

《抽样技术》上机任务说明

抽样设计性能的模拟比较研究

黄娟

2026 年 4 月 16 日

本次上机任务的定位

今天不是做一道普通的抽样计算题，而是完成一个小型方法研究。

你们要做的不只是：

- 抽一次样
- 算一个均值
- 比较几个点估计

你们真正要做的是：

- 提出研究问题
- 设计多种抽样方案
- 进行重复抽样模拟
- 比较精度、偏差、稳定性和成本
- 分析结论在不同情景下是否成立

关键词：研究问题、研究假设、重复抽样、情景分析、偏差分析、成本约束

某市拟开展“居民家庭消费与数字生活方式调查”，调查目标包括：

- ① 估计全市家庭月消费均值
- ② 估计不同区域家庭月消费均值
- ③ 估计高收入家庭占比
- ④ 估计在线消费支出占比

教师将提供：

- 家庭层总体框架数据
- 可选的社区层抽样框
- 不同总体结构或偏差机制情景

本次任务要求你基于给定总体，比较不同抽样设计在复杂调查条件下的表现。

为什么这是研究型任务

本任务的核心不是问“怎么写代码”，而是问：

- 哪种抽样设计更优？
- 优势体现在哪些指标上？
- 这种优势依赖于哪些总体结构条件？
- 成本与非抽样误差会不会改变结论？

因此，最后要提交的不是一堆代码，而是一组有证据支持的研究结论。

本次任务的总体要求

你必须完成以下内容：

- ① 提出一个明确的研究问题
- ② 提出至少 2 个研究假设
- ③ 设计至少 3 种抽样方案
- ④ 进行重复抽样模拟
- ⑤ 比较 Bias、标准差、MSE 等指标
- ⑥ 至少做 1 个情景分析
- ⑦ 至少做 1 个偏差分析
- ⑧ 在成本约束下比较方案
- ⑨ 给出最终推荐方案及研究结论

数据文件说明

教师提供的数据文件通常包括：

- household_population.csv
- community_frame.csv (可选)

典型变量包括：

- hh_id: 家庭编号
- district_type: 区域类型
- community_id: 社区编号
- community_size: 社区家庭数
- income_level: 收入等级
- family_size: 家庭人口数
- elderly_flag: 是否有老年人
- total_consume: 家庭月消费总支出
- online_consume: 在线消费支出
- response_prob: 理论响应概率

注意：总体直值在本任务中可用于评价设计效果。

第一步：提出研究问题

请围绕“抽样设计比较”提出一个明确、具体、可验证的研究问题。

示例：

- 区域差异较大时，分层抽样是否优于简单随机抽样？
- 社区聚集性较强时，两阶段抽样的 MSE 是否更高？
- 社区规模差异较大时，PPS 是否更有效？
- 非应答存在时，哪种设计更稳健？

不要只写：“比较几种抽样方法”

这太泛，缺乏研究型任务应有的聚焦性。

第二步：提出研究假设

请提出至少 2 个具有方向性的研究假设。

示例：

- 假设 1：区域异质性较强时，按区域分层抽样的 MSE 低于 SRS。
- 假设 2：社区内同质性较高时，两阶段整群抽样的设计效应显著大于 1。
- 假设 3：高收入家庭非应答较严重时，加权可降低偏差，但可能提高方差。

要求：

- 假设应与抽样理论有关
- 假设应能被后续模拟结果检验
- 最终结论要明确回答“假设是否被支持”

第三步：设计抽样方案

你必须至少构建并比较 3 种抽样设计：

- ① 简单随机抽样 (SRS)
- ② 分层抽样
- ③ 两阶段整群抽样或 PPS 抽样

也可扩展：

- 分层 + 两阶段抽样
- PPS 两阶段设计
- 对稀有群体过度抽样

每种设计必须说明：

- 抽样单位与阶段
- 样本量与分配方式
- 入样概率与权重
- 理论优点
- 潜在风险

方案设计时要思考什么

设计不是套模板，请认真思考：

- 哪个变量适合做分层变量？
- 哪个变量适合做规模变量？
- 你的目标是总体估计，还是子总体估计？
- 你的设计是否有利于远郊或老年家庭估计？
- 权重是否会变得很不稳定？

没有统一标准答案，但必须逻辑清楚。

第四步：做重复抽样模拟

本任务的关键特征之一：**不是抽一次，而是重复抽样。**

基本要求：

- 每个设计至少重复抽样 50 次
- 推荐重复 100 次
- 每次都估计目标参数

至少估计以下参数：

- ① 全市家庭月消费均值
- ② 各区域家庭月消费均值
- ③ 高收入家庭占比
- ④ 在线消费支出占比

核心评价指标

对每种设计、每个目标参数，你至少要比较：

- 平均估计值
- 偏差 (Bias)
- 经验标准差
- 均方误差 (MSE)

你需要回答：

- 哪种设计平均上最接近真值？
- 哪种设计波动最小？
- 哪种设计更稳定？
- 是否存在“对均值好、对子总体不一定好”的情况？

结果展示要求

请至少提供 1 张结果图。

推荐图形：

- 箱线图：比较不同设计估计值分布
- 柱状图：比较不同设计的 MSE
- 误差条图：展示平均估计与波动
- 分区域结果图：比较子总体估计表现

注意：图表不是装饰，而是服务于结论。

第五步：做情景分析

单一总体上的结论可能并不稳健，因此你至少要完成 1 个情景分析。

可选情景包括：

- 区域异质性强 vs 弱
- 社区聚集性强 vs 弱
- 社区规模差异大 vs 小

你需要回答：

- 哪种设计表现最稳定？
- 哪些结论依赖于总体结构？
- 是否存在“没有统一最优设计”的证据？

为什么要做情景分析

研究型任务要强调结论的条件性。

你需要判断：

- 你的结论只是对当前这个总体成立吗？
- 还是对一类总体结构都成立？
- 总体结构变化后，设计优劣会不会反转？

这一步是“研究味道”的关键。

第六步：做偏差分析

在实际调查中，误差不只有抽样误差。

请至少完成 1 个非抽样误差分析，例如：

- 高收入家庭更容易非应答
- 远郊家庭更难联系
- 老年家庭对在线消费存在系统低报

你需要回答：

- 偏差机制如何影响估计结果？
- 哪类指标受影响最大？
- 哪种设计更稳健？
- 权重调整是否有效？代价是什么？

提醒：抽样误差不是全部误差

真实调查中的误差来源还包括：

- 覆盖误差
- 非应答误差
- 测量误差
- 数据处理误差

一个抽样设计“统计上好”，并不意味着它在现实调查中一定最优。

第七步：加入成本约束

本任务不仅比较统计精度，还要考虑调查成本。

默认成本函数如下：

- 进入一个新社区：100 元
- 完成一户调查：20 元
- 远郊户额外访问成本：15 元/户
- 跟访成本：10 元/户

你需要思考：

- 精度最好的设计是否一定最值得推荐？
- 在固定预算下，推荐方案是否改变？
- 若预算减少 30%，你的方案会如何调整？

最终要提交的不是“数字”，而是“建议”

你的研究结论必须回答以下问题：

- ① 哪种方案在什么条件下更优？
- ② 哪些假设得到了支持？
- ③ 哪些结果与你预期不一致？
- ④ 如果你是调查负责人，你会推荐哪种方案？
- ⑤ 成本和偏差会不会改变推荐结果？

研究结论必须由模拟结果支撑。

AI 可以用，但不能替代思考

本任务允许使用 AI，但必须**批判性使用**。

AI 可以帮助：

- 写代码
- 查公式
- 整理图表
- 生成分析框架

AI 不能替你决定：

- 抽样设计是否合理
- 权重是否正确
- 为什么出现某种结果
- 结论是否真的成立

你必须提交你的 **AI 交互记录**。

本次任务最容易出错的地方

请特别注意以下常见问题：

- 只抽一次样，没有做重复模拟
- 只报估计值，没有比较 Bias 和 MSE
- 只描述结果，没有解释机制
- 把加权当作万能纠偏方法
- 忽视子总体估计需求
- 忽视成本因素
- 研究问题与结果脱节
- 直接照搬 AI 给出的公式或结论

本次评分不以“代码长短”为核心，而重点看：

- 研究问题是否明确
- 假设是否合理
- 设计是否完整
- 模拟是否规范
- 结果解释是否有深度
- 是否能识别偏差与局限
- 是否能进行综合决策

核心原则：结论必须与证据一致。

提交内容

请提交以下材料：

- 研究摘要
- 研究问题与研究假设
- 抽样方案设计说明
- 核心结果表
- 至少 1 张图
- 情景分析结果
- 偏差分析结果
- 成本比较结果
- 研究结论
- AI 使用说明

可附：

- 代码文件
- 补充图表
- 额外分析说明