詳細解答

STEP I 1994，第 13 題

概率與統計 • 指示變量

> **STEP I 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-s1-q13

關鍵思路

把一個計數寫成若干指示變量（indicator variable，每個取值 1 或 0）之和，可以把難算的期望值（expectation）變成易算的，因為無論各部分是否獨立（independent），期望值都是線性的。這裏各次抽取肯定不獨立，但對稱性告訴我們每個 X_i 的期望值都相同。然後我們用第二種方法計算 $\mathrm{E}(X)$，即直接由 X 的分佈出發；令兩個答案相等，幾乎不費工夫就得到一條二項式係數（binomial coefficient）恆等式。

解答

全文設 $1 \le n \le M$，並約定當 $b > a$ 或 $b < 0$ 時 $\binom{a}{b} = 0$。

第 (i) 部分

若第 i 個取出的代幣是紅色，X_i 為 1，否則為 0。因此在和 $X_1 + X_2 + \cdots + X_n$ 之中，每個取出的紅色代幣恰好貢獻 1，每個非紅色代幣貢獻 0。所以這個和數出的正是取出的 n 個代幣中紅色代幣的數目，即 X。

第 (ii) 部分

由於 X_i 只取 0 和 1 兩個值，

$$\mathrm{E}(X_i) = 0 \cdot \mathrm{P}(X_i = 0) + 1 \cdot \mathrm{P}(X_i = 1) = \mathrm{P}(\text{第 } i \text{ 個取出的代幣是紅色}).$$

? **想一想** 第 i 個代幣是紅色的機會，難道不取決於之前抽到甚麼嗎？

就條件概率而言，是的；就無條件概率而言，不是。試想像把代幣編號，然後逐一全部抽出：M 個代幣的每一種排列次序都同樣可能出現。因此，位置 i 上的代幣同樣可能是 M 個代幣中的任何一個，正如位置 1 上的代幣一樣。抽了 n 次便停下來，並不會改變位置 $i \le n$ 上發生的事。由於 M 個代幣中有 m 個是紅色，

$$\mathrm{E}(X_i) = \mathrm{P}(X_i = 1) = \frac{m}{M} \qquad \text{對每個} i = 1, \ldots, n.$$

（對於 $i = 2$，你可以用較長的方法驗證：$\frac{m}{M} \cdot \frac{m-1}{M-1} + \frac{M-m}{M} \cdot \frac{m}{M-1} = \frac{m(M-1)}{M(M-1)} = \frac{m}{M}$。）

第 (iii) 部分

由第 (i) 部分及期望值的線性性質（無需獨立性），

$$\mathrm{E}(X) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n) = n \cdot \frac{m}{M} = \frac{mn}{M},$$

正是所求。

第 (iv) 部分

取出的 n 個代幣所成的集合，同樣可能是 $\binom{M}{n}$ 個大小為 n 的子集中的任何一個。要恰好有 k 個紅色代幣，我們從 m 個紅色代幣中選 k 個，再從其餘 $M - m$ 個中選 $n - k$ 個，所以

$$\mathrm{P}(X = k) = \frac{\binom{m}{k}\binom{M-m}{n-k}}{\binom{M}{n}}, \qquad k = 0, 1, \ldots, n.$$

按照關於二項式係數的約定，當 $k > m$（紅色代幣不足）或 $n - k > M - m$（其他代幣不足）時，上式正確地等於 0。這就是超幾何分佈（hypergeometric distribution）。

第 (v) 部分

直接由第 (iv) 部分計算 $\mathrm{E}(X)$；$k = 0$ 的一項為零，所以

$$\mathrm{E}(X) = \sum_{k=1}^{n} k\,\mathrm{P}(X = k) = \frac{1}{\binom{M}{n}} \sum_{k=1}^{n} k \binom{m}{k}\binom{M-m}{n-k}.$$

令此式與第 (iii) 部分的結果相等，

$$\sum_{k=1}^{n} k \binom{m}{k}\binom{M-m}{n-k} = \frac{mn}{M}\binom{M}{n}.$$

最後，

$$\frac{n}{M}\binom{M}{n} = \frac{n}{M} \cdot \frac{M!}{n!\,(M-n)!} = \frac{(M-1)!}{(n-1)!\,(M-n)!} = \binom{M-1}{n-1},$$

所以右邊等於 $m\binom{M-1}{n-1}$，這就是所求的恆等式。

驗算

- 取 $M = 5$，$m = 2$，$n = 3$。第 (iv) 部分給出 $\mathrm{P}(X = 0, 1, 2) = \frac{1}{10}, \frac{6}{10}, \frac{3}{10}$，三者之和為 1，而 $\mathrm{E}(X) = \frac{6+6}{10} = \frac{6}{5} = \frac{mn}{M}$。恆等式則為 $1 \cdot 2 \cdot 3 + 2 \cdot 1 \cdot 3 = 12 = 2\binom{4}{2}$。
- 若 $n = M$（全部取出），則必有 $X = m$，而確實 $\mathrm{E}(X) = mM/M = m$。
- 我們對所有 $M \leq 6$ 的袋子，列出代幣的每一種排列次序，驗證了 $\mathrm{E}(X_i)$、$\mathrm{E}(X)$ 和 $\mathrm{P}(X = k)$；並對每個 $M < 30$ 驗證了 (v) 中的恆等式。

常見錯誤

- 聲稱 $\mathrm{E}(X) = \sum \mathrm{E}(X_i)$ 需要獨立性。其實不需要：期望值的線性性質永遠成立。
- 把各次抽取當作獨立，在 (iv) 中寫出二項分佈（binomial distribution）$\mathrm{B}(n, \frac{m}{M})$。這樣得出的平均值是對的，但概率卻是錯的，因為抽取是不放回的。

STEP 之外

指示變量法是概率論中最有用的工具之一：它可以求出隨機置換（random permutation）的不動點（fixed point）數目的期望值、擲硬幣時連串（run）數目的期望值、隨機圖（random graph）中孤立頂點數目的期望值，而且還可藉 $\mathrm{Var}(\sum X_i) = \sum \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$ 推廣至方差（variance）。以兩種方法計算同一個量來證明組合恆等式，這種手法同樣普遍適用。Blitzstein 與 Hwang 所著的 *Introduction to Probability*（概率論入門）全書貫徹使用指示變量，是觀摩這兩種想法實際運用的絕佳去處。

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。