詳細解答

STEP III 2004，第 12 題

概率與統計 • 指示變量與錯排

> **STEP III 2004，第 12 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/04-s3-q12

關鍵思路

把一個計數寫成指示變量（indicator variable）之和，是概率論中最有用的技巧之一。指示變量的期望值（expectation）其實就是概率，所以 $\mathrm{E}(N)$ 唾手可得；方差（variance）則需要用到每一對變量之間的協方差（covariance），這些協方差很小但不是零，因為球衣並不是獨立地分配的。一旦知道 N 的均值和方差都等於 1，參數為 1 的泊松分佈（Poisson distribution）自然就是值得一試的模型。

解答

單一指示變量 C_1

球衣的 $m!$ 種分配方式每一種都等可能，所以球衣 1 分給 m 名球員中任何一人的可能性都相同。因此 $\mathrm{P}(C_1 = 1) = \frac{1}{m}$，而由於 C_1 只取 0 和 1 兩個值，

$$\mathrm{E}(C_1) = 1 \cdot \mathrm{P}(C_1 = 1) = \frac{1}{m},$$

正如所求。因為 $C_1^2 = C_1$（因 $0^2 = 0$ 及 $1^2 = 1$），我們亦有 $\mathrm{E}(C_1^2) = \frac{1}{m}$，所以

$$\mathrm{Var}(C_1) = \frac{1}{m} - \frac{1}{m^2} = \frac{m-1}{m^2}.$$

C_1 與 C_2 的協方差

這裏我們需要 $m \geq 2$。乘積 C_1C_2 恰好在球員 1 和 2 都拿到自己的球衣時等於 1，否則等於 0。已知球員 1 拿到球衣 1，餘下的 $m-1$ 件球衣便隨機分給其餘 $m-1$ 名球員，所以

球員 2 拿到球衣 2 的概率是 $\frac{1}{m-1}$。因此

$$\mathrm{E}(C_1C_2) = \mathrm{P}(C_1 = 1,\, C_2 = 1) = \frac{1}{m} \cdot \frac{1}{m-1},$$

而

$$\begin{aligned}\mathrm{Cov}(C_1, C_2) &= \mathrm{E}(C_1C_2) - \mathrm{E}(C_1)\,\mathrm{E}(C_2) = \frac{1}{m(m-1)} - \frac{1}{m^2} \\ &= \frac{m-(m-1)}{m^2(m-1)} = \frac{1}{m^2(m-1)}.\end{aligned}$$

這個值是正數：知道球員 1 拿對了球衣，就等於從球員 2 可能拿到的球衣中剔除了一件錯的，令其拿對的機會略為提高。

想一想 為甚麼協方差不是零？各球員的球衣不是分開選的嗎？

並不是。球衣是每人分配一件的，所以球員 1 拿到甚麼，會影響其他人剩下甚麼可拿。各個 C_i 是相依的，這正是計算 $\mathrm{Var}(N)$ 時必須計入協方差項的原因。

N 的均值與方差

根據對稱性，每個 C_i 的分佈都與 C_1 相同，而每一對 C_i, C_j（其中 $i \neq j$）的聯合分佈（joint distribution）都與 C_1, C_2 相同（把球員和球衣一併重新編號，並不會改變這個隨機分配）。利用期望值的線性性質（linearity of expectation），而這性質並不需要獨立性，

$$\mathrm{E}(N) = \sum_{i=1}^{m} \mathrm{E}(C_i) = m \cdot \frac{1}{m} = 1.$$

至於和的方差，我們把所有方差加起來，再加上所有有序對 $i \neq j$ 的協方差，這樣的有序對共有 $m(m-1)$ 個：

$$\begin{aligned}\mathrm{Var}(N) &= \sum_{i=1}^{m} \mathrm{Var}(C_i) + \sum_{i \neq j} \mathrm{Cov}(C_i, C_j) \\ &= m \cdot \frac{m-1}{m^2} + m(m-1) \cdot \frac{1}{m^2(m-1)} \\ &= \frac{m-1}{m} + \frac{1}{m} = 1.\end{aligned}$$

所以 $\mathrm{E}(N) = \mathrm{Var}(N) = 1$，正如所求。（若 $m = 1$，則必有 $N = 1$，於是 $\mathrm{Var}(N) = 0$；題目其實隱含假設了 $m \geq 2$，一如計算協方差時已經需要的那樣。）

用哪一種近似？

無論 m 取甚麼值，正態近似（Normal approximation）的效果都很差。N 只取整數值 $0, 1, 2, \ldots$，而由於均值為 1、標準差為 1，它的概率幾乎全部集中在 $0, 1, 2, 3$ 這三四個

值上。這個分佈在 0 處被截斷，那裏只比均值低一個標準差，所以它明顯向右偏斜；而 $\mathrm{N}(1,1)$ 分佈是對稱的，並且在 $-\frac{1}{2}$ 以下有不可忽視的概率（即使作了連續性校正（continuity correction），也約有 7%），但 N 根本不可能取到那裏的值。最關鍵的是，增大 m 也無濟於事：對每一個 m，均值和方差都維持在 1，所以分佈的形狀永遠不會展開成鐘形。

相反，泊松近似則合情合理。N 所數的，是在為數甚多的 m 個事件之中「成功」的個數，每個事件的概率 $\frac{1}{m}$ 都很小，而這些事件之間只有很弱的相依性（協方差 $\frac{1}{m^2(m-1)}$ 極小）。計數近似服從泊松分佈，正是在這種情況下發生的。此外，泊松變量的均值等於方差，而 N 恰好具有這個性質，所以 $\mathrm{Po}(1)$ 是自然的選擇。

$m = 4$ 的情況

共有 $4! = 24$ 種等可能的分配方式。把一種分配寫成球員 $1,2,3,4$ 所穿球衣的號碼序列，則沒有人穿上自己球衣的分配是

$$2143,\quad 2341,\quad 2413,\quad 3142,\quad 3412,$$
$$3421,\quad 4123,\quad 4312,\quad 4321.$$

（有系統地列舉的方法：球員 1 穿的是球衣 2、3 或 4，而每種情況下都恰好有三種方法完成餘下的分配，所以共有 $3 \times 3 = 9$ 種。）因此

$$\mathrm{P}(N = 0) = \frac{9}{24} = \frac{3}{8} = 0.375.$$

參數為 1 的泊松近似給出 $\mathrm{P}(N = 0) \approx \mathrm{e}^{-1}$。取 $\mathrm{e} \approx 2.72$，

$$\mathrm{e}^{-1} \approx \frac{100}{272} = \frac{25}{68}.$$

以真確值為基準的相對誤差（relative error）是

$$\frac{\frac{3}{8} - \frac{25}{68}}{\frac{3}{8}} = 1 - \frac{25}{68} \cdot \frac{8}{3} = 1 - \frac{200}{204} = \frac{1}{51} \approx 0.0196,$$

即大約 2%，與題目所述相符。（如果改為以近似值為基準來量度誤差，則得 $\frac{1}{50}$，恰好是 2%；無論用哪一種量法，對這麼小的 m 來說，吻合程度都好得驚人。）

驗算

- 當 $m = 2$ 時，只有兩種可能：兩人都拿對（$N = 2$）或兩人都拿錯（$N = 0$），概率各為 $\frac{1}{2}$：均值為 1，方差為 1。吻合。
- 用電腦枚舉 $m = 2, \ldots, 7$ 的所有排列（permutation），確認了 $\mathrm{E}(C_1) = \frac{1}{m}$、$\mathrm{Var}(C_1) = \frac{m-1}{m^2}$、$\mathrm{Cov}(C_1, C_2) = \frac{1}{m^2(m-1)}$ 及 $\mathrm{E}(N) = \mathrm{Var}(N) = 1$ 全部準確成立，並找到 4 件物件的錯排（derangement）共 9 個，所以 $\mathrm{P}(N = 0) = 0.375$，對比 $1/2.72 \approx 0.3676$。

- 留意 $N = m - 1$ 是不可能的：如果除一人外其他球員都拿對了球衣，最後那一人也必定拿對。這再次顯示 N 與正態分佈相去甚遠。

常見錯誤

- 把各個 C_i 當作獨立，結果得出 $\mathrm{Var}(N) = \frac{m-1}{m}$ 而不是 1。
- 只數了 $\binom{m}{2}$ 個協方差項而不是 $m(m-1)$ 個，即忘記了 $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$ 中的因子 2。
- 列舉錯排時粗心大意：有系統地數（3×3）可避免遺漏。

STEP 之外

這是經典的「配對」問題，又稱錯排問題。利用容斥原理（inclusion–exclusion）可以證明 $\mathrm{P}(N = 0) = \sum_{k=0}^{m} \frac{(-1)^k}{k!}$，這個值極快地趨向 e^{-1}，而事實上當 $m \to \infty$ 時，N 依分佈收斂（converges in distribution）於 $\mathrm{Po}(1)$。更廣泛的原理是：由許多罕見而弱相依的事件所構成的計數，近似服從泊松分佈；Chen–Stein 方法把這個原理嚴格地表述出來。Blitzstein 與 Hwang 的 *Introduction to Probability*（概率論入門教材）一書，反覆以指示變量和配對問題作為例子。

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。