详细解答

STEP I 1995，第 13 题

概率与统计 · 泊松稀疏化

STEP I 1995，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

step.maths.org/questions/95-s1-q13

关键思路

第 (i) 部分是泊松分布（Poisson distribution）的稀疏化（thinning）性质：先以癌细胞的真实数目为条件，则被标记的细胞数目服从二项分布（binomial distribution）。再对未知的真实数目求和，结果恰好化简成一个指数级数（exponential series）。至于第 (ii) 部分，“第二个标记落在第 k 块玻片上”这个事件，取决于头 $k-1$ 块玻片上合共有多少个标记（零个或一个），而第 (i) 部分连同独立泊松变量的可加性（additivity），正好告诉我们这个总数的分布。

解答

全文记 $q = 1-p$ 及 $\lambda = \mu p$。我们假设不同玻片上的癌细胞数目互相独立（independent）（这是对“随机”一词最自然的理解），并假设科学家每次能否察觉细胞亦互相独立，且与细胞数目独立。

第 (i) 部分

在一块玻片上，设 N 为癌细胞的数目，M 为她标记的数目，故 $N \sim \mathrm{Po}(\mu)$。给定 $N = n$，这 n 个细胞各自独立地以概率 p 被发现，所以 $M \mid N = n \sim \mathrm{B}(n, p)$。对于 $m = 0, 1, 2, \ldots$，由全概率公式（law of total probability）（只有 $n \geq m$ 的项有贡献），

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

其中先用 $\binom{n}{m}/n! = 1/(m!\,(n-m)!)$，再以 $j = n - m$ 套用指数级数，最后用 $-\mu + \mu q = -\mu p$。因此 $M \sim \mathrm{Po}(\mu p)$，正如所求。

第 (ii) 部分

设 M_i 为第 i 块玻片上被标记的细胞数目。由第 (i) 部分及各玻片之间的独立性，$M_1, M_2, \ldots$ 是互相独立的 $\mathrm{Po}(\lambda)$ 变量。设 $S = M_1 + \cdots + M_{k-1}$ 为头 $k-1$ 块玻片上的标记总数；独立泊松变量之和仍服从泊松分布，其均值为各均值之和，所以 $S \sim \mathrm{Po}((k-1)\lambda)$，而且 S 与 M_k 独立。

> **想一想** 要令第二个标记落在第 k 块玻片上，究竟必须发生甚么事?

在第 k 块玻片之前，她最多只作了一个标记；而到第 k 块玻片完结时，她至少已作了两个标记。所以有两种互斥（disjoint）的情况：



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

（当 $k = 1$ 时，前面没有玻片，$S = 0$ 自动成立，以下公式依然适用。）利用 S 与 M_k 的独立性：

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

两式相加，

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

为了化成题目要求的形式，写 $1 + (k-1)\lambda = (1 + k\lambda) - \lambda$：

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

代入 $\lambda = \mu p$ 即得

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

这正是题目所述的结果。

验算

▶ 当 $k = 1$ 时，公式给出 $1 - \mathrm{e}^{-\lambda} - \lambda \mathrm{e}^{-\lambda} = \mathrm{P}(M_1 \geq 2)$，理应如此。

- 对 $k = 1, 2, \ldots$ 求和，各概率之和应为 1（她最终会标记第二个细胞的概率为 1）；取若干个 λ 的值作数值求和，结果确认了这一点。
- 以 $\mu = 1.6$、$p = 0.45$（故 $\lambda = 0.72$）进行模拟，在 200 000 次试验中，第二个标记落在第 2 块玻片上的约占 26.0%；公式则给出 $Q \approx 0.2591$。

常见错误

- 遗漏了“第 k 块玻片之前有一个标记，第 k 块玻片上至少有一个”这种情况，或误把 $S = 1, M_k = 0$ 也计算在内。
- 逐块玻片地以“第一块有标记的玻片”来论证，这样做会变得十分繁复；头 $k-1$ 块玻片上的总数 S 才是最有效率的切入对象。
- 没有说明在哪里用到独立性（各玻片之间的独立性，以及 S 与 M_k 之间的独立性）。

STEP 之外

稀疏化是泊松过程（Poisson process）的其中一面：若事件按泊松过程发生，而每个事件各自独立地以概率 p 被保留，则被保留的事件仍然构成一个泊松过程，其发生率（rate）乘以 p。这样看来，第 (ii) 部分就是“等待第二个事件发生所需时间”的离散版本，而在连续时间下，这个等待时间服从伽玛分布（gamma distribution）。Grimmett 与 Welsh 所著的 *Probability: An Introduction*（概率论入门教材）详细讲述了泊松过程及这些等待时间的分布。

Beyond STEP 详细解答：https://beyondstep.keithmcnulty.org/zh-hans/。解答文字采用 CC0 1.0；题目本身版权属 Cambridge University Press & Assessment 所有（©），本文件不予转载。由机器从英文翻译而成；尚未经审阅。