詳細解答

STEP I 1995，第 13 題

概率與統計 • 泊松稀疏化

STEP I 1995，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/95-s1-q13

關鍵思路

第 (i) 部分是泊松分佈（Poisson distribution）的稀疏化（thinning）性質：先以癌細胞的真實數目為條件，則被標記的細胞數目服從二項分佈（binomial distribution）。再對未知的真實數目求和，結果恰好化簡成一個指數級數（exponential series）。至於第 (ii) 部分，「第二個標記落在第 k 塊玻片上」這個事件，取決於頭 $k-1$ 塊玻片上合共有多少個標記（零個或一個），而第 (i) 部分連同獨立泊松變量的可加性（additivity），正好告訴我們這個總數的分佈。

解答

全文記 $q = 1-p$ 及 $\lambda = \mu p$。我們假設不同玻片上的癌細胞數目互相獨立（independent）（這是對「隨機」一詞最自然的理解），並假設科學家每次能否察覺細胞亦互相獨立，且與細胞數目獨立。

第 (i) 部分

在一塊玻片上，設 N 為癌細胞的數目，M 為她標記的數目，故 $N \sim \mathrm{Po}(\mu)$。給定 $N = n$，這 n 個細胞各自獨立地以概率 p 被發現，所以 $M \mid N = n \sim \mathrm{B}(n,p)$。對於 $m = 0,1,2,\ldots$，由全概率公式（law of total probability）（只有 $n \geq m$ 的項有貢獻），

$$\begin{aligned}\mathrm{P}(M=m) &= \sum_{n=m}^{\infty} \mathrm{e}^{-\mu}\frac{\mu^n}{n!}\binom{n}{m}p^m q^{n-m} = \mathrm{e}^{-\mu}\frac{(\mu p)^m}{m!}\sum_{n=m}^{\infty}\frac{(\mu q)^{n-m}}{(n-m)!} \\ &= \mathrm{e}^{-\mu}\frac{(\mu p)^m}{m!}\mathrm{e}^{\mu q} = \mathrm{e}^{-\mu p}\frac{(\mu p)^m}{m!},\end{aligned}$$

其中先用 $\binom{n}{m}/n! = 1/(m!\,(n-m)!)$，再以 $j = n - m$ 套用指數級數，最後用 $-\mu + \mu q = -\mu p$。因此 $M \sim \mathrm{Po}(\mu p)$，正如所求。

第 (ii) 部分

設 M_i 為第 i 塊玻片上被標記的細胞數目。由第 (i) 部分及各玻片之間的獨立性，$M_1, M_2, \ldots$ 是互相獨立的 $\mathrm{Po}(\lambda)$ 變量。設 $S = M_1 + \cdots + M_{k-1}$ 為頭 $k-1$ 塊玻片上的標記總數；獨立泊松變量之和仍服從泊松分佈，其均值為各均值之和，所以 $S \sim \mathrm{Po}((k-1)\lambda)$，而且 S 與 M_k 獨立。

想一想 要令第二個標記落在第 k 塊玻片上，究竟必須發生甚麼事？

在第 k 塊玻片之前，她最多只作了一個標記；而到第 k 塊玻片完結時，她至少已作了兩個標記。所以有兩種互斥（disjoint）的情況：



$S = 0$ 且 $M_k \geq 2$，或 $S = 1$ 且 $M_k \geq 1$

（當 $k = 1$ 時，前面沒有玻片，$S = 0$ 自動成立，以下公式依然適用。）利用 S 與 M_k 的獨立性：

$$\mathrm{P}(S = 0,\ M_k \geq 2) = \mathrm{e}^{-(k-1)\lambda}\left(1 - \mathrm{e}^{-\lambda} - \lambda \mathrm{e}^{-\lambda}\right),$$
$$\mathrm{P}(S = 1,\ M_k \geq 1) = (k-1)\lambda\, \mathrm{e}^{-(k-1)\lambda}\left(1 - \mathrm{e}^{-\lambda}\right).$$

兩式相加，

$$Q = \mathrm{e}^{-(k-1)\lambda}\left\{\left(1 + (k-1)\lambda\right)\left(1 - \mathrm{e}^{-\lambda}\right) - \lambda \mathrm{e}^{-\lambda}\right\}.$$

為了化成題目要求的形式，寫 $1 + (k-1)\lambda = (1 + k\lambda) - \lambda$：

$$\begin{aligned}\left(1 + (k-1)\lambda\right)\left(1 - \mathrm{e}^{-\lambda}\right) - \lambda \mathrm{e}^{-\lambda} &= (1 + k\lambda)\left(1 - \mathrm{e}^{-\lambda}\right) - \lambda + \lambda \mathrm{e}^{-\lambda} - \lambda \mathrm{e}^{-\lambda} \\ &= (1 + k\lambda)\left(1 - \mathrm{e}^{-\lambda}\right) - \lambda.\end{aligned}$$

代入 $\lambda = \mu p$ 即得

$$Q = \mathrm{e}^{-\mu p(k-1)}\left\{(1 + k\mu p)\left(1 - \mathrm{e}^{-\mu p}\right) - \mu p\right\},$$

這正是題目所述的結果。

驗算

▶ 當 $k = 1$ 時，公式給出 $1 - \mathrm{e}^{-\lambda} - \lambda \mathrm{e}^{-\lambda} = \mathrm{P}(M_1 \geq 2)$，理應如此。

- 對 $k = 1, 2, \ldots$ 求和，各概率之和應為 1（她最終會標記第二個細胞的概率為 1）；取若干個 λ 的值作數值求和，結果確認了這一點。
- 以 $\mu = 1.6$、$p = 0.45$（故 $\lambda = 0.72$）進行模擬，在 200 000 次試驗中，第二個標記落在第 2 塊玻片上的約佔 26.0%；公式則給出 $Q \approx 0.2591$。

常見錯誤

- 遺漏了「第 k 塊玻片之前有一個標記，第 k 塊玻片上至少有一個」這種情況，或誤把 $S = 1, M_k = 0$ 也計算在內。
- 逐塊玻片地以「第一塊有標記的玻片」來論證，這樣做會變得十分繁複；頭 $k - 1$ 塊玻片上的總數 S 才是最有效率的切入對象。
- 沒有說明在哪裏用到獨立性（各玻片之間的獨立性，以及 S 與 M_k 之間的獨立性）。

STEP 之外

稀疏化是泊松過程（Poisson process）的其中一面：若事件按泊松過程發生，而每個事件各自獨立地以概率 p 被保留，則被保留的事件仍然構成一個泊松過程，其發生率（rate）乘以 p。這樣看來，第 (ii) 部分就是「等待第二個事件發生所需時間」的離散版本，而在連續時間下，這個等待時間服從伽瑪分佈（gamma distribution）。Grimmett 與 Welsh 所著的 *Probability: An Introduction*（概率論入門教材）詳細講述了泊松過程及這些等待時間的分佈。

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。