详细解答

STEP II 1993，第 16 题

概率与统计 · 泊松过程

> **STEP II 1993，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/93-s2-q16

关键思路

这里有两个互相独立的随机来源：巴士带来一个几何分布（geometric distribution）的等候时间，而乘客则构成一个泊松过程（Poisson process）。先对第一班巴士到站的时间取条件（condition）。一旦知道这个时间，候车人数就只是一段已知长度时间内的泊松计数。第 (iii) 部分接着运用全概率定理（law of total probability），而真正的问题是：只凭统计表和计算机，怎样把所得的无穷和计算至两位有效数字。

解答

以 8:00 之后的分钟数来量度时间。设 T 为第一班巴士到站的时间，并设 X_t 为在 $(0,t]$ 内到达的乘客人数。乘客过程是速率为每分钟 4 人的泊松过程，所以 $X_t \sim \mathrm{Po}(4t)$。巴士只可能在 $t = 1,2,3,\ldots$ 到站（“刚过八时”表示可能在 8:00 到站的那班巴士已经来过并开走了），而且巴士与乘客互相独立。

第 (i) 部分

第一班巴士在 8 时 n 分到站，当且仅当第 $1,2,\ldots,n-1$ 分钟都没有巴士到站，而第 n 分钟有巴士到站。这些事件互相独立，所以

$$\mathrm{P}(T = n) = \left(\tfrac{3}{4}\right)^{n-1}\tfrac{1}{4}, \qquad n = 1,2,3,\ldots$$

因此 T 服从参数为 $\frac{1}{4}$ 的几何分布。（验算一下：这些概率之和为 $\frac{1}{4}/(1-\frac{3}{4}) = 1$。）

第 (ii) 部分

在 8:05 候车的人，正是在 $(0,5]$ 内到达的那些人。没有人离开过，因为 8:05 之前没有巴士到站。所以我们要求的是在 $T=5$ 的条件下 X_5 的分布。

> **想一想** 知道第一班巴士在 8:05 到站，会否改变乘客人数的分布？

不会。事件 $T=5$ 只取决于巴士，而巴士与乘客互相独立。所以对 $T=5$ 取条件并不改变 X_5 的分布，即 $X_5 \sim \mathrm{Po}(4\times 5)=\mathrm{Po}(20)$。因此

$$\mathrm{P}(m \text{ 人候车} \mid T=5)=\frac{\mathrm{e}^{-20}\,20^m}{m!},\qquad m=0,1,2,\ldots$$

第 (iii) 部分

巴士未能接载所有人，当且仅当候车人数超过 25，即 $X_T \geq 26$。设 F 为这个事件。对 T 取条件，并如 (ii) 般运用独立性，

$$\mathrm{P}(F)=\sum_{n=1}^{\infty}\mathrm{P}(T=n)\,\mathrm{P}(X_n\geq 26)=\sum_{n=1}^{\infty}\left(\tfrac{3}{4}\right)^{n-1}\tfrac{1}{4}\,\mathrm{P}(\mathrm{Po}(4n)\geq 26).$$

现在我们需要一个方案，把这个无穷和化为有限的工作量，而误差要远低于 0.005。我们要的是两位有效数字，而答案最后会是 0.1 这个数量级。

第 1 步：改为处理补事件，令尾项容易处理。 记 $w_n=\left(\tfrac{3}{4}\right)^{n-1}\tfrac{1}{4}$ 及 $q_n=\mathrm{P}(\mathrm{Po}(4n)\leq 25)$。由于 $\sum w_n=1$，

$$\mathrm{P}(F)=1-\sum_{n=1}^{\infty}w_n\,q_n.$$

这个形式较好，因为当 n 很大时 q_n 极小。当 $n\geq 11$ 时，平均值 $4n\geq 44$ 远高于 25。由于 q_n 随 n 增加而递减（时间愈长，乘客愈多），

$$\sum_{n\geq 11}w_nq_n\leq q_{11}\sum_{n\geq 11}w_n=q_{11}\left(\tfrac{3}{4}\right)^{10}.$$

用正态近似（normal approximation）可得 $q_{11}\approx\Phi((25.5-44)/\sqrt{44})=\Phi(-2.79)\approx 0.003$，所以这个尾项小于 0.0002，可以略去。这样便只剩下十项。

第 2 步：最初几项。 当 $n=1$ 和 $n=2$ 时（平均值为 4 和 8），有 26 名或以上乘客的可能性极低，在远超所需的准确度之内 $q_1=q_2=1$，所以这两项的贡献是 $w_1+w_2=\tfrac{1}{4}+\tfrac{3}{16}$。

第 3 步：中间各项查表。 当 $n=3,4,5$ 时（平均值为 12、16、20），直接从累积泊松分布表读出 $q_n=\mathrm{P}(\mathrm{Po}(4n)\leq 25)$；这类表通常涵盖至平均值 20。如果表中也列出平均值 24 和 28，则 $n=6,7$ 亦可查表。

第 4 步：较大的平均值用正态近似。 对于余下直至 10 的 n（平均值最大为 40），使用 $\mathrm{Po}(\mu) \approx \mathrm{N}(\mu, \mu)$ 并作连续性校正（continuity correction）：

$$q_n \approx \Phi\left(\frac{25.5 - 4n}{\sqrt{4n}}\right),$$

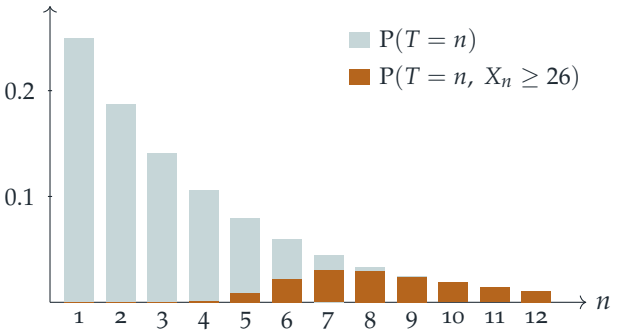
其中 Φ 的值从正态分布表读出。这些平均值已经够大，近似效果良好。更重要的是，权重 w_n 本身已经很小（$w_8 \approx 0.033$，$w_{10} \approx 0.019$），所以 q_n 即使有千分之几的误差，对 $\mathrm{P}(F)$ 的影响也小于 0.001。

第 5 步：合并。 用计算机算出 $w_1, \ldots, w_{10}$，并计算

$$\mathrm{P}(F) \approx 1 - \sum_{n=1}^{10} w_n q_n,$$

然后舍入至两位有效数字。每个被略去或作近似的部分最多约为 0.001，所以除非总和刚好非常接近舍入的分界，否则舍入是稳妥的。万一真的如此，这个方法也清楚显示应该改进哪几项。

题目并没有要求我们实际计算，但算出来的结果是 $\mathrm{P}(F) \approx 0.19$。下图显示这个概率来自何处：大部分来自第一班巴士在 8:06 至 8:10 到站的情况，这时预期的候车人数为 24 至 40，接近或超过载客量。



验算

- 在 (i) 中，各概率之和为 1，而且 $\mathrm{P}(T = 1) = \frac{1}{4}$，与预期相符。
- 在 (ii) 中，候车人数的平均值为 20，即每分钟四人、共五分钟。
- 在 (iii) 中，精确的和（由电脑计算）为 0.1904。以蒙地卡罗模拟（Monte Carlo simulation）模拟 2×10^5 个早上，得 0.189。上述查表加正态近似的方法得 0.1912。三者舍入后都是 0.19。

常见错误

- 分界值差了一。只有当候车人数为 26 或以上时，巴士才未能接载所有人，所以我们需要的概率是 $1 - \mathrm{P}(\mathrm{Po}(4n) \leq 25)$，而不是 $1 - \mathrm{P}(\mathrm{Po}(4n) \leq 24)$。
- 把候车人数当作 $\mathrm{Po}(4\,\mathrm{E}[T]) = \mathrm{Po}(16)$。等候时间是随机的，所以必须把泊松概率对 T 取平均，而不是把平均时间代入。
- 让和式保持无穷，却没有说明为何可以截断。你必须证明被略去的尾项小得不足以影响第二位有效数字。

STEP 之外

候车人数 X_T 是一个平均值本身随机的泊松计数，即泊松分布的一个混合（mixture）。由于 T 服从几何分布，X_T 的母函数（generating function）原来有一个简洁的闭合形式；而像这样的取条件论证，正是排队论（queueing theory）的基础，也是大学第一门概率课程中复合分布（compound distribution）和混合分布（mixed distribution）的基础。怎样截断一个和式并估计误差的上界，这个问题就是数值分析（numerical analysis）的起点。Grimmett 与 Welsh 的 *Probability: An Introduction*（概率论入门教材）以合适的程度讲解泊松过程和母函数。

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。