13.5 答えにくい質問をする

ワークシート

問題

STEP II 2023、問 11

先に進む前に、まず問題を読んでおこう。問題は、OCR が配布している 2023 年の STEP 過去問一式（zip ファイル）のうち、STEP II の問題用紙にある：

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

ヒント

まずは自力で問題に取り組んでみよう。行き詰まったら、ヒントを**一つずつ**読み、そのたびに問題に戻ること。**手を止めて、やってみよう**と記したヒントが重要な段階である。各ヒントの後には計算用の余白がある。

? **考えてみよう** Y は確率 p で X_1、確率 q で X_2 である。$\mathrm{E}(Y)$ は何か。$\mathrm{E}(Y^2)$ は何か。

段階 1 混合分布の分散（variance of a mixture）

扱う内容：小問 (i)

考えてみよう どちらの変数が選ばれるかで条件付け（conditioning）をして、$\mathrm{E}(Y)$ と $\mathrm{E}(Y^2)$ を求めよ。$\mathrm{E}(X_i^2) = \sigma_i^2 + \mu_i^2$ と書くとよい。

手を止めて、やってみよう $(\mathrm{E}(Y))^2$ を引き、$p - p^2 = pq$ と $q - q^2 = pq$ を用いて μ を含む項をまとめよ。

段階 2　独立な無作為化（randomisation）

扱う内容：小問 (ii)

考えてみよう　$\mathrm{P}(B = 1)$ はいくらか？n 個の回答は互いに独立（independent）か？それならば Z_1 の分布（distribution）は何か？

考えてみよう　「平均（mean）の 10% 以内」を、標準偏差を単位とする標準化した尺度（standard units）で書き表せ。$n \to \infty$ のとき何が起こるか？

段階 3 全員に共通の 1 枚の硬貨

扱う内容：小問 (iii)

手を止めて、やってみよう 今度は 1 枚の硬貨が n 個のさいころすべてについてまとめて決める。表が出たとき、Z_2 の分布は何か？裏が出たときはどうか？平均と分散は小問 (i) を用いて求めよ。

考えてみよう Z_2 の確率の大部分はどこに集まっているか？$\frac{n}{2}$ の近くには多くの確率があるか？それぞれの山が $\frac{n}{2}$ から標準偏差（standard deviation）何個分離れているかを見積もれ。

背後にある数学

無作為化回答法

Warner の方式では、各回答者は確率 p で「あなたはグループ A に属していますか？」に答え、確率 $1-p$ で「あなたはグループ A に属して**いない**のですか？」に答える。どちらに答えるかは本人だけが知っている。A に属する人の割合を π とすると、「はい」と答える確率は

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

であるから、観測された「はい」の割合 $\hat{\lambda}$ から、$\hat{\pi} = \frac{\hat{\lambda} - (1-p)}{2p-1}$ と推定できる（ただし $p \neq \frac{1}{2}$ とする）。問題の小問 (ii) はまさにこれであり、各人が自分のさいころで決めている。プライバシーの代償は精度である。$2p-1$ で割ることで分散が大きくなるため、無作為化した調査では直接たずねる調査よりも多くの回答者が必要になる。しかし各人が独立に無作為化するので、大数の法則（law of large numbers）がはたらく。小問 (iii) は、無作為化が共有されると何がまずいのかを示している。ただ 1 枚の硬貨についての不確かさは、決して平均化されて消えることがないのである。

差分プライバシー

無作為化回答法は、**差分プライバシー**（differential privacy）の祖先にあたる。差分プライバシーとは、2006 年に Cynthia Dwork らが導入したプライバシーの数学的な定義であり、ある手続きがプライベートであるとは、任意の 1 人のデータを加えたり取り除いたりしても、どの出力の確率もたかだか小さな倍率しか変わらないことをいう。テクノロジー企業は利用統計の収集にこれを用いており、アメリカ合衆国国勢調査局は 2020 年の国勢調査でこれを用いた。公表する数値に注意深く調整したランダムなノイズを加え、どの個人の回答も復元できないようにしたのである。

全分散の法則

小問 (i) の恒等式は、**全分散の法則**（law of total variance）$\mathrm{Var}(Y) = \mathrm{E}(\mathrm{Var}(Y \mid C)) + \mathrm{Var}(\mathrm{E}(Y \mid C))$ の特別な場合である。ここで C は隠れた選択である。この法則は**分散分析**（analysis of variance、ANOVA）の基礎になっている。分散分析は、1920 年代にフィッシャー（Ronald Fisher）がロザムステッド（Rothamsted）での農業実験のために導入したもので、作物の収量のばらつきを、肥料や品種によって説明される部分と、説明されずに残る部分とに分解する。

次の一歩

1. **双子。** STEP II 2005、問 12（step.maths.org/questions/05-s2-q12）に挑戦せよ。これは姿を変えた無作為化回答法のパズルである。回答が硬貨について何も教えてくれないのはどのようなときか。

2. **π の推定。** $p = \frac{3}{4}$ の Warner 方式の調査で、400 人が回答し、160 人が「はい」と答えた。グループ A に属する人の割合 π を推定し、その推定値のおおよその標準誤差（standard error）を求めよ。

3. **全分散。** いくつかの値をとる離散的な隠れた変数 C について、全分散の法則を証明せよ。

4 **共有された硬貨。** 小問 (iii) において、$Z_2 > \frac{n}{2}$ となる確率は何か。Z_2 が $\frac{n}{2}$ に近い値をとることはほとんどないにもかかわらず、n が大きいときこの確率が $\frac{1}{2}$ に近い理由を説明せよ。

さらに学ぶために

- **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***（オンラインで無料公開）。現代の理論であり、最初の数章は読みやすい。
- **David Spiegelhalter, *The Art of Statistics***（Pelican）。調査とその落とし穴について。

詳しい解説付きの解答は第 13.5 節にある：

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

毎週 1 問の STEP の問題を丁寧に解き明かす The STEP Up。メールで無料配信中：keithmcnulty.substack.com