详细解答

STEP III 2000，第 12 题

概率与统计 · 泊松近似

> **STEP III 2000，第 12 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/00-s3-q12

关键思路

一旦知道自己的彩票中奖，余下的随机性就只在其他玩家的彩票上。他们每人各自独立地以概率 $\frac{1}{N}$ 选中中奖号码，所以其他中奖者的人数服从二项分布（binomial distribution）。当玩家众多而成功概率很小时，这个二项分布接近泊松分布（Poisson distribution），而题目给出的极限 $\left(1-\frac{a}{n}\right)^n \approx \mathrm{e}^{-a}$ 正是令这个近似成立的关键。

解答

精确概率

假设你持有一张中奖彩票。设 X 为其余 $2N-1$ 位玩家中同样持有中奖号码的人数，则中奖者总数为 $W = 1 + X$。

机器为每位玩家选号时，是均匀（uniform）地抽选，与其他人的号码互相独立（independent），亦与开奖结果独立。因此，给定中奖号码（不论是哪一个）及你的彩票后，其余 $2N-1$ 位玩家每人以概率 $\frac{1}{N}$ 选中该号码，且彼此独立。所以

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

中奖者不多于三人即 $X \le 2$，故精确概率为

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

完整写出来就是

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

近似值

对每个固定的 $k \in \{0,1,2\}$，把幂拆开：

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

在题目给出的结果中取 $a=1$ 及 $n=N$，方括号内的式子约为 e^{-1}。最后一个因子是一个非常接近 1 的数的固定小次幂，所以约为 1。因此每个幂都约为 e^{-2}。

至于二项式系数（binomial coefficient），k 最大为 2 而 N 很大，所以

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

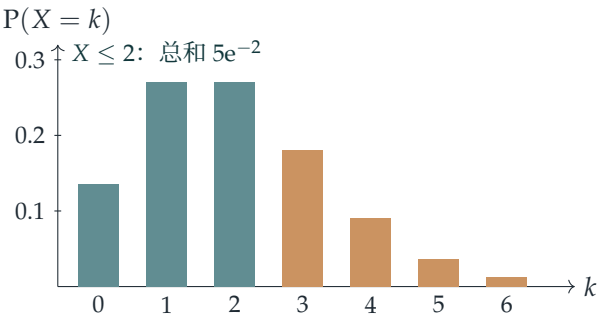
综合以上各项，$k=0,1,2$ 的项分别约为 e^{-2}、$2\mathrm{e}^{-2}$ 和 $2\mathrm{e}^{-2}$，所以

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

其中用了 $\mathrm{e}^2 \approx 7.39$。这就是所求的结果。

> **想一想** 为甚么 X 的表现有如一个平均值为 2 的泊松变量?

上面的计算就是二项分布的标准泊松极限。试验次数 $2N-1$ 很大，成功概率 $\frac{1}{N}$ 很小，而两者之积始终接近 2。在这种情况下，$\binom{n}{k}p^k(1-p)^{n-k} \to \mathrm{e}^{-\lambda}\lambda^k/k!$，其中 $\lambda = np$，而这里 $\lambda \to 2$。



X 的极限分布是平均值为 2 的泊松分布。蓝绿色的长条（$X \le 2$，即总共最多三位中奖者）约占总和的三分之二。

玩家自行选号

如果由玩家自行选号，号码就不再均匀分布：人们偏爱生日（不超过 31 的号码）、幸运号码和有规律的号码。假设每位玩家以概率 p_i 选择号码 i，其中 $\sum p_i = 1$，而中奖号码

仍是均匀抽出。那么，在你中奖的条件下，你的号码是 i 的概率与 $p_i \cdot \frac{1}{N}$ 成正比，即概率为 p_i。所以热门号码在中奖彩票中所占的比例偏高。若你的号码是 i，其他中奖者的人数服从 $\mathrm{B}(2N-1, p_i)$，平均值为 $(2N-1)p_i$。对 i 取平均，其他中奖者人数的期望值（expectation）为

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

这是根据柯西－施瓦兹不等式（Cauchy–Schwarz inequality）$(\sum p_i)^2 \leq N\sum p_i^2$ 得出的，等号仅在每个 $p_i = \frac{1}{N}$ 时成立。所以选号不均匀时，平均而言会有更多人与你分享奖金，而中奖者不多于三人的概率通常会下降。（个别玩家若刻意选择冷门号码，表现会比平均好，但玩家整体而言则较差。）

中奖者人数的期望值

由于 $X \sim \mathrm{B}\left(2N-1, \frac{1}{N}\right)$，其平均值为 $np = (2N-1)\cdot\frac{1}{N}$。（等价地，可把 X 写成 $2N-1$ 个指示变量（indicator variable）之和，每位其他玩家对应一个，各自的平均值为 $\frac{1}{N}$。）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。注意这是精确结果，并非近似值。

验算

- 当 $N=2$ 和 $N=3$ 时，列举 $2N$ 位玩家所有可能的号码分配及所有可能的中奖号码，所得结果与二项式表达式完全吻合，期望值亦恰为 $3-\frac{1}{N}$。
- 精确概率在 $N=10$ 时为 0.7054，$N=100$ 时为 0.6794，$N=10\,000$ 时为 0.6767，趋近 $5\mathrm{e}^{-2} = 0.67668$。
- 以 $N=50$ 进行模拟（10^5 个星期），得出 0.683，而精确值为 0.682；中奖者人数的平均值为 2.981，而 $3-\frac{1}{50} = 2.98$。
- 当 $N=10$ 而其中一半号码的受欢迎程度是另一半的两倍时，概率由 0.705 降至 0.641，中奖者人数的期望值则由 2.9 升至 3.11。

常见错误

- 用了 $2N$ 次试验而非 $2N-1$ 次：你自己的彩票已知中奖，所以只有其他玩家是随机的。若用 $2N$ 次试验，期望值会算成 $1+2=3$，而不是 $3-\frac{1}{N}$。
- 用了无条件平均值：就所有星期而言，中奖者人数的期望值是 $2N\cdot\frac{1}{N} = 2$，但知道你中奖后，便多了一位必然的中奖者，余下 $2N-1$ 位其他玩家。

► 遗漏了 $k = 2$ 的项，或把它写成 2^2e^{-2} 而漏掉 $\frac{1}{2!}$。

STEP 之外

这里的近似就是稀有事件定律（law of rare events）：试验次数多而成功概率小的二项分布会收敛（converge）至泊松分布。最后一部分触及检查悖论（inspection paradox），又称大小偏差抽样（size-biased sampling）。以你的彩票中奖为条件，会令多人选择的号码更有可能出现，这正是条件平均值是 3 而不是 2 的原因。这两个概念在 Grimmett 与 Welsh 的 *Probability: An Introduction*（概率入门教材）以及 Blitzstein 与 Hwang 的 *Introduction to Probability*（概率导论教材）中都有很好的论述。

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。