詳細解答

STEP I 2012，第 13 題

概率與統計 • 以指示變量求期望值

> **STEP I 2012，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/12-s1-q13

關鍵思路

首先要把條件樣本空間（conditional sample space）弄清楚：一旦知道哪些數字被排除，這個數就只是一串五個數字，每個數字都從 $\{1,2,3,4,5\}$ 中選取，而每一串出現的可能性都相同。接着，與其求出「不同數字的個數」的整個分佈（distribution），不如把這個個數寫成指示變量（indicator variable）之和，每個可能出現的數字對應一個。期望值的線性性質（linearity of expectation）把問題化為求一個概率，而這個概率考慮其補事件（complement）便很容易求得。

解答

條件樣本空間

由 10000 至 99999 的每個整數都是一串五個數字 $d_1d_2d_3d_4d_5$，其中 $d_1 \neq 0$，而每一串都恰好出現一次。條件說明每個數字都屬於 $\{1,2,3,4,5\}$。由於 0 本來就已被排除，限制 $d_1 \neq 0$ 並沒有增加任何約束，因此符合條件的整數恰好對應於

$$5^5 = 3125$$

串由 $\{1,2,3,4,5\}$ 中數字組成的五位數字串。

原本的選取是均勻的。以某事件為條件時，該事件內的各個結果仍然等可能，所以在給定條件下，這 3125 串中每一串的概率都是 $\frac{1}{3125}$。換句話說，這五個數字互相獨立（independent），而且每個都在 $\{1,2,3,4,5\}$ 上均勻分佈（uniform）：這 3125 個結果中，每一個的概率都是五個因子 $\frac{1}{5}$ 的乘積。

指示變量

設 N 為所選的數中不同數字的個數。對 $j = 1,2,3,4,5$，定義

$$I_j = \begin{cases} 1 & \text{若數字} j \text{ 至少出現一次，} \\ 0 & \text{否則。} \end{cases}$$

每個出現了的數字對總和恰好貢獻 1，因此

$$N = I_1 + I_2 + I_3 + I_4 + I_5.$$

想一想 各個 I_j 並不互相獨立。我們仍然可以把它們的期望值相加嗎？

可以。期望值的線性性質 $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$ 對任何隨機變量（random variable）都成立，不論它們是否獨立。因此

$$\mathrm{E}(N) = \sum_{j=1}^{5} \mathrm{E}(I_j) = \sum_{j=1}^{5} \mathrm{P}(\text{數字} j \text{ 出現}).$$

數字 j 不出現，當且僅當五個獨立的數字每一個都是其餘四個值之一，而每個數字發生這情況的概率都是 $\frac{4}{5}$。因此

$$\mathrm{P}(\text{數字} j \text{ 出現}) = 1 - \left(\tfrac{4}{5}\right)^5 = 1 - \frac{1024}{3125} = \frac{2101}{3125}.$$

由對稱性，這個概率對 j 的五個值都相同，所以

$$\mathrm{E}(N) = 5 \cdot \frac{2101}{3125} = \frac{2101}{625} = 3.3616,$$

即為所求。

另一方法：完整分佈

如果你較喜歡直接計數，恰好使用 k 個不同數字的字串數目，等於 $\binom{5}{k}$（選出那些數字）乘以填滿五個位置而使所選的 k 個數字全部出現的方法數。後者可由容斥原理（inclusion–exclusion）求得。例如當 $k = 3$ 時，它等於 $3^5 - 3 \cdot 2^5 + 3 \cdot 1^5 = 150$。

k	1	2	3	4	5	總數
字串數	$5 \cdot 1$	$10 \cdot 30$	$10 \cdot 150$	$5 \cdot 240$	$1 \cdot 120$	
	5	300	1500	1200	120	3125

於是

$$\mathrm{E}(N) = \frac{1 \cdot 5 + 2 \cdot 300 + 3 \cdot 1500 + 4 \cdot 1200 + 5 \cdot 120}{3125} = \frac{10505}{3125} = 3.3616.$$

這方法行得通，但較長，出錯的地方也較多。指示變量的方法只需要求一個概率。

驗算

- 表中各數目之和為 $3125 = 5^5$，證實每一串都恰好被計算了一次。
- 答案合理：N 介乎 1 與 5 之間，而從五個值中抽取五次，預期會有相當多重複，因此一個略大於 3 的值是可信的。作為一般性的驗算，若從 m 個值中抽取 n 次，同樣的論證給出 $m\left(1-\left(1-\frac{1}{m}\right)^n\right)$，當 $n=1$ 時它等於 1，正如所料。
- 用電腦列舉由 10000 至 99999 的所有整數，保留只使用數字 1–5 的那 3125 個，再求不同數字個數的平均值，得到的恰好是 $\frac{2101}{625} = 3.3616$。

常見錯誤

- 把樣本空間當作全部 90000 個整數。條件把它縮減為 5^5 串等可能的字串。
- 以為期望值的線性性質需要獨立性。其實不需要，而這正是指示變量如此有用的原因。
- 以對五個位置各加 $\frac{1}{5}$ 的方式計算 P(數字j 出現)，結果得出 1，因為這樣會把 j 出現多於一次的字串重複計算了多次。應該使用補事件。

STEP 之外

這是經典的「佔位」(occupancy) 問題，又稱集券問題 (coupon-collector problem)：把 n 個球隨機投進 m 個盒子，N 為最終有球的盒子數目。指示變量的技巧 $\mathrm{E}(N) = m\left(1-\left(1-\frac{1}{m}\right)^n\right)$ 是大學一年級概率論中，以至分析雜湊 (hashing) 等算法時最有用的工具之一。精確的分佈牽涉第二類斯特靈數 (Stirling numbers of the second kind)，它們已在上面的容斥計數中出現過。Blitzstein 與 Hwang 的 *Introduction to Probability*（《概率導論》）全書貫徹使用指示隨機變量，是絕佳的下一步讀物。

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。