解答例

STEP I 1991、第 15 問

確率・統計 ・ 正規近似

> **STEP I 1991、問 15**
>
> 先に進む前に、まず問題を読んでおこう。問題は STEP Support Programme の問題データベースにある：
>
> step.maths.org/questions/91-s1-q15

鍵となる考え方

各 X_j は平均（mean）0、分散（variance）1 をもち、各回の投げは独立（independent）であるから、S_n は平均 0、分散 n をもつ。n が大きいときは、中心極限定理（central limit theorem）により S_n を近似的に正規分布（normal distribution）に従うものとして扱える。最後の小問も見かけは似ているが、落とし穴がある。どの Y_k にも**同じ 1 回**目の投げ X_1 が含まれているので、Y_k たちは独立ではなく、その共有された 1 回の投げが結局は合計全体を支配してしまうのである。

解答

各 X_j はそれぞれ確率 $\frac{1}{2}$ で値 ± 1 をとるので、

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

各回の投げは独立であるから、$\mathrm{E}(S_n) = 0$ かつ $\mathrm{Var}(S_n) = n$ である。中心極限定理により、n が大きいとき

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

（同じことであるが、表の出た回数を H とすると $H \sim \mathrm{B}(n, \frac{1}{2})$ であり、$S_n = H - (n - H) = 2H - n$ である。）

小問 (i)

$n = 10\,000$ のとき、S_n の標準偏差（standard deviation）は $\sqrt{10\,000} = 100$ であるから、

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

もう少し丁寧に扱いたければ、H に連続修正（continuity correction）を施せばよい。$S_n > 100$ は $2H - 10\,000 > 100$、すなわち $H \geq 5051$ を意味する。H の平均は 5000、標準偏差は 50 であるから、

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

いずれにせよ、答えはおよそ 0.16 である。

小問 (ii)

標準化すると、

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

これは n が増加すると減少し、0.01 より小さくなるのはちょうど $\Phi(0.01\sqrt{n}) > 0.99$ のときである。数表より $\Phi(2.326) = 0.99$ であるから、

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

が必要である。よって、そのような最小の n はおよそ 54 000 である。（代わりに 2.33 を用いるとおよそ 54 300 となる。この程度の近似では、5.4×10^4 に近い答えであればどれも正しい。）

新しい得点の付け方

$2 \leq k \leq n$ に対して、1 回目の投げと k 回目の投げの組合せは 4 通りある。

1 回目の投げ	k 回目の投げ	得点 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	$-1 + 1 = 0$

最後の 2 列はすべての場合で一致するので、求めるとおり $Y_k = X_1 + X_k$ である。

平均と分散。 線形性（linearity）により、$\mathrm{E}(Y_k) = \mathrm{E}(X_1) + \mathrm{E}(X_k) = 0$ である。$k \geq 2$ であるから、X_1 と X_k は異なる回の投げであり、したがって独立である。よって

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

（直接計算してもよい。Y_k はそれぞれ確率 $\frac{1}{4}, \frac{1}{4}, \frac{1}{2}$ で 2、-2、0 をとるので、$\mathrm{E}(Y_k^2) = 2$ である。）

極限。 $m = n-1$、$T = Y_2 + Y_3 + \cdots + Y_n$ とおく。すると

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

考えてみよう 各 Y_k は、定数倍した X_j とちょうど同じように、平均 0、分散 2 をもつ。それなのに、なぜ小問 (ii) の議論をそのまま写すことができないのだろうか。

それは、小問 (ii) の議論では各項が独立であることが必要だったのに、Y_k たちは独立ではないからである。これらはすべて X_1 を共有している。実際、$j \neq k$ のとき $\mathrm{Cov}(Y_j, Y_k) = \mathrm{Var}(X_1) = 1$ であり、$\mathrm{Var}(T) = m^2 + m$ となって $2m$ にはならない。そこで、X_1 の 2 つの値を別々に調べる。S' はそれぞれ ± 1 である m 個の項の和なので、$-m \leq S' \leq m$ であることに注意する。

- 1 回目の投げが裏ならば、$X_1 = -1$ かつ $T = -m + S' \leq 0$ である。よって $T > 0.01m$ となることはありえない。
- 1 回目の投げが表ならば、$X_1 = 1$ かつ $T = m + S'$ であるから、$T > 0.01m$ となるのはちょうど $S' > -0.99m$ のときである。小問 (ii) と同じ正規近似により、$S' \approx \mathrm{N}(0, m)$ として

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

  （実は近似は必要ない。チェビシェフ（Chebyshev）の不等式（Chebyshev's inequality）により、この確率は高々 $m/(0.99m)^2 \to 0$ である。）

S' は X_1 と独立であるから、

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

よって、この主張は**偽**である。確率は 0 ではなく $\frac{1}{2}$ に収束する。言葉で言えば、n が大きいとき、合計が正になるのは本質的に 1 回目の投げが表のときであり、そうでなければ正にはならない。

検算

- 小問 (i)：$H \sim \mathrm{B}(10\,000, \frac{1}{2})$ に対する厳密な二項分布（binomial distribution）の確率 $\mathrm{P}(H \geq 5051)$ は 0.15625 であり、連続修正をした値 0.156 と一致する。
- 小問 (ii)：$\mathrm{P}(S_n > 0.01n)$ の厳密な二項分布の値は、$n = 50\,000$ で 0.0125、$n = 54\,000$ 付近でおよそ 0.0100、$n = 55\,000$ で 0.0094 である。したがって、境目は確かに 54 000 に近い。

- 最後の小問：$n \leq 14$ についてすべての 2^n 通りの結果を数え上げると、$\frac{1}{2}\mathrm{P}(S' > -0.99m)$ と厳密に一致する。$n = 14$ ですでに確率は 0.49994 であり、$n = 100$ でのシミュレーションでは 0.497 となった。

よくある間違い

- $Y_2, \ldots, Y_n$ を独立として扱い、$\mathrm{Var}(T) = 2(n-1)$ と結論してしまうこと。こうすると確率が 0 に収束するように見えてしまう。
- 得点の分散として、n の代わりに $\mathrm{Var}(S_n) = n/4$（表の出た回数の分散）を用いてしまうこと。

STEP のその先へ

小問 (ii) は大数の弱法則（weak law of large numbers）を定量的に表したものである。n が大きくなれば、S_n/n が 0 から大きく離れることはほとんど起こりえない。最後の小問は、各項が強く従属しているときにこの法則がどのように破れるかを示している。ここではすべての項が 1 回目の投げを共有しているので、平均 $T/m = X_1 + S'/m$ は平均値 0 ではなく、確率的な量である X_1 に収束する。従属な列に対する大数の法則（エルゴード定理（ergodic theorems）、交換可能な列（exchangeable sequences））は、大学の確率論における豊かな話題である。Grimmett and Welsh, *Probability: An Introduction*（確率論入門）では、大数の弱法則とチェビシェフの不等式が扱われている。

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。