Basic probability and statistics

Fabrice Gamboa & Bertrand Iooss

Course written by Alberto Pasanisi

Probability / Statistics

- Probability Theory:
  - Allows modeling random phenomena, ruled by hazard
  - It is an axiomatic mathematical theory (out of touch with any phisical reality)
  - It is mathematical tool for representing uncertainty
  - It is the basic mathematical tool for statistical analysis

- Statistical analysis:
  - Observation and analysis of real data/phenomena
  - Establishing general conclusions under the basis of limited-size samples, i.e. a given number of observations of a real phenomenon

- Other representation of uncertainties (≠ probability) exist …

Random experiments and events

- Random experiments: hazard acts and makes the result unforeseeable (e.g. dice rolling)
  - NB It is often a "modeling choice", when underlying physics is too complex
- Let us consider the set of all possible results: "Sample Space" : $\Omega = \{1, 2, 3, 4, 5, 6\}$
- "Event": assertion related to the result of an experiment
- The event is associated to a subset A of possible values
  - Ex 1: get an even number → $A = \{2, 4, 6\}$
  - Ex. 2: get a number ≤ 2 → $A = \{1, 2\}$
  - The event occurs (or not) with a given "probability"
  - Thus, the probability is associated to each of the subsets A
    - … which are expected to obey some properties

Can we really establish a physical model for dice rolling?

Probability

- We are interested in subsets of Ω which belong to a class Ψ such as:

$$\Omega \in \Psi$$

$$A \in \Psi \Rightarrow \bar{A} \in \Psi$$ → The complement of A is in à Ψ

$$A_1, A_2 \ldots A_n \in \Psi \Rightarrow \bigcup_{i=1}^{i=n} A_i \in \Psi$$ → The union of elements of Ψ is in Ψ

  - The sample space Ω, with the set Ψ of all possible events is "probabilisable" → We may associate a probability to each events

  - The "probability measure" (or simply "probability") is a mapping from A to [0,1] obeying the three axioms :

$$1)\ \ \forall A \in \Psi : \mathbb{P}(A) \in [0,1]$$

$$2)\ \ \mathbb{P}(\Omega) = 1$$

$$3)\ \ A_i \ldots A_n \in \Psi ; \forall (i,j) A_i \cap A_j = \varnothing \Rightarrow \mathbb{P}\left(\bigcup_{i=1}^{i=n} A_i\right) = \sum_{i=1}^{i=n} \mathbb{P}(A_i)$$

Andrey Nikolayevich Kolmogorov (1903-1987)

Probability... beyond mathematical formalism

- Our starting point was a random experience:
  - We have defined some events (which occur or not)
  - And we associated to each of the events a probability measure contained between 0 (impossible event) et 1 (certain event)
  - We also had to impose some mathematical constraints to events …
- The probability is just a mathematical object. **What interpretation?**
- Classical "frequentist" interpretation of probability:
  - Probability is seen as the limit frequency of a set of results over an infinite number of trials
  - This interpretation is suited to events which are (at least in principle) repeatable
    - NB Founders of probability calculation were historically interested in hazard games (e.g. Fermat and Bernoulli 1654 / Law of large numbers, Bernoulli, Poisson)
    - **But what about non-repeatable events?**

Probability... beyond mathematical formalism

- "Subjective" interpretation of probability
  - Probability is seen as a numerical quantification of a state of knowledge. This "translation" is not arbitrary but obeys some rationality principles.
  - Subjective probability is associated to the idea of odd. The probability of an event depends on the amount that a rational individual is ready to bet on it.

Bruno de Finetti
(1906-1985)

Let us suppose that an individual is obliged to evaluate the rate p at which he would be ready to exchange the possession of an arbitrary sum S (positive or negative) dependent on the occurrence of a given event E, for the possession of the sum pS; we will say by definition that this number p is the measure of the degree of probability attributed by the individual considered to the event E, or, more simply, that p is the probability of E

Bruno de Finetti, 1937,

"La Prévision: Ses Lois Logiques, Ses Sources Subjectives", Annales de l'Institut Henri Poincaré, 7: 1–68; translated as "Foresight. Its Logical Laws, Its Subjective Sources", in Studies in Subjective Probability, H. E. Kyburg, Jr. and H. E. Smokler (eds.), Krieger Publishing, 1980.

Cf. http://plato.stanford.edu/entries/probability-interpret/

Different interpretations but only one mathematical object, defined hereinbefore

Probability: some properties

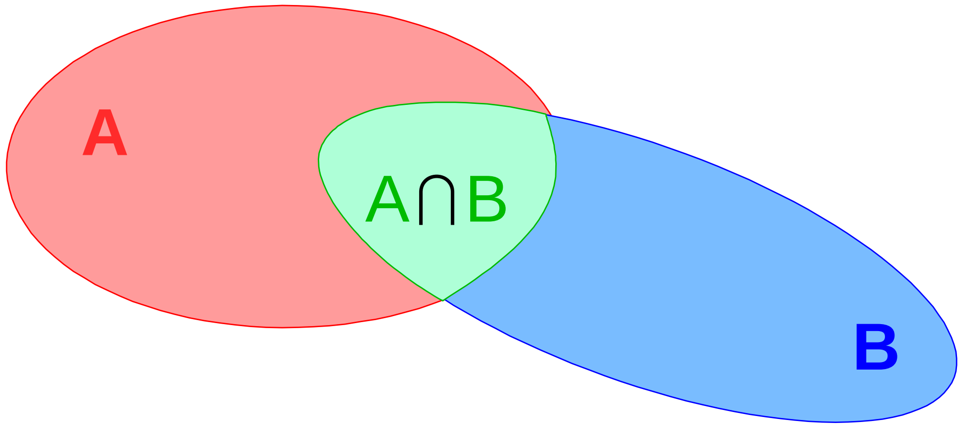
- Basic properties

$\mathbb{P}(\varnothing) = 0$ → Probability of the “null” event

$\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$ → Probability of the complementary event

$A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$ → Prob. of an event included into another

$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ → Probability of the union of events



Conditional probability and independence (1/2)

- Definition (1) : conditional probability of A, given B, (with P(B) ≠ 0)

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

- Definition (2) : independent events

$$\text{A et B indep. : } \mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

- The actual question: Knowing that B occurred, has (or not) an impact on the probability of A?
  - No → A et B are independent
  - Yes → A et B are dependent

Conditional probability and independence (2/2)

- If A et B are independent:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A) \cdot \mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$$

The fact that B has occurred does not change the probability that A will occur

Attention: Dependence ≠ Causality !

- Some examples
  - Dependence between the number of ice-creams sold and the number of deaths by drowning
  - Dependence between shoe-size of children and their language skill
  - In both cases, a third underlying variable explains these probabilistic dependences

Bayes formula

- Inverse conditioning relationship: from A|B to B|A
  - Starting point: definition of conditional probability

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \Rightarrow \left| \begin{array}{l} \mathbb{P}(A \cap B) = \mathbb{P}(A|B) \cdot \mathbb{P}(B) \\ \mathbb{P}(B \cap A) = \mathbb{P}(B|A) \cdot \mathbb{P}(A) \end{array} \right.$$

  - If we replace at numerator P(A ∩ B) with the expression of P(B ∩ A) :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)}$$

  - We also have: $\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B) \cdot \mathbb{P}(B)}{\mathbb{P}(A)}$

Thomas Bayes (1702 -1761)

Law of total probability

- Let B_1, B_2, … B_n be a partition of Ω : $\cup(B_1, B_2 \dots B_n) = \Omega$
  - Then: $\mathbb{P}(A) = \mathbb{P}(A \cap B_1) + \mathbb{P}(A \cap B_2) + \dots \mathbb{P}(A \cap B_n) =$

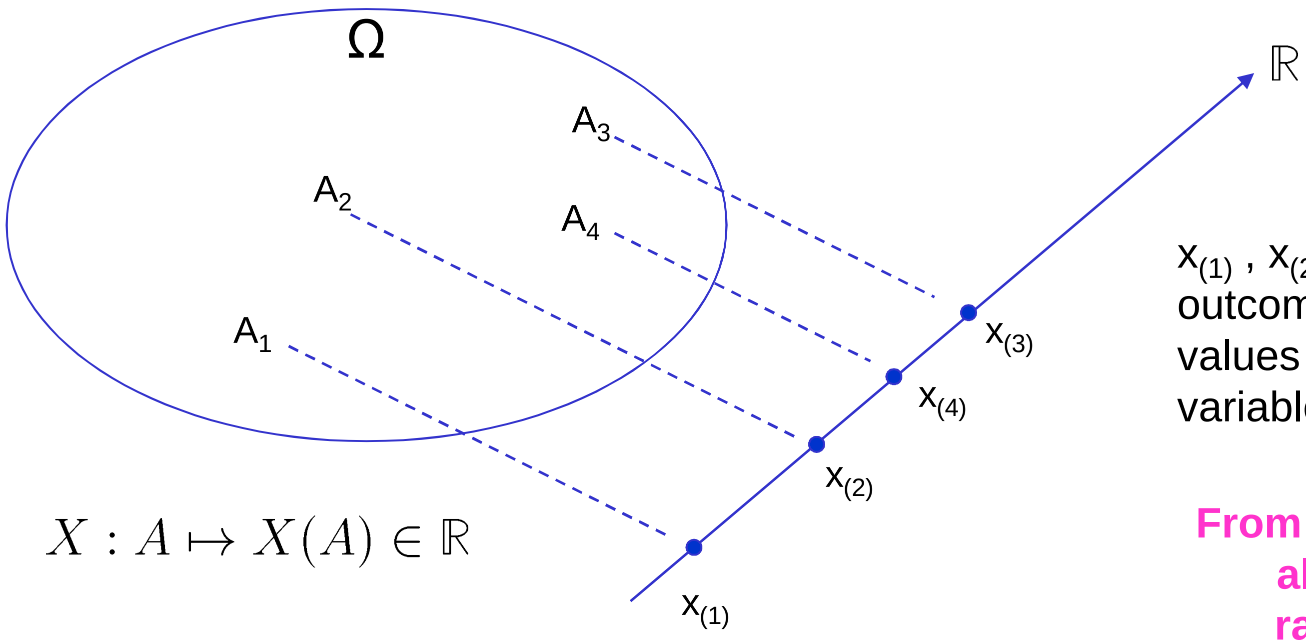
$$\mathbb{P}(A|B_1) \cdot P(B_1) + \mathbb{P}(A|B_2) \cdot P(B_2) + \dots \mathbb{P}(A|B_n) \cdot P(B_n)$$

- New expression of the Bayes formula:

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j) \cdot \mathbb{P}(B_j)}{\sum_{i=1}^{n} \mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i)}$$

Random variable

- Last mathematical item for completing this reminder on probability
- The problem: we defined the probabilities of events, but it is easier to cope with numbers!
  - We simply let a real number x corresponds to each of the events



$$X : A \mapsto X(A) \in \mathbb{R}$$

$x_{(1)}$, $x_{(2)}$ … are outcomes (i.e. possible values for the random variable X)

From now on, we will always work with random variables

Discrete Random variables

- Variables taking a discrete number of values
  - Example. Coin tossing
    - X=1 if the outcome is "head"
    - X=0 if the outcome is "tail"

- Distribution of probability of a discrete r.v.
  - Function associating to each of the possible outcomes of X, (x_1, x_2, … x_n) its probability

$$x_i \mapsto \mathbb{P}(x_{(i)})$$

$$\sum_{i=1}^{n} \mathbb{P}(x_{(i)}) = 1$$

  - For instance, for coin tossing:

$$\mathbb{P}(0) = 0.5$$
$$\mathbb{P}(1) = 0.5$$

1974 World Cup FInal

Continuous random variables

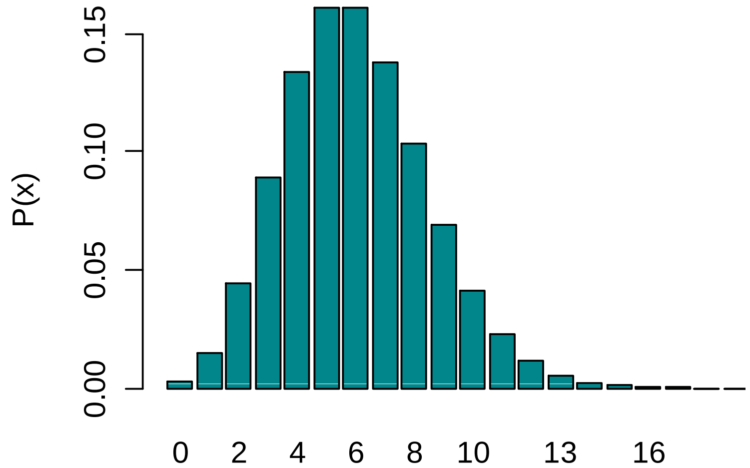
- Variables taking values in an uncountable set (in practice, intervals)
  - Example: the Seine water level in Chatou

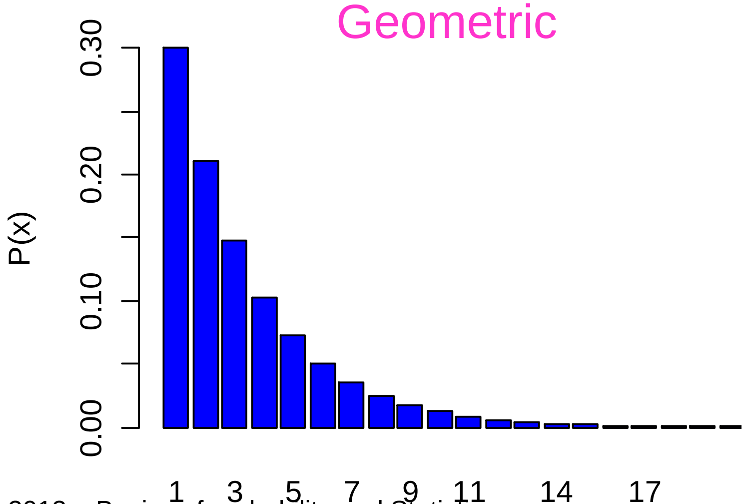
- Distribution of probability of a continuous random variable
  - Associates to each interval (a, b), the probability for the r.v. to be between a and b
  - NB In the case of discrete r.v., a probability is associated to each value of X. In the case of continuous r.v., a probability is associated to each interval of values of X.

Some probability distributions usually employed in common practice – discrete

Binomial

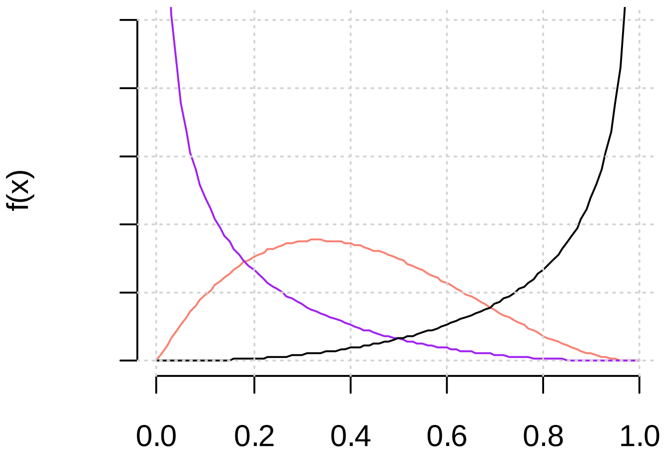
Poisson





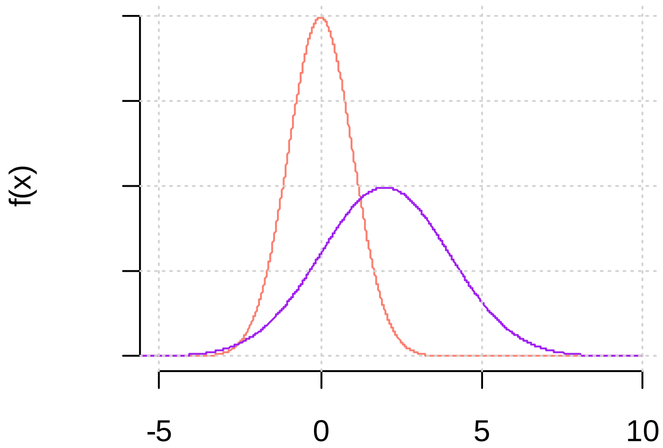
Some probability distributions usually employed in common practice – continuous (1/2)

Beta

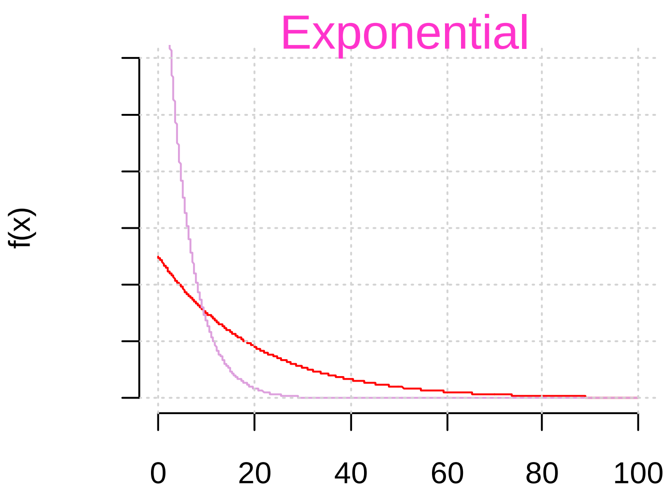


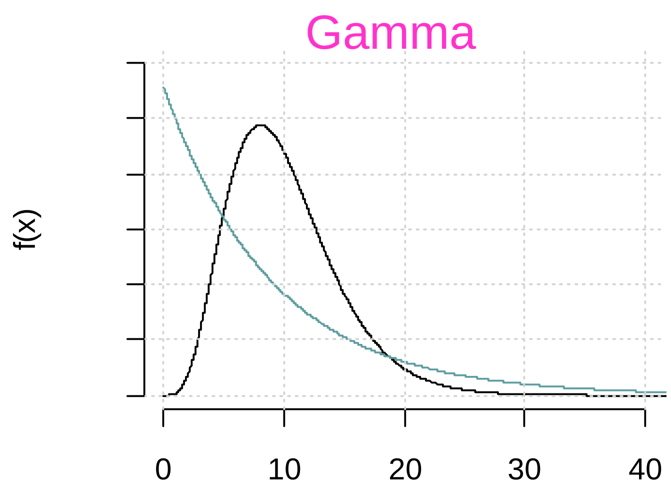
Log-Normal

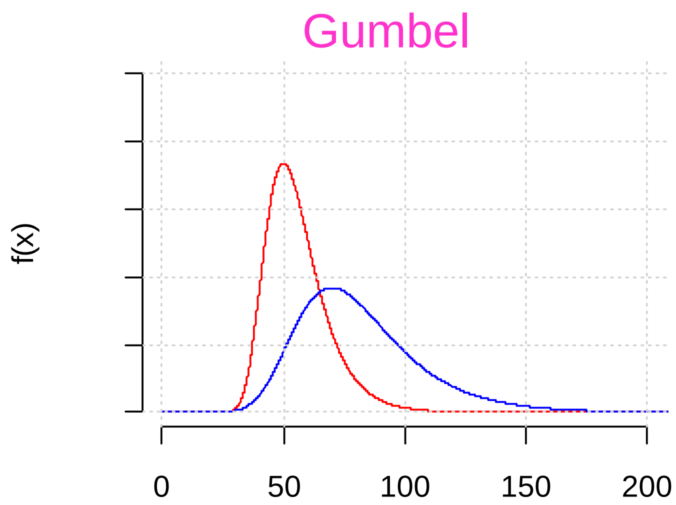
Normal

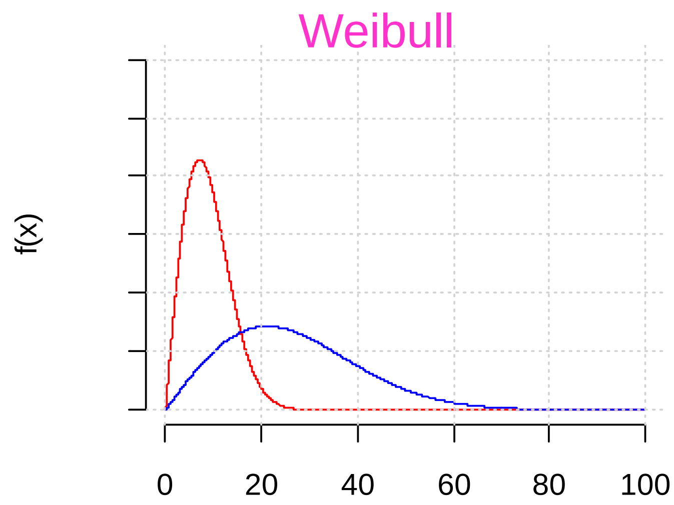


Some probability distributions usually employed in common practice – continuous (2/2)









Cumulative distribution function (CDF)

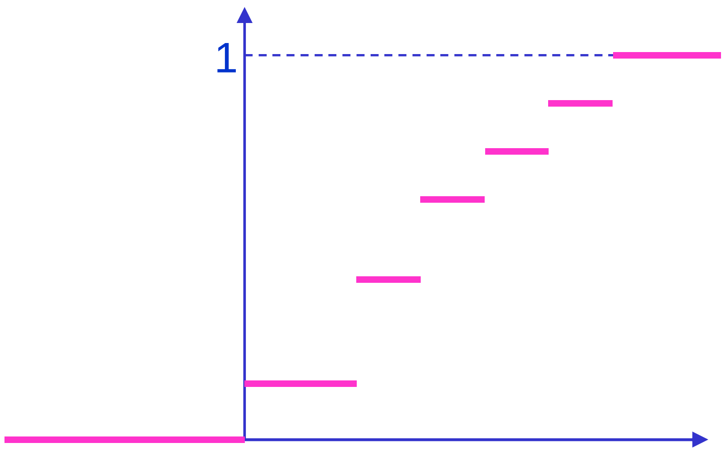
- Cumulative distribution function F
- Associates to each real number x, the prob. for the r.v. X be ≤ x

$$F(x) = \mathbb{P}(X \leq x)$$

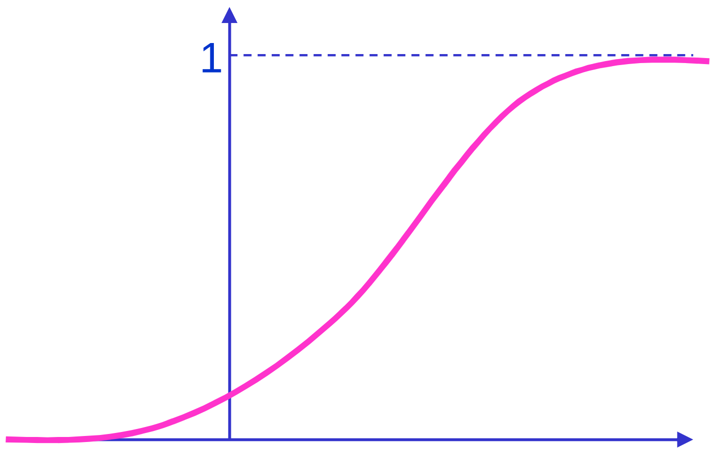
$$\mathbb{P}(a \leq X \leq b) = F(b) - F(a)$$

Monotonous increasing function, with values in [0,1)

Discrete case (Step function)



Continuous case



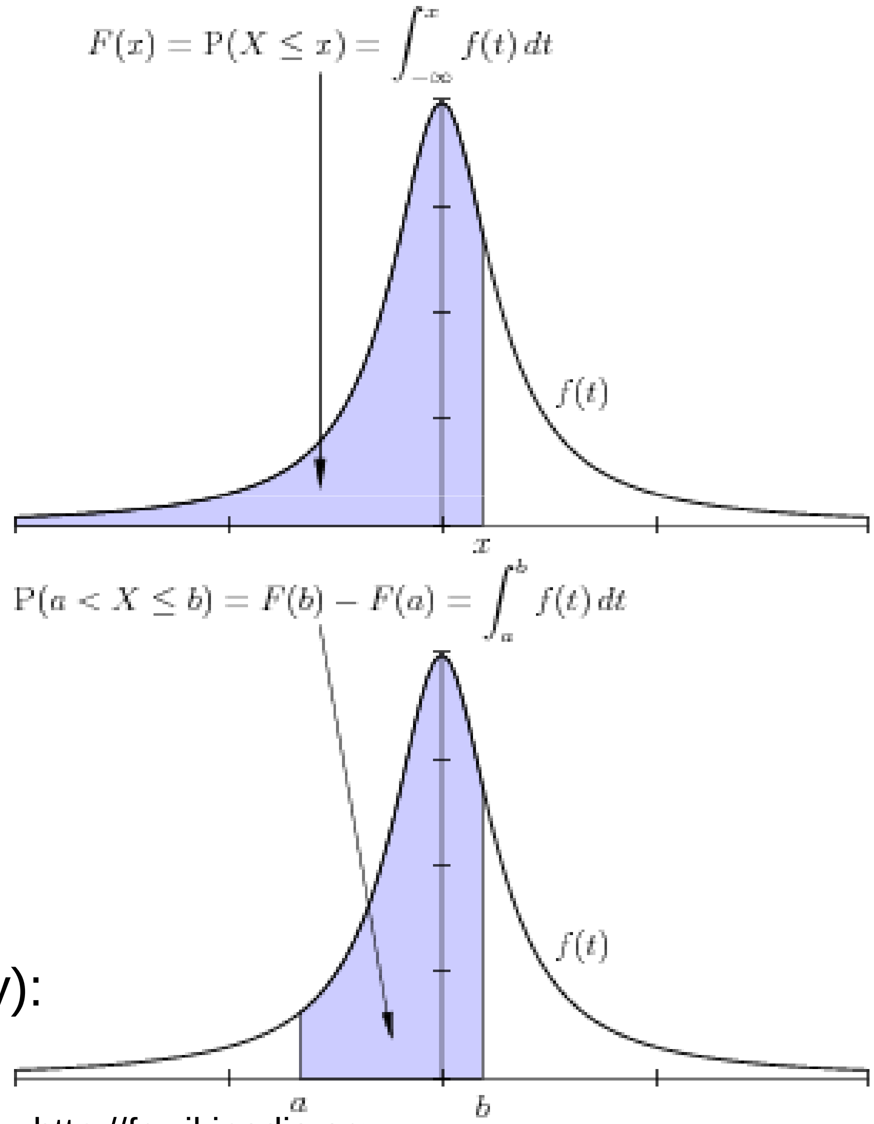
Probability density function (continuous r.v.)

- Defined by the relation:

$$F(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f(t) \cdot dt$$

  – So then: $\mathbb{P}(a \le X \le b) = \int_{a}^{b} f(t) \cdot dt$

  – Formally, it is the derivative of the cumulated distr. function



http://fr.wikipedia.org

- Properties: $f(x) \ge 0 \quad \forall x$

$$\int_{-\infty}^{+\infty} f(t) \cdot dt = 1$$

- Intuitive Interpretation (limit probability):

$$\mathbb{P}(t \le X \le t + dt) = f(t) \cdot dt$$

Expected value

- Expected value (or mean): “Central tendency” value of a r.v., defined by the expressions:

$$\mathbb{E}(x) = \int_{-\infty}^{+\infty} x \cdot f(x) dx$$

Discrete case

$$\mathbb{E}(x) = \sum_{i=1}^{n} x_i \cdot \mathbb{P}(x_i)$$

Continuous case

- Properties:

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

$$\mathbb{E}(aX) = a \cdot \mathbb{E}(X)$$

$$\mathbb{E}(a) = a$$

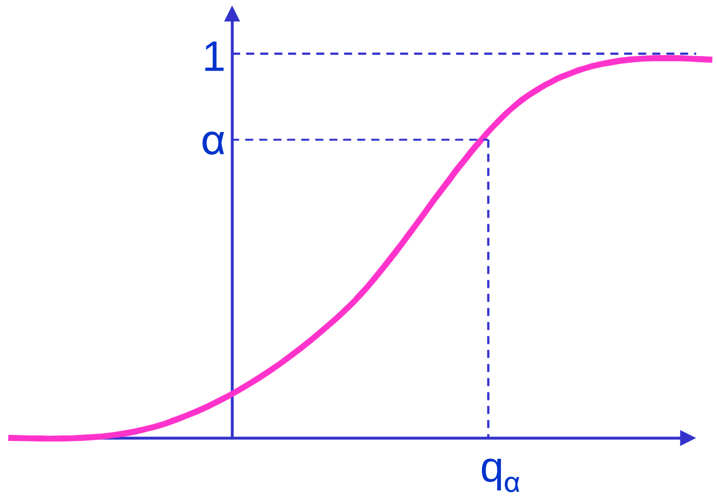
Linear Operator

Median and quantiles

- Quantile of probability α: value of X, having a probability α for not being exceeded, i.e. such that:

$$\mathbb{P}(X \leq q_\alpha) = \alpha$$

- If α = ½, this quantile is called the median value of X



Variance and standard deviation

- Variance: Expected value of the random variable: $(X - \mathbb{E}(x))^2$

  $\mathbb{V}(X) = \mathbb{E}\left((X - \mathbb{E}(x))^2\right)$

  Squared deviation from the expected value

- Standard deviation: $\sigma_X = \sqrt{\mathbb{V}(X)}$

- Properties : $\mathbb{V}(X) = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2$

  $$\mathbb{V}(a \cdot X + b) = a^2 \cdot \mathbb{V}(X)$$

- If X et Y are independent: $\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y)$

Covariance and linear correlation coefficient

- Quantity involving two random variables X and Y
- Definition :

$$\mathrm{cov}(X,Y) = \mathbb{E}\left((X-\mathbb{E}(X))\cdot(Y-\mathbb{E}(Y))\right) \longrightarrow \mathrm{cov}(X,X) = \mathbb{V}(X)$$

  – Intuitively, the covariance is a measure of the simultaneous variation of two r.v. A high (absolute value) covariance means that X et Y vary "in the same way" (positive relation, direct, increasing) or in the opposite way (negative relation, inverse, decreasing).

- Properties: $-\infty \leq \mathrm{cov}(X,Y) \leq +\infty$

$$\mathrm{cov}(X,Y) = \mathbb{E}(X\cdot Y) - = \mathbb{E}(X)\cdot\mathbb{E}(Y)$$

$$\text{X et Y indep.} \Rightarrow \mathrm{cov}(X,Y) = 0$$ → the reverse is not true

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\mathrm{cov}(X,Y)$$ → Variance of the sum of two r.v.

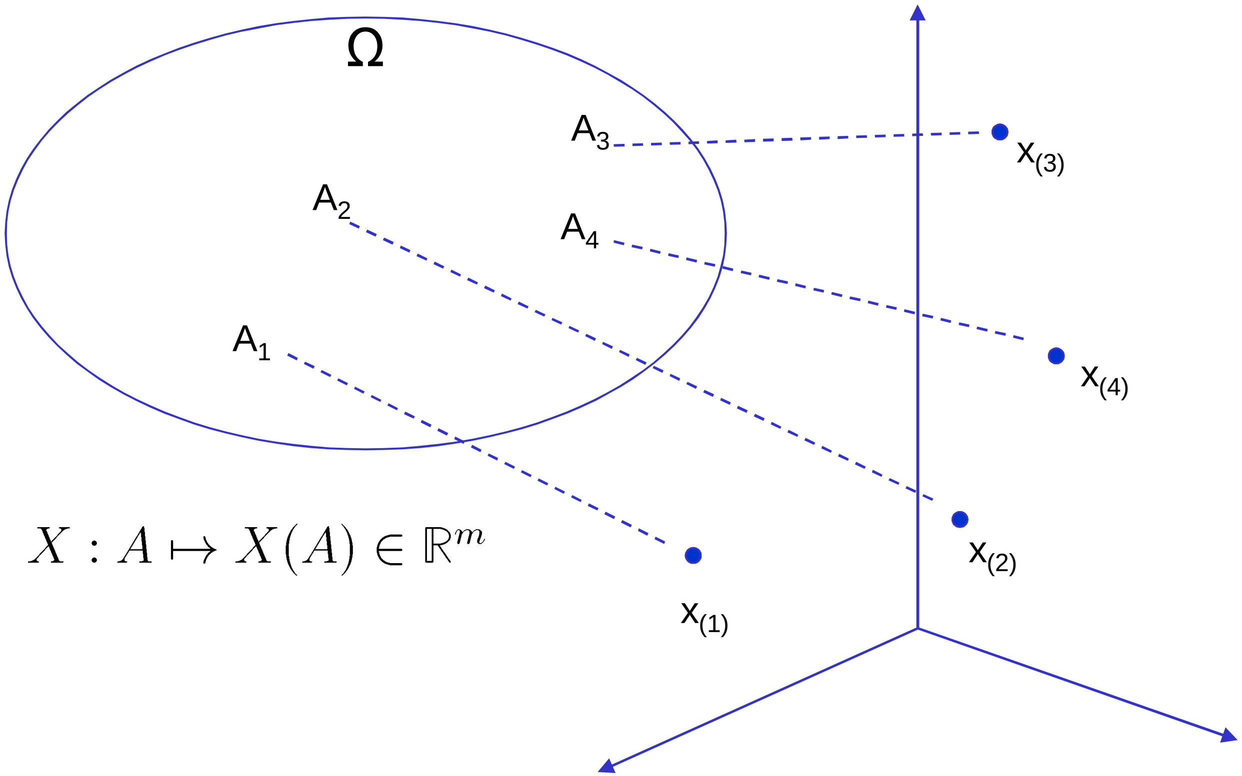
- Linear correlation coefficient: $\varrho(X,Y) = \dfrac{\mathrm{cov}(X,Y)}{\sqrt{\mathbb{V}(X)\cdot\mathbb{V}(Y)}} \in [-1,1]$

Expected value and variance of some usual laws

Name of the law (parameters)	Possible values	Analytical expression of the distribution	Expected value	Variance
Binomial(n,p), $n \geq 0$ et $0 \leq p \leq 1$	{0 ; 1 ; ... ; n}	$\mathrm{Prob}(X=k) = C_n^k p^k (1-p)^{n-k}$	np	np(1-p)
Poisson(λ), $\lambda \geq 0$	0 ; 1 ; 2 ; ...	$\mathrm{Prob}(X=k) = \exp(-\lambda)\frac{\lambda^k}{k!}$	λ	λ
Normal(μ,σ), $\sigma > 0$	]-∞ ; + ∞[	$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$	μ	σ^2
χ^2(n), n entier	[0 ; + ∞[	$f(x;n) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1}\exp\left(-\frac{x}{2}\right)$	n	2n
Log-Normal(μ,σ), $\sigma > 0$	[0 ; + ∞[	$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}x}\exp\left[-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2\right]$	$\exp\left(\mu+\frac{\sigma^2}{2}\right)$	$\exp(2\mu+\sigma^2)\left[\exp(\sigma^2)-1\right]$
Uniform(a,b)	[a ; b]	$f(x;a,b) = \frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential(μ,λ), $\lambda > 0$	[μ ; + ∞[	$f(x;\mu,\lambda) = \lambda\exp[-\lambda(x-\mu)]$	$\mu+\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Weibull(μ,η,β), η et $\beta > 0$	[μ ; + ∞[	$f(x;\mu,\eta,\beta) = \frac{\beta}{\eta}\left(\frac{t-\mu}{\eta}\right)^{\beta-1}\exp\left[-\left(\frac{t-\mu}{\eta}\right)^{\beta}\right]$	$\mu+\eta\Gamma\left(1+\frac{1}{\beta}\right)$	$\eta^2\left[\Gamma\left(1+\frac{2}{\beta}\right)-\left\{\Gamma\left(1+\frac{1}{\beta}\right)\right\}^2\right]$
Gamma(α,β), α et $\beta > 0$	[0 ; + ∞[	$f(x;\alpha;\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x)$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Gumbel(m,s), s > 0	]-∞ ; + ∞[	$f(x;m,s) = \frac{1}{s}\exp\left[-\left(\frac{x-m}{s}\right)\right]\exp\left[-\exp\left\{-\left(\frac{x-m}{s}\right)\right\}\right]$	$m+\gamma s$ $\gamma = 0{,}577222$	$\frac{1}{6}\pi^2 s^2$

Multi-dimensional random variable

- Random vector: generalization of the notion of real r.v.



$$X : A \mapsto X(A) \in \mathbb{R}^m$$

$X_1, X_2, \ldots, X_m$: component of the random vector X (real random variables)

Multi-dimensional random variable

- Let us stay in the case m = 2. Let X et Y be the two components
  - Multi-dimensional cumulated distrib. function: $F(x,y) = \mathbb{P}(X \leq x, Y \leq y)$
  - Density : $f(x,y) = \dfrac{\partial^2 F(x,y)}{\partial X \partial Y}$
  - Variance-covariance matrix: $\begin{bmatrix} \mathbb{V}(X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(Y,X) & \mathbb{V}(Y) \end{bmatrix}$

Marginal and conditional distributions

- Marginal distribution: distribution of a component regardless of the other:

$$f_X(x) = \int f(x,y) \cdot dy \qquad f_Y(y) = \int f(x,y) \cdot dx$$

  – Univariate distributions of the components, taken "one at a time"

- Conditional distribution: "generalization" of the notion of conditional probability

$$f(x|y) = \frac{f(x,y)}{f_Y(y)}$$

$f(x,y)$ → Joint distribution of (x,y)

$f_Y(y)$ → Marginal distribution of y

- Independence : $f(x,y) = f_X(x) \cdot f_Y(y)$

$$\text{X et Y indep. : } f(x|y) = f_X(x)$$

How modeling, in practice, a multi-dimensional random variable?

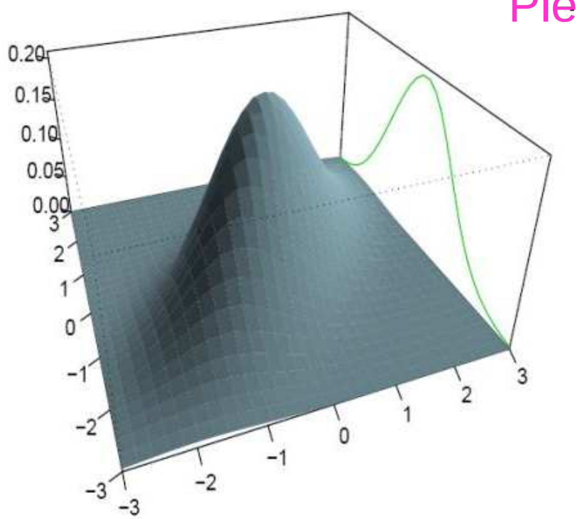
- Using “standard” distribution
- Conditioning
- Copulas

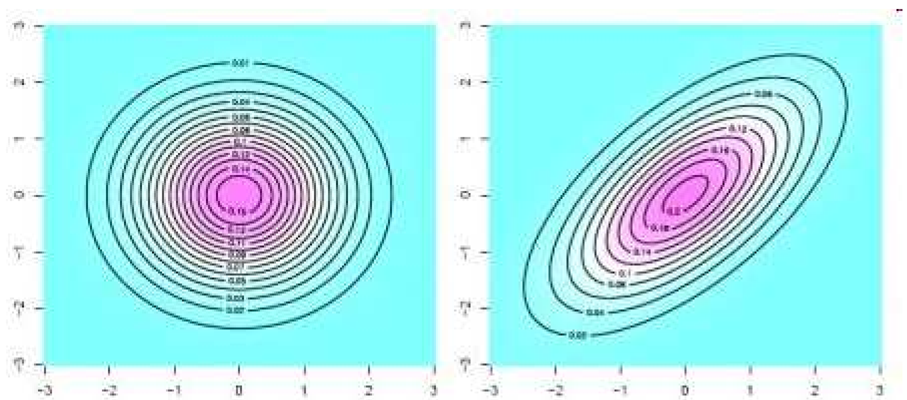
“Usual” multi-dimensional distribution

- There are not a lot of distributions! Moreover, uneasy manipulation!
- Example: Multivariate Normal distribution → generalization of the normal distribution in R^n :
  - Density : $$f(x|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(\tfrac{1}{2}(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)\right)$$

mean   Var-covar. mat. n x n   x and μ are vectors of R^n

Plenty of good properties …



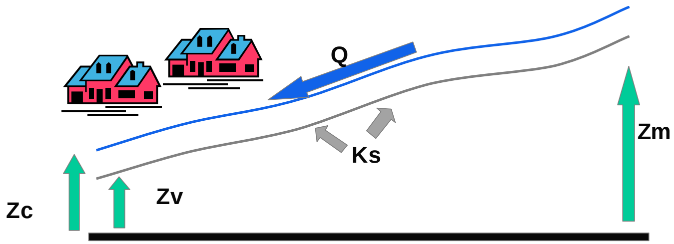


http://blogperso.univ-rennes1.fr/arthur.charpentier/

Conditioning

- Reminder: definition of conditional distribution: $f(x|y) = \dfrac{f(x,y)}{f_Y(y)}$
  - The joint distribution can be written as the product of two distributions: $f(x,y) = f(x|y) \cdot f_Y(y)$
  - This modeling approach is often linked to a given "expert" knowledge allowing a kind of "hierarchy" between the variables

Example: relation between the river bottom levels in two different points (upstream and downstream)



$$Z_v = Z_m - (L \cdot I) + \epsilon$$

"expert" vision → Probabilistic modeling vision

Downstr. level — Upstream level — Length and average slope — Disturbing term depending on local variation

$$Z_m \sim \mathcal{T}(a,b,c)$$
$$I \sim \mathcal{U}(I_{\min}, I_{\max})$$
$$Z_v | Z_m, I \sim \mathcal{N}(Z_m - L \cdot I, \sigma)$$

Copulas-based modeling

- Somehow, a "descriptive" approach
  - Idea: using two different mathematical objects for describing:
    - The uncertainty tainting the two components of the vector taken "one at a time"

      $f_X(x), \quad f_Y(y)$ → Marginal densities of (x,y)

      $F_X(x), \quad F_Y(y)$ → Marginal CDF of X and Y

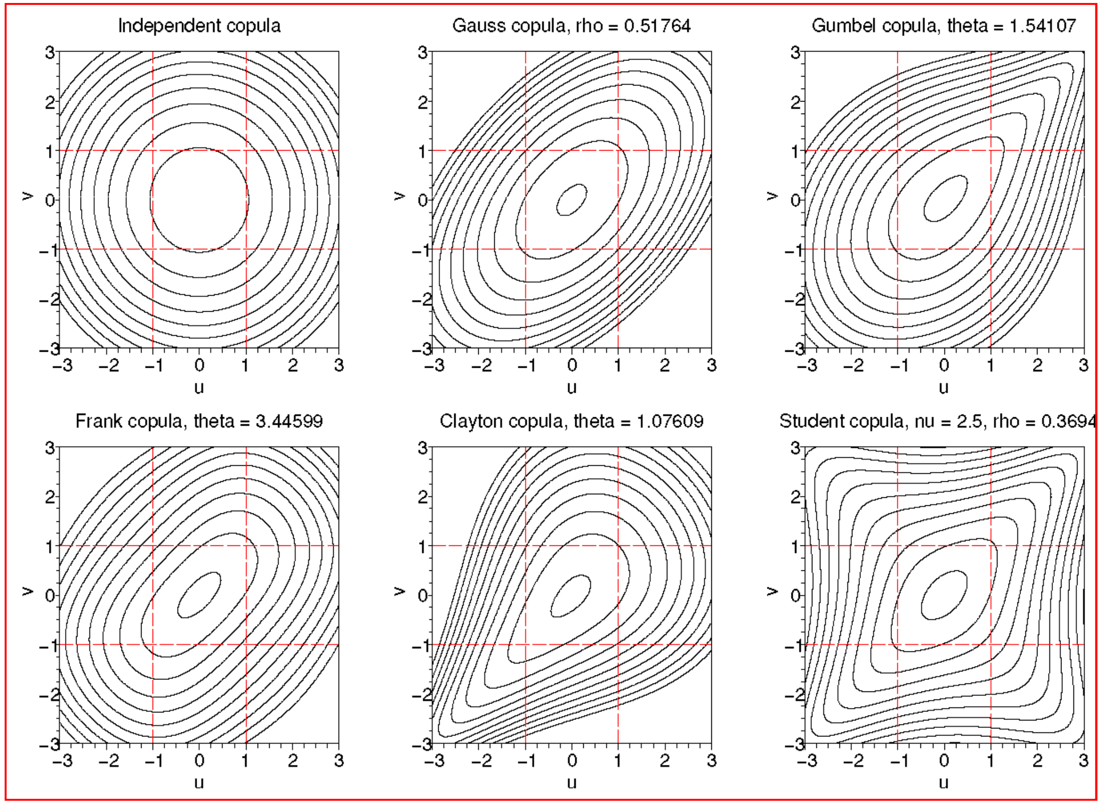
    - The dependence structure :
      - Function $\mathcal{C}$ (copula), such as: $F(x,y) = \mathcal{C}(F_X(x), F_Y(y))$
      - $\mathcal{C}$ is a cumulated distribution functions: $[0,1]^m \mapsto [0,1]$

    - A theoretical result (Sklar theorem, 1959) states that any joint distribution can be written using its marginal and a copula. Under mild conditions, the copula is unique.

Copulas-based modeling (more)

- Pragmatic choice (en practice we prefer working with 1D distribution)
  - Several copulas available → very varied modeling of the dependence
    - That is the choice made by Open TURNS developers

Example : All bivariate distribution here have the same marginal distributions (standard normal). Only the copula changes.



Beware of **implicit choices** made by popular software tools

Principal component analysis