9.11 集齊一套

工作紙

題目

STEP II 1994，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/94-s2-q13

提示

請先自行嘗試解題。如果遇到困難，請逐一閱讀提示，每讀完一個便回到題目再試。標有停一停，試一試的提示是關鍵步驟。每個提示之後均留有空位供你演算。

> **想一想**　如果每包各自獨立地以概率 p 開出一張新卡，那麼直至開出第一張新卡為止（連這一包計算在內），你開包數目的期望值（expectation）是多少？

階段 1　等待新卡

涵蓋：等待新卡

> **想一想**　已經有 r 張不同的卡時，下一包抽到新卡的概率是多少？利用幾何分佈（geometric distribution）的期望等待次數。

階段 2 比較和與積分

涵蓋：積分不等式

想一想 在 $r \le x \le r+1$ 上，$\frac{1}{x}$ 介乎哪兩個值之間？在這個長度為 1 的區間上積分。

停一停，試一試 把你的不等式由 $r = 1$ 加到 $r = n-1$。各個積分會連接成一個積分：是哪一個？

階段 3　集齊全套

涵蓋：集齊整套卡

? **想一想**　所需的總包數，就是每張新卡的等待次數之和。把它們的期望值（expectation）相加，再利用第二階段的結果，把這個和與 $\ln n$ 比較。

背後的數學

這個估計有多準確？

這個比較還可以再精確一些。圖 9.11.1（見網上該節的完整版）中曲線與上方長方形之間的陰影空隙，總面積是有限的，所以 $H_n - \ln n$ 趨向一個極限，

$$\gamma = \lim_{n\to\infty}(H_n - \ln n) = 0.5772156649\ldots,$$

即歐拉常數。（沒有人知道 γ 是否有理數。）因此包數的期望值是 $n\ln n + \gamma n + \frac{1}{2} + \cdots$。以 600 張貼紙的貼紙簿為例，$600 \times (\ln 600 + 0.577) \approx 4185$，這就是引子中所引述的數字。

時間都花到哪裏去了？各個階段極不平均。集齊全套的前一半只需大約 $n\ln 2 \approx 0.69n$ 包，而單是最後一張卡，平均就要 n 包。大部分工夫都花在最後幾張卡上，這正是收藏者要互相交換重複卡的原因。

集中現象

期望值只是個開始：總數的變動有多大？總數的方差（variance）是 $\sum nr/(n-r)^2$，約為 $\frac{\pi^2}{6}n^2$（Section 2.4 中的和 $\sum 1/k^2 = \pi^2/6$ 在這裏出現了）。所以標準差（standard deviation）約為 $1.28n$，當 n 很大時遠小於平均值 $n\ln n$：總數集中在其平均值附近。艾狄胥（Paul Erdős）和雷尼（Alfréd Rényi）在 1961 年證明了一個更精確的結果：總數減去 $n\ln n$ 再除以 n，會有一個極限分佈（limiting distribution），即耿貝爾分佈（Gumbel distribution），這個分佈也描述了許多隨機量之中最大的一個（Section 12.25）。

無處不在的集券問題

這個問題可追溯至棣莫弗（Abraham de Moivre，1712 年）和拉普拉斯（Laplace）。如今它以許多不同的面貌出現。要檢查一個系統的每個組件，需要做多少次隨機測試？在一個網絡上進行隨機漫步（random walk），要多久才會走遍每個節點？電腦的雜湊表（hash table）要探查多少次？每一種情況的答案都涉及 $n \ln n$。在生物學中，同樣的計算可用來估計要有多少個定序「讀段」（reads）才能覆蓋整個基因組。

延伸練習

1 **半套。**證明集齊 $\frac{n}{2}$ 張不同的卡（n 為偶數）所需包數的期望值是 $n(H_n - H_{n/2})$，並且這個值約為 $n \ln 2$。

2 **不同的卡。**證明開了 k 包之後，你手上不同的卡的數目的期望值是 $n(1 - (1 - \frac{1}{n})^k)$。（提示：對每一張卡，求出你擁有它的概率，然後相加。）大約要開多少包，才能得到全套的 95%？

3 **兩位收藏者。**兩個小孩收集同一套卡，並共用所開的各包，因此對其中一人來說是重複的卡，對另一人來說可能是新的。證明集齊兩套所需包數的期望值約為 $n(\ln n + \ln \ln n)$。（這題較難；先說服自己這個數目遠小於 $2n \ln n$。）

4 **更精確的界。**改用梯形法則（trapezium rule）的圖像而不用長方形，證明 $H_n - \ln n$ 是遞減的，而 $H_n - \ln n - \frac{1}{2n}$ 是遞增的。由較小的 n 值推導出 $0.5 < \gamma < 0.7$。

延伸閱讀

- **Geoffrey Grimmett & David Stirzaker, *One Thousand Exercises in Probability***（Oxford University Press）。集券問題的許多變化。
- **Julian Havil, *Gamma: Exploring Euler's Constant***（Princeton University Press）。關於 γ 和調和級數的一切，寫得深入淺出。
- **Michael Mitzenmacher & Eli Upfal, *Probability and Computing***（Cambridge University Press）。集券問題及其在電腦科學中的應用。

完整的引導式解答見第 9.11 節：

https://beyondstep.keithmcnulty.org/zh-hant/ch09/coupon-collector.html

每星期一道 STEP 題目，完整講解：The STEP Up，以電郵免費發送，訂閱網址：keithmcnulty.substack.com