详细解答

STEP III 1992，第 16 题

概率与统计 · 泊松稀疏化

> **STEP III 1992，第 16 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/92-s3-q16

关键思路

一旦错字的总数固定下来，每个错字都是一次独立试验（independent trial），我要么发现它，要么看漏，所以第 (i) 部分是一个二项概率（binomial probability）。第 (ii) 部分是条件概率（conditional probability），而分母，即我恰好发现 m 个错字的概率，需要我们对所有可能的看漏数目逐一相加。这个和原来是一个指数级数（exponential series），正因如此，整个式子才会化简成简洁的泊松（Poisson）形式。

解答

全篇记 $q = 1 - p$。设 N 为该期报纸的错字总数，S 为我发现的错字数目，$U = N - S$ 为我未能发现的数目。题目告诉我们

$$\mathrm{P}(N = k) = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

第 (i) 部分

已知 $N = m + n$ 时，这 $m + n$ 个错字各自以概率 p 被发现，彼此独立。因此，在 $N = m + n$ 的条件下，被发现的错字数目服从二项分布（binomial distribution）$\mathrm{B}(m + n, p)$。任意指定哪 m 个错字被发现、哪 n 个被看漏，其概率都是 $p^m q^n$，而这样的选法共有 $\binom{m+n}{m}$ 种，所以

$$\mathrm{P}(S = m \mid N = m + n) = \binom{m+n}{m} p^m (1-p)^n.$$

第 (ii) 部分

我们要求 $\mathrm{P}(U = n \mid S = m)$。事件 $\{S = m,\ U = n\}$ 与 $\{N = m+n,\ S = m\}$ 是同一事件，所以由第 (i) 部分可得

$$\begin{aligned}\mathrm{P}(S = m,\ U = n) &= \mathrm{P}(N = m+n)\,\mathrm{P}(S = m \mid N = m+n)\\ &= \frac{\mathrm{e}^{-\lambda}\lambda^{m+n}}{(m+n)!}\cdot\frac{(m+n)!}{m!\,n!}\,p^m q^n\\ &= \mathrm{e}^{-\lambda}\,\frac{(\lambda p)^m}{m!}\cdot\frac{(\lambda q)^n}{n!}.\end{aligned}$$

因子 $(m+n)!$ 刚好约去，这正是本题得以顺利解出的关键。

> **想一想** 我们不知道总共有多少个错字，那么怎样求 $\mathrm{P}(S = m)$ 呢?

如果我恰好发现 m 个，看漏的数目可以是 $0, 1, 2, \ldots$ 中的任何一个，而这些情况互斥（mutually exclusive）。因此我们把联合概率（joint probability）对 U 的所有取值 j 求和：

$$\begin{aligned}\mathrm{P}(S = m) &= \sum_{j=0}^{\infty}\mathrm{e}^{-\lambda}\,\frac{(\lambda p)^m}{m!}\cdot\frac{(\lambda q)^j}{j!} = \mathrm{e}^{-\lambda}\,\frac{(\lambda p)^m}{m!}\,\mathrm{e}^{\lambda q}\\ &= \frac{(\lambda p)^m}{m!}\,\mathrm{e}^{-\lambda p},\end{aligned}$$

这里用到 $\sum_{j\ge 0} x^j/j! = \mathrm{e}^x$ 及 $-\lambda + \lambda q = -\lambda p$。（由此可见，我发现的错字数目本身也服从泊松分布（Poisson distribution），平均值（mean）为 λp。）这个概率是正数，所以可以用它作除数。于是

$$\begin{aligned}\mathrm{P}(U = n \mid S = m) &= \frac{\mathrm{P}(S = m,\ U = n)}{\mathrm{P}(S = m)} = \frac{\mathrm{e}^{-\lambda}\,\dfrac{(\lambda p)^m}{m!}\cdot\dfrac{(\lambda q)^n}{n!}}{\mathrm{e}^{-\lambda p}\,\dfrac{(\lambda p)^m}{m!}}\\ &= \frac{(\lambda q)^n}{n!}\,\mathrm{e}^{-\lambda+\lambda p} = \frac{(1-p)^n\lambda^n}{n!}\,\mathrm{e}^{-(1-p)\lambda},\end{aligned}$$

这正是所要证明的结果。

答案完全不依赖 m：知道我发现了多少个错字，并不能告诉我看漏了多少个。无论 m 是多少，看漏的数目都服从平均值为 $(1-p)\lambda$ 的泊松分布。

验算

- 把最后的答案对 $n = 0, 1, 2, \ldots$ 求和，得 $\mathrm{e}^{(1-p)\lambda}\mathrm{e}^{-(1-p)\lambda} = 1$，条件分布理应如此。
- 当 $p \to 1$ 时，我会发现所有错字，而答案在 $n = 0$ 时趋向 1，在 $n \ge 1$ 时趋向 0。吻合。
- 取 $\lambda = 3$、$p = 0.6$、$m = 2$、$n = 1$，公式给出 $1.2\,\mathrm{e}^{-1.2} \approx 0.3614$。模拟 400 000 份报纸，只保留恰好发现两个错字的那些，所得结果为 0.360。

常见错误

- 在第 (i) 部分漏掉二项式系数（binomial coefficient）$\binom{m+n}{m}$：被发现的 m 个错字可以是 $m+n$ 个之中的任意 m 个。
- 把 $\mathrm{P}(S=m)$ 当成 $\mathrm{P}(N=m)$，或当成第 (i) 部分的答案，而没有对所有可能的看漏数目求和。

STEP 之外

这就是泊松分布的稀疏化（thinning，又称分裂 splitting）性质：如果有 Poisson(λ) 个物件，每个各以概率 p 被保留，那么被保留的数目与被舍弃的数目是独立的泊松变量，平均值分别为 λp 和 $\lambda(1-p)$。我们求得的联合概率恰好分解成这两个部分。稀疏化是泊松过程（Poisson process）理论中的重要工具，它让你把一串随机事件分拆成若干串互相独立的子事件流。Grimmett 与 Welsh 合著的 *Probability: An Introduction*（概率论入门教材）有讲述这个性质，也介绍了只需一行便能证明它的母函数（generating function）方法。

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。