13.5 問尷尬的問題

工作紙

題目

STEP II 2023，第 11 題

繼續閱讀之前，請先看題目（原題為英文）。題目載於 OCR 提供下載的 2023 年 STEP 試卷（zip 檔案）內的 STEP II 試卷：

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

提示

請先自行嘗試解題。如果遇到困難，請逐一閱讀提示，每讀完一個便回到題目再試。標有停一停，試一試的提示是關鍵步驟。每個提示之後均留有空位供你演算。

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

階段 1 混合分佈的方差

涵蓋：第 (i) 部分

想一想 按選中的是哪一個變量來取條件（condition），求出 $\mathrm{E}(Y)$ 和 $\mathrm{E}(Y^2)$。寫出 $\mathrm{E}(X_i^2) = \sigma_i^2 + \mu_i^2$。

停一停，試一試 減去 $(\mathrm{E}(Y))^2$，再把含 μ 的項合併，當中要用到 $p - p^2 = pq$ 和 $q - q^2 = pq$。

階段 2　獨立的隨機化

涵蓋：第 (ii) 部分

? **想一想**　$\mathrm{P}(B = 1)$ 是多少？那 n 個答案是否獨立（independent）？那麼 Z_1 服從甚麼分佈（distribution）？

? **想一想**　把「與平均值（mean）相差不超過 10%」用標準單位（standard units）寫出來。當 $n \to \infty$ 時會怎樣？

階段 3 所有人共用一枚硬幣

涵蓋：第 (iii) 部分

停一停，試一試 現在由一枚硬幣決定全部 n 顆骰子的做法。已知擲出正面，Z_2 服從甚麼分佈？已知擲出反面呢？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裏？在 $\frac{n}{2}$ 附近的概率多嗎？估算每個峯與 $\frac{n}{2}$ 相距多少個標準差（standard deviation）。

背後的數學

隨機化回答

在 Warner 的方案中，每位受訪者私下作出選擇，以概率 p 回答「你是否屬於 A 組？」，以概率 $1-p$ 回答「你是否不屬於 A 組？」。若屬於 A 的人所佔比例為 π，則答「是」的概率為

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

因此由觀察所得答「是」的比例 $\hat{\lambda}$，我們可以估計 $\hat{\pi} = \frac{\hat{\lambda}-(1-p)}{2p-1}$（前提是 $p \neq \frac{1}{2}$）。本題第 (ii) 部分正是這種情況，由每個人自己的骰子來決定。私隱的代價是精確度：除以

$2p - 1$ 會令方差增大，所以隨機化調查所需的受訪者比直接調查更多。不過每個人都是獨立地隨機化的，所以大數定律（law of large numbers）能夠發揮作用。第 (iii) 部分顯示了共用隨機化時會出甚麼問題：那一枚硬幣所帶來的不確定性，永遠不會被平均抵銷。

差分私隱

隨機化回答是*差分私隱*（differential privacy）的前身。差分私隱是 Cynthia Dwork 及其同事於 2006 年提出的一個關於私隱的數學定義：如果加入或移除任何一個人的數據，只會令任何輸出的概率改變至多一個很小的倍數，該程序便算是保障私隱的。科技公司用它來收集使用統計數據，美國人口普查局亦在 2020 年人口普查中採用了它，在公佈的數字上加入經仔細校準的隨機雜訊，令任何個人的答案都無法被還原。

全方差定律

第 (i) 部分的恆等式是*全方差定律*（law of total variance）的一個特例：$\mathrm{Var}(Y) = \mathrm{E}(\mathrm{Var}(Y \mid C)) + \mathrm{Var}(\mathrm{E}(Y \mid C))$，其中 C 是隱藏的選擇。它是*方差分析*（analysis of variance，ANOVA）的基礎。方差分析由費雪（Ronald Fisher）於 1920 年代在 Rothamsted 為農業實驗而引入，它把農作物產量的變異分拆為兩部分：可由肥料或品種解釋的部分，以及未能解釋的部分。

延伸練習

1 **雙胞胎。**試做 STEP II 2005，第 12 題（step.maths.org/questions/05-s2-q12）。這其實是一道喬裝了的隨機化回答謎題：在甚麼情況下，一個答案完全不能透露硬幣的結果？

2 **估計 π。**在一項 $p = \frac{3}{4}$ 的 Warner 調查中，有 400 人作答，其中 160 人答「是」。試估計 A 組所佔的比例 π，並求出你的估計值的近似標準誤差（standard error）。

3 **全方差。**設 C 為取多個值的離散隱藏變量，證明全方差定律。

4 **共用硬幣。**在第 (iii) 部分中，$Z_2 > \frac{n}{2}$ 的概率是多少？試解釋為何當 n 很大時，這個概率接近 $\frac{1}{2}$，儘管 Z_2 幾乎從不接近 $\frac{n}{2}$。

延伸閱讀

► **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***（可於網上免費閱讀）。現代理論；開首幾章不難讀懂。
- **David Spiegelhalter, *The Art of Statistics***（Pelican），談調查及其陷阱。

完整的引導式解答見第 13.5 節：
https://beyondstep.keithmcnulty.org/zh-hant/ch13/awkward-questions.html

每星期一道 STEP 題目，完整講解：The STEP Up，以電郵免費發送，訂閱網址：keithmcnulty.substack.com