



中国人民大学夏令营试题汇编

数学学院励志书院数学试验班师生联合党支部 整理人：李怡龙 (25)



中国人民大学统计学院
2022 年夏令营招生考试试题（统计学）

1. 计算定积分 $\int_0^1 x \ln(1+x) dx$.

解答：分部积分，令 $u = \ln(1+x)$, $dv = x dx$, 则 $du = \frac{1}{1+x} dx$, $v = \frac{1}{2} x^2$.

$$\begin{aligned} \int_0^1 x \ln(1+x) dx &= \left. \frac{1}{2} x^2 \ln(1+x) \right|_0^1 - \int_0^1 \frac{1}{2} x^2 \cdot \frac{1}{1+x} dx \\ &= \frac{1}{2} \ln 2 - \frac{1}{2} \int_0^1 \frac{x^2 - 1 + 1}{1+x} dx \\ &= \frac{1}{2} \ln 2 - \frac{1}{2} \int_0^1 \left(x - 1 + \frac{1}{1+x} \right) dx \\ &= \frac{1}{2} \ln 2 - \frac{1}{2} \left(\frac{1}{2} x^2 - x + \ln(1+x) \right) \Big|_0^1 \\ &= \frac{1}{2} \ln 2 - \frac{1}{2} \left[\left(\frac{1}{2} - 1 + \ln 2 \right) - 0 \right] \\ &= \frac{1}{2} \ln 2 + \frac{1}{4} - \frac{1}{2} \ln 2 = \frac{1}{4}. \end{aligned}$$

2. 设数列 a_n 满足条件 $a_0 = 1, a_1 = 1, a_n = a_{n-1} + a_{n-2} (n = 2, 3, \dots)$. 判断级数 $\sum_{n=1}^{\infty} \frac{1}{a_n}$ 的收敛性.

解答：该数列为斐波那契数列。 a_n 严格递增，

$$\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} = \lim_{n \rightarrow \infty} \frac{a_n + a_{n-1}}{a_n} = 1 + \lim_{n \rightarrow \infty} \frac{a_{n-1}}{a_n} = \frac{1 + \sqrt{5}}{2} = \phi > 1.$$

$$\lim_{n \rightarrow \infty} \frac{\frac{1}{a_{n+1}}}{\frac{1}{a_n}} = \lim_{n \rightarrow \infty} \frac{a_n}{a_{n+1}} = \frac{\sqrt{5} - 1}{2} < 1.$$

由比值判别法，级数 $\sum_{n=1}^{\infty} \frac{1}{a_n}$ 收敛。

3. 设矩阵 $A = (\alpha_1, \alpha_2, \alpha_3)$ 有三个不同的特征值，其中 $\alpha_1, \alpha_2, \alpha_3$ 为 A 的列向量且满足 $\alpha_3 = \alpha_1 + 2\alpha_2$.

(1) 证明： $\text{rank}(A) = 2$.

证明：由条件 $\alpha_3 - \alpha_1 - 2\alpha_2 = \mathbf{0}$ ，说明列向量组线性相关，故 $\text{rank}(A) \leq 2$ 。矩阵有三个不同特征值，则对应三个特征向量线性无关；不同特征值的特征向量必线性无关，因此矩阵至少有 2 个线性无关特征向量， $\text{rank}(A) \geq 2$ 。综上 $\text{rank}(A) = 2$ 。

(1) 已知向量 $\beta = \alpha_1 + \alpha_2 - 3\alpha_3$ ，求解线性方程组 $Ax = \beta$.

解答：

$$A = (\alpha_1, \alpha_2, \alpha_3), \alpha_3 = \alpha_1 + 2\alpha_2 \implies A \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} = \mathbf{0},$$

即 $\xi_0 = (1, 2, -1)^\top$ 是齐次方程组 $Ax = \mathbf{0}$ 的一个非零解。

$$\beta = \alpha_1 + \alpha_2 - 3\alpha_3 = A \begin{pmatrix} 1 \\ 1 \\ -3 \end{pmatrix},$$

得到特解 $\eta^* = (1, 1, -3)^\top$ 。因为 $\text{rank}(A) = 2$ ，未知数 3 个，齐次基础解系含 $3 - 2 = 1$ 个向量。通解：

$$x = \begin{pmatrix} 1 \\ 1 \\ -3 \end{pmatrix} + k \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \quad k \in \mathbb{R}.$$



4. 现有两枚硬币, 硬币 A 正面和反面出现的概率均为 $\frac{1}{2}$, 硬币 B 正面和反面出现的概率分别为 $\frac{3}{4}$ 和 $\frac{1}{4}$. 随机选择一枚硬币进行抛掷, 若前三次结果均为正面, 求第四次抛掷出现正面的概率.

解答: 记事件 A : 选硬币 A ; B : 选硬币 B ; C_3 : 前三次均正面; C_4 : 第四次正面. $P(A) = P(B) = \frac{1}{2}$.

$$P(C_3|A) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}, \quad P(C_3|B) = \left(\frac{3}{4}\right)^3 = \frac{27}{64}.$$

由贝叶斯公式:

$$P(A|C_3) = \frac{P(C_3|A)P(A)}{P(C_3|A)P(A) + P(C_3|B)P(B)} = \frac{\frac{1}{8} \cdot \frac{1}{2}}{\frac{1}{8} \cdot \frac{1}{2} + \frac{27}{64} \cdot \frac{1}{2}} = \frac{8}{35},$$

$$P(B|C_3) = 1 - P(A|C_3) = \frac{27}{35}.$$

第四次结果与前三次条件独立,

$$P(C_4|C_3) = P(C_4|A)P(A|C_3) + P(C_4|B)P(B|C_3) = \frac{1}{2} \cdot \frac{8}{35} + \frac{3}{4} \cdot \frac{27}{35} = \frac{113}{140}.$$

5. 有 m 种不同的卡片, 每次抽到第 i 种卡片的概率为 p_i , 且满足 $\sum_{i=1}^m p_i = 1$. 现有放回地抽取 n 次, 定义随机变量 X 为抽到的不同卡片种类数, 求 $E(X)$.

解答: 引入示性随机变量

$$I_i = \begin{cases} 1, & \text{第 } i \text{ 种卡片被抽到过,} \\ 0, & \text{第 } i \text{ 种卡片从未被抽到.} \end{cases} \quad i = 1, 2, \dots, m$$

则 $X = \sum_{i=1}^m I_i$.

$$P(I_i = 0) = (1 - p_i)^n, \quad P(I_i = 1) = 1 - (1 - p_i)^n, \quad E(I_i) = 1 - (1 - p_i)^n.$$

利用期望线性性:

$$E(X) = \sum_{i=1}^m E(I_i) = \sum_{i=1}^m [1 - (1 - p_i)^n].$$

6. 设总体 X 服从均匀分布 $U\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$, 令 $X_1, \dots, X_n$ 为从总体 X 中抽取的简单样本。

- (1) 证明: 均值 $\hat{\theta}_1 = \bar{X}$ 和 $\hat{\theta}_2 = \frac{1}{2}(X_{(1)} + X_{(n)})$ 均为 θ 的无偏估计。

证明: 总体 $X \sim U\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$, $E(X) = \theta$ 。

$$E(\hat{\theta}_1) = E(\bar{X}) = E(X) = \theta,$$

故 $\hat{\theta}_1$ 无偏。做变换 $Y_i = X_i - \theta$, 则 $Y_i \sim U\left(-\frac{1}{2}, \frac{1}{2}\right)$ 。 $Y_{(1)} = X_{(1)} - \theta$, $Y_{(n)} = X_{(n)} - \theta$ 。

$$E\left(\frac{1}{2}(X_{(1)} + X_{(n)})\right) = \theta + \frac{1}{2}(E(Y_{(1)}) + E(Y_{(n)})).$$

$Y_{(1)}$ 与 $-Y_{(n)}$ 同分布, 故 $E(Y_{(1)}) = -E(Y_{(n)})$, 得到 $E(Y_{(1)}) + E(Y_{(n)}) = 0$ 。

$$E(\hat{\theta}_2) = \theta + 0 = \theta,$$

$\hat{\theta}_2$ 也是 θ 的无偏估计。

- (1) 求 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 的方差, 并说明哪一个更有效。

解答: 总体方差

$$\text{Var}(X) = \frac{\left[(\theta + \frac{1}{2}) - (\theta - \frac{1}{2})\right]^2}{12} = \frac{1}{12}.$$



$$\text{Var}(\hat{\theta}_1) = \text{Var}(\bar{X}) = \frac{1}{n} \text{Var}(X) = \frac{1}{12n}.$$

$Y \sim U(-1/2, 1/2)$, 次序统计量 $Y_{(1)}, Y_{(n)}$ 的联合密度:

$$f_{Y_{(1)}, Y_{(n)}}(y_1, y_n) = n(n-1)(y_n - y_1)^{n-2}, \quad -\frac{1}{2} < y_1 < y_n < \frac{1}{2}.$$

$$\begin{aligned} \text{Var}(\hat{\theta}_2) &= \text{Var}\left(\frac{1}{2}(X_{(1)} + X_{(n)})\right) = \text{Var}\left(\frac{1}{2}(Y_{(1)} + Y_{(n)})\right) \\ &= \frac{1}{4}[\text{Var}(Y_{(1)}) + \text{Var}(Y_{(n)}) + 2\text{Cov}(Y_{(1)}, Y_{(n)})]. \end{aligned}$$

由对称性 $\text{Var}(Y_{(1)}) = \text{Var}(Y_{(n)})$, 计算得到

$$\text{Var}(\hat{\theta}_2) = \frac{1}{2(n+1)(n+2)}.$$

$$\frac{1}{12n} - \frac{1}{2(n+1)(n+2)} = \frac{(n+1)(n+2) - 6n}{12n(n+1)(n+2)} = \frac{(n-1)(n-2)}{12n(n+1)(n+2)} \geq 0, \quad n \geq 1.$$

当 $n > 2$ 时, $\text{Var}(\hat{\theta}_2) < \text{Var}(\hat{\theta}_1)$, $\hat{\theta}_2$ 更有效。



中国人民大学统计学院
2022 年夏令营招生考试试题（应用统计）

1. 计算积分 $\int_a^{+\infty} 4x^2 e^{-2x} dx$.

解答：反复分部积分，记 $I = \int 4x^2 e^{-2x} dx$

$$\begin{aligned}\int 4x^2 e^{-2x} dx &= -2x^2 e^{-2x} + \int 4x e^{-2x} dx \\ &= -2x^2 e^{-2x} - 2x e^{-2x} + \int 2e^{-2x} dx \\ &= -2x^2 e^{-2x} - 2x e^{-2x} - e^{-2x} + C \\ &= -e^{-2x} (2x^2 + 2x + 1) + C.\end{aligned}$$

$$\int_a^{+\infty} 4x^2 e^{-2x} dx = -e^{-2x} (2x^2 + 2x + 1) \Big|_a^{+\infty} = e^{-2a} (2a^2 + 2a + 1).$$

2. 计算二重积分 $\iint_D y dx dy$, 其中积分区域 D 由直线 $x = -2$, $y = 2$, $y = 0$ 和半圆 $x^2 + (y-1)^2 = 1$ ($x < 0$) 共同围成.

解答：区域：左侧矩形减去左半圆。矩形区域： $-2 \leq x \leq 0$, $0 \leq y \leq 2$ ；半圆： $x^2 + (y-1)^2 \leq 1$, $x < 0$ 。

$$\iint_D y dx dy = \iint_{\text{矩形}} y dx dy - \iint_{\text{左半圆}} y dx dy.$$

矩形部分：

$$\int_0^2 y dy \int_{-2}^0 dx = 2 \cdot \int_0^2 y dy = 2 \cdot \left[\frac{y^2}{2} \right]_0^2 = 4.$$

半圆极坐标：令 $y - 1 = r \sin \theta$, $x = r \cos \theta$, $r \in [0, 1]$, $\theta \in [\pi/2, 3\pi/2]$, $y = 1 + r \sin \theta$ 。

$$\begin{aligned}\iint_{\text{左半圆}} y dx dy &= \int_{\pi/2}^{3\pi/2} \int_0^1 (1 + r \sin \theta) \cdot r dr d\theta \\ &= \int_{\pi/2}^{3\pi/2} \left(\frac{1}{2} + \frac{1}{3} \sin \theta \right) d\theta = \frac{\pi}{2} - \frac{2}{3}.\end{aligned}$$

$$\iint_D y dx dy = 4 - \left(\frac{\pi}{2} - \frac{2}{3} \right) = \frac{14}{3} - \frac{\pi}{2}.$$

3. 设矩阵 $A = \begin{pmatrix} 2 & 2 & -2 \\ 2 & 5 & -4 \\ -2 & -4 & a \end{pmatrix}$ 有一个特征值为 10.

(1) 求参数 a 的值.

解答：代入 $\lambda = 10$, $|10I - A| = 0$:

$$\begin{vmatrix} 8 & -2 & 2 \\ -2 & 5 & 4 \\ 2 & 4 & 10 - a \end{vmatrix} = 0 \implies a = 2.$$

(1) 求 A 的全部特征值.

解答：此时 $A = \begin{pmatrix} 2 & 2 & -2 \\ 2 & 5 & -4 \\ -2 & -4 & 2 \end{pmatrix}$, $\text{tr}(A) = 9$ 。

$$|\lambda I - A| = (\lambda - 10)(\lambda - 1)(\lambda + 1) = 0.$$

全部特征值： $\lambda_1 = 10$, $\lambda_2 = 1$, $\lambda_3 = -1$.



(1) 求正交矩阵 P 使得 $P^{-1}AP$ 为对角矩阵.

解答: 分别求各特征值对应的特征向量, 再施密特正交单位化。 $\lambda = 10$: 解 $(10I - A)x = 0$, 得 $\xi_1 = (1, 2, -2)^T$; $\lambda = 1$: 解 $(I - A)x = 0$, 得 $\xi_2 = (2, -1, 0)^T$; $\lambda = -1$: 解 $(-I - A)x = 0$, 得 $\xi_3 = (2, 4, 5)^T$ 。

这三个向量已经两两正交, 单位化:

$$p_1 = \frac{1}{3} \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}, \quad p_2 = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix}, \quad p_3 = \frac{1}{3\sqrt{5}} \begin{pmatrix} 2 \\ 4 \\ 5 \end{pmatrix}.$$

$$P = (p_1, p_2, p_3), \quad P^T AP = \text{diag}\{10, 1, -1\}.$$

P 即为所求正交矩阵。

4. 现有两枚硬币, 硬币 A 正面和反面出现的概率均为 $\frac{1}{2}$, 硬币 B 正面和反面出现的概率分别为 $\frac{3}{4}$ 和 $\frac{1}{4}$. 随机选择一枚硬币进行抛掷, 若前三次结果均为正面, 求第四次抛掷出现正面的概率.

解答: 贝叶斯公式。记事件 A : 选硬币 A ; B : 选硬币 B ; C_3 : 前三次都是正面。 $P(A) = P(B) = \frac{1}{2}$.

$$P(C_3|A) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}, \quad P(C_3|B) = \left(\frac{3}{4}\right)^3 = \frac{27}{64}.$$

$$P(A|C_3) = \frac{P(C_3|A)P(A)}{P(C_3|A)P(A) + P(C_3|B)P(B)} = \frac{\frac{1}{8} \cdot \frac{1}{2}}{\frac{1}{8} \cdot \frac{1}{2} + \frac{27}{64} \cdot \frac{1}{2}} = \frac{8}{35}.$$

$$P(B|C_3) = 1 - P(A|C_3) = \frac{27}{35}.$$

第四次正面概率:

$$P(\text{第4次正面}|C_3) = P(A|C_3) \cdot \frac{1}{2} + P(B|C_3) \cdot \frac{3}{4} = \frac{8}{35} \cdot \frac{1}{2} + \frac{27}{35} \cdot \frac{3}{4} = \frac{113}{140}.$$

5. 有 m 种不同的卡片, 每次抽到第 i 种卡片的概率为 p_i , 且满足 $\sum_{i=1}^m p_i = 1$ 。现有放回地抽取 n 次, 定义随机变量 X 为抽到的不同卡片种类数, 求 $E(X)$.

解答: 引入示性变量

$$I_i = \begin{cases} 1, & \text{第} i \text{种卡片被抽到过,} \\ 0, & \text{未被抽到} \end{cases}, \quad X = \sum_{i=1}^m I_i.$$

$$P(I_i = 0) = (1 - p_i)^n, \quad P(I_i = 1) = 1 - (1 - p_i)^n.$$

$$E(I_i) = 1 - (1 - p_i)^n.$$

$$E(X) = \sum_{i=1}^m E(I_i) = \sum_{i=1}^m [1 - (1 - p_i)^n].$$

6. 设总体 X 服从均匀分布 $U\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$, 令 $X_1, \dots, X_n$ 为从总体 X 中抽取的简单样本.

(1) 证明: 均值 $\hat{\theta}_1 = \bar{X}$ 和 $\hat{\theta}_2 = \frac{1}{2}(X_{(1)} + X_{(n)})$ 均为 θ 的无偏估计.

证明: 总体 $X \sim U\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$, $E(X) = \theta$ 。

$$E(\hat{\theta}_1) = E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \theta,$$

故 $\hat{\theta}_1$ 无偏。



令 $Y_i = X_i - \theta$, 则 $Y_i \sim U(-\frac{1}{2}, \frac{1}{2})$ 。 $Y_{(1)} = X_{(1)} - \theta$, $Y_{(n)} = X_{(n)} - \theta$ 。 $Y_{(1)}$ 与 $-Y_{(n)}$ 同分布, $\mathbb{E}(Y_{(1)}) = -\mathbb{E}(Y_{(n)})$ 。

$$\mathbb{E}\left(\frac{1}{2}(X_{(1)} + X_{(n)})\right) = \theta + \frac{1}{2}(\mathbb{E}(Y_{(1)}) + \mathbb{E}(Y_{(n)})) = \theta.$$

$\hat{\theta}_2$ 也是无偏估计。

(1) 求 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 的方差, 并说明哪一个更有效.

解答: 总体方差 $\text{Var}(X) = \frac{1}{12}$.

$$\text{Var}(\hat{\theta}_1) = \text{Var}(\bar{X}) = \frac{1}{12n}.$$

$$\text{Var}\left(\frac{X_{(1)} + X_{(n)}}{2}\right) = \frac{1}{2(n+1)(n+2)}.$$

$$\text{Var}(\hat{\theta}_2) = \frac{1}{2(n+1)(n+2)}.$$

比较: 对 $n \geq 2$, $\frac{1}{2(n+1)(n+2)} < \frac{1}{12n}$, $\text{Var}(\hat{\theta}_2) < \text{Var}(\hat{\theta}_1)$, 故 $\hat{\theta}_2$ 更有效。



中国人民大学统计学院
2023 年夏令营招生考试试题（应用统计）

注：考试时间 120 分钟。

1. 求极限 $\lim_{x \rightarrow 0} \left(\frac{e^x + e^{2x} + \cdots + e^{nx}}{n} \right)^{\frac{e}{x}}$ ，其中 n 是给定的正整数。

解答：取对数，令

$$L = \lim_{x \rightarrow 0} \frac{e}{x} \ln \left(\frac{e^x + e^{2x} + \cdots + e^{nx}}{n} \right).$$

泰勒展开 $e^{kx} = 1 + kx + \frac{1}{2}(kx)^2 + o(x)$,

$$\sum_{k=1}^n e^{kx} = n + \frac{n(n+1)}{2}x + o(x).$$

$$\frac{1}{n} \sum_{k=1}^n e^{kx} = 1 + \frac{n+1}{2}x + o(x).$$

利用 $\ln(1+t) \sim t$ ($t \rightarrow 0$):

$$\ln \left(\frac{1}{n} \sum_{k=1}^n e^{kx} \right) \sim \frac{n+1}{2}x.$$

$$L = \lim_{x \rightarrow 0} \frac{e}{x} \cdot \frac{n+1}{2}x = \frac{e(n+1)}{2}.$$

$$\text{原极限} = e^L = \exp \left(\frac{e(n+1)}{2} \right).$$

2. 求幂级数 $\sum_{n=0}^{\infty} \frac{x^{2n+2}}{(n+1)(2n+1)}$ 的和函数。

解答：令 $t = x^2$ ，级数化为 $S(t) = \sum_{n=0}^{\infty} \frac{t^{n+1}}{(n+1)(2n+1)}$ ，收敛域 $t \in [0, 1]$ ，即 $|x| \leq 1$ 。

$$S'(t) = \sum_{n=0}^{\infty} \frac{t^n}{2n+1}, \quad \sqrt{t} S'(t) = \sum_{n=0}^{\infty} \frac{t^{n+\frac{1}{2}}}{2n+1}.$$

求导：

$$(\sqrt{t} S'(t))' = \sum_{n=0}^{\infty} t^n = \frac{1}{1-t}, \quad 0 < t < 1.$$

积分：

$$\sqrt{t} S'(t) = \int_0^t \frac{1}{1-u} du = -\ln(1-t), \quad S'(t) = -\frac{\ln(1-t)}{\sqrt{t}}.$$

$$S(t) = \int_0^t -\frac{\ln(1-u)}{\sqrt{u}} du.$$

作换元 $u = v^2$,

$$S(t) = 2 \int_0^{\sqrt{t}} -\ln(1-v^2) dv = 2 \int_0^{\sqrt{t}} (-\ln(1-v) - \ln(1+v)) dv.$$

$$\int -\ln(1-v) dv = (1-v) \ln(1-v) + v, \quad \int -\ln(1+v) dv = (1+v) \ln(1+v) - v.$$

$$S(t) = 2 \left[(1-\sqrt{t}) \ln(1-\sqrt{t}) + (1+\sqrt{t}) \ln(1+\sqrt{t}) \right].$$

代回 $t = x^2$:

$$S(x^2) = 2 \left[(1-|x|) \ln(1-|x|) + (1+|x|) \ln(1+|x|) \right], \quad |x| \leq 1.$$



3. 已知矩阵

$$A = \begin{pmatrix} 1 & -a & -a \\ -a & 1 & -a \\ -a & -a & 1 \end{pmatrix}.$$

(1) 求 A 的特征值和特征向量。

解答:

$$A = (1+a)I - a\mathbf{1}\mathbf{1}^\top.$$

矩阵 $\mathbf{1}\mathbf{1}^\top$ 的特征值: 3 (对应特征向量 $\xi_1 = (1, 1, 1)^\top$), 0 (二重, 正交于 $(1, 1, 1)^\top$)。故 A 特征值:

$$\lambda_1 = 1 - 2a, \quad \lambda_2 = \lambda_3 = 1 + a.$$

$\lambda_1 = 1 - 2a$: 特征向量 $k_1(1, 1, 1)^\top$, $k_1 \neq 0$; $\lambda_{2,3} = 1 + a$: 所有满足 $x_1 + x_2 + x_3 = 0$ 的非零向量, 例如 $\xi_2 = (1, -1, 0)^\top$, $\xi_3 = (1, 1, -2)^\top$ 。

(1) 当 $a = 0.6$ 时, 判断是否存在随机向量 (X, Y, Z) 使其协方差矩阵恰好为 A 。解答: 代入 $a = 0.6$:

$$\lambda_1 = 1 - 2 \times 0.6 = -0.2 < 0, \quad \lambda_2 = \lambda_3 = 1.6 > 0.$$

协方差矩阵必须半正定, 所有特征值 ≥ 0 , 此处出现负特征值, **不存在**这样的随机向量。

4. 设矩阵 A 结构如下

$$A = \begin{pmatrix} 1 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 & 1 \\ x_{100} & x_{99} & x_{98} & \cdots & x_2 & x_1 \end{pmatrix},$$

项目的收益 $Y = |A|$, 随机变量 X_i ($i = 1, 2, \dots, 100$) 独立同分布于标准正态分布 $N(0, 1)$ 。现需确定储备资金 M 的最小值, 使得项目不破产的概率满足 $\mathbb{P}(Y + M \geq 0) \geq 0.95$, 求满足上述条件的 M 的最小值 (标准正态分布的 0.95 分位数取 1.645)。

解答: 按最后一行展开行列式。该矩阵为下 Hessenberg 矩阵, 递推: 记 D_k 为 k 阶同结构行列式,

$$D_k = x_1 D_{k-1} - x_2 D_{k-2} + x_3 D_{k-3} - \cdots + (-1)^{k-1} x_k.$$

这里 $k = 100$, 每一项都是 X_i 的线性组合, 所以 Y 是独立标准正态变量的线性组合, Y 服从正态分布。观察: $D_1 = x_1$, $D_2 = x_1 - x_2$, $D_3 = x_1 D_2 - x_2 D_1 + x_3$, $\mathbb{E}[X_i] = 0$, 故 $\mathbb{E}(Y) = 0$ 。

$$Y \sim N(0, \sigma_Y^2).$$

条件:

$$\mathbb{P}(Y + M \geq 0) = \mathbb{P}(Y \geq -M) = \Phi\left(\frac{M}{\sigma_Y}\right) \geq 0.95 \implies \frac{M}{\sigma_Y} \geq 1.645 \implies M \geq 1.645\sigma_Y.$$

由递推结构, 可求得 $\text{Var}(D_k) = k$, $\text{Var}(Y) = 100$, $\sigma_Y = 10$ 。

$$M_{\min} = 1.645 \times 10 = \mathbf{16.45}.$$

5. 在一个有 n 个人参加的晚会上, 每个人带了一件礼物, 且假定各人带的礼物都不相同。晚会期间各人从放在一起的 n 件礼物中随机抽取一件, 试求抽中自己礼品的人数 X 的均值和方差。



解答：引入示性变量

$$I_i = \begin{cases} 1, & \text{第} i \text{个人抽到自己礼物} \\ 0, & \text{其他} \end{cases}, \quad X = \sum_{i=1}^n I_i.$$

$$P(I_i = 1) = \frac{1}{n}, \quad \mathbb{E}(I_i) = \frac{1}{n}, \quad \mathbb{E}(X) = \sum_{i=1}^n \mathbb{E}(I_i) = 1.$$

$$\text{Var}(I_i) = \frac{1}{n} \left(1 - \frac{1}{n} \right).$$

当 $i \neq j$,

$$P(I_i = 1, I_j = 1) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}.$$

$$\text{Cov}(I_i, I_j) = \mathbb{E}(I_i I_j) - \mathbb{E}(I_i) \mathbb{E}(I_j) = \frac{1}{n(n-1)} - \frac{1}{n^2} = \frac{1}{n^2(n-1)}.$$

$$\text{Var}(X) = \sum_{i=1}^n \text{Var}(I_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(I_i, I_j) = n \cdot \frac{n-1}{n^2} + 2 \cdot \binom{n}{2} \cdot \frac{1}{n^2(n-1)} = 1.$$

6. 设某场景中有两种型号的电池，其使用时长分别服从参数为 λ_1 和 λ_2 的指数分布。随机选取电池时，选中型号 1 的概率为 p_1 ，选中型号 2 的概率为 p_2 （满足 $p_1 + p_2 = 1$ ）。已知在时刻 t 电池仍能正常工作，求该电池后续能继续使用至少 s 小时的条件概率。

解答：记事件： A_1 ：选型号 1； A_2 ：选型号 2； T ：电池寿命。指数分布生存函数 $P(T > u | A_i) = e^{-\lambda_i u}$ 。由贝叶斯公式：

$$P(A_1 | T > t) = \frac{p_1 e^{-\lambda_1 t}}{p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t}}, \quad P(A_2 | T > t) = \frac{p_2 e^{-\lambda_2 t}}{p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t}}.$$

指数无记忆性 $P(T > t + s | T > t, A_i) = P(T > s | A_i) = e^{-\lambda_i s}$ 。

$$\begin{aligned} P(T > t + s | T > t) &= P(A_1 | T > t) e^{-\lambda_1 s} + P(A_2 | T > t) e^{-\lambda_2 s} \\ &= \frac{p_1 e^{-\lambda_1(t+s)} + p_2 e^{-\lambda_2(t+s)}}{p_1 e^{-\lambda_1 t} + p_2 e^{-\lambda_2 t}}. \end{aligned}$$

7. 设随机变量 X 服从对数正态分布，即其对数变换 $Z = \ln X$ 服从正态分布 $N(\mu, \sigma^2)$ 。

(1) 计算均值 $\mathbb{E}(X)$ 。

解答： $X = e^Z$, $Z \sim N(\mu, \sigma^2)$ ，利用正态矩母函数：

$$\mathbb{E}(X) = \mathbb{E}(e^Z) = \exp\left(\mu + \frac{1}{2}\sigma^2\right).$$

(1) 已知 $x_1, x_2, \dots, x_n$ 是从该对数正态分布中抽取的独立同分布样本，求 μ 和 σ^2 的最大似然估计。

解答：令 $z_i = \ln x_i$ ，则 $z_1, \dots, z_n$ 是来自 $N(\mu, \sigma^2)$ 的样本。

$$\hat{\mu}_{\text{MLE}} = \bar{z} = \frac{1}{n} \sum_{i=1}^n \ln x_i, \quad \hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum_{i=1}^n (\ln x_i - \hat{\mu})^2.$$

(1) 当 $n = 20$ 时，给定具体样本值

$$\begin{array}{cccccccccc} 52 & 24 & 15 & 67 & 15 & 22 & 63 & 26 & 16 & 32 \\ 7 & 33 & 28 & 14 & 7 & 29 & 10 & 6 & 59 & 30 \end{array}$$

求均值 $\mathbb{E}(X)$ 的最大似然估计。

解答：MLE 具有不变性： $\mathbb{E}(X) = \exp(\mu + \frac{1}{2}\sigma^2)$,

$$\widehat{\mathbb{E}(X)}_{\text{MLE}} = \exp\left(\hat{\mu}_{\text{MLE}} + \frac{1}{2}\hat{\sigma}_{\text{MLE}}^2\right).$$

对样本取对数，计算 $\hat{\mu}, \hat{\sigma}^2$ ，代入上式即可得到数值结果。



中国人民大学统计与大数据研究院
2023 年夏令营招生考试试题（应用统计）

1. (1) 证明: $\left(\int_a^b f(x)dx\right)^2 = 2 \int_a^b f(x) \left(\int_a^x f(t)dt\right) dx$.

证明: 令 $F(x) = \int_a^x f(t)dt$, 则 $F'(x) = f(x)$, $F(a) = 0$, $F(b) = \int_a^b f(t)dt$. 右侧:

$$2 \int_a^b f(x)F(x)dx = 2 \int_a^b F(x) dF(x) = F^2(x)|_a^b = F^2(b) - F^2(a) = F^2(b) = \left(\int_a^b f(x)dx\right)^2,$$

等于左侧, 等式得证。

- (1) 判断 $f(x) = \frac{x^3}{1+x^2}$ 在 $\mathbb{R}$ 上的连续性和一致连续性.

解答: $f(x)$ 是有理函数, 分母 $1+x^2 > 0$, 故 $f(x)$ 在 $\mathbb{R}$ 上处处连续。

$$f(x) = x - \frac{x}{1+x^2}, \quad \lim_{x \rightarrow +\infty} f(x) = +\infty, \quad \lim_{x \rightarrow -\infty} f(x) = -\infty.$$

求导:

$$f'(x) = \frac{3x^2(1+x^2) - x^3 \cdot 2x}{(1+x^2)^2} = \frac{x^2(3+x^2)}{(1+x^2)^2}.$$

$\lim_{x \rightarrow \infty} f'(x) = 1$, 导数趋于 1, 函数斜率不会趋于 0, 故 $f(x)$ 在 $\mathbb{R}$ 上不一致连续。

- (1) 判断下列命题的正确性, 并给出证明或反例.

- i. 若方阵 A 满足 $A^2 = O$, 则 $A = O$.

解答: 命题错误。反例: $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, $A^2 = O$, 但 $A \neq O$ 。

- i. 若矩阵 A, B, C 满足 $AB = AC$, $A \neq O$, 则 $B = C$.

解答: 命题错误。反例: $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $B = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$, $C = O$, $A \neq O$, $AB = AC = O$, 但 $B \neq C$ 。

2. 设 $\Phi(x)$ 和 $\varphi(x)$ 分别为标准正态分布 $N(0, 1)$ 的分布函数和密度函数. 证明下列命题:

- (1) $\varphi'(x) + x\varphi(x) = 0$.

证明: 标准正态密度 $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ 。

$$\varphi'(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \cdot (-x) = -x\varphi(x),$$

移项即得 $\varphi'(x) + x\varphi(x) = 0$.

- (1) 当 $x > 0$ 时, $\varphi(x) \left(\frac{1}{x} - \frac{1}{x^3}\right) \leq 1 - \Phi(x) \leq \varphi(x) \left(\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5}\right)$.

证明: 正态尾部不等式 (Mill 比率界)。由分部积分:

$$1 - \Phi(x) = \int_x^{+\infty} \varphi(t)dt = \varphi(t) \left(-\frac{1}{t}\right) \Big|_x^{+\infty} - \int_x^{+\infty} \varphi(t) \frac{1}{t^2} dt = \frac{\varphi(x)}{x} - \int_x^{+\infty} \frac{\varphi(t)}{t^2} dt.$$

再分部积分剩余积分:

$$\int_x^{+\infty} \frac{\varphi(t)}{t^2} dt = \frac{\varphi(x)}{x^3} - 3 \int_x^{+\infty} \frac{\varphi(t)}{t^4} dt.$$

$$1 - \Phi(x) = \varphi(x) \left(\frac{1}{x} - \frac{1}{x^3}\right) + 3 \int_x^{+\infty} \frac{\varphi(t)}{t^4} dt.$$

余项积分 > 0 , 得到左边不等式:

$$1 - \Phi(x) \geq \varphi(x) \left(\frac{1}{x} - \frac{1}{x^3}\right).$$



继续再做一次分部积分，余项积分大于 0，得到上界：

$$1 - \Phi(x) \leq \varphi(x) \left(\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} \right).$$

证毕。

3. f 为正定二次函数 $f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top \mathbf{A} \mathbf{x} + \mathbf{b}^\top \mathbf{x}$, $\mathbf{d}^k$ 为下降方向, $\mathbf{x}^k$ 为当前迭代点.

(1) 试求出精确线搜索步长 $\alpha_k = \arg \min_{\alpha > 0} f(\mathbf{x}^k + \alpha \mathbf{d}^k)$.

解答：

$$f(\mathbf{x}^k + \alpha \mathbf{d}^k) = \frac{1}{2} (\mathbf{x}^k + \alpha \mathbf{d}^k)^\top \mathbf{A} (\mathbf{x}^k + \alpha \mathbf{d}^k) + \mathbf{b}^\top (\mathbf{x}^k + \alpha \mathbf{d}^k).$$

对 α 求导并置 0:

$$\frac{df}{d\alpha} = (\mathbf{x}^k + \alpha \mathbf{d}^k)^\top \mathbf{A} \mathbf{d}^k + \mathbf{b}^\top \mathbf{d}^k = 0.$$

整理：

$$((\mathbf{x}^k)^\top \mathbf{A} + \mathbf{b}^\top) \mathbf{d}^k + \alpha (\mathbf{d}^k)^\top \mathbf{A} \mathbf{d}^k = 0.$$

注意梯度 $\nabla f(\mathbf{x}^k) = \mathbf{A} \mathbf{x}^k + \mathbf{b}$,

$$\nabla f(\mathbf{x}^k)^\top \mathbf{d}^k + \alpha_k (\mathbf{d}^k)^\top \mathbf{A} \mathbf{d}^k = 0,$$

$$\alpha_k = - \frac{\nabla f(\mathbf{x}^k)^\top \mathbf{d}^k}{(\mathbf{d}^k)^\top \mathbf{A} \mathbf{d}^k}.$$

(1) 证明：最速下降法的步长 $\alpha_k = \frac{\|\nabla f(\mathbf{x}^k)\|^2}{\nabla f(\mathbf{x}^k)^\top \mathbf{A} \nabla f(\mathbf{x}^k)}$.

证明：最速下降方向取 $\mathbf{d}^k = -\nabla f(\mathbf{x}^k)$, 代入 (1) 结果：

$$\alpha_k = - \frac{\nabla f(\mathbf{x}^k)^\top (-\nabla f(\mathbf{x}^k))}{(-\nabla f(\mathbf{x}^k))^\top \mathbf{A} (-\nabla f(\mathbf{x}^k))} = \frac{\nabla f(\mathbf{x}^k)^\top \nabla f(\mathbf{x}^k)}{\nabla f(\mathbf{x}^k)^\top \mathbf{A} \nabla f(\mathbf{x}^k)} = \frac{\|\nabla f(\mathbf{x}^k)\|^2}{\nabla f(\mathbf{x}^k)^\top \mathbf{A} \nabla f(\mathbf{x}^k)}.$$

证毕。

4. 设样本 $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} B(1, p_1)$, $Y_1, Y_2, \dots, Y_m \stackrel{\text{i.i.d.}}{\sim} B(1, p_2)$, 其中 $B(1, p)$ 表示 0-1 二项分布. 定义统计量

$$n_1 = \sum_{i=1}^n \mathbb{I}\{X_i = 1\}, \quad n_0 = \sum_{i=1}^n \mathbb{I}\{X_i = 0\}, \quad m_1 = \sum_{j=1}^m \mathbb{I}\{Y_j = 1\}, \quad m_0 = \sum_{j=1}^m \mathbb{I}\{Y_j = 0\}.$$

(1) 设随机变量序列 $\{X_n\}$ 满足

$$\sqrt{n}(X_n - \mu) \xrightarrow{\mathcal{L}} N(0, \sigma^2),$$

其中 $\sigma^2 > 0$ 和 μ 均为有限数. $g(x)$ 的一阶导函数 $g'(x)$ 在 $x = \mu$ 处连续且非零, 证明:

$$\sqrt{n}[g(X_n) - g(\mu)] \xrightarrow{\mathcal{L}} N(0, [g'(\mu)]^2 \sigma^2).$$

证明：Delta 方法。泰勒展开：

$$g(X_n) = g(\mu) + g'(\mu)(X_n - \mu) + o_p(|X_n - \mu|).$$

$$\sqrt{n}(g(X_n) - g(\mu)) = g'(\mu) \cdot \sqrt{n}(X_n - \mu) + \sqrt{n} o_p(|X_n - \mu|).$$

由条件 $\sqrt{n}(X_n - \mu) = O_p(1)$, 故余项 $\sqrt{n} o_p(|X_n - \mu|) = o_p(1)$ 。依分布收敛性质：

$$g'(\mu) \cdot \sqrt{n}(X_n - \mu) \xrightarrow{\mathcal{L}} N(0, [g'(\mu)]^2 \sigma^2),$$

得到要证结论。



(1) 针对以下两个参数，构造其点估计并给出统计量的极限分布.

$$OR = \ln \left(\frac{p_1}{1-p_1} \right) - \ln \left(\frac{p_2}{1-p_2} \right), \quad RR = \ln \left(\frac{p_1}{p_2} \right).$$

解答：点估计用样本比例：

$$\hat{p}_1 = \frac{n_1}{n}, \quad \hat{p}_2 = \frac{m_1}{m}.$$

$$\widehat{OR} = \ln \frac{\hat{p}_1}{1-\hat{p}_1} - \ln \frac{\hat{p}_2}{1-\hat{p}_2}, \quad \widehat{RR} = \ln \frac{\hat{p}_1}{\hat{p}_2}.$$

$$\sqrt{n}(\hat{p}_1 - p_1) \xrightarrow{\mathcal{L}} N(0, p_1(1-p_1)), \quad \sqrt{m}(\hat{p}_2 - p_2) \xrightarrow{\mathcal{L}} N(0, p_2(1-p_2)).$$

设 $n, m \rightarrow \infty$, $\frac{n}{n+m} \rightarrow \lambda \in (0, 1)$ 。对 $g_1(p) = \ln \frac{p}{1-p}$, $g'_1(p) = \frac{1}{p(1-p)}$; 对 $g_2(p) = \ln p$, $g'_2(p) = \frac{1}{p}$ 。

$$\widehat{OR} \overset{\text{asympt}}{\sim} N \left(OR, \frac{1}{np_1(1-p_1)} + \frac{1}{mp_2(1-p_2)} \right),$$

$$\widehat{RR} \overset{\text{asympt}}{\sim} N \left(RR, \frac{1}{np_1} + \frac{1}{mp_2} \right).$$

(1) 基于 (2) 中结果求 OR 和 RR 的置信水平为 95% 的置信区间.

解答：95% 正态分位数 $z_{0.975} = 1.96$ 。

$$OR: \quad \widehat{OR} \pm 1.96 \sqrt{\frac{1}{n\hat{p}_1(1-\hat{p}_1)} + \frac{1}{m\hat{p}_2(1-\hat{p}_2)}},$$

$$RR: \quad \widehat{RR} \pm 1.96 \sqrt{\frac{1}{n\hat{p}_1} + \frac{1}{m\hat{p}_2}}.$$



中国人民大学统计与大数据研究院
2024 年夏令营招生考试试题（统计学）

注：考试时间 120 分钟。

1. (1) 求极限

$$\lim_{x \rightarrow +\infty} \left[\sqrt[3]{x^3 + x^2 + x + 2025} - \sqrt{x^2 + x + 2024} \cdot \frac{\ln(e^x + x)}{x} \right].$$

解答：先化简对数项：

$$\ln(e^x + x) = x + \ln(1 + xe^{-x}) \implies \frac{\ln(e^x + x)}{x} = 1 + o(1), \quad x \rightarrow +\infty.$$

泰勒展开立方根：

$$\sqrt[3]{x^3 + x^2 + x + 2025} = x \left(1 + \frac{1}{x} + \frac{1}{x^2} + o\left(\frac{1}{x^2}\right) \right)^{1/3} = x + \frac{1}{3} + o(1).$$

泰勒展开平方根：

$$\sqrt{x^2 + x + 2024} = x \left(1 + \frac{1}{x} + o\left(\frac{1}{x}\right) \right)^{1/2} = x + \frac{1}{2} + o(1).$$

$$\sqrt{x^2 + x + 2024} \cdot \frac{\ln(e^x + x)}{x} = x + \frac{1}{2} + o(1).$$

$$\lim_{x \rightarrow +\infty} \left[\left(x + \frac{1}{3} + o(1) \right) - \left(x + \frac{1}{2} + o(1) \right) \right] = -\frac{1}{6}.$$

(1) 计算二重积分 $\iint_{|x|+|y|\leq 2} f(x, y) dx dy$ ，其中函数

$$f(x, y) = \begin{cases} x^2, & |x| + |y| \leq 1 \\ \frac{1}{\sqrt{x^2 + y^2}}, & 1 < |x| + |y| \leq 2 \end{cases}.$$

解答：利用区域对称性，计算第一象限部分再乘以 4。

$$I = 4 \left(\iint_{\substack{x \geq 0, y \geq 0 \\ x+y \leq 1}} x^2 dx dy + \iint_{\substack{x \geq 0, y \geq 0 \\ 1 < x+y \leq 2}} \frac{1}{\sqrt{x^2 + y^2}} dx dy \right) = 4(I_1 + I_2).$$

$$I_1 = \int_0^1 x^2 dx \int_0^{1-x} dy = \int_0^1 x^2(1-x) dx = \frac{1}{12}.$$

极坐标变换 $x = r \cos \theta$, $y = r \sin \theta$, $x + y = r(\cos \theta + \sin \theta)$ 。

$$I_2 = \int_0^{\frac{\pi}{2}} \int_{\frac{1}{\cos \theta + \sin \theta}}^2 \frac{1}{r} \cdot r dr d\theta = \int_0^{\frac{\pi}{2}} \frac{d\theta}{\cos \theta + \sin \theta} = \sqrt{2} \ln(1 + \sqrt{2}).$$

$$I = 4 \left(\frac{1}{12} + \sqrt{2} \ln(1 + \sqrt{2}) \right) = \frac{1}{3} + 4\sqrt{2} \ln(1 + \sqrt{2}).$$

(1) 设 $f(x)$ 在 $[0, +\infty)$ 内连续，证明：

$$\int_0^x f(t)(x-t) dt = \int_0^x \left(\int_0^t f(u) du \right) dt.$$

证明：记 $F(t) = \int_0^t f(u) du$ ，则 $F'(t) = f(t)$ 。对左侧分部积分：

$$\int_0^x f(t)(x-t) dt = \int_0^x (x-t) dF(t) = (x-t)F(t)|_0^x + \int_0^x F(t) dt = \int_0^x \left(\int_0^t f(u) du \right) dt.$$

等式得证。也可以交换二重积分次序证明。



- (1) 证明：行满秩矩阵具有右消去律，即若矩阵 A 行满秩且 $AC = C$ ，则 $A = I$ 。

证明：本题条件下 A 是方阵，行满秩方阵必可逆。

$$AC = C \implies (A - I)C = 0.$$

该等式对矩阵 C 成立，取 $C = I$ ，代入得到 $(A - I)I = 0$ ，即 $A - I = 0$ ， $A = I$ 。

2. 考虑线性混合模型

$$y_{ij} = \mu + \alpha_j + \beta x_{ij} + \varepsilon_{ij}, \quad \alpha_j \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma_\alpha^2), \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma_\varepsilon^2)$$

其中所有 α_j 与 ε_{ij} 相互独立， $i = 1, 2, \dots, n$ ， $j = 1, 2, \dots, m$ 。

- (1) 给出 $\mathbf{y}_j = (y_{1j}, y_{2j}, \dots, y_{nj})^\top$ 的分布。

解答：

$$\mathbf{y}_j = \mu \mathbf{1}_n + \beta \mathbf{x}_j + \mathbf{1}_n \alpha_j + \boldsymbol{\varepsilon}_j,$$

$$\mathbb{E} \mathbf{y}_j = \mu \mathbf{1}_n + \beta \mathbf{x}_j, \quad \text{协方差矩阵 } \text{Cov}(\mathbf{y}_j) = \sigma_\alpha^2 \mathbf{1}_n \mathbf{1}_n^\top + \sigma_\varepsilon^2 \mathbf{I}_n.$$

$$\mathbf{y}_j \sim N(\mu \mathbf{1}_n + \beta \mathbf{x}_j, \sigma_\alpha^2 \mathbf{1}_n \mathbf{1}_n^\top + \sigma_\varepsilon^2 \mathbf{I}_n).$$

- (1) 若 σ_ε^2 已知，求 $\boldsymbol{\theta} = (\mu, \beta)^\top$ 的最大似然估计。

解答：记 $\mathbf{V}_j = \sigma_\alpha^2 \mathbf{1}_n \mathbf{1}_n^\top + \sigma_\varepsilon^2 \mathbf{I}_n$ ，设计矩阵 $\mathbf{X}_j = \begin{pmatrix} \mathbf{1}_n & \mathbf{x}_j \end{pmatrix}$ 。对数似然最大化等价广义最小二乘：

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \left(\sum_{j=1}^m \mathbf{X}_j^\top \mathbf{V}_j^{-1} \mathbf{X}_j \right)^{-1} \left(\sum_{j=1}^m \mathbf{X}_j^\top \mathbf{V}_j^{-1} \mathbf{y}_j \right).$$

- (1) 若 σ_ε^2 未知，写出给定 $\mathbf{y}_j$ 时 α_j 的条件分布，并求 $\mathbb{E}(\alpha_j | \mathbf{y}_j)$ 和 $\text{Var}(\alpha_j | \mathbf{y}_j)$ 。

解答：联合正态，条件分布为正态：

$$\alpha_j | \mathbf{y}_j \sim N(\mathbb{E}(\alpha_j | \mathbf{y}_j), \text{Var}(\alpha_j | \mathbf{y}_j)).$$

$$\mathbb{E}(\alpha_j | \mathbf{y}_j) = \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2 + n\sigma_\alpha^2} \mathbf{1}_n^\top (\mathbf{y}_j - \mu \mathbf{1}_n - \beta \mathbf{x}_j),$$

$$\text{Var}(\alpha_j | \mathbf{y}_j) = \frac{\sigma_\alpha^2 \sigma_\varepsilon^2}{\sigma_\varepsilon^2 + n\sigma_\alpha^2}.$$

- (1) 若 σ_ε^2 未知，设计算法估计 β 和 σ_ε^2 的值。

解答：可使用 EM 算法：

- E 步：给定当前参数，计算随机效应 α_j 的条件期望 $\mathbb{E}(\alpha_j | \mathbf{y}_j)$ ；
- M 步：把条件期望当作已知观测，更新固定效应 μ, β 以及方差参数 $\sigma_\varepsilon^2, \sigma_\alpha^2$ ；
- 循环 E 步 M 步直到参数收敛，输出 $\hat{\beta}, \hat{\sigma}_\varepsilon^2$ 。

也可以采用 REML 限制极大似然直接数值优化得到估计。

3. 设 $X_1, X_2, \dots, X_n$ 为从总体 X 中抽取的简单样本，定义统计量

$$Y_k = \frac{X_1 + \dots + X_k}{X_1 + \dots + X_n}, \quad k = 1, 2, \dots, n.$$

另设 $U_{(1)} \leq U_{(2)} \leq \dots \leq U_{(n)}$ 是来自均匀分布 $U(0, 1)$ 总体的次序统计量。

- (1) 证明： $\mathbb{E}(Y_k) = \frac{k}{n}$ 。

证明：记 $S_n = \sum_{i=1}^n X_i$ 。由于样本独立同分布，

$$\mathbb{E}\left(\frac{X_1}{S_n}\right) = \mathbb{E}\left(\frac{X_2}{S_n}\right) = \dots = \mathbb{E}\left(\frac{X_n}{S_n}\right).$$



$$\sum_{i=1}^n \mathbb{E} \left(\frac{X_i}{S_n} \right) = \mathbb{E} \left(\frac{S_n}{S_n} \right) = 1 \implies \mathbb{E} \left(\frac{X_i}{S_n} \right) = \frac{1}{n}.$$

$$\mathbb{E}(Y_k) = \mathbb{E} \left(\frac{X_1 + \cdots + X_k}{S_n} \right) = \sum_{i=1}^k \mathbb{E} \left(\frac{X_i}{S_n} \right) = \frac{k}{n}.$$

- (1) 设总体分布为指数分布 $\text{Exp}(1)$, 判断 $U_{(k)}$ 与 Y_k 是否具有相同的分布。

解答: 二者分布相同。指数样本的部分和比例服从均匀次序统计量的分布。

- (1) 设总体分布为指数分布 $\text{Exp}(1)$, 证明: Y_k 与 $\sum_{i=1}^n X_i$ 相互独立。

证明: 令 $S_n = \sum_{i=1}^n X_i$, 做变换 $(X_1, \dots, X_n) \mapsto (Y_1, \dots, Y_{n-1}, S_n)$ 。雅可比行列式 $J = S_n^{n-1}$, 联合密度可以分解为:

$$f(y_1, \dots, y_{n-1}, s) = \underbrace{(n-1)!}_{Y_1, \dots, Y_{n-1} \text{ 的联合密度}} \cdot \underbrace{\frac{s^{n-1} e^{-s}}{\Gamma(n)}}_{S_n \text{ 的 Gamma 密度}}.$$

变量分离, $(Y_1, \dots, Y_{n-1})$ 与 S_n 独立。 Y_k 是 $Y_1, \dots, Y_{n-1}$ 的函数, 故 Y_k 与 S_n 独立。

- (1) 证明: $n(1 - U_{(n)}) \xrightarrow{L} \text{Exp}(1)$ 。

证明: $U_{(n)}$ 的分布函数: $P(U_{(n)} \leq t) = t^n, 0 < t < 1$ 。令 $Z_n = n(1 - U_{(n)})$,

$$P(Z_n \leq x) = P\left(U_{(n)} \geq 1 - \frac{x}{n}\right) = 1 - \left(1 - \frac{x}{n}\right)^n \rightarrow 1 - e^{-x}, n \rightarrow \infty, x > 0.$$

极限分布是 $\text{Exp}(1)$, 依分布收敛成立。



中国人民大学统计与大数据研究院
2024 年夏令营招生考试试题（应用统计）

注：考试时间 120 分钟。

1. (1) 求极限 $\lim_{n \rightarrow \infty} \frac{n}{n^2+1} + \frac{n}{n^2+2^2} + \cdots + \frac{n}{n^2+n^2}$.

解答：转化为定积分定义：

$$\sum_{k=1}^n \frac{n}{n^2+k^2} = \sum_{k=1}^n \frac{1}{1+\left(\frac{k}{n}\right)^2} \cdot \frac{1}{n}.$$

$$\lim_{n \rightarrow \infty} \sum_{k=1}^n \frac{1}{1+\left(\frac{k}{n}\right)^2} \cdot \frac{1}{n} = \int_0^1 \frac{1}{1+x^2} dx = \arctan x \Big|_0^1 = \frac{\pi}{4}.$$

原极限 = $\frac{\pi}{4}$.

- (1) 证明：函数 $f(x) = \frac{|\sin x|}{x}$ 在区间 $[-\pi, 0)$ 和 $(0, \pi]$ 均为一致连续，但在 $[-\pi, 0) \cup (0, \pi]$ 上不一致连续。

证明：1. 在 $(0, \pi]$: $f(x) = \frac{\sin x}{x}$, $\lim_{x \rightarrow 0^+} f(x) = 1$, 将 f 在 $x=0$ 补充定义为 1, 得到闭区间 $[0, \pi]$ 上连续函数, 闭区间连续必一致连续, 故限制在 $(0, \pi]$ 上一致连续。

2. 在 $[-\pi, 0)$: $|\sin x| = |\sin(-x)|$, $f(x) = \frac{|\sin x|}{x} = -\frac{|\sin(-x)|}{-x}$, 令 $t = -x \in (0, \pi]$, $f(x) = -f(t)$ 。
 $\lim_{x \rightarrow 0^-} f(x) = -1$, 补充定义后在 $[-\pi, 0]$ 连续, 故 $[-\pi, 0)$ 上一致连续。

3. 在集合 $[-\pi, 0) \cup (0, \pi]$: 取序列 $x_n = \frac{1}{n} > 0$, $y_n = -\frac{1}{n} < 0$,

$$|x_n - y_n| = \frac{2}{n} \rightarrow 0, \quad |f(x_n) - f(y_n)| = \left| \frac{|\sin \frac{1}{n}|}{\frac{1}{n}} - \frac{|\sin(-\frac{1}{n})|}{-\frac{1}{n}} \right| \rightarrow |1 - (-1)| = 2.$$

两点距离趋于 0, 但函数值差不趋于 0, 故在 $[-\pi, 0) \cup (0, \pi]$ 上不一致连续。

2. 设随机变量 X 服从 Pareto 分布, 概率密度函数

$$f(x) = \frac{\beta \alpha^\beta}{x^{\beta+1}} \mathbb{I}\{x \geq \alpha\} \quad (\alpha > 0, \beta > 2).$$

- (1) 计算 $\mathbb{E}(X)$ 和 $\text{Var}(X)$.

解答：

$$\mathbb{E}(X) = \int_{\alpha}^{+\infty} x \cdot \frac{\beta \alpha^\beta}{x^{\beta+1}} dx = \beta \alpha^\beta \int_{\alpha}^{+\infty} x^{-\beta} dx = \frac{\beta \alpha}{\beta - 1}.$$

$$\mathbb{E}(X^2) = \int_{\alpha}^{+\infty} x^2 \cdot \frac{\beta \alpha^\beta}{x^{\beta+1}} dx = \beta \alpha^\beta \int_{\alpha}^{+\infty} x^{-(\beta-1)} dx = \frac{\beta \alpha^2}{\beta - 2}.$$

$$\text{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \frac{\beta \alpha^2}{\beta - 2} - \left(\frac{\beta \alpha}{\beta - 1} \right)^2 = \frac{\beta \alpha^2}{(\beta - 1)^2 (\beta - 2)}.$$

- (1) 求随机变量 $Y = 2\beta(\ln X - \ln \alpha)$ 的分布。

解答： $Y = 2\beta \ln \frac{X}{\alpha} \implies X = \alpha \exp\left(\frac{Y}{2\beta}\right)$, $Y \geq 0$. 做变量替换, $x \geq \alpha \iff y \geq 0$, 且

$$\frac{dx}{dy} = \alpha \cdot \exp\left(\frac{y}{2\beta}\right) \cdot \frac{1}{2\beta}.$$

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = \frac{\beta \alpha^\beta}{(\alpha e^{y/(2\beta)})^{\beta+1}} \cdot \alpha e^{y/(2\beta)} \cdot \frac{1}{2\beta} = \frac{1}{2} e^{-y/2}, \quad y \geq 0.$$

即 $Y \sim \chi^2(2)$ (自由度为 2 的卡方分布, 等价于参数为 $\lambda = \frac{1}{2}$ 的指数分布)。

3. 设独立观测样本满足以下条件：



(a) $X_i^{(1)}$ 和 $X_i^{(0)}$ 分别服从参数为 λ_1, λ_0 的 Poisson 分布 $P(\lambda_1), P(\lambda_0)$.

(b) T_i 服从两点分布 $B(1, p)$, 且与 $X_i^{(1)}, X_i^{(0)}$ 独立。

观测值为 $X_i = T_i X_i^{(1)} + (1 - T_i) X_i^{(0)}$.

(1) 求 λ_0 和 λ_1 的无偏估计量 $\hat{\lambda}_0$ 和 $\hat{\lambda}_1$, 并证明其无偏性。

解答: $\mathbb{E}(X_i) = \mathbb{E}(T_i X_i^{(1)} + (1 - T_i) X_i^{(0)}) = p\lambda_1 + (1 - p)\lambda_0$. 按来源分组: $\mathbb{E}(X_i | T_i = 1) = \lambda_1$, $\mathbb{E}(X_i | T_i = 0) = \lambda_0$. 记 $n_T = \sum_{i=1}^n T_i$, $n - n_T = \sum_{i=1}^n (1 - T_i)$, 取

$$\hat{\lambda}_1 = \frac{1}{n_T} \sum_{i:T_i=1} X_i, \quad \hat{\lambda}_0 = \frac{1}{n - n_T} \sum_{i:T_i=0} X_i.$$

由条件期望, $\mathbb{E}(\hat{\lambda}_1 | \{T_i\}) = \lambda_1$, 故 $\mathbb{E}(\hat{\lambda}_1) = \lambda_1$; 同理 $\mathbb{E}(\hat{\lambda}_0) = \lambda_0$.

(1) 推导 $\hat{\lambda}_1 - \hat{\lambda}_0$ 和 $\frac{\hat{\lambda}_1}{\hat{\lambda}_0}$ 的渐近分布.

解答: 当 $n \rightarrow \infty$, $n_T/n \xrightarrow{P} p$, $(n - n_T)/n \xrightarrow{P} 1 - p$,

$$\hat{\lambda}_1 \stackrel{\text{asympt}}{\sim} N\left(\lambda_1, \frac{\lambda_1}{np}\right), \quad \hat{\lambda}_0 \stackrel{\text{asympt}}{\sim} N\left(\lambda_0, \frac{\lambda_0}{n(1-p)}\right),$$

由渐近独立性得

$$\hat{\lambda}_1 - \hat{\lambda}_0 \stackrel{\text{asympt}}{\sim} N\left(\lambda_1 - \lambda_0, \frac{\lambda_1}{np} + \frac{\lambda_0}{n(1-p)}\right).$$

对 $g(u, v) = u/v$ 使用 Delta 方法:

$$\frac{\hat{\lambda}_1}{\hat{\lambda}_0} \stackrel{\text{asympt}}{\sim} N\left(\frac{\lambda_1}{\lambda_0}, \frac{1}{n} \left(\frac{\lambda_1}{p\lambda_0^2} + \frac{\lambda_1^2}{\lambda_0^3(1-p)}\right)\right).$$

(1) 确定 p 的取值, 使得 $\hat{\lambda}_1 - \hat{\lambda}_0$ 和 $\frac{\hat{\lambda}_1}{\hat{\lambda}_0}$ 的渐近方差最小.

解答: (1) 对 $\hat{\lambda}_1 - \hat{\lambda}_0$, 等价于最小化 $V_1(p) = \frac{\lambda_1}{p} + \frac{\lambda_0}{1-p}$, $p \in (0, 1)$. 令导数为 0:

$$-\frac{\lambda_1}{p^2} + \frac{\lambda_0}{(1-p)^2} = 0 \implies \frac{1-p}{p} = \sqrt{\frac{\lambda_0}{\lambda_1}} \implies p = \frac{\sqrt{\lambda_1}}{\sqrt{\lambda_0} + \sqrt{\lambda_1}}.$$

(2) 对比值 $\frac{\hat{\lambda}_1}{\hat{\lambda}_0}$, 渐近方差可写成 $\frac{1}{n}$ 的倍数:

$$V_2(p) = \frac{\lambda_1}{p\lambda_0^2} + \frac{\lambda_1^2}{\lambda_0^3(1-p)} = \frac{C_1}{p} + \frac{C_2}{1-p}, \quad C_1, C_2 > 0.$$

$$-\frac{C_1}{p^2} + \frac{C_2}{(1-p)^2} = 0 \implies \frac{1-p}{p} = \sqrt{\frac{C_2}{C_1}} = \sqrt{\frac{\lambda_1}{\lambda_0}} \implies p = \frac{\sqrt{\lambda_0}}{\sqrt{\lambda_0} + \sqrt{\lambda_1}}.$$

(1) 构造 $\hat{\lambda}_1 - \hat{\lambda}_0$ 和 $\frac{\hat{\lambda}_1}{\hat{\lambda}_0}$ 的置信水平为 95% 的近似置信区间.

解答: 取标准正态分布的分位数 $z_{0.975} = 1.96$:

$$\hat{\lambda}_1 - \hat{\lambda}_0 \pm 1.96 \sqrt{\frac{\hat{\lambda}_1}{np} + \frac{\hat{\lambda}_0}{n(1-p)}}.$$

$$\frac{\hat{\lambda}_1}{\hat{\lambda}_0} \pm 1.96 \sqrt{\frac{1}{n} \left(\frac{\hat{\lambda}_1}{p\hat{\lambda}_0^2} + \frac{\hat{\lambda}_1^2}{\hat{\lambda}_0^3(1-p)} \right)}.$$