# [AI2613 随机过程][第六讲] 泊松过程

张驰豪

最后更新：*2025* 年 *4* 月 *14* 日

## 目录



## 1 泊松分布

考虑一个场景：某餐厅过去五天的顾客数量分别是 100、120、80、75 和 110。为了确保明天的食材准备足够，我们需要根据前几天的顾客数量预测明天的顾客数量。一个常见的方法是计算顾客数量的平均值（在本例中为 97）。然而，仅仅使用期望的信息往往是不够的，有可能造成大量天数出现食材短缺的情况。因此，我们尝试建模每天顾客到来人数的分布。

为了得到这个分布，我们需要做一些假设。一个最常见的假设是假设顾客的到来是均匀且独立的。我们可以假设一天被分为 $n$ 个等长的时间段，每个时间段足够短，以至于在该时间段内最多只能有一位顾客进入餐厅，并且在每个时间段以 $p$ 的概率独立有一位顾客进入餐厅。

形式化地表示，对于 $i \in [n]$，我们令 $X_i = \mathbb{1}\left[\text{第} i \text{ 个时间段有顾客进入}\right]$，则 $X_i \sim \mathrm{Ber}(p)$，并且 $X_i$ 相互独立。设

$$Z_n = \sum_{i=1}^{n} X_i, \quad \lambda = \mathbf{E}[Z_n] = p \cdot n.$$

现在我们计算顾客数量 $Z_n$ 的分布。对于任何常数 $k \in \mathbb{N}$，有：

$$\mathbb{P}[Z_n = k] = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

由于我们假设 $k$ 和 $\lambda$ 是常数，当 $n \to \infty$，上述表达式收敛于：

$$\mathbb{P}[Z_n = k] = \frac{\lambda^k}{k!} e^{-\lambda}.$$

我们在学习中心极限定理的时候，也证明过二项式分布的和会收敛到正态分布。和这里不同的点在于，在中心极限定理中，我们假设 $p$ 是一个常数。

我们便把概率质量函数满足任意 $k \in \mathbb{N}$，$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ 的分布称为均值为 $\lambda$ 的泊松分布，并把这个分布记作 $\mathrm{Pois}(\lambda)$。

设 $X \sim \mathrm{Pois}(\lambda)$。由于我们在上面是通过取极限的方式定义了泊松分布，因此需要验证它确实是一个分布，并且其均值确实为 $\lambda$：

- 验证分布性质

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = 1.$$

  因此它确实是一个分布。

- 验证均值

$$\mathbf{E}[X] = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \lambda.$$

两个独立的满足泊松分布的随机变量之和依旧满足泊松分布。

**命题 1.** *假设 $X_1 \sim \mathrm{Pois}(\lambda_1)$ 并且 $X_2 \sim \mathrm{Pois}(\lambda_2)$ 是两个独立的随机变量。那么：*

$$X_1 + X_2 \sim \mathrm{Pois}(\lambda_1 + \lambda_2).$$

这个结论可以推广到任意 $n$ 个满足泊松分布的随机变量：
如果 $X_1, X_2, \ldots, X_n$ 是 $n$ 个相互独立的随机变量，且 $X_i \sim \mathrm{Pois}(\lambda_i)$，则：

$$\sum_{i=1}^{n} X_i \sim \mathrm{Pois}\left(\sum_{i=1}^{n} \lambda_i\right).$$

证明. 对于任意的 $n \geq 0$，

$$\begin{aligned}\mathbb{P}\left[X_1 + X_2 = n\right] &= \sum_{m=0}^{n} \mathbb{P}\left[X_1 = m\right] \cdot \mathbb{P}\left[X_2 = n - m\right] \\ &= \sum_{m=0}^{n} \frac{\lambda_1^m}{m!} e^{-\lambda_1} \cdot \frac{\lambda_2^{n-m}}{(n-m)!} e^{-\lambda_2} \\ &= e^{-(\lambda_1+\lambda_2)} \cdot \sum_{m=0}^{n} \binom{n}{m} \frac{\lambda_1^m \lambda_2^{n-m}}{n!} \\ &= e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1+\lambda_2)^n}{n!}.\end{aligned}$$

□

# 2 泊松过程（Poisson Process）

## 2.1 泊松过程的定义

我们刚才说了，均值为 $\lambda$ 的泊松分布可以用来描述单位时间内平均顾客数为 $\lambda$ 人的时候，顾客人数的分布。如果我们统计一段时间内的顾客数量（例如从时间 $t_1$ 到时间 $t_2$），假设 $t_2 - t_1$ 是整数，并且顾客的到来依然是均匀的，那么按照我们上述的多个泊松分布变量之和的结论，这段时间内来的顾客人数应该符合 $\mathsf{Pois}((t_2 - t_1)\lambda)$ 分布。我们可以使用泊松过程来描述一段时间内到来的顾客人数。我们说一族随机变量 $\{N(s)\}_{s\geq 0}$ 为速率为 $\lambda$ 的泊松过程，当且仅当其满足以下条件：

我们之前说过，对于一个随机过程，如果它的时间戳 $t$ 是取 $0, 1, 2\ldots$ 这样离散的值，我们把它称为“链”；如果它的时间戳 $t$ 是取比如 $\mathbb{R}_{\geq 0}$ 这样连续的值，我们把它称为“过程”。我们可以自然的把马尔可夫链的定义推广到马尔科夫过程（即“无后效/记忆性”）。容易验证，泊松过程是一个马尔科夫过程。

1. $N(0) = 0$；
2. 对于任意 $t, s \geq 0$，有：
$$N(t+s) - N(s) \sim \mathsf{Pois}(\lambda t);$$
3. 对于任意 $t_0 \leq t_1 \leq \cdots \leq t_n$，随机变量：
$$N(t_1) - N(t_0),\ N(t_2) - N(t_1),\ \ldots,\ N(t_n) - N(t_{n-1})$$
相互独立。

实际上，满足条件的这样一族随机变量的概率空间长什么样子，这些随机变量作为概率空间上的可测函数为什么存在，如何构造，并不是一件容易的事情。在我们这个课上，我们不妨假设泊松过程是存在的，然后研究它的性质。

我们可以从另外一个角度来描述一个泊松过程，也就是考虑相邻两个顾客到达的间隔时间。事实上，对于一个速率为 $\lambda$ 的泊松过程，两名顾客之间的间隔时间满足速率为 $\lambda$ 的指数分布 $\mathsf{Exp}(\lambda)$。为了说明这一点，我们先来考虑指数分布的一些性质。

## 2.2 指数分布

回忆到我们曾经说过，速率为 $\lambda > 0$ 的指数分布的概率密度函数为：

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0; \\ 0, & x < 0. \end{cases}$$

它的累积分布函数为：

$$F(t) = \int_0^t \lambda e^{-\lambda x} dx = 1 - e^{-\lambda t}.$$

指数分布可以用来建模顾客到来的时间。比如说，速率为 $\lambda$ 的指数分布可以用来表示，在一个速率为 $\lambda$ 的泊松过程里，从 0 时刻开始，第一个顾客在 $t$ 时刻还未到来的概率是 $1 - F(t) = e^{-\lambda t}$。

使用定义可以容易地计算出，对于 $X \sim \mathsf{Exp}(\lambda)$，有：

$$\mathbf{E}[X] = \frac{1}{\lambda}, \quad \mathbf{Var}[X] = \frac{1}{\lambda^2}.$$

指数分布的一个重要性质就是所谓的“无记忆性”，直观上来说，就是“假设已经等了第一个顾客 $s$ 时间之后，他还没来，那此时还需等待的他到来时间的分布与一开始他到来时间的分布相同”。 万恶之源

**命题 2** (指数分布的无记忆性)**.** *设 $X \sim \mathsf{Exp}(\lambda)$。那么对于任何 $t, s > 0$，*

$$\mathbb{P}[X > t + s \mid X > s] = \mathbb{P}[X > t].$$

*证明.*

$$\begin{aligned} \mathbb{P}[X > t + s \mid X > s] &= \frac{\mathbb{P}[X > t + s \wedge X > s]}{\mathbb{P}[X > s]} \\ &= \frac{\mathbb{P}[X > t + s]}{\mathbb{P}[X > s]} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\ &= e^{-\lambda t}. \end{aligned}$$

□

指数分布另外一个有趣的性质是所谓“指数竞赛”。假设有两家商店，第一家商店第一位顾客到来时间服从 $\mathsf{Exp}(\lambda_1)$ 分布，第二家商店第一位顾客到来时间服从 $\mathsf{Exp}(\lambda_2)$ 分布。那么，这两家店同时开门后，第一位顾客的到来时间符合什么分布？答案是 $\mathsf{Exp}(\lambda_1 + \lambda_2)$。

**命题 3** (指数竞赛)**.** *设 $X_1 \sim \mathsf{Exp}(\lambda_1), X_2 \sim \mathsf{Exp}(\lambda_2)$ 为两个独立的随机变量。那么 $Y := X_1 \wedge X_2 \sim \mathsf{Exp}(\lambda_1 + \lambda_2)$。*

这个性质可以自然推广到 $n$ 个独立指数分布，即如果 $\forall i \in [n]$, $X_i \sim \mathrm{Exp}(\lambda_i)$，那么 $\min\{X_1, \ldots, X_n\} \sim \mathrm{Exp}(\lambda_1 + \cdots + \lambda_n)$。

证明.

$$\mathbb{P}[Y > t] = \mathbb{P}[X_1 > t \wedge X_2 > t] = \mathbb{P}[X_1 > t]\,\mathbb{P}[X_2 > t] = e^{-(\lambda_1+\lambda_2)t}.$$

□

我们接下来考虑“谁赢了竞赛”的问题，也就是说 $X_1 < X_2$ 的概率。我们用 $f_\lambda$ 来表示 $\mathrm{Exp}(\lambda)$ 的概率密度函数，那么使用全概率公式，我们有：

这个结论同样可以自然推广到 $n$ 个指数分布的场合：$X_i$ 是 $X_1, \ldots, X_n$ 中最小的概率为 $\frac{\lambda_i}{\sum_{j=1}^n \lambda_j}$。

$$\mathbb{P}[X_1 < X_2] = \int_0^\infty f_{\lambda_1}(t) \cdot \mathbb{P}[X_2 > t]\,\mathrm{d}t = \int_0^\infty \lambda_1 e^{-\lambda_1 t} e^{-\lambda_2 t}\,\mathrm{d}t = \frac{\lambda_1}{\lambda_1+\lambda_2}.$$

## 2.3 泊松过程的指数分布刻画

我们接下来证明一个重要的结论，它说明我们一开始定义的泊松过程可以用指数分布的间隔来描述。

**定理 4.** *设 $\tau_1, \tau_2, \ldots, \tau_n$ 为一系列独立的指数分布随机变量，满足 $\tau_i \sim \mathrm{Exp}(\lambda)$。设 $T_n = \sum_{i=1}^n \tau_i$。对于 $t \geq 0$，定义 $N(t) := \max\{n \mid T_n \leq t\}$。那么 $\{N(t)\}_{t\geq 0}$ 是一个泊松过程。*

根据定理描述，我们可以想象 $\tau_i$ 就是第 $i-1$ 个顾客和第 $i$ 个顾客到达的时间间隔。$T_n$ 就表示第 $n$ 个顾客到达的时间。那么 $N(t)$ 就表示 $t$ 时刻之前到达了多少顾客。

证明. 我们首先说明，$T_n$ 满足所谓的 Gamma 分布 $\Gamma(n, \lambda)$，其概率密度为

$$g_n(t) = \begin{cases} \lambda e^{-\lambda t} \cdot \frac{(\lambda t)^{n-1}}{(n-1)!}, & t \geq 0; \\ 0, & t < 0. \end{cases}$$

我们通过对 $n$ 进行归纳来证明。当 $n = 1$ 时，$T_1 = \tau_1 \sim \mathrm{Exp}(\lambda) = \Gamma(1, \lambda)$。假设 $T_n \sim \Gamma(n, \lambda)$ 对于 $n \geq 1$ 成立。根据 $T_n$ 和 $\tau_{n+1}$ 的独立性，对于每一个 $t \geq 0$，我们有

$$\begin{aligned} g_{n+1}(t) &= \int_0^t g_n(s) \cdot f_\lambda(t-s)\,\mathrm{d}s \\ &= \int_0^t \lambda e^{-\lambda s} \cdot \frac{(\lambda s)^{n-1}}{(n-1)!} \cdot \lambda e^{-\lambda(t-s)}\,\mathrm{d}s \\ &= \lambda e^{-\lambda t} \frac{\lambda^n}{(n-1)!} \int_0^t s^{n-1}\,\mathrm{d}s \\ &= \lambda e^{-\lambda t} \frac{\lambda^n}{(n-1)!} \cdot \frac{t^n}{n} \\ &= \lambda e^{-\lambda t} \cdot \frac{(\lambda t)^n}{n!}. \end{aligned}$$

于是，我们便可以使用全概率公式来计算 $N(t)$ 的分布。

$$\begin{aligned}\mathbb{P}\left[N(t)=n\right] &= \mathbb{P}\left[T_n \le t \wedge T_{n+1} > t\right] \\ &= \int_0^t g_n(s)\cdot \mathbb{P}\left[\tau_{n+1} > t-s\right] \mathrm{d}s \\ &= \int_0^t \lambda e^{-\lambda s}\cdot \frac{(\lambda s)^{n-1}}{(n-1)!}\cdot e^{-\lambda(t-s)}\,\mathrm{d}s \\ &= \lambda^n e^{-\lambda t}\cdot \frac{t^n}{n!}.\end{aligned}$$

因此，$N(t) \sim \mathsf{Pois}(\lambda t)$。我们接着来一一验证 $\{N(t)\}_{t\ge 0}$ 满足泊松过程定义的三个条件。

首先，$N(0)=0$ 是显然的。根据指数分布的无记忆性，我们知道 $N(s+t)-N(s)$ 的分布和 $N(t)-N(0)$ 一样是 $\mathsf{Pois}(\lambda t)$，因此第二个条件也满足。我们同样可以使用指数分布的无记忆性验证第三个独立性条件。 □

## 3 泊松过程的稀疏化 (Thinning)

在我们用泊松过程建模顾客到店的例子里，我们可以分别考虑男顾客和女顾客的到来。假设男女顾客的比例是 $1:1$，我们可以等价地建模为：对于每一个到达的顾客，投掷一个公平的硬币，若是正面，则该顾客为男性；若是反面，则该顾客为女性。泊松过程的稀疏化定理说明，男女顾客人数分别构成两个独立的速率为 $\frac{\lambda}{2}$ 的泊松过程。

严格地说，给定一个泊松过程 $\{N(t)\}_{t\ge 0}$，我们可以给第 $i$ 位到达的顾客引入一个独立同分布的随机变量 $Y_i$，表示该顾客的种类，从而可以将泊松过程划分为若干子过程。我们假设 $Y_i$ 的取值是有限的非负整数，并且设 $p_k = \mathbb{P}\left[Y_i = k\right]$ 为其概率质量函数。对于每一个 $Y_i$ 的可能取值 $k$，我们用 $N_k(t)$ 表示截止时间 $t$ 时具有种类 $k$ 的顾客到达数量。那么 $\{N_k(t)\}_{t\ge 0}$ 被称为泊松过程的一个“稀疏化”。有如下一个有用且令人惊讶的命题。

**命题 5.** *对于每个 $k$，$\{N_k(t)\}_{t\ge 0}$ 是一个参数为 $\lambda p_k$ 的泊松过程。此外，所有过程的集合*

$$\left\{\{N_k(t)\}_{t\ge 0} : k \in \mathsf{Im}(Y_1)\right\}$$

*是相互独立的。*

注意到，我们说两个随机过程 $N_1(t)$ 和 $N_2(t)$ 独立，指的是对于任意 $t_0 \le t_1 \le \cdots \le t_n$，随机变量 $\bigcup_{i\in[n]}\{N_1(t_i)-N_1(t_{i-1}), N_2(t_i)-N_2(t_{i-1})\}$ 是相互独立的。

*证明.* 我们只需要证明 $Y_i$ 取两种值的情况即可，多种值的情况可以类似证明。假设 $Y_i \in \{0,1\}$。由于泊松分布本身的独立增量性质，为了验证独立性，我们只需要验证对于任意 $s,t\ge 0$，$N_0(t+s)-N_0(s)$ 和 $N_1(t+s)-N_1(s)$ 独立即可。

假设我们有两个速率为 $\lambda_1$ 和 $\lambda_2$ 的泊松过程，类似于有两个商店，他们到来的顾客是独立的。我们可以考察它们到来的顾客的总和，那这个总和也是一个泊松过程，容易验证它的速率是 $\lambda_1+\lambda_2$。稀疏化可以看成这个总和过程的逆操作，因此，两个过程是独立的并不那么令人惊讶。但实际上，我们未来会看到，稀疏化里的这个独立性会有产生很多神奇的结论。从另外一个角度来说，稀疏化的独立性也有不直观的地方。假设我们有一家商店的顾客满足 Pois(1)，即平均一分钟有一位顾客。然后每一个到来的顾客都以 50% 的概率是男生，50% 的概率是女生，可以等价的看成，每一位顾客到来后，我们投掷一个公平硬币，来决定顾客的性别。那么，假设在已知这一分钟来了 10 名男生顾客的情况下，女生顾客平均应该来多少呢？直观上，由于性别是用投掷公平硬币决定的，在男生到来的数量大大超过平均数的情况下，应该是发生了来的顾客人数偏离平均数的事件，所以女生到来的人数应该也远高于平均数。但实际上，稀疏化告诉我们，女生人数不受男生人数的影响，是独立的。

我们下面计算 $N_0(t)$ 和 $N_1(t)$ 的联合分布，并验证每一个 $N_i(t)$ 都是泊松过程。由泊松过程的无记忆性，便可以得到独立性结论：

$$\begin{aligned}\mathbb{P}\left[N_0(t)=m \wedge N_1(t)=n\right] &= \mathbb{P}\left[N_0(t)=m \wedge N(t)=m+n\right] \\ &= \mathbb{P}\left[N(t)=m+n\right] \cdot \mathbb{P}\left[N_0(t)=m \mid N(t)=m+n\right] \\ &= \frac{e^{-\lambda t}(\lambda t)^{m+n}}{(m+n)!} \cdot \binom{m+n}{m} p_0^m p_1^n \\ &= \frac{e^{-\lambda p_0 t}(\lambda p_0 t)^m}{m!} \cdot \frac{e^{-\lambda p_1 t}(\lambda p_1 t)^n}{n!}.\end{aligned}$$

□

## 3.1 稀疏化的一个应用

我们来看泊松过程在最大似然估计（maximum likelihood estimation, MLE）中的一个应用。

假设有两位编辑在阅读一本 300 页的书。编辑 A 在书中发现了 100 个错别字，编辑 B 发现了 120 个错别字，其中 80 个是两人都发现的。

假设作者的错别字遵循一个参数为 $\lambda$ 的泊松过程，每页的错误率未知。两位编辑分别以未知的成功概率 $p_A$ 和 $p_B$ 抓到错别字。我们想知道总共有多少错别字。这可以通过估计 $\lambda$、$p_A$ 和 $p_B$ 来解决这个问题。

显然，总共有四种类型的错别字：

- **类型 1**：未被两位编辑发现的错别字。发生概率为 $p_1 = (1-p_A)(1-p_B)$。
- **类型 2**：仅被编辑 A 发现的错别字。发生概率为 $p_2 = p_A(1-p_B)$。
- **类型 3**：仅被编辑 B 发现的错别字。发生概率为 $p_3 = (1-p_A)p_B$。
- **类型 4**：被两位编辑都发现的错别字。发生概率为 $p_4 = p_A p_B$。

因此，每种类型的错别字发生过程是一个独立的泊松过程，参数为 $\lambda p_i$。令 $N_1, N_2, N_3, N_4$ 分别表示书中对应类型的错别字数量，则 $N_i \sim \text{Pois}(300\lambda p_i)$。

书中有 20 个类型 2 的错别字，40 个类型 3 的错别字，80 个类型 4 的错别字。我们声称最可能的参数值满足以下方程组：

$$\begin{cases} 300\lambda p_A(1-p_B) = 20, \\ 300\lambda(1-p_A)p_B = 40, \\ 300\lambda p_A p_B = 80. \end{cases}$$

换句话说，我们在得到一个 $X \sim \mathrm{Pois}(\mu)$ 的样本 $x$ 之后，认为最有可能的 $\mu$ 的值就是 $x$。这是泊松分布的最大似然估计，我们马上进行验证。假设承认这件事情，我们可以立刻解得：

$$\lambda = \frac{1}{2}, \quad p_A = \frac{2}{3}, \quad p_B = \frac{4}{5}.$$

## 3.2 泊松分布的最大似然估计验证

我们最后验证上述最大似然估计是否成立。假设 $N \sim \mathrm{Pois}(\lambda)$，其中 $\lambda$ 未知。给定 $N = n$，我们的目标是找到：

$$\arg\max_{\lambda} \mathbb{P}[N = n] \quad \text{provided } N \sim \mathrm{Pois}(\lambda).$$

注意到在已知 $N \sim \mathrm{Pois}(\lambda)$ 时，$\mathbb{P}[N = n] = \frac{e^{-\lambda}\lambda^n}{n!}$，且

$$\log \mathbb{P}[N = n] = -\lambda + n \log \lambda - \log n!.$$

因此，最大化的目标等价于：

$$\arg\max_{\lambda} \{-\lambda + n \log \lambda\}.$$

对上式求导并令其为 0，得到：

$$-\frac{\partial}{\partial \lambda}\{-\lambda + n \log \lambda\} = -1 + \frac{n}{\lambda} = 0.$$

解得 $\lambda = n$，即泊松分布的最大似然估计值是样本的观测值 $n$。