Probability distributions

Contents

1 To do in this document

Description of famous probability distributions (pdf, cdf, mean, variance), if they are a generalization of another one, how they are linked to each other if a link exist, in which case they are useful (Bayesian framework, sociology) or which phenomena are modelled by a certain distribution.
Décrire les propriétés générales des distributions de probabilités classiques, leurs liens entre elles (généralisation, limiting distribution, égalité en distribution), leur interprétation, leur utilisation (Bayesian, conjugate priori,... ou encore dans quel domaine).
Distribution : Bernouilli, Binomial, Multinomial, Negative Binomial, Poisson, Géométrique, Dirichlet, normal, gamma, beta, uniform, t distribution, chi 2, erlang, Fisher

2 Discrete probability distributions

2.1 Bernouilli, Binomial and Multinomial distributions

2.1.1 Bernouilli

`https://fr.wikipedia.org/wiki/Loi_de_Bernoulli`

Description/use : Probability distribution of a random variable which takes the value 1 with probability p and 0 with probability $1-p$. It is used to model an experiment which has only two outcomes (yes/no, 0/1).

Formulas : We note $X \sim \mathrm{B}(p)$ and we have $\mathrm{P}(X=1) = p$ and $\mathrm{P}(X=0) = 1-p$. The probability mass function (**pmf**) can be written as :

$$f(x) = p^x(1-p)^{1-x} \quad \text{for } x \in \{0,1\}$$

Support : $\{0,1\}$

Expected value : $\mathrm{E}(X) = p$

Variance : $\mathrm{V}(X) = p(1-p)$

Links with the binomial distribution : It can be seen as a special case of the binominal distribution where a single trial is conducted $n = 1$.

2.1.2 Binomial

`https://en.wikipedia.org/wiki/Binomial_distribution`

Description / use : It is a distribution which represents the following situation : we have a n independant experiments and for each experiments we have two possibilities (yes/no, 0/1) with probability p to have one outcome (yes or 1 for example) and $1-p$ to have the other (no or 0). The binomial distribution models the number of successes k (i.e number of yes or 1 in our example) after these n independant experiments.

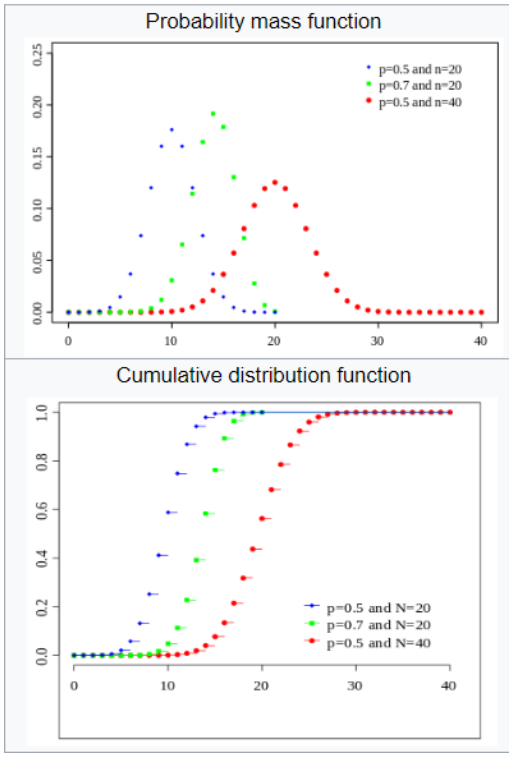
Formulas : For $k = 0, 1, 2, ..., n$, $\mathrm{P}(X = k) = \binom{n}{k} \cdot p^k (1-p)^{n-k}$ (and 0 otherwise)

Interpretation : As we have exactly k successes and n experiments, we have tht term $p^k(1-p)^{n-k}$. And there are exactly $\binom{n}{k}$ ways to have to obtain k successes in n experiments (it can be seen as the number of paths on a probability trees with k times in one direction or the number of subset of cardinal k in a set of cardinal n).

Support : $\{0, 1, ..., n\}$

Expected value : $\mathrm{E}(X) = np$

Variance : $\mathrm{V}(X) = np(1-p)$



Link with the bernouilli distribution : If $Y \sim \mathrm{B}(n, p)$ and $X_i \overset{iid}{\sim} \mathrm{B}(p)$, we have $\sum_{k=1}^{n} X_i \overset{d}{=} Y$

Link with the Poisson distribution : The Poisson is the limiting distribution of $\mathrm{B}(n, p)$ as $n \to \infty$ and $p \to 0$ with $\lambda = np$

2.1.3 Multinomial distribution

`https://en.wikipedia.org/wiki/Multinomial_distribution`

Description/use : It is a generalization of the binomial distribution. Instead of having only two possibilities for each independant experiment, we have k possibilities each associated to a probability p_i and we observe how many times we get the first possibility, the second one,... and the k-th one after n experiments.
Formulas : We need to have $\sum_{i=1}^{k} p_i = 1$. We observe the vector $X = (X_1, ..., X_k)$ where X_i indice the number of times outcome number i is observed over the n trials.

The probability mass function is :

$$f(x_1, ..., x_k; n, p_1, ..., p_k) = \begin{cases} \frac{n!}{x_1!...x_k!} p_1^{x_1} \times ... \times p_k^{x_k} & \text{when } \sum_{i=1}^{k} x_k = n \\ 0 \quad \text{otherwise} \end{cases}$$

when $x_1, ...x_k$ are non negative intergers.

Support : $x_i \in \{0, ..., n\}, \quad i \in \{1, ..., k\}$ with $\sum x_i = n$

Expectation : $\mathrm{E}(X) = (np_1, np_2, ..., np_k)$

Variance :$\mathrm{Var}(X_i) = np_i(1 - p_i)$
$\mathrm{Cov}(X_i, X_j) = -np_ip_j \quad (i \neq j)$

Link with others distributions : When k is 2 and n is 1, it is the Bernouilli distribution and when n is bigger than 1 it is the binomial distribution.
When k is bigger than 2 and n is 1, it is the categorical distribution. In some fields such as NLP, categorical and multinomial distributions are synonymous and it is common to speak of a multinomial distribution when a categorical distribution is actually meant.

2.2 Poisson distribution

`https://en.wikipedia.org/wiki/Poisson_distribution`

Description/use : Poisson distribution expressed the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.

Simple examples :

- Number of mail arriving each day, noticing that they receive 4 letters per day
- Number of phone calls received by a call center per hour
- Number of decay events per second from a radioactive source

An example which violates the Poisson assumptions : number of student arriving at the student union. No constant rate (class time/break), no independence between the arrival of the students (students tend to arrive in groups).

Some applications :

- Telecommunication : telphone calls arriving in a system
- Astronomy : photons arriving at a telescope
- Biology : number of mutation of strand of DNA per unit lenght

Formulas : We say that X has the Poisson distribution with parameter λ, and we write $X \sim \text{Poi}(\lambda)$, if

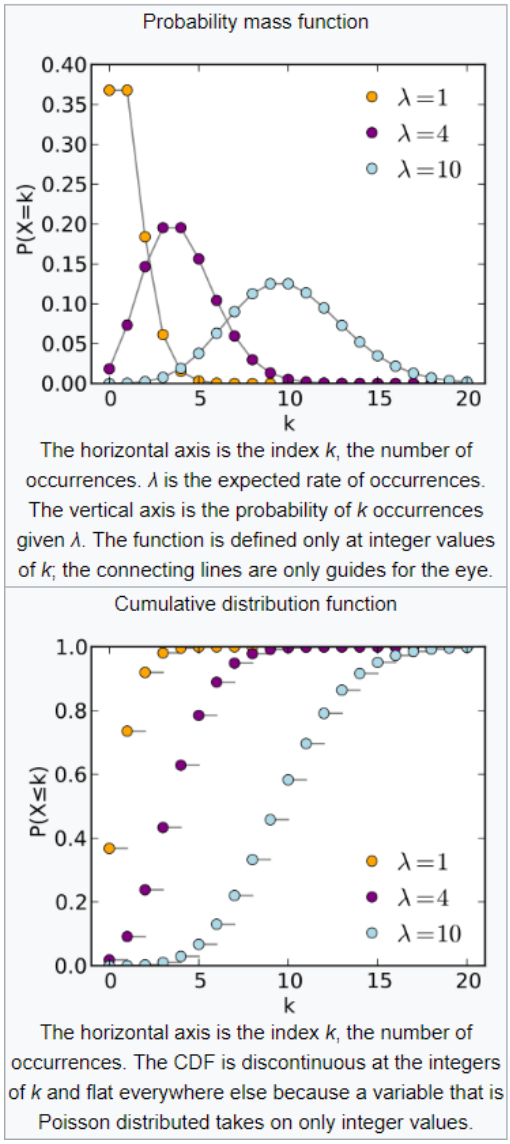
$$\mathbb{P}(X = k) = \begin{cases} \frac{e^{-\lambda}\lambda^k}{k!} & \text{if } k \in \{0, 1, 2, ...\} \\ 0 & \text{otherwise} \end{cases}$$

Support : $\mathbb{Z}^+ = \{0, 1, 2, ...\}$

Expectation : $\mathbb{E}(X) = \lambda$

Variance : $\text{Var}(X) = \lambda$

Bayesian inference : The conjugate prior for the parameter λ of the Poisson distribution (as a likelihood distribution) is the gamma distribution.



2.3 Negative binomial

`https://en.wikipedia.org/wiki/Negative_binomial_distribution`

Description/use : This distribution models the number of failures in a sequence of i.i.d distributed Bernoulli trials before a specified number of successes(denoted r) occurs.
The Pascal distribution and Polya distribution are special cases of the negative binomial distribution. A convention is to use "negative binomial" or "Pascal"

for the of an integer-value stopping-time parameter r and use "Polya" for the real-valued case.

Example : For occurrences of associated discrete events, like tornado outbreaks, the Polya distributions can be used to give more accurate models than the Poisson distribution by allowing the mean and variance to be different, unlike the Poisson.

Formulas : When $X \sim \mathrm{NB}(r,p)$, the probability mass function is :

$$f(k;r,p) = \binom{k+r-1}{k} (1-p)^k p^r$$

for $k \in \{0,1,2,...\}$

Support : $k \in \{0,1,2,...\}$

Expectation : $\mathbb{E}(X) = \frac{r(1-p)}{p}$

Variance : $\mathrm{Var}(X) = \frac{r(1-p)}{p^2}$

Link with the Geometric distribution : The geometric distribution can be seen as a special case with $r = 1$.

Link with gamma-Poisson distribution : By placing a gamma distribution prior with shape r and scale $p/(1-p)$ on λ, a negative binomial (NB) distribution $y \sim \mathrm{NB}(r,p)$ can be generated as $f_Y(y) = \int_0^\infty \mathrm{Pois}\,(y;\lambda)\ \mathrm{Gamma}\left(\lambda; r, \frac{p}{1-p}\right) d\lambda = \frac{\Gamma(r+y)}{y!\Gamma(r)}(1-p)^r p^y$, where $\Gamma(\cdot)$ denotes the gamma function, r is the nonnegative dispersion parameter and p is a probability parameter.
Therefore, the NB distribution is also known as the gamma-Poisson distribution. It has a variance $rp/(1-p)^2$ larger than the mean $rp/(1-p)$, and thus it is usually favored over the Poisson distribution for modeling overdispersed counts.

2.4 Geometric

`https://en.wikipedia.org/wiki/Geometric_distribution`

Description/use : In probability theory and statistics, the geometric distribution is either of two discrete probability distributions :

- The probability distribution of the number X of Bernoulli trials needed to get one success, support on the set $\{1,2,3,...\}$
- The probability distribution of the number $Y = X - 1$ of failures before the first success, supported on the set $\{0,1,2,...\}$

Formulas : The pmf is

$$f(k;p) = \begin{cases} (1-p)^k p & \text{if we choose the first interpretation} \\ (1-p)^{k-1} p & \text{if we choose the second interpretation} \end{cases}$$

Characterization of memoryless : The mathematics characterization of memoryless is

$$\forall (s,t) \in \mathbb{R}_+^2, \quad \mathbb{P}\left(X > t+s \mid X > t\right) = \mathbb{P}\left(X > s\right)$$

Link with the exponential distribution : The geometric and exponential are discrete and continuous analogues. They are the ***unique*** « memoryless » distributions as characterized above.

2.5 Hypergeometric

3 Continuous probability distribution

3.1 Normal distribution

3.1.1 1D

`https://en.wikipedia.org/wiki/Normal_distribution`

Description/use : We say that

3.2 Exponential distribution

`https://en.wikipedia.org/wiki/Exponential_distribution`
`https://fr.wikipedia.org/wiki/Loi_exponentielle#Loi_g%C3%A9om%C3%A9trique`

Description/use : The exponential distribution models the lifetime of memoryless phenomena, i.e the probability that the events lasts at least $s+t$ (hours) given it has already lasted t (hours) is equal to the probability to last s (hours). It is the probability distribution of the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate.
It is a particular case of the gamma distribution. It is the continuous analogue of the geometric distribution, and it has the key property of being memoryless.

Characterization of memoryless : The mathematics characterization of memoryless is

$$\forall (s,t) \in \mathbb{R}_+^2, \quad \mathbb{P}\left(X > t+s \mid X > t\right) = \mathbb{P}\left(X > s\right)$$

This property implies that the random variable follows an exponential distribution (*see Wikipedia in French for the proof*). Conversely the exponential distribution verifies this property.

Link with the geometric distribution : The geometric and exponential are discrete and continuous analogues. They are the ***unique*** « memoryless » distributions as characterized above.

Formulas : We note $X \sim \mathcal{E}(\lambda)$ and the PDF is

$$f(x) = \mathbb{1}\{x \geq 0\}\, \lambda e^{-\lambda x}$$

Support : $x \in [0, \infty)$

Expectation : $\mathrm{E}(X) = \frac{1}{\lambda}$

Variance : $\mathrm{Var}(X) = \frac{1}{\lambda^2}$

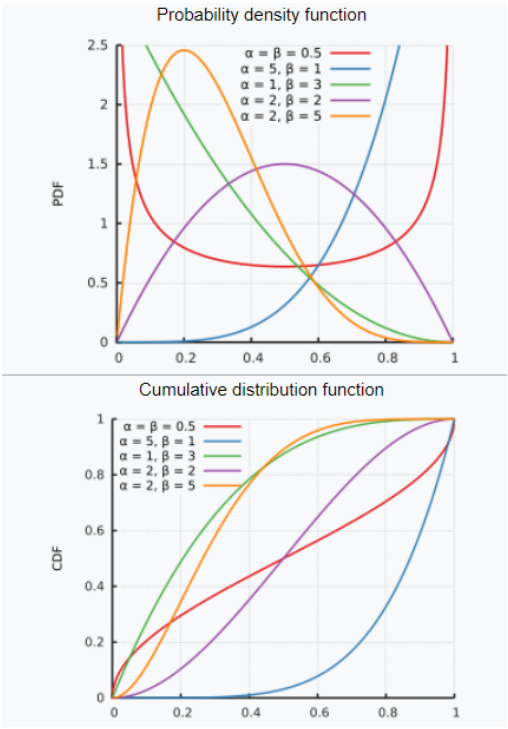
3.3 Gamma distribution

3.4 Beta and Dirichlet distributions

3.4.1 Beta

`https://en.wikipedia.org/wiki/Beta_distribution`

Description/use : Family of continuous probability distributions defined on [0,1]. It has been applied to model the behavior of random variables limited to intervals of finite length in a wide variety of dsiciplines.
In Bayesian inference, the beta distribution is the conjugate prior probability distribution for the Bernoulli, binomial, negative binomial and geometric distributions.



3.4.2 Dirichlet

`https://en.wikipedia.org/wiki/Dirichlet_distribution`
`https://towardsdatascience.com/dirichlet-distribution-a82ab942a879`

Description/use : Family of continuous multivariate probability distributions parametrized by a vector α of postitive reals and denoted $\text{Dir}(\alpha)$. It is a multivariate generalization of the beta distribution (hence it is also called **multivariate beta distribution (MBD)**). They are commonly used as a prior distribution in Bayesian statistics and one reason is that the Dirichlet distribution is the conjugate prior of the categorical and multinomial distribution (it makes the maths a lot easier to have a conjugate prior).

Formulas : For $K \geq 2$ with parameters $\alpha_1, ..., \alpha_K > 0$, the PDF is :

$$f(x_1, x_2, ..., x_K; \alpha_1, \alpha_2, ..., \alpha_k) = \frac{1}{\text{B}(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

where $\{x_k\}_{k=1}^{K}$ belong to the standard $K-1$ simplex i.e $\sum_{i=1}^{K} x_i = 1$ and $x_i \geq 0$ for all $i \in [1, K]$.

The normalization factor is the multavariate beta function $\text{B}(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}$

Support : $x_1, ..., x_k$ where $x_i \in [0, 1]$ and $\sum_{i=1}^{K} x_i = 1$

Link with the Gamma distribution : The Dirichlet distribution can be derived from the Gamma distribution as the Beta distribution.

Special cases : A common special case is the symmetric Dirichlet distribution where $\alpha_1 = \alpha_2 = ... = \alpha_K$. It might be useful, for example, when a Dirichlet prior is used but without prior knowledge favoring one component over another.

Example of use in Bayesian Statistics

$$\begin{aligned}
\boldsymbol{\alpha} &= (\alpha_1, \ldots, \alpha_K) &&= \text{concentration hyperparameter} \\
\mathbf{p} \mid \boldsymbol{\alpha} &= (p_1, \ldots, p_K) &&\sim \text{Dir}(K, \boldsymbol{\alpha}) \\
\mathbb{X} \mid \mathbf{p} &= (\mathbf{x}_1, \ldots, \mathbf{x}_K) &&\sim \text{Cat}(K, \mathbf{p})
\end{aligned}$$

then the following holds:

$$\begin{aligned}
\mathbf{c} &= (c_1, \ldots, c_K) &&= \text{number of occurrences of category } i \\
\mathbf{p} \mid \mathbb{X}, \boldsymbol{\alpha} &\sim \text{Dir}(K, \mathbf{c} + \boldsymbol{\alpha}) &&= \text{Dir}(K, c_1 + \alpha_1, \ldots, c_K + \alpha_K)
\end{aligned}$$

Relation to Dirichlet-multinomial distribution : to be completed

4 Stochastic processes

Poisson process, Brownian Motion / Wiener process

5 Conjugate priors

https://en.wikipedia.org/wiki/Conjugate_prior