解答例

STEP II 2002、第 12 問

確率・統計 ・ 正規近似とポアソン近似

> **STEP II 2002、問 12**
>
> 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/02-s2-q12

鍵となる考え方

この問題には、3 層のランダム性が入れ子になって積み重なっている。1 枚の硬貨の中での表の回数、1 日の中での硬貨、そして 1 週間の中での日である。各層は二項分布（binomial distribution）に従う個数であり、その「成功」確率は 1 つ下の層から与えられる。与えられた式は、最下層を連続性補正（continuity correction）つきの正規分布（normal distribution）で近似し、中間層をポアソン分布（Poisson distribution）で、さらに正規分布で近似し、最上層は厳密な二項分布のまま残すことで組み立てられている。3 つの層が見えてしまえば、やるべきことは、式の中の各表式がどの近似から生じたのかを特定し、そのうえで与えられた数値に対してそれぞれの近似が妥当かどうかを問うことである。

解答

1 日における 1 枚の硬貨

ある 1 日とある 1 枚の硬貨を固定する。その硬貨を M 回投げたときの表の回数を X とする。各回の投げは独立（independent）で、表の出る確率はいずれも p であるから

$$X \sim \mathrm{B}(M, p), \qquad \mathrm{E}(X) = Mp, \qquad \mathrm{Var}(X) = Mp(1-p).$$

M が大きいとき、X を正規確率変数 $X' \sim \mathrm{N}(Mp,\, Mp(1-p))$ で近似する。X は整数値をとるので、「表が m 回未満」とは $X \le m-1$ のことであり、連続性補正を施すとこれは $X' < m - \frac{1}{2}$ となる。よって

$$r = \mathrm{P}(X < m) = \mathrm{P}(X \le m-1) \approx \Phi\left(\frac{m - \frac{1}{2} - Mp}{\sqrt{Mp(1-p)}}\right) = \Phi\left(\frac{2m - 1 - 2Mp}{2\sqrt{Mp(1-p)}}\right).$$

問題文と比べると、これは $h \approx Lr$ であることを意味している。

1 日におけるすべての硬貨

その日に表が m 回未満となる硬貨の枚数（L 枚のうち）を Y とする。各硬貨は独立に投げられ、どれも同じように振る舞うので、各硬貨は独立に確率 r で「成功」し、

$$Y \sim \mathrm{B}(L,r), \qquad \mathrm{E}(Y) = Lr = h, \qquad \mathrm{Var}(Y) = Lr(1-r) = h(1-r).$$

考えてみよう q の式では分母に $\sqrt{h}$ がある。標準偏差（standard deviation）が $\sqrt{h(1-r)}$ ではなく $\sqrt{h}$ なのはなぜか。

この式は r が小さい場合を想定して作られている。すなわち、1 枚の硬貨にとって表が m 回未満となるのはまれな事象であり、一方で L は大きい。この状況では $1-r \approx 1$ なので分散（variance）はほぼ平均に等しく、これはまさにポアソン近似 $Y \approx \mathrm{Po}(h)$ が成り立つ状況である。さらに h もある程度大きければ、$\mathrm{Po}(h)$ はさらに $\mathrm{N}(h,h)$ で近似できる。（同じことだが、$\mathrm{B}(L,r)$ を直接正規分布で近似し、$1-r$ を 1 に置き換えてもよい。）

そこで $Y' \sim \mathrm{N}(h,h)$ をとる。「l 枚より多い」とは $Y \geq l+1$ のことであり、連続性補正を施すと $Y' > l + \frac{1}{2}$ となる。したがって

$$q = \mathrm{P}(Y > l) \approx 1 - \Phi\left(\frac{l+\frac{1}{2}-h}{\sqrt{h}}\right) = \Phi\left(\frac{h-l-\frac{1}{2}}{\sqrt{h}}\right) = \Phi\left(\frac{2h-2l-1}{2\sqrt{h}}\right),$$

ここで標準正規分布の対称性 $1-\Phi(z) = \Phi(-z)$ を用いた。これが与えられた q である。

すべての日

その日の硬貨のうち表が m 回未満となるものが l 枚より多いとき、その日を**よい日**と呼ぶ。どの日もよい日となる確率は同じ q であり、異なる日は異なる投げに関わるので、各日は独立である。したがって K 日のうちのよい日の数は $\mathrm{B}(K,q)$ に従い、

$$\mathrm{P}(\text{よい日がちょうど } k \text{ 日}) = \binom{K}{k} q^k (1-q)^{K-k},$$

ただし q は上のように近似する。これが求める結果である。

数値の場合

ここでは $Mp = 60$、$Mp(1-p) = 24$ であるから

$$\frac{2m-1-2Mp}{2\sqrt{Mp(1-p)}} = \frac{96-1-120}{2\sqrt{24}} = \frac{-25}{9.798} \approx -2.55,$$

となり、$r \approx \Phi(-2.55) \approx 0.0054$、$h = 500r \approx 2.68$ を得る。すると

$$\frac{2h - 2l - 1}{2\sqrt{h}} \approx \frac{5.36 - 9}{3.27} \approx -1.11, \qquad q \approx \Phi(-1.11) \approx 0.133,$$

であり、近似式は $\binom{7}{2} q^2 (1-q)^5 \approx 21 \times 0.0178 \times 0.490 \approx 0.18$ を与える。

考えてみよう どの近似も標準的なものである。そのうち、信頼できる条件の外で使っているものはどれか。

3 つの層を順に見ていく。

- **最上層**。$\mathrm{B}(7, q)$ の段階は（q を与えれば）厳密なので、それ自体は誤差を生まない。しかし $k = 2$ なので、q の誤差を 2 乗して増幅する。
- **最下層**。$M = 100$ で $Mp = 60$、$M(1-p) = 40$ であれば、二項分布の正規近似が通常用いられる範囲に十分入っている。しかしここではそれを裾の 2.55 標準偏差の位置で使い、約 0.005 という確率を求めている。正規近似は分布の主要部分では**絶対**誤差（absolute error）の意味で良い近似であるが、裾ではその**相対**誤差（relative error）が大きくなりうる。とくに $p \neq \frac{1}{2}$ のため二項分布は歪んでいる。r の相対誤差はすべてそのまま $h = Lr$ に伝わり、しかも q は h に敏感である。
- **中間層**。$\mathrm{B}(500, r)$ を $\mathrm{Po}(h)$ で置き換えるのは非常に良い近似である。r はごく小さく L は大きいからである。しかし、続いて $h \approx 2.7$ の $\mathrm{Po}(h)$ を正規分布で近似することは正当化できない。平均が 3 未満のポアソン分布は著しく歪んでおり、通常の目安では平均が少なくともおよそ 10（15 や 20 とする人もいる）必要とされる。

したがって、この近似が正確であるとは期待できない。疑わしいのは、$\mathrm{B}(100, 0.6)$ の裾に対する正規近似と、平均の小さいポアソン分布に対する正規近似の 2 段階である。

実際、r の厳密な値は $\mathrm{P}(X \leq 47) \approx 0.00576$ で、正規近似による推定値より約 7% 大きい。よって h は約 2.88 となるべきであり、そのとき厳密な $\mathrm{P}(Y > 4)$ は 0.133 ではなく約 0.165 となる。最終的な確率は 0.18 ではなく約 0.23 である。（この特定の数値では、$\mathrm{Po}(2.68)$ の正規近似はたまたまポアソン分布そのものとほぼ同じ裾確率を与え、0.133 に対して 0.134 である。したがって誤差の大部分は最下層で生じている。しかし、この幸運を前もって当てにすることはできなかったはずである。）

検算

- 連続性補正の向きは正しい。「m 未満」は $X \leq m - 1$ なので $m - \frac{1}{2}$ で止め、「l より多い」は $Y \geq l + 1$ なので $l + \frac{1}{2}$ から始める。与えられた式にはちょうど $2m - 1$ と $2l + 1$ が現れている。

- 数値的には、$\Phi(-2.5516) = 0.005362$、$h = 2.681$、$q = \Phi(-1.1108) = 0.1333$ であり、0.1825 を得る。二項分布の厳密な和を用いると $r = 0.005761$、$q = 0.1645$（2×10^4 日分のシミュレーションでも 0.164 となり確認できた）であり、最終的な答えは 0.2314 となる。

よくある間違い

- $m - \frac{1}{2}$ の代わりに m を（あるいは $l + \frac{1}{2}$ の代わりに l を）使うこと。そうすると式が一致しない。
- $\sqrt{h(1-r)}$ を使ってしまい、$\sqrt{h}$ にたどり着けなくなること。問題の式で因子 $1 - r$ が落ちているのは、実質的にポアソン近似を用いているからであり、そのことを明記すべきである。
- 「$Mp > 5$ かつ $M(1-p) > 5$」だけで精度を判断すること。この目安は分布の中心部分についてのものであり、0.005 という裾確率についてのものではない。

STEP のその先へ

この問題は、近似を連鎖させたときに誤差がどのように伝わるか、そして裾における絶対誤差と相対誤差の違いを扱っている。大学では、中心極限定理（central limit theorem）はベリー–エッセンの定理（Berry–Esseen theorem）によって精密化される。この定理は正規近似の絶対誤差を評価するものである。裾確率には別の道具が必要であり、たとえばチェルノフ限界（Chernoff bounds）や大偏差理論（large-deviation theory）は、裾確率がどれほど速く小さくなるかを記述する。Blitzstein と Hwang の *Introduction to Probability*（確率論入門）は、ポアソン近似と正規近似、およびチェルノフ型の不等式をわかりやすく扱っている。Grimmett と Welsh の *Probability: An Introduction*（確率論入門）は良い副読本である。

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。