詳細解答

# STEP II 2023，第 11 題

概率與統計 • 混合分佈

**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

本題在 *Beyond STEP* 第 13.5 節 *問心尷尬的問題* 中有完整的引導式講解：https://beyondstep.keithmcnulty.org/zh-hant/ch13/awkward-questions.html

## 關鍵思路

$Y$ 是一個*混合分佈*（mixture）：先隨機選定一個分佈，再從該分佈中抽取數值。它的平均值（mean）是兩個平均值的加權平均，但它的方差（variance）多了一項，因為兩個平均值之間的差距本身也會帶來變異。第 (ii) 和第 (iii) 部分都是先擲硬幣，再擲骰子，但兩者的表現截然不同。在 (ii) 中，每擲一次骰子都重新擲一次硬幣，所以混合的效果會被平均掉。在 (iii) 中，一枚硬幣決定一切，而那額外的一項最終會佔主導地位。

## 解答

### 第 (i) 部分

選用哪個隨機變量，與 $X_1$ 和 $X_2$ 的取值互相獨立（independent），所以

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

因此

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

同理，$\mathrm{E}(Y^2) = p\,\mathrm{E}(X_1^2) + q\,\mathrm{E}(X_2^2) = p(\sigma_1^2+\mu_1^2) + q(\sigma_2^2+\mu_2^2)$。所以

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

當中用到 $1-p=q$ 及 $1-q=p$。這就是所求的結果。

**第 (ii) 部分**

設 $H$ 為硬幣擲出正面的事件，$S$ 為骰子擲出六點的事件。硬幣與骰子互相獨立，所以

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

所以每個 $B$ 取 0 或 1 的概率各為 $\frac{1}{2}$。$Z_1$ 是 $n$ 個獨立的 $B$ 之中 1 的個數，而每個 $B$ 成功的概率同樣是 $\frac{1}{2}$。這正是二項分佈（binomial distribution）的設定，所以 $Z_1 \sim \mathrm{B}(n, \frac{1}{2})$，平均值為 $\frac{n}{2}$，方差為 $\frac{n}{4}$。

當 $n$ 很大時，我們以 $\mathrm{N}(\frac{n}{2}, \frac{n}{4})$ 近似 $Z_1$，其標準差（standard deviation）為 $\frac{\sqrt{n}}{2}$。與平均值相差不超過 10%，即 $|Z_1 - \frac{n}{2}| \le \frac{n}{20}$。設 $Z \sim \mathrm{N}(0,1)$，進行標準化（standardising），得

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

當 $n \to \infty$ 時，$\sqrt{n}/10 \to \infty$，所以 $\Phi(\sqrt{n}/10) \to 1$，而該概率趨向 $2-1=1$。（連續性校正（continuity correction）只會令界限改變 $\frac{1}{\sqrt{n}}$ 個標準差，並不影響極限。）

**第 (iii) 部分**

設 $X_1$ 為 $n$ 粒骰子中擲出六點的數目，$X_2$ 為不是六點的數目。則 $X_1 \sim \mathrm{B}(n, \frac{1}{6})$ 及 $X_2 \sim \mathrm{B}(n, \frac{5}{6})$。現在 $Z_2$ 以概率 $p = \frac{1}{2}$ 取 $X_1$ 的值，以概率 $q = \frac{1}{2}$ 取 $X_2$ 的值，而且硬幣與骰子互相獨立。這正是第 (i) 部分的情況，其中

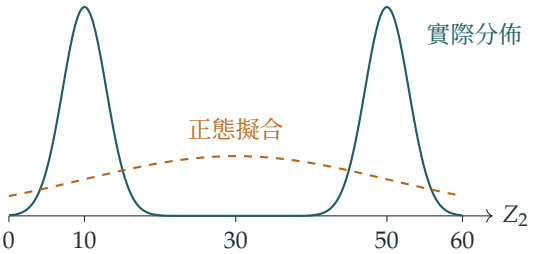
$$\mu_1 = \tfrac{n}{6}, \quad \mu_2 = \tfrac{5n}{6}, \quad \sigma_1^2 = \sigma_2^2 = n\cdot\tfrac{1}{6}\cdot\tfrac{5}{6} = \tfrac{5n}{36}.$$

所以

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

**想一想** 當 $n$ 很大時，$Z_2$ 的分佈實際上是甚麼樣子？

$Z_2$ 有 $\frac{1}{2}$ 的概率接近 $\frac{n}{6}$，另有 $\frac{1}{2}$ 的概率接近 $\frac{5n}{6}$。這兩個峯各自的標準差為 $\frac{\sqrt{5n}}{6}$，與兩者之間的距離 $\frac{2n}{3}$ 相比極小。所以這個分佈是雙峯（bimodal）的：兩個狹窄的峯，而在平均值 $\frac{n}{2}$ 附近幾乎沒有概率。正態分佈（normal distribution）只在平均值處有單一個峯，所以在這裏恰恰錯了：它把最多的概率放在 $Z_2$ 幾乎從不落入的地方。另外還有第二個問題。標準差約為 $\frac{n}{3}$，所以 $\mathrm{N}(\frac{n}{2}, \frac{n^2}{9})$ 給負值的概率約為 $\mathrm{P}(Z < -1.5) \approx 0.067$，給大於 $n$ 的值的概率也一樣，但 $0 \le Z_2 \le n$ 總是成立。下圖展示 $n = 60$ 時的兩幅圖像。




現在處理最後一部分。當 $0.45n \le Z_2 \le 0.55n$ 時，$Z_2$ 與其平均值相差不超過 10%。以硬幣的結果為條件，

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

對於第一項，$\mathrm{P}(0.45n \le X_1 \le 0.55n) \le \mathrm{P}(X_1 \ge 0.45n)$，而 $0.45n - \frac{n}{6} = \frac{17n}{60}$。利用正態近似（normal approximation）$X_1 \approx \mathrm{N}(\frac{n}{6}, \frac{5n}{36})$，其標準差為 $\frac{\sqrt{5n}}{6}$，得

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

第二項是第一項的鏡像：$X_2 = n - X_1$，所以 $0.45n \le X_2 \le 0.55n$ 當且僅當 $0.45n \le X_1 \le 0.55n$，因此第二項等於第一項。故此，$Z_2$ 與其平均值相差不超過 10% 的概率至多約為 $1 - \Phi\big(\frac{17\sqrt{n}}{10\sqrt{5}}\big)$，當 $n \to \infty$ 時趨向 0。

這一步其實並不依賴近似。切比雪夫（Chebyshev）不等式可準確地給出

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

## 驗算

- 在 (i) 中，若 $\mu_1 = \mu_2$，額外的一項便會消失；若 $\sigma_1 = \sigma_2 = 0$，則 $Y$ 只取 $\mu_1$ 或 $\mu_2$ 兩個值，方差為 $pq(\mu_1 - \mu_2)^2$，與經縮放的伯努利（Bernoulli）隨機變量相同。

- 在 (ii) 中，$n = 400$ 時準確的二項概率為 0.9598，而正態近似值為 $2\Phi(2) - 1 = 0.9545$。$n = 1000$ 時，兩者分別為 0.9986 和 0.9984。
- 在 (iii) 中，以電腦模擬 $n = 30$ 的 $Z_2$，得出平均值 15.00 及方差 104.2，而公式給出的是 15 和 104.17。當 $n = 100$ 時，與平均值相差不超過 10% 的準確概率約為 $3 \times 10^{-11}$。

## 常見錯誤

- 寫出 $\mathrm{Var}(Y) = p\sigma_1^2 + q\sigma_2^2$，忘記了在兩個平均值之間作選擇本身就是變異的來源。
- 把 (iii) 中的 $Z_2$ 當作 (ii) 中的 $Z_1$ 處理。在 (ii) 中，每粒骰子都重新擲一次硬幣；但在 (iii) 中，一次擲硬幣便決定了全部 $n$ 粒骰子，所以各結果並不獨立。
- 只用「方差很大」來解釋擬合不佳。真正的問題在於形狀：兩個分開的峯，而平均值附近的概率極少。

## STEP 之外

第 (i) 部分是*全方差定律*（law of total variance）$\mathrm{Var}(Y) = \mathrm{E}\big(\mathrm{Var}(Y \mid C)\big) + \mathrm{Var}\big(\mathrm{E}(Y \mid C)\big)$ 最簡單的情況，其中 $C$ 代表硬幣。第 (ii) 部分是弱大數定律（weak law of large numbers），大學的入門課程會利用切比雪夫不等式嚴格證明它。第 (iii) 部分說明了為何這條定律需要獨立性：只要一枚共用的硬幣，便足以令它失效。像 $Z_2$ 這樣的混合分佈是統計建模的常用工具。Blitzstein 與 Hwang 所著的 *Introduction to Probability*（《概率導論》）以大量例子涵蓋了以上三個概念。

Beyond STEP 詳細解答：https://beyondstep.keithmcnulty.org/zh-hant/。解答文字採用 CC0 1.0；題目本身版權屬 Cambridge University Press & Assessment 所有（©），本文件不予轉載。由機器從英文翻譯而成；尚未經審閱。