詳細解答

STEP II 1994，第 13 題

概率與統計 • 贈券收集問題

> **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

本題在 *Beyond STEP* 第 9.11 節集齊一套中有完整的引導式講解：https://beyondstep.keithmcnulty.org/zh-hant/ch09/coupon-collector.html

關鍵思路

這是經典的「優惠券收集問題」（coupon collector problem），雖然資料庫把它歸入別的類別，它其實是純粹的概率（probability）題。把集齊全套所需的漫長等待拆成 n 段較短的等待：先等第一張新卡，再等第二張，如此類推。每段短等待都服從幾何分佈（geometric distribution），所以平均值很容易求得，再用期望值的線性性質（linearity of expectation）把它們加起來。總和是 n 乘以一個調和和（harmonic sum），而題中的積分不等式正是用來告訴我們這個和有多大。

解答

等待一張新卡

我們假設每包內有 n 張卡的其中一張，每張出現的概率均為 $\frac{1}{n}$，而且各包之間互相獨立（independent）。

假設我已經擁有 r 張不同的卡，其中 $0 \le r \le n-1$。打開一包得到新卡的概率是

$$p = \frac{n-r}{n} > 0,$$

得到舊卡的概率則是 $q = 1 - p = \frac{r}{n}$。設 X 為我直至第一次得到新卡為止（包括該包在內）所打開的包數。由於各包互相獨立，$X = k$ 表示先有 $k-1$ 張舊卡，然後是一張新

卡，所以

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

因此

$$\mathrm{E}(X)=\sum_{k=1}^{\infty}kpq^{k-1}=p\,(1-q)^{-2}=\frac{p}{p^2}=\frac{1}{p},$$

其中用到當 $|q|<1$ 時 $\sum_{k\geq 1}kq^{k-1}=(1-q)^{-2}$（把幾何級數（geometric series）求導，或者用二項式定理展開 $(1-q)^{-2}$）。這裏 $0\leq q<1$，因為 $r\leq n-1$。所以

$$\mathrm{E}(X)=\frac{n}{n-r},$$

這正是所要證明的結果。

（一個更簡潔的做法：第一包要麼給出新卡，我們就此停下；要麼以概率 q 令我們回到起點，所以 $\mathrm{E}(X)=1+q\,\mathrm{E}(X)$，由此得 $\mathrm{E}(X)=1/(1-q)=1/p$。這個做法假設了 $\mathrm{E}(X)$ 是有限的，而上面的級數已確認了這一點。）

積分不等式

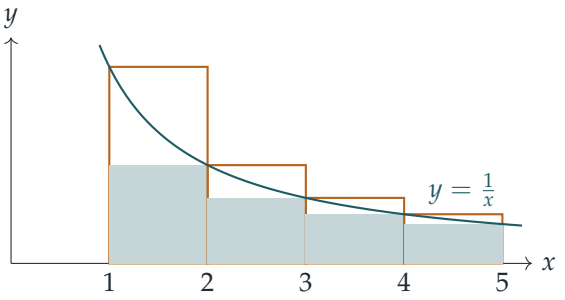
設 $r\geq 1$。當 $r\leq x\leq r+1$ 時，我們有

$$\frac{1}{r+1}\leq\frac{1}{x}\leq\frac{1}{r},$$

因為 $1/x$ 在 $x>0$ 時是遞減的。在長度為 1 的區間上積分，這些不等式仍然成立，所以

$$\frac{1}{r+1}\leq\int_r^{r+1}\frac{1}{x}\,\mathrm{d}x\leq\frac{1}{r}.$$

在圖中，曲線 $y=1/x$ 在 $x=r$ 與 $x=r+1$ 之間的面積，介乎高度為 $\frac{1}{r+1}$ 的矮長方形與高度為 $\frac{1}{r}$ 的高長方形之間。



（為清楚起見，縱軸的比例已經拉長。）塗上陰影的長方形位於曲線之下；只畫出外框的長方形則高於曲線。

現在把 $r = 1, 2, \ldots, n-1$ 的不等式加起來，其中 $n \geq 2$。各個積分會連接成一個：

$$\sum_{r=1}^{n-1} \frac{1}{r+1} \leq \int_1^n \frac{1}{x}\,\mathrm{d}x \leq \sum_{r=1}^{n-1} \frac{1}{r}.$$

中間的一項是 $\ln n - \ln 1 = \ln n$，而在左邊的和式中，我們把 $r+1$ 改記為 r，它由 2 取到 n。所以

$$\sum_{r=2}^{n} \frac{1}{r} \leq \ln n \leq \sum_{r=1}^{n-1} \frac{1}{r},$$

這正是所要證明的結果。

集齊全套

設 X_r 為我恰好擁有 r 張不同的卡期間所打開的包數，以給出第 $r+1$ 張卡的那一包作結。集齊全套所需的總包數是

$$T = X_0 + X_1 + \cdots + X_{n-1}.$$

? **想一想** 各個 X_r 取決於之前發生過的事。我們仍然可以直接把它們的期望值相加嗎？

可以。期望值的線性性質 $\mathrm{E}(A+B) = \mathrm{E}(A) + \mathrm{E}(B)$ 完全不需要獨立性。（事實上，無論之前的經過如何，X_r 都服從 $p = \frac{n-r}{n}$ 的幾何分佈，所以第一部分的結果適用於每一個 X_r。）因此

$$\mathrm{E}(T) = \sum_{r=0}^{n-1} \frac{n}{n-r} = n\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{1}\right) = n\sum_{k=1}^{n} \frac{1}{k},$$

其中記 $k = n - r$。

記 $H_n = \sum_{k=1}^{n} \frac{1}{k}$。不等式的兩邊分別給出

$$H_n = 1 + \sum_{k=2}^{n} \frac{1}{k} \leq 1 + \ln n, \qquad H_n = \sum_{k=1}^{n-1} \frac{1}{k} + \frac{1}{n} \geq \ln n + \frac{1}{n}.$$

所以

$$n\ln n + 1 \leq \mathrm{E}(T) \leq n\ln n + n.$$

上下界之間的差距小於 n，當 n 很大時，這與 $n\ln n$ 相比微不足道（兩者之比 $1/\ln n \to 0$）。所以當 n 很大時，我預計大約要吃

$$n\ln n \text{ 包。}$$

例如，100 位數學家就意味着大約 460 包 Cruncho。

驗算

- $r = 0$：第一包必定是新卡，而 $n/(n-0) = 1$。$r = n-1$：只欠某一張指定的卡，它出現的概率是 $\frac{1}{n}$，所以我們預計需要 n 包，與 $n/1$ 吻合。
- $n = 2$：$\mathrm{E}(T) = 2(1 + \frac{1}{2}) = 3$，而上下界分別給出 $2\ln 2 + 1 \approx 2.39$ 和 $2\ln 2 + 2 \approx 3.39$。
- 以 $n = 20$ 模擬 10^5 次完整的收集過程，平均需要 71.9 包，而 $20H_{20} \approx 71.95$；上下界是 60.9 和 79.9。

常見錯誤

- 由 $r = 0$ 開始把積分不等式相加：界 $\frac{1}{r}$ 在該處不成立，而且 $\int_0^1 \frac{1}{x}\,\mathrm{d}x$ 發散（diverge）。求和必須由 $r = 1$ 開始。
- 以為各個 X_r 必須互相獨立，才可以把期望值相加。線性性質並不需要獨立性。
- 答 $\ln n$ 或 n：包數的期望值的增長速度與 $n\ln n$ 相若，比 n 快。

STEP 之外

差 $H_n - \ln n$ 遞減至一個極限，即歐拉－馬斯刻若尼常數（Euler–Mascheroni constant）$\gamma \approx 0.5772$，所以更精確的答案是 $\mathrm{E}(T) \approx n(\ln n + \gamma)$；當 $n = 100$ 時，這給出大約 519 包，而不是 460 包。像這樣把和與積分作比較，就是級數的積分審斂法（integral test），在大學第一門分析課程的初段就會遇到。優惠券收集問題本身亦會在演算法分析中再次出現，也會在研究要隨機洗牌多少次才能把一副牌洗勻時出現。Grimmett 與 Welsh 合著的 *Probability: An Introduction*（概率入門教科書）是一個好去處，可以看到幾何等待時間與期望值的線性性質如何一併運用。

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。