详细解答

STEP II 2023，第 11 题

概率与统计 · 混合分布

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

本题在 *Beyond STEP* 第 13.5 节问尴尬的问题中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch13/awkward-questions.html

关键思路

Y 是一个混合分布（mixture）：先随机选定一个分布，再从该分布中抽取数值。它的平均值（mean）是两个平均值的加权平均，但它的方差（variance）多了一项，因为两个平均值之间的差距本身也会带来变异。第 (ii) 和第 (iii) 部分都是先掷硬币，再掷骰子，但两者的表现截然不同。在 (ii) 中，每掷一次骰子都重新掷一次硬币，所以混合的效果会被平均掉。在 (iii) 中，一枚硬币决定一切，而那额外的一项最终会占主导地位。

解答

第 (i) 部分

选用哪个随机变量，与 X_1 和 X_2 的取值互相独立（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

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

当中用到 $1-p=q$ 及 $1-q=p$。这就是所求的结果。

第 (ii) 部分

设 H 为硬币掷出正面的事件，S 为骰子掷出六点的事件。硬币与骰子互相独立，所以

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

所以每个 B 取 0 或 1 的概率各为 $\frac{1}{2}$。Z_1 是 n 个独立的 B 之中 1 的个数，而每个 B 成功的概率同样是 $\frac{1}{2}$。这正是二项分布（binomial distribution）的设置，所以 $Z_1 \sim \mathrm{B}(n, \frac{1}{2})$，平均值为 $\frac{n}{2}$，方差为 $\frac{n}{4}$。

当 n 很大时，我们以 $\mathrm{N}(\frac{n}{2}, \frac{n}{4})$ 近似 Z_1，其标准差（standard deviation）为 $\frac{\sqrt{n}}{2}$。与平均值相差不超过 10%，即 $|Z_1 - \frac{n}{2}| \le \frac{n}{20}$。设 $Z \sim \mathrm{N}(0,1)$，进行标准化（standardising），得

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

当 $n \to \infty$ 时，$\sqrt{n}/10 \to \infty$，所以 $\Phi(\sqrt{n}/10) \to 1$，而该概率趋向 $2-1=1$。（连续性校正（continuity correction）只会令界限改变 $\frac{1}{\sqrt{n}}$ 个标准差，并不影响极限。）

第 (iii) 部分

设 X_1 为 n 粒骰子中掷出六点的数目，X_2 为不是六点的数目。则 $X_1 \sim \mathrm{B}(n, \frac{1}{6})$ 及 $X_2 \sim \mathrm{B}(n, \frac{5}{6})$。现在 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，而且硬币与骰子互相独立。这正是第 (i) 部分的情况，其中

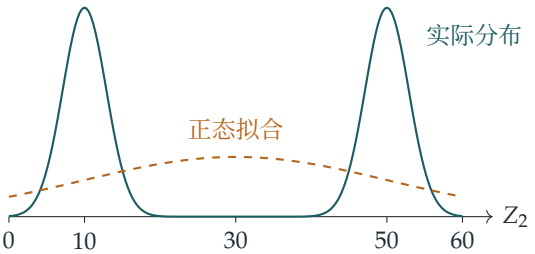
$$\mu_1 = \tfrac{n}{6}, \quad \mu_2 = \tfrac{5n}{6}, \quad \sigma_1^2 = \sigma_2^2 = n\cdot\tfrac{1}{6}\cdot\tfrac{5}{6} = \tfrac{5n}{36}.$$

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 当 n 很大时，Z_2 的分布实际上是甚么样子?

Z_2 有 $\frac{1}{2}$ 的概率接近 $\frac{n}{6}$，另有 $\frac{1}{2}$ 的概率接近 $\frac{5n}{6}$。这两个峰各自的标准差为 $\frac{\sqrt{5n}}{6}$，与两者之间的距离 $\frac{2n}{3}$ 相比极小。所以这个分布是双峰（bimodal）的：两个狭窄的峰，而在平均值 $\frac{n}{2}$ 附近几乎没有概率。正态分布（normal distribution）只在平均值处有单一个峰，所以在这里恰恰错了：它把最多的概率放在 Z_2 几乎从不落入的地方。另外还有第二个问题。标准差约为 $\frac{n}{3}$，所以 $\mathrm{N}(\frac{n}{2}, \frac{n^2}{9})$ 给负值的概率约为 $\mathrm{P}(Z < -1.5) \approx 0.067$，给大于 n 的值的概率也一样，但 $0 \le Z_2 \le n$ 总是成立。下图展示 $n = 60$ 时的两幅图像。



现在处理最后一部分。当 $0.45n \le Z_2 \le 0.55n$ 时，Z_2 与其平均值相差不超过 10%。以硬币的结果为条件，

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

对于第一项，$\mathrm{P}(0.45n \le X_1 \le 0.55n) \le \mathrm{P}(X_1 \ge 0.45n)$，而 $0.45n - \frac{n}{6} = \frac{17n}{60}$。利用正态近似（normal approximation）$X_1 \approx \mathrm{N}(\frac{n}{6}, \frac{5n}{36})$，其标准差为 $\frac{\sqrt{5n}}{6}$，得

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

第二项是第一项的镜像：$X_2 = n - X_1$，所以 $0.45n \le X_2 \le 0.55n$ 当且仅当 $0.45n \le X_1 \le 0.55n$，因此第二项等于第一项。故此，Z_2 与其平均值相差不超过 10% 的概率至多约为 $1 - \Phi\big(\frac{17\sqrt{n}}{10\sqrt{5}}\big)$，当 $n \to \infty$ 时趋向 0。

这一步其实并不依赖近似。切比雪夫（Chebyshev）不等式可准确地给出

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

验算

- 在 (i) 中，若 $\mu_1 = \mu_2$，额外的一项便会消失；若 $\sigma_1 = \sigma_2 = 0$，则 Y 只取 μ_1 或 μ_2 两个值，方差为 $pq(\mu_1 - \mu_2)^2$，与经缩放的伯努利（Bernoulli）随机变量相同。

- 在 (ii) 中，$n = 400$ 时准确的二项概率为 0.9598，而正态近似值为 $2\Phi(2) - 1 = 0.9545$。$n = 1000$ 时，两者分别为 0.9986 和 0.9984。
- 在 (iii) 中，以电脑模拟 $n = 30$ 的 Z_2，得出平均值 15.00 及方差 104.2，而公式给出的是 15 和 104.17。当 $n = 100$ 时，与平均值相差不超过 10% 的准确概率约为 3×10^{-11}。

常见错误

- 写出 $\mathrm{Var}(Y) = p\sigma_1^2 + q\sigma_2^2$，忘记了在两个平均值之间作选择本身就是变异的来源。
- 把 (iii) 中的 Z_2 当作 (ii) 中的 Z_1 处理。在 (ii) 中，每粒骰子都重新掷一次硬币；但在 (iii) 中，一次掷硬币便决定了全部 n 粒骰子，所以各结果并不独立。
- 只用“方差很大”来解释拟合不佳。真正的问题在于形状：两个分开的峰，而平均值附近的概率极少。

STEP 之外

第 (i) 部分是*全方差定律*（law of total variance）$\mathrm{Var}(Y) = \mathrm{E}\big(\mathrm{Var}(Y \mid C)\big) + \mathrm{Var}\big(\mathrm{E}(Y \mid C)\big)$ 最简单的情况，其中 C 代表硬币。第 (ii) 部分是弱大数定律（weak law of large numbers），大学的入门课程会利用切比雪夫不等式严格证明它。第 (iii) 部分说明了为何这条定律需要独立性：只要一枚共用的硬币，便足以令它失效。像 Z_2 这样的混合分布是统计建模的常用工具。Blitzstein 与 Hwang 所著的 *Introduction to Probability*（《概率导论》）以大量例子涵盖了以上三个概念。

Beyond STEP 详细解答：https://beyondstep.keithmcnulty.org/zh-hans/。解答文字采用 CC0 1.0；题目本身版权属 Cambridge University Press & Assessment 所有（©），本文件不予转载。由机器从英文翻译而成；尚未经审阅。