解答例

STEP II 2023、第 11 問

確率・統計 ・ 混合分布

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

この問題は、*Beyond STEP* の第 13.5 節「答えにくい質問をする」で詳しく解説されている：https://beyondstep.keithmcnulty.org/ja/ch13/awkward-questions.html

鍵となる考え方

Y は**混合分布**（mixture）に従う確率変数である。すなわち、まず分布を無作為に一つ選び、次にその分布から値を取り出す。Y の平均は二つの平均の加重平均になるが、分散には余分な項が加わる。二つの平均どうしの隔たりそのものが、ばらつきを生むからである。小問 (ii) と (iii) はどちらもさいころを使う前に硬貨を投げるが、その振る舞いはまったく異なる。(ii) では一回振るたびに新たに硬貨を投げるので、混合の効果は平均化されて消える。(iii) では一回の硬貨投げがすべてを決めるので、その余分な項が支配的になる。

解答

小問 (i)

どちらの確率変数を使うかの選択は X_1 と X_2 の値とは独立に行われるので、

$$\mathrm{P}(Y = x_i) = p\,a_i + q\,b_i \qquad (i = 1, \ldots, n).$$

である。したがって

$$\mathrm{E}(Y) = \sum_i x_i(pa_i + qb_i) = p\sum_i x_i a_i + q\sum_i x_i b_i = p\mu_1 + q\mu_2.$$

同様に、$\mathrm{E}(Y^2) = p\,\mathrm{E}(X_1^2) + q\,\mathrm{E}(X_2^2) = p(\sigma_1^2+\mu_1^2) + q(\sigma_2^2+\mu_2^2)$ である。よって

$$\begin{aligned}\mathrm{Var}(Y) &= p\sigma_1^2 + q\sigma_2^2 + p\mu_1^2 + q\mu_2^2 - (p\mu_1+q\mu_2)^2\\ &= p\sigma_1^2 + q\sigma_2^2 + p(1-p)\mu_1^2 + q(1-q)\mu_2^2 - 2pq\mu_1\mu_2\\ &= p\sigma_1^2 + q\sigma_2^2 + pq\mu_1^2 + pq\mu_2^2 - 2pq\mu_1\mu_2\\ &= p\sigma_1^2 + q\sigma_2^2 + pq(\mu_1-\mu_2)^2,\end{aligned}$$

となる。ここで $1-p=q$ と $1-q=p$ を用いた。これが求める結果である。

小問 (ii)

硬貨が表を出す事象を H、さいころが 6 の目を出す事象を S とする。硬貨とさいころは独立であるから、

$$\mathrm{P}(B=1) = \mathrm{P}(H)\mathrm{P}(S) + \mathrm{P}(H^c)\mathrm{P}(S^c) = \tfrac{1}{2}\cdot\tfrac{1}{6} + \tfrac{1}{2}\cdot\tfrac{5}{6} = \tfrac{1}{2}.$$

である。よって各 B はそれぞれ確率 $\frac{1}{2}$ で 0 または 1 をとる。Z_1 は、B の n 個の独立なコピーのうち 1 となるものの個数であり、各コピーは同じ確率 $\frac{1}{2}$ で成功となる。これはまさに二項分布（binomial distribution）の設定であるから、$Z_1 \sim \mathrm{B}(n, \frac{1}{2})$ であり、平均は $\frac{n}{2}$、分散は $\frac{n}{4}$ である。

n が大きいとき、Z_1 を $\mathrm{N}(\frac{n}{2}, \frac{n}{4})$ で近似する。その標準偏差（standard deviation）は $\frac{\sqrt{n}}{2}$ である。平均との差が平均の 10% 以内であるとは、$|Z_1 - \frac{n}{2}| \le \frac{n}{20}$ ということである。$Z \sim \mathrm{N}(0,1)$ として標準化（standardising）すると、

$$\mathrm{P}\big(|Z_1 - \tfrac{n}{2}| \le \tfrac{n}{20}\big) \approx \mathrm{P}\Big(|Z| \le \frac{n/20}{\sqrt{n}/2}\Big) = \mathrm{P}\Big(|Z| \le \frac{\sqrt{n}}{10}\Big) = 2\Phi\Big(\frac{\sqrt{n}}{10}\Big) - 1.$$

$n \to \infty$ のとき $\sqrt{n}/10 \to \infty$ であるから $\Phi(\sqrt{n}/10) \to 1$ となり、確率は $2-1=1$ に収束する。（連続修正（continuity correction）を施しても、境界は標準偏差の $\frac{1}{\sqrt{n}}$ 倍しか変わらず、極限には影響しない。）

小問 (iii)

n 個のさいころのうち、6 の目の個数を X_1、6 以外の目の個数を X_2 とする。すると $X_1 \sim \mathrm{B}(n, \frac{1}{6})$、$X_2 \sim \mathrm{B}(n, \frac{5}{6})$ である。Z_2 は確率 $p = \frac{1}{2}$ で X_1 の値を、確率 $q = \frac{1}{2}$ で X_2 の値をとり、硬貨はさいころと独立である。これはまさに小問 (i) の状況であり、

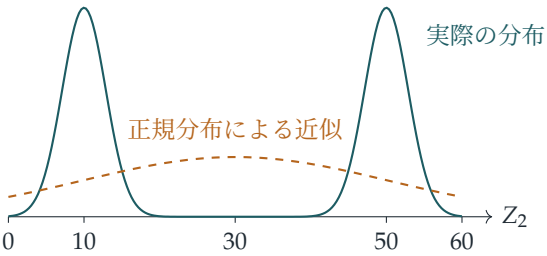
$$\mu_1 = \tfrac{n}{6}, \quad \mu_2 = \tfrac{5n}{6}, \quad \sigma_1^2 = \sigma_2^2 = n\cdot\tfrac{1}{6}\cdot\tfrac{5}{6} = \tfrac{5n}{36}.$$

である。よって

$$\begin{aligned}\mathrm{E}(Z_2) &= \tfrac{1}{2}\cdot\tfrac{n}{6} + \tfrac{1}{2}\cdot\tfrac{5n}{6} = \tfrac{n}{2},\\ \mathrm{Var}(Z_2) &= \tfrac{1}{2}\cdot\tfrac{5n}{36} + \tfrac{1}{2}\cdot\tfrac{5n}{36} + \tfrac{1}{4}\Big(\tfrac{4n}{6}\Big)^2 = \frac{5n}{36} + \frac{n^2}{9}.\end{aligned}$$

考えてみよう n が大きいとき、Z_2 の分布は実際にはどのような形をしているだろうか。

確率 $\frac{1}{2}$ で Z_2 は $\frac{n}{6}$ の近くにあり、確率 $\frac{1}{2}$ で $\frac{5n}{6}$ の近くにある。これら二つの山はそれぞれ標準偏差 $\frac{\sqrt{5n}}{6}$ をもち、これは山どうしの隔たり $\frac{2n}{3}$ に比べてごく小さい。したがってこの分布は**二峰性**（bimodal）であり、二つの細い山があって、平均 $\frac{n}{2}$ の近くにはほとんど確率がない。正規分布は平均のところに山を一つだけもつから、ここではまさに正反対の結果を与える。Z_2 がほとんど決してとらない値のところに最も多くの確率を置いてしまうのである。問題はもう一つある。標準偏差はおよそ $\frac{n}{3}$ であるから、$\mathrm{N}(\frac{n}{2}, \frac{n^2}{9})$ は負の値に約 $\mathrm{P}(Z < -1.5) \approx 0.067$ の確率を与え、n を超える値にも同じだけの確率を与える。しかし常に $0 \le Z_2 \le n$ である。下図は $n = 60$ の場合の二つの分布の様子を示している。



最後の部分に移る。Z_2 がその平均の 10% 以内にあるのは $0.45n \le Z_2 \le 0.55n$ のときである。硬貨の結果で条件付けると、

$$\mathrm{P}(0.45n \le Z_2 \le 0.55n) = \tfrac{1}{2}\,\mathrm{P}(0.45n \le X_1 \le 0.55n) + \tfrac{1}{2}\,\mathrm{P}(0.45n \le X_2 \le 0.55n).$$

第 1 項について、$\mathrm{P}(0.45n \le X_1 \le 0.55n) \le \mathrm{P}(X_1 \ge 0.45n)$ であり、$0.45n - \frac{n}{6} = \frac{17n}{60}$ である。標準偏差 $\frac{\sqrt{5n}}{6}$ をもつ正規近似 $X_1 \approx \mathrm{N}(\frac{n}{6}, \frac{5n}{36})$ を用いると、

$$\mathrm{P}(X_1 \ge 0.45n) \approx \mathrm{P}\Big(Z \ge \frac{17n/60}{\sqrt{5n}/6}\Big) = 1 - \Phi\Big(\frac{17\sqrt{n}}{10\sqrt{5}}\Big) \longrightarrow 1 - 1 = 0.$$

第 2 項は鏡像である。$X_2 = n - X_1$ であるから、$0.45n \le X_2 \le 0.55n$ となるのはちょうど $0.45n \le X_1 \le 0.55n$ となるときであり、第 2 項は第 1 項に等しい。したがって、Z_2 がその平均の 10% 以内にある確率は高々およそ $1 - \Phi(\frac{17\sqrt{n}}{10\sqrt{5}})$ であり、これは $n \to \infty$ のとき 0 に収束する。

この段階は、実は近似に依存していない。チェビシェフ（Chebyshev）の不等式により、厳密に

$$\mathrm{P}\big(|X_1 - \tfrac{n}{6}| \ge \tfrac{17n}{60}\big) \le \frac{5n/36}{(17n/60)^2} = \frac{500}{289\,n} \longrightarrow 0.$$

が成り立つ。

検算

- (i) において、$\mu_1 = \mu_2$ ならば余分な項は消える。また $\sigma_1 = \sigma_2 = 0$ ならば Y は値 μ_1 または μ_2 のみをとり、その分散は $pq(\mu_1 - \mu_2)^2$ となる。これはスケールを変えたベルヌーイ（Bernoulli）確率変数の場合と同じである。
- (ii) において、$n = 400$ のときの二項分布による厳密な確率は 0.9598 であり、正規近似による値 $2\Phi(2) - 1 = 0.9545$ と比べられる。$n = 1000$ のときは、二つの値はそれぞれ 0.9986 と 0.9984 である。
- (iii) において、$n = 30$ として Z_2 をコンピュータでシミュレーションしたところ、平均 15.00、分散 104.2 が得られ、公式による値 15 と 104.17 に一致した。$n = 100$ のとき、平均の 10% 以内にある厳密な確率はおよそ 3×10^{-11} である。

よくある間違い

- 二つの平均のどちらを選ぶかということ自体がばらつきの源であることを忘れて、$\mathrm{Var}(Y) = p\sigma_1^2 + q\sigma_2^2$ と答えてしまう。
- (iii) の Z_2 を (ii) の Z_1 と同じように扱ってしまう。(ii) ではさいころ一つごとに硬貨を投げ直すが、(iii) では一回の硬貨投げが n 個のさいころすべてを支配するので、結果は独立ではない。
- 当てはまりの悪さを「分散が大きいから」とだけ説明してしまう。本当の問題は分布の形にある。二つの離れた山があり、平均の近くにはほとんど確率がないのである。

STEP のその先へ

小問 (i) は**全分散の公式**（law of total variance）$\mathrm{Var}(Y) = \mathrm{E}\big(\mathrm{Var}(Y \mid C)\big) + \mathrm{Var}\big(\mathrm{E}(Y \mid C)\big)$ の最も簡単な場合である。ここで C は硬貨を表す。小問 (ii) は大数の弱法則（weak law of large numbers）であり、大学の最初の講義ではチェビシェフの不等式を用いて厳密に証明される。小問 (iii) は、この法則になぜ独立性が必要なのかを示している。共有される硬貨が一枚あるだけで、法則は成り立たなくなるのである。Z_2 のような混合分布は統計モデリングの定番である。Blitzstein と Hwang による *Introduction to Probability*（確率論入門）は、これら三つの考え方すべてを多くの例とともに扱っている。

Beyond STEP 解答集：https://beyondstep.keithmcnulty.org/ja/。解答本文は CC0 1.0 である。問題そのものは © Cambridge University Press & Assessment に帰属し、ここには掲載していない。英語版から機械翻訳したものであり、まだ校閲を経ていない。