详细解答

# STEP III 1997，第 12 题

概率与统计 • 赠券收集问题

> **STEP III 1997，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/97-s3-q12

## 关键思路

第 (i) 部分是几何分布（geometric distribution），而它的概率生成函数（probability generating function）只需求导，便可给出平均值和方差：$\mathrm{E}(K) = \mathrm{G}'(1)$ 及 $\mathrm{E}(K(K-1)) = \mathrm{G}''(1)$。在第 (ii) 部分，窍门是把集齐全套的漫长等待拆成 52 段短等待，每段对应一张新牌。每段短等待都是一个“掷硬币直至首次出现正面”的实验，所以第 (i) 部分已经把工夫全做好了。

## 解答

### 第 (i) 部分

首次正面出现在第 $k$ 次投掷，当且仅当头 $k-1$ 次投掷都是反面，而第 $k$ 次是正面。各次投掷互相独立（independent），所以

$$\mathrm{P}(K = k) = q^{k-1}p, \qquad k = 1, 2, 3, \ldots$$

因此

$$\mathrm{G}(s) = \sum_{k=1}^{\infty} pq^{k-1}s^k = ps\sum_{j=0}^{\infty}(qs)^j = \frac{ps}{1-qs},$$

这里用了等比级数（geometric series）求和，该级数在 $|qs| < 1$ 时收敛（特别是当 $p > 0$ 时，对接近 1 的 $s$ 收敛）。这就是所要求的结果。

求导得

$$G'(s) = \frac{p(1-qs)+psq}{(1-qs)^2} = \frac{p}{(1-qs)^2},$$
$$G''(s) = \frac{2pq}{(1-qs)^3}.$$

由于 $G'(s) = \sum k\,P(K=k)s^{k-1}$ 及 $G''(s) = \sum k(k-1)P(K=k)s^{k-2}$，代入 $s=1$ 并利用 $1-q=p$，可得

$$E(K) = G'(1) = \frac{p}{p^2} = \frac{1}{p}, \qquad E(K(K-1)) = G''(1) = \frac{2pq}{p^3} = \frac{2q}{p^2}.$$

所以

$$\begin{aligned} \mathrm{Var}(K) &= E(K(K-1)) + E(K) - (E(K))^2 \\ &= \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2q+p-1}{p^2} = \frac{q}{p^2}. \end{aligned}$$

故此 $E(K) = \dfrac{1}{p}$ 及 $\mathrm{Var}(K) = \dfrac{q}{p^2}$。

**第 (ii) 部分**

设 $N_i$ 为已经见过恰好 $i-1$ 张不同的牌之后所抽的次数，计至并包括抽出第 $i$ 张不同的牌的那一次。于是 $N_1 = 1$（第一张牌必定是新的），而

$$N = N_1 + N_2 + \cdots + N_{52}.$$

**想一想** 为甚么每个 $N_i$ 都是像 $K$ 那样的“首次正面”变量?

一旦见过 $i-1$ 张不同的牌，52 张牌之中仍有 $53-i$ 张未见过。由于抽牌后放回（with replacement），每次抽牌都与过去的结果独立，而且抽出未见过的牌的概率都相同，即

$$p_i = \frac{53-i}{52}.$$

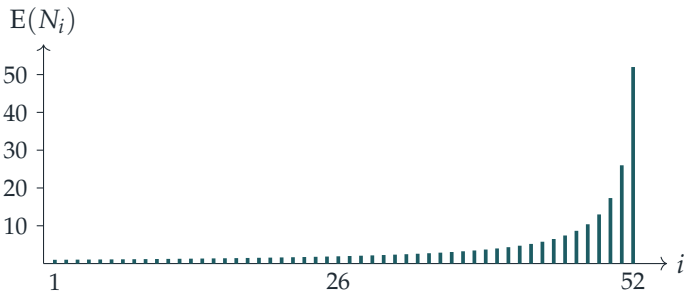
把这样的一次抽牌称为“正面”。那么 $N_i$ 就是计至并包括首次正面的抽牌次数，正正是第 (i) 部分中的变量 $K$，其中 $p = p_i$。因此

$$E(N_i) = \frac{1}{p_i} = \frac{52}{53-i}.$$

期望值（expectation）是线性的（这里毋须各个 $N_i$ 互相独立），所以

$$\begin{aligned} E(N) &= \sum_{i=1}^{52} \frac{52}{53-i} = 52\left(\frac{1}{52} + \frac{1}{51} + \cdots + \frac{1}{2} + \frac{1}{1}\right) \\ &= 52\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{52}\right). \end{aligned}$$

早段的等待很短，最后几段才是主导：单是最后一张牌，平均就要抽 52 次。




要作估算，可用 $1+\frac{1}{2}+\cdots+\frac{1}{n}\approx 0.5+\ln n$，取 $n=52$。至于 $\ln 52$：$2.7^2=7.29$ 而 $7.29^2\approx 53.1$，所以 $e^4\approx 53\approx 52$，得出 $\ln 52\approx 4$。因此

$$\mathrm{E}(N)\approx 52\,(0.5+4)=52\times 4.5=234.$$

所以平均而言，我们大约要抽 234 次，即约为整副牌张数的四倍半，才能见齐每一张牌。

## 验算

- 在 (i) 中，$\mathrm{G}(1)=p/(1-q)=1$，任何概率生成函数都必须如此。若硬币是公平的 $(p=\frac{1}{2})$，便得 $\mathrm{E}(K)=2$ 及 $\mathrm{Var}(K)=2$；而当 $p\to 1$ 时，方差 $q/p^2\to 0$，这正是第一次投掷几乎必然是正面时应有的结果。
- 准确值是 $52H_{52}\approx 235.98$，而模拟 $10^5$ 次完整收集所得的平均值是 235.7。我们的估计值 234 相当接近；那少许差额来自近似式 $0.5+\ln n$（较精确的是 $\ln n+0.577$），以及把 $\ln 52\approx 3.95$ 上舍入为 4。

## 常见错误

- 把 $N_i$ 的成功概率弄颠倒：抽到新牌的机会是 $(53-i)/52$，它随着收集到的牌增多而减小。
- 取 $\mathrm{Var}(K)=\mathrm{G}''(1)-\mathrm{G}'(1)^2$，忘记了 $+\mathrm{G}'(1)$ 这一项；$\mathrm{G}''(1)$ 是 $\mathrm{E}(K(K-1))$，而不是 $\mathrm{E}(K^2)$。
- 以为各个 $N_i$ 必须互相独立，才可以把期望值相加。无论独立与否，期望值的线性性质都成立（不过在这里它们其实是独立的）。

## STEP 之外

这就是*优惠券收集问题*（coupon collector's problem）：若有 $n$ 种出现机会均等的类型，集齐全部所需的期望时间是 $nH_n\approx n\ln n$。大学本科课程会更进一步，证明这个时间高

度集中在 $n \ln n$ 附近，波动的量级为 $n$；而概率生成函数亦会成为处理这类独立变量之和的标准工具。Grimmett 与 Welsh 的 *Probability: An Introduction* 仔细讲解生成函数和几何分布；Blitzstein 与 Hwang 的 *Introduction to Probability* 则讨论优惠券收集问题及其多种变化。

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。