华东师范大学期中试卷

2022-2023 学年第 2 学期

课程名称：数据科学与工程算法 课程性质：专业必修

专业：_____ 年级：_____

姓名：_____ 学号：_____

/	一	二	三	四	五	总分
得分						

一. 填空题 （本大题共有 8 题，满分 48 分，每题 6 分.）

1. 采用圆形等距抽样算法从总体 1-21 中抽取样本. 抽样间距为 4，第一个被抽样的元素编号为 15，请问接下去被抽样的四个元素编号依次是 _____ .
2. 假设抛一枚正面向上概率为 $\frac{3}{4}$ 的硬币 1000 次，随机变量 X 定义为硬币正面朝上的次数，使用 Chebyshev 不等式估算 $X > 900$ 的概率上界为 _____ .
3. 假设抛一枚正面向上概率为 $\frac{1}{5}$ 的硬币 800 次，随机变量 X 定义为硬币正面朝上的次数，使用 Chernoff 不等式估算 $X < 40$ 的概率上界为 _____ .
4. 当哈希函数 $h(x) = (3x + 1)mod5$ 被用于行排列变换时，集合 $A = \{0, 1, 4\}$ 和 $B = \{2, 3, 4\}$ 的最小哈希值分别为 _____ .
5. 设 $k = 3$，使用 Misra Gries 算法求得输入数据流 $< a, b, b, c, c, a, a, d >$ 中的频繁元素为 _____ .
6. 对于输入数据流 $< 0, 1, 1, 2, 3, 3 >$，假设给定哈希函数 $h_1(x) = (2x + 1)mod3$ 和 $h_2(x) = (x + 1)mod3$，用 CM Sketch 估计元素 0 的频度为 _____ .
7. 对于数据流 $< 0, 0, 1, 2, 2, 3, 3, 3 >$, 假设给定哈希函数 $h(x) = (7x + 2)mod3$ 和 $g(x) = \begin{cases} +1 & \text{if x mod } 2 = 0 \\ -1 & \text{if x mod } 2 = 1 \end{cases}$，用 Count Sketch 估计元素 1 的频度为 _____ .
8. 对于转移概率矩阵为 $\mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{3}{4} \\ p & 1-p & 0 \end{pmatrix}$ 的马尔可夫链，已知其平稳分布为 $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$，则参数 p 的取值是 _____ .

二. （本题满分 20 分，其中每题 10 分.）

9. 设一组独立随机变量 $x_{ij}(i = 1, ..., k; j = 1, ..., n)$ 服从参数为 p 的伯努利分布. 定义随机变量 $X_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}, i = 1, ..., k.$

   (1) 设随机变量 $Y = min_{1\le i\le k} X_i$，计算事件 $Y > (1+\epsilon)p$ 的概率上界；

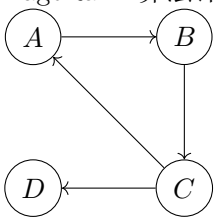
   (2) 设随机变量 $Z = median_{1\le i\le k} X_i$，计算事件 $|Z - p| > \epsilon p$ 的概率上界.

三. （本题满分 12 分.）

10. 给定两个集合 A,B 各包含 50 亿个元素，每个元素占用 $64B$。当内存使用被限制在 $4 \times 10^9 B$ 时，设计恰当的方案计算集合 A 和 B 的交集，并分析方案的误差.

四. （本题满分 10 分.）

11. 若存在四个网站 A、B、C、D，其链接关系如图所示. 使用随机跳转参数 $p = 0.2$ 的改进版 PageRank 算法计算每个网站的 PageRank 值.



五. （本题满分 10 分.）

12. 使用 Flajolet-Martin 算法估算数据流 $< 3, 1, 4, 1, 8, 2, 9 >$ 中不同元素的个数，使用哈希函数 $h_1(x) = (2x + 1) mod 16$ 和 $h_2(x) = (4x) mod 16$. 请给出算法运行结果.