MAE 206 : Information Engineering : Basics on probability and statistics

Meryem Benammar

ISAE-Supaéro, DEOS

Where is Information ? ...

Where is information contained ?

- Numerical : digital data, text, DNA string,
- Audio : music (digital or analog), probe signals (rovers)
- Image : camera images, snapshots, paintings
- Video : high motion videos, live streams,

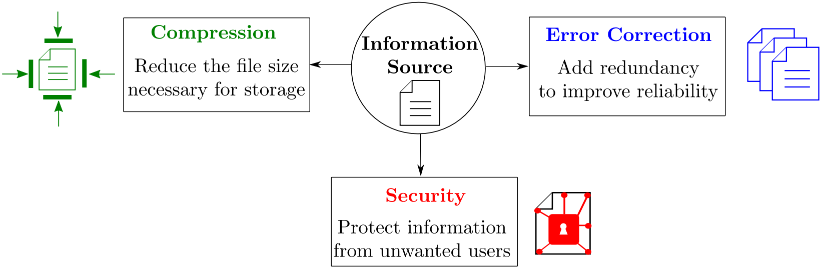
FIGURE – Source : icon-icons.com

Types of information : Semantic or quantitative

We focus on quantitative information !

What is Information Engineering (IE) ?

Three families of algorithms



Mathematics behind IE

Where there is uncertainty $\Rightarrow$ there is information !

Randomness (uncertainty) :
Probabilities and statistics

FIGURE – Source : www.sci-highs.com/

Information measurement :
Information theory !

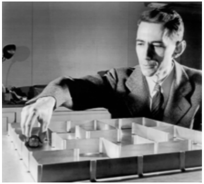
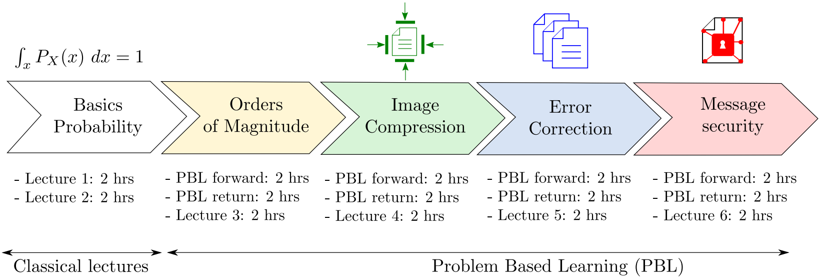


FIGURE – Source : www.rfi.fr/

In this course ...



Problem Based Learning :

- Mini-projects of 5 to 6 students
- PBL forward and return sessions
- Autonomy, rigor, organization, ...

In this course ...

- List and identify basic probability distributions (discrete and continuous)
- Master the operations on probability distributions
- List and analyze basic results from statistics
- Identify the notion of information to a measure of randomness
- Describe basic information engineering tools
- Evaluate information engineering tools from a probabilistic point of view
- Criticize and suggest improvements to these tools

Context

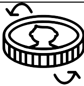
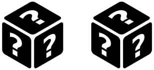
Probability spaces

Discrete random variables and vectors

Continuous random variables and vectors

Experiments

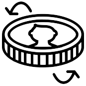
- Consider an *experiment* for which the output is unknown
- Output belongs to an *alphabet* set Ω :

Experiment	Alphabet Ω
	$\Omega = \{H, T\}$
	$\Omega = \{1, 2, 3, 4, 5, 6\}$
	$\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$

Events as sets : definition

Definition (Alphabet and Events)

- The set of possible outcomes Ω is called an Alphabet
- An event E is a subset of Ω : $E \subseteq \Omega$
- An event space $\mathcal{E}$ is the set of all possible events

Experiment	Alphabet Ω	Examples of events E
	$\Omega = \{1, 2, 3, 4, 5, 6\}$	• Result is 1 : $E_1 = \{1\}$ • Result is even : $E_2 = \{2, 4, 6\}$ • Result is ≥ 3 : $E_3 = \{3, 4, 5, 6\}$ • Result is ≤ 6 : $E_4 = \{1, 2, 3, 4, 5, 6\}$
	$\Omega = \{H, T\}$	• Result is H : $E_1 = \{H\}$ • Result is T : $E_2 = \{T\}$ • Result is whatever : $E_3 = \{H, T\}$ • Coin was lost : $E_4 = \{\emptyset\}$

Events as sets : axioms

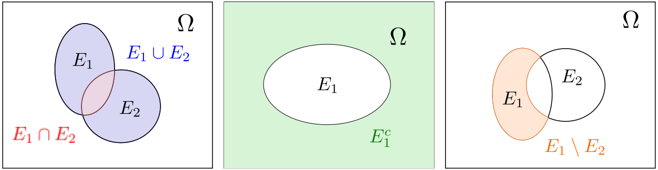


FIGURE – Set and event operations

Operation on event	Operation on sets
Event 1 and Event 2	$E_1 \cap E_2$
Event 1 or Event 2	$E_1 \cup E_2$
Not event 2	$E_2^c = \Omega \setminus E_2$
Event 1 except Event 2	$E_1 \setminus E_2$
Null event	$\emptyset$
Trivial event	Ω

TABLE – Set axioms

Probability measure

Definition (Probability measure)

Let Ω be an alphabet, $\mathcal{E}$ be its event space. A probability measure $\mathbb{P}$ is defined as

$$\begin{aligned} \mathbb{P} : \mathcal{E} &\rightarrow [0,1] \\ E &\rightarrow \mathbb{P}(E) \end{aligned}$$

where $\mathbb{P}$ verifies the following axioms

- Exhaustivity : $\mathbb{P}(\Omega) = \sum_{w \in \Omega} \mathbb{P}(w) = 1$
- Additivity : if two events E_1 and E_2 are disjoint, then

$$E_1 \cap E_2 = \emptyset \qquad \Rightarrow \qquad \mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2).$$

Value $w \in \Omega$	H	T
$\mathbb{P}(w)$	0.5	0.5

$\Rightarrow$

Event $E \subset \Omega$	$\{H\}$	$\{T\}$	$\{H,T\}$	$\emptyset$
$\mathbb{P}(E)$	0.5	0.5	1	0

Table – Example : Fair coin toss

Probability measure : axioms

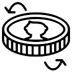
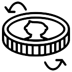
Event complement	$\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$
Union of events	$\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2)$
Exustivity	$\sum_{w \in \Omega} \mathbb{P}(w) = 1$
Null event	$\mathbb{P}(\emptyset) = 0$
Elementary measures	$\mathbb{P}(E) = \sum_{w \in E} \mathbb{P}(w)$
Subsets	$E_1 \subset E_2 \quad \Rightarrow \quad \mathbb{P}(E_1) \leq \mathbb{P}(E_2)$

Probability space

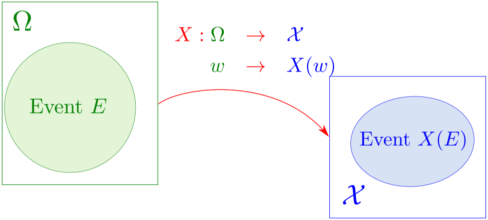
Definition (Probability space)

A probability space $\mathcal{P} = (\Omega, \mathcal{E}, \mathbb{P})$ is defined by three components

- An alphabet set Ω representing a set of all possible outputs
- An event set $\mathcal{E}$ consists in all possible subsets of the alphabet Ω
- A probability measure $\mathbb{P}$

Experiment	Probability measure $\mathbb{P}$	Probability space
Fair coin flip 	• $\mathbb{P}_1(H) = 0.5$ • $\mathbb{P}_1(T) = 0.5$	$(\{H, T\}, \mathcal{P}(\{H, T\}), \mathbb{P}_1)$
Rigged coin flip 	• $\mathbb{P}_2(H) = 0.7$ • $\mathbb{P}_2(T) = 0.3$	$(\{H, T\}, \mathcal{P}(\{H, T\}), \mathbb{P}_2)$

Random variables



Definition

Let us the a probability space $\mathcal{P} = (\Omega, \mathcal{E}, \mathbb{P})$ and $(\mathcal{X}, \mathcal{E}_X)$ a new sample. A random variable X is a mapping from Ω to $\mathcal{X}$

$$\begin{aligned} X : \quad \Omega &\rightarrow \mathcal{X} \\ w &\rightarrow X(w). \end{aligned}$$

- The probability measure $\mathbb{P}_X$ is defined, for all events E_x in $\mathcal{E}_X$

$$\mathbb{P}_X(E_x) = \mathbb{P}(X^{-1}(E_X)) = \mathbb{P}(\{w \in \Omega, X(w) \in E_x\}$$

- The probability space associated with X is $\mathcal{P}_X = (\mathcal{X}, \mathcal{E}_X, \mathbb{P}_X)$.

Random variables : examples

Consider a fair dice throw with the probability space :

$$\Omega = \{1, ..., 6\} \quad , \quad \mathcal{E} = \mathcal{P}(\Omega) \quad , \quad \mathbb{P} = \left\{\frac{1}{6}, \ldots, \frac{1}{6}\right\}$$

We can define a variety of random variables $X : w \to X(w)$

Event w	Random variable $X(w)$	Alphabet $\mathcal{X}$	Probability $\mathbb{P}_X$
Value w	Output w itself : $X(w) = w$	$\mathcal{X} = \Omega$	$\mathbb{P}_X = \left\{\frac{1}{6}, \ldots, \frac{1}{6}\right\}$
Value w	Parity (even/odd) : $X(w) = w[2]$	$\mathcal{X} = \{0, 1\}$	$\mathbb{P}_X = \{0.5, 0.5\}$
Value w	Threshold value : $X(w) = (w \geq 3)?$	$\mathcal{X} = \{0, 1\}$	$\mathbb{P}_X = \{0.33, 0.666\}$
Value w	Square : $X(w) = w^2$	$\mathcal{X} = \Omega^2$	$\mathbb{P}_X = \left\{\frac{1}{6}, \ldots, \frac{1}{6}\right\}$

Context

Probability spaces

Discrete random variables and vectors

Continuous random variables and vectors

Probability mass function

Definition (Probability mass function (pmf))

The probability mass function of a discrete random variable X with associated probability measure $\mathbb{P}_X$ is defined by

$$\begin{array}{rccl} P_X : & \mathcal{X} & \to & [0:1] \\ & x & \to & P_X(x) = \mathbb{P}_X(X = x) = \mathbb{P}_X(\{x\}) \end{array}$$

We have by definition of the pmf that

$$\sum_{x \in \mathcal{X}} P_X(x) = 1.$$

$\boxed{\text{The probability mass function } P_X(\cdot) \text{ RULES them all!}}$

Classical discrete random variables

1. Bernoulli of parameter p ($\mathrm{Bern}(p)$) : Coin toss with Head probability p

2. Discrete uniform over the interval $[1:K]$ ($\mathrm{Unif}\left([1:K]\right)$) : Fair dice throw

3. Binomial with parameter (n,p) ($\mathrm{Binom}(n,p)$) : number of heads in n coin flips

4. Constant variable equal to K ($\mathrm{Const}(K)$) : stale dice at value K

Moments : expectation and variance

Definition (Moments)

To each random variable X with pmf P_X are associated

- An expected (average) value $\mathbb{E}(X)$ (first order moment)

$$\mathbb{E}(X) \triangleq \sum_{x \in \mathcal{X}} x.P_X(x)$$

- A variance (squared standard deviation) $\mathbb{V}(X)$ (second order moment)

$$\mathbb{V}(X) \triangleq \mathbb{E}\left(X^2\right) - \mathbb{E}^2(X).$$

Examples :

Random law	Expectation $\mathbb{E}(X)$	Variance $\mathbb{V}(X)$
Bern(p)	p	$p(1-p)$
Unif ($[1 : K]$)	$\frac{K+1}{2}$	$\frac{K^2-1}{2}$
Const(K)	K	0

Pair of random variables : joint pmf

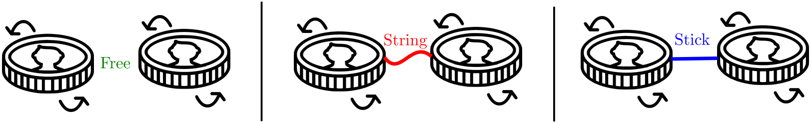
Let (X, Y) be a pair of random variables resulting from a joint experiment.

Definition (Joint pmf)

The joint pmf $P_{X,Y}$ associated with the pair (X, Y) is given by

$$\begin{array}{rcl} \mathcal{X} \times \mathcal{Y} & \rightarrow & [0,1] \\ (x, y) & \rightarrow & P_{X,Y}(x,y) = \mathbb{P}(X = x \text{ and } Y = y) \end{array}$$

Examples : Tossing two coins



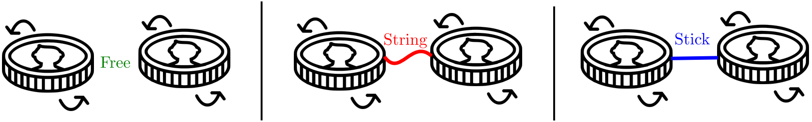
Pair of random variables : marginal pmfs

Definition (Marginal pmfs)

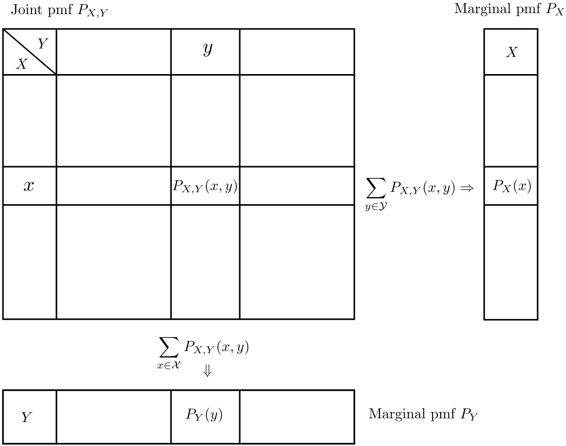
To the joint pmf $P_{X,Y}$ are associated two marginal pmfs P_X and P_Y defined by

$$P_X(x) = \sum_{y \in \mathcal{Y}} P_{X,Y}(x,y) \ , \ P_Y(y) = \sum_{x \in \mathcal{X}} P_{X,Y}(x,y)$$

Examples : Tossing two coins



Pair of random variables : joint and marginal pmfs



$$\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y) = 1 \text{ and } \sum_{x\in\mathcal{X}} P_X(x) = \sum_{y\in\mathcal{Y}} P_Y(y) = 1$$

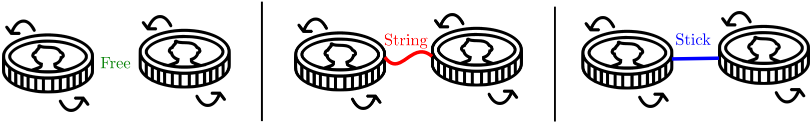
Pair of random variables : conditional pmfs

Definition (Conditional pmfs)

The conditional pmfs associated with $P_{X,Y}$ are defined by

$$
\begin{aligned}
P_{X|Y}(x|y) &= \frac{P_{X,Y}(x,y)}{P_Y(y)} = \mathbb{P}(X=x|Y=y) \\
P_{Y|X}(y|x) &= \frac{P_{X,Y}(x,y)}{P_X(x)} = \mathbb{P}(Y=y|X=x)
\end{aligned}
$$

Examples : Tossing two coins



Pairs of random variables : conditional pmfs

Conditional pmf $P_{X|Y}$

$X \backslash Y$		y	
x		$P_{X\|Y}(x\|y)$	

$\Downarrow$

$$\sum_{x \in \mathcal{X}} P_{X|Y}(x|y) = 1$$

Bayes' formula

Definition

Let (X, Y) be a pair of random variables with joint pmf $P_{X,Y}$. Assume that we only know P_X and $P_{Y|X}$, then Bayes' formulae write as

$$P_{X|Y}(x|y) = \frac{P_X(x)P_{Y|X}(y|x)}{\sum_{x'} P_X(x')P_{Y|X}(y|x')}$$

Very useful in machine learning, signal processing (estimation, detection), communication engineering....

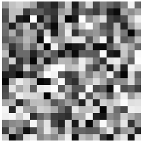
The joint pmf $P_{X,Y}$ RULES them all!

Random vectors

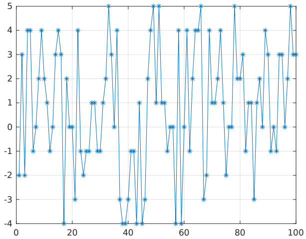
- A binary stream of $n = 10$ bits iid Bern(0.5)

$$U^k = (0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1)$$

- A random image of $n = 20 \times 20$ gray-level iid pixels



- $n = 100$ consecutive iid samples of a random noise $Unif[-5:5]$



Random vectors : joint pmf

Generalize to a vector of random variables $(X_1, ..., X_n)$.

Definition (Joint and marginal pmfs)

The joint pmf of the vector can be defined as $P_{X_1,\ldots,X_n}$

$$\begin{array}{ccc} \mathcal{X}_1 \times \ldots \mathcal{X}_n & \to & [0,1] \\ (x_1, \ldots, x_n) & \to & P_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \end{array}$$

To the joint pmf $P_{X_1,\ldots,X_n}$ are associated n marginal pdfs

$$P_{X_i}(x_i) = \sum_{(x_1,\ldots,x_n)\backslash x_i} P_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$$

Example : Throwing a dice n consecutive times

Random vectors : chain rule

Definition (The chain rule)

The joint pmf can be expanded using the so-called chain rule as follows

$$
\begin{aligned}
P_{X_1,\ldots,X_n}(x_1,\ldots,x_n) &= \prod_{i=1}^{n} P_{X_i|X_1,\ldots,X_{i-1}}(x_i|x_1,\ldots,x_{i-1}) \\
&= \prod_{i=1}^{n} \frac{P_{X_1,\ldots,X_i}(x_1,\ldots,x_i)}{P_{X_1,\ldots,X_{i-1}}(x_1,\ldots,x_{i-1})}
\end{aligned}
$$

Example :

1. A set of variables $(X_1,\ldots,X_n)$ are pairwise independent

$$
P_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} P_{X_i}(x_i),
$$

2. If, further, the variables are identically distributed (iid), i.e., they follow the same law P_X, then

$$
P_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} P_X(x_i).
$$

Probability density function : pdf

A continuous random variables X takes values over an interval $\mathcal{X}$ in $\mathbb{R}$.

Definition (Probability density function)

Let X be a continous random variable ($X : \omega \to \mathbb{R}$), then the associated probability density function is denoted by f_X and verifies the following properties :

- Positivity :
$$\forall x \in \mathbb{R}, \quad f_X(x) \geq 0$$
- Exhaustivity :
$$\int_{x \in \mathcal{X}} f_X(x)\ dx = 1$$
- Interval probability :
$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\ dx$$

Moments : expectation, variance, ...

Definition (Moments)

To each random variable X with pmf P_X / pdf f_X are associated

- An expected value $\mathbb{E}(X)$ (first order moment)

$$\int_{x \in \mathcal{X}} x.f_X(x)\, dx.$$

- A variance $\mathbb{V}(X)$ (second order moment)

$$\mathbb{V}(X) \triangleq \mathbb{E}\left(X^2\right) - \mathbb{E}^2(X).$$

Moments : expectation, variance, ...

Properties (Moments)

1. *The expectation is linear, i.e.,*

$$\mathbb{E}(f(X)) = f\left(\mathbb{E}(X)\right)$$

for any linear transformation f.

2. *For any constant α, we have that*

$$\mathbb{V}(\alpha.X) = \alpha^2\mathbb{V}(X).$$

3. *For any constant c, we have that*

$$\mathbb{V}(X + c) = \mathbb{V}(X).$$

Cumulative distribution function (cdf)

Definition (Cumulative distribution function (cdf))

Let X be a continuous random variable with associated pdf f_X. Then, to X is associated a cumulative distribution function (cdf) F_X defined by

$$F_X(a) = \mathbb{P}(X \leq a) = \int_{-\infty}^{a} f_X(x)\, dx$$

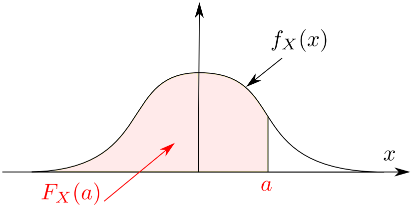
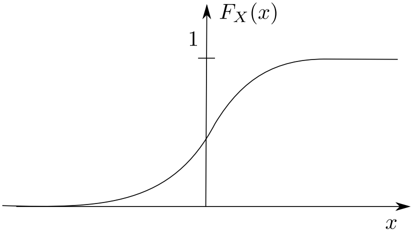


Figure – Cumulative distribution function : integration

Cumulative distribution function (cdf)

Properties (Cumulative distribution function (cdf))

1. *Exhaustivity*
$$F_X(\infty) = 1$$
2. *Increasing function*
$$a_1 \leq a_2 \Rightarrow F_X(a_1) \leq F_X(a_2)$$
3. *Positivity*
$$F_X(-\infty) = 0$$
4. *Relation to pdf*
$$f_X(x) = \frac{\partial F_X}{\partial x}(x)$$



Continuous random variables : examples

- Uniform over an interval $[a : b]$ ($\sim$ Unif $[a, b]$)

- Gaussian with mean μ and variance σ^2 ($\sim \mathcal{N}(\mu, \sigma^2)$)

Continuous random vectors

Definition (Joint and marginal pds)

The joint pdf of the vector can be defined as $f_{X_1,\ldots,X_n}$

$$\begin{array}{ccc} \mathcal{X}_1 \times \ldots \mathcal{X}_n & \to & [0,1] \\ (x_1,\ldots,x_n) & \to & f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) \end{array}$$

To the joint pdf $f_{X_1,\ldots,X_n}$ are associated n marginal pdfs

$$f_{X_i}(x_i) = \int_{(x_1,\ldots,x_n)\setminus x_i} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\, dx_1 \ldots dx_n$$

Chain rule

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) &= \prod_{i=1}^{n} f_{X_i|X_1,\ldots,X_{i-1}}(x_i|x_1,\ldots,x_{i-1}) \\ &= \prod_{i=1}^{n} \frac{f_{X_1,\ldots,X_i}(x_1,\ldots,x_i)}{f_{X_1,\ldots,X_{i-1}}(x_1,\ldots,x_{i-1})} \end{aligned}$$

Law of large numbers

Theorem (Law of Large Numbers (LLN))

Let $(X_1, \ldots, X_n)$ *be* n *iid random variables, with pmf* P_X */pdf* f_X*, and let* $\mu = \mathbb{E}(X)$ *be the expectation of* X*.*
The empirical mean $\bar{X}_n$ *of* $(X_1, \ldots, X_n)$*, defined by*

$$\bar{X}_n \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i$$

converges in probability, as n*, grow infinite, to* μ*, i.e.,*

$$\mathbb{P}\left(\lim_{n \to \infty} \bar{X}_n = \mu \right) = 1.$$

Central limit theorem

Theorem (Central limit theorem (CLT))

Let $(X_1, \ldots, X_n)$ be n iid random variables, with pmf P_X/pdf f_X, and let $\mu = \mathbb{E}(X)$ be the expectation of X and σ^2 be its variance.
The random variable $Z_n = \sqrt{n}(\bar{X}_n - \mu)$ defined by

$$Z_n = \frac{\sqrt{n}}{n} \sum_{i=1}^{n} X_i - \mu$$

converges in distribution, as n, grow infinite, to a normal Gaussian distribution

$$\lim_{n\to\infty} \mathbb{P}\left(Z_n \leq z\right) = \Phi(z).$$

where $\Phi(z)$ is the cdf of a normal Gaussian distribution.

Kullback-Leibler divergence

Definition (Kullback-Leibler (KL) divergence)

Let P_X and Q_X be two probability distributions on $\mathcal{X}$.
The KL divergence between P_X and Q_X is defined as

$$D_{KL}(P_X||Q_X) \triangleq \sum_{x\in\mathcal{X}} P_X(x) \log\left(\frac{P_X(x)}{Q_X(x)}\right).$$

The KL is extensively used in statistical information engineering

Properties

The KL divergence verifies a certain set of properties :

1. *Asymmetry :* $D_{KL}(P_X||Q_X) \neq D_{KL}(Q_X||P_X)$.
2. *Null element :* $D_{KL}(P_X||P_X) = 0$.
3. *Positivity :* $D_{KL}(P_X||Q_X) \geq 0$, *for all laws* (P_X, Q_X).