详细解答

STEP II 2003，第 13 题

概率与统计 · 截尾泊松分布

> **STEP II 2003，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/03-s2-q13

关键思路

这是一个剔除了 0 这个值的泊松分布（Poisson distribution），余下的概率按比例放大，使它们的总和仍然是 1。因此，我们需要的每一个和，都是缺了 $k=0$ 一项的泊松和；而在计算平均值（mean）和方差（variance）时，$k=0$ 一项本来就没有任何贡献。所以我们可以借用熟悉的泊松分布结果 $\sum k\,p_k=\lambda$ 和 $\sum k^2 p_k=\lambda+\lambda^2$，然后乘以 A 就可以了。

解答

常数 A

记 $p_k=\dfrac{\lambda^k \mathrm{e}^{-\lambda}}{k!}$ 为普通 $\mathrm{Po}(\lambda)$ 的概率。由于 $\sum_{k\ge 0}\lambda^k/k!=\mathrm{e}^{\lambda}$，

$$\sum_{k=1}^{\infty} p_k = \mathrm{e}^{-\lambda}(\mathrm{e}^{\lambda}-1) = 1-\mathrm{e}^{-\lambda}.$$

概率 $\mathrm{P}(X=k)=Ap_k$ 的总和必须是 1，所以 $A(1-\mathrm{e}^{-\lambda})=1$，即

$$A=(1-\mathrm{e}^{-\lambda})^{-1},$$

这正是题目要求的结果。留意当 $\lambda>0$ 时 $0<1-\mathrm{e}^{-\lambda}<1$，所以 $A>1$：剩下来的概率 p_0 已经分摊给 $k\ge 1$ 的各个值。

平均值

$\sum k\,p_k$ 中 $k=0$ 的一项是零，所以

$$\sum_{k=1}^{\infty} k\,p_k = \sum_{k=0}^{\infty} k\,p_k = \lambda,$$

即 $\mathrm{Po}(\lambda)$ 的平均值。（直接计算：$\sum_{k\geq 1} k\lambda^k \mathrm{e}^{-\lambda}/k! = \lambda \sum_{k\geq 1} \lambda^{k-1}\mathrm{e}^{-\lambda}/(k-1)! = \lambda$。）因此

$$\mu = \mathrm{E}(X) = A\lambda = \frac{\lambda}{1-\mathrm{e}^{-\lambda}}.$$

方差

同样地，$\sum k^2 p_k$ 中 $k=0$ 的一项是零，而对于 $\mathrm{Po}(\lambda)$，我们有 $\mathrm{E}(Y^2) = \mathrm{Var}(Y) + (\mathrm{E}\,Y)^2 = \lambda + \lambda^2$。如果你不想直接引用这个结果，可以写 $k^2 = k(k-1) + k$：

$$\sum_{k\geq 2} k(k-1)\frac{\lambda^k \mathrm{e}^{-\lambda}}{k!} = \lambda^2 \sum_{k\geq 2} \frac{\lambda^{k-2}\mathrm{e}^{-\lambda}}{(k-2)!} = \lambda^2, \qquad \text{所以} \qquad \sum_{k\geq 1} k^2 p_k = \lambda^2 + \lambda.$$

于是

$$\begin{aligned} \mathrm{E}(X^2) &= A\lambda(1+\lambda) = \mu(1+\lambda), \\ \mathrm{Var}(X) &= \mathrm{E}(X^2) - \mu^2 = \mu(1+\lambda) - \mu^2 = \mu(1-\mu+\lambda), \end{aligned}$$

这正是题目要求的结果。

不等式

> ? **想一想** 严格不等式 $\mu < 1+\lambda$ 可以从哪里来?

从方差而来。X 取 $1, 2, 3, \ldots$ 中每一个值的概率都是正数，所以它不是常数，而一个非常数的随机变量（random variable）的方差必定严格大于零（因为 $\mathrm{Var}(X) = \mathrm{E}((X-\mu)^2)$，而 $(X-\mu)^2 > 0$ 的概率是正数）。所以

$$\mu(1-\mu+\lambda) > 0.$$

由于 $\mu > 0$(它是一个取正值的随机变量的平均值),我们可以两边除以 μ,得到 $1-\mu+\lambda > 0$，即 $\mu < 1+\lambda$。

至于下界，$\mu = A\lambda$，其中 $A > 1$ 而 $\lambda > 0$，所以 $\mu > \lambda$。两者合起来，

$$\lambda < \mu < 1+\lambda.$$

$\lambda = 100$ 时的正态近似

e^{-100} 小得惊人（大约是 4×10^{-44}），所以在任何合理的准确度下

$$\mu = \frac{100}{1-\mathrm{e}^{-100}} \approx 100, \qquad \mathrm{Var}(X) = \mu(101-\mu) \approx 100 \times 1 = 100.$$

所以我们用 $Y \sim \mathrm{N}(100, 10^2)$ 来近似 X，这就是正态近似（normal approximation）。由于 X 只取整数值，我们要使用连续性校正（continuity correction）：事件 $X = 100$ 对应

$99.5 < Y < 100.5$。以 $Z = (Y - 100)/10$ 标准化，

$$\begin{aligned} \mathrm{P}(X = 100) &\approx \mathrm{P}(-0.05 < Z < 0.05) \\ &= 2\Phi(0.05) - 1 \\ &\approx 2(0.51994) - 1 = 0.0399, \end{aligned}$$

所以 $\mathrm{P}(X = 100) \approx 0.04$（准确至小数点后 2 位）。

验算

- 当 $\lambda \to 0$ 时，$\mu = \lambda/(1 - \mathrm{e}^{-\lambda}) \to 1$，而方差 $\to 0$：几乎所有概率都集中在 $k = 1$，这是理所当然的。这也与不等式 $\lambda < \mu < 1 + \lambda$ 把 μ 夹向 1 的情况吻合。
- 对许多随机选取的 λ 以数值方法求级数的和，所得结果与 A、μ 和 $\mu(1 - \mu + \lambda)$ 一致，准确至机器精度。
- 当 $\lambda = 100$ 时，准确的概率是 $\dfrac{100^{100}\mathrm{e}^{-100}}{100!\,(1 - \mathrm{e}^{-100})} \approx 0.03986$，而正态近似给出 0.03988。两者四舍五入后都是 0.04。

常见错误

- 忘记连续性校正，算出 $\mathrm{P}(Y = 100) = 0$，或者未加解释就使用密度函数的值。
- 只证明了 $\mu \leq 1 + \lambda$：要得到严格不等式，必须指出 X 不是常数。
- 除以 μ 时没有说明 $\mu > 0$。

STEP 之外

X 服从零截尾泊松（zero-truncated Poisson）分布，统计学家会在无法观察到计数为零的情况下使用它（例如经过检查站的汽车的乘载人数：没有一辆车的乘载人数是零）。正态近似之所以行得通，是因为一个 $\mathrm{Po}(100)$ 变量是 100 个独立的 $\mathrm{Po}(1)$ 变量之和，所以中心极限定理（central limit theorem）适用；在这个尺度下，截尾的影响根本看不出来。Grimmett and Welsh, *Probability: An Introduction*（概率论入门教材）讲解了母函数（generating functions）和中心极限定理，它们正是这道题前后两部分背后的理论。

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。