信息学中的概率统计：作业七

截止日期：2025 年 12 月 24 日（周三）下午 3 点前。请务必通过教学网提交电子版。可下课前同时提交纸质版。

与本次作业相关的事宜，请发邮件给我（ruosongwang@pku.edu.cn），抄送研究生助教叶昊洋（yhyfhgs@gmail.com），以及负责本次作业的本科生助教方嘉聪（jiacong_fang@stu.pku.edu.cn）。

注意：本次作业第五题为**附加题**，正确解决该题目可以得到额外 10% 的分数。

第一题

令总体 X 服从概率密度函数如下的连续分布，其中 $\theta > 0$ 为未知参数，

$$f(x) = \begin{cases} \frac{\theta}{x^2} & x \geq \theta \\ 0 & x < \theta \end{cases}。$$

给定简单随机样本 $X_1, X_2, \ldots, X_n$，本题中，我们假设样本量 $n > 2$。给出 θ 的最大似然估计量 $\hat{\theta}$，判断 $\hat{\theta}$ 是否为无偏估计量及渐进无偏估计量，是否为一致估计量，并计算其均方误差。基于 $\hat{\theta}$，构造置信水平为 $1 - \alpha$ 的置信区间。

第二题

令总体 $X \sim \mathrm{Exp}(\lambda)$, 其中参数 λ 为未知参数。给定简单随机样本 $X_1, X_2, \cdots, X_n$，在课上，我们给出了 λ 的一个矩法估计量 $\hat{\lambda}_0 = 1/\overline{X}$。本题中，我们假设样本量 $n > 2$。

(1) 判断 $\hat{\lambda}_0$ 是否为无偏估计量及渐进无偏估计量。提示: 回顾伽玛函数及伽玛分布的定义和性质。

(2) 基于上一问的结果，构造 λ 的无偏估计量 $\hat{\lambda}_1$。

(3) 计算 $\hat{\lambda}_1$ 的均方误差, 并判断 $\hat{\lambda}_1$ 是否为一致估计量。

(4) 给定 $0 < \alpha < 1$, 基于样本均值 $\overline{X}$，构造统计量 $\hat{\lambda}_L$ 与 $\hat{\lambda}_R$ 使得 $P\left(\hat{\lambda}_L \leq \lambda \leq \hat{\lambda}_R\right) = 1 - \alpha$，也即构造 λ 的置信水平为 $1 - \alpha$ 的置信区间。最终的结果应依赖于 χ^2 分布的分布函数的逆函数 F^{-1}。(提示: 给出 $2\lambda X_i$ 服从的分布)

第三题

令总体 X 服从概率密度函数如下的连续分布,

$$f(x) = \begin{cases} (\theta + 1)x^{\theta} & 0 < x < 1 \\ 0 & x \notin (0, 1) \end{cases},$$

其中 $\theta > 0$ 为未知参数. 给定简单随机样本 $X_1, X_2, \cdots, X_n$。本题中，我们假设样本量 $n > 2$。

(1) 计算总体 X 的分布函数 $F_X(x) = P(X \leq x)$.

(2) 计算总体 X 的数学期望 $E(X)$。用 $E(X)$ 表示未知参数 θ, 并将 $E(X)$ 替换为样本均值 $\overline{X}$，从而给出 θ 的矩估计量 $\hat{\theta}_1$。

(3) 计算 θ 的最大似然估计量 $\hat{\theta}_2$。判断 $\hat{\theta}_2$ 是否为无偏估计量及渐进无偏估计量。(提示: 给出 $-\ln X_i$ 服从的分布)

(4) 计算 $\hat{\theta}_2$ 的均方误差, 并判断 $\hat{\theta}_2$ 是否为一致估计量.。

第四题

令总体 $X \sim \pi(\lambda)$，也即参数为 λ 的泊松分布，λ 为未知参数。给定简单随机样本 $X_1, X_2, \ldots, X_n$，本题中，我们将考虑 $p = e^{-\lambda}$ 的两个不同的估计量。

(1) 考虑 p 的矩法估计量 $\hat{p}_1 = e^{-\overline{X}}$。这里，$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 为样本均值。判断 $\hat{p}_1$ 是否为 $p = e^{-\lambda}$ 的最大似然估计（简要说明原因，无需严格证明），判断 $\hat{p}_1$ 是否为无偏估计量，渐进无偏估计量，一致估计量，并计算 $\hat{p}_1$ 的均方误差。提示：参考作业二第四题。

(2) 令 $\hat{p}_2 = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i=0}$。这里

$$1_{X_i=0} = \begin{cases} 1 & X_i = 0 \\ 0 & X_i > 0 \end{cases}。$$

判断 $\hat{p}_2$ 是否为无偏估计量，渐进无偏估计量，一致估计量，并计算 $\hat{p}_2$ 的均方误差。

第五题（附加题）

在课上，我们考虑了下述模型：给定 n 台游戏机，第 i 台游戏机的中奖概率为 $0 \le p_i \le 1$，且 p_i 均为未知参数。在第 t 轮中，选择一台游戏机 $1 \le i \le n$，并观测到结果 $X_t \sim B(1, p_i)$。这里 $X_1, X_2, \ldots$ 相互独立。

在课上，我们考虑了下述均匀采样策略：对每台游戏机进行 N 次观测，并返回样本均值最大的游戏机。若取 $N = O(\ln n/\epsilon^2)$，则有 $P(p_o \ge \max p_i - \epsilon) \ge 2/3$，这里 $1 \le o \le n$ 为策略返回的选择。注意到，该策略需要的总样本数量为 $nN = O(n \ln n/\epsilon^2)$。在本题中，我们将给出更好的算法，将总样本数量改进至 $O(n/\epsilon^2)$。

新的算法共有若干轮。对于第 ℓ 轮（$\ell \ge 1$），我们将所有的候选游戏机（也即未被移除的游戏机）进行均匀采样，每台游戏机的采样数量为 $C \cdot \ln(1/\delta_\ell)/\epsilon_\ell^2$，这里 $C > 0$ 为某个常数，并计算每台游戏机这一轮采样结果的样本均值。我们将样本均值落在后一半的游戏机移除，并进入到下一轮。具体来说，若第 ℓ 轮开始前未被移除的游戏机数量为 n_ℓ，计算这一轮采样结果的样本均值后，我们将全部游戏机按照样本均值排序，并移除均值排序最小的 $\lceil n_\ell/2 \rceil$ 台游戏机。当只剩一个候选游戏机时，我们将该台游戏机的编号作为 o 返回。

(1) 对于任意 $\ell \ge 1$，令 S_ℓ 为第 ℓ 轮结束后的候选游戏机（也即未被移除的游戏机）。令 $S_0 = \{1, 2, \ldots, n\}$。对于任意 $1 \le \ell \le L$，证明

$$P\left(\max_{i \in S_\ell} p_i \ge \max_{i \in S_{\ell-1}} p_i - \epsilon_\ell\right) \ge 1 - \delta_\ell。$$

(2) 对于任意 $\ell \ge 1$，定义合适的 ϵ_ℓ 和 δ_ℓ，使得 $P(p_o \ge \max p_i - \epsilon) \ge 2/3$，且算法的总采样次数为 $O(n/\epsilon^2)$。

本题为附加题，正确解决本题可以得到额外 10% 的分数。