詳細解答

STEP III 1988，第 16 題

概率與統計 • 超幾何分佈

> **STEP III 1988，第 16 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/88-s3-q16

關鍵思路

不放回抽取（drawing without replacement）令各次抽取互相依賴，逐次分析會十分繁瑣。有兩個想法可以繞過這個困難。第一，數算不計次序的選法，從而得到紅球數目的分佈。第二，利用對稱性（symmetry）：假想把整個甕以隨機次序全部抽空，則每個球出現在任何指定位置的機會均等。把 X 寫成指示變量（indicator）X_i 之和，便可求得期望值（mean）和方差（variance），完全不需要知道分佈。我們只需要各對變量之間的協方差（covariance）。

解答

恰好有 k 個紅球的概率

把各個球視為可區分的。抽出的 n 個球組成一個大小為 n 的集合，而每個這樣的集合出現的機會均等，所以共有 $\binom{M}{n}$ 個等可能的結果。要組成一個恰好含 k 個紅球的集合，須從 m 個紅球中選 k 個，並從 $M-m$ 個綠球中選 $n-k$ 個。因此

$$\mathrm{P}(\text{恰好}k\text{ 個紅球}) = \frac{\binom{m}{k}\binom{M-m}{n-k}}{\binom{M}{n}},$$

其中沿用慣常的約定：若 $b<0$ 或 $b>a$，則 $\binom{a}{b}=0$。（這已涵蓋不可能發生的情況，例如 $n-k>M-m$。）這就是超幾何（hypergeometric）分佈。

第 (i) 部分

假想一直抽下去，直至甕中再沒有球為止。M 個（可區分的）球被抽出的次序是一個均勻隨機的排列：全部 $M!$ 種次序出現的機會均等。抽了 n 次便停下來並不會改變首 n 個位置，所以我們可以用完整的排列來處理。

固定一個位置 i。某個指定的球位於位置 i 的排列共有 $(M-1)!$ 個，這個數目對每個球都相同。所以位置 i 上的球是 M 個球中任何一個的機會均等，而

$$\mathrm{P}(X_i = 1) = \frac{m\,(M-1)!}{M!} = \frac{m}{M},$$

正是所求。

想一想 第二個球的顏色，不是明明取決於第一個球是甚麼顏色嗎？

在已知第一個球的情況下，確實如此。但我們要求的是無條件（unconditional）概率。當 $i = 2$ 時，由全概率定律（law of total probability）可得

$$\mathrm{P}(X_2 = 1) = \frac{m}{M}\cdot\frac{m-1}{M-1} + \frac{M-m}{M}\cdot\frac{m}{M-1} = \frac{m(M-1)}{M(M-1)} = \frac{m}{M}.$$

對稱性論證則一次過為每個 i 完成了同樣的計算。

第 (ii) 部分

固定兩個位置 $i \neq j$（所以 $M \geq 2$）。數算這兩個位置上都是紅球的排列。位置 i 上的球有 m 個選擇，然後位置 j 有 $m-1$ 個選擇，其餘的球則有 $(M-2)!$ 種排法。因此

$$\mathrm{P}(X_i = 1 \text{ 及} X_j = 1) = \frac{m(m-1)(M-2)!}{M!} = \frac{m(m-1)}{M(M-1)},$$

正是所求。換句話說，位置 (i, j) 上的有序球對，是 $M(M-1)$ 個由相異的球組成的有序對中任何一個的機會均等。

X 的期望值和方差

X 數算首 n 個抽出的球當中紅球的數目。根據期望值的線性性質（linearity of expectation）——這並不需要獨立性（independence）——再加上 $\mathrm{E}(X_i) = \mathrm{P}(X_i = 1)$，

$$\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(X_i) = \frac{nm}{M}.$$

至於方差，可用 $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$。把和的平方展開，並利用 $X_i^2 = X_i$（因為 X_i 只取 0 或 1），

$$X^2 = \sum_{i=1}^{n} X_i^2 + \sum_{i\neq j} X_i X_j = X + \sum_{i\neq j} X_i X_j,$$

其中第二個和遍歷滿足 $i \neq j$ 的 $n(n-1)$ 個有序對 (i,j)。而 $X_iX_j = 1$ 當且僅當兩個球都是紅球，所以由第 (ii) 部分得 $\mathrm{E}(X_iX_j) = \dfrac{m(m-1)}{M(M-1)}$。因此，當 $M \geq 2$ 時，

$$\begin{aligned}\mathrm{Var}(X) &= \frac{nm}{M} + \frac{n(n-1)m(m-1)}{M(M-1)} - \frac{n^2m^2}{M^2} \\ &= \frac{nm}{M^2(M-1)}\Big[M(M-1) + (n-1)(m-1)M - nm(M-1)\Big].\end{aligned}$$

方括號內的式子展開後為

$$\begin{aligned}&M^2 - M + nmM - nM - mM + M - nmM + nm \\ &\quad = M^2 - nM - mM + nm = (M-m)(M-n).\end{aligned}$$

因此

$$\mathrm{Var}(X) = \frac{nm(M-m)(M-n)}{M^2(M-1)}.$$

（若 $M = 1$，則 $n \leq 1$，根本沒有球對，而 X 是常數，所以 $\mathrm{Var}(X) = 0$。）

設 $p = m/M$ 會令結果更具啟發性。這時

$$\mathrm{E}(X) = np, \qquad \mathrm{Var}(X) = np(1-p)\cdot\frac{M-n}{M-1}.$$

這正是二項分佈（binomial distribution）的方差乘以一個修正因子，而當 $n \geq 2$ 時，這個因子小於 1。

驗算

- $n = M$：我們把所有球都抽出，所以必定有 $X = m$。公式給出期望值 m 和方差 0，後者源於因子 $M - n$。
- $n = 1$：X 是單獨一個伯努利（Bernoulli）變量，其 $p = m/M$，而方差化簡為 $p(1-p)$。
- $M = 6$、$m = 2$、$n = 3$：公式給出 $\mathrm{Var}(X) = \frac{3\cdot2\cdot4\cdot3}{36\cdot5} = \frac{2}{5}$。把全部 6! 種次序逐一列舉，結果完全吻合；$M \leq 6$ 的每一個情況亦然。

常見錯誤

- 把各個 X_i 當作互相獨立，於是得出二項分佈的方差 $np(1-p)$。其實它們是負相關（negatively correlated）的：早段抽到紅球，會令其後抽到紅球的機會降低。
- 忘記了當球對是有序的時候，二重和 $\sum_{i\neq j}$ 共有 $n(n-1)$ 項，而不是 $\binom{n}{2}$ 項。

STEP 之外

把一個計數寫成指示變量之和，是概率論中最強大的技巧之一。即使分佈本身根本無從寫出，這個方法仍能給出期望值和方差。因子 $\frac{M-n}{M-1}$ 就是抽樣調查中所用的有限總體修正（finite population correction）。它說明了為何一個訪問 1000 人的民意調查，對一個國家所能提供的資訊，與對一個無限總體所能提供的幾乎一樣多。Blitzstein 與 Hwang 所著的 *Introduction to Probability*（概率論入門教材）全書貫徹使用指示變量，並詳細討論超幾何分佈。

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。