… solutions for robust engineering

Probability theory basics

Gaëtan Blondet, Phimeca Engineering SA

'HPC and Uncertainty Treatment – Examples with Open TURNS and Uranie'

EDF – Phimeca – Airbus Group – IMACS – CEA

PRACE Advanced Training Center – May, 10-12 2021

Motivation

- Uncertainty includes variability, randomness and lack-of-knowledge.

  - Aleatory uncertainty
    - Lack of control over environmental variability and test settings, errors made during testing.
    - Can be better characterized but cannot be reduced with more measurements or simulations.

  - Epistemic uncertainty
    - Lack-of-knowledge and assumptions made during testing and modeling.
    - Can be reduced by collecting more information and evidence.

These sources of uncertainty can be modeled thanks to probability theory

Note: *Other theories have been developed to represent epistemic uncertainty such as Imprecise Theory (IP), Possibility theory, Fuzzy sets and fuzzy logic.*

Outline

General definitions

Random variables

- Definitions
- Cumulative distribution function and probability density function
- Moments
- Confidence intervals (CI)

Random vectors

- Definitions
- Moments
- Copulas

Outline

- General definitions

- Random variables
  - Definitions
  - Cumulative distribution function and probability density function
  - Moments
  - Confidence intervals (CI)

- Random vectors
  - Definitions
  - Moments
  - Copulas

Definitions

- Random experiment

  Repeatable procedure leading to possible outcomes.

- Sample space Ω

  Set of all possible outcomes of the experiment.

- Event

  Set of outcomes of an experiment (a subset of Ω).

Definitions

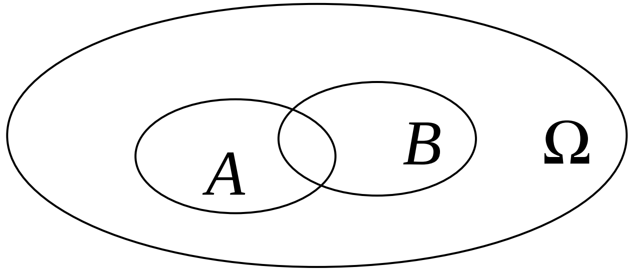
Probability

- Measure between 0 and 1 applied to events: $\mathbb{P}[A] \in [0,1]$
- Satisfies Kolmogorov axioms

Throwing 2 dice	
Event A_i	$\mathbb{P}[A_i]$
Do an even number	$\frac{1}{2}$
Do more than 2	$\frac{35}{36}$

Common properties:

- $\mathbb{P}[\emptyset] = 0\,, \mathbb{P}[\Omega] = 1$
- $\mathbb{P}[\bar{A}] = 1 - \mathbb{P}[A]$
- $\mathbb{P}[A \setminus B] = \mathbb{P}[A] - \mathbb{P}[A \cap B]$
- $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$
- $A \subseteq B \implies \mathbb{P}[A] \leq \mathbb{P}[B]$



Definitions

Conditional probability

- probability of A given B

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

Independence

- B does not affect the probability of A, and vice versa:

$$\mathbb{P}[A|B] = \mathbb{P}[A] \Longrightarrow \mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$$

Definitions

Bayes' theorem

- Shows the probability of A updated by the knowledge of B
- Defined from the conditional probability definition

Initial probability of *A*

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]}{\mathbb{P}[B]} \mathbb{P}[A]$$

Influence of the information contained in B

Definitions

Frequentist interpretation of probabilities

- Probabilities can be estimated by N observations

$$\mathbb{P}[A] = \lim_{N \to \infty} \frac{N_A}{N}$$

Outline

- General definitions
- Random variables
  - Definitions
  - Cumulative distribution function and probability density function
  - Moments
  - Confidence intervals (CI)
- Random vectors
  - Definitions
  - Moments
  - Copulas

Random variables

Definition

- A random variable (r.v.) X is a measurable function

$$\begin{aligned} X \ : \Omega \ &\longrightarrow \mathcal{D}_X \\ \omega \ &\longmapsto x = X(\omega) \end{aligned}$$

- Can be discrete ($\mathcal{D}_X \subseteq \mathbb{Z}$) or continuous ($\mathcal{D}_X \subseteq \mathbb{R}$)

Random variables

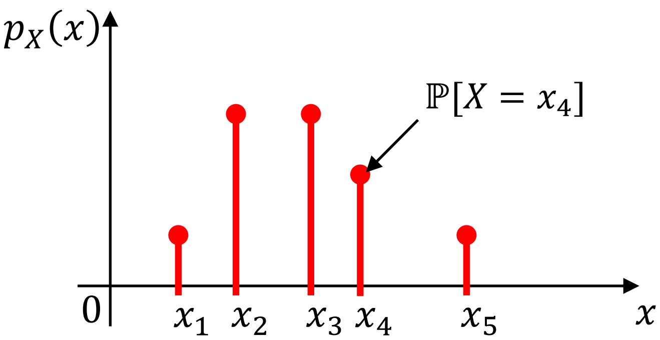
Probability Density Function (PDF)

- Discrete: probability mass function

$$p_X(x_i) = \mathbb{P}[X = x_i]$$

$$\forall x_i, 0 \leq p_X(x_i) \leq 1$$

$$\sum_{x_i} p_X(x_i) = 1$$

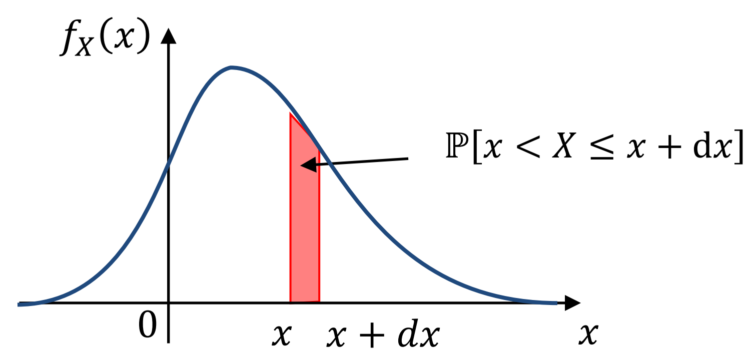


- Continuous: probability density function

$$f_X(x)\ \mathrm{dx} = \mathbb{P}[x < X + \mathrm{dx}]$$

$$\forall x, f_X(x) \geq 0$$

$$\int_{x \in \mathbb{X}} f_X(x)\ \mathrm{d}x = 1$$



Random variables

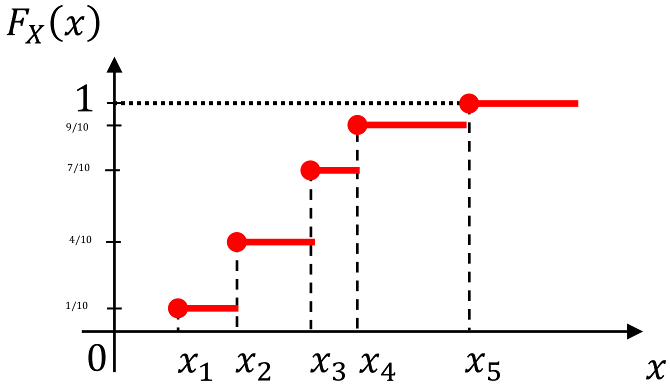
Cumulative Distribution Function (CDF)

- Discrete

$$F_X(x) = \mathbb{P}[X \leq \mathrm{x}] = \sum_{x \leq x_i} p_X(x_i)$$

$$\lim_{x \to \sup \mathcal{D}_X} F_X(x) = 1$$

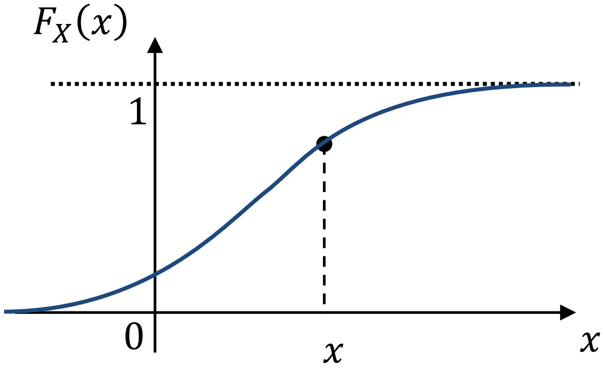
$$\lim_{x \to \inf \mathcal{D}_X} F_X(x) = 0$$



- Continuous

$$F_X(x) = \mathbb{P}[X \leq \mathrm{x}] = \int_{-\infty}^{x} f_X(x)\mathrm{d}x$$

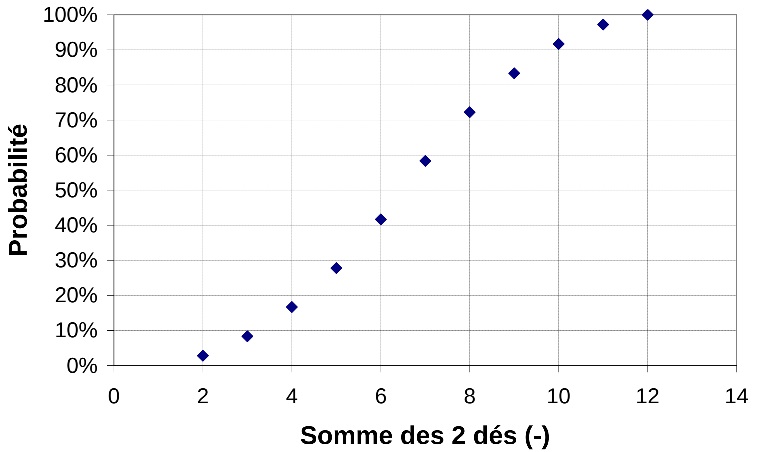
$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}$$



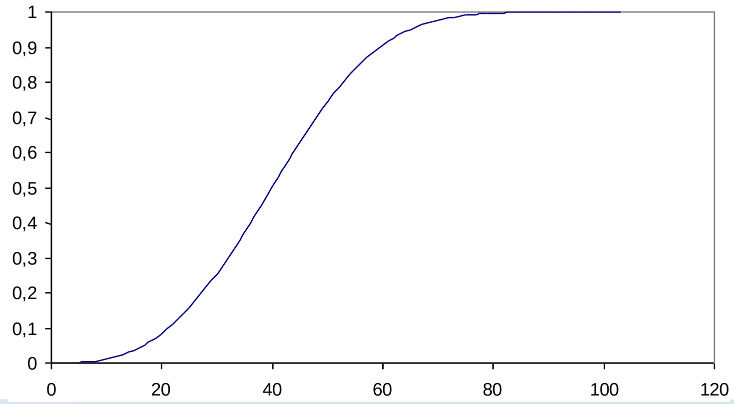
Random variables

Cumulative Distribution Function (CDF)

- Discrete: Sum of 2 dice: $\Omega \rightarrow \{2,\ldots,12\}$



- Continuous: Wind speed: $\Omega \longrightarrow \mathbb{R}^+$



Random variables

Link between several r.v.

- Let S the sum of two independent continuous r.v. X and Y :

$$S = X + Y$$

➢ The distribution of S can be deduced by convolution or characteristics functions

Measure = true value + error

- Composition

$$Y = \varphi(X)$$

➢ In some cases, the distribution of Y can be computed analytically. Otherwise, we must generate a sample.

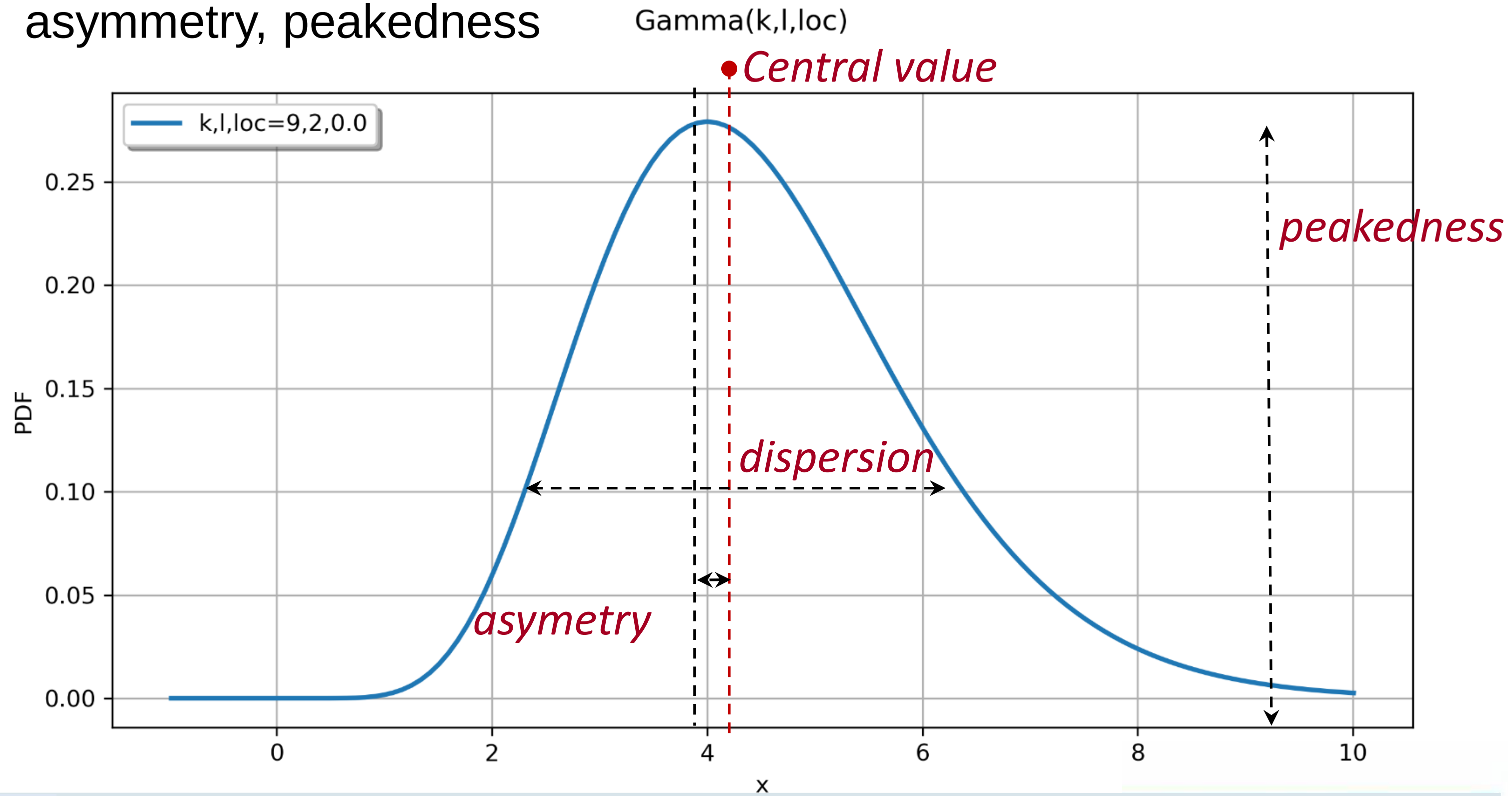
Output = f(input)

Random variables

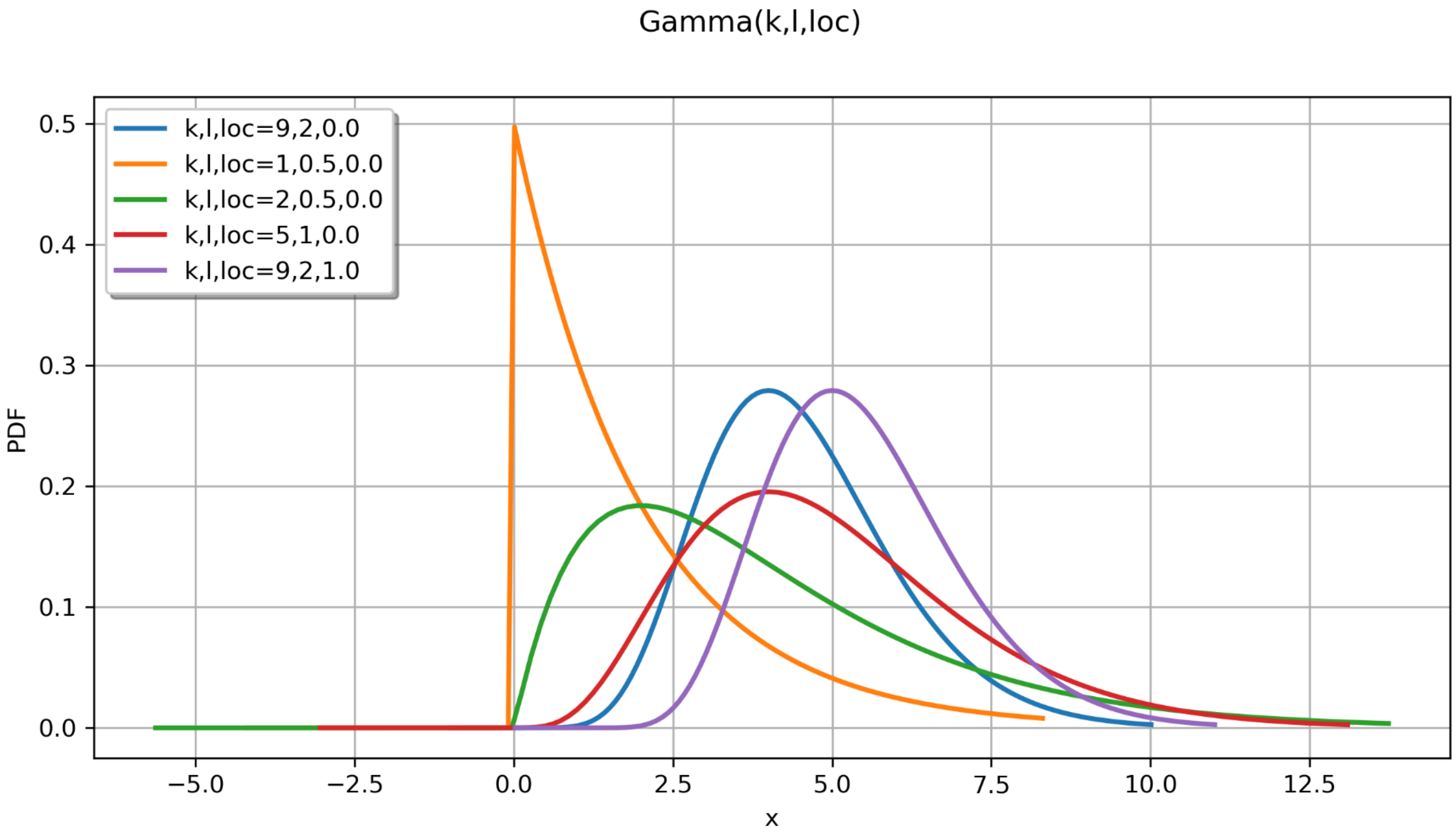
Characterization of a random variable

A distribution is characterized by its **moments** :

- central value
- dispersion
- asymmetry, peakedness



Random variables



Random variables

Moments of order $r (> 0)$

$$\mu^r_{X\,centered,standardized} = \mathbb{E}\left[\frac{(X-\mu_X)^r}{\sigma_X^r}\right]$$

r = 1	r = 2	r = 3	r = 4
Mean	**Variance**	**skewness**	**kurtosis**
central value	Dispersion	asymmetry	flattening
μ_X	σ_X^2	δ_X	κ_X

Coefficient of variation

$$\text{c.o.v.} = \frac{\sigma_X}{|\mu_X|}, \quad \mu_X \neq 0$$

Random variables

Expected value (Mean value)

Given X and Y, two r.v. and a and b two reals.

- For discrete r.v : $\mathbb{E}[X] = \sum_{x_i} x_i p_X(x_i)$
- For continuous r.v. : $\mathbb{E}[X] = \int_{x \in \mathbb{X}} x f_X(x) \mathrm{d}x$
- Linearity : $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$
- only if X and Y are independent: $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

Random variables

Variance (dispersion around the mean)

$$\sigma_X^2 = \mathrm{Var}\,[X] = \mathbb{E}[(X-\mu_X)^2]$$ *(if it exists, cf Cauchy distribution)*

- König-Huyghens formula: $Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mu_X^2$
- $Var[aX+b] = a^2\mathrm{Var}[X]$
- $Var[X+Y] = var[X] + Var\,[Y] + 2\underbrace{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}_{Cov[X,Y]}$
- If X and Y are independent:
  - $Var[XY] = Var[X]\,Var[Y] + Var[X]\mathbb{E}[Y]^2 + Var[Y]\mathbb{E}[X]^2$

Random variables

Quantiles

The quantile $\mathbf{x}_\alpha$ at probability level α, is

$$F_X(x_\alpha) = \alpha \quad \Rightarrow \quad x_\alpha = F_X^{-1}(\alpha), \quad 0 \leq \alpha \leq 1$$

First quartile	α = 25 %
median	α = 50 %
Third quartile	α = 75 %

Confidence intervals

Estimated moments are also random variables

To sum up the variability of a r.v. bounded by two quantiles centered on the median.

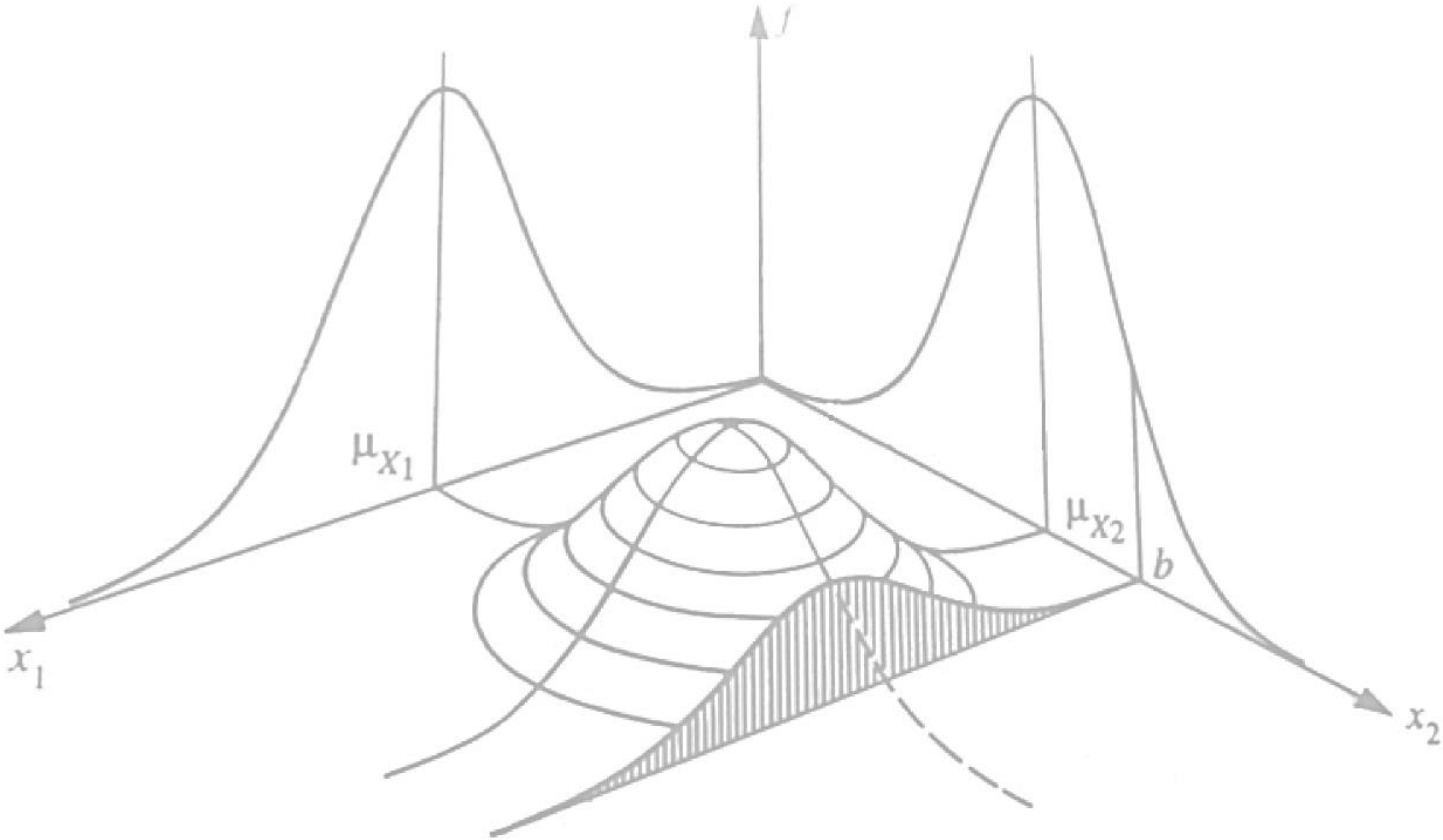
Confidence interval at the probability level of $1 - \alpha$:

$$\left[x_{\alpha/2};\ x_{1-\alpha/2}\right] = [F_X^{-1}(\alpha/2); F_X^{-1}(1-\alpha/2)\ ], \quad 0 \leq \alpha \leq 1$$

Outline

- General definitions
- Random variables
  - Definitions
  - Cumulative distribution function (CDF) and probability density function (PDF)
  - Discrete / continuous random variables
  - Statistical moments
  - Confidence intervals (CI)
- Random vectors
  - Definitions
  - Moments
  - Copulas

Random vectors



Random vectors

Definition

- A random vector is a measurable function

$$\begin{aligned}\mathbf{X} &: \Omega \to \mathbb{X} \subseteq \mathbb{R}^n \\ &\omega \mapsto \mathbf{x} = \mathbf{X}(\omega) = \left(X_1(\omega), \dots, X_n(\omega)\right)^t\end{aligned}$$

- Defined by:
  - Its joint cumulative distribution function

$$F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}\left[\bigcap_{i=1}^{n} X_i \leq x_i\right]$$

  - Its joint probability density function

$$f_X(\mathbf{x}) = \frac{\mathbb{P}[\cap_{i=1}^{n} x_i \leq X_i \leq x_i + dx_i]}{\prod_{i=1}^{n} dx_i} = \frac{\partial F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \dots \partial x\_n}$$

Random vectors

Complements

- marginal PDF: If $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^t$, the marginal density of $\mathbf{X}_1$ (in $\mathbf{X}$) is given by:

$$f_{\mathbf{X}_1}(\mathbf{x}_1) = \int_{\mathbf{x}_2 \in \mathbb{X}_2} f_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2) \mathrm{d}\mathbf{x}_2$$

- conditional PDF: If $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^t$ the conditional PDF of $\mathbf{X}_1$ given $\mathbf{x}_2 = \mathrm{b}$ is:

$$f_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_1|\mathbf{x}_2 = b) = \frac{f_{\mathbf{X}}(\mathbf{x}_1, b)}{\int_{\mathbf{x}_1 \in \mathbb{X}_1} f_{\mathbf{X}}(\mathbf{x}_1, b) \mathrm{d}\mathbf{x}_1} = \frac{f_{\mathbf{X}}(\mathbf{x}_1, b)}{f_{\mathbf{X}_2}(b)}$$

- Copula: a stochastic dependence structure, in case of r.v. are correlated.
  - Sklar Theorem : $F_{\mathbf{X}}(\mathbf{x}) = C(F_{X_1}(x_1), F_{X_2}(x_2))$

Random vectors

Moments

- Expected value: vector of expected values of random variables

$$\mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_i], i = 1, \dots, n)^t$$

- Covariance matrix:

$$\sigma_{ij} = \mathrm{Cov}[X_i, X_j] = \mathbb{E}\left[(X_i - \mu_{X_i})(X_j - \mu_{X_j})\right], \quad i, j = 1, \dots, n$$

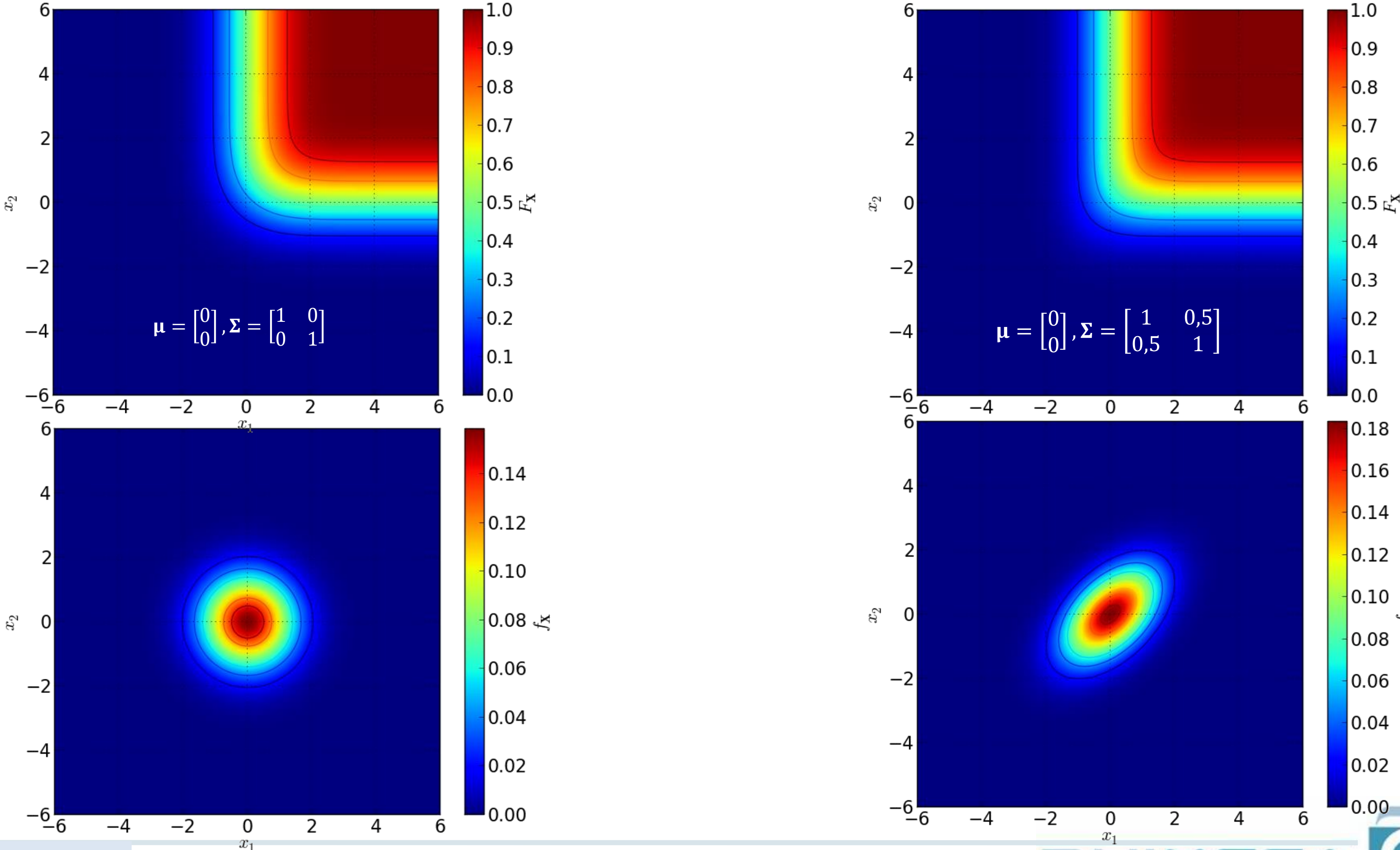
	X_1	X_2
X_1	${\sigma_1}^2$	$\mathrm{Cov}[X_1, X_2]$
X_2	$\mathrm{Cov}[X_2, X_1]$	${\sigma_2}^2$

Correlation matrix: $\rho_{ij} = \frac{\mathrm{Cov}[x_i, x_j]}{\sqrt{\mathrm{Var}[x_i]\mathrm{Var}[x_j]}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}, \quad i, j = 1, \dots, n$

Random vectors

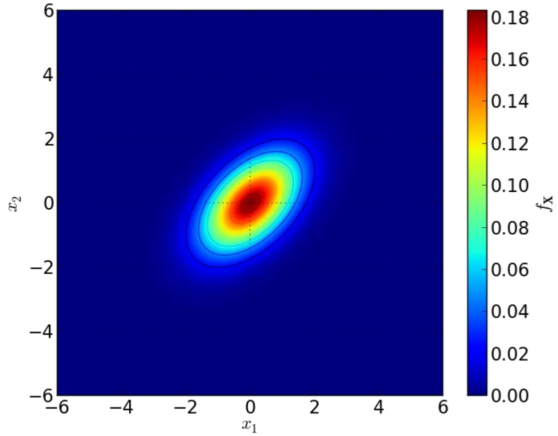
Multivariate normal distribution

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \sim \mathcal{N}_n \left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^{\mathrm{T}} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$

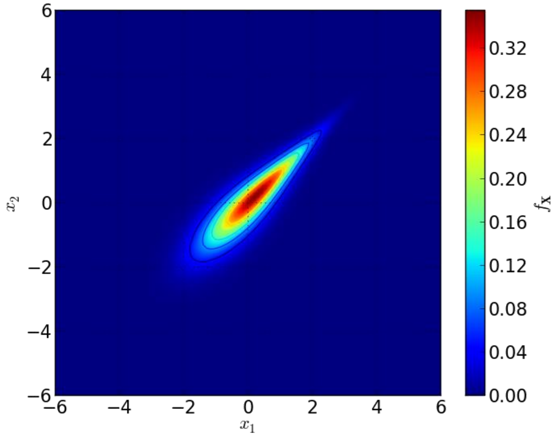


Random vectors

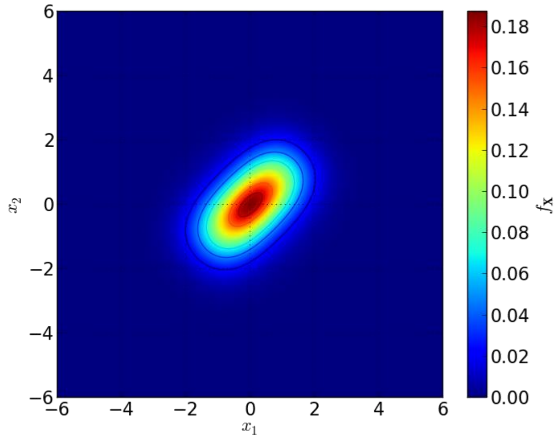
- Multivariate normal distribution with copulas



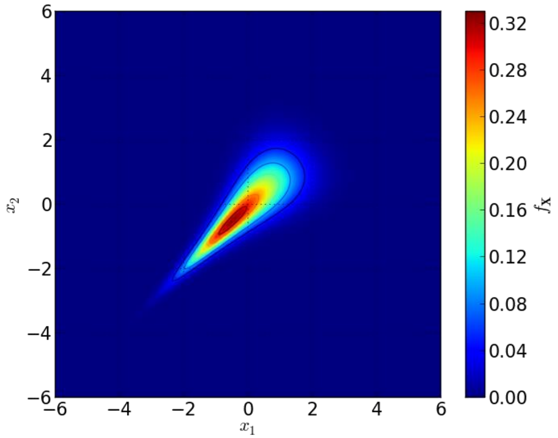
Gaussienne (ρ_0 = 0,5)



Gumbel (θ = 3)



Frank (θ = 3)



Clayton (θ = 3)

Random vectors

Synthesis

- Defined by a joint distribution …
- … or a collection of marginal distributions, and a copula if required.

- Used for multi-dimensional problems.

Some references

- Probability, Random Variables and Stochastic Processes, 4th Edition International Edition, Athanasios Papoulis, S. Unnikrishna Pillai, Mc Graw Hill (2002) (www.mhhe.com/engcs/electrical/Papoulis)

- Nelsen, Roger B. An introduction to copulas. Springer Science & Business Media, 2007.