13.5 问尴尬的问题

工作纸

题目

STEP II 2023，第 11 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2023 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

www.ocr.org.uk/Images/707102-2023-past-papers.zip

提示

请先自行尝试解题。如果遇到困难，请逐一阅读提示，每读完一个便回到题目再试。标有停一停，试一试的提示是关键步骤。每个提示之后均留有空位供你演算。

想一想 Y 以概率 p 等于 X_1，以概率 q 等于 X_2。$\mathrm{E}(Y)$ 是甚么？$\mathrm{E}(Y^2)$ 又是甚么？

阶段 1 混合分布的方差

涵盖：第 (i) 部分

想一想 按选中的是哪一个变量来取条件（condition），求出 $\mathrm{E}(Y)$ 和 $\mathrm{E}(Y^2)$。写出 $\mathrm{E}(X_i^2) = \sigma_i^2 + \mu_i^2$。

停一停，试一试 减去 $(\mathrm{E}(Y))^2$，再把含 μ 的项合并，当中要用到 $p - p^2 = pq$ 和 $q - q^2 = pq$。

阶段 2 独立的随机化

涵盖：第 (ii) 部分

? **想一想** $\mathrm{P}(B = 1)$ 是多少？那 n 个答案是否独立（independent）？那么 Z_1 服从甚么分布（distribution）？

? **想一想** 把“与平均值（mean）相差不超过 10%”用标准单位（standard units）写出来。当 $n \to \infty$ 时会怎样？

阶段 3 **所有人共用一枚硬币**

涵盖：第 (iii) 部分

停一停，试一试 现在由一枚硬币决定全部 n 颗骰子的做法。已知掷出正面，Z_2 服从甚么分布？已知掷出反面呢？利用第 (i) 部分求平均值和方差（variance）。

想一想 Z_2 的概率大部分集中在哪里？在 $\frac{n}{2}$ 附近的概率多吗？估算每个峰与 $\frac{n}{2}$ 相距多少个标准差（standard deviation）。

背后的数学

随机化回答

在 Warner 的方案中，每位受访者私下作出选择，以概率 p 回答“你是否属于 A 组?”，以概率 $1-p$ 回答“你是否不属于 A 组?”。若属于 A 的人所占比例为 π，则答“是”的概率为

$$\lambda = p\pi + (1-p)(1-\pi),$$

因此由观察所得答“是”的比例 $\hat{\lambda}$，我们可以估计 $\hat{\pi} = \frac{\hat{\lambda}-(1-p)}{2p-1}$（前提是 $p \neq \frac{1}{2}$）。本题第 (ii) 部分正是这种情况，由每个人自己的骰子来决定。隐私的代价是精确度：除以 $2p-1$ 会令方差增大，所以随机化调查所需的受访者比直接调查更多。不过每个人都是独立地

随机化的，所以大数定律（law of large numbers）能够发挥作用。第 (iii) 部分显示了共用随机化时会出甚么问题：那一枚硬币所带来的不确定性，永远不会被平均抵销。

差分隐私

随机化回答是*差分隐私*（differential privacy）的前身。差分隐私是 Cynthia Dwork 及其同事于 2006 年提出的一个关于隐私的数学定义：如果加入或移除任何一个人的数据，只会令任何输出的概率改变至多一个很小的倍数，该程序便算是保障隐私的。科技公司用它来收集使用统计数据，美国人口普查局亦在 2020 年人口普查中采用了它，在公布的数字上加入经仔细校准的随机杂讯，令任何个人的答案都无法被还原。

全方差定律

第 (i) 部分的恒等式是*全方差定律*（law of total variance）的一个特例：$\mathrm{Var}(Y) = \mathrm{E}(\mathrm{Var}(Y \mid C)) + \mathrm{Var}(\mathrm{E}(Y \mid C))$，其中 C 是隐藏的选择。它是*方差分析*（analysis of variance，ANOVA）的基础。方差分析由费雪（Ronald Fisher）于 1920 年代在 Rothamsted 为农业实验而引入，它把农作物产量的变异分拆为两部分：可由肥料或品种解释的部分，以及未能解释的部分。

延伸练习

1 **双胞胎。**试做 STEP II 2005，第 12 题（step.maths.org/questions/05-s2-q12）。这其实是一道乔装了的随机化回答谜题：在甚么情况下，一个答案完全不能透露硬币的结果？

2 **估计 π。**在一项 $p = \frac{3}{4}$ 的 Warner 调查中，有 400 人作答，其中 160 人答“是”。试估计 A 组所占的比例 π，并求出你的估计值的近似标准误差（standard error）。

3 **全方差。**设 C 为取多个值的离散隐藏变量，证明全方差定律。

4 **共用硬币。**在第 (iii) 部分中，$Z_2 > \frac{n}{2}$ 的概率是多少？试解释为何当 n 很大时，这个概率接近 $\frac{1}{2}$，尽管 Z_2 几乎从不接近 $\frac{n}{2}$。

延伸阅读

- **Stanley Warner, “Randomized response: a survey technique for eliminating evasive answer bias”**，*Journal of the American Statistical Association* 60 (1965)。
- **Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy***（可于网上免费阅读）。现代理论；开首几章不难读懂。
- **David Spiegelhalter, *The Art of Statistics***（Pelican），谈调查及其陷阱。

完整的引导式解答见第 13.5 节：

https://beyondstep.keithmcnulty.org/zh-hans/ch13/awkward-questions.html

每星期一道 STEP 题目，完整讲解：The STEP Up，以电子邮件免费发送，订阅网址：keithmcnulty.substack.com