详细解答

STEP III 2008，第 13 题

概率与统计 · 指示变量之和

> **STEP III 2008，第 13 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/08-s3-q13

本题在 *Beyond STEP* 第 12.10 节把绳头系起来中有完整的引导式讲解：https://beyondstep.keithmcnulty.org/zh-hans/ch12/loose-ends.html

关键思路

无论某一步发生甚么事，绳子的段数总是刚好减少一：要么有一个绳圈退出游戏，要么两段绳子合成一段。因此每一步之后，不管之前发生过甚么，我们都回到同样的处境，只是少了一段绳子。这样就可以把绳圈的总数写成一些独立（independent）指示变量（indicator variable）之和，于是平均值和方差（variance）都只是简单的求和。

解答

第一步

有 n 段绳子时，共有 $2n$ 个自由端，所以选取两个端点的方法有

$$\binom{2n}{2} = n(2n-1)$$

种，而且各种选法的可能性相等。当且仅当两个端点属于同一段绳子时，才会形成绳圈，而这样的配对共有 n 个（每段绳子一个）。因此

$$\mathrm{P}(\text{第一步形成绳圈}) = \frac{n}{n(2n-1)} = \frac{1}{2n-1}.$$

换个说法：无论我们先选哪一个端点，余下的 $2n-1$ 个端点之中，恰好有一个是它的同伴。设 X_n 在第一步形成绳圈时取值 1，否则取值 0，则第一步所形成绳圈数目的期望值

（expectation）是

$$\mathrm{E}(X_n) = \frac{1}{2n-1}.$$

绳圈数目的期望值

第一步之后，剩下 $n-1$ 段绳子，每段各有两个自由端。两种情况下都是如此：形成绳圈会令一段绳子退出游戏，而把两段不同的绳子接起来则令两段变成一段。绳子的长短无关重要，要紧的只是它们的端点。

> **想一想** 现在有些绳子比其他的长，又或者之前有没有形成过绳圈，这些有影响吗？

没有影响。唯一会影响选中哪一对端点的，就是还剩下多少段绳子（因而还剩下多少个端点）。所以当剩下 k 段绳子时，不论之前各步的结果如何，这一步形成绳圈的概率都是 $\frac{1}{2k-1}$。

以每一步开始前的绳子段数 $k = n, n-1, \ldots, 1$ 来标记各步，并设 X_k 为该步形成绳圈的指示变量。于是 $\mathrm{P}(X_k = 1) = \frac{1}{2k-1}$，而整个过程在 $k = 1$ 那一步之后结束（这一步必定形成绳圈，即 $X_1 = 1$）。绳圈的总数是

$$R = X_1 + X_2 + \cdots + X_n,$$

所以由期望值的线性性质（linearity of expectation）可得

$$\mathrm{E}(R) = \sum_{k=1}^{n} \frac{1}{2k-1} = 1 + \frac{1}{3} + \frac{1}{5} + \cdots + \frac{1}{2n-1}.$$

方差

求方差时，单靠线性性质并不足够：我们还需要各个 X_k 互相独立。它们确实如此。无论之前各步的结果是甚么，有 k 段绳子的那一步形成绳圈的概率仍然是 $\frac{1}{2k-1}$。所以在已知之前所有 X 的条件下，X_k 的条件分布（conditional distribution）与它的无条件分布相同，因此各个 X_k 互相独立。

每个 X_k 都服从伯努利分布（Bernoulli distribution），其中 $p_k = \frac{1}{2k-1}$，所以 $\mathrm{Var}(X_k) = p_k(1-p_k)$；而对于独立变量，方差可以相加：

$$\begin{aligned}\mathrm{Var}(R) &= \sum_{k=1}^{n} \frac{1}{2k-1}\left(1 - \frac{1}{2k-1}\right) = \sum_{k=1}^{n} \frac{2k-2}{(2k-1)^2} \\ &= \sum_{k=1}^{n} \frac{1}{2k-1} - \sum_{k=1}^{n} \frac{1}{(2k-1)^2}.\end{aligned}$$

$n = 40\,000$ 的情况

记 $H_m = 1 + \frac{1}{2} + \cdots + \frac{1}{m}$。直至 $2n-1$ 为止的奇数倒数之和，等于直至 $2n$ 为止的所有倒数之和，减去其中偶数的倒数：

$$\sum_{k=1}^{n} \frac{1}{2k-1} = H_{2n} - \left(\frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2n}\right) = H_{2n} - \tfrac{1}{2}H_n.$$

利用 m 很大时 $H_m \approx \ln m$，

$$\mathrm{E}(R) \approx \ln(2n) - \tfrac{1}{2}\ln n = \tfrac{1}{2}\left(2\ln 2n - \ln n\right) = \tfrac{1}{2}\ln(4n).$$

当 $n = 40\,000$ 时，$4n = 160\,000 = 20^4$，所以

$$\mathrm{E}(R) \approx \tfrac{1}{2}\ln\left(20^4\right) = 2\ln 20 \approx 6.$$

因此，由 40 000 段绳子开始，我们预期只会得到大约六个绳圈。

验算

- $n = 1$：只有一段绳子，必定得到一个绳圈。公式给出 $\mathrm{E}(R) = 1$ 和 $\mathrm{Var}(R) = 0$，正如所料。
- $n = 2$：第一步形成绳圈的概率是 $\frac{1}{3}$，而最后一步必定形成绳圈，所以 R 取值 2 的概率是 $\frac{1}{3}$，否则取值 1。由此得平均值为 $\frac{4}{3}$，方差为 $\frac{2}{9}$，与公式给出的 $1 + \frac{1}{3}$ 和 $\frac{2}{9}$ 吻合。
- 对 $n \le 5$ 精确地列举所有可能的选择序列，并对 $n = 6$ 进行模拟，结果都与两条公式一致。当 $n = 40\,000$ 时，精确的和是 $\mathrm{E}(R) \approx 6.28$。比我们的估计多出的 0.29，来自 $H_m \approx \ln m$ 这个粗略的近似（见下文）。

常见错误

- 把形成绳圈的概率当作 $\frac{1}{n}$（选两段绳子），而不是 $\frac{1}{2n-1}$（选两个端点）。
- 没有先说明各个 X_k 互相独立，就把方差相加。求平均值时有线性性质已经足够，求方差则不然。
- 没有把偶数项分拆出来，就以 $\ln n$ 或 $\frac{1}{2}\ln n$ 来近似 $\sum \frac{1}{2k-1}$。

STEP 之外

更精确的估计是 $H_m = \ln m + \gamma + o(1)$，其中 $\gamma \approx 0.5772$ 是欧拉－马斯刻若尼常数（Euler–Mascheroni constant）。由此可得 $\mathrm{E}(R) \approx \frac{1}{2}\ln(4n) + \frac{1}{2}\gamma$，以及 $\mathrm{Var}(R) \approx \frac{1}{2}\ln(4n) + \frac{1}{2}\gamma -$

$\frac{\pi^2}{8}$，因为 $\sum 1/(2k-1)^2 = \pi^2/8$。同样的“概率为 $1/k$ 的独立指示变量之和”结构，也可以用来数算随机置换（random permutation）的循环（cycle）数目，以及随机数列中破纪录（record）的次数；两种情况下，这个数目的增长都与 $\ln n$ 相若。想进一步了解指示变量和这些随机结构，Grimmett 与 Welsh 的 *Probability: An Introduction*（概率论入门教材）是个很好的起点。

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。