詳細解答

STEP III 2000，第 12 題

概率與統計 • 泊松近似

> **STEP III 2000，第 12 題**
>
> 繼續閱讀之前，請先看題目（原題為英文）。題目載於 STEP Support Programme 的試題資料庫：
>
> step.maths.org/questions/00-s3-q12

關鍵思路

一旦知道自己的彩票中獎，餘下的隨機性就只在其他玩家的彩票上。他們每人各自獨立地以概率 $\frac{1}{N}$ 選中中獎號碼，所以其他中獎者的人數服從二項分佈（binomial distribution）。當玩家眾多而成功概率很小時，這個二項分佈接近泊松分佈（Poisson distribution），而題目給出的極限 $\left(1-\frac{a}{n}\right)^n \approx \mathrm{e}^{-a}$ 正是令這個近似成立的關鍵。

解答

精確概率

假設你持有一張中獎彩票。設 X 為其餘 $2N-1$ 位玩家中同樣持有中獎號碼的人數，則中獎者總數為 $W = 1 + X$。

機器為每位玩家選號時，是均勻（uniform）地抽選，與其他人的號碼互相獨立（independent），亦與開獎結果獨立。因此，給定中獎號碼（不論是哪一個）及你的彩票後，其餘 $2N-1$ 位玩家每人以概率 $\frac{1}{N}$ 選中該號碼，且彼此獨立。所以

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

中獎者不多於三人即 $X \le 2$，故精確概率為

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

完整寫出來就是

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

近似值

對每個固定的 $k \in \{0,1,2\}$，把冪拆開：

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

在題目給出的結果中取 $a=1$ 及 $n=N$，方括號內的式子約為 e^{-1}。最後一個因子是一個非常接近 1 的數的固定小次冪，所以約為 1。因此每個冪都約為 e^{-2}。

至於二項式係數（binomial coefficient），k 最大為 2 而 N 很大，所以

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

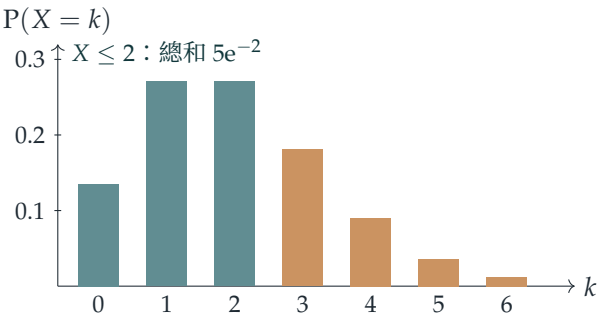
綜合以上各項，$k=0,1,2$ 的項分別約為 e^{-2}、$2\mathrm{e}^{-2}$ 和 $2\mathrm{e}^{-2}$，所以

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

其中用了 $\mathrm{e}^2 \approx 7.39$。這就是所求的結果。

想一想 為甚麼 X 的表現有如一個平均值為 2 的泊松變量？

上面的計算就是二項分佈的標準泊松極限。試驗次數 $2N-1$ 很大，成功概率 $\frac{1}{N}$ 很小，而兩者之積始終接近 2。在這種情況下，$\binom{n}{k}p^k(1-p)^{n-k} \to \mathrm{e}^{-\lambda}\lambda^k/k!$，其中 $\lambda = np$，而這裏 $\lambda \to 2$。



X 的極限分佈是平均值為 2 的泊松分佈。藍綠色的長條（$X \le 2$，即總共最多三位中獎者）約佔總和的三分之二。

玩家自行選號

如果由玩家自行選號，號碼就不再均勻分佈：人們偏愛生日（不超過 31 的號碼）、幸運號碼和有規律的號碼。假設每位玩家以概率 p_i 選擇號碼 i，其中 $\sum p_i = 1$，而中獎號碼

仍是均勻抽出。那麼，在你中獎的條件下，你的號碼是 i 的概率與 $p_i \cdot \frac{1}{N}$ 成正比，即概率為 p_i。所以熱門號碼在中獎彩票中所佔的比例偏高。若你的號碼是 i，其他中獎者的人數服從 $\mathrm{B}(2N-1, p_i)$，平均值為 $(2N-1)p_i$。對 i 取平均，其他中獎者人數的期望值（expectation）為

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

這是根據柯西－施瓦茲不等式（Cauchy–Schwarz inequality）$(\sum p_i)^2 \leq N\sum p_i^2$ 得出的，等號僅在每個 $p_i = \frac{1}{N}$ 時成立。所以選號不均勻時，平均而言會有更多人與你分享獎金，而中獎者不多於三人的概率通常會下降。（個別玩家若刻意選擇冷門號碼，表現會比平均好，但玩家整體而言則較差。）

中獎者人數的期望值

由於 $X \sim \mathrm{B}\left(2N-1, \frac{1}{N}\right)$，其平均值為 $np = (2N-1)\cdot\frac{1}{N}$。（等價地，可把 X 寫成 $2N-1$ 個指示變量（indicator variable）之和，每位其他玩家對應一個，各自的平均值為 $\frac{1}{N}$。）因此

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

正如所求。注意這是精確結果，並非近似值。

驗算

- 當 $N = 2$ 和 $N = 3$ 時，列舉 $2N$ 位玩家所有可能的號碼分配及所有可能的中獎號碼，所得結果與二項式表達式完全吻合，期望值亦恰為 $3 - \frac{1}{N}$。
- 精確概率在 $N = 10$ 時為 0.7054，$N = 100$ 時為 0.6794，$N = 10\,000$ 時為 0.6767，趨近 $5\mathrm{e}^{-2} = 0.67668$。
- 以 $N = 50$ 進行模擬（10^5 個星期），得出 0.683，而精確值為 0.682；中獎者人數的平均值為 2.981，而 $3 - \frac{1}{50} = 2.98$。
- 當 $N = 10$ 而其中一半號碼的受歡迎程度是另一半的兩倍時，概率由 0.705 降至 0.641，中獎者人數的期望值則由 2.9 升至 3.11。

常見錯誤

- 用了 $2N$ 次試驗而非 $2N-1$ 次：你自己的彩票已知中獎，所以只有其他玩家是隨機的。若用 $2N$ 次試驗，期望值會算成 $1+2=3$，而不是 $3 - \frac{1}{N}$。
- 用了無條件平均值：就所有星期而言，中獎者人數的期望值是 $2N \cdot \frac{1}{N} = 2$，但知道你中獎後，便多了一位必然的中獎者，餘下 $2N-1$ 位其他玩家。

► 遺漏了 $k = 2$ 的項，或把它寫成 2^2e^{-2} 而漏掉 $\frac{1}{2!}$。

STEP 之外

這裏的近似就是稀有事件定律（law of rare events）：試驗次數多而成功概率小的二項分佈會收斂（converge）至泊松分佈。最後一部分觸及檢查悖論（inspection paradox），又稱大小偏差抽樣（size-biased sampling）。以你的彩票中獎為條件，會令多人選擇的號碼更有可能出現，這正是條件平均值是 3 而不是 2 的原因。這兩個概念在 Grimmett 與 Welsh 的 *Probability: An Introduction*（概率入門教材）以及 Blitzstein 與 Hwang 的 *Introduction to Probability*（概率導論教材）中都有很好的論述。

Beyond STEP 詳細解答：https://beyondstep.keithmcnulty.org/zh-hant/。解答文字採用 CC0 1.0；題目本身版權屬 Cambridge University Press & Assessment 所有（©），本文件不予轉載。由機器從英文翻譯而成；尚未經審閱。