詳細解答

STEP III 1987，第 16 題

概率與統計 • 順序統計量與泊松分佈

> **STEP III 1987，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/87-s3-q16

關鍵思路

在第 (i) 部分，事件「$A \le k$ 且 $B \ge 1-k$」直接處理起來頗為棘手，但它的補集（complement）卻是兩個簡單事件的併集：「所有 X_i 都大於 k」和「所有 X_i 都小於 $1-k$」。餘下的工作交給容斥原理（inclusion–exclusion）便可。在第 (ii) 部分，得獎人數服從二項分佈（binomial distribution），其中 n 極大而 p 極小，這正是泊松分佈（Poisson distribution）大派用場的地方。因此我們先求出恰有一位或兩位得獎者的泊松概率的最大值，然後再確保近似所帶來的誤差不會把答案推過 60%。

解答

第 (i) 部分

設 E 為事件 $A > k$，F 為事件 $B < 1-k$。那麼 $A > k$ 表示每個 X_i 都大於 k，而 $B < 1-k$ 表示每個 X_i 都小於 $1-k$。由獨立性（independence），

$$\mathrm{P}(E) = (1-k)^n, \qquad \mathrm{P}(F) = (1-k)^n, \qquad \mathrm{P}(E \cap F) = (1-2k)^n,$$

其中最後一式成立，是因為 $E \cap F$ 表示每個 X_i 都落在 $(k, 1-k)$ 之內，而這個區間的長度為 $1-2k > 0$。

我們要求的事件是 $E \cup F$ 的補集，所以由容斥原理得

$$p_n = 1 - \mathrm{P}(E \cup F) = 1 - 2(1-k)^n + (1-2k)^n.$$

由於 $0 < 1-2k < 1-k < 1$，兩個冪都趨向 0，因此

$$p_n \to 1 \quad \text{當} n \to \infty.$$

評註。無論把 k 取得多小，只要觀測值夠多，我們便幾乎可以肯定：有某個觀測值與 0 相距不超過 k，又有另一個觀測值與 1 相距不超過 k。換言之，樣本最小值收斂於 0，樣本最大值收斂於 1（均指依概率收斂），所以樣本全距（sample range）$B - A$ 趨向 1，即整個區間的長度。這正是為甚麼（舉例說）A 和 B 是未知均勻分佈（uniform distribution）兩個端點的自然估計量（estimator）。

第 (ii) 部分

設 $N = 4\,000\,000$，並設 X 為得獎人數。各讀者是否得獎互相獨立，每人得獎的概率為 p，所以 $X \sim \mathrm{B}(N, p)$。記 $q = 1 - p$。你能保住工作，當且僅當 $X = 1$ 或 $X = 2$，其概率為

$$K(p) = Npq^{N-1} + \binom{N}{2} p^2 q^{N-2}.$$

*從泊松分佈的角度看。*由於 N 極大而 p 很小，X 近似服從平均值為 $\lambda = Np$ 的泊松分佈，所以

$$K(p) \approx h(\lambda), \qquad \text{其中} h(\lambda) = \mathrm{e}^{-\lambda}\left(\lambda + \frac{\lambda^2}{2}\right).$$

求導得

$$h'(\lambda) = \mathrm{e}^{-\lambda}\left(1 + \lambda - \lambda - \frac{\lambda^2}{2}\right) = \mathrm{e}^{-\lambda}\left(1 - \frac{\lambda^2}{2}\right),$$

它在 $0 < \lambda < \sqrt{2}$ 時為正，在 $\lambda > \sqrt{2}$ 時為負。所以 h 在 $\lambda = \sqrt{2}$ 處取得最大值：

$$h(\sqrt{2}) = \mathrm{e}^{-\sqrt{2}}(\sqrt{2} + 1) \approx 0.2431 \times 2.4142 \approx 0.5869.$$

你的最佳選擇是 $p = \sqrt{2}/N \approx 3.5 \times 10^{-7}$，但即使如此，你保住工作的機會也只有約 58.7%。這就是題目所要求的推理。

> ? **想一想** 泊松分佈只是一個近似。真正的二項概率會不會悄悄越過 0.6？

不會。我們可以直接為 $K(p)$ 找出上界。由於 $q \le 1$ 及 $\binom{N}{2} \le \frac{1}{2}N^2$，

$$K(p) = q^{N-2}\left(Npq + \binom{N}{2}p^2\right) \le q^{N-2}\left(\lambda + \frac{\lambda^2}{2}\right).$$

對所有 p 都有 $1 - p \le \mathrm{e}^{-p}$（直線 $1 - p$ 是凸曲線 e^{-p} 在 $p = 0$ 處的切線），所以 $q^{N-2} \le \mathrm{e}^{-(N-2)p} = \mathrm{e}^{-\mu}$，其中 $\mu = (N-2)p$。設 $c = N/(N-2)$，則 $\lambda = c\mu$，且 $c > 1$。於是

$$K(p) \le \mathrm{e}^{-\mu}\left(c\mu + \frac{c^2\mu^2}{2}\right) \le c^2\,\mathrm{e}^{-\mu}\left(\mu + \frac{\mu^2}{2}\right) = c^2 h(\mu) \le c^2 h(\sqrt{2}).$$

這裏 $c^2 = (1 - \frac{2}{N})^{-2} < 1.000002$，所以無論 p 取甚麼值，

$$K(p) < 1.000002 \times 0.58694 < 0.587 < 0.6.$$

因此，無論你怎樣選擇 p，保住工作的機會都少於 60%，這正是所要證明的。

驗算

- 在第 (i) 部分，取 $n = 1$ 得 $p_1 = 1 - 2(1-k) + (1-2k) = 0$。這是對的：當 $k < \frac{1}{2}$ 時，單獨一個觀測值不可能既 $\le k$ 又 $\ge 1-k$。取 $n = 2$ 則得 $p_2 = 2k^2$，即兩點分別落在兩端的窄帶內（次序不拘）的概率。
- 以 $n = 10$、$k = 0.3$ 進行模擬，結果為 0.944，而公式給出 0.9436。
- 在 $\lambda = Np$ 的細密網格上計算精確的二項概率 $K(p)$，得出最大值為 0.58694，在 $\lambda \approx 1.414$ 處取得，與泊松值在小數點後五位吻合。

常見錯誤

- 寫成 $p_n = \mathrm{P}(A \le k)\,\mathrm{P}(B \ge 1-k)$。同一個樣本的最小值和最大值並不獨立。
- 直接對 p 求二項概率的最大值。這樣做是可行的，但代數運算相當繁瑣；改用 $\lambda = Np$ 來處理，問題便化為一行就能完成的微積分題。
- 做到泊松近似便停筆，沒有留意到 0.587 與 0.6 相當接近，值得補上一些論證。

STEP 之外

第 (i) 部分讓我們初嘗次序統計量（order statistics）和依概率收斂（convergence in probability）的滋味：樣本的最小值和最大值會收斂於支集（support）的兩端。第 (ii) 部分是泊松的「稀有事件定律」（law of rare events），而我們所用的上界，正是以下一般事實的一個簡單例子：當 p 很小時，二項分佈與泊松分佈在全變差（total variation）的意義下非常接近。這兩個想法在 Grimmett 與 Welsh 的 *Probability: An Introduction*，以及 Blitzstein 與 Hwang 的 *Introduction to Probability* 中都有詳細闡述。

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。