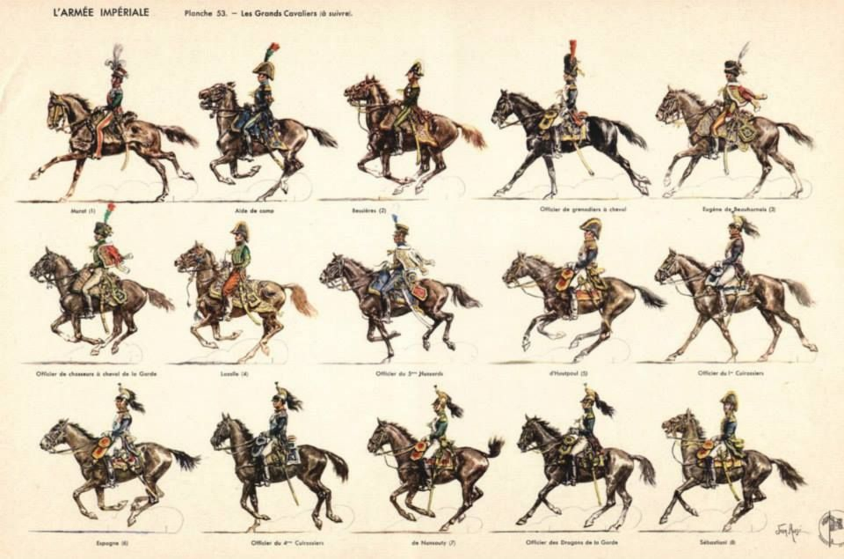
Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#433 Memoryless Process (Geometric, Exponential)

été 2020

… and the Poisson distribution



The Geometric Distribution

The Geometric $\mathcal{G}(p)$, $p \in (0,1)$

$$\mathbb{P}(X = k) = p\,(1-p)^{k-1} \text{ for } k = 1, 2, \cdots$$

with cdf $\mathbb{P}(X \leq k) = 1 - p^k$.

Observe that this distribution satisfies the following relationship

$$\frac{\mathbb{P}(X = k+1)}{\mathbb{P}(X = k)} = 1 - p \text{ (= constant) for } k \geq 1$$

First moments are here

$$\mathbb{E}(X) = \frac{1}{p} \text{ and } \mathrm{Var}(X) = \frac{1-p}{p^2}.$$

The Period of Return & Memoryless

A return period, also known as a recurrence interval or repeat interval, is an average time or an estimated average time between events such as earthquakes, floods, landslides, or a river discharge flows to occur

$$\mathbb{E}(X) = \frac{1}{p} \text{ or } p = \frac{1}{\mathbb{E}(X)}$$

A 100-year flood is a flood event that has a 1 in 100 chance (1% probability) of being equaled or exceeded in any given year (wikipedia)

Note that $\mathbb{P}(X \geq h) = (1-p)^h$, then

$$\mathbb{P}(X \geq k+h | X \geq h) = \frac{\mathbb{P}(X \geq k+h)}{\mathbb{P}(X \geq h)} = \frac{(1-p)^{k+h}}{(1-p)^h} = (1-p)^k$$

i.e. $\mathbb{P}(X \geq k)$

The Exponential Distribution

The Exponential $\mathcal{E}(\lambda)$, $\lambda > 0$

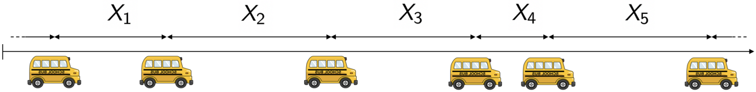
$$\mathbb{P}(X > x) = e^{-\lambda x} \text{ for } x \geq 0$$

with density $\lambda e^{-\lambda x}$. $\mathbb{E}(X) = \dfrac{1}{\lambda}$ and $\text{Var}(X) = \dfrac{1}{\lambda^2}$

Note: if $X \sim \mathcal{E}(\lambda)$, $Y = \lfloor X \rfloor \sim \mathcal{G}(p)$ with $p = 1 - e^{-\lambda}$.

The geometric and the exponential distribution are memoryless

$$\mathbb{P}(X > t + h | X > t) = \mathbb{P}(X > h).$$



The Poisson Distribution

The Poisson distribution $\mathcal{P}(\lambda)$, $\lambda > 0$, has distribution

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \text{ where } k = 0, 1, \cdots$$

Then $\mathbb{E}(X) = \lambda$ and $\text{Var}(X) = \lambda$ $(= \mathbb{E}(X))$.
Further, if $X_1 \sim \mathcal{P}(\lambda_1)$ and $X_2 \sim \mathcal{P}(\lambda_2)$ are independent, then

$$X_1 + X_2 \sim \mathcal{P}(\lambda_1 + \lambda_2)$$

Observe that a recursive equation can be obtained

$$\frac{\mathbb{P}(X = k+1)}{\mathbb{P}(X = k)} = \frac{\lambda}{k+1} \text{ pour } k \geq 1$$

Note: $\mathbb{P}(N = 0) = e^{-\lambda}$, e.g. if $\lambda = 1$, $\mathbb{P}(N = 0) \simeq 36.788\%$
(and $\mathbb{P}(N > 0) \simeq 63.212\%$)

The Poisson Approximation

Let $X_i \sim \mathcal{B}(p)$,

$$\mathbb{P}(X=0) = 1-p \text{ and } \mathbb{P}(X=1) = p.$$

then $X = X_1 + \cdots + X_n \sim \mathcal{B}(n,p)$ (binomial distribution)

$$\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ where } k = 0,1,\cdots,n, \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

If $n \cdot p \simeq \lambda$, $X \simeq \mathcal{P}(\lambda)$

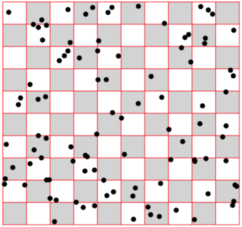
$$\mathbb{P}(X=0) = (1-p)^n \simeq \left(1-\frac{\lambda}{n}\right)^n \simeq e^{-\lambda}$$

	Number of years without catastrophes				
	10	20	50	100	200
10	65.1%	40.1%	18.3%	9.6%	4.9%
20	87.8%	64.2%	33.2%	18.2%	9.5%
50	99.5%	92.3%	63.6%	39.5%	22.5%
100	99.9%	99.4%	86.7%	63.4%	39.5%
200	99.9%	99.9%	98.2%	86.6%	63.3%

The Poisson Distribution

Consider some 10×10 chess-board. Threw $n = 100$ stones on it, and count the number of stones in each square.

```
> data.frame(N,F=table(nb_cell),P=c
    (dpois(0:4,1),1-ppois(4,1)))
   N  F      P
1  0 36  36.78
2  1 39  36.78
3  2 16  18.39
4  3  7   6.13
5  4  2   1.53
6 5+  0   0.37
```

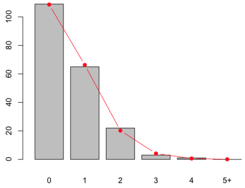


Consider some square, and let X denote the number of

The Poisson Distribution

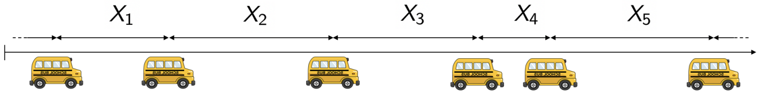
Nombre de soldats de cavaliers morts par ruade de cheval, entre 1875 et 1894, dans 10 corps (soit 200 corps annuels) Bortkiewicz (1898)

```
> data.frame(N,F=table(ruades),P=c(
    dpois(0:4,mean(ruades)),1-ppois
    (4,mean(ruades))))
   N   F      P
1  0 109 108.67
2  1  65  66.21
3  2  22  20.22
4  3   3   4.11
5  4   1   0.63
6 5+   0   0.08
```



The Poisson Process

Consider some bus arrivals, at times T_i, and let $X_i = T_i - T_{i-1}$, assume that $X_i \sim \mathcal{E}(\lambda)$ are independent.



The number of buses in the time interval $[0, t]$ is $N_t \sim \mathcal{P}(\lambda t)$.

See also queuing theory

The Records Process

Let $X_1, X_2, \cdots$ denote some yearly observed value (maximum temperature, etc).

Let $T_1 = \Delta_1 = 1$, $T_n = \Delta_1 + \cdots + \Delta_n$, and

$$\Delta_{n+1} = \min_k \{X_{T_n+k} > X_{T_n}\}$$

If X_t's are independent and identically distributed,

$$\frac{\log T_n - n}{\sqrt{n}} \to \mathcal{N}(0,1)$$

hence $\mathbb{E}(\log T_n) \simeq n$
Note: similarly $\mathbb{E}(\log \Delta_n) \simeq n$