}{(m+n)^{m+n}}\left(m+nF^{-1}\right)^m(n+mF)^n\end{aligned}$$

其中 $F:=F(\boldsymbol{x},\boldsymbol{y})=c\overline{x}/\overline{y}$ ．注意到 Λ 关于 F 先递减后递增，因此检验的拒绝域的形式为

$$R=\left\{(\boldsymbol{X},\boldsymbol{Y}):F(\boldsymbol{X},\boldsymbol{Y})<c_1\text{ 或}F(\boldsymbol{X},\boldsymbol{Y})>c_2\right\}.$$

注意到 $2m\lambda_1\overline{X}\sim\chi^2_{2m},2n\lambda_2\overline{Y}\sim\chi^2_{2n}$ ，所以在 H_0 下，

$$F(\boldsymbol{X},\boldsymbol{Y})=\frac{c\overline{X}}{\overline{Y}}\sim F_{2m,2n}$$

由显著性水平 α 要求知 $c_1=F_{2m,2n}(1-\alpha/2),c_2=F_{2m,2n}(\alpha/2)$ ．

2.3 23 数理统计期末残卷

1. 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 $N(\mu,\sigma^2)$ ．
(1）求包月客户平均使用时间和标准差的 95% 置信区间。
(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？
(3）假设总体标准差为 90 分钟，取 μ 的先验分布为无信息先验，求 μ 的后验分布并据此给出 μ 的可信系数为 95% 的可信区间。
(4）解释（1）和（3）中关于平均使用时间所得区间的含义与区别．
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

$$[88.15, 92.02]$$

(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
(3）位置参数的无信息先验为

$$\pi(\mu)=1$$

样本联合密度为

$$\begin{aligned} f(x_1,\cdots,x_n;\mu) &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right\} \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n \left(x_i-\overline{X}\right)^2 + n(\overline{X}-\mu)^2\right\} \end{aligned}$$

则

$$\pi(\mu \mid \vec{X}) \propto \exp\left\{-\frac{n}{2\sigma^2}\left(\overline{X}-\mu\right)^2\right\}$$

也就是

$$\mu \mid \vec{X} \sim N\left(\overline{X}, \frac{\sigma^2}{n}\right)$$

则可信区间为

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}}u_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}}u_{\alpha/2}\right]$$

代入数值为

$$[214.12, 225.88]$$

（4）置信区间：经过多次重复实验，μ 落在置信区间的**频率**趋于 95%，参数是一个真实的固定值
可信区间：相当于把 μ 视为**随机变量**，一次试验后 μ 落在可信区间的**概率**为 95%

2. 设 $X_1,\dots,X_n$ i.i.d.$\sim N(\theta,1), Y_1,\dots,Y_n$ i.i.d.$\sim N(2\theta,1), \theta$ 为未知参数. 又设合样本 $X_1,\dots,X_n,Y_1,\dots,Y_n$ 独立. 试
（1）求 θ 的 UMVUE.
（2）如果要求 θ 的置信系数为 $1-\alpha$ 的置信区间宽不超过指定的 d，则样本量 n 应该至少多大?
解：（1）样本联合密度为

$$\begin{aligned} f\left(X_1,\cdots,X_n,Y_1,\cdots,Y_n;\theta\right) &= \left(\frac{1}{\sqrt{2\pi}}\right)^{-2n} \cdot e^{\frac{-\sum_{i=1}^n (X_i-\theta)^2 - \sum_{i=1}^n (Y_i-2\theta)^2}{2}} \\ &= \left(\frac{1}{2\pi}\right)^n \cdot e^{-\frac{5n\theta^2}{2}} \cdot e^{\theta\cdot\left(\sum_{i=1}^n (X_i+2Y_i)\right)} \cdot e^{-\frac{1}{2}\sum_{i=1}^n \left(X_i^2+Y_i^2\right)} \end{aligned}$$

显然自然参数空间有内点，则 $T(\boldsymbol{X}) = \sum\limits_{i=1}^n (X_i+2Y_i)$ 为充分完全统计量
注意独立性，有

$$T(\boldsymbol{X}) \sim N(5n\theta, 5n)$$

则有无偏估计

$$\frac{T(\boldsymbol{X})}{5n} = \frac{\overline{X}+2\overline{Y}}{5}$$

它也是充分完全统计量的函数，则其为 UMVUE
（2）取

$$\frac{\frac{\overline{X}+2\overline{Y}}{5}-\theta}{\frac{1}{\sqrt{5n}}} \sim N(0,1)$$

作为枢轴变量
则区间长度

$$\frac{2u_{\alpha/2}}{\sqrt{5n}} \le d$$

得到

$$n \ge \left\lceil \frac{4u_{\alpha/2}^2}{5d^2} \right\rceil$$

3. 随机地从一批铁钉中抽取 16 枚，测得它们的平均长度（单位：cm)$\bar{x}=3.035$ ．已知铁钉长度 X 服从正态分布 $N\left(\mu, 0.1^2\right)$ ．考虑假设检验问题

$$H_0: \mu \le 3 \longleftrightarrow H_1: \mu > 3 \tag{$\star$}$$

（1）在 $\alpha=0.05$ 水平下对检验问题 $(\star)$ 进行检验，并给出检验的 p 值。
（2）如果行动 $a=0$ 和 $a=1$ 分别表示接受 H_0 和拒绝 H_0, μ 的先验分布为 $N\left(3, 0.1^2\right)$ ，损失函数取为

$$L(a, \mu)=\begin{cases}11, & a=1, \mu \le 3, \\ 1, & a=0, \mu>3, \\ 0, & \text{其他.}\end{cases}$$

基于后验风险最小原则给出检验问题 $(\star)$ 的最优决策行动，并与（1）中检验结论进行对比。
解：（1）给出检验统计量

$$U(\boldsymbol{X})=\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \stackrel{H_0}{\sim} N(0,1)$$

由检验问题的形式知拒绝域为

$$D=\{\boldsymbol{X}: U(\boldsymbol{X})>u_\alpha\}$$

现有 $\bar{x}=3.035, \mu_0=3, \sigma=0.1, n=16$ ，查表得 $u_{0.05}=1.645$ ，因此

$$u(\boldsymbol{x})=\frac{3.035-3}{0.1/\sqrt{16}}=1.4<1.645$$

因此不能拒绝原假设.
检验的 p 值为

$$p=P(U(\boldsymbol{X})>u(\boldsymbol{x}))=1-\Phi(1.4)=0.0808$$

（2）在先验分布 $\mu \sim N\left(\mu_0, \tau^2\right)$ 下，正确计算出 μ 的后验分布

$$\mu \,\Big|\, \boldsymbol{x} \sim N\left(\frac{n\tau^2\bar{x}+\sigma^2\mu_0}{n\tau^2+\sigma^2}, \frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}\right) \stackrel{\text{代入数据}}{=} N\left(\frac{1289}{425}, \frac{1}{1700}\right).$$

后验风险

$$R\left(a_0, \mu\right)=\mathbb{P}(\mu>3 \mid \boldsymbol{x})=0.9131 \quad R\left(a_1, \mu\right)=11\mathbb{P}(\mu \le 3 \mid \boldsymbol{x})=0.9559$$

结论：按后验风险最小原则，应采取行动 a_0 ，即接受原假设 H_0 。
这与（1）中检验结果一致，这是因为我们对拒绝 H_0 赋予了较大的惩罚。

4. 设总体 X 的密度函数为

$$f(x;\theta)=\begin{cases}\theta, & 0<x<1,\\ 1-\theta, & 1\le x<2,\\ 0, & \text{其他},\end{cases}$$

其中 $0<\theta<1$ 为未知参数．现从此总体中抽出一样本量为 n 的样本 $X_1,\dots,X_n$ ．

（1）试求参数 θ 的充分统计量，并说明它是否为完全统计量？

（2）求假设检验问题 $H_0:\theta=\theta_0 \leftrightarrow H_1:\theta\neq\theta_0$ 的水平 α 似然比检验，其中 $0<\theta_0,\alpha<1$ 已知。

（3）上述假设是否存在 UMPT？为什么？

解：（1）样本联合密度

$$f(x_1,\cdots,x_n;\theta)=\theta^{n_1}\cdot(1-\theta)^{n_2}\mathbb{I}_{\left\{0<X_{(1)}<X_{(n)}<2\right\}}$$

其中

$$n_1=\sum_{i=1}^{n}\mathbb{I}_{\{0<X_i<1\}}\quad n_2=n-n_1$$

则

$$n_1\sim \mathrm{B}(n,\theta)$$

将联合密度改写成指数族形式

$$\begin{aligned}f(x_1,\cdots,x_n;\theta)&=e^{n_1\ln\theta+n_2\ln(1-\theta)}\mathbb{I}_{\left\{0<X_{(1)}<X_{(n)}<2\right\}}\\&=(1-\theta)^n\cdot e^{n_1\ln\frac{\theta}{1-\theta}}\cdot h(\boldsymbol{X})\end{aligned}$$

由因子分解定理和自然参数空间有内点，$T(\boldsymbol{X})=n_1$ 为充分完全统计量

（2）对全空间 $0<\theta<1$，易得

$$\hat{\theta}_{MLE}=\frac{n_1}{n}$$

则

$$\sup_{0<\theta<1}f(x_1,\cdots,x_n;\theta)=\left(\frac{n_1}{n}\right)^{n_1}\left(1-\frac{n_1}{n}\right)^{n-n_1}\mathbb{I}_{\left\{0<X_{(1)}<X_{(n)}<2\right\}}$$

另一方面

$$\sup_{\theta=\theta_0}f(x_1,\cdots,x_n;\theta)=f(x_1,\cdots,x_n;\theta_0)$$

则似然比为

$$\lambda(\boldsymbol{X})=\frac{\sup\limits_{0<\theta<1}f(x_1,\cdots,x_n;\theta)}{\sup\limits_{\theta=\theta_0}f(x_1,\cdots,x_n;\theta)}=\left(\frac{n_1}{\theta_0 n}\right)^{n_1}\left(\frac{n-n_1}{(1-\theta_0)\,n}\right)^{n-n_1}$$

则

$$2\log\lambda(\boldsymbol{X})\xrightarrow{H_0}\chi^2(1)$$

拒绝域为

$$\left\{(X_1,\cdots,X_n):2\log\lambda(\boldsymbol{X})>\chi_1^2(\alpha)\right\}$$

（3）设 φ_1 为假设检验问题

$$H_0:\theta=\theta_0\leftrightarrow H_1':\theta=\theta_1$$

的 UMPT，这里 $\theta_1 > \theta_0$
则由 N-P 引理

$$\varphi_1(\boldsymbol{x}) = \begin{cases} 1 & f(x,\theta_1)/f(x,\theta_0) > c_1 \\ r_1 & f(x,\theta_1)/f(x,\theta_0) = c_1 \\ 0 & f(x,\theta_1)/f(x,\theta_0) < c_1 \end{cases}$$

同理对假设检验问题

$$H_0 : \theta = \theta_0 \leftrightarrow H_1'' : \theta = \theta_2$$

这里 $\theta_2 < \theta_0$
有 UMPT

$$\varphi_2(\boldsymbol{x}) = \begin{cases} 1 & f(x,\theta_2)/f(x,\theta_0) > c_2 \\ r_2 & f(x,\theta_2)/f(x,\theta_0) = c_2 \\ 0 & f(x,\theta_2)/f(x,\theta_0) < c_2 \end{cases}$$

另一方面

$$\begin{aligned} f(\vec{x},\theta_1)/f(\vec{x},\theta_0) &= \left(\frac{\theta_1}{\theta_0}\right)^{n_1} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-n_1} \\ &= \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n} \quad \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{n_1} \end{aligned}$$

关于 n_1 单调递增
则 $\varphi_1(\boldsymbol{x})$ 可以改写为

$$\varphi_1(\boldsymbol{x}) = \begin{cases} 1 & n_1 > c_1 \\ r_1 & n_1 = c_1 \\ 0 & n_1 < c_1 \end{cases}$$

同理

$$f(\vec{x},\theta_2)/f(\vec{x},\theta_0) = \left(\frac{1-\theta_2}{1-\theta_0}\right)^{n} \cdot \left(\frac{\theta_2-\theta_2\theta_0}{\theta_0-\theta_0\theta_2}\right)^{n_1}$$

关于 n_1 单调递减
则 $\varphi_1(\boldsymbol{x})$ 可以改写为

$$\varphi_2(\boldsymbol{x}) = \begin{cases} 1 & n_1 < c_2 \\ r_2 & n_1 = c_2 \\ 0 & n_1 > c_2 \end{cases}$$

如果假设检验问题

$$H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0$$

的 UMPT 存在，记为 $\varphi_0(\boldsymbol{x})$
则由 N-P 引理的唯一性

$$\begin{aligned} &\theta_1 > \theta_0 \text{ 时} \quad \varphi_1 = \varphi_0 \quad a.e. \\ &\theta_2 < \theta_0 \text{ 时} \quad \varphi_2 = \varphi_0 \quad a.e. \end{aligned}$$

但这与 φ_1, φ_2 的形式矛盾，也就不存在假设检验问题 $H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0$ 的 UMPT

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