泊松分布：从 0 到 1

陈普　　整理

伯努利分布　如果 X 满足

$$X = \begin{cases} 1, & \text{概率为}p \\ 0, & \text{概率为}1-p \end{cases}$$

则 X 服从伯努利分布。易知其期望和方差为 p 和 $p(1-p)$。

二项分布　若同时进行 n 个相同的伯努利实验，定义事件 $A_i = \{\text{对第}i\text{个实验有}X=1\}$，那么可以定义随机变量，

$$Y = n\text{个实验中成功实验的数目}$$

即 n 个实验 y 个成功，$n-y$ 个不成功，则事件 $\{Y=y\}$ 发生了。这样的事件发生，必然意味着一个形如这样的事件 $A_1 \cap A_2 \cap A_3^C \cdots A_n^C$ 产生（其中 A_i^C 是 A_i 的补集），而产生一个这样的事件的概率为，

$$P(A_1 \cap A_2 \cap A_3^C \cdots A_n^C) = pp(1-p)\cdots(1-p) = p^y(1-p)^{n-y}$$

但上述概率只是说有 y 个成功发生，至于是哪 y 个成功发生是不知道的，那么要穷尽所有可能的 y 个成功事件，无非是从 n 个挑 y 个的组合数，因此，

$$P(Y=y|n,y) = \binom{n}{y} p^y(1-p)^{n-y}$$

上述概率就是二项分布的概率，其中，$\binom{n}{y} = C_y^n = \frac{n!}{y!(n-y)!}$。

泊松分布　泊松分布就是给定时间内事物出现的次数所服从的分布。可用一个例子描述，比如一小时内银行客户到达的数目，这个数目可以服从泊松分布。该分布的概率公式为，

$$P(X=x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

要得到这个概率分布可如下思考，考虑一段时间内银行客户的到达数。将这段时间区间分成 n 段，每段时间正好有一个客户到达的概率与这段时间的长度 $\frac{1}{n}$ 成正比（这也是泊松分布的一个重要假设），比例为 λ，即 $P = \frac{\lambda}{n}$。同时认为因为时间太短，两个客户到达是不可能事件，于是没有客户到达的概率为 $1-\frac{\lambda}{n}$。这样，这 n 个时间段的每个小段内，要么到达一个，要么没有人到达。这不就是一个二项分布吗？那么在这个时间段，到达 x 个客户的概率就可以利用二项分布公式，

$$P(X=x) = C_x^n \cdot \left(\frac{\lambda}{n}\right)^x \cdot \left(1-\frac{\lambda}{n}\right)^{n-x}$$

当 $n \to \infty$，有，

$$\frac{C_x^n}{n^x} \to \frac{1}{x!}, \qquad \left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}$$

因此，

$$P(X=x) = C_x^n \cdot \left(\frac{\lambda}{n}\right)^x \cdot \left(1-\frac{\lambda}{n}\right)^{n-x} \to \frac{e^{-\lambda}\lambda^x}{x!}$$

泊松分布的期望和方差都是 λ，λ 也称为泊松到达率，即这段时间内到达客户的平均数目。

例子 1 可以举个例子，比如平均来看，三个小时内到达的客户是 5 人，即 $\lambda = 5$，那么三个小时内，到达 $0, 1, 2, \cdots, 10$ 人的概率就可用 R 语言计算如下，

```
> data.frame(n = 0:9, p= lapply(0:9, dpois, lambda = 5) %>% unlist)
```

输出结果为，

客户到达数目	概率
0	0.0067379472
1	0.0336897353
2	0.0842243374
3	0.1403738965
4	0.1754673706
5	0.1754673707
6	0.1462228088
7	0.1044448639
8	0.06527803910
9	0.036265577

可以看到 $\lambda = 5$ 的时候概率最大。

例子 2 公司门口有 10 个停车位，公司有 100 个上班的员工，每个员工早上 8 点之前开车来上班的概率是 10%，他们每天什么时候来公司不仅是随机的，而且彼此无关，不存在两个人商量之后一起到的情况，而且也不存在头一天来晚了没抢到停车位，第二天找到的可能性。现在你是这家公司的新员工，早上 8 点整开车到了公司，请问停车场还有车位的概率是多大?

第一步计算，平均来看，早上八点会来几台车。这个可以用二项分布搞定，算二项分布期望的代码如下，

```
sum((lapply(1:100, dbinom, size = 100, prob = 0.1) %>% unlist()) * (1:100))
```

这个值为 10，即平均来看是 10 台。

第二步计算，基于 $\lambda = 10$，用泊松分布计算 0 到 10 的概率，然后累加 0 到 10 的概率就是有车位的概率。

```
data.frame(n = 0:10, p= lapply(0:10, dpois, lambda = 10) %>% unlist)
```