Data Science Fundamentals

Part I: Probability theory

ISEP 2nd year
2023-2024

Based on the course given by Nathalie Colin & Jean-Claude Guillerot

Probability theory

➢ **2nd session (October 6th 2023):**

Chapter 3 : REAL-VALUED RANDOM VARIABLE

- 3.1Definition of random variable
- 3.2 Cumulative distribution function (CDF)
- 3.3 Probability density function
- 3.4 Continuous random variable
- 3.5 Discrete random variable

Introduction to real-valued random variables

Until now :

An experiment ⇒ Ω space of all possible outcomes

$$\left.\begin{array}{l} \text{Coin : } \Omega = \{\text{Heads, tails}\} \\ \text{Die : } \Omega = \{1, 2, 3, 4, 5, 6\} \end{array}\right\} \text{Identification of outcomes}$$

NOW : Consider the set of real numbres $\mathbb{R}$

Identify each outcome ω_i and associate to this one a real number:

$$x_i : x_i = X(\omega_i)$$

Interest : A unique support to describe diverse random experiments.

Example 1 :

Experiment : Rolling a poker die: Ω = {Ace , King , Queen, Jack, ten, nine}

To each outcome, we associate a number.

Outcomes : ω_i	**Real variable:** $x_i = X(\omega_i)$
Ace	1
King	2
Queen	3
Jack	4
Ten	5
Nine	6

ω_i = {Queen} $\Leftrightarrow$ $x_i = 3$

Example 2 :

Experiment : Tossing a coin 3 times

Random variable : number of times the outcome is a *tail* (out of 3 trials)

Outcomes : ω_i	**Real variable:** $x_i = X(\omega_i)$
ω_1 = { T T T }	3
ω_2 = { T T H }	2
ω_3 = { T H T }	2
ω_4 = { T H H }	1
ω_5 = { H T T }	2
ω_6 = { H T H }	1
ω_7 = { H H T }	1
ω_8 = { H H H }	0

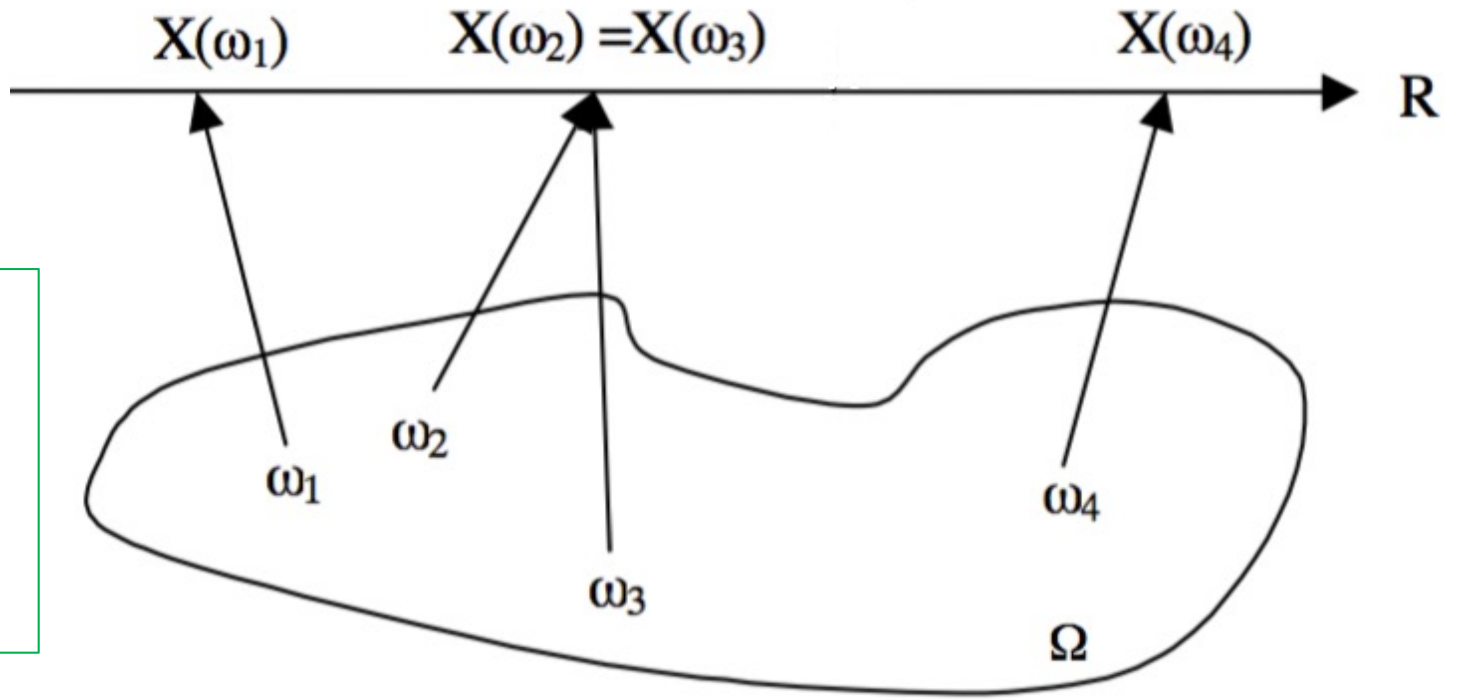
Many outcomes results have the same image by the function X (.)

Definition of real-valued random variable*:

A function defined on a probability space that maps from the sample space Ω to the set of real numbers.

$$\text{function } X(.) : \Omega \rightarrow \mathbb{R}$$

We project the elements of Ω onto R by the function X(.) :



Reminder:

Real function:

All outcomes have exactly one image. 2 results can have the same image, the opposite is not true.

*usually abreviated r. v.

How to associate a probability to a random variable?

Given the fact that X (.) is a mapping from Ω to $\mathbb{R}$, we can match the subsets of $\mathbb{R}$ to the subsets of Ω to find out the probability.

For example:

- Let A be a subset of Ω ($A \subset \Omega$) with probability P(A),
- Let X(A) be the set of x that are the image of A by the function X(.) in $\mathbb{R}$.
- Then, the probability of X(A) will be:

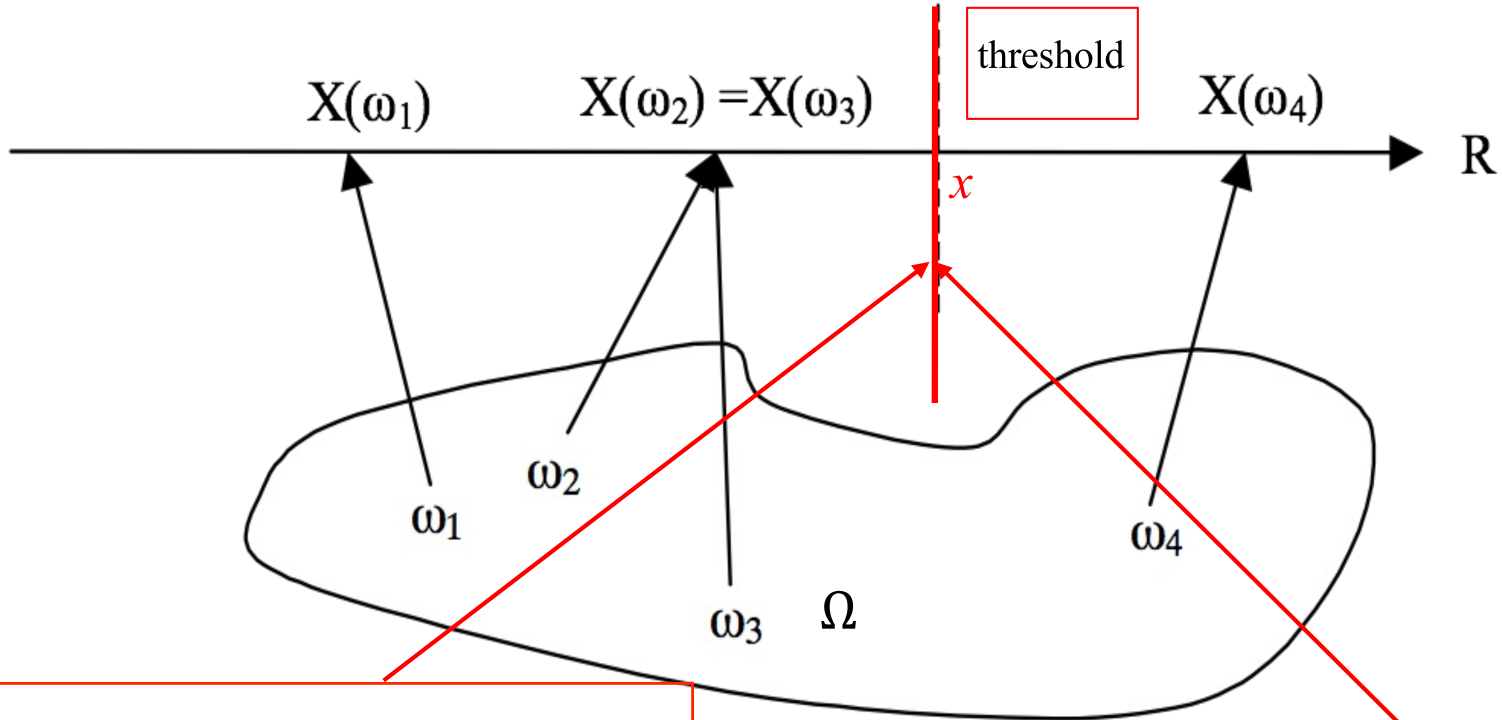
$$\boxed{P(X(A))=P(\{\omega \text{ such that } X(\omega) \in X(A)\})}$$

illustration

In $\mathbb{R}$, let the subset {X<x} correspond to the image of all the outcomes ω such that their image by X(.) is less than the threshold x:

<u>In $\mathbb{R}$</u> <u>In Ω</u>

$$\{X < x\} = \{\omega | X(\omega) < x\}$$



The condition {X<x} corresponds to subset $\{\omega_1, \omega_2, \omega_3\}$ de Ω

P(X<x)= P(ω I X(ω) < x)

General definition of real-valued random variable :

A random variable X is a function X(.):

$X : \Omega \longrightarrow \mathbb{R}$ such that

a) The set of points ω that satisfy the condition $\{X(\omega) < x\}$ and denoted $\{X < x\}$ constitutes an event for all x.

b) The probability of the events $\{X=+\infty\}$ and $\{X=-\infty\}$ is null.

In practice : X is a random variable

if $P(X<x)$ $\forall$ x real is known

Remark : Probability of an interval $C = \{x_1 \leq X < x_2\}$

X being a real random variable, consider the events A and B:

$$A = \{X < x_1\} \qquad B = \{X < x_2\}$$

B=A∪C= $\{ X < x_1 \} \cup \{ x_1 \leq X < x_2 \}$

A and C being disjoint, by the axiom 3 (additivity) we obtain:

$$\boxed{P(\{ x_1 \leq X < x_2 \}) = P(B) - P(A) = P(\{ X < x_2 \}) - P(\{ X < x_1 \})}$$

Cumulative distribution function (CDF)

Definition of Cumulative distribution function (CDF) :

The cumulative distribution function (CDF) of a real random variable

X is the probability of the event $\{ X \leq x \}$, denoted:

$$\boxed{F_X(x) = P(\{ X \leqslant x \})}$$

Notation :

F : Cumulative distribution function

X : Random variable (subscript of F)

x : real threshold

P : Probability

Example :

Experiment : Tossing a coin 3 times.

Random variable (r. v.) X : Number of times the outcome is *tail* out of 3 trials.

$$P(TTT)=P(THT)=\ldots\ldots.=P(HHH) = 1/8$$

Values taken by X and the corresponding probability:

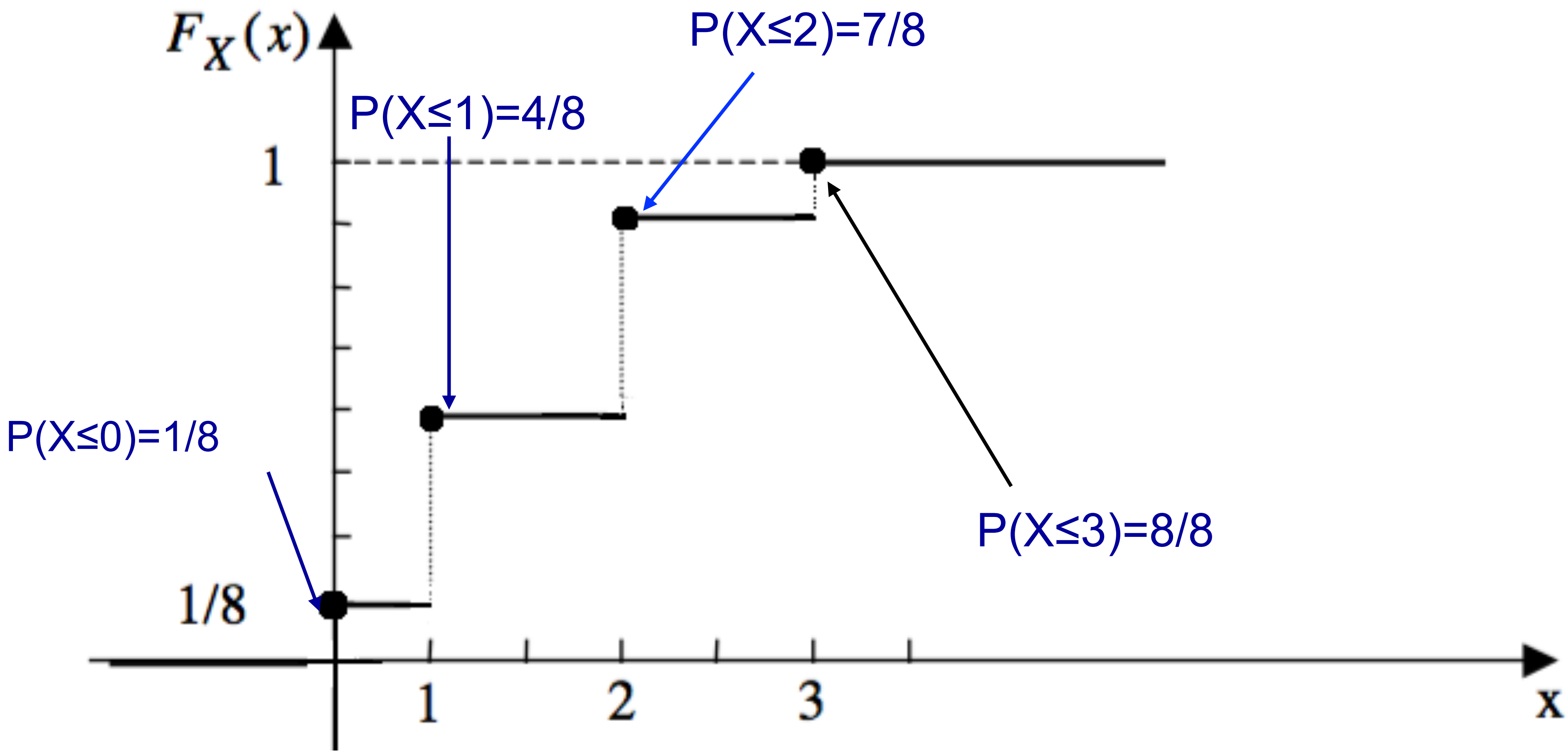
$P(X = 0) = P(HHH) = 1/8$

$P(X = 1) = P(HHT \text{ or } HTH \text{ or } THH) = 3/8$

$P(X = 2) = P(HTT \text{ or } THT \text{ or } TTH) = 3/8$

$P(X = 3) = P(TTT) = 1/8$

Figure: Cumulative Distribution function (CDF)



threshold x

Properties of the cumulative distribution function (CDF) :

The cumulative distribution function verifies the following properties :

a) It is bounded and normalized $0 \leq F_X(x) \leq 1$

b) It is monotonically non-decreasing $F_X(x+\varepsilon) \geq F_X(x), \varepsilon \geq 0$

c) It is right continuous (the limit of F(x) when x approaches x_0 from the right (values greater than x_0) is F(x_0).

d) The CDF's limits are :

$$\lim_{x \to -\infty} F_X(x) = 0 \quad \lim_{x \to +\infty} F_X(x) = 1$$

Probability density function :

$$\boxed{f_X(x) = \frac{dF_X(x)}{dx}}$$

Notation :

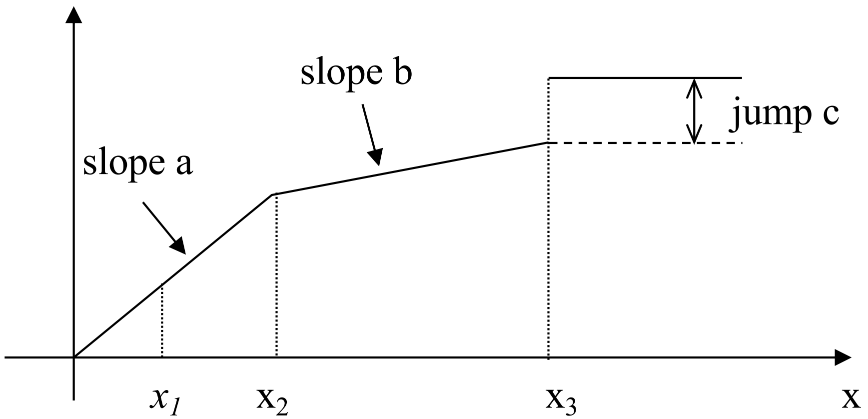
f : probability density function

X : random variable (subscript of f)

x : real threshold

d./dx : derivative with respect to x (threshold)

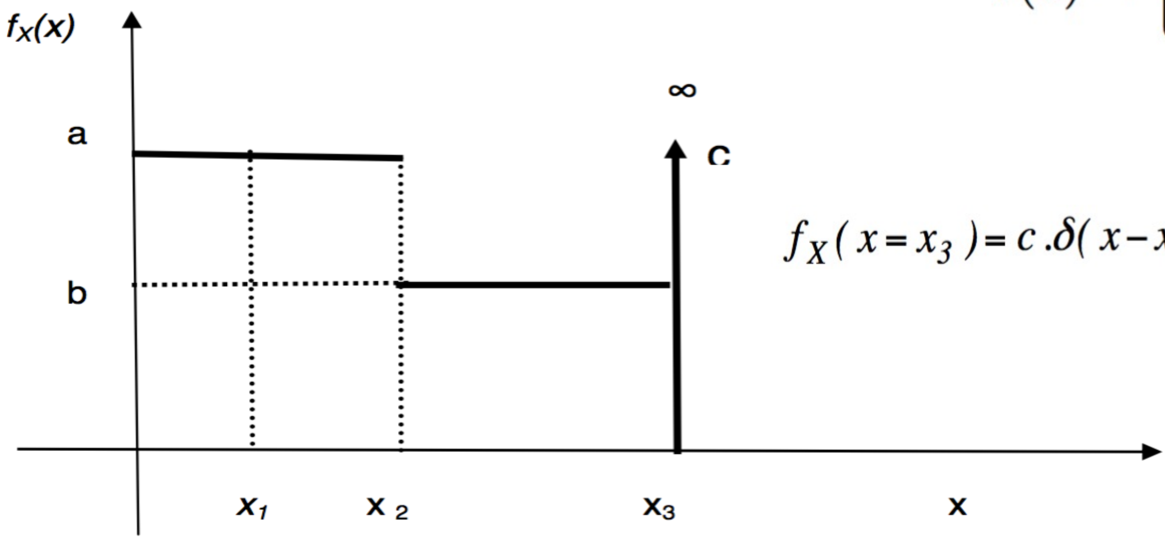
Example, of discontinuous cumulative distribution function :



Reminder:

$$\delta(x) = \begin{cases} 0, & x \neq 0 \\ +\infty, & x = 0 \end{cases}$$

The density:



$$f_X(x = x_3) = c\,.\delta(x - x_3)$$

Properties of the probability density function

1) The area under the curve is normalized : $\boxed{F_X(x) = \int_{-\infty}^{x} f_X(u)du}$

and in particular

$$\boxed{F_X(+\infty) = \int_{-\infty}^{+\infty} f_X(u)\,du = 1}$$

2) The probability density function is nonnegative everywhere : $\boxed{f_X(x) \geq 0}$

(because it is the derivative of a monotonically non-decreasing function)

Remark : If a function f(x) is nonnegative and its integral is equal to 1, then, it can be considered to the probability density function of a random variable.

Continuous random variables

Definition: A random variable is continuous if its cumulative distribution function is continuous everywhere.

Interpretation of the density of a continuous random variable

$$F_X(x_2) - F_X(x_1) = P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f_X(x)dx$$

If $x_1 = x$ and $x_2 = x + dx$ we obtain: $P(x < X \leq x + dx) = f_X(x)dx$

Then, $f_X(x)dx$ can be seen as the probability that the variable X belongs to the interval $[x, x+dx]$.

The density can be interpreted as the factor of proportionality between: the probability of belongging to a given interval and the length of this interval.
So:

- $f_X(x)$ is not a probability.
- $P(X = x) = 0$ (for a continuous r.v.)

Discrete random variables

X : a random variable that can take on either a finite or at most a countably infinite set of discrete values (for example, the set of integer numbers).

For each value x_i we have:

$$P(X = x_i) = p_i \neq 0 \quad \text{and} \quad \sum_i p_i = 1$$

The cumulative distribution function is then discontinuous and has a staircase-like appearance

$$\boxed{F_X(x) = \sum_i P(X = x_i).H(x - x_i)}$$

$$\boxed{f_X(x) = \sum_i P_i \,.\delta(x - x_i)}$$

Where:

- H (x) : Heaviside step function $\forall x \in \mathbb{R},\ H(x) = \begin{cases} 0 & \text{si} \quad x < 0 \\ 1 & \text{si} \quad x \geq 0. \end{cases}$
- δ (x) : Dirac delta function $\delta(x) = \begin{cases} 0, & x \neq 0 \\ +\infty, & x = 0 \end{cases}$

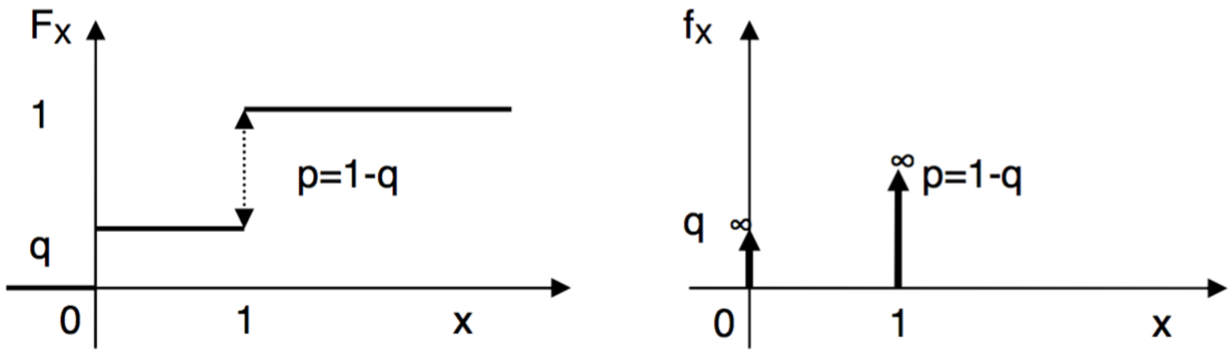
Example: tossing a coin

Two outcomes $\begin{cases} P(tail) = p \\ P(head) = q \end{cases}$ with $p + q = 1$

Random variable (indicator) X: $\Omega \rightarrow \mathbb{R}$ $\begin{cases} X(tail) = 1 \\ X(head) = 0 \end{cases}$

Cumulative distribution function $F_X(x) = q.H(x) + p.H(x-1)$

Probability density function $\boxed{f_X(x) = q.\delta(x) + p.\delta(x-1)}$



The probability varies between *0 and 1.* The density tends to infinity. The limit is a dirac–delta function.

Examples of well-known discrete probability distributions

- Bernoulli distribution: X takes on 2 values: 0 and 1

$$P\{X=1\}=p \qquad P\{X=0\}=q=1-p$$

- Binomial distribution with parameters n and p :
  - X: number of successes in a sequence of n Bernoulli experiments.
  - X takes on integer values X: 0,1,…, n

$$0 \le k \le n \quad P\{X=k\}=C_n^k p^k q^{n-k} \quad \text{with } p+q=1$$

- Poisson distribution with parameter λ ≥ 0

X: number of events occurring in a fixed interval of time (or space).

Assumptions : the events occur at a mean rate **λ** and are independent of the time since the last event.

X takes on integer values X: 0,1,…, example: number of calls per hour.

$$k \ge 0 \quad P(X=k)=e^{-\lambda}\cdot\frac{\lambda^k}{k!}$$

Examples of well-known continuous probability distributions (1/4)

- Uniform distribution with parameters a and b :

  X takes on values in the interval [a,b]

$$f_X(x) = \frac{1}{b-a} \quad \text{if } a \leq x \leq b$$

$$f_X(x) = 0 \text{ elsewhere}$$

- Exponential distribution with parameter $\boldsymbol{\lambda}$:

  X: time between events in a Poisson process. Example: time between two calls.

$$f_X(x) = \lambda e^{-\lambda x} \qquad \lambda > 0; x \geq 0$$

Cauchy distribution with x_0 parameter of position and α >0 parameter of scale

$$f(x) = \frac{1}{\pi\alpha\left[1+\left(\frac{x-x_0}{\alpha}\right)^2\right]}; \quad x \in \mathbb{R}$$

Examples of well-known continuous probability distributions (2/4)

- Normal distribution (also called Laplace-Gauss distribution, or simply Gaussian distribution) with parameters $m \in \mathbb{R}$, $\sigma > 0$.

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \boldsymbol{e}^{-\frac{(\boldsymbol{x}-\boldsymbol{m})^2}{2\sigma^2}} \quad \text{with } x \in \mathbb{R}$$

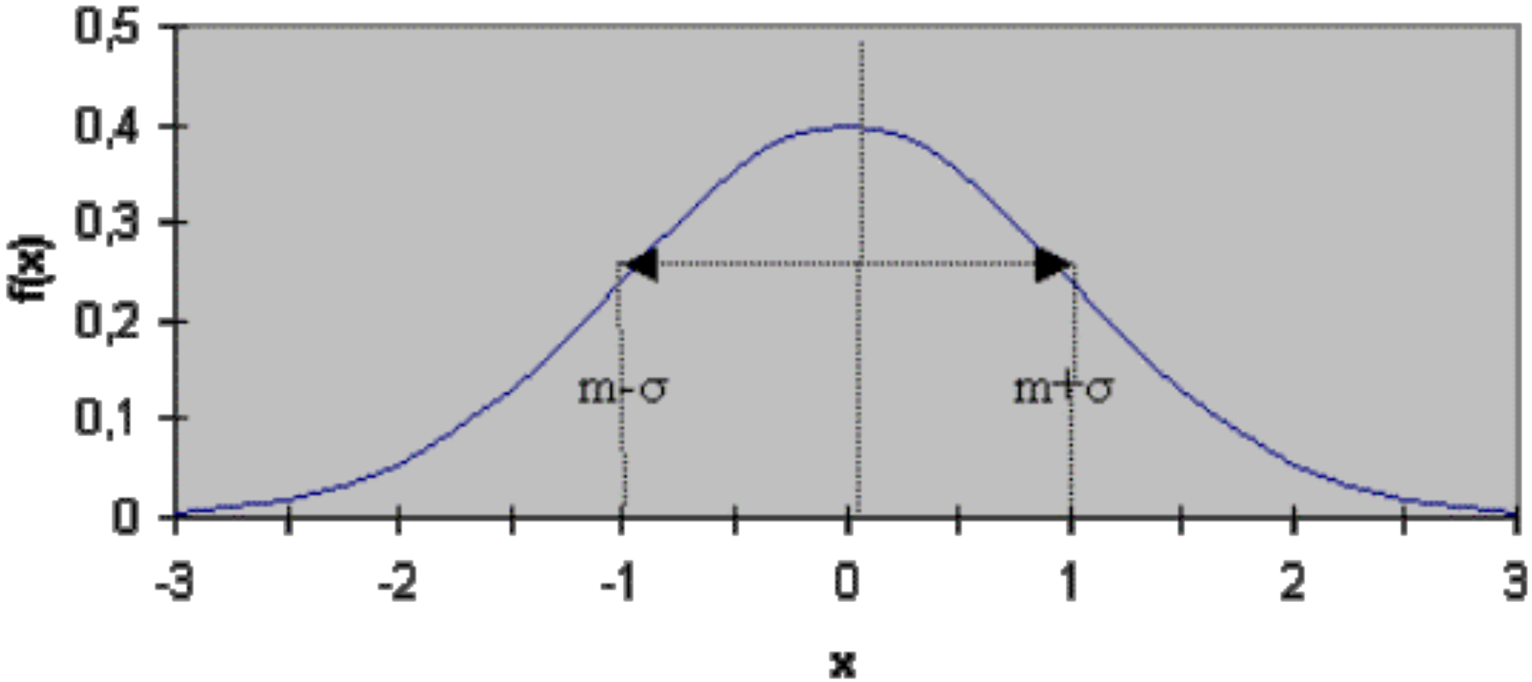
- We denote $X \sim N(m, \sigma^2)$
- If $m = 0$, X is a centered random variable
- If $m = 0$ and $\sigma = 1$ the distribution is called **normal standard:**

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \boldsymbol{e}^{-\frac{\boldsymbol{x}^2}{2}}$$

Examples of well-known continuous probability distributions (3/4)

Probability density of the Standard normal distribution

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



Examples of well-known continuous probability distributions (4/4)

Table of the Cumulative Distribution Function CDF $\boldsymbol{F}(.)$ of the *Standard normal distribution.*

$$\boldsymbol{Z \sim N(0,1)}$$

$$\boldsymbol{F(z) = \int_{-\infty}^{z} e^{-\frac{u^2}{2}} du}$$

Example $P(Z \leq 1.96) = 0{,}975$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

How to read the table of CDF of the Standard Gaussian distribution?

First, standardize the variable X :

If $X \sim N(m, \sigma^2)$, then $Z = \frac{X-m}{\sigma} \sim N(0,1)$.

Next, calculate $F_X(x) = (P(X \leq x) = P(Z \leq z) = F_Z(z)$ where $z = \frac{x-m}{\sigma}$

If z = 0,12 ; we have : F(0,12) =0,5478

If z = 0,05 ; we have : F F(0,05) =0,5199

Hundredth of z

Ones and tenths of z

	0	1	2	3	4	5	6
0,0						F(0,05)	
0,1			0,5478				
0,2					F(z)		
0,3							
0,4							