详细解答

STEP I 1991，第 15 题

概率与统计 · 正态近似

> **STEP I 1991，第 15 题**
>
> 继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：
>
> step.maths.org/questions/91-s1-q15

关键思路

每个 X_j 的平均值(mean)为 0、方差(variance)为 1，而且各次掷币互相独立(independent)，所以 S_n 的平均值为 0、方差为 n；当 n 很大时，中心极限定理（central limit theorem）容许我们把 S_n 视为近似正态分布（normal distribution）。最后一部分看似相仿，却藏着一个陷阱：每个 Y_k 都包含同一次首掷 X_1，所以各个 Y_k 并不独立，而这一次共用的掷币最终支配了整个总分。

解答

每个 X_j 取值 ± 1，概率各为 $\frac{1}{2}$，所以

$$\mathrm{E}(X_j) = 0, \qquad \mathrm{E}(X_j^2) = 1, \qquad \mathrm{Var}(X_j) = 1.$$

各次掷币互相独立，所以 $\mathrm{E}(S_n) = 0$ 及 $\mathrm{Var}(S_n) = n$。根据中心极限定理，当 n 很大时，

$$S_n \approx \mathrm{N}(0,\, n).$$

（等价地，若 H 为掷得正面的次数，则 $H \sim \mathrm{B}(n, \frac{1}{2})$，且 $S_n = H - (n - H) = 2H - n$。）

第 (i) 部分

当 $n = 10\,000$ 时，S_n 的标准差（standard deviation）为 $\sqrt{10\,000} = 100$，所以

$$\mathrm{P}(S_n > 100) \approx \mathrm{P}\left(Z > \frac{100}{100}\right) = 1 - \Phi(1) \approx 0.16.$$

如果想再谨慎一点，可对 H 使用连续性校正（continuity correction）：$S_n > 100$ 即 $2H - 10\,000 > 100$，亦即 $H \geq 5051$。由于 H 的平均值为 5000、标准差为 50，

$$\mathrm{P}(H \geq 5051) \approx \mathrm{P}\left(Z > \frac{5050.5 - 5000}{50}\right) = 1 - \Phi(1.01) \approx 0.156.$$

无论用哪一种方法，答案都约为 0.16。

第 (ii) 部分

标准化（standardising）后，

$$\mathrm{P}(S_n > 0.01n) \approx \mathrm{P}\left(Z > \frac{0.01n}{\sqrt{n}}\right) = 1 - \Phi\left(0.01\sqrt{n}\right).$$

这个值随 n 增大而减小，而且它小于 0.01 当且仅当 $\Phi(0.01\sqrt{n}) > 0.99$。查表得 $\Phi(2.326) = 0.99$，所以我们需要

$$0.01\sqrt{n} > 2.326 \quad \Longleftrightarrow \quad \sqrt{n} > 232.6 \quad \Longleftrightarrow \quad n > 54\,100 \text{（约数）}.$$

所以符合条件的最小 n 大约是 54 000。（若改用 2.33，则得出约 54 300；在这个近似程度下，任何接近 5.4×10^4 的答案都算正确。）

新的计分方法

对于 $2 \leq k \leq n$，首掷与第 k 掷共有四种情况：

首掷	第 k 掷	得分 Y_k	$X_1 + X_k$
H	H	2	$1 + 1 = 2$
T	T	-2	$-1 - 1 = -2$
H	T	0	$1 - 1 = 0$
T	H	0	$-1 + 1 = 0$

最后两栏在每种情况下都相同，所以 $Y_k = X_1 + X_k$，正如题目所求。

平均值与方差。 由线性性质（linearity），$\mathrm{E}(Y_k) = \mathrm{E}(X_1) + \mathrm{E}(X_k) = 0$。由于 $k \geq 2$，X_1 与 X_k 来自不同的两次掷币，因此互相独立，于是

$$\mathrm{Var}(Y_k) = \mathrm{Var}(X_1) + \mathrm{Var}(X_k) = 2.$$

（直接计算：Y_k 取值 2、-2、0 的概率分别为 $\frac{1}{4}, \frac{1}{4}, \frac{1}{2}$，所以 $\mathrm{E}(Y_k^2) = 2$。）

极限。 设 $m = n - 1$ 及 $T = Y_2 + Y_3 + \cdots + Y_n$。那么

$$T = \sum_{k=2}^{n} (X_1 + X_k) = mX_1 + S', \qquad \text{其中} S' = X_2 + X_3 + \cdots + X_n.$$

想一想 每个 Y_k 的平均值为 0、方差为 2，就像一个经过缩放的 X_j。为甚么我们不能直接照搬第 (ii) 部分的论证?

因为第 (ii) 部分的论证需要各项互相独立，而各个 Y_k 并不独立：它们全都共用 X_1。事实上，对于 $j \neq k$，$\mathrm{Cov}(Y_j, Y_k) = \mathrm{Var}(X_1) = 1$，而 $\mathrm{Var}(T) = m^2 + m$，并不是 $2m$。所以我们把 X_1 的两个取值分开考虑。留意 S' 是 m 个取值为 ± 1 的项之和，所以 $-m \leq S' \leq m$。

- 若首掷是反面，则 $X_1 = -1$ 且 $T = -m + S' \leq 0$。所以 $T > 0.01m$ 不可能发生。
- 若首掷是正面，则 $X_1 = 1$ 且 $T = m + S'$，所以 $T > 0.01m$ 当且仅当 $S' > -0.99m$。沿用第 (ii) 部分的正态近似，取 $S' \approx \mathrm{N}(0, m)$，

$$\mathrm{P}(S' \leq -0.99m) \approx \Phi\left(-0.99\sqrt{m}\right) \to 0 \quad \text{当} m \to \infty.$$

（其实无须使用近似：根据切比雪夫不等式（Chebyshev's inequality），这个概率至多为 $m/(0.99m)^2 \to 0$。）

由于 S' 与 X_1 独立，

$$\mathrm{P}(T > 0.01m) = \tfrac{1}{2} \cdot 0 + \tfrac{1}{2}\,\mathrm{P}(S' > -0.99m) \;\to\; \tfrac{1}{2} \quad \text{当} n \to \infty.$$

所以该命题是**错误**的：这个概率趋向 $\frac{1}{2}$，而不是 0。用文字来说，当 n 很大时，总分为正基本上就等同于首掷是正面，否则总分便不是正数。

验算

- 第 (i) 部分：对于 $H \sim \mathrm{B}(10\,000, \frac{1}{2})$，精确的二项概率 $\mathrm{P}(H \geq 5051)$ 为 0.15625，与经连续性校正的数值 0.156 吻合。
- 第 (ii) 部分：$\mathrm{P}(S_n > 0.01n)$ 的精确二项数值在 $n = 50\,000$ 时为 0.0125，在 $n = 54\,000$ 附近约为 0.0100，在 $n = 55\,000$ 时为 0.0094，所以临界值确实接近 54 000。
- 最后部分：对 $n \leq 14$ 列举全部 2^n 个结果，所得与 $\frac{1}{2}\,\mathrm{P}(S' > -0.99m)$ 完全一致；在 $n = 14$ 时概率已达 0.49994，而以 $n = 100$ 进行的一次模拟则得出 0.497。

常见错误

- 把 $Y_2, \ldots, Y_n$ 当作互相独立，从而得出 $\mathrm{Var}(T) = 2(n-1)$，令概率看似趋向 0。
- 计算得分的方差时，用了 $\mathrm{Var}(S_n) = n/4$（即正面次数的方差）而不是 n。

STEP 之外

第 (ii) 部分是弱大数定律（weak law of large numbers）的一个定量形式：只要 n 够大，S_n/n 就极不可能远离 0。最后一部分则显示，当各项之间有很强的相依关系时，这条定律会如何失效：这里每一项都共用首掷，所以平均值 $T/m = X_1 + S'/m$ 趋向 X_1 这个随机量，而不是趋向平均值 0。相依序列的大数定律（遍历定理（ergodic theorems）、可交换序列（exchangeable sequences））是大学本科概率论中内容丰富的课题；Grimmett 与 Welsh 的 *Probability: An Introduction*（概率论入门教材）涵盖了弱大数定律和切比雪夫不等式。

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。