詳細解答

STEP II 2000，第 13 題

概率與統計 • 捕捉再捕捉估計

STEP II 2000，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/00-s2-q13

本題在 *Beyond STEP* 第 13.2 節點算田鼠中有完整的引導式講解：https://beyondstep.keithmcnulty.org/zh-hant/ch13/voles.html

關鍵思路

這是捉放法（capture–recapture）：第二次抽樣是從 N 隻田鼠中隨機選出 200 隻，而當中有 200 隻帶有標記，所以樣本中帶標記田鼠的數目服從超幾何分佈（hypergeometric distribution）。要找出 p_N 在何處最大，不要求導（N 的階乘（factorial）並沒有簡單意義下的導數）；應該改為透過比值 p_N / p_{N-1} 比較相鄰兩項，這時幾乎所有東西都會約去。最後的恆等式則不費吹灰之力：第三次抽樣的各個概率之和必定是 1。

解答

概率 p_N

把第二次抽樣看成從 N 隻田鼠中選出 200 隻，每個大小為 200 的子集出現的機會均等。這樣的子集共有 $\binom{N}{200}$ 個。在 N 隻田鼠中，200 隻帶有標記，$N-200$ 隻沒有。一個恰好含 11 隻帶標記田鼠的樣本，就是選出 11 隻帶標記的和 189 隻沒有標記的，所以

$$p_N = \frac{\binom{200}{11}\binom{N-200}{189}}{\binom{N}{200}}.$$

把含有 N 的兩個二項式係數（binomial coefficient）寫成階乘，

$$\begin{aligned} p_N &= \binom{200}{11} \cdot \frac{(N-200)!}{189!\,(N-389)!} \cdot \frac{200!\,(N-200)!}{N!} \\ &= k\,\frac{\left((N-200)!\right)^2}{N!\,(N-389)!}, \qquad \text{其中} k = \binom{200}{11}\frac{200!}{189!}, \end{aligned}$$

而這個常數與 N 無關。這正是所要求的形式。（這裏需要 $N \geq 389$，因為兩次抽樣合共包含 $200 + 189 = 389$ 隻不同的田鼠。）

求 p_N 的最大值

對於 $N \geq 390$，把 p_N 除以 p_{N-1}。利用 $(N-200)! = (N-200)\,(N-201)!$、$N! = N\,(N-1)!$ 和 $(N-389)! = (N-389)\,(N-390)!$，所有東西都會約去，只餘下

$$\frac{p_N}{p_{N-1}} = \frac{(N-200)^2}{N(N-389)}.$$

分子和分母都是正數，所以

$$\begin{aligned} p_N > p_{N-1} &\iff (N-200)^2 > N(N-389) \\ &\iff N^2 - 400N + 40000 > N^2 - 389N \\ &\iff 40000 > 11N \\ &\iff N < 3636.\overline{36}. \end{aligned}$$

想一想 這告訴我們數列（sequence）$p_{389}, p_{390}, \ldots$ 的形狀是怎樣的？

當 $390 \leq N \leq 3636$ 時，比值大於 1，所以數列嚴格遞增，直至 p_{3636}。當 $N \geq 3637$ 時，比值小於 1（等號成立需要 $11N = 40000$，而這沒有整數解），所以數列自此嚴格遞減。因此 p_N 有唯一的最大值，位於

$$N = 3636.$$

這就是 $\frac{200\times200}{11}$ 的整數部分。這也正是你會猜到的答案：只要認為樣本中帶標記的比例 $\frac{11}{200}$ 應該等於總體中帶標記的比例 $\frac{200}{N}$ 便可。

第三次抽樣

取總體大小為 3636，即剛才求得的估計值。第二次放回之後，帶標記的田鼠是原來的 200 隻加上新標記的 189 隻，所以有 389 隻帶有標記，$3636 - 389 = 3247$ 隻沒有標記。按照與之前完全相同的推理，一個大小為 200 的隨機樣本中恰好有 j 隻帶標記田鼠的概

率是

$$\frac{\binom{389}{j}\binom{3247}{200-j}}{\binom{3636}{200}}, \qquad j = 0, 1, \ldots, 200.$$

由於 $389 \geq 200$ 且 $3247 \geq 200$，由 0 至 200 的每一個 j 都有可能出現，而這些就是全部可能的結果。它們互斥（mutually exclusive），所以它們的概率之和是 1：

$$\sum_{j=0}^{200} \frac{\binom{389}{j}\binom{3247}{200-j}}{\binom{3636}{200}} = 1.$$

乘以 $\binom{3636}{200}$ 便得

$$\sum_{j=0}^{200} \binom{389}{j}\binom{3247}{200-j} = \binom{3636}{200},$$

這正是所要證明的。

驗算

- 用整數二項式係數準確計算 p_N：對於若干個 N 的值（例如 389、1000 和 5000），結果與階乘公式一致；而在 $389 \leq N < 20000$ 的範圍內搜尋，最大值位於 $N = 3636$。
- 在轉折點處，$p_{3636}/p_{3635} = 3436^2/(3636 \times 3246) \approx 1.0003 > 1$，但 $p_{3637}/p_{3636} < 1$。
- 最後的恆等式兩邊以整數準確計算，結果相等。留意 $389 + 3247 = 3636$，這是必然的。

常見錯誤

- 把 N 當作連續變量來求導。應該改用相鄰兩項的比值，並準確地判定不等式。
- 忘記了第二次放回之後，帶標記的田鼠是 389 隻，而不是 200 隻。
- 說「最大值位於 $N = 3636.36$」：總體大小必須是整數，而且你應該檢查相鄰的兩個整數中哪一個勝出。

STEP 之外

選取使觀測數據出現的概率最大的參數值，就是*最大似然法*（maximum likelihood），這是大學一年級統計學的核心思想。在這裏，它給出捉放法的 Lincoln–Petersen 估計量

（estimator）。最後的恆等式是范德蒙德恆等式（Vandermonde's identity），而所用的論證方法，即用兩種方法數算同一樣東西，是組合數學（combinatorics）中一種強而有力的技巧。Grimmett and Welsh 的 *Probability: An Introduction*（概率入門教材）涵蓋了超幾何分佈和這類論證方法。

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。