詳細解答

STEP II 2014，第 13 題

概率與統計 • 離散隨機變量

> **STEP II 2014，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/14-s2-q13

關鍵思路

這其實是喬裝了的生日問題（birthday problem）。事件 $X \geq k$ 不過是說首 $k-1$ 個整數互不相同，而 $X = r$ 則是說首 $r-1$ 個互不相同，而第 r 個與其中一個相同。兩個概率（probability）都是一些簡單因子的乘積。至於那兩條奇特的恆等式，則來自以兩種方法計算同一個量：第 (i) 部分是總概率等於 1，第 (iv) 部分是 $\mathrm{E}(X)$ 的兩條公式相等。

解答

全文中記

$$Q_m = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{m-1}{n}\right) \qquad (m \geq 1),$$

其中 $Q_1 = 1$（空積(empty product)）。由於每個整數都與之前的整數獨立(independent)，並在 $\{1, \ldots, n\}$ 上均勻分佈（uniform），Q_m 就是 $I_1, \ldots, I_m$ 互不相同的概率：已知首 j 個互不相同，下一個避開它們全部的概率是 $1 - \frac{j}{n}$。

$\mathrm{P}(X = 4)$ 的值

我們需要 I_1, I_2, I_3 互不相同，然後 I_4 等於這三個之中的一個：

$$\mathrm{P}(X = 4) = 1 \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdot \frac{3}{n} = \frac{3(n-1)(n-2)}{n^3}.$$

第 (i) 部分

同樣地，$X = r$ 表示 $I_1, \ldots, I_{r-1}$ 互不相同（概率為 Q_{r-1}），而 I_r 是這 $r-1$ 個值之一（概率為 $\frac{r-1}{n}$，且獨立）。所以

$$\mathrm{P}(X = r) = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{r-2}{n}\right)\frac{r-1}{n} = Q_{r-1}\,\frac{r-1}{n}, \qquad 2 \le r \le n+1.$$

想一想 為甚麼這些概率加起來等於 1？

X 必定存在，並且介乎 2 與 $n+1$ 之間：第一個整數不可能是重複，所以 $X \ge 2$；而 $I_1, \ldots, I_{n+1}$ 共有 $n+1$ 個整數，卻只取 n 個值，所以由抽屜原理（pigeonhole principle），到第 $n+1$ 個時必有某個值出現了兩次，故 $X \le n+1$。（數列永不重複的機會是零，所以沒有遺漏任何情況。）因此

$$\sum_{r=2}^{n+1} \mathrm{P}(X = r) = 1.$$

把 $r = 2, 3, 4, \ldots$ 的項逐一寫出，得

$$\frac{1}{n} + \left(1 - \frac{1}{n}\right)\frac{2}{n} + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\frac{3}{n} + \cdots = 1,$$

這正是所求的恆等式。級數（series）在 $r = n+1$ 終止，之後的任何項都是零，因為它們含有因子 $1 - \frac{n}{n}$。證畢。令 $m = r - 1$，上式可寫成

$$\sum_{m=1}^{n} \frac{m}{n}\,Q_m = 1. \qquad (*)$$

第 (ii) 部分

由期望值（expectation）的定義，

$$\mathrm{E}(X) = \sum_{r=2}^{n+1} r\,\frac{r-1}{n}\left(1 - \frac{1}{n}\right)\cdots\left(1 - \frac{r-2}{n}\right) = \sum_{m=1}^{n} \frac{m(m+1)}{n}\,Q_m.$$

第 (iii) 部分

$X \ge k$ 恰好在 $I_1, \ldots, I_{k-1}$ 之中沒有重複時發生，即這 $k-1$ 個整數互不相同。所以

$$\mathrm{P}(X \ge k) = Q_{k-1} = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{k-2}{n}\right), \qquad 2 \le k \le n+2,$$

而 $\mathrm{P}(X \ge 1) = \mathrm{P}(X \ge 2) = 1$。（當 $k = n+2$ 時，乘積含有 $1 - \frac{n}{n} = 0$，理應如此。）

第 (iv) 部分

記 $p_r = \mathrm{P}(Y = r)$。由於 $r = \sum_{k=1}^{r} 1$，所以

$$\mathrm{E}(Y) = \sum_{r=1}^{N} r\,p_r = \sum_{r=1}^{N}\sum_{k=1}^{r} p_r = \sum_{k=1}^{N}\sum_{r=k}^{N} p_r = \sum_{k=1}^{N} \mathrm{P}(Y \geq k).$$

由於是有限和，交換求和次序並無問題：兩個二重和都遍歷所有滿足 $1 \leq k \leq r \leq N$ 的數對 (k, r)。證畢。

想一想 我們應該比較 $\mathrm{E}(X)$ 的哪兩個表達式？

把這結果應用於 $Y = X$，它在 $\{1, \ldots, N\}$ 中取值，其中 $N = n+1$（取值 1 的概率只是 0 而已）。利用第 (iii) 部分，

$$\mathrm{E}(X) = \sum_{k=1}^{n+1} \mathrm{P}(X \geq k) = 1 + \sum_{k=2}^{n+1} Q_{k-1} = 1 + \sum_{m=1}^{n} Q_m.$$

現在令這式等於第 (ii) 部分的結果，把 $\frac{m(m+1)}{n} = \frac{m^2}{n} + \frac{m}{n}$ 拆開，並利用 $(*)$：

$$\sum_{m=1}^{n} \frac{m(m+1)}{n} Q_m = \sum_{m=1}^{n} \frac{m^2}{n} Q_m + \underbrace{\sum_{m=1}^{n} \frac{m}{n} Q_m}_{=1 \text{ 由}(*)}.$$

所以 $1 + \sum_{m=1}^{n} Q_m = \sum_{m=1}^{n} \frac{m^2}{n} Q_m + 1$，即

$$\sum_{m=1}^{n} Q_m \left(1 - \frac{m^2}{n}\right) = 0.$$

$m = 1, 2, 3$ 的項分別是 $\left(1 - \frac{1^2}{n}\right)$、$\left(1 - \frac{1}{n}\right)\left(1 - \frac{2^2}{n}\right)$ 和 $\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\left(1 - \frac{3^2}{n}\right)$，所以這正是所求的恆等式。與第 (i) 部分一樣，$m = n$ 之後的每一項都含有因子 $1 - \frac{n}{n} = 0$。證畢。

驗算

- 當 $n = 1$ 時，$X = 2$ 必然成立。第 (i) 部分給出 $\frac{1}{1} = 1$，而最後的和只有一項 $1 - 1 = 0$。
- 當 $n = 4$ 時，最後的和是 $\frac{3}{4} + \frac{3}{4} \cdot 0 + \frac{3}{4} \cdot \frac{1}{2} \cdot \left(-\frac{5}{4}\right) + \frac{3}{4} \cdot \frac{1}{2} \cdot \frac{1}{4} \cdot (-3) = \frac{24-15-9}{32} = 0$，而兩條公式都給出 $\mathrm{E}(X) = \frac{103}{32}$。
- 當 $n = 365$（生日）時，$\mathrm{E}(X) = 1 + \sum Q_m \approx 24.62$。以電腦模擬 10^5 個數列，得平均值 24.61；而對 $n \leq 5$ 作完全窮舉（exact enumeration），亦證實了上述每一條公式。

常見錯誤

- 寫 $\mathrm{P}(X = r)$ 時用了 $r-1$ 個形如 $1 - \frac{j}{n}$ 的因子，而不是 $r-2$ 個。只有第二至第 $(r-1)$ 個整數須避開之前的值。
- 在 $\sum \mathrm{P}(X \geq k)$ 中漏掉 $k = 1$ 的項。這一項貢獻 1，正好與 $(*)$ 互相抵消。

STEP 之外

尾和公式（tail-sum formula）$\mathrm{E}(Y) = \sum_k \mathrm{P}(Y \geq k)$ 對任何取非負整數值的隨機變量（random variable）都成立，而它的連續版本 $\mathrm{E}(T) = \int_0^\infty \mathrm{P}(T > t)\,\mathrm{d}t$ 是大學一年級概率論的標準工具。當 n 很大時，$\mathrm{E}(X) \approx \sqrt{\pi n/2}$，這解釋了為何雜湊函數（hash function）的碰撞（collision）只需約 $\sqrt{n}$ 次嘗試便會出現（即「生日攻擊」（birthday attack））。Grimmett 與 Welsh 的 *Probability: An Introduction*（《概率導論》）涵蓋了這兩個概念。

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。