Analyses statistiques des expériences numériques

Cours 1 : Introduction et rappels de base

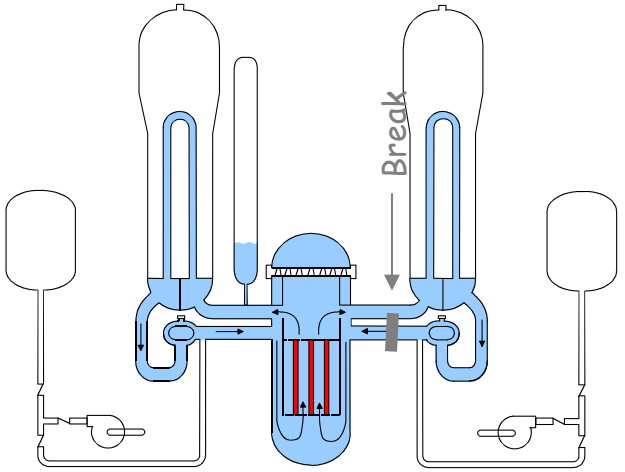
Bertrand Iooss

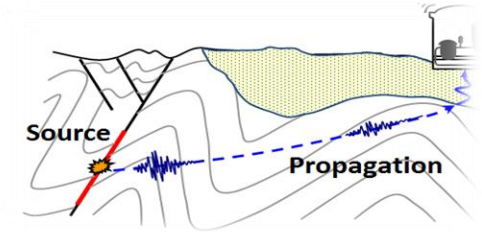
Polytech Nice Sophia

Décembre 2025

CHANGER L'ÉNERGIE ENSEMBLE

Décision sous incertitudes



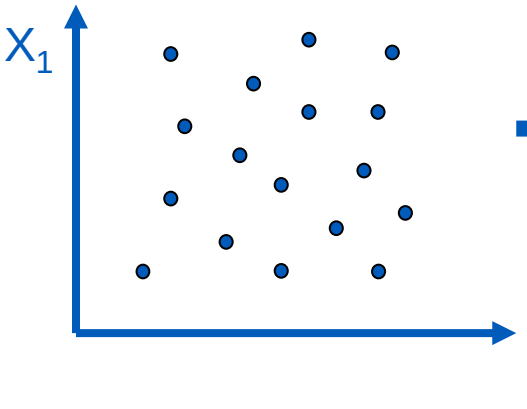


Utilisation de codes de calcul
Exploration de modèles

Plan d'expériences numériques

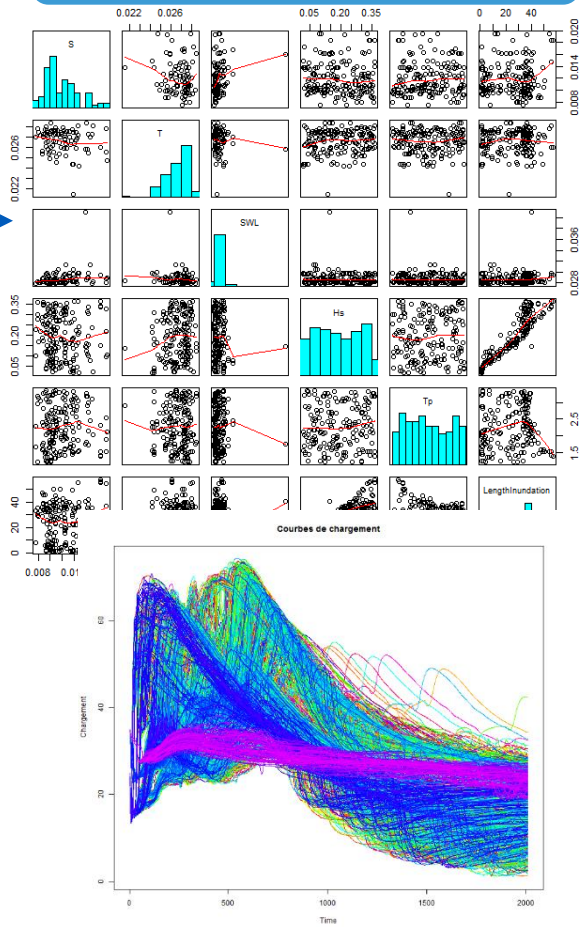
Simulations numériques

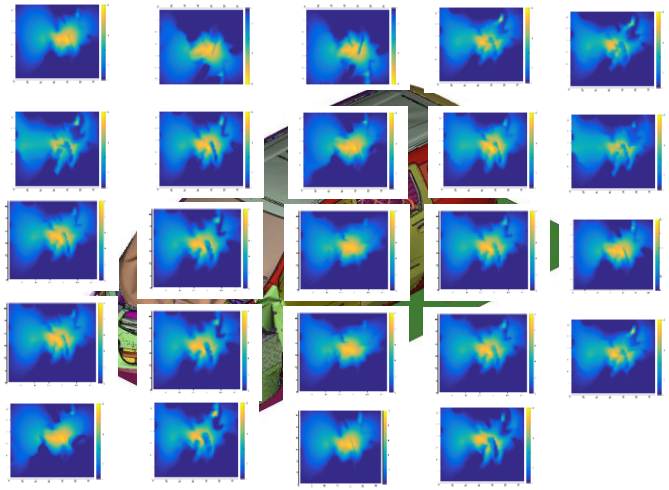
Analyse des sorties



Code de calcul

$Y = G(X)$





Analyse automatique d'un grand nombre de résultats, potentiellement volumineux

=> Technologies HPC, Big Data & IA

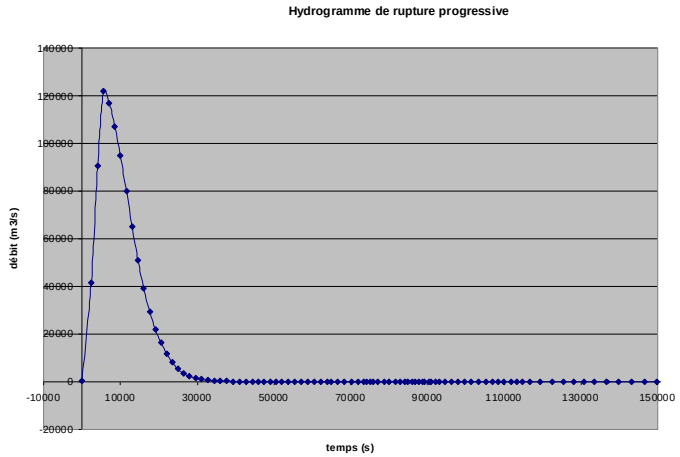
Exemple : simulation de rupture d'un barrage en terre

Juin 1976 – Teton dam - https://en.wikipedia.org/wiki/Teton_Dam

Exemple : simulation de rupture d'un barrage (1/2)

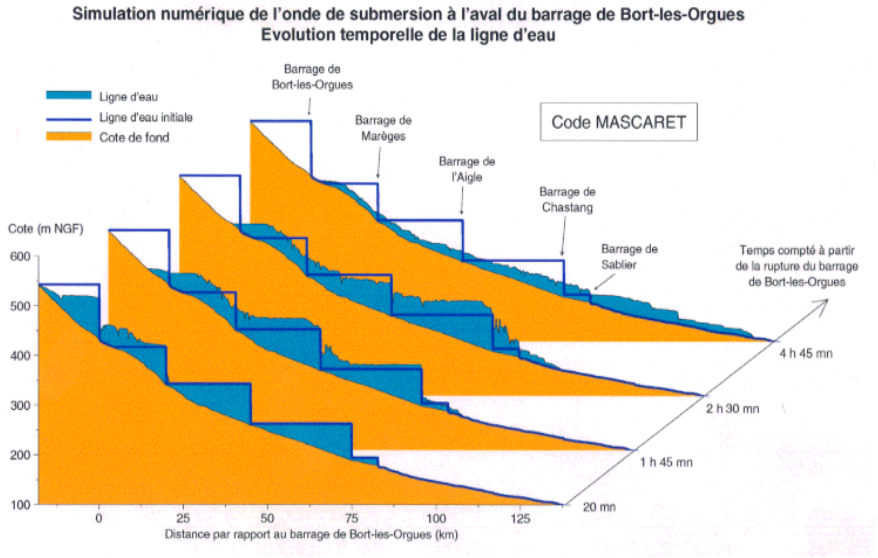
- L'objectif : évaluer la côte maximale de l'eau et le temps d'arrivée de l'onde de submersion

1. Les paramètres **fixes** : les caractéristiques du barrage (longueur/hauteur/épaisseur/volume d'eau etc.)
2. Les variables **aléatoires** :
   - La rugosité du fond de la rivière (modélisée par dire d'expert)
   - Les paramètres de l'hydrogramme de brèche (débit $Q_0(t)$) :
     - Temps de montée T_m
     - débit maximum Q_m

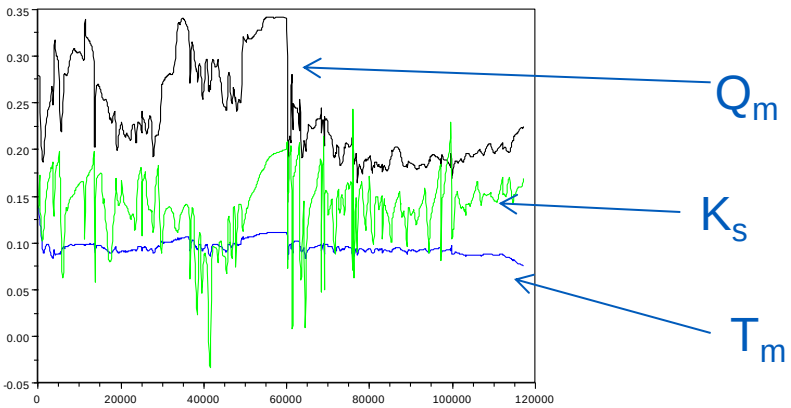


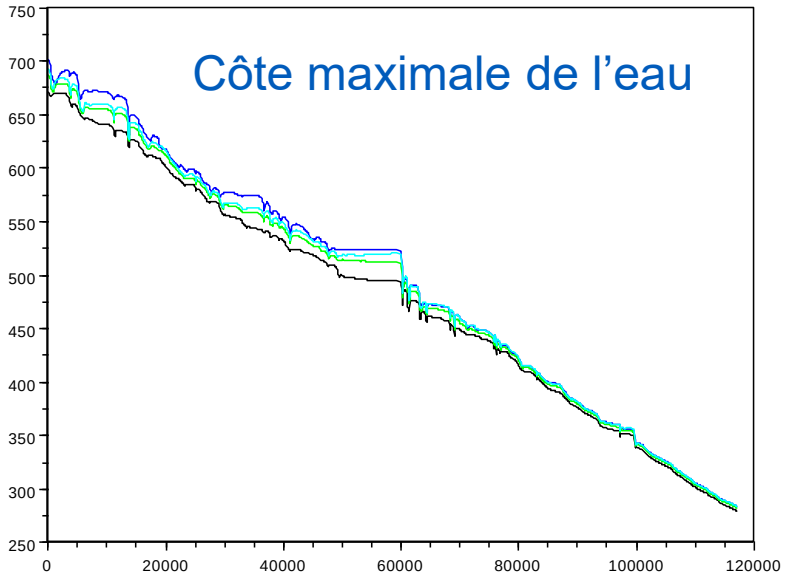
Exemple : simulation de rupture d'un barrage (2/2)

- Utilisation d'un **code de calcul** simulant l'hydraulique de l'inondation



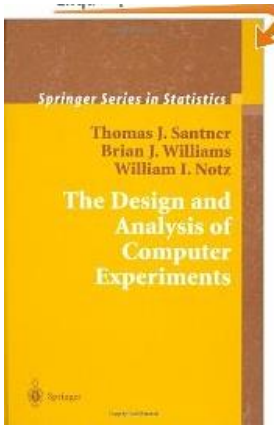
- Les données de sortie ou résultats :
  1. Calcul avec valeurs pessimistes, optimistes et de référence
  2. Calculs de quantiles et de probabilités de dépassement de seuil
  3. Analyse de sensibilité : influence des variables aléatoires sur l'incertitude que l'on a sur la côte maximale de l'eau



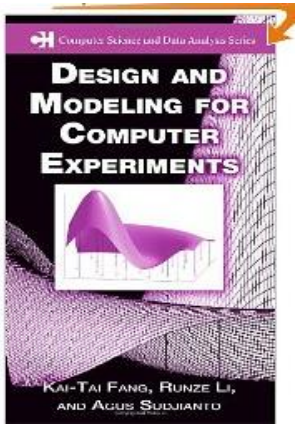


Domaine scientifique « Computer experiments »

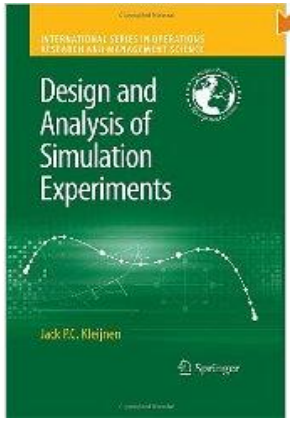
- **Design, analysis and modeling of computer experiments**



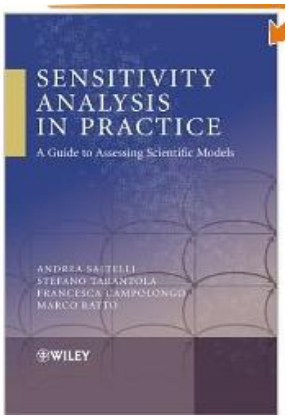
2003



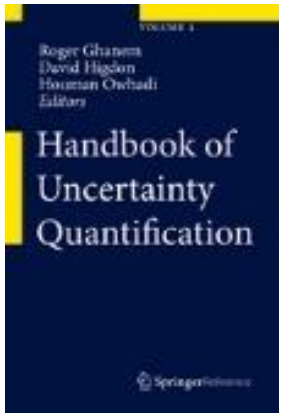
2006



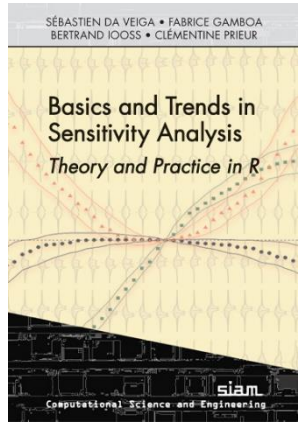
2008



2008



2017

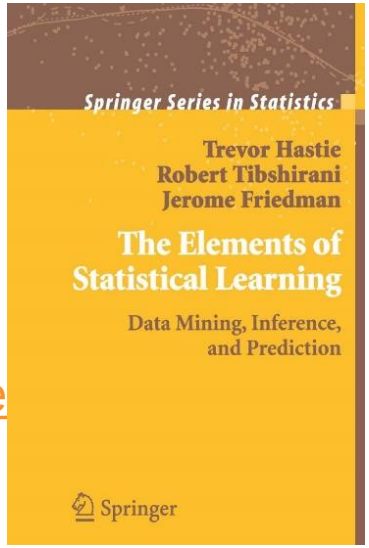


2021

- Communauté qui vient des maths numériques, optimisation, théorie de l'approximation, calcul scientifique, fiabilité des structures, statistiques industrielles et académiques, recherche opérationnelle, apprentissage, géostatistique, assimilation de données, …

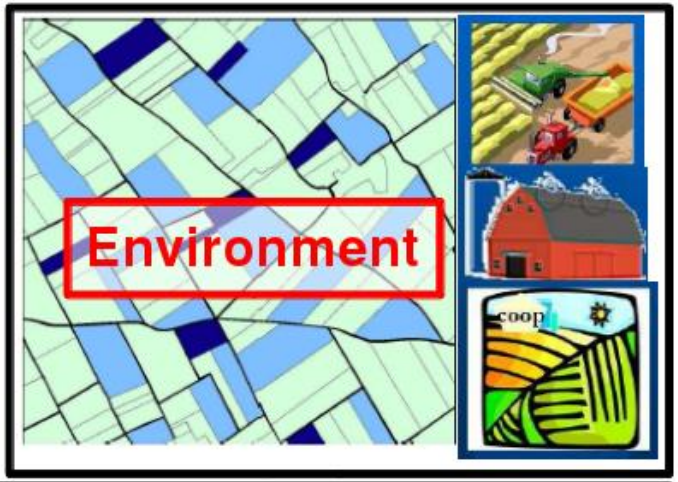
- Conf annuelle française RT-UQ (~ 120 pers.) - RT-UQ : Research ne Uncertainty Quantification [RT UQ - GdR MASCOT-NUM]

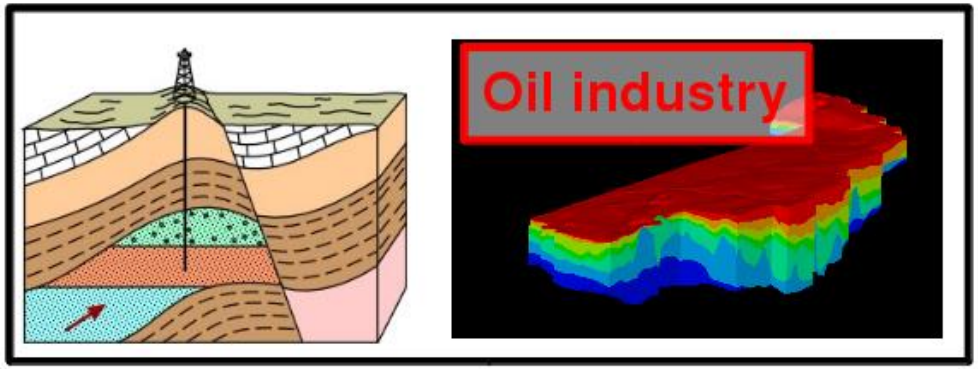
- Conf américaine SIAM tous les 2 ans (~ 800 pers. en 2022)

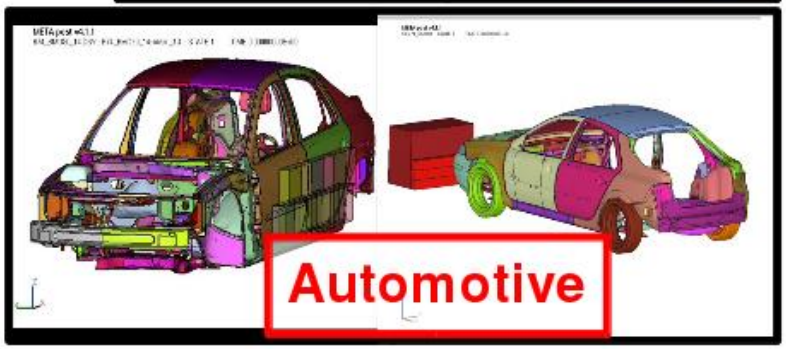


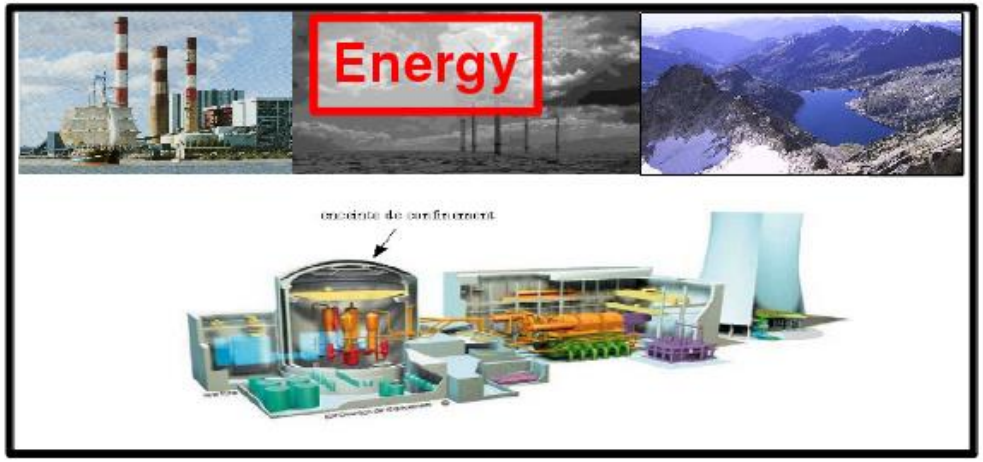
2001

Une problématique multi-sectorielle



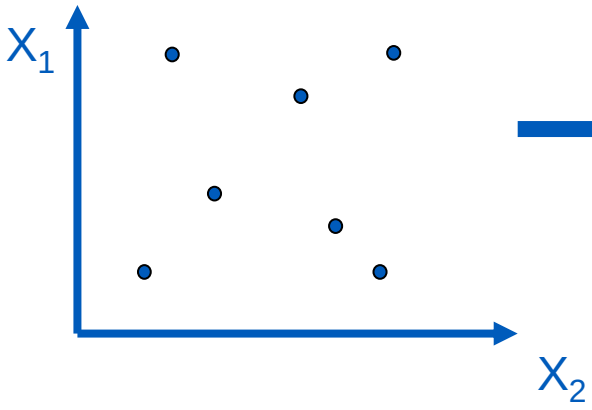






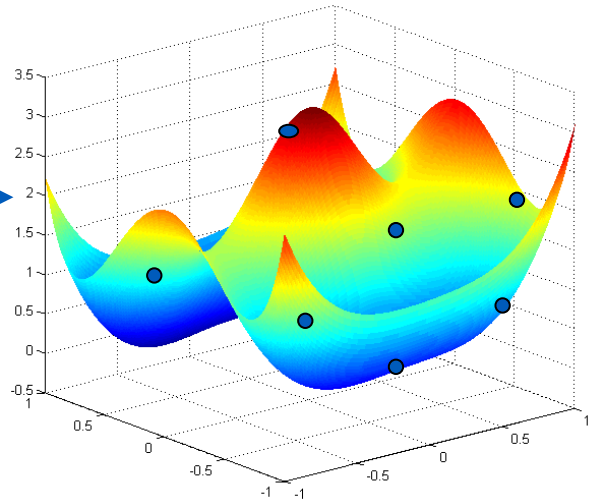
Simulation Analytics - Métamodèles

Plan d'expériences numériques → Simulations numériques → Analyse des sorties



Code de calcul

$Y = G(X)$



Problème du coût important
($€, cpu, tps d'études…)
=> Big & Small DATA

Métamodèle

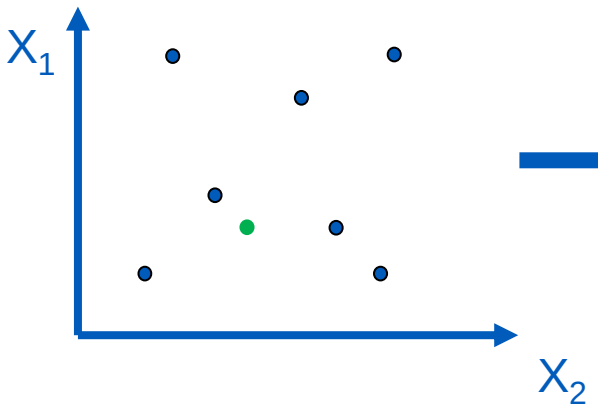
$Y_{app} = g(X)$

Apprentissage statistique (polynômes, processus gaussiens (ou krigeage), forêts aléatoires, réseaux de neurones, …) ou Réduction de modèles

Simulation Analytics - Métamodèles

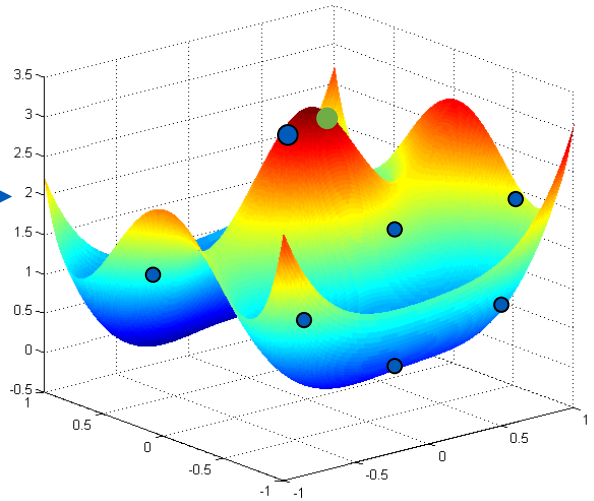
Spécification de nouveaux calculs pour l'objectif :
Exploration globale, Fiabilité (sûreté), optimisation (conception), …

Plan d'expériences numériques → Simulations numériques → Analyse des sorties



Code de calcul

$Y = G(X)$



Problème du coût important
($€, cpu, tps d'études…)
=> Big & Small DATA

Métamodèle

$Y_{app} = g(X)$

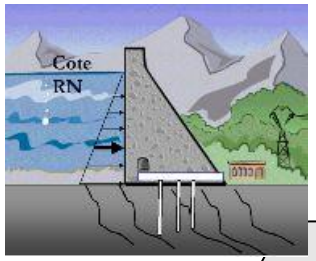
Apprentissage statistique (polynômes, processus gaussiens (ou krigeage), forêts aléatoires, réseaux de neurones, …) ou Réduction de modèles

Taxonomie des modèles d'apprentissage stat.

Methods	Applications	Pros	Cons
Linear models: Linear and logistic regressions	classification, regression	Good mathematical properties (proof), ~~easily interpretable~~	Poor performance in complex problems
Neighborhood models: KNN, Kmeans, Kernel density, Gaussian process	classification, regression, clustering, density estim.	Easy to use	Not efficient in high dimension (because distance-based)
Trees: decision trees, regression trees	classification, regression	Fast, applicable on any type of data Easily interpretable	Unstable, unable to cope with missing data
Bayesian models: naive Bayesian, Bayesian network	classification, density estimation	Probabilistic model (confidence in outputs), cope with missing data	Difficult with numerical data, complex to train
Combination of models: Random Forest, Adaboost, gradient boosting (XGboost)	classification, regression, clustering, density estim.	Efficient, with any type of data	Some of these approaches have several tuning parameters, not efficient on high-dimensional inputs (such as image)
Neural networks (NN), Deep Learning (DNN)	classification, regression	Easily adaptable, effective on high-dimensional data	Many parameters, computational complexity

Présence d'incertitudes dans toute la chaîne de modélisation

Crédibilité des résultats issus d'un modèle physico-numérique ?



Simplifications

incertitudes de modèle

Phénomène physique

physicien

Modèle physique

Approximation numérique

incertitudes numériques

mathématicien

Objet de ce COURS

incertitudes stochastiques

Données d'entrée

Bruit numérique

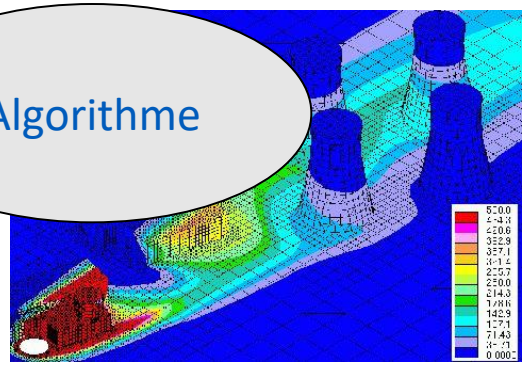
Variable d'intérêt

statisticien

Modèle numérique

informaticien

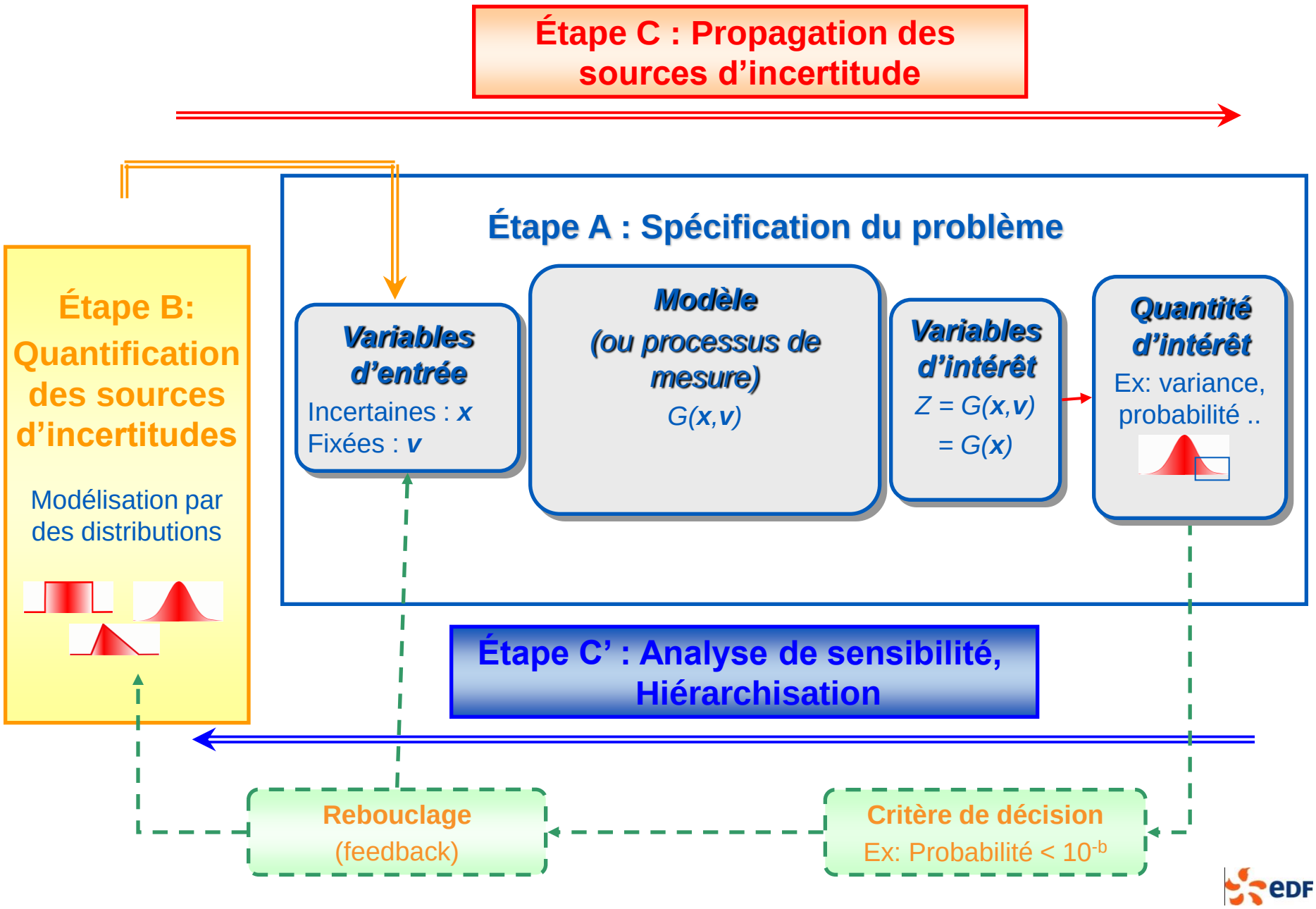
Algorithme



incertitudes épistémiques

Paramètres

Approches quantitatives : schéma générique introductif



Plan du cours 1

1. Introduction
2. **Rappels stats/probas**
3. Modélisation des sources d'incertitudes

Nous nous restreignons au cas où les sources d'incertitudes sont modélisées par des distributions de probabilité

Terminologie de base

Échantillon : sous-ensemble (de taille *n*) de la population sur lequel sont réalisées les observations

Variable X : $\Omega \rightarrow \Omega'$ (caractéristique définie sur la population) ;

- **Quantitative ($\Omega'=\Re$)**
  discrète *(ex : âge)* ou continue *(ex : poids)*
- **Qualitative ($\Omega'=V$)**
  nominale *(ex : sexe)* ou ordinale *(ex : mention)*

Données : ensemble des individus observés, des variables considérées et des observations de ces variables sur ces individus

Variable aléatoire

- **Variable aléatoire (X : $\Omega \rightarrow \Omega'$)** :

  Grandeur dépendant du résultat d'une expérience aléatoire (dont le résultat est non prévisible)

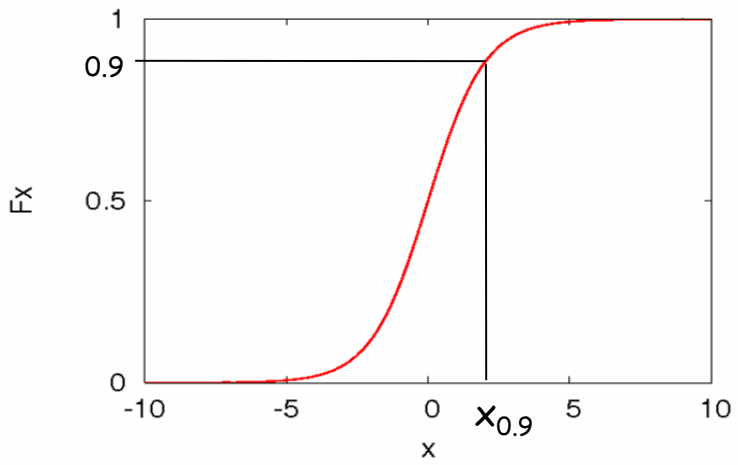
  Ex : choisir une caisse au supermarché, X = son temps d'attente

- **Réalisation :** *x* est une *réalisation* de *X* (valeur prise par *X*)

- **Fonction de répartition de *X* :**

$$F_X : \Omega' \rightarrow [0\ ;\ 1]$$
$$x \rightarrow F_X(x) = \mathrm{P}(X \leq x)$$

$$\lim_{x \rightarrow -\infty} F_X(x) = 0\ ;\ \lim_{x \rightarrow +\infty} F_X(x) = 1$$



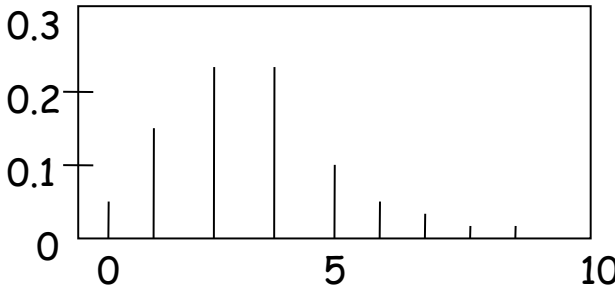
- **Quantile (ou fractile) d'ordre q :**

  x_q tel que $\mathrm{P}(X \leq x_q) = q$ $\Leftrightarrow$ $F_X(x_q) = q$

Variable aléatoire discrète

➢ Variable aléatoire discrète : ne prend qu'un nombre fini ou dénombrable de valeurs

$$X \in \{x_k, k \in K \subset \mathbb{N}\}$$

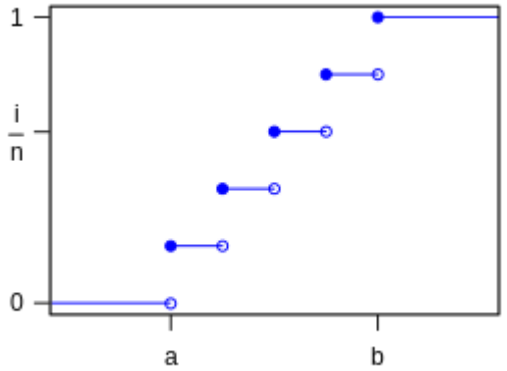


➢ Probabilité de chaque valeur $p_k = \mathrm{P}(X = x_k)$

Représentation graphique par diagramme en bâtons

Déterminer la loi d'une variable aléatoire discrète, c'est :

1. Déterminer l'ensemble des valeurs que peut prendre X
2. Calculer $\mathrm{P}(X = x_k)$ pour chacune de ces valeurs x_k



➢ Fonction de répartition de X :

$$F_X(x) = \mathrm{P}(X \leq x) = \sum p_k \, 1_{x_k \leq x}$$

➢ Espérance X : *la valeur que l'on s'attend à trouver, en moyenne, si l'on répète un grand nombre de fois la même expérience aléatoire, valeur « espérée »*

$$\mathrm{E}[X] = \sum p_k x_k$$

Variable aléatoire continue

- **$P(a < X \leq b) = F_X(b) - F_X(a)$**

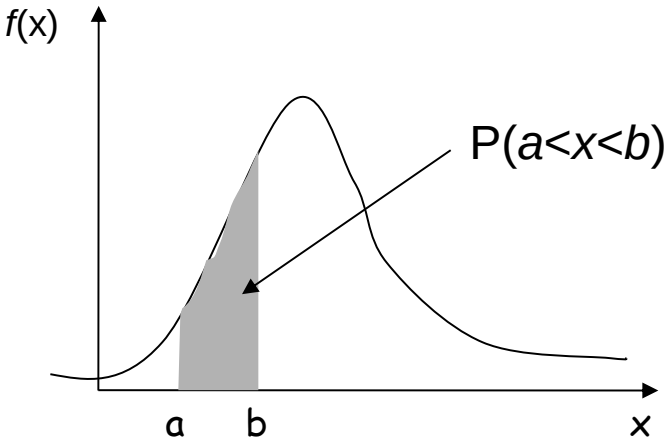
  Densité moyenne de probabilité sur $[a,b]$: $f_X(a,b)=[F_X(b)-F_X(a)]/(b-a)$

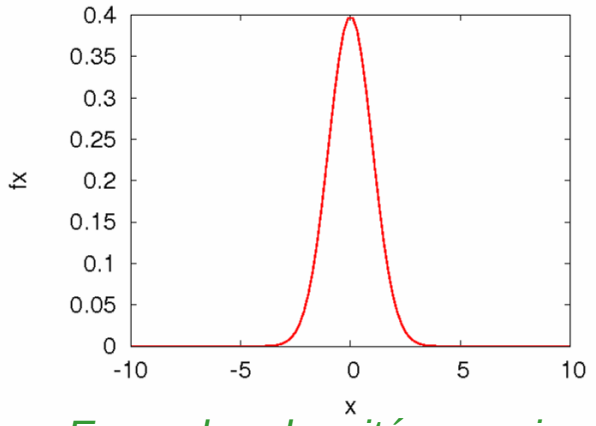
- **<u>Densité de probabilité f_X</u>** = dérivée de la fonction F_X

$$P(X \in I) = \int_I f_X(x)dx$$ pour tout intervalle I de $\Re$

- f_X est une fonction positive telle que $\int_{-\infty}^{\infty} f_X(x)dx = 1$ et $\lim_{x\to\pm\infty} f_X(x) = 0$

- Sa représentation graphique met en évidence les **zones à + forte probabilité.**

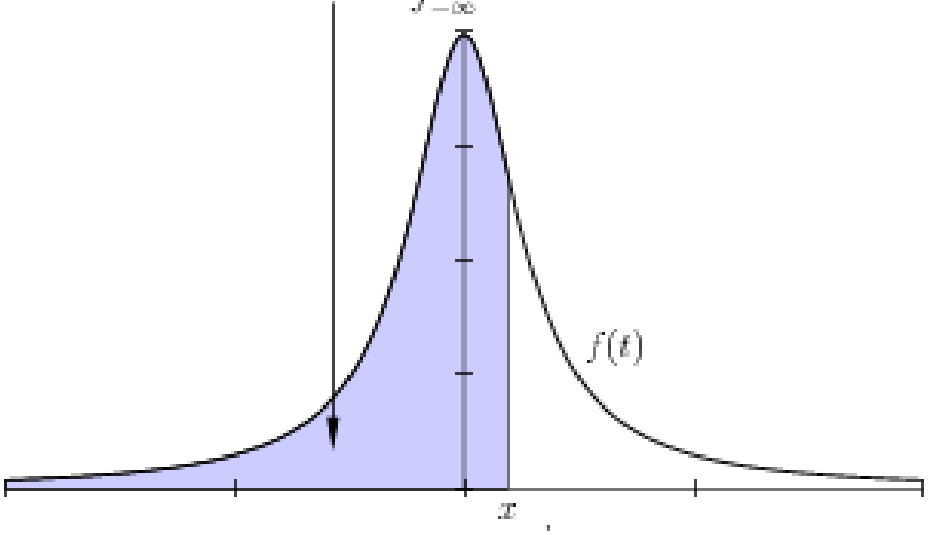




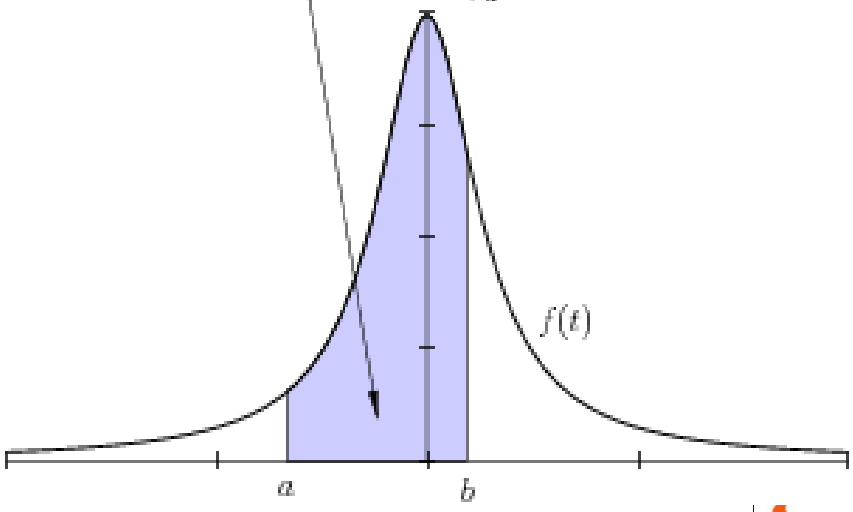
Exemple : densité gaussienne

Variable aléatoire continue

$$F(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$



$$\mathrm{P}(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(t)\,dt$$



Moments des variables aléatoires

➢ **Espérance mathématique** d'une v.a. continue :

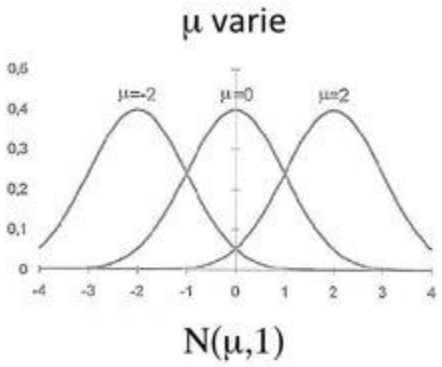
$$\mu = \mathrm{E}(X) = \int x f_X(x) dx$$

Propriétés : Indicateur de tendance centrale

$$\mathrm{E}[aX + b] = a\,\mathrm{E}[X] + b$$

Remarque : l'existence de E(X) n'est pas garantie

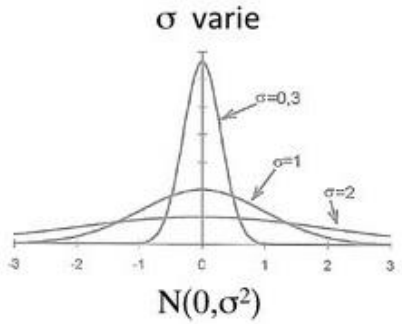
(ex : $f(x) = \dfrac{1}{\pi(x^2+1)}$ *)*



➢ **Variance** d'une variable aléatoire :

$$\sigma^2 = Var(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right]$$

$$\sigma^2 = \int [x-\mu]^2 f_X(x) dx = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$$



Propriétés : Indicateur de dispersion

$$\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X) \quad ; \quad \mathrm{Var}(X + b) = \mathrm{Var}(X)$$

Remarque : variance nulle ⟺ *variable aléatoire certaine*

Covariance et coefficient de corrélation linéaire

- Quantité impliquant deux variables aléatoires (v.a.) *X* et *Y*
- $\mathrm{cov}(X,Y) = \mathbb{E}\left((X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))\right) \longrightarrow \mathrm{cov}(X,X) = \mathbb{V}(X)$
  - Intuitivement, la covariance est une mesure des variations simultanées de v.a. Une grande valeur (en valeur absolue) de covariance signifie que *X* et *Y* varient “de la même manière” (relation positive, directe, croissante) ou dans le sens opposé (relation négative, inverse, décroissante).
- Propriétés:

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

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

$$\mathrm{X\ et\ Y\ indep.} \Rightarrow \mathrm{cov}(X,Y) = 0 \longrightarrow$$ L’inverse n’est pas vraie

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\mathrm{cov}(X,Y) \longrightarrow$$ Variance of la somme de 2 v.a.

- Coefficient de corrélation linéaire:

$$\varrho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathbb{V}(X) \cdot \mathbb{V}(Y)}} \in [-1,1]$$

Principales lois discrètes utilisées

➢ **Loi uniforme :** $X = \{1,2,\ldots,n\}$ avec $P(X=k) = 1/n$

$$E(X) = \frac{n+1}{2} \; ; \; \mathrm{var}(X) = \frac{n^2-1}{12}$$

Exemple : lancement d'un dé

➢ **Loi de Bernouilli $\mathcal{B}(p)$:**

$$E(X) = p \; ; \; \mathrm{var}(X) = p(1-p)$$

$$X = \begin{cases} 1 \text{ avec une proba } p & \textit{(succès)} \\ 0 \text{ avec une proba } 1\text{-}p & \textit{(échec)} \end{cases}$$

➢ **Loi binomiale $\mathcal{B}(n,p)$:** n répétitions indépendantes d'une Bernouilli

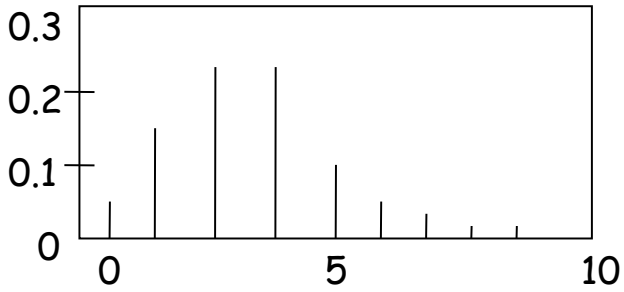
$$X = \sum_{i=1}^{n} X_i \quad \Rightarrow \quad P(X=k) = C_n^k p^k (n-p)^{n-k}$$

Exemple : sondage (OUI=1, NON=0)

p faible, n grand

➢ **Loi de Poisson $\mathcal{P}(\lambda)$** : loi du nombre d'occurrences d'événements « rares », sans mémoire et dans un intervalle de temps donné

$$P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!} \; ; \; E(X) = \mathrm{var}(X) = \lambda$$

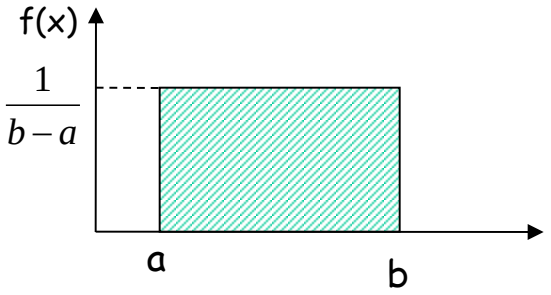


Exemples :
nombre de personnes dans une file d'attente,
nombre d'appels à un standard

Principales lois continues utilisées (1/3)

➢ Loi uniforme $\mathcal{U}$ [a,b]

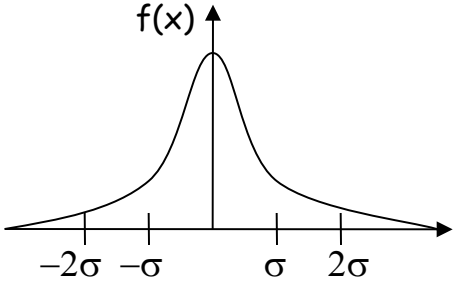
$$f(x) = \frac{1}{b-a} \text{ si } a \le x \le b \; ; \; f(x) = 0 \text{ ailleurs}$$

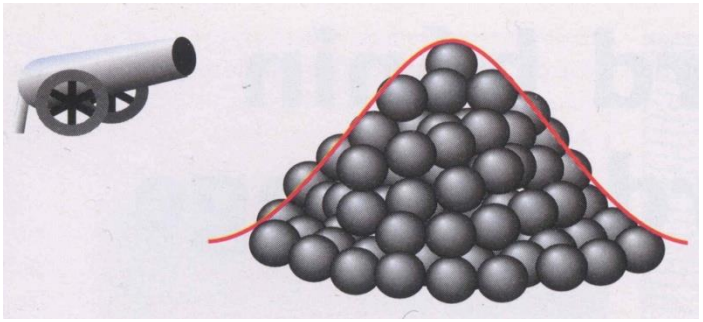


➢ Loi normale $\mathcal{N}(\mu,\sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\mathrm{E}(X) = \mu \; ; \; \mathrm{var}(X) = \sigma^2$$





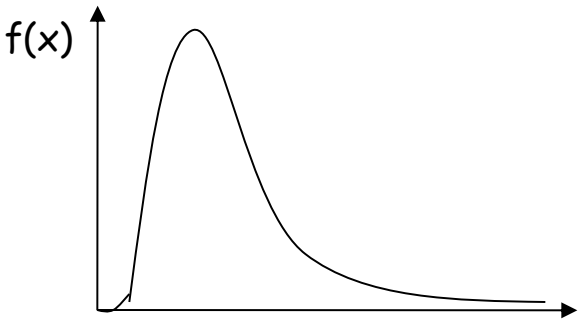
$P(\mu - \sigma < X < \mu + \sigma) = 0.68$
$P(\mu - 1.64\sigma < X < \mu + 1.64\sigma) = 0.90$
$P(\mu - 1.96\sigma < X < \mu + 1.96\sigma) = 0.95$
$P(\mu - 3.09\sigma < X < \mu + 3.09\sigma) = 0.998$

Exemples : impacts des boulets de canon (Jouffret, 1872), incertitude de mesure

Principales lois continues utilisées (2/3)

➢ **Loi du Chi-deux** : si $X_i \sim \mathcal{N}(0,1)$ pour i=1,…,n alors $\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$

➢ **Loi lognormale $\mathcal{LN}(\mu,\sigma^2)$:** $\ln(X) \sim \mathcal{N}(\mu,\sigma^2)$



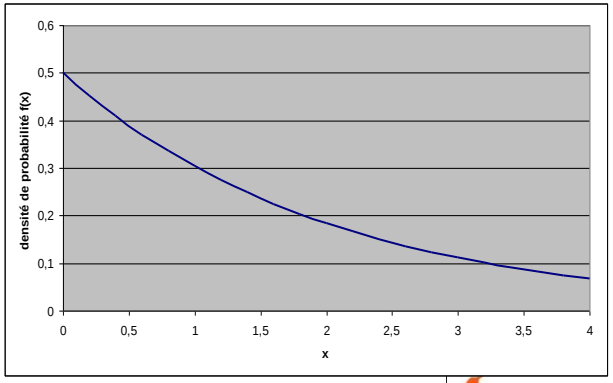
Le produit de v.a. $\xrightarrow{\mathcal{L}}$ $\mathcal{LN}$

Exemples : variables positives et asymétriques (poids, salaires, …), résolution d'un instrument (sources d'erreur = multiplication d'un grand nombre de petits facteurs indépendants)

➢ **Loi exponentielle $\mathcal{E}(\lambda)$:** $f(x) = \lambda \exp(-\lambda x)$ si $x \geq 0$;

$$E(X) = \frac{1}{\lambda} \; ; \; \mathrm{var}(X) = \frac{1}{\lambda^2}$$

Exemples : temps d'attente, durée de vie de systèmes sans usure ⇨ i.e. la proportion de matériels défaillants est chaque année la même

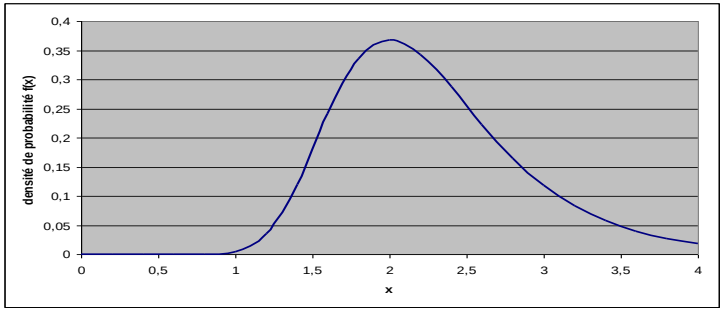


Principales lois continues utilisées (3/3)

➢ **Loi de Gumbel $\mathcal{G}(m,s)$:** $f(x) = \frac{1}{s}\exp\left(-\frac{x-\mu}{s}\right)\exp\left(-\exp\left(-\frac{x-\mu}{s}\right)\right)$

- Densité de probabilité fortement asymétrique autour du mode *m*

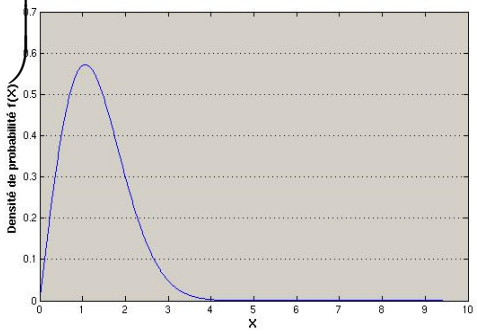
- les fortes valeurs restent probables



Exemple : modélisation des phénomènes climatiques extrêmes (débit d'une rivière)

➢ **Loi de Weibull $\mathcal{W}(x_0,\alpha,\beta)$:** $f(x) = \frac{\alpha}{\beta}\left(\frac{x-x_0}{\beta}\right)^{\alpha-1}\exp\left(-\left(\frac{x-x_0}{\beta}\right)^{\alpha}\right)$

Généralisation de la loi exponentielle



Exemples en mécanique : durée de vie d'un matériel qui :

- *se dégrade pour $\alpha>1$ (structure acier)*
- *ou se bonifie pour $\alpha<1$ (résistance du béton en début de vie)*

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}	$\text{Prob}(X = k) = C_n^k p^k (1-p)^{n-k}$	np	np(1-p)
Poisson(λ), $\lambda \geq 0$	0 ; 1 ; 2 ; ...	$\text{Prob}(X = k) = \exp(-\lambda)\frac{\lambda^k}{k!}$	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	]-∞ ; + ∞[	$f(x;m,s) = \frac{1}{s}\exp\left[-\left(\frac{x-m}{s}\right)\right]\exp\left[-\exp\left\{-\left(\frac{x-m}{s}\right)\right\}\right]$	$m + \gamma s$ $\gamma = 0{,}577222$	$\frac{1}{6}\pi^2 s^2$

Plan du cours 1

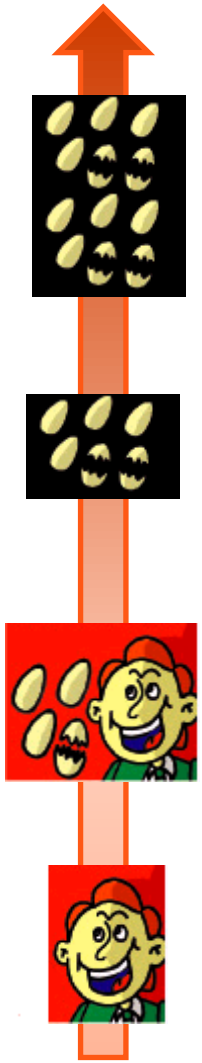
1. Introduction

2. Rappels stats/probas

3. **Modélisation des sources d'incertitudes**

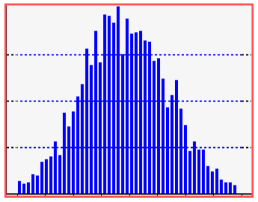
Nous nous restreignons au cas où les sources d'incertitudes sont modélisées par des distributions de probabilité

OUTILS DE QUANTIFICATION / NOMBRE DE DONNÉES DISPONIBLES



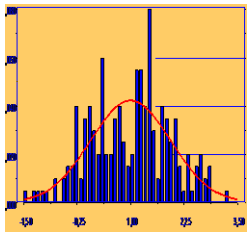
Enormément de données : approches **non-paramétriques**

- Estimation de la loi par **histogramme** / **méthode à noyaux**
- Nécessite **beaucoup plus de données** que les approches paramétriques car ne nécessite aucune hypothèse sur la loi



Beaucoup de données : approches **fréquentistes paramétriques**

- Le plus souvent, estimation ponctuelle des paramètres par **maximum de vraisemblance**, complétée par des int. de confiance
- Tests d'adéquation pour 'valider' le choix d'un modèle paramétrique

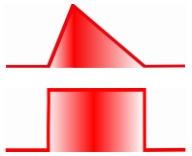


Peu de données : approches **bayésiennes paramétriques**

- Combinaison d'information **objective** (les données) et **subjective** (l'avis d'expert)
- Fournit une distribution d'incertitudes sur les valeurs possibles des paramètres définissant la loi

Rév. Thomas Bayes, 1701 - 1761

Pas de données : quantification de la loi d'incertitude par **élicitation d'avis d'experts**



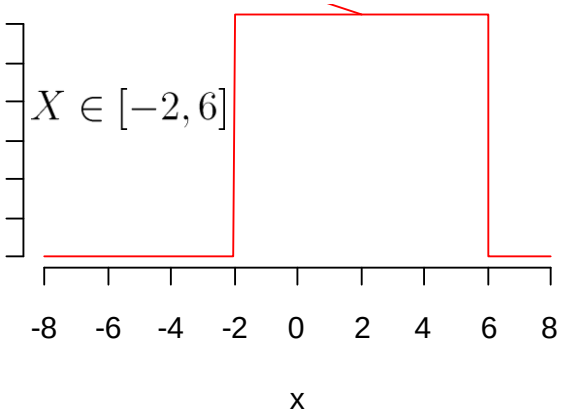
Cas sans données - Application du Max d'entropie (1/2)

$$H(X) = -\int_{\mathcal{X}} f(x)\log\left(f(x)\right)dx$$

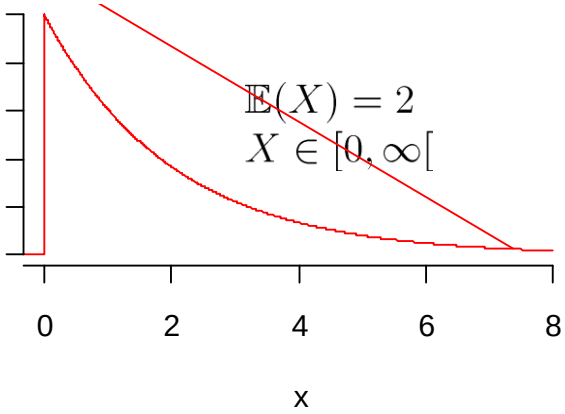
Principe du maximum d'entropie : parmi toutes les lois possibles, choisir celle qui apporte le minimum d'information (« objectivité ») → celle qui maximise l'entropie

Information fournie	Distribution maximisant l'entropie
$X \in [a,b]$	Loi uniforme $X \sim \mathcal{U}(a,b)$
$\mathbb{E}(X) = \mu$ $X \in [0,\infty[$	Loi exponentielle $X \sim \mathcal{E}(1/\mu)$
$\mathbb{E}(X) = \mu$ $\mathbb{V}(X) = \sigma^2$	Loi gaussienne $X \sim \mathcal{N}(\mu,\sigma)$

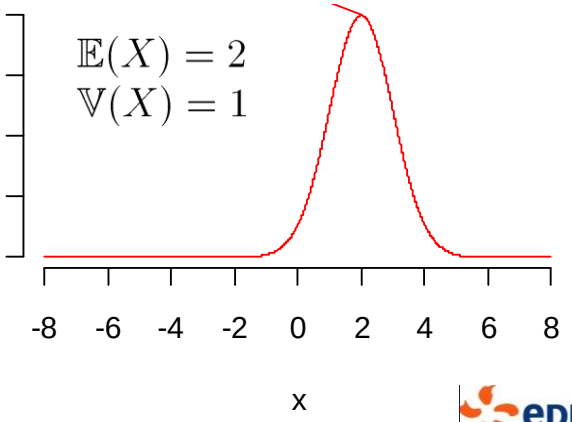
Loi uniforme



Loi exponentielle

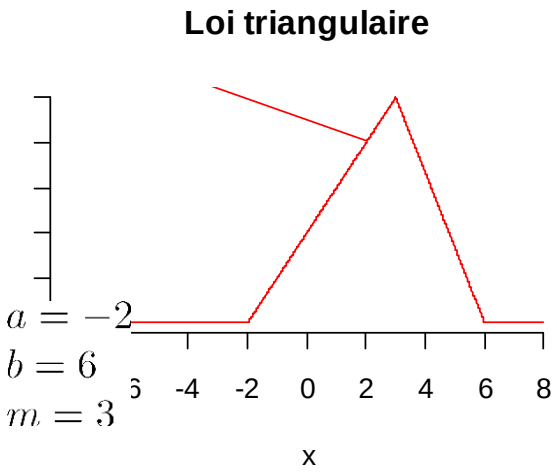


Loi normale



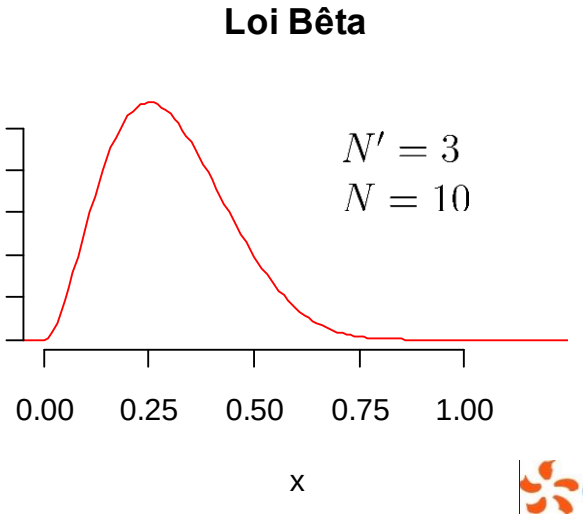
Cas sans données - Application du Max d'entropie (2/2)

- Loi triangulaire $\mathcal{T}(a, b, m)$
  - Quand l'expert fournit un intervalle et un mode *m* (valeur la plus probable)



- Loi Bêta $\mathcal{B}(\alpha, \beta)$
  - Si la v.a. est la probabilité d'un événement
  - L'expert fournit un nombre de « succès » *N'* sur la base de *N* expériences « virtuelles »

$$\alpha = N'$$
$$\beta = N - N'$$



Cas avec données

- Problème :
  - A partir d'un échantillon i.i.d. de la v.a. *X* : $x^{(1)}, x^{(2)}, ..., x^{(n)}$
  - Reconstruire la loi de probabilité de *X*, pour :
    - Prédire des moments, des quantiles de *X*
    - Simuler aléatoirement la v.a. *X* (ex. Monte Carlo)
    - …
  - On se focalisera sur des v.a. uni-dimensionnelles.

  Le problème spécifique des variables multi-dimensionnelles et de la modélisation de la dépendance, qui fait appel à la notion de copules, ne sera pas vu ici

- Ajustement non paramétrique
- Ajustement paramétrique
- Contrôle de la qualité de l'ajustement

Crédits & Bibliographie

- Tutoriel « Incertitudes », EDF R&D
- De Rocquigny, Devictor & Tarantola (eds), Uncertainty in industrial practice, Wiley, 2008
- Y. Dodge, Premiers pas en statistique, Springer, 2001
- G. Saporta, Probabilités, Analyse des données et Statistique, Ed.Technip, 1990
- M. Lejeune, Statistique : la théorie et ses applications, Springer Verlag, 2004

Annexe : Ajustement de lois de probabilité

Cas avec données : Ajustement non-paramétrique

Intéressant

- en présence d'une grande quantité de données
- en présence de loi de forme non usuelle, par ex. avec plusieurs modes

- Fonction de répartition empirique
- Histogramme empirique
  - Outils « basiques » pour l'ingénieur
- Reconstruction de la densité par noyaux

Fonction de répartition empirique

- Échantillon i.i.d. de taille *n* de *X* : $x^{(1)}, x^{(2)}, ..., x^{(n)}$

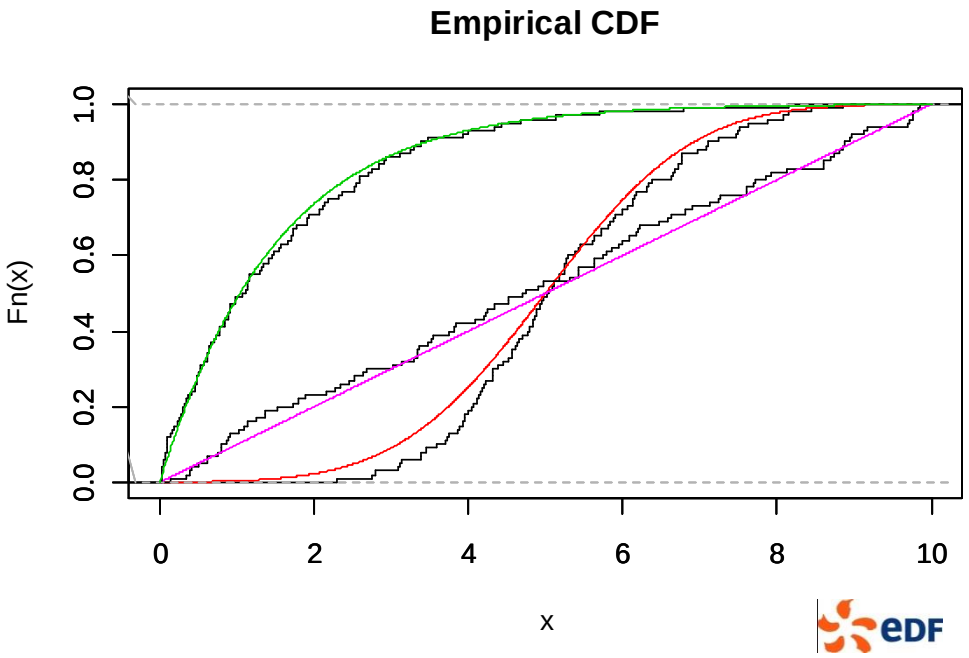
- Fonction de répartition empirique :
  - Proportion des observations ≤ à une valeur fixée x de la v.a. *X*

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x^{(i)} \leq x\}}$$

$$\hat{F}_n(x) \to F(x) \quad \text{p.s.}$$

- « Inversion » de la fonction de répartition empirique
  - Quantile empirique :

$$\hat{x}_p = \inf \left(z : \hat{F}_n(x) \geq p \right)$$



Estimation par histogramme de la densité

- Diviser le domaine de *X* en *m* intervalles de longueurs égales *h*

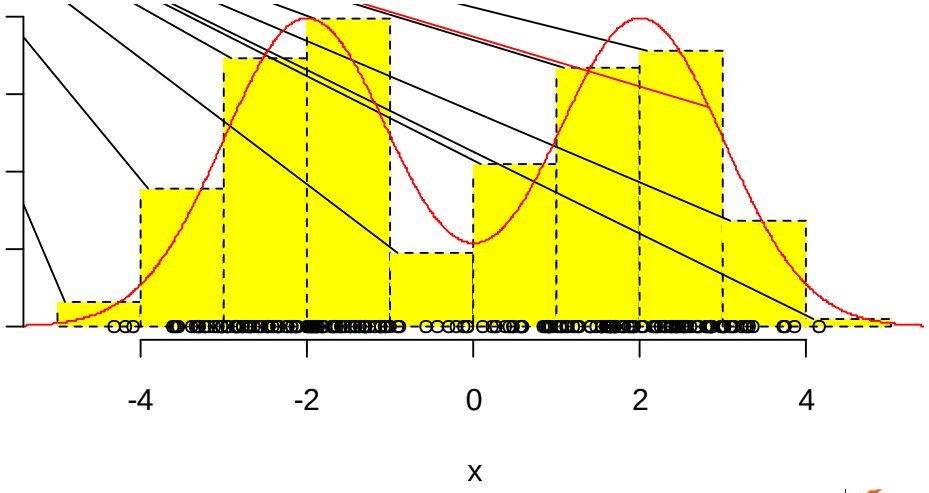
$$]x^* + jh, x^* + (j+1)h] \quad j \in \mathbb{N}$$

- Estimation de la densité de *X* par la fonction en escalier :

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}_{\{x^{(i)} \in \mathcal{I}(x)\}} \qquad \mathcal{I}(x) \ \text{ intervalle contenant } x$$

Nombre de points de
l'échantillon qui se trouvent dans
le même intervalle que x

- L'estimation par noyaux est inspirée par l'estimation par histogramme

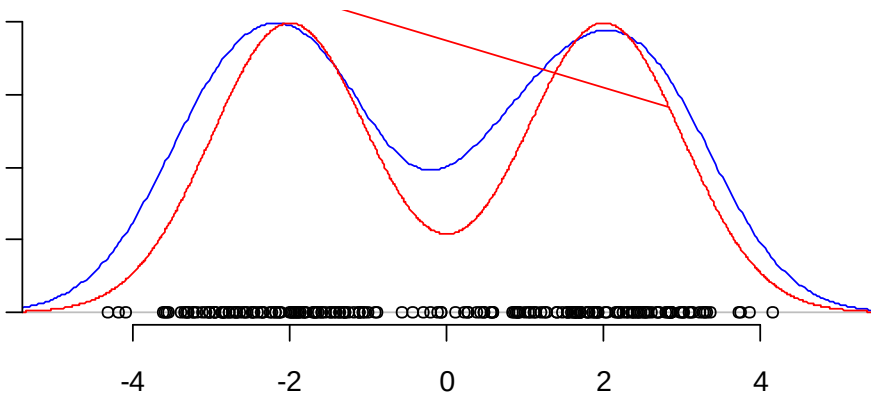


Estimation par noyaux

- Estimation de la densité de *X* : $\hat{f}_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^{N} K\left(\frac{x - x^{(i)}}{h}\right)$
  - h est appelé « largeur de bande »
    - Paramètre de lissage, plus h est grand, plus la densité est « lisse »
  - K est une fonction, dite « noyau », positive et telle que : $\int_{\mathcal{X}} K(x)dx = 1$
  - Le noyau est, en général, une densité symétrique, p.ex. une loi normale $\mathcal{N}(0,1)$ ce qui donne :

$$K\left(\frac{x - x^{(i)}}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h^2}}$$

  - D'autres noyaux : triangulaire, rectangulaire, Epanechnikov …



Cas avec données : Ajustement paramétrique
Estimation des paramètres d'une loi

En général, si possible, mieux vaut utiliser l'ajustement paramétrique car les lois obtenues sont plus facilement manipulables

- **Une loi de type donnée (normale, gamma, Weibull,…) est caractérisée par un ensemble de paramètres.**

- **Au vu des données disponibles, quelles valeurs numériques choisir pour ces paramètres?**

- **Quelques méthodes**
  - Maximum de Vraisemblance
  - Méthode des moments

- **Robustesse des estimations obtenues?**
  - Incertitude inhérente au nombre de données utilisées
  - Contrôle via des intervalles de confiance, + ou - difficiles à obtenir selon la situation (formules analytiques, résultats asymptotiques, bootstrap…)

Estimation par maximum de vraisemblance (1/2)

- **Fonction de vraisemblance**

$$\mathcal{L}(x^{(1)}, \ldots, x^{(n)}|\theta) = \prod_{i=1}^{n} f_\theta(x^{(i)}) \quad \text{cas continu}$$

$$\mathcal{L}(x^{(1)}, \ldots, x^{(n)}|\theta) = \prod_{i=1}^{n} P_\theta(x^{(i)}) \quad \text{cas discret}$$

- **Estimateur par max. de vraisemblance (ML)**
  - Valeur de θ qui maximise la vraisemblance (ou la log-vraisemblance)
  - De manière intuitive, on cherche la valeur qui maximise la « probabilité » d'observer l'échantillon donné
  - C'est un problème d'optimisation

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta}{\mathrm{ArgMax}} \left[\mathcal{L}(x^{(1)}, \ldots, x^{(n)}|\theta)\right] = \underset{\theta}{\mathrm{ArgMax}} \left[\log\left(\mathcal{L}(x^{(1)}, \ldots, x^{(n)}|\theta)\right)\right]$$

Estimation par maximum de vraisemblance (2/2)

- **Si la vraisemblance est dérivable (2 fois)**

$$\left.\frac{\partial \mathcal{L}}{\partial \theta}\right|_{\theta=\hat{\theta}_{\mathrm{ML}}} = 0 \qquad \left.\frac{\partial^2 \mathcal{L}}{\partial \theta^2}\right|_{\theta=\hat{\theta}_{\mathrm{ML}}} \leq 0$$

- **Quelques exemples classiques**

Normale $\mathcal{N}(\mu, \sigma)$	$\hat{\mu}_{\mathrm{ML}} = \overline{x}, \quad \hat{\sigma}_{\mathrm{ML}} = (1/n)\sum_i (x^{(i)} - \overline{x})^2$ avec $\overline{x} = (1/n)\sum_i x^{(i)}$
Exponentielle $\mathcal{E}(\lambda)$	$\hat{\lambda}_{\mathrm{ML}} = 1/\overline{x}$
Uniforme $\mathcal{U}(a, b)$	$\hat{a}_{\mathrm{ML}} = \min(x^{(i)}), \quad \hat{b}_{\mathrm{ML}} = \max(x^{(i)})$ $i = 1 \ldots n$ Cas où la dérivée ne s'annule jamais

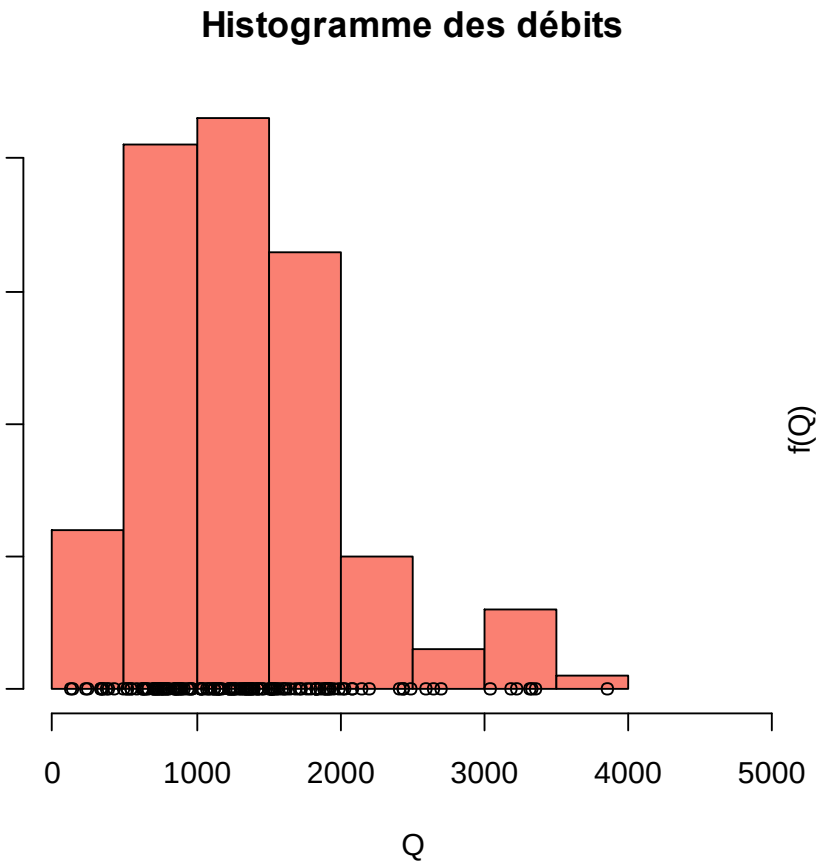
Contrôle de la qualité de l'ajustement

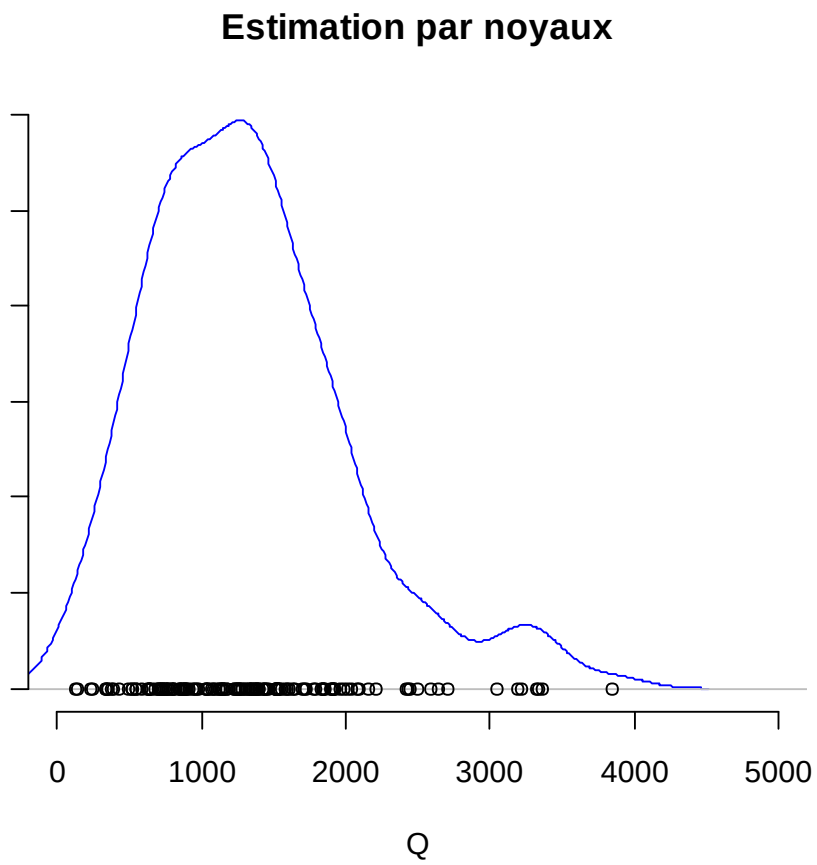
- Ajustement graphique
  - Superposition des fonctions de répartition théoriques et empiriques
  - QQ plot

- Tests d'adéquation
  - Kolmogorov - Smirnov
  - Cramer – Von Mises
  - Anderson – Darling
  - …

- Exemple : ajuster une loi de probabilité sur 149 données de maxima de débits annuels d'une rivière

Exemple d'ajustement (1/2)



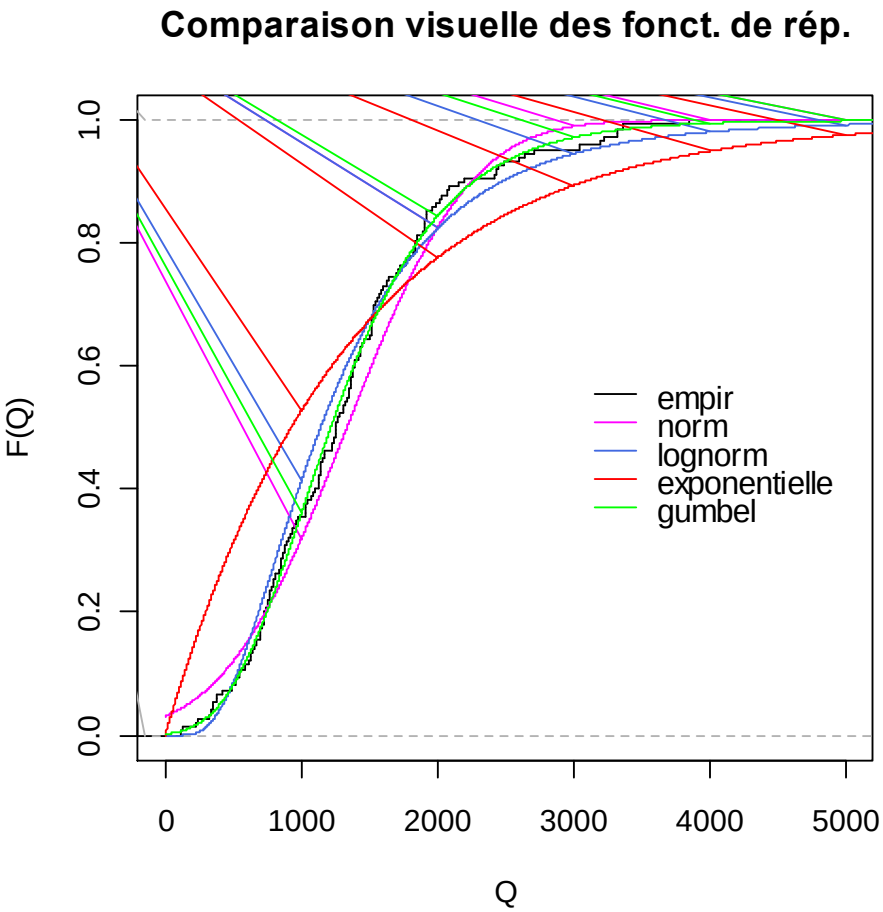


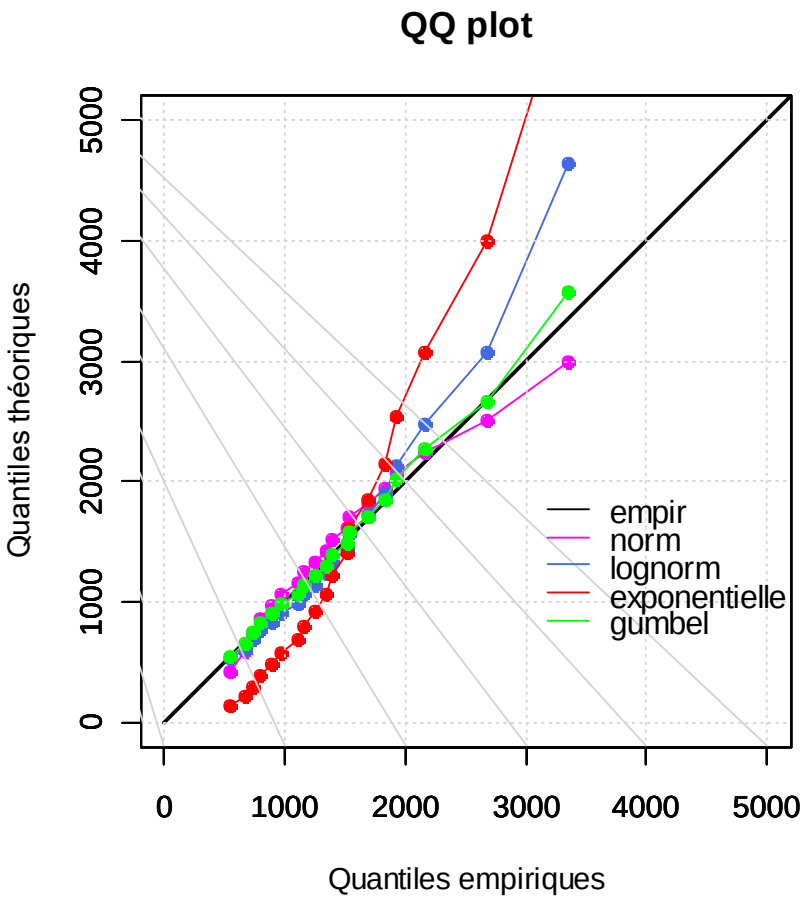
Exemple d'ajustement (2/2) - Estimation par maximum de vraisemblance

Ajustement paramétrique (max de vraisemblance)

- Loi normale
  - $(\hat{\mu}, \hat{\sigma}) = (1335, 711)$
- Loi log-normale
  - $(\hat{\mu}_{\log}, \hat{\sigma}_{\log}) = (7.0, 0.60)$
- Loi exponentielle
  - $\hat{\lambda} = 1/1335$
- Loi de Gumbel
  - $(\hat{\mu}, \hat{\beta}) = (1013, 557)$

Contrôle « visuel » de la qualité de l'ajustement





Contrôle de la qualité de l'ajustement - tests (1/3)

- Test d'adéquation
  Hypothèse H_0 : l'échantillon est une réalisation de la loi donnée
  Hypothèse H_1 : H_0 est fausse

- Ce test se base sur l'évaluation d'une fonction des données (statistique de test) qui, sous l'hypothèse H_0, suit une loi connue

- Niveau de signification α : probabilité de rejeter à tort l'hypothèse H_0

- Pour un test unilatéral (à droite), la règle de décision est :

$$\text{Accepter } H_0 \text{ si } \tau(x^{(1)}, \ldots, x^{(n)}) \leq \tau_{1-\alpha}$$

Valeur de la stat. de test pour l'échantillon donné

Quantile d'ordre 1-α de la stat. de test, sous l'hypothèse H_0 → Cette quantité est connue (tables, logiciels stat)

Contrôle de la qualité de l'ajustement - tests (2/3)

- Quelques tests
  - Kolmogorov – Smirnov (écart maximal entre fcts théorique et empirique)
    - $\tau_{KS} = \sup_x \sqrt{n}\,|F_n(x) - F(x)|$
  - Cramer – Von Mises (bien pour ajustement global)
    - $\tau_{CM} = \int_{-\infty}^{+\infty} [F_n(x) - F(x)]^2\, dF(x) =$
      $\frac{1}{12n} \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - F(x^{(i)})\right]^2$ ← Après avoir ordonné l'échantillon
  - Anderson – Darling (bien pour queues de distribution)
    - $\tau_{AD^2} = n \int_{-\infty}^{+\infty} \frac{[F_n(x) - F(x)]^2}{F(x) \cdot (1 - F(x))} dF(x) =$
      $-n - \frac{1}{n} \sum_i \left[\log(F(x^{(i)}) + \log(1 - F(x^{(n-i+1)})\right]$

Contrôle de la qualité de l'ajustement - tests (3/3)

	Loi normale	Loi log-normale	Loi Gumbel
Kolmogorov – Smirnov $\tau_{95\%} = 0.11$	$\tau_{KS} = 0.091$ p-val = 0.17	$\tau_{KS} = 0.087$ p-val = 0.20	$\tau_{KS} = 0.043$ p-val = 0.94
Cramer – Von Mises $\tau_{95\%} = 0.46$	$\tau_{CM} = 0.29$ p-val = 0.17	$\tau_{CM} = 0.23$ p-val = 0.21	$\tau_{CM} = 0.038$ p-val = 0.94
Anderson Darling	$\tau_{AD} = 2.08$ p-val = 0	$\tau_{AD} = 1.44$ p-val = 0.02	$\tau_{AD} = 0.25$ p-val = 1

pvalue = probabilité de rejeter à tort l'hypothèse H_0

1 - pvalue = niveau de confiance avec lequel on peut rejeter l'hypothèse H_0

- Préférence pour la loi de Gumbel
- Attention au faible pouvoir discriminant des tests d'adéquation
- Autres critères de sélection (basés sur le rapport des vraisemblances) :

$$\mathrm{AIC} = 2k - 2\log(\mathcal{L})$$
$$\mathrm{BIC} = k\log(n) - 2\log(\mathcal{L})$$