详细解答

STEP I 2016，第 13 题

概率与统计 · 指数分布的顺序统计量

> **STEP I 2016，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/16-s1-q13

关键思路

对于第一封和第二封邮件的到达，要问的是它们何时尚未发生。第一封电子邮件在时间 t 之后才到达，当且仅当全部 n 封都在 t 之后到达；第二封在 t 之后才到达，当且仅当到那时至多只有一封已到达。利用独立性（independence），这两个事件都很容易计算，再把生存概率（survival probability）求导，便得到概率密度函数（density）。

解答

设 $X_1, \ldots, X_n$ 为各封邮件的到达时间。它们互相独立，而且每一个都满足

$$\mathrm{P}(X_i > t) = \int_t^\infty \lambda \mathrm{e}^{-\lambda s}\,\mathrm{d}s = \mathrm{e}^{-\lambda t}, \qquad \mathrm{P}(X_i \le t) = 1 - \mathrm{e}^{-\lambda t}$$

其中 $t \ge 0$。我们会用到标准积分 $\int_0^\infty t\,k\mathrm{e}^{-kt}\,\mathrm{d}t = \frac{1}{k}$（$k > 0$），这正是参数为 k 的指数分布（exponential distribution）的平均值。（用分部积分法（integration by parts）：$[-t\mathrm{e}^{-kt}]_0^\infty + \int_0^\infty \mathrm{e}^{-kt}\,\mathrm{d}t = 0 + \frac{1}{k}$。）

第 (i) 部分

第一封邮件在 t 之后才到达，当且仅当每一封邮件都在 t 之后才到达。由独立性，

$$\mathrm{P}(T > t) = \prod_{i=1}^{n} \mathrm{P}(X_i > t) = \left(\mathrm{e}^{-\lambda t}\right)^n = \mathrm{e}^{-n\lambda t}.$$

所以累积分布函数（cumulative distribution function）为 $F(t) = 1 - \mathrm{e}^{-n\lambda t}$，其中 $t \ge 0$。求导便得到密度函数

$$f(t) = F'(t) = n\lambda\,\mathrm{e}^{-n\lambda t} \qquad (t \ge 0),$$

正如所求。这是参数为 $n\lambda$ 的指数密度函数，所以取 $k = n\lambda$ 代入标准积分，得

$$\mathrm{E}(T) = \frac{1}{n\lambda}.$$

第 (ii) 部分

设 S 为第二封邮件的到达时间（因此需要 $n \geq 2$）。第二封邮件在 t 之后才到达，当且仅当到时间 t 为止至多一封邮件已到达：要么一封也没有，要么恰好一封。

? **想一想** 到时间 t 为止恰好有一封邮件已到达，有多少种可能情况?

n 封邮件中任何一封都可以是那封提早到达的。那一封已到达（概率为 $1 - \mathrm{e}^{-\lambda t}$），而其余 $n-1$ 封尚未到达（每封概率为 $\mathrm{e}^{-\lambda t}$）。由独立性，并且由于这 n 种情况互斥（mutually exclusive），

$$\mathrm{P}(S > t) = \mathrm{e}^{-n\lambda t} + n\left(1 - \mathrm{e}^{-\lambda t}\right)\mathrm{e}^{-(n-1)\lambda t}.$$

展开第二项得 $n\mathrm{e}^{-(n-1)\lambda t} - n\mathrm{e}^{-n\lambda t}$，所以

$$\mathrm{P}(S > t) = n\,\mathrm{e}^{-(n-1)\lambda t} - (n-1)\,\mathrm{e}^{-n\lambda t}.$$

简单验证一下：当 $t = 0$ 时它等于 1，而当 $t \to \infty$ 时它趋于 0。

S 的密度函数 g 是 $1 - \mathrm{P}(S > t)$ 的导数：

$$\begin{aligned} g(t) &= -\frac{\mathrm{d}}{\mathrm{d}t}\left(n\,\mathrm{e}^{-(n-1)\lambda t} - (n-1)\,\mathrm{e}^{-n\lambda t}\right) \\ &= n(n-1)\lambda\,\mathrm{e}^{-(n-1)\lambda t} - n(n-1)\lambda\,\mathrm{e}^{-n\lambda t} \\ &= n(n-1)\lambda\left(\mathrm{e}^{-(n-1)\lambda t} - \mathrm{e}^{-n\lambda t}\right) \qquad (t \geq 0). \end{aligned}$$

为求期望值（expectation），把 g 写成指数密度函数的组合：

$$g(t) = n \cdot (n-1)\lambda\,\mathrm{e}^{-(n-1)\lambda t} \;-\; (n-1) \cdot n\lambda\,\mathrm{e}^{-n\lambda t}.$$

分别取 $k = (n-1)\lambda$ 和 $k = n\lambda$ 代入标准积分（这是成立的，因为 $n \geq 2$，所以两种情况下都有 $k > 0$），

$$\begin{aligned} \mathrm{E}(S) = \int_0^\infty t\,g(t)\,\mathrm{d}t &= \frac{n}{(n-1)\lambda} - \frac{n-1}{n\lambda} \\ &= \frac{1}{\lambda} \cdot \frac{n^2 - (n-1)^2}{n(n-1)} = \frac{1}{\lambda} \cdot \frac{2n-1}{n(n-1)}. \end{aligned}$$

由于 $\dfrac{1}{n-1} + \dfrac{1}{n} = \dfrac{n + (n-1)}{n(n-1)} = \dfrac{2n-1}{n(n-1)}$，所以

$$\mathrm{E}(S) = \frac{1}{\lambda}\left(\frac{1}{n-1} + \frac{1}{n}\right),$$

这正是所求的结果。

想一想 为甚么答案是两个简单项之和?

指数分布具有无记忆性（memoryless）。第一封邮件在一段速率为 $n\lambda$ 的指数时间后到达，平均值为 $\frac{1}{n\lambda}$。之后，其余 $n-1$ 封邮件重新开始，就好像时钟刚被重设一样，所以额外的等待时间是速率为 $(n-1)\lambda$ 的指数分布，平均值为 $\frac{1}{(n-1)\lambda}$。把两段等待时间相加，立即得到答案，这也是记住它的好方法。

验算

- 当 $n=2$ 且 $\lambda=1$ 时，第二封邮件的到达时间是两个独立指数时间中较晚的一个。它的平均值应为 $1+\frac{1}{2}=1.5$，而以电脑模拟 10^5 次试验得到 1.501。
- 密度函数 g 的积分为 1：$\int_0^\infty g = n-(n-1)=1$。
- 当 $n=5$ 且 $\lambda=1$ 时，公式给出 $\mathrm{E}(S)=\frac{1}{4}+\frac{1}{5}=0.45$，与 $\int_0^\infty t\,g(t)\,\mathrm{d}t$ 的数值计算结果吻合。

常见错误

- 在恰好一封邮件已到达的情况中漏掉因子 n。n 封中任何一封都可以是提早到达的那封。
- 把 $\mathrm{P}(S>t)$ 写成恰好一封已到达的概率，这样便漏掉了一封也未到达的情况。

STEP 之外

T 和 S 是一个指数样本的首两个顺序统计量（order statistics）。上面的无记忆性论证可以推广，证明第 k 封邮件的到达时间平均值为 $\frac{1}{\lambda}\left(\frac{1}{n}+\frac{1}{n-1}+\cdots+\frac{1}{n-k+1}\right)$，而这正是泊松过程（Poisson process）和排队论（queueing theory）的基础。Blitzstein 和 Hwang 的 *Introduction to Probability*（《概率导论》）对顺序统计量和无记忆性都有清晰的讲解。

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。