詳細解答

STEP I 1989，第 14 題

概率與統計 • 條件分佈

> **STEP I 1989，第 14 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/89-s1-q14

關鍵思路

一切都建基於一組概率：種子落在每一條闊一里格的地帶內的機會。種子當且僅當落在園林或農地時才會長成樹，所以樹的數目服從二項分佈（binomial distribution）。樹生長之處的距離 Y，其實就是在「落在陸地上」這個條件下的 L。它的密度函數（density）就是把 f 只保留陸地地帶的部分，再按比例放大，使總面積等於 1。最後一部分是假設檢驗（hypothesis test）：對一個二項計數而言，農地上有 23 棵樹是否多得出奇？

解答

各地帶的概率

L 的累積分佈函數（cumulative distribution function）在 $0 \le x \le 4$ 時為

$$\mathrm{F}(x) = \int_0^x \left(\tfrac{1}{2} - \tfrac{1}{8}t\right)\mathrm{d}t = \frac{x}{2} - \frac{x^2}{16},$$

所以 $\mathrm{F}(1) = \frac{7}{16}$、$\mathrm{F}(2) = \frac{3}{4}$、$\mathrm{F}(3) = \frac{15}{16}$ 及 $\mathrm{F}(4) = 1$。因此

$$\mathrm{P}(\text{園林}) = \mathrm{P}(0 \le L \le 1) = \tfrac{7}{16}, \qquad \mathrm{P}(\text{湖}) = \tfrac{3}{4} - \tfrac{7}{16} = \tfrac{5}{16},$$
$$\mathrm{P}(\text{農地}) = \mathrm{P}(2 \le L \le 3) = \tfrac{15}{16} - \tfrac{3}{4} = \tfrac{3}{16}, \qquad \mathrm{P}(\text{海}) = \tfrac{1}{16}.$$

（由於 L 是連續的，各邊界的概率為零，所以邊界屬於哪一條地帶並無影響。）一顆種子長成樹的概率為

$$p = \tfrac{7}{16} + \tfrac{3}{16} = \tfrac{5}{8}.$$

設 T 為一年內長出的樹的數目。由於 96 顆種子各自獨立地被風吹走，每顆長成樹的概率為 $\frac{5}{8}$，故 $T \sim \mathrm{B}(96, \frac{5}{8})$，而每年新長出的發燒樹數目的平均值為

$$\mathrm{E}(T) = 96 \times \tfrac{5}{8} = 60.$$

第 (i) 部分

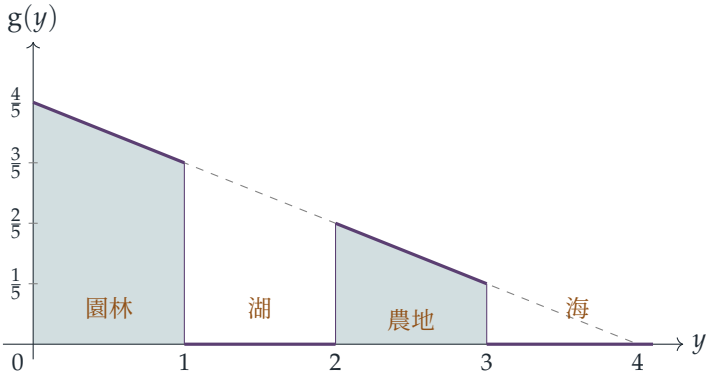
Y 的分佈，就是已知種子落在陸地上時 L 的分佈。以 $A = [0,1] \cup [2,3]$ 表示陸地。對於 $y \in A$，

$$\mathrm{P}(y \le Y \le y + \delta y) = \frac{\mathrm{P}(y \le L \le y + \delta y)}{\mathrm{P}(L \in A)} \approx \frac{\mathrm{f}(y)\,\delta y}{5/8},$$

而 Y 永遠不會取 A 以外的值。所以 Y 的密度函數為

$$\mathrm{g}(y) = \begin{cases} \frac{8}{5}(\frac{1}{2} - \frac{1}{8}y) = \dfrac{4-y}{5} & \text{當} 0 \le y \le 1 \text{ 或} 2 \le y \le 3, \\ 0 & \text{其他情況。} \end{cases}$$

驗算一下，它的面積是 $\frac{8}{5}(\frac{7}{16} + \frac{3}{16}) = 1$。圖像是同一條下降直線上的兩段：在園林上由 $\frac{4}{5}$ 降至 $\frac{3}{5}$，在湖上為零，然後在農地上由 $\frac{2}{5}$ 降至 $\frac{1}{5}$。



虛線是 $\frac{8}{5}\mathrm{f}$，畫出來以作比較；Y 的密度函數只在陸地上才與它重合。

至於平均值，$\int y(4-y)\,\mathrm{d}y = 2y^2 - \frac{1}{3}y^3$，所以

$$\begin{aligned} \mathrm{E}(Y) &= \frac{1}{5}\Big[2y^2 - \tfrac{1}{3}y^3\Big]_0^1 + \frac{1}{5}\Big[2y^2 - \tfrac{1}{3}y^3\Big]_2^3 \\ &= \frac{1}{5}\Big(\frac{5}{3}\Big) + \frac{1}{5}\Big(9 - \frac{16}{3}\Big) = \frac{1}{5}\cdot\frac{5}{3} + \frac{1}{5}\cdot\frac{11}{3} = \frac{16}{15}. \end{aligned}$$

所以新樹平均生長在城堡以南 $\frac{16}{15}$ 里格處。這個位置剛過了園林，落在湖中，但這並無不妥：平均值不一定是變量可以取到的值。

第 (ii) 部分

設 X 為一年內在農地長出的樹的數目。若有依從傳統，則放出的種子有 96 顆，每顆獨立地以概率 $\frac{3}{16}$ 落在農地，所以

$$X \sim \mathrm{B}\left(96, \tfrac{3}{16}\right), \qquad \mathrm{E}(X) = 18, \qquad \mathrm{Var}(X) = 96 \cdot \tfrac{3}{16} \cdot \tfrac{13}{16} = \tfrac{117}{8} \approx 14.6,$$

由此得標準差（standard deviation）約為 3.82。

想一想 我們究竟在檢驗甚麼？朝哪個方向檢驗？

零假設（null hypothesis）是傳統得到依從：種子有 96 顆，所以 $X \sim \mathrm{B}\left(96, \frac{3}{16}\right)$。農地上的樹比平常多，最自然的懷疑是放出的種子多於 96 顆，所以我們採用單尾檢驗（one-tailed test），問在零假設下出現 23 棵或以上的可能性有多大。

由於 $np = 18$ 和 $n(1-p) = 78$ 都夠大，我們採用正態近似（normal approximation）$X \approx \mathrm{N}(18, 14.625)$，並作連續性修正（continuity correction）：

$$\mathrm{P}(X \geq 23) \approx \mathrm{P}\left(Z \geq \frac{22.5 - 18}{3.824}\right) = \mathrm{P}(Z \geq 1.18) \approx 0.12.$$

（二項分佈的精確值是 0.121。）所以在正常情況下，農地上有 23 棵或以上的樹，大約每八年便有一年出現。這遠遠未達 5% 水平（甚至 10% 水平）的顯著性。這個觀察結果與傳統得到依從的說法相當吻合，單憑這項證據，國王的懷疑並無充分理由。

作為比較，即使在 5% 水平下，數目也要達到至少 25（精確值 $\mathrm{P}(X \geq 25) \approx 0.049$），國王才算有合理的根據。假如他同時掌握園林的樹木數目，便可以檢驗樹的總數 $T \sim \mathrm{B}\left(96, \frac{5}{8}\right)$，這對於判斷放出了多少顆種子，會是好得多的證據。

驗算

- 四個地帶的概率 $\frac{7}{16}, \frac{5}{16}, \frac{3}{16}, \frac{1}{16}$ 相加得 1，而 Y 的密度函數的總面積為 1。
- 用中點法則作數值積分得 $\mathrm{E}(Y) = 1.06667$，與 $\frac{16}{15}$ 吻合。模擬 400 000 顆種子（取 $L = 4 - 4\sqrt{1-U}$，其中 U 在 $[0,1]$ 上均勻分佈），結果有 62.5% 落在陸地上，而這些種子的平均落點距離為 1.069。
- 把二項概率逐項精確相加，得 $\mathrm{P}(X \geq 23) = 0.1213$，與正態估計值 0.120 相近。

常見錯誤

- 直接把 f 當作 Y 的密度函數。它在陸地上的面積只有 $\frac{5}{8}$，所以必須除以 $\frac{5}{8}$。
- 計算了 $\mathrm{E}(L\,;\, L \in A) = \int_A y\,\mathrm{f}(y)\,\mathrm{d}y = \frac{2}{3}$，卻忘記除以 $\mathrm{P}(L \in A)$。

- 檢驗 $\mathrm{P}(X = 23)$ 而非尾部概率 $\mathrm{P}(X \geq 23)$。X 取任何一個個別數值的可能性本來就不大。

STEP 之外

第 (i) 部分是條件分佈（conditional distribution）的一個例子：把隨機變量限制在某個事件上，再重新歸一化。密度函數形式的貝葉斯定理（Bayes' theorem），以及統計學中隨處可見的截尾分佈（truncated distribution），背後都是同一種運算。第 (ii) 部分讓你初嘗假設檢驗和 p 值，大學一年級的課程會用奈曼－皮爾遜（Neyman–Pearson）框架把這些概念嚴格化。Blitzstein 與 Hwang 的 *Introduction to Probability*（概率論入門教材）講解條件分佈時，附有大量像本題這樣帶故事情境的例題。

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。