# [AI2613 随机过程][第七讲] 泊松近似

张驰豪

最后更新：*2025* 年 *4* 月 *14* 日

## 目录



## 1 非均匀奖券收集（Coupon Collector）问题

我们在第一次课讲过奖券收集问题，现在复述如下：

> 考虑玩一个抽卡手游。现在总共有 $n$ 种不同类型的卡，每一抽可以均匀的得到其中一种。现在想问平均要抽多少次，可以集齐一套，即 $n$ 种卡每种至少一张。

我们使用期望的线性性证明了，如果每一次抽卡每一种类型的卡出现的概率都是等概率的 $\frac{1}{n}$，那么我们期望需要抽 $nH_n$ 次就能收集到所有类型的卡，其中 $H_n = \sum_{k=1}^{n} \frac{1}{k} \xrightarrow{n\to\infty} n\log n + \gamma n$ 是调和级数。然而在实际中，手游公司往往会对每一种卡有稀有度的设定，比如，对于每一抽，第 $i$ 种卡被开出来的概率是 $p_i$ （$\sum_{i=1}^{n} p_i = 1$）。那么，在这样一种设定下，集齐一套平均要抽多少次呢？稍微想一下就会明白，因为现在每种卡不再有对称性，我们之前基于期望的线性性的简单技巧不再有效了。

令 $N_i$ 表示第一次获得第 $i$ 种卡片需要抽卡的次数，那么 $N_i$ 服从参数为 $p_i$ 的几何分布。令 $N$ 表示收集到所有 $n$ 种类型卡片所需的抽卡次数，那么 $N = \max_{i\in[n]} N_i$。我们问题便是计算 $\mathbf{E}[N]$。然而，由于 $N_i$ 之间不是独立的，直接计算 $\mathbf{E}[N]$ 并不容易。

我们可以脑补如下一个抽卡的方式，和我们要研究的问题是等价的：

> 玩家在一个线下的商店柜台购买卡包进行抽卡。每一分钟过来一位顾客，进行一次抽卡。现在问平均第几位顾客（也就是第几分钟）抽完之后，前面所有的顾客抽的卡放在一起能够集齐所有 $n$ 种。

**泊松抽卡法** 我们现在稍稍修改上面这个场景，假设柜台的顾客并不是严格的每一分钟过来一位，而是按照速率为 1 的泊松过程过来。同样，每一位过来的顾客随机抽一张卡。我们同样考虑当所有顾客抽的卡放在一起集齐全套的时间 $T$。注意这里 $T \in \mathbb{R}$ 是一个实数。

回忆我们在上一讲讨论过的泊松过程的稀疏化（thinning）。令 $X_i(t)$ 表示在时间区间 $[0,t]$ 内，通过这个泊松过程收集到的类型 $i$ 卡片的数量。那么 $\{X_i(t)\}_{t\geq 0}$ 是一个速率为 $p_i$ 的泊松过程，并且所有这些 $X_i(t), i \in [n]$ 是相互独立的。换句话说，我们看那些抽到了第 $i$ 种卡片的人的队伍，它们的人数是各自独立的泊松过程！

对于 $i \in [n]$，令 $T_i = \min\{t \mid X_i(t) = 1\}$ 表示第一次开出类型 $i$ 的卡片的时间。显然 $T_i$ 的分布是 $\mathsf{Exp}(p_i)$，并且 $T = \max_{i\in[n]} T_i$。于是，由于独立性，我们可以计算得到

$$\mathbf{E}[T] = \int_0^\infty \mathbb{P}[T \geq t]\,\mathrm{d}t = \int_0^\infty \left(1 - \prod_{i=1}^n (1 - e^{-p_i t})\right)\mathrm{d}t.$$

这里通过泊松分布的稀疏化神奇的引入了独立性质，从而使简单的计算变得可能。那么它和我们实际想分析的抽卡过程有什么关系呢？直观上来说，由于是速率为 1 的泊松过程，平均一分钟抽一张卡，所以，从平均的意义上来看，$\mathbf{E}[T]$ 很有可能和 $\mathbf{E}[N]$，即固定一分钟抽一张卡的平均集齐时间是很接近的。接下来，我们通过耦合（coupling）的方法说明事实上 $\mathbf{E}[T] = \mathbf{E}[N]$。

我们这里使用了对于任何非负随机变量 $\mathbf{E}[T] = \int_0^\infty \mathbb{P}[T \geq t]\,\mathrm{d}t$ 这个公式。它的验证如下：

$$\begin{aligned}\int_0^\infty \mathbb{P}[T \geq t]\,\mathrm{d}t &= \int_0^\infty \mathbf{E}[\mathbb{1}[T \geq t]]\,\mathrm{d}t \\ \text{(Fubini)} \qquad &= \mathbf{E}\left[\int_0^\infty \mathbb{1}[T \geq t]\,\mathrm{d}t\right] \\ &= \mathbf{E}[T].\end{aligned}$$

**泊松抽卡法与正常抽卡法的比较** 想象现在有两个柜台，壹号柜台是固定一分钟来一个顾客抽卡，贰号柜台是按照速率为 1 的泊松过程到来顾客抽卡。我们想象，两个柜台的各自第 $i$ 位顾客总是抽到同样的卡片（这种定义联合分布的思想实验便叫做耦合）。显然，假设壹号柜台上第 $N$ 位顾客抽完卡后集齐了一套（$N$ 是随机变量），那么在贰号柜台上，也是第 $N$ 位顾客抽完卡后集齐一套。如果我们用 $\tau_i$ 表示贰号柜台第 $i-1$ 位顾客和第 $i$ 位顾客到达的间隔时间，那么 $T$ 和 $\sum_{i=1}^N \tau_i$ 有同样的分布。于是，

$$\mathbf{E}[T] = \mathbf{E}\left[\sum_{i=1}^N \tau_i\right].$$

由于 $\tau_i \sim \mathsf{Exp}(1)$，所以 $\mathbf{E}[\tau_i] = 1$。在上面的式子里，如果 $N = n$ 是一个常数，那么就有期望的线性性 $\mathbf{E}\left[\sum_{i=1}^N \tau_i\right] = \sum_{i=1}^n \mathbf{E}[\tau_i] = n$ 成立。但是，我们这里 $N$ 是一个随机变量，在最一般的情况下，期望和求和是不一定可以交换的。但在我们这儿，$N$ 和 $\{\tau_i\}$ 是独立的，Wald's

Equation 可以保证这儿交换是成立的。对于最一般的 Wald's equation，我们要学习了鞅相关的知识后才能够比较方便地证明。但在现在这个特殊情况，我们可以直接使用定义证明。

**命题 1.** 在我们上述例子里

$$\mathbf{E}\left[\sum_{i=1}^{N}\tau_i\right]=\mathbf{E}\left[N\right].$$

证明.

$$\mathbf{E}\left[\sum_{i=1}^{N}\tau_i\right]=\mathbf{E}\left[\sum_{i=1}^{\infty}\tau_i\cdot\mathbb{I}_{[i\le N]}\right]\overset{\text{(Fubini)}}{=}\sum_{i=1}^{\infty}\mathbf{E}\left[\tau_i\cdot\mathbb{I}_{[i\le N]}\right].$$

由于 $\tau_i$ 和 $[i\le N]$ 独立，所以 $\mathbf{E}\left[\tau_i\cdot\mathbb{I}_{[i\le N]}\right]=\mathbf{E}\left[\tau_i\right]\cdot\mathbb{P}\left[i\le N\right]$。于是

$$\mathbf{E}\left[\sum_{i=1}^{N}\tau_i\right]=\sum_{i=1}^{\infty}\mathbb{P}\left[N\ge i\right]=\mathbf{E}\left[N\right].$$

□

这便说明了，对于非均匀的奖券收集问题，平均集齐一套的时间

$$\mathbf{E}\left[N\right]=\mathbf{E}\left[T\right]=\int_0^{\infty}\left(1-\prod_{i=1}^{n}(1-e^{-p_it})\right)\mathrm{d}t.$$

上述式子看起来比较吓人，我们接着进行一个合理性检查，即对每一个 $p_i=\frac{1}{n}$ 的时候计算一下这个积分。于是乎，

神奇

$$\begin{aligned}
\mathbf{E}\left[N\right]&=\int_0^{\infty}1-\prod_{i=1}^{n}\left(1-e^{-\frac{t}{n}}\right)\mathrm{d}t\\
&\overset{(x=e^{-\frac{t}{n}})}{=}-n\int_0^1 1-(1-x)^n\,\mathrm{d}\log x\\
&=-n\int_0^1\frac{1}{x}-\frac{(1-x)^n}{x}\,\mathrm{d}x\\
&=-n\int_0^1\sum_{k=1}^{n}\frac{(1-x)^{k-1}}{x}-\frac{(1-x)^k}{x}\,\mathrm{d}x\\
&\overset{\text{(Fubini)}}{=}-n\sum_{k=1}^{n}\int_0^1(1-x)^{k-1}\,\mathrm{d}x\\
&=n\sum_{k=1}^{n}\frac{1}{k}=nH_n.
\end{aligned}$$

# 2 泊松近似与投球入箱问题

我们在第一节课研究过投球入箱（balls-into-bins）的模型，即将 $m$ 个相同的球随机投放到 $n$ 个箱子中的随机试验。对任意 $i \in [n]$，令 $X_i$ 表示第 $i$ 个箱子中的球的数量。如果投放是均匀随机的，那么我们有 $X_i \sim \mathrm{Binom}\left(m, \frac{1}{n}\right)$ 且 $\mathbf{E}[X_i] = \frac{m}{n}$。

这个模型可以用来建模很多概率问题，比如我们以前讨论过的奖券收集问题以及生日悖论等。在计算机科学中，它也很自然地被用来建模随机映射的哈希表。为了理解哈希表中的冲突，我们自然会关注 $\max_{i\in[n]} X_i$ 的值，这个被称为最大负载（maxload）。然而，最大负载 $\max_{i\in[n]} X_i$ 并不是一个特别容易计算的量，原因在于 $X_i$ 之间不是相互独立的。然而，我们可以使用泊松分布来近似计算它。我们将要发展一个研究投球入箱问题的很一般化的工具。

## 2.1 泊松近似定理

**定理 2.** *设 $\forall i \in [n], Y_i \sim \mathrm{Pois}(\lambda)$ 是一组独立的泊松分布，其中 $\lambda > 0$ 为任意固定常数。那么，在 $\sum_{i=1}^n Y_i = m$ 的条件下，$(Y_1, \ldots, Y_n)$ 和 $(X_1, \ldots X_n)$ 具有相同的分布。*

*换句话说，对于任何 $a_1, \ldots, a_n$，*

$$\mathbb{P}\left[(Y_1, \ldots, Y_n) = (a_1, \ldots, a_n) \,\middle|\, \sum_{i=1}^n Y_i = m\right] = \mathbb{P}\left[(X_1, \ldots, X_n) = (a_1, \ldots, a_n)\right].$$

*证明.* 对于任意给定的 $(a_1, \ldots, a_n) \in \mathbb{N}^n$ 满足 $\sum_{i=1}^n a_i = m$，我们有

$$\mathbb{P}\left[(X_1, \ldots, X_n) = (a_1, \ldots, a_n)\right] = \frac{1}{n^m} \cdot \frac{m!}{a_1! a_2! \cdots a_n!}.$$

另一方面，

$$\begin{aligned}
&\mathbb{P}\left[(Y_1, \ldots, Y_n) = (a_1, \ldots, a_n) \,\middle|\, \sum_{i=1}^n Y_i = m\right] \\
&= \frac{\mathbb{P}\left[(Y_1, \ldots, Y_n) = (a_1, \ldots, a_n)\right]}{\mathbb{P}\left[\sum_{i=1}^n Y_i = m\right]} \\
&= \frac{\prod_{i=1}^n \mathbb{P}\left[Y_i = a_i\right]}{\mathbb{P}\left[\sum_{i=1}^n Y_i = m\right]} \\
&= \frac{\prod_{i=1}^n e^{-\lambda} \frac{\lambda^{a_i}}{a_i!}}{e^{-\lambda n} \frac{(\lambda n)^m}{m!}} = \frac{1}{n^m} \cdot \frac{m!}{a_1! a_2! \cdots a_n!}.
\end{aligned}$$

□

上面这个等分布的结论说明，当我们想计算 $(X_1, \ldots, X_n)$ 的某个性质的时候，我们可以转而计算独立的 $(Y_1, \ldots, Y_n)$ 的性质。当然了，我们等分布的结论需要 $\sum_i Y_i = m$ 这个条件，因此处理的方法便是把所有不满足这个条件的贡献全部扔掉。

**推论 3** (泊松近似公式)**.** *设 $f : \mathbb{N}^n \to \mathbb{N}$ 是可测函数，且 $Y_1, Y_2, \ldots, Y_n$ 是独立的泊松随机变量，其速率为 $\lambda = \frac{m}{n}$。则有：*

$$\mathbf{E}\left[f(X_1, X_2, \ldots, X_n)\right] \le e\sqrt{m} \cdot \mathbf{E}\left[f(Y_1, Y_2, \ldots, Y_n)\right].$$

*证明.* 根据全期望公式，我们有：

$$\mathbf{E}\left[f(Y_1, \ldots, Y_n)\right] = \sum_{k=0}^{\infty} \mathbf{E}\left[f(Y_1, \ldots, Y_n) \,\middle|\, \sum_{i=1}^{n} Y_i = k\right] \cdot \mathbb{P}\left[\sum_{i=1}^{n} Y_i = k\right].$$

由于 $f$ 是非负函数，我们扔掉所有 $k \neq m$ 的项，可以得到

$$\mathbf{E}\left[f(Y_1, \ldots, Y_n)\right] \ge \mathbf{E}\left[f(Y_1, \ldots, Y_n) \,\middle|\, \sum_{i=1}^{n} Y_i = m\right] \cdot \mathbb{P}\left[\sum_{i=1}^{n} Y_i = m\right].$$

我们知道，假设 $Y_i \sim \mathsf{Pois}(\lambda)$，则 $\sum_{i=1}^{n} Y_i \sim \mathsf{Pois}(\lambda n)$，并且上述等式对于任意 $\lambda$ 均成立。我们希望 $\mathbb{P}\left[\sum_{i=1}^{n} Y_i = m\right]$ 尽量的大，根据我们计算过的泊松分布的最大似然原理，我们取 $\lambda = \frac{m}{n}$，于是根据 Stirling 公式（需要对常数进行仔细的讨论），有

$$\mathbb{P}\left[\sum_{i=1}^{n} Y_i = m\right] = e^{-m}\frac{m^m}{m!} > \frac{1}{e\sqrt{m}}.$$

所以

$$\mathbf{E}\left[f(Y_1, \ldots, Y_n)\right] \ge \frac{1}{e\sqrt{m}}\mathbf{E}\left[f(Y_1, \ldots, Y_n) \,\middle|\, \sum_{i=1}^{n} Y_i = m\right] = \frac{1}{e\sqrt{m}}\mathbf{E}\left[f(X_1, \ldots, X_n)\right].$$

□

## 2.2 最大负载的上下界

我们现在研究在 $m = n$ 时候的最大负载问题。我们将证明，$m = n$ 时，最大负载 $X = \max_{i \in [n]} X_i$ 以 $1 - o(1)$ 的概率，满足

$$X = \Theta\left(\frac{\log n}{\log \log n}\right).$$

**上界** 首先证明上界，即存在常数 $c_1 > 0$，使得 $\mathbb{P}\left[X \geq c_1 \frac{\log n}{\log\log n}\right] = o(1)$。令 $k = \frac{c_1 \log n}{\log\log n}$。通过 union-bound，我们有：

$$\mathbb{P}[X \geq k] = \mathbb{P}[\exists i \in [n], X_i \geq k] \leq \sum_{i=1}^{n} \mathbb{P}[X_i \geq k] = n \cdot \mathbb{P}[X_1 \geq k].$$

再次使用 union-bound，可以得到

$$\mathbb{P}[X \geq k] \leq n \cdot \binom{n}{k} n^{-k} \leq n \left(\frac{e}{k}\right)^k.$$

注意到

$$k \log k = \frac{c_1 \log n}{\log\log n} (\log\log n - \log\log\log n + \log c_1).$$

取 $c_1 = 6$，我们有

$$\log n + k - k \log k < -\log n.$$

于是，$\mathbb{P}[X \geq k] \leq n \left(\frac{e}{k}\right)^k < \frac{1}{n} = o(1)$。

**下界** 我们接着使用泊松近似公式证明下界，即存在常数 $c_2 > 0$，使得：

$$\mathbb{P}\left[X \leq \frac{c_2 \log n}{\log\log n}\right] = o(1).$$

设 $h = \frac{c_2 \log n}{\log\log n}$。我们定义函数 $f(X_1, \ldots, X_n) := \mathbb{I}_{[X \leq h]} = \mathbb{I}_{[\max_i X_i \leq h]}$。于是，根据泊松近似公式

$$\mathbb{P}[X \leq h] = \mathbf{E}[f(X_1, \ldots, X_n)] \leq e\sqrt{n}\mathbf{E}[f(Y_1, \ldots, Y_n)] = e\sqrt{n} \cdot \mathbb{P}\left[\max_{i \in [n]} Y_i \leq h\right].$$

根据 $Y_i$ 的定义，我们有

$$\begin{aligned}\mathbb{P}\left[\max_{i \in [n]} Y_i \leq h\right] &= (\mathbb{P}[Y_1 \leq h])^n = (1 - \mathbb{P}[Y_1 > h])^n \\ &\leq (1 - \mathbb{P}[Y_1 = h+1])^n = \left(1 - \frac{1}{(h+1)!e}\right)^n \leq e^{-\frac{n}{e(h+1)!}}.\end{aligned}$$

注意到

$$\begin{aligned}\log(h+1)! &= \sum_{i=1}^{h+1} \log i < \int_1^{h+2} \log x \, \mathrm{d}x \\ &= (h+2)\log(h+2) - h - 1 \leq (h+2)\log h - h + 3 \\ &= \frac{c_2 \log n + 2\log\log n}{\log\log n} (\log\log n - \log\log\log n + \log c_2) - \frac{c_2 \log n}{\log\log n} + 3 \\ &\leq c_2 \log n - \log\log n - 2.\end{aligned}$$

设 $c_2 = 1$, 我们有 $\log(h+1)! \le \log n - \log\log n - 2$。因此

$$e(h+1)! \le \frac{n}{e\log n}.$$

所以

$$\mathbb{P}\left[\max_{i\in[n]} Y_i \le h\right] \le e^{-\frac{n}{(h+1)!e}} \le e^{-e\log n} = n^{-e}.$$

综上所述，我们证明了 $m = n$ 时，最大负载 $X$ 满足

$$X = \Theta\left(\frac{\log n}{\log\log n}\right),$$

且该结果以 $1 - o(1)$ 的概率成立。