詳細解答

STEP III 2008，第 13 題

概率與統計 • 指示變量之和

> **STEP III 2008，第 13 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/08-s3-q13

本題在 *Beyond STEP* 第 12.10 節把繩頭繫起來中有完整的引導式講解：https://beyondstep.keithmcnulty.org/zh-hant/ch12/loose-ends.html

關鍵思路

無論某一步發生甚麼事，繩子的段數總是剛好減少一：要麼有一個繩圈退出遊戲，要麼兩段繩子合成一段。因此每一步之後，不管之前發生過甚麼，我們都回到同樣的處境，只是少了一段繩子。這樣就可以把繩圈的總數寫成一些獨立（independent）指示變量（indicator variable）之和，於是平均值和方差（variance）都只是簡單的求和。

解答

第一步

有 n 段繩子時，共有 $2n$ 個自由端，所以選取兩個端點的方法有

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

種，而且各種選法的可能性相等。當且僅當兩個端點屬於同一段繩子時，才會形成繩圈，而這樣的配對共有 n 個（每段繩子一個）。因此

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

換個說法：無論我們先選哪一個端點，餘下的 $2n-1$ 個端點之中，恰好有一個是它的同伴。設 X_n 在第一步形成繩圈時取值 1，否則取值 0，則第一步所形成繩圈數目的期望值

（expectation）是

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

繩圈數目的期望值

第一步之後，剩下 $n-1$ 段繩子，每段各有兩個自由端。兩種情況下都是如此：形成繩圈會令一段繩子退出遊戲，而把兩段不同的繩子接起來則令兩段變成一段。繩子的長短無關重要，要緊的只是它們的端點。

> **想一想** 現在有些繩子比其他的長，又或者之前有沒有形成過繩圈，這些有影響嗎？

沒有影響。唯一會影響選中哪一對端點的，就是還剩下多少段繩子（因而還剩下多少個端點）。所以當剩下 k 段繩子時，不論之前各步的結果如何，這一步形成繩圈的概率都是 $\frac{1}{2k-1}$。

以每一步開始前的繩子段數 $k = n, n-1, \ldots, 1$ 來標記各步，並設 X_k 為該步形成繩圈的指示變量。於是 $\mathrm{P}(X_k = 1) = \frac{1}{2k-1}$，而整個過程在 $k = 1$ 那一步之後結束（這一步必定形成繩圈，即 $X_1 = 1$）。繩圈的總數是

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

所以由期望值的線性性質（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

求方差時，單靠線性性質並不足夠：我們還需要各個 X_k 互相獨立。它們確實如此。無論之前各步的結果是甚麼，有 k 段繩子的那一步形成繩圈的概率仍然是 $\frac{1}{2k-1}$。所以在已知之前所有 X 的條件下，X_k 的條件分佈（conditional distribution）與它的無條件分佈相同，因此各個 X_k 互相獨立。

每個 X_k 都服從伯努利分佈（Bernoulli distribution），其中 $p_k = \frac{1}{2k-1}$，所以 $\mathrm{Var}(X_k) = p_k(1-p_k)$；而對於獨立變量，方差可以相加：

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

$n = 40\,000$ 的情況

記 $H_m = 1 + \frac{1}{2} + \cdots + \frac{1}{m}$。直至 $2n-1$ 為止的奇數倒數之和，等於直至 $2n$ 為止的所有倒數之和，減去其中偶數的倒數：

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

利用 m 很大時 $H_m \approx \ln m$，

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

當 $n = 40\,000$ 時，$4n = 160\,000 = 20^4$，所以

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

因此，由 40 000 段繩子開始，我們預期只會得到大約六個繩圈。

驗算

- $n = 1$：只有一段繩子，必定得到一個繩圈。公式給出 $\mathrm{E}(R) = 1$ 和 $\mathrm{Var}(R) = 0$，正如所料。
- $n = 2$：第一步形成繩圈的概率是 $\frac{1}{3}$，而最後一步必定形成繩圈，所以 R 取值 2 的概率是 $\frac{1}{3}$，否則取值 1。由此得平均值為 $\frac{4}{3}$，方差為 $\frac{2}{9}$，與公式給出的 $1 + \frac{1}{3}$ 和 $\frac{2}{9}$ 吻合。
- 對 $n \le 5$ 精確地列舉所有可能的選擇序列，並對 $n = 6$ 進行模擬，結果都與兩條公式一致。當 $n = 40\,000$ 時，精確的和是 $\mathrm{E}(R) \approx 6.28$。比我們的估計多出的 0.29，來自 $H_m \approx \ln m$ 這個粗略的近似（見下文）。

常見錯誤

- 把形成繩圈的概率當作 $\frac{1}{n}$（選兩段繩子），而不是 $\frac{1}{2n-1}$（選兩個端點）。
- 沒有先說明各個 X_k 互相獨立，就把方差相加。求平均值時有線性性質已經足夠，求方差則不然。
- 沒有把偶數項分拆出來，就以 $\ln n$ 或 $\frac{1}{2}\ln n$ 來近似 $\sum \frac{1}{2k-1}$。

STEP 之外

更精確的估計是 $H_m = \ln m + \gamma + o(1)$，其中 $\gamma \approx 0.5772$ 是歐拉－馬斯刻若尼常數（Euler–Mascheroni constant）。由此可得 $\mathrm{E}(R) \approx \frac{1}{2}\ln(4n) + \frac{1}{2}\gamma$，以及

$\mathrm{Var}(R) \approx \frac{1}{2}\ln(4n) + \frac{1}{2}\gamma - \frac{\pi^2}{8}$，因為 $\sum 1/(2k-1)^2 = \pi^2/8$。同樣的「概率為 $1/k$ 的獨立指示變量之和」結構，也可以用來數算隨機置換（random permutation）的循環（cycle）數目，以及隨機數列中破紀錄（record）的次數；兩種情況下，這個數目的增長都與 $\ln n$ 相若。想進一步了解指示變量和這些隨機結構，Grimmett 與 Welsh 的 *Probability: An Introduction*（概率論入門教材）是個很好的起點。

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。