解答例

STEP III 2000、第 12 問

確率・統計 ・ ポアソン近似

> **STEP III 2000、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/00-s3-q12

鍵となる考え方

自分の券が当たっているとわかってしまえば、残る偶然性は他のプレイヤーの券だけである。他のプレイヤーはそれぞれ独立に確率 $\frac{1}{N}$ で当選番号と一致するので、**他の**当選者の人数は二項分布（binomial distribution）に従う。プレイヤーが多く成功確率が小さいとき、この二項分布はポアソン分布（Poisson distribution）に近く、与えられた極限 $\left(1-\frac{a}{n}\right)^n \approx \mathrm{e}^{-a}$ こそがこの近似を成り立たせるものである。

解答

厳密な確率

自分が当たり券を持っているとする。他の $2N-1$ 人のプレイヤーのうち当選番号も持っている人数を X とすると、当選者の総数は $W = 1 + X$ である。

機械は各プレイヤーの番号を一様に、他の全員の番号とは独立に、また抽選とも独立に選ぶ。したがって、当選番号（それが何であれ）と自分の券が与えられたとき、他の $2N-1$ 人はそれぞれ互いに独立に確率 $\frac{1}{N}$ でそれと一致する。よって

$$X \sim \mathrm{B}\left(2N-1,\ \tfrac{1}{N}\right).$$

当選者が 3 人以下であることは $X \le 2$ を意味するので、厳密な確率は

$$\mathrm{P}(W \le 3) = \sum_{k=0}^{2} \binom{2N-1}{k} \left(\frac{1}{N}\right)^k \left(1-\frac{1}{N}\right)^{2N-1-k}.$$

すべて書き出すと、これは

$$\left(1-\frac{1}{N}\right)^{2N-1} + \frac{2N-1}{N}\left(1-\frac{1}{N}\right)^{2N-2} + \frac{(2N-1)(2N-2)}{2N^2}\left(1-\frac{1}{N}\right)^{2N-3}.$$

近似

固定した各 $k \in \{0,1,2\}$ について、累乗を次のように分ける。

$$\left(1-\frac{1}{N}\right)^{2N-1-k} = \left[\left(1-\frac{1}{N}\right)^{N}\right]^{2}\left(1-\frac{1}{N}\right)^{-(1+k)}.$$

与えられた事実を $a=1$、$n=N$ として用いると、角括弧の中はおよそ e^{-1} である。最後の因子は 1 に非常に近い数の固定された小さな累乗なので、およそ 1 である。よって各累乗はおよそ e^{-2} である。

二項係数については、k は高々 2 で N は大きいので

$$\binom{2N-1}{k}\frac{1}{N^k} = \frac{(2N-1)(2N-2)\cdots(2N-k)}{k!\,N^k} \approx \frac{(2N)^k}{k!\,N^k} = \frac{2^k}{k!}.$$

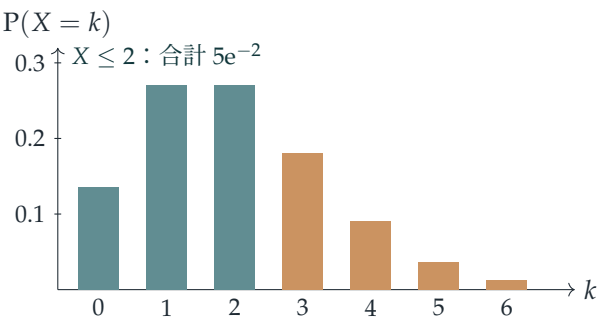
これらを合わせると、$k=0,1,2$ の項はそれぞれおよそ e^{-2}、$2\mathrm{e}^{-2}$、$2\mathrm{e}^{-2}$ となるので

$$\mathrm{P}(W \le 3) \approx \mathrm{e}^{-2}\left(1+2+\frac{2^2}{2!}\right) = 5\mathrm{e}^{-2} \approx 0.677 \approx \frac{2}{3},$$

ここで $\mathrm{e}^2 \approx 7.39$ を用いた。これが求める結果である。

考えてみよう なぜ X は平均 2 のポアソン変数のように振る舞うのだろうか。

上の計算は、二項分布の標準的なポアソン極限である。試行回数 $2N-1$ は大きく、成功確率 $\frac{1}{N}$ は小さく、その積は 2 に近いまま保たれる。このような状況では $\lambda = np$ として $\binom{n}{k}p^k(1-p)^{n-k} \to \mathrm{e}^{-\lambda}\lambda^k/k!$ が成り立ち、ここでは $\lambda \to 2$ である。



X の極限分布は平均 2 のポアソン分布である。青緑色の棒（$X \le 2$、すなわち当選者が全部で 3 人以下）は全体のおよそ 3 分の 2 を占める。

プレイヤーが自分で番号を選ぶ場合

プレイヤーが番号を選ぶ場合、番号はもはや一様ではない。人は誕生日（31 までの数）、ラッキーナンバー、規則的な並びを好むからである。プレイヤーが番号 i を確

率 p_i（$\sum p_i = 1$）で選び、当選番号は依然として一様に抽選されるとしよう。このとき、自分が当選したという条件のもとで、自分の番号が i である確率は $p_i \cdot \frac{1}{N}$ に比例し、すなわち p_i である。したがって、人気のある番号は当たり券の中に多く現れることになる。自分の番号が i のとき、他の当選者の人数は $\mathrm{B}(2N-1, p_i)$ に従い、その平均は $(2N-1)p_i$ である。i について平均をとると、他の当選者の人数の期待値は

$$(2N-1)\sum_i p_i^2 \;\geq\; (2N-1)\cdot\frac{1}{N},$$

となる。これはコーシー（Cauchy）・シュワルツ（Schwarz）の不等式（Cauchy–Schwarz inequality）$(\sum p_i)^2 \leq N\sum p_i^2$ による。等号はすべての $p_i = \frac{1}{N}$ のときに限り成り立つ。したがって、番号の選び方が一様でないと、平均的には当選を分け合う人が増え、当選者が 3 人以下である確率は通常**減少する**。（人気のない番号を意図的に選ぶ一人のプレイヤーは平均より有利になるが、プレイヤー全体としては不利になる。）

当選者数の期待値

$X \sim \mathrm{B}\left(2N-1, \frac{1}{N}\right)$ であるから、その平均は $np = (2N-1)\cdot\frac{1}{N}$ である。（同じことだが、X を他のプレイヤー一人ひとりに対応する $2N-1$ 個の指示変数（indicator variable）の和として書けば、それぞれの平均は $\frac{1}{N}$ である。）したがって

$$\mathrm{E}(W) = 1 + \mathrm{E}(X) = 1 + \frac{2N-1}{N} = 3 - \frac{1}{N} = 3 - N^{-1},$$

となり、求める結果が得られた。これは近似ではなく厳密な値であることに注意しよう。

検算

- $N = 2$ および $N = 3$ について、$2N$ 人のプレイヤーへの番号の割り当てと当選番号をすべて数え上げると、二項分布の式とちょうど一致し、期待値もちょうど $3 - \frac{1}{N}$ となる。
- 厳密な確率は $N = 10$ で 0.7054、$N = 100$ で 0.6794、$N = 10\,000$ で 0.6767 であり、$5\mathrm{e}^{-2} = 0.67668$ に近づいていく。
- $N = 50$ のシミュレーション（10^5 週分）では、厳密値 0.682 に対して 0.683 が得られ、当選者数の平均は $3 - \frac{1}{50} = 2.98$ に対して 2.981 であった。
- $N = 10$ で、番号の半分が残りの半分の 2 倍人気がある場合、確率は 0.705 から 0.641 に下がり、当選者数の期待値は 2.9 から 3.11 に上がる。

よくある間違い

- 試行回数を $2N-1$ ではなく $2N$ とすること。自分の券が当たっていることはすでにわかっているので、偶然に左右されるのは他のプレイヤーだけである。試行回数を $2N$ とすると、期待値は $3-\frac{1}{N}$ ではなく $1+2=3$ になってしまう。
- 条件なしの平均を使うこと。すべての週を通じた当選者数の期待値は $2N\cdot\frac{1}{N}=2$ であるが、自分が当選していると知ることで確実な当選者が 1 人加わり、残りは他の $2N-1$ 人となる。
- $k=2$ の項を忘れること、あるいはそれを $\frac{1}{2!}$ を付けずに 2^2e^{-2} と書くこと。

STEP のその先へ

ここでの近似は少数の法則（law of rare events）である。すなわち、試行回数が多く成功確率が小さい二項分布はポアソン分布に収束（converge）する。最後の小問は、検査のパラドックス（inspection paradox）、あるいはサイズバイアス抽出（size-biased sampling）に関わっている。自分の券が当たったという条件を付けると混み合った番号が選ばれやすくなり、それが条件付き平均が 2 ではなく 3 になる理由である。どちらの考え方も、Grimmett と Welsh の *Probability: An Introduction*（確率論入門）や、Blitzstein と Hwang の *Introduction to Probability*（確率入門）でよく扱われている。

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。