25 数理统计期中

$\mathscr{NULIOUS}$

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：

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

再有

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

$$X_{n+1}-\overline{X} \sim N\left(\mu, \sigma^2\right)-N\left(\mu, \frac{\sigma^2}{n}\right)=N\left(0, \frac{n+1}{n}\sigma^2\right) \tag{3}$$

(3) 式来自两个变量的独立性

则

$$\sqrt{\frac{n}{n+1}}\left(X_{n+1}-\overline{X}\right)/S \sim t_{n-1} \tag{4}$$

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

$$\frac{(X_{n+1}) - (\overline{X})}{S} = \frac{(X_{n+1} - \mu) - (\overline{X} - \mu)}{S} \tag{5}$$

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

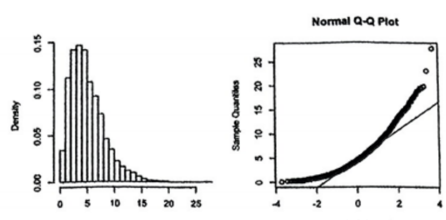
（D）估计值 1.5 是 θ 的相合估计

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:

$$f(\vec{x};\alpha,\beta)=\frac{\beta^{n\alpha}}{\Gamma(\alpha)^n}\prod_{i=1}^{n}x_i^{\alpha-1}e^{-\beta\sum_{i=1}^{n}x_i} \tag{6}$$

把密度写成指数族的形式：

$$f(\vec{x};\alpha,\beta)=\frac{\beta^{n\alpha}}{\Gamma(\alpha)^n}e^{(\alpha-1)\sum_{i=1}^{n}\ln(x_i)-\beta\sum_{i=1}^{n}x_i} \tag{7}$$

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：

$$f(\vec{x};\theta)=\mathbb{I}_{\theta<X_{(1)}<X_{(n)}<\theta+1} \tag{8}$$

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：

$$\begin{aligned}\frac{f(\vec{x};\theta)}{f(\vec{y};\theta)}=C(\vec{x},\vec{y}) &\iff \frac{\mathbb{I}_{\theta<X_{(1)}<X_{(n)}<\theta+1}}{\mathbb{I}_{\theta<Y_{(1)}<Y_{(n)}<\theta+1}}=C(\vec{x},\vec{y})\\ &\iff (X_{(1)},X_{(n)})=(Y_{(1)},Y_{(n)})\end{aligned} \tag{9}$$

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

$$f(\vec{x};\theta) = \begin{cases} 1, & X_{(n)} - 1 < \theta < X_{(1)}, \\ 0, & \text{其他}. \end{cases} \tag{10}$$

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 $0<t<1$

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

而

$$X_{(1)}-\theta\sim\beta(1,n),\tag{11}$$
$$X_{(n)}-\theta\sim\beta(n,1).\tag{12}$$

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

$$\ln f(\vec{x};\theta) = n_1\ln\theta + n_2\ln(2\theta) + n_3\ln(1-3\theta) \tag{13}$$

$$\frac{\partial \ln f(\vec{x};\theta)}{\partial\theta} = \frac{n_1}{\theta} + \frac{n_2}{\theta} - \frac{3n_3}{1-3\theta} = 0 \tag{14}$$

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

$$f(\vec{x};\theta) = e^{n_2\ln 2}e^{n_1\ln\theta + n_2\ln\theta + n_3\ln(1-3\theta)} \tag{15}$$

$$= e^{n_2\ln 2}e^{\ln\theta}e^{n_3\ln(\frac{1-3\theta}{\theta})} \tag{16}$$

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 $\Longleftrightarrow$ 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：

$$\mathrm{Var}(\hat{\theta}_{\mathrm{MLE}}) = \frac{\mathrm{Var}(n_3)}{9n^2} = \frac{3n\theta(1-3\theta)}{9n^2} = \frac{\theta(1-3\theta)}{3n}, \tag{17}$$

$$\begin{aligned} n\text{个样本的 Fisher 信息量：} I(\theta) &= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(\vec{x};\theta)\right] \\ &= -\mathbb{E}[-\frac{n_1}{\theta^2} - \frac{n_2}{\theta^2} - \frac{9n_3}{(1-3\theta)^2}] \\ &= \frac{3n}{\theta(1-3\theta)} \end{aligned} \tag{18}$$

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：

$$f(\vec{x};\mu,\sigma^2)=(2\pi\sigma^2)^{-\frac{n}{2}}e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}} \tag{19}$$

后验联合密度为：

$$\pi(\mu,\sigma^2|\vec{x})\propto f(\vec{x};\mu,\sigma^2)\cdot\pi\left(\mu,\sigma^2\right)\propto(\sigma^2)^{-\frac{n}{2}-1}e^{-\frac{\sum(x_i-\overline{X})^2+n(\overline{X}-\mu)^2}{2\sigma^2}} \tag{20}$$

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

$$\begin{aligned}\pi(\sigma^2|\vec{x})&=\int_{-\infty}^{+\infty}\pi(\mu,\sigma^2|\vec{x})d\mu\\&=(\sigma^2)^{-\frac{n}{2}-1}e^{-\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2}{2\sigma^2}}\int_{-\infty}^{+\infty}e^{-\frac{n(\overline{X}-\mu)^2}{2\sigma^2}}d\mu\\&=(\sigma^2)^{-\frac{n}{2}-\frac{1}{2}}e^{-\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2}{2\sigma^2}}\end{aligned} \tag{21}$$

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

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

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：

$$\begin{aligned}
\pi(\mu|\vec{x}) &= \int_0^{+\infty} \pi(\mu,\sigma^2|\vec{x}) d\sigma^2 \\
&= \int_0^{+\infty} (\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{\sum(x_i-\overline{X})^2+n(\overline{X}-\mu)^2}{2\sigma^2}} d\sigma^2 \\
&\overset{t=\frac{1}{\sigma^2}}{=} \int_0^{+\infty} (t)^{\frac{n}{2}-1} e^{-\frac{t(\sum(x_i-\overline{X})^2+n(\overline{X}-\mu)^2)}{2}} dt
\end{aligned} \tag{22}$$

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

$$\begin{aligned}\pi(\mu|\vec{x}) &\propto \frac{\Gamma(\alpha)}{\beta^{\alpha}} \\ &= \frac{\Gamma(\frac{n}{2})}{(\frac{\sum_{i=1}^{n}(x_i-\overline{X})^2-n(\overline{X}-\mu)^2}{2})^{\frac{n}{2}}}\end{aligned} \tag{23}$$

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：

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