Traitement des incertitudes en simulation numérique

Cours 1 : Introduction, modélisation et propagation d' incertitudes

Bertrand looss

Module INSA Toulouse/GMM 5
Planification, risque et incertitudes

14 novembre 2012

EDF
CHANGER L'ÉNERGIE ENSEMBLE

Module Traitement des incertitudes en simulation numérique

Depuis une trentaine d'années, l'industrie a développé des processus et des codes de calcul parfois très lourds pour modéliser des phénomènes complexes !
La plupart des ingénieurs sont amenés à manipuler ces codes & processus

1) Il est nécessaire d'**optimiser leur utilisation pour prendre des décisions !**
=> *Analyse de sensibilité, planification d'expérience, développement de modèles réduits*

2) La **validation de leurs résultats** est un problème crucial lorsqu'ils sont utilisés dans des cycles industriels (conception, sûreté, prévision, etc.)
=> *Gestion des incertitudes, calculs fiabilistes*

3 cours de 3h15 pour INSA GMM 5 & Master Pro 2 UPS
1. Cours 1 : Introduction, modélisation et propagation d'incertitudes
2. Cours 2 : Planification et analyse d'expériences numériques
3. Cours 3 : Modélisation d'expériences numériques, krigeage

3 séances de TP pour INSA GMM 5
1. TP 1 :Exercices en R
2. TP 2 : Exercices en R
3. TP 3 : Exercices en R

Une note sera délivrée via des compte-rendus réalisés à l'issue des TPs

Plan du cours 1

1. **Introduction**
2. Modélisation des sources d'incertitudes
3. Propagation des incertitudes

Contexte général : management des risques

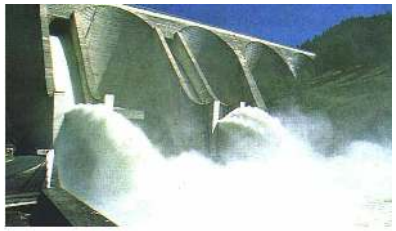
1. Divers cadres : conception, opération, maintenance et contrôle d'installations industrielles

A EDF: 58 centrales nucléaires, 14 centrales thermiques, 220 barrages, …

2. Un objectif industriel général : maximiser la production en **minimisant les risques**, sous des contraintes fortes de sûreté, disponibilité, qualité (au sens coût et fiabilité du kW)

3. Malgré des incertitudes importantes…

- La plupart des données sont aléatoires : niveau de la demande, niveau d'eau dans les barrages, prévisions météorologiques (+/-1°C = une variation importante de la demande!)



- La connaissance des systèmes - donc de leur comportement et leur fiabilité - est incomplète : problématique de vieillissement des composants, variabilité des caractéristiques des matériaux, …

- Des agressions internes - conditions d'exploitation anormales, défaillance de systèmes, incident d'exploitation - ou externes - inondations, tempêtes, séisme, crue - peuvent survenir et perturber le cycle normal d'exploitation d'une tranche

⇒ **Analyser, mesurer, quantifier les incertitudes pour réduire les risques et maîtriser leur impact**

Enjeux EDF : La sûreté, la disponibilité & la performance

- **Augmenter la sûreté des installations** en enrichissant les démonstrations de sûreté :
  - identification des scénarios accidentels pénalisants,
  - prise en compte du retour d'expérience événementiel,
  - calcul de données de fiabilité de composants importants pour la sûreté

- **Augmenter la disponibilité des installations** ⇒ **quantifier et hiérarchiser les risques** pour optimiser les performances par une **meilleure évaluation des marges**

- Gestion des accidents nucléaires ou hydrauliques

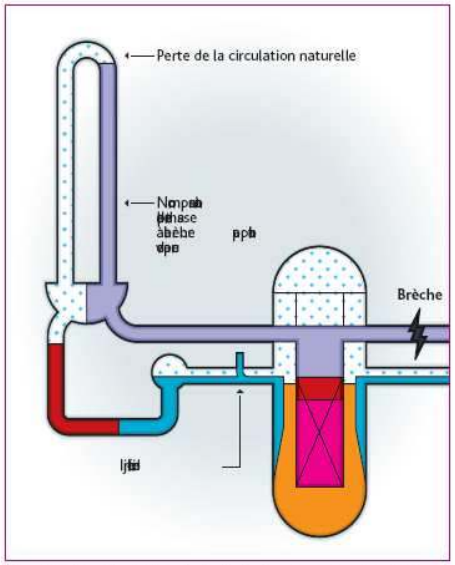
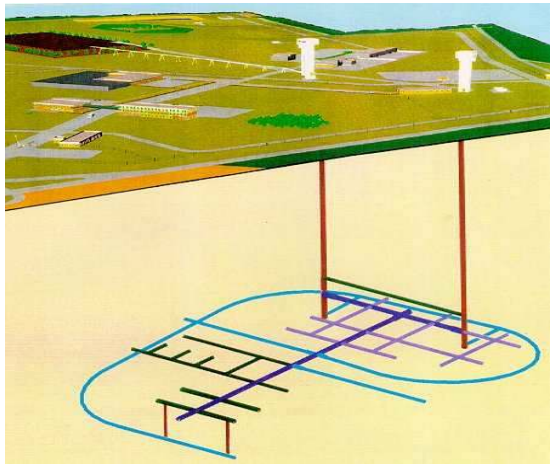
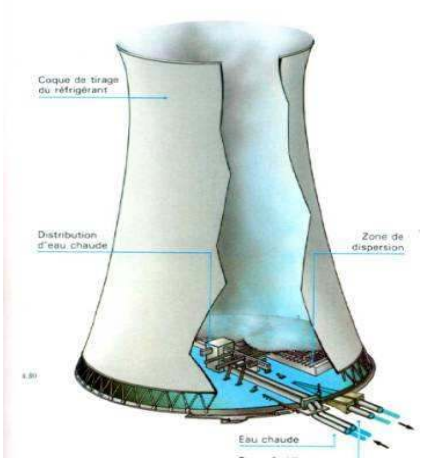


Figure 1 : Schéma d'une brèche du circuit primaire entraînant un choc froid diphasique.

- Démantèlement, stockage des déchets radioactifs



- Vieillissement : génie civil

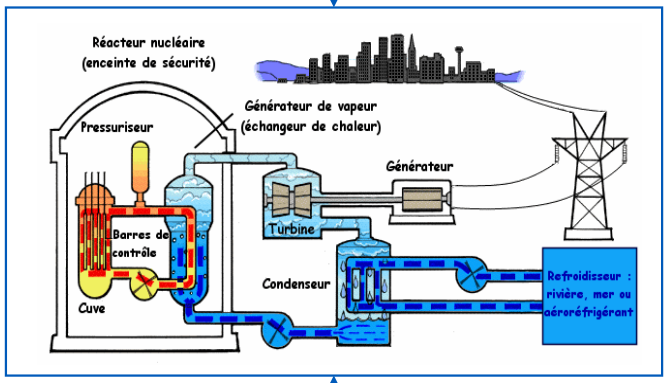


La problématique : des données aux méthodes

Ex : probabilité de défaillance d'une pompe

Un REX de défaillances « important »

STATISTIQUE CLASSIQUE OU FREQUENTIELLE
(loi des grands nombres)



Un REX « faible » mais des avis d'expert

POINT DE VUE BAYESIEN
(lois a priori et a posteriori)

Ex : probabilité de rupture d'une tuyauterie

Ex : probabilité de rupture d'un barrage

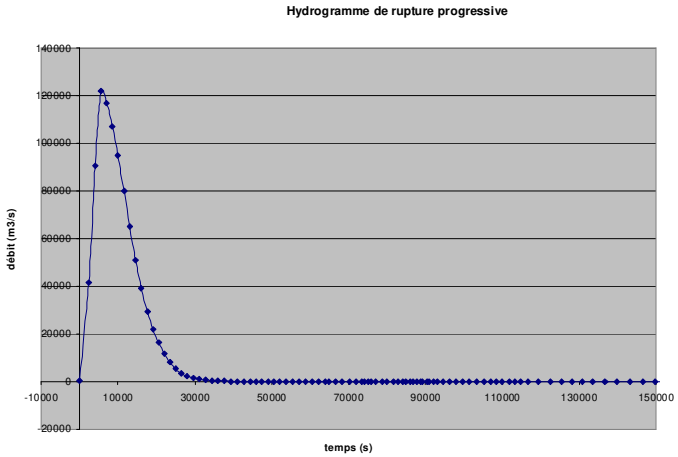
Pas de REX de défaillance mais un modèle physique

INCERTITUDES EN SIMULATION NUMERIQUE

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

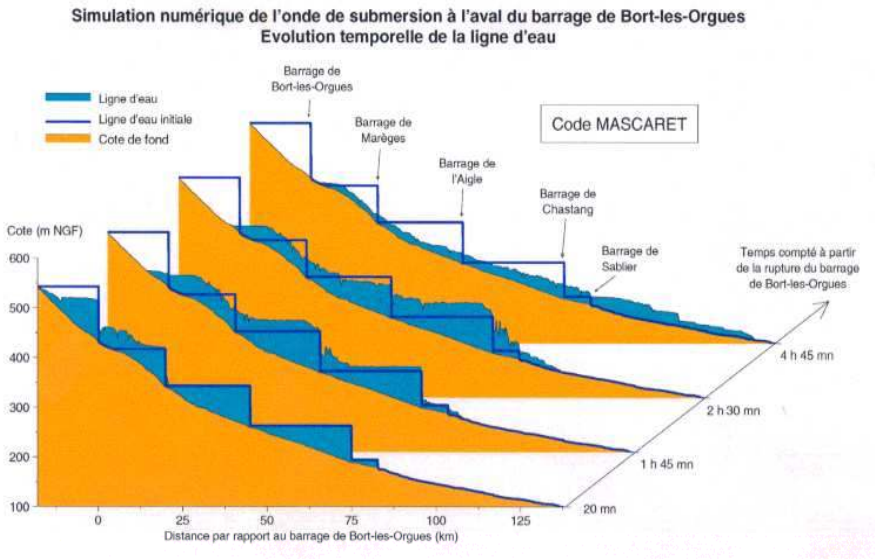
- L'objectif : évaluer la côte maximale de l'eau et le temps d'arrivée de l'onde de submersion

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

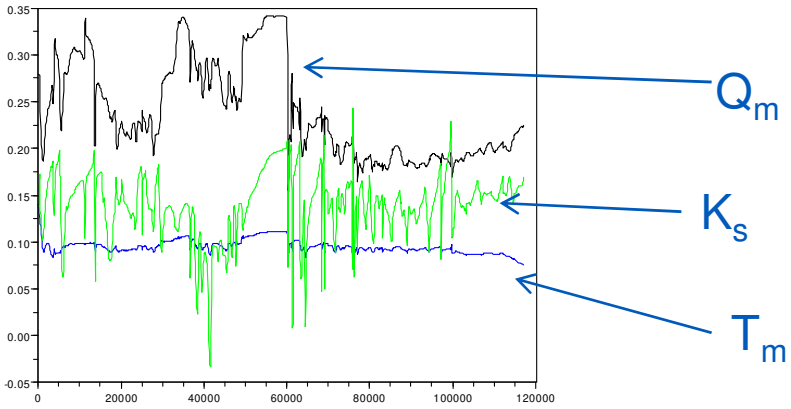


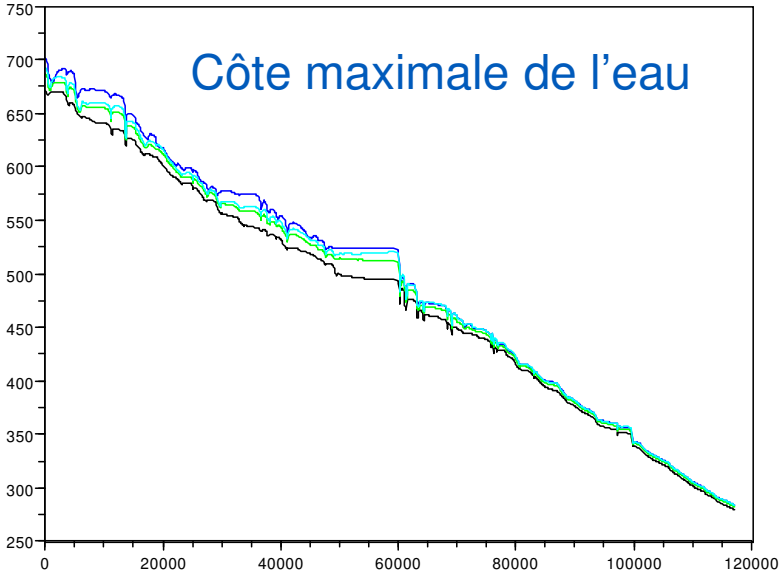
Exemple 1 : simulation de rupture d'un barrage (2/2)

- 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





Exemple 2 : dispersion de particules dans l'atmosphère (1/3)

Scénario d'accident de rejet radioactif

Domaine d'étude : 10 km autour d'un site industrielle

Deux sources arbitraires (au niveau du sol) :
* source 1 : traceur (gaz)
* source 2 : iode (particules)

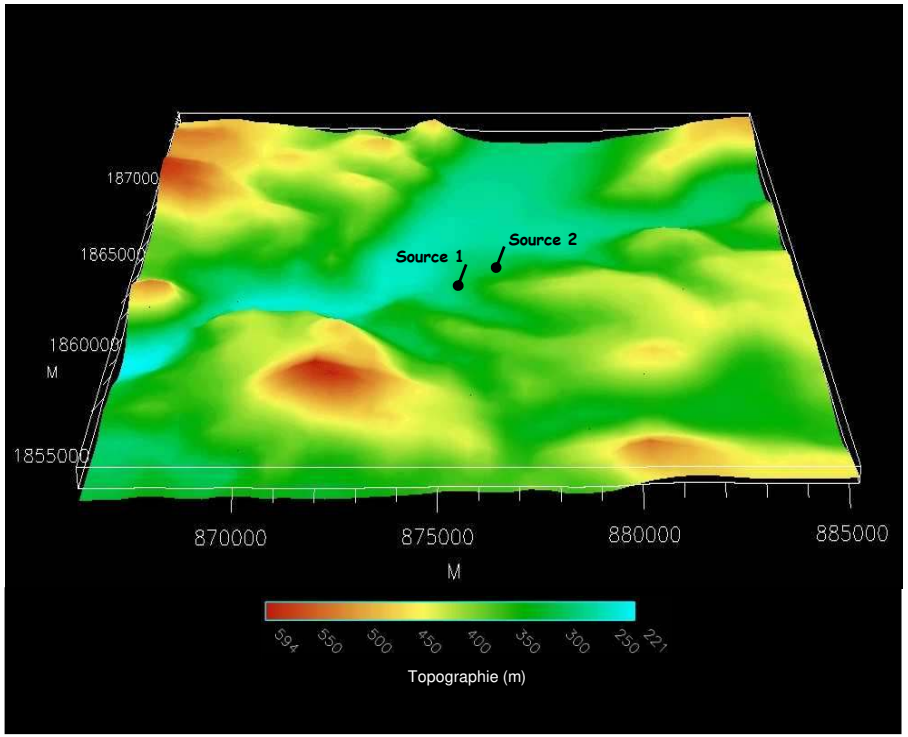
Quantités émises : 1.10^{13} Bq sur 10 heures

Calcul sur 4 jours

Données Météo : vent, température, humidité, pluie

Rugosité du sol (végétation)

Topographie et localisation des sources



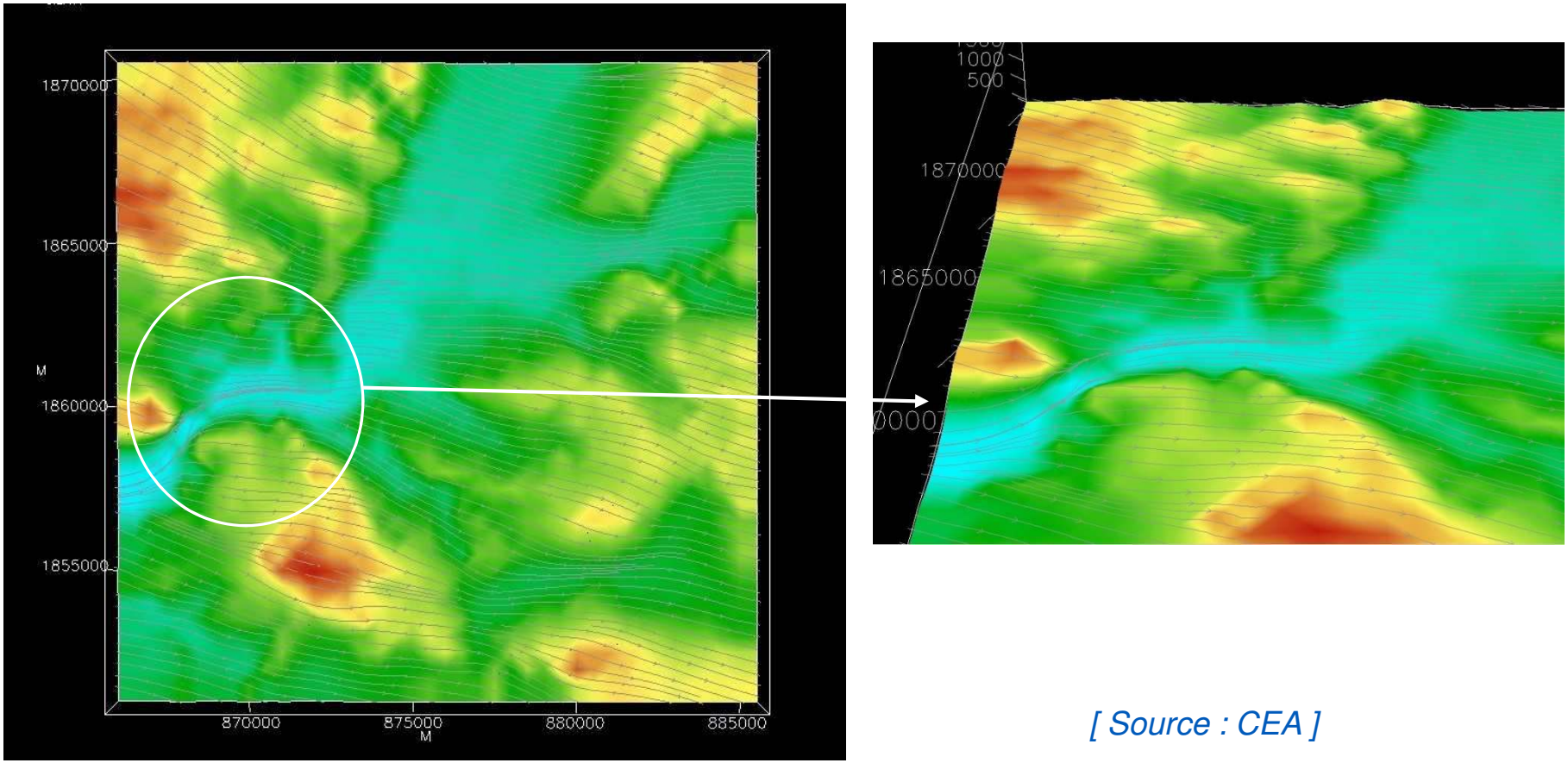
[Source : CEA]

Exemple 2 : dispersion de particules dans l'atmosphère (2/3)

Calcul des champs de vent (direction et amplitude)

Visualisation du vent sous forme de ligne de courant

Mise en évidence des circulations de vallées

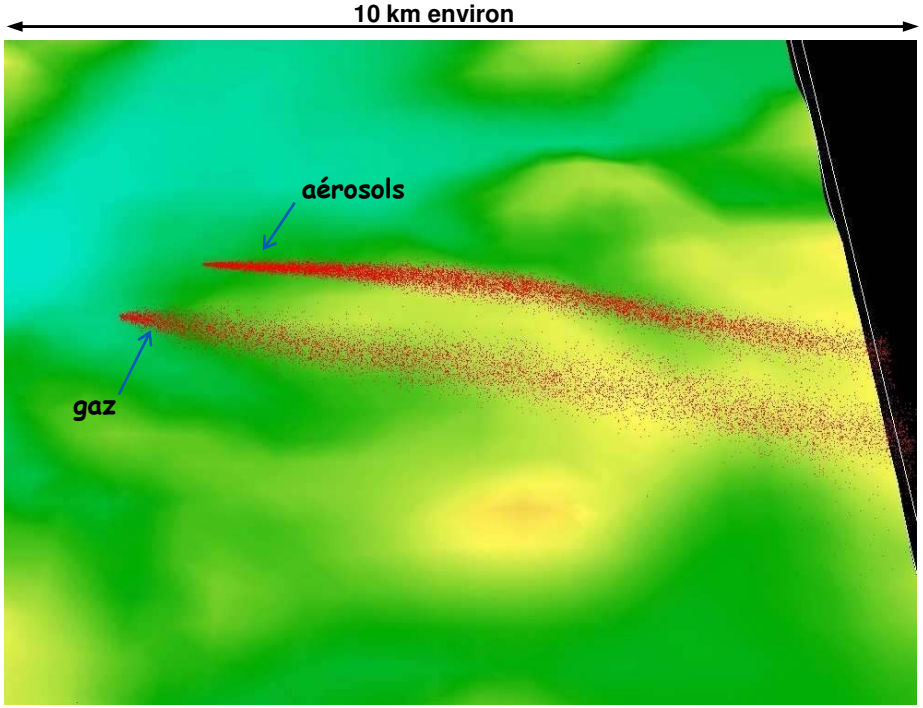


[Source : CEA]

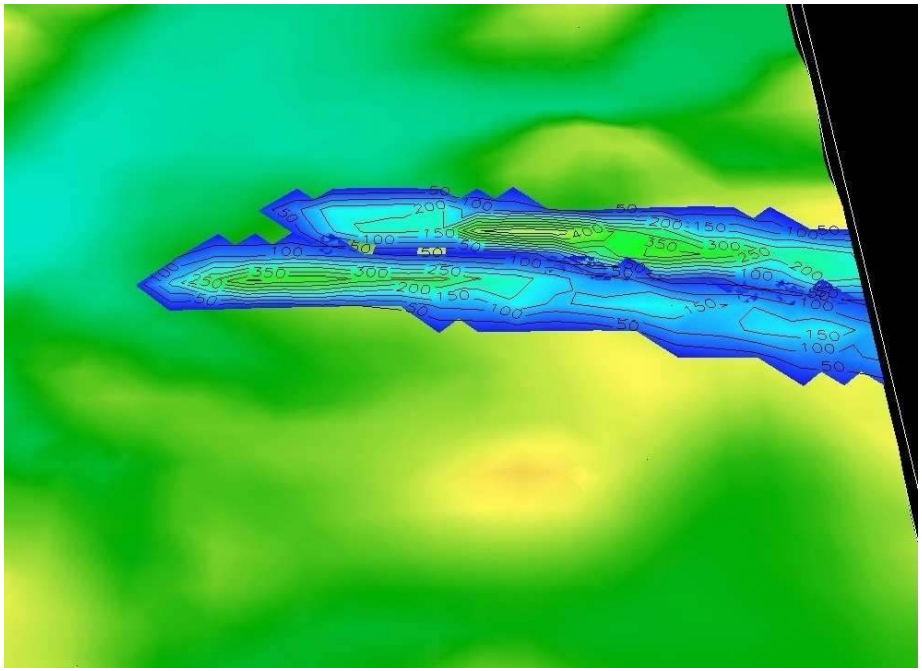
Exemple 2 : dispersion de particules dans l'atmosphère (3/3)

Utilisation d'un **code de calcul** de dispersion particulaire lagrangienne (résout les «équations de la mécanique des fluides)

Visualisation des concentrations en gaz et en aérosols après 5h de rejet



Panache sous forme particulaire



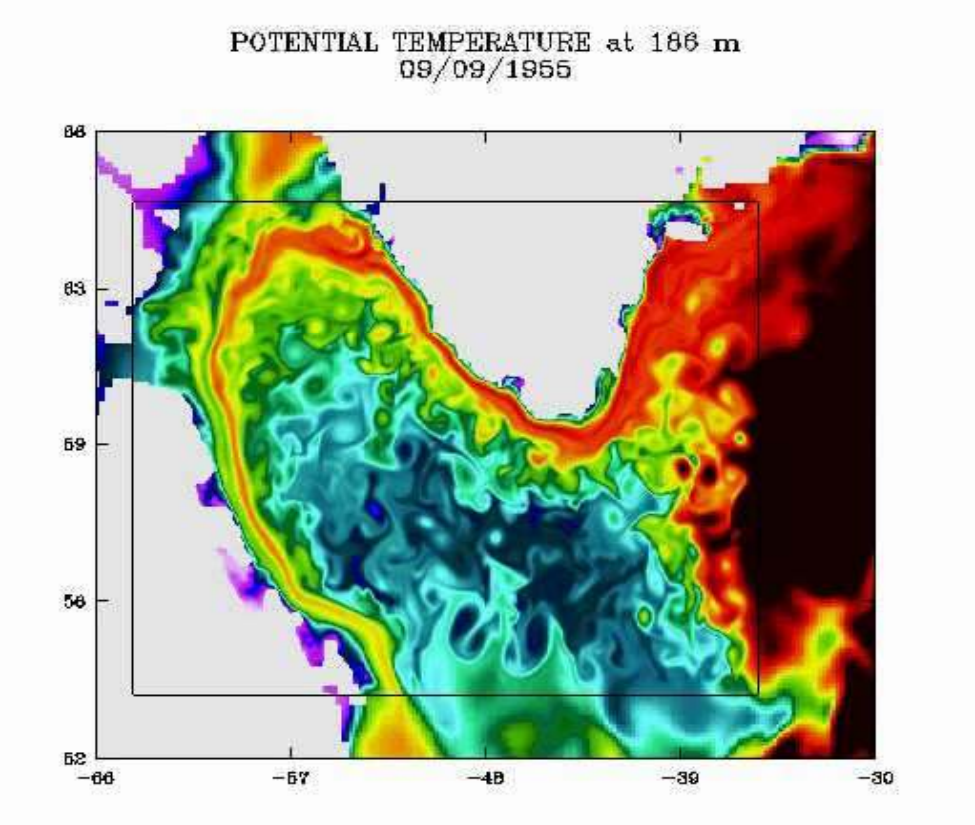
Panache de concentration (à 10 m du sol)

Les résultats sont très sensibles aux données météo

[Source : CEA]

Exemple 3 : simulations de dynamiques spatio-temporelles

Calculs de circulation océanique pour étudier les flux thermiques océaniques
(étude de l'impact du changement climatique)



Mer du Labrador

Modèle NEMO
(Nucleaus for
European Modelling
of the Ocean)

Influence importante des conditions initiales

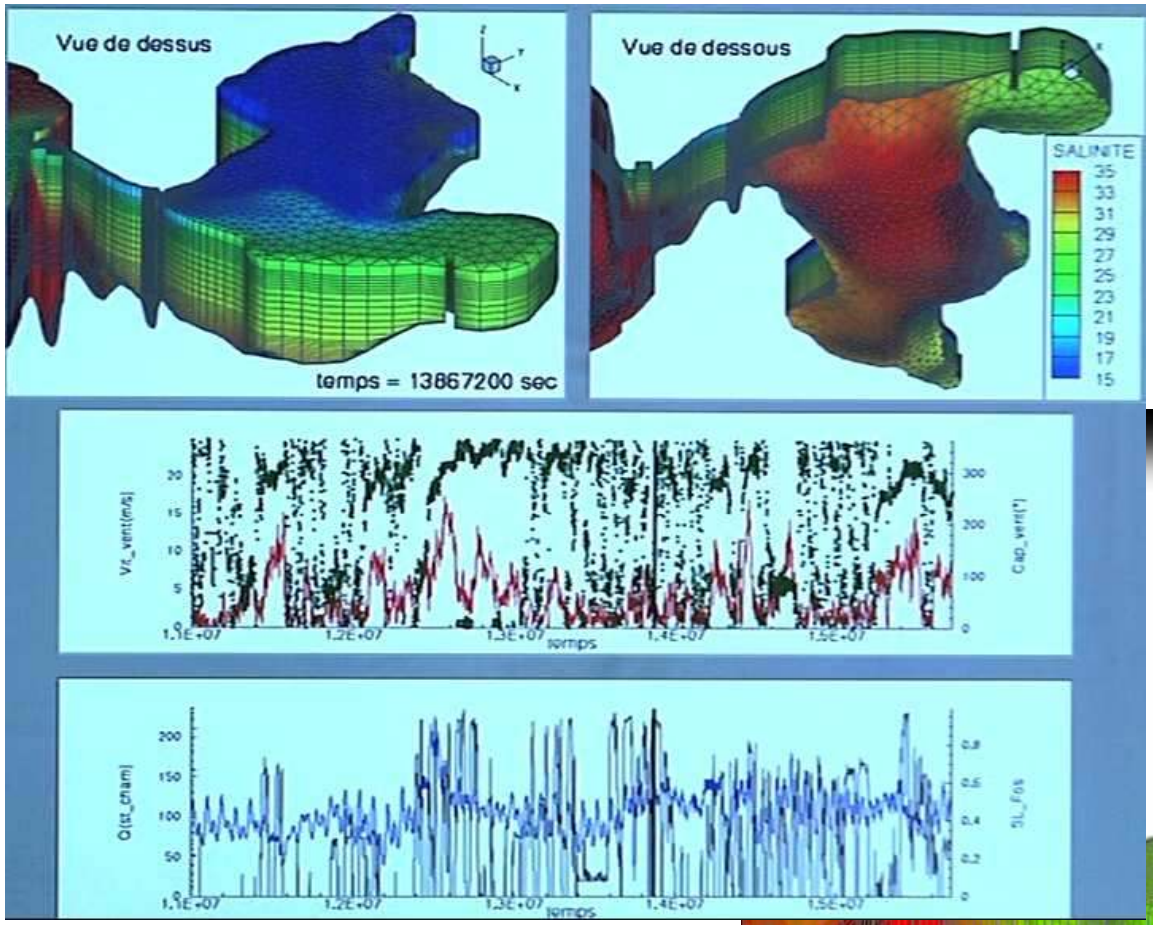
[Source : INRIA]

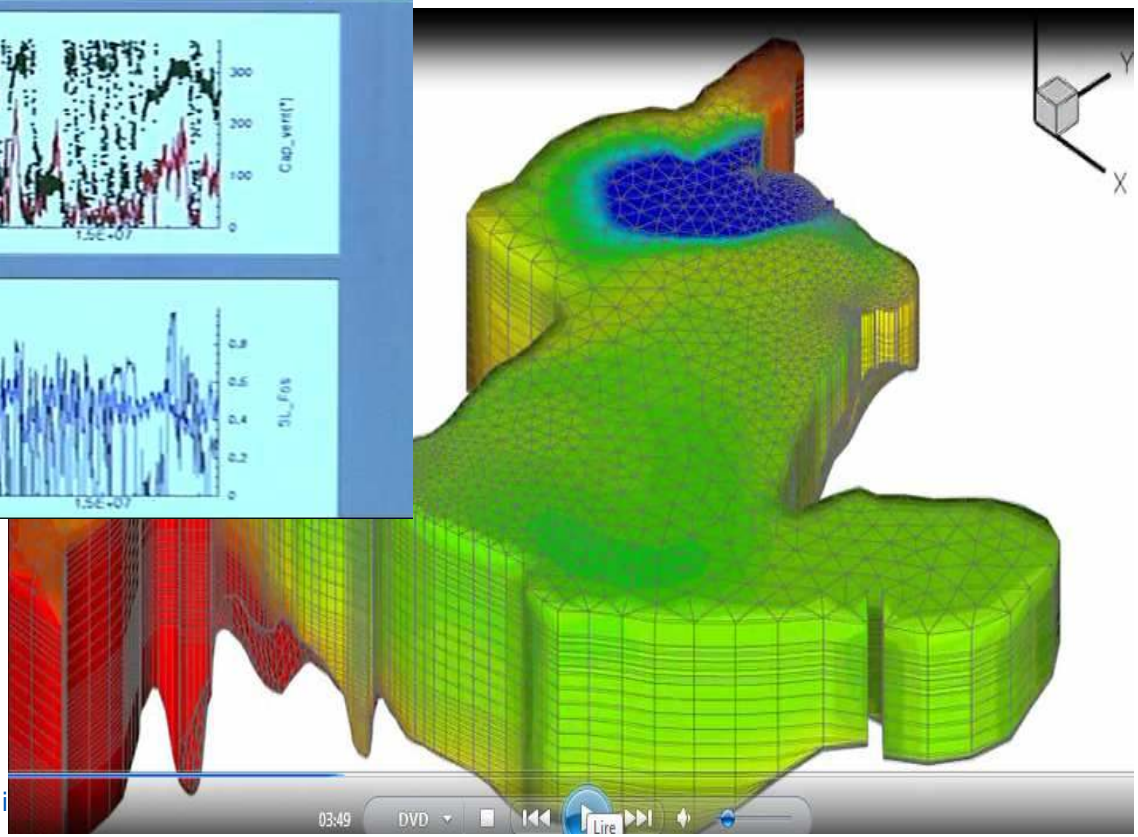
Exemple 4 : simulations hydrodynamiques (1/3)

<u>Objectif :</u> compréhension de l'impact des rejets d'eau douce d'une usine hydroélectrique d'EDF sur la salinité d'une lagune méditerranéenne (étang de Berre)

Exemple 4 : simulations hydrodynamiques (2/3)

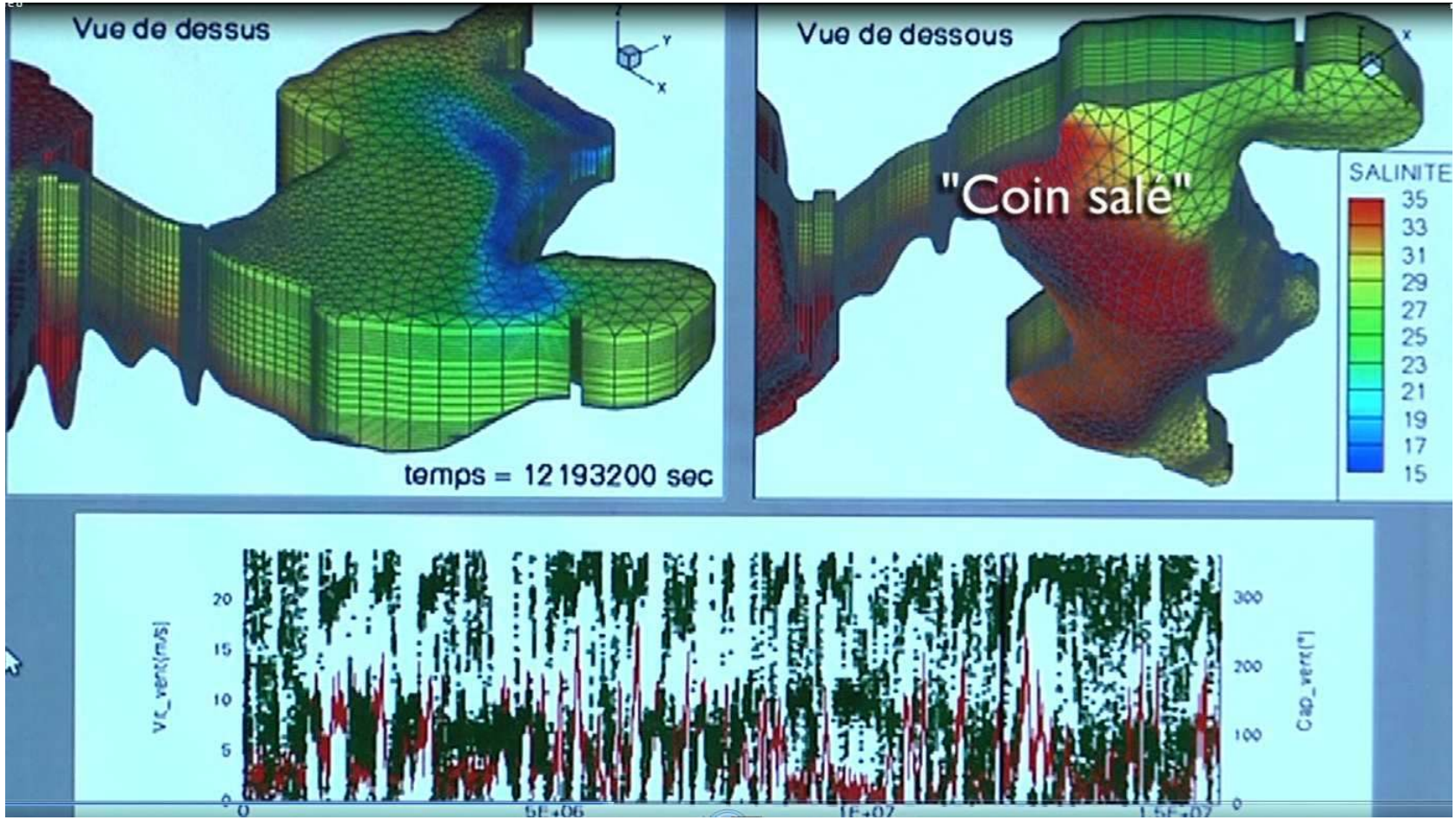
Constitution d'un modèle numérique hydrodynamique





[Source : EDF/LNHE]

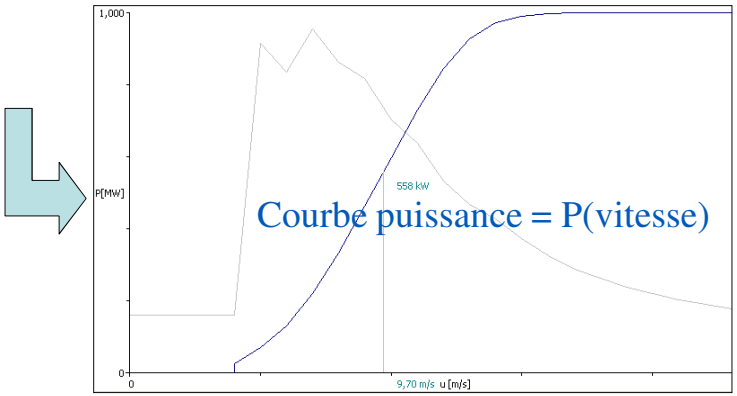
Exemple 4 : simulations hydrodynamiques (3/3)



Sensibilité aux données météo

[Source : EDF/LNHE]

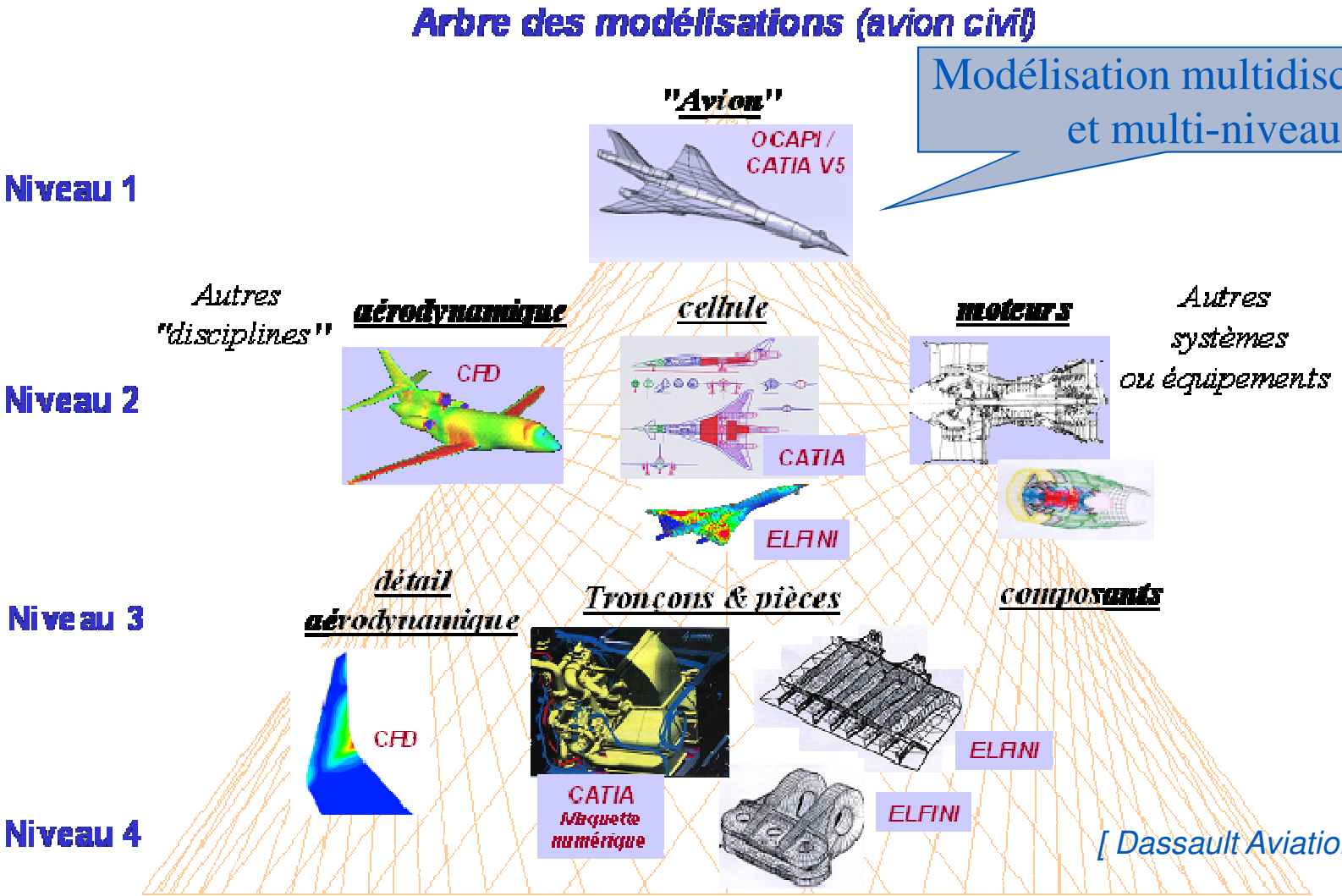
Exemple 5 : production éolienne



Finalité	**Comprendre** le risque de mésestimation du productible – **hiérarchiser** les incertitudes – **optimiser** le lieu des éoliennes
Critère	%inc sur productible
Sources /modèle	Erreurs d'estimation du vent (erreur spatiale, échantill. Temporel …); Inc. de modèle Modèle intrinsèquement statistique du vent / du productible
Quantification / propagation	**Estimation statistique (moments, max. vraisemblance …)** **Propagation par cumul quadratique simple**

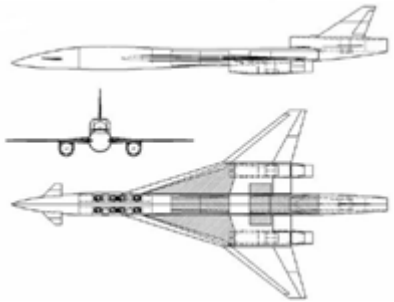
Exemple 6 : dimensionnement avant-projet d'un avion d'affaires supersonique (1/2)

Dimensionner l'avion pour que son rayon d'action soit supérieur à 7400 km

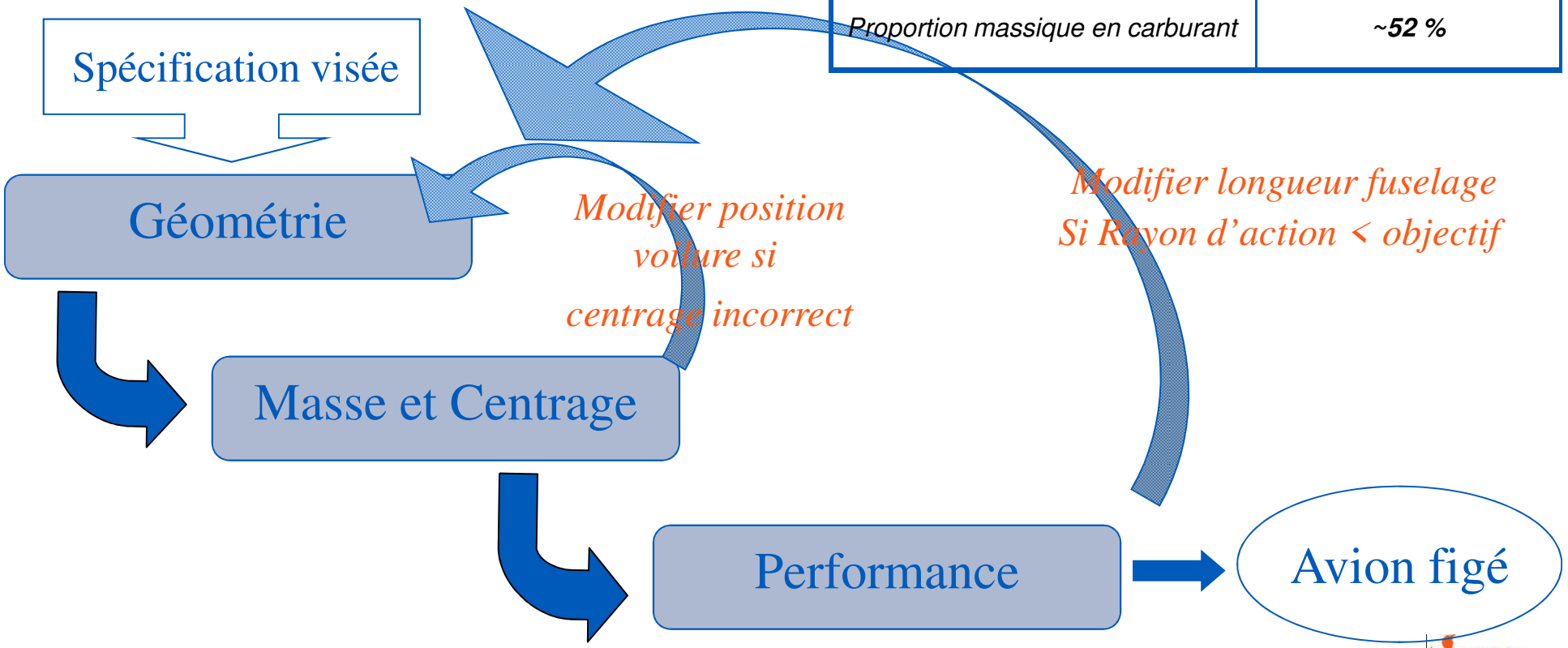


[Dassault Aviation – EADS]

Exemple 6 : dimensionnement avant-projet d'un avion d'affaires supersonique (2/2)

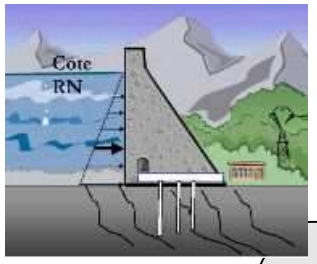


Distance franchissable	***4000 NM / 7400 km***
Mach de croisière	***1.6***
Nb. passagers	***8***
Longueur	***39 m***
Envergure	***18 m***
MTOW	***~50000 kg***
Proportion massique en carburant	***~52 %***



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

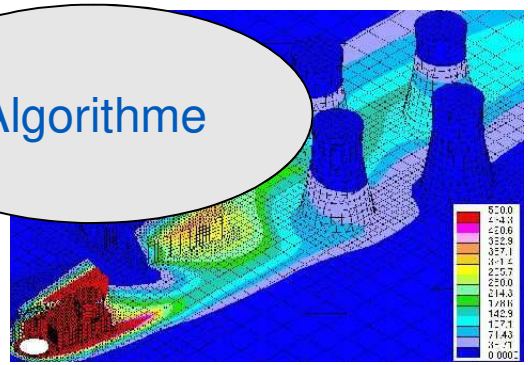
Variable d'intérêt

statisticien

Modèle numérique

informaticien

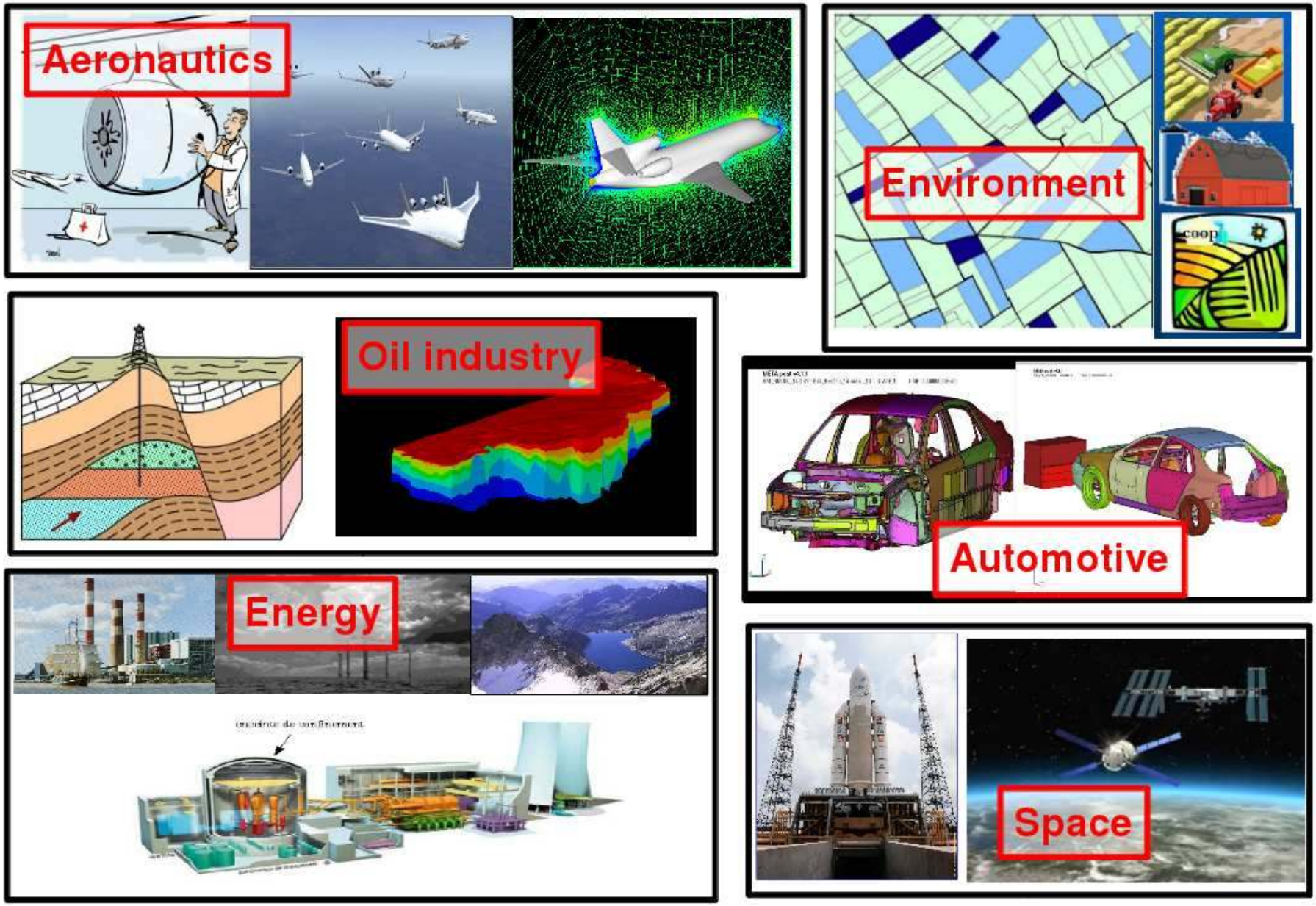
Algorithme



incertitudes épistémiques

Paramètres

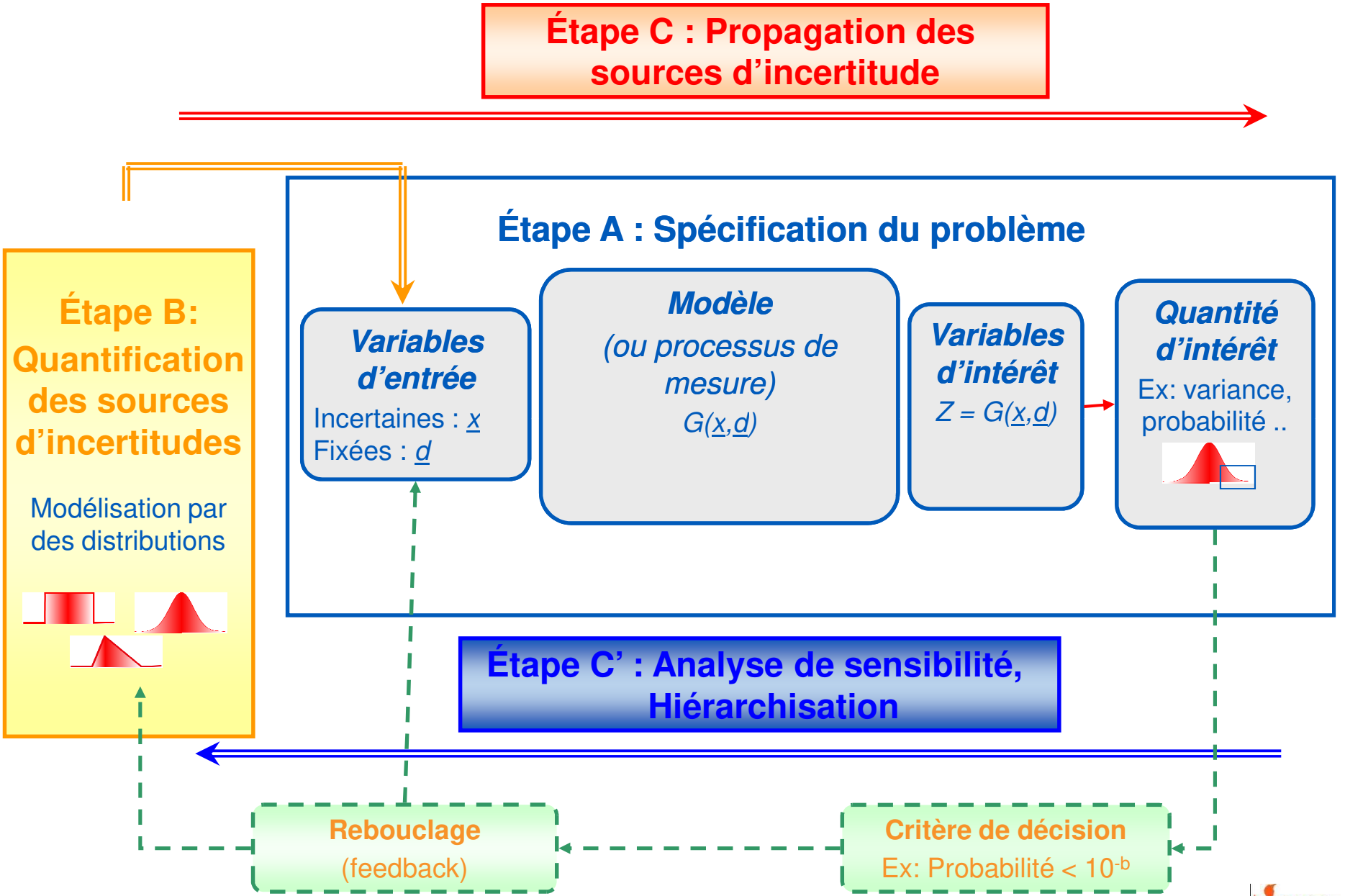
Une problématique multi-sectorielle



Incertitudes en simulation numérique : les enjeux

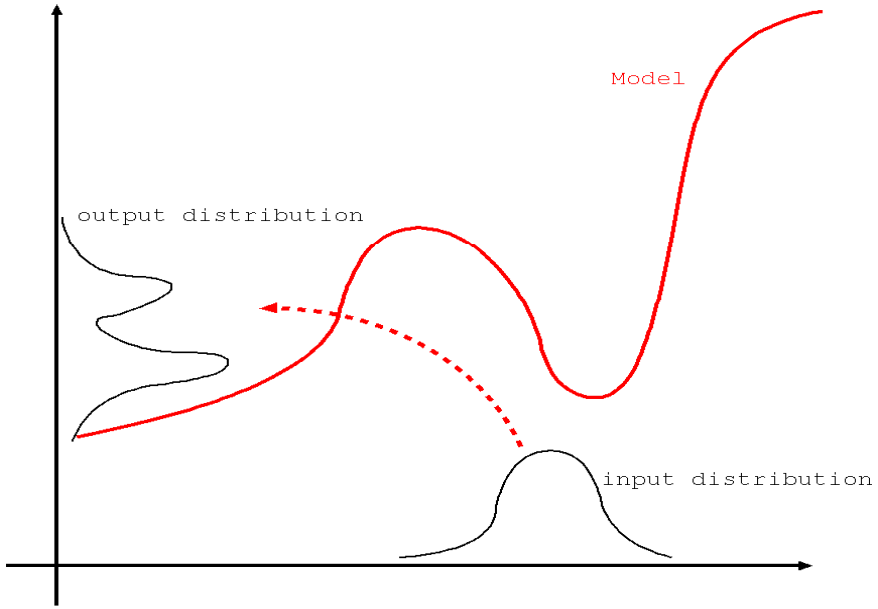
- Modélisation :
  - Explorer au mieux différentes combinaisons des entrées
  - Identifier les données influentes pour prioriser la R&D
  - Améliorer le modèle
- Validation :
  - Réduire l'incertitude de prédiction
  - Calibrer les paramètres du modèle
- Utilisation :
  - Études de sûreté : calculer un risque de défaillance (Fiabilité des structures - événements rares), calculer des **marges** (par rapport à une réglementation)
  - Conception : optimiser les performances d'un système

Approches quantitatives : schéma générique introductif



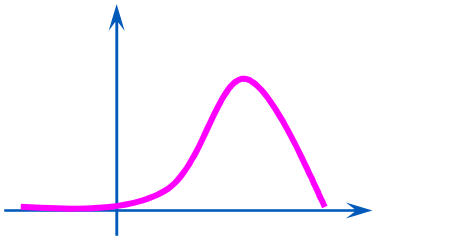
Propagation d'incertitudes

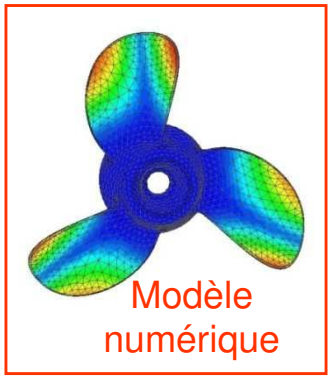
- Transfert des incertitudes de **X** ϵ $\Re^d$ vers Z ϵ $\Re$, via la fonction déterministe G(•)
- **X** (noté aussi $\underline{X}$ ou X) est un vecteur aléatoire, avec une certaine mesure de proba
- Z = G (**X**) devient un vecteur aléatoire, avec une mesure de proba à déterminer

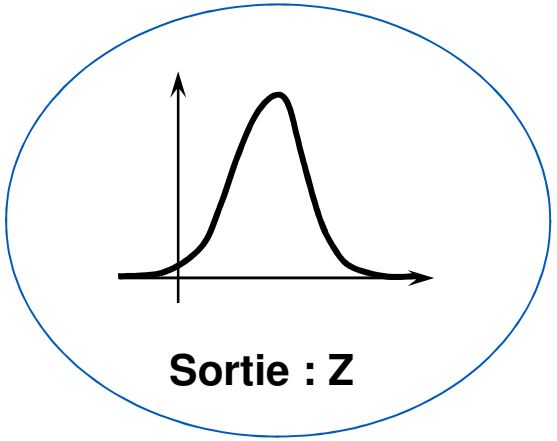


  - Problème conceptuellement « simple » mais (parfois) de mise en œuvre complexe
  - Le choix de la méthode dépend très fortement de la « quantité d'intérêt » de l'étude
  - … d'où l'importance de l'étape A, peu mathématique mais essentielle en pratique

Etape A : la quantité d'intérêt (1/2)

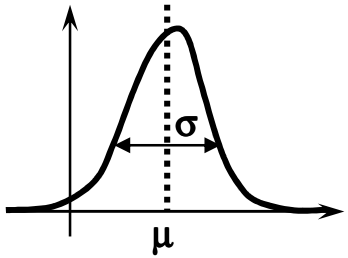




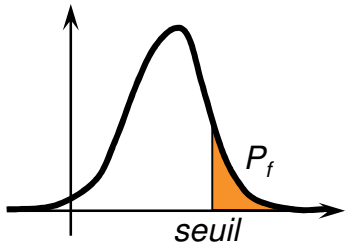


Entrées : X

Qu'est ce qui est (vraiment) intéressant pour notre étude ?



Moyenne, médiane, variance, (moments) de Z



Quantiles (extrêmes), probabilité de dépassement d'un seuil fixé, distribution complète

Etape A : la quantité d'intérêt (2/2)

- La quantité d'intérêt est liée à des enjeux décisionnels
  - Du point de vue de la propagation, on distingue deux types de problèmes :

- Tendance centrale (ex. moyenne) ou dispersion (variance)
  - Exemple : métrologie

→ Méthodes analytiques (parfois) envisageables

- Quantile extrême, « probabilité de défaillance »
  - Point de vue « exploitant » → justification d'un critère de sûreté

→ Méthodes numériques (optimisation, échantillonnage Monte Carlo)

Le système est dans un bon état de fonctionnement si la valeur de Z (par ex. température, hauteur d'eau) est en dessous (ou en dessus) d'un seuil de sécurité

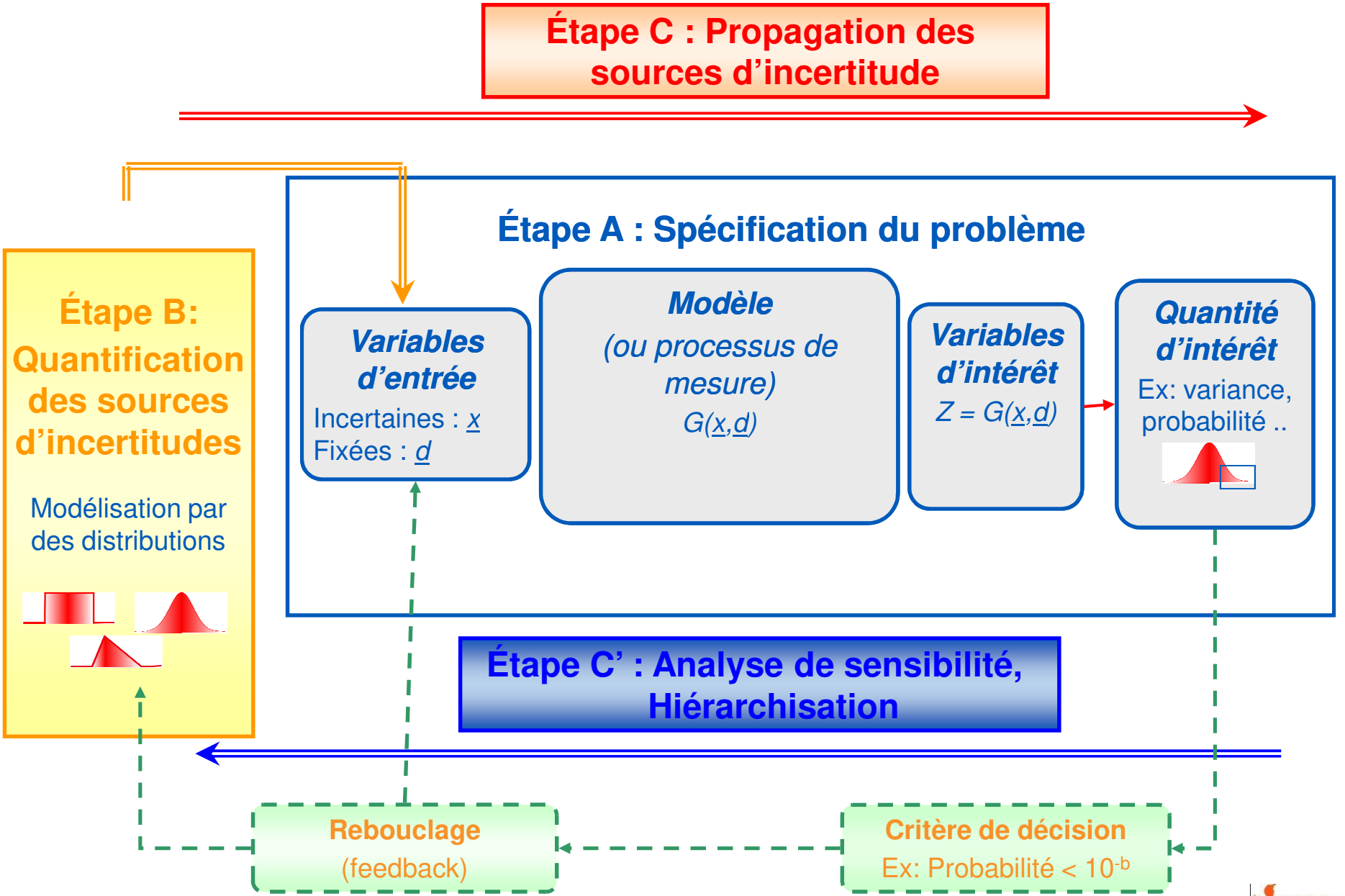
L'évènement « défaillance » est associé au dépassement de ce seuil

Probabilité de dépassement = Probabilité de défaillance $P_f = \mathrm{P}(Z \geq z^*)$

Plan du cours 1

1. Introduction
2. **Modélisation des sources d'incertitudes**
3. Propagation des incertitudes

Schéma générique



Étape B - Introduction

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

  X est une variable aléatoire multi-dimensionnelle, décrite par une loi jointe

- Situations dans la pratique industrielle :
  - Données disponibles → Ajustement de lois paramétriques ou non-paramétriques
  - Données « indirectement » observées → modèles inverses
  - Absence de données → Expertise pour la construction de la loi de **X**

- Absence de données => méthodes d'élicitation
  - Traduction formelle de l'avis d'expert en une distribution de probabilité
  - Une manière de construire des lois de probabilité à partir d'informations « minimalistes » : la méthode du maximum d'entropie
  - Méthodes plus rigoureuses basées sur la théorie bayésienne

2.1 Illustration de l'élicitation d'intervalles de confiance - Questionnaire

2.2 Entropie statistique (1/2)

- Définition donnée par Shannon (1948), puis formalisée par Jaynes (1957)

- Cas discret : X est une v.a. discrète de loi $P_X = \{p_1, p_2, \ldots, p_k\}$

$$H(X) = -\sum_{i=1}^{k} p_i \log(p_i)$$

- Propriétés :
  - $H(X) \geq 0$
    $H(X) = 0 \Leftrightarrow \exists! p_i : p_i = 1, \quad \forall i \neq j \quad p_j = 0$

    ← Toujours positive, sauf dans un cas particulier (minimum = 0)
  - $H(X) \leq \log(k)$

    ← Maximum de H
  - Le maximum, égal à log(k), est atteint dans le cas d'une distribution uniforme :

$$p_i = \frac{1}{k} \forall i \Rightarrow H(X) = -\sum_{i=1}^{k} \frac{1}{k} \log\left(\frac{1}{k}\right) = -\frac{1}{k}\, k\, log(\frac{1}{k}) = \log(k)$$

Entropie statistique (2/2)

- Interprétation (intuitive) de l'entropie
  - Minimale en cas d'information « parfaite » → aucun doute sur la valeur de X parmi les k valeurs possibles
  - Maximale dans le cas où l'information apportée par la loi de prob. est la plus vague possible → chaque valeur possible de X est équiprobable
  - Entropie : mesure (inverse) de l'information apportée sur X par sa loi de probabilité

- Extension à une v.a. continue : $H(X) = -\int_{\mathcal{X}} f(x) \log\left(f(x)\right) dx$

- Principe du maximum d'entropie
  - Parmi toutes les lois possibles, choisir celle qui apporte le minimum d'information → celle qui maximise l'entropie
  - Justification : Recherche « d'objectivité » :
    - Ne rajouter aucune information, mise à part celle fournie par l'expert

Application du Maximum d'entropie

- Application triviale : un expert nous dit que X est une v.a. discrète pouvant prendre k valeurs → choix de la loi discrète uniforme : $p_i = 1/k$

- Plus généralement, supposons qu'un expert nous donne N informations sur la loi de X sous la forme :

$$\int_{\mathcal{X}} g_j(x) f(x) dx = c_j$$

  - le problème du maximum d'entropie consiste en rechercher une fonction f(x) qui maximise H(X) et qui respecte les N +1 conditions :

$$\begin{cases} \int_{\mathcal{X}} f(x) dx = 1 \\ \int_{\mathcal{X}} g_j(x) f(x) dx = c_j \quad j = 1...N \end{cases}$$

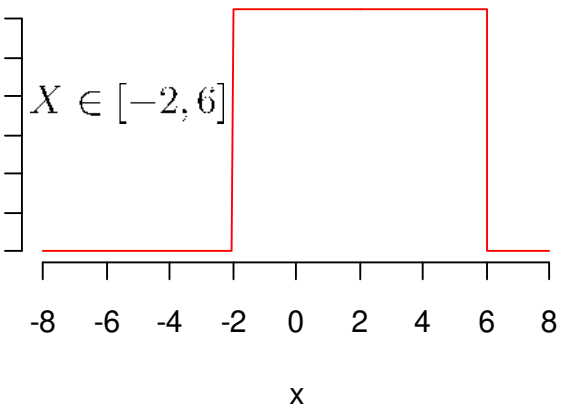
Optimisation sous contraintes

  - Justification : parmi toutes les densités compatibles avec l'information disponible, on choisit celle qui apporte le minimum d'information sur X

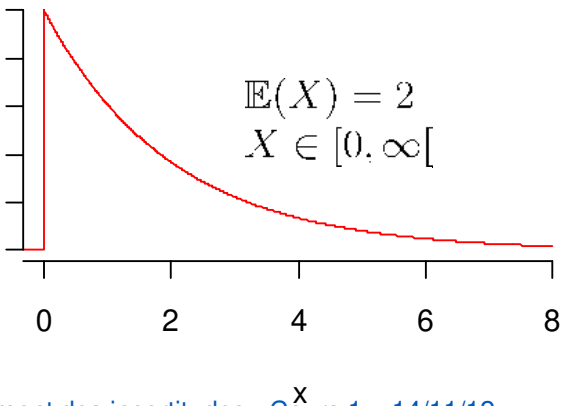
Application du Max d'entropie - exemples (1/2)

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 $X \sim \mathcal{N}(\mu, \sigma)$

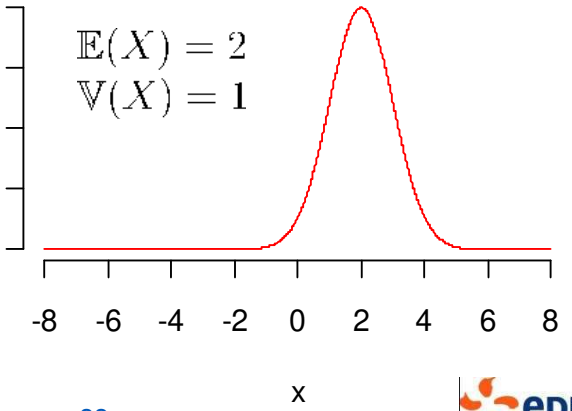
Loi uniforme



Loi exponentielle

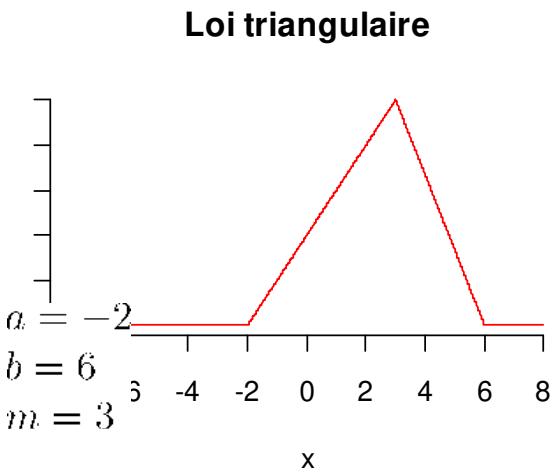


Loi normale



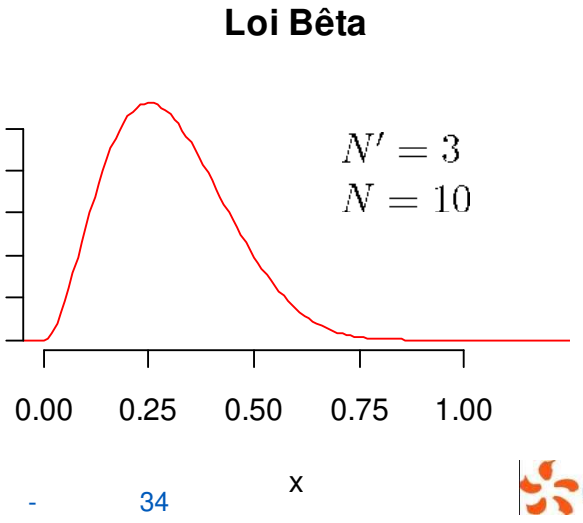
Application du Max d'entropie - exemples (2/2)

- 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 »

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



2.3 Quantification de l'incertitude en présence de données

- Problème :
  - A partir d'un échantillon i.i.d. de la v.a. X : $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$
  - 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

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 : $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$

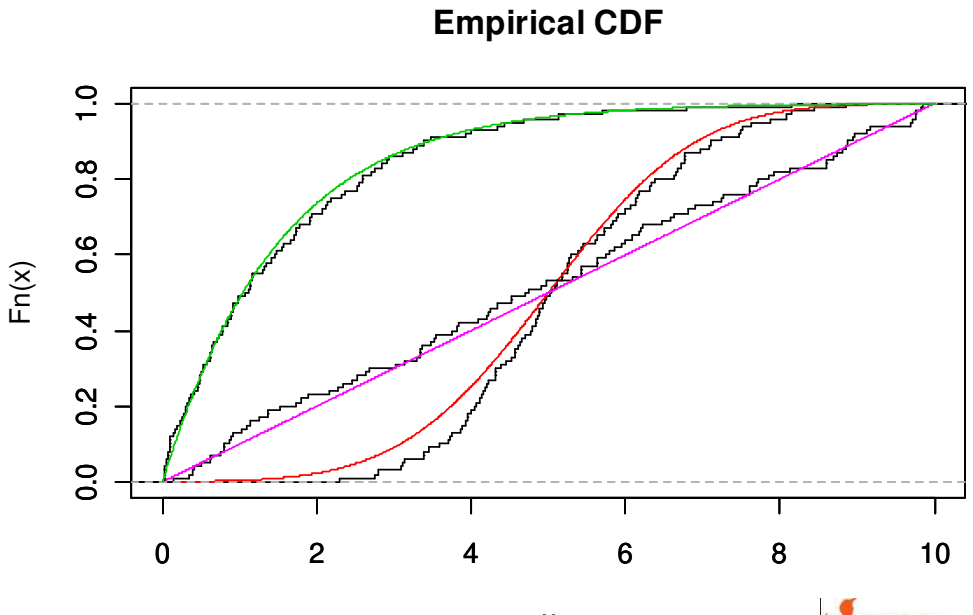
- Fonction de répartition empirique :
  - Proportion des observations ≤ à une valeur fixée x de la v.a. X

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

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

- « 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

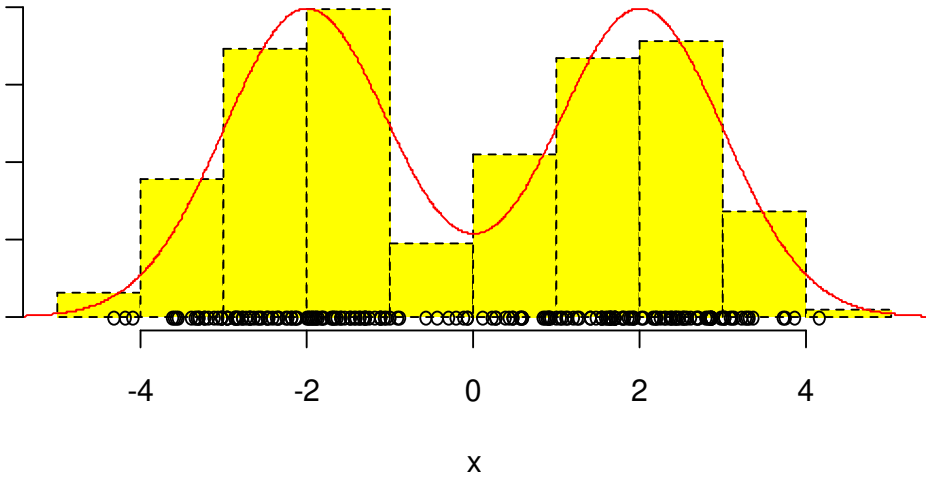
$$]x^* + jh, x^* + (j+1)h] \quad j \in \mathbb{N}$$

- 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

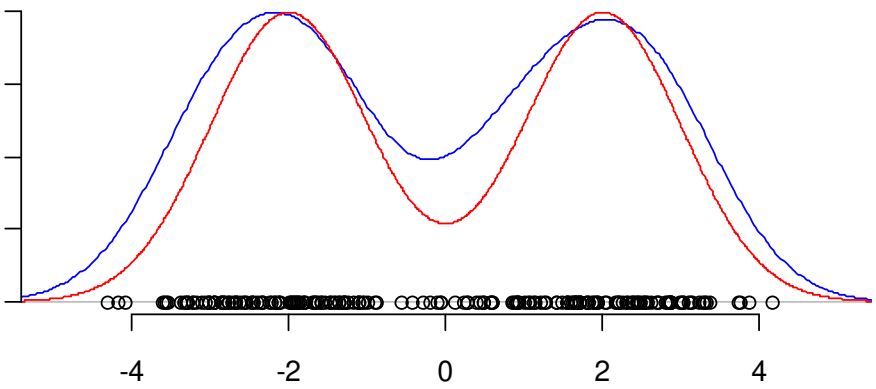


Estimation par noyaux

- 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 : $\int_X K(x)dx = 1$
  - 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 …



Ajustement paramétrique à une densité de probabilité

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

Quelques considérations sur le support des lois

Positive Continue $(0,+\infty)$	Illimité Continue $(-\infty,+\infty)$	Limité Continue (a,b)
Exponentielle Gamma/Erlang Lognormale Weibull Chi-deux F (Fisher-Snedecor) Log-Laplace Log-logistique Pareto …	Normale Cauchy Loi des Extrêmes A,B Laplace Logistique Student …	Bêta Triangulaire Uniforme …

Principales lois discrètes utilisées

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

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

Exemple : lancement d'un dé

➢ **Loi de Bernouilli B(p)** :

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

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

➢ **Loi binomiale B(n,p)** : *n* répétitions indépendantes d'une Bernouilli

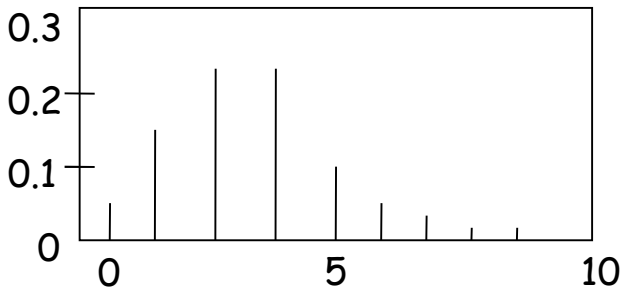
$$X = \sum_{i=1}^{n} X_i \quad \Rightarrow \quad \mathrm{P}(X=k) = C_n^k p^k (n-p)^{n-k}$$

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

p faible, n grand

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

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

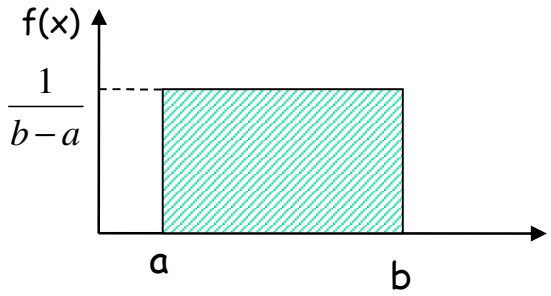


Ex emples :
nombre de personnes dans une file d'attente,
nombre d'appels à un standard

Principales lois continues utilisées (1/3)

➢ Loi uniforme U[a,b]

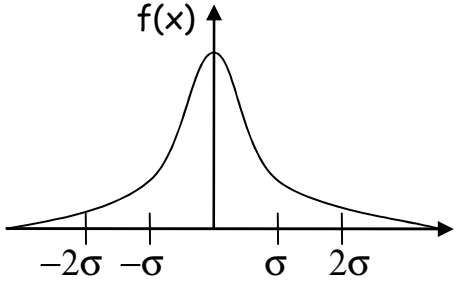
$$f(x) = \frac{1}{b-a} \text{ si } a \le x \le b \; ; \; f(x) = 0 \text{ ailleurs}$$

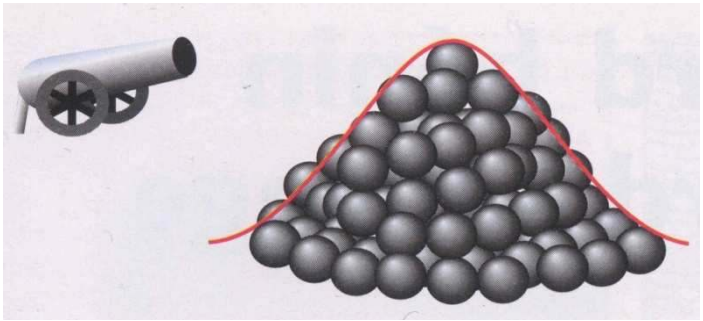


➢ Loi normale N(μ,σ²)

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

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





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

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

Principales lois continues utilisées (2/3)

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

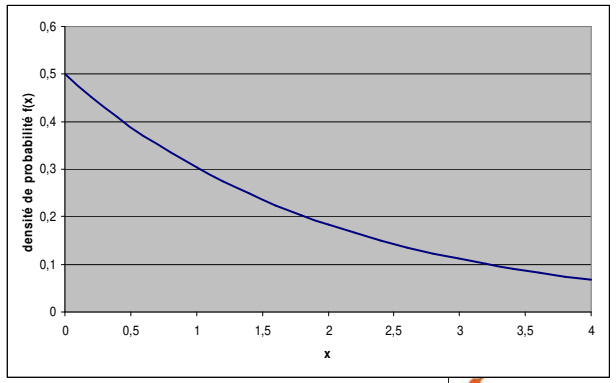
➢ **Loi lognormale LN(μ,σ²)** : $\ln(X) \sim \mathrm{N}\ (\mu, \sigma^2)$

Le produit de v.a. $\xrightarrow{\ \mathrm{L}\ }$ 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 E(λ)** : $\mathrm{f(x)} = \lambda \exp(-\lambda x)$ si $x \geq 0$;

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



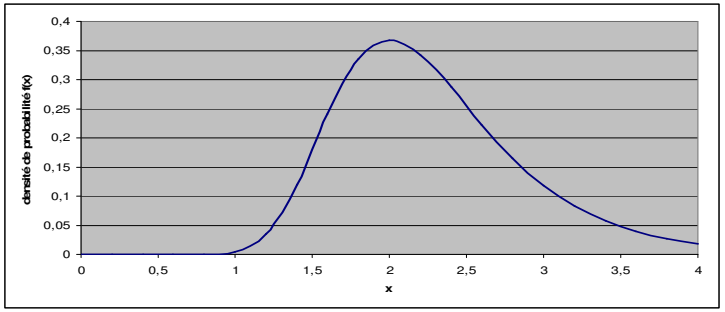
Exemples : temps d'attente, durée de vie de systèmes sans usure ⇨ i.e. la proportion de matériels défaillants est chaque année la même.

Principales lois continues utilisées (3/3)

➢ **Loi de Gumbel 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*

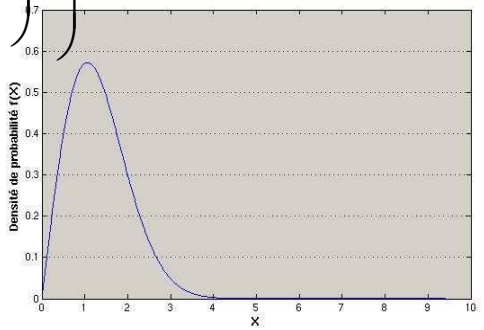
- les fortes valeurs restent probables



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

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

Généralisation de la loi exponentielle

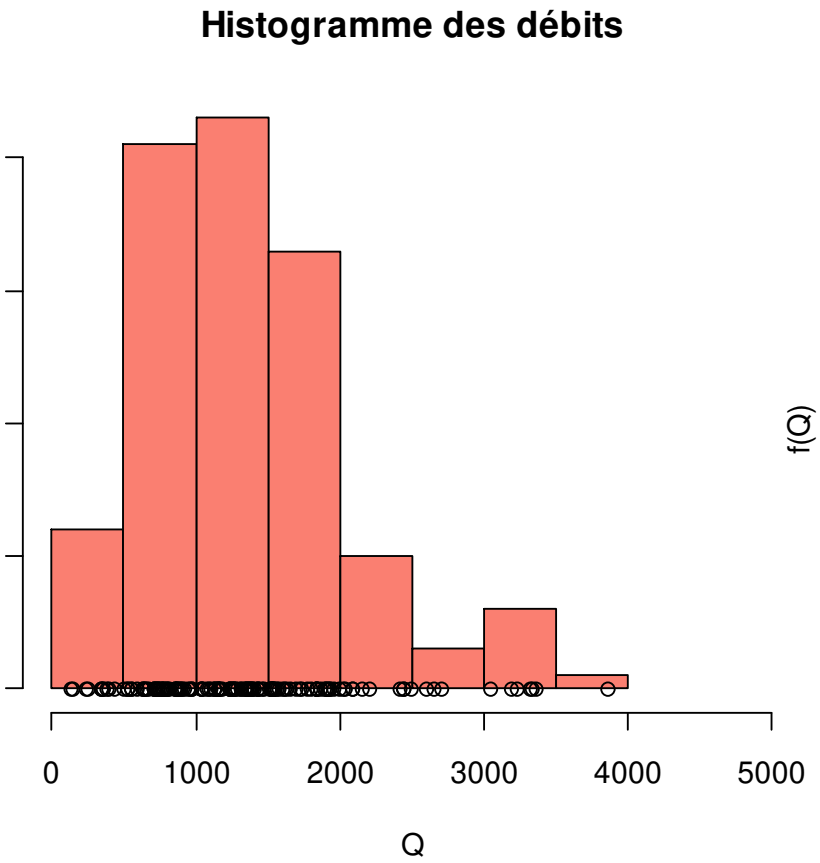


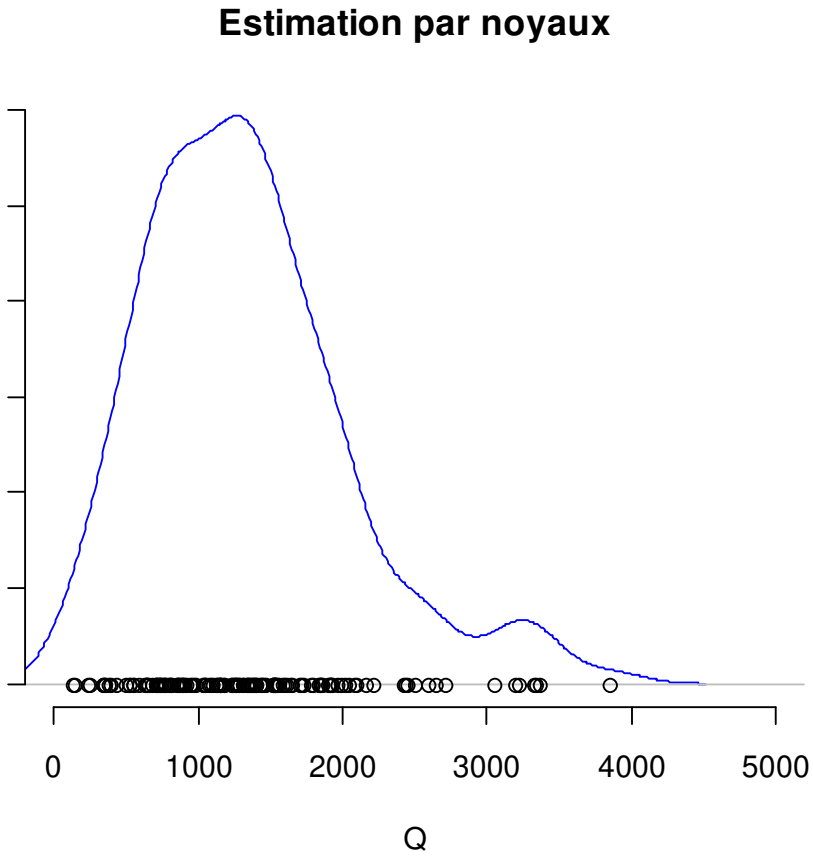
Exemples en mécanique : durée de vie d'un matériel qui :
- *se dégrade pour α>1 (structure acier)*
- *ou se bonifie pour α<1 (résistance du béton en début de vie)*

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)



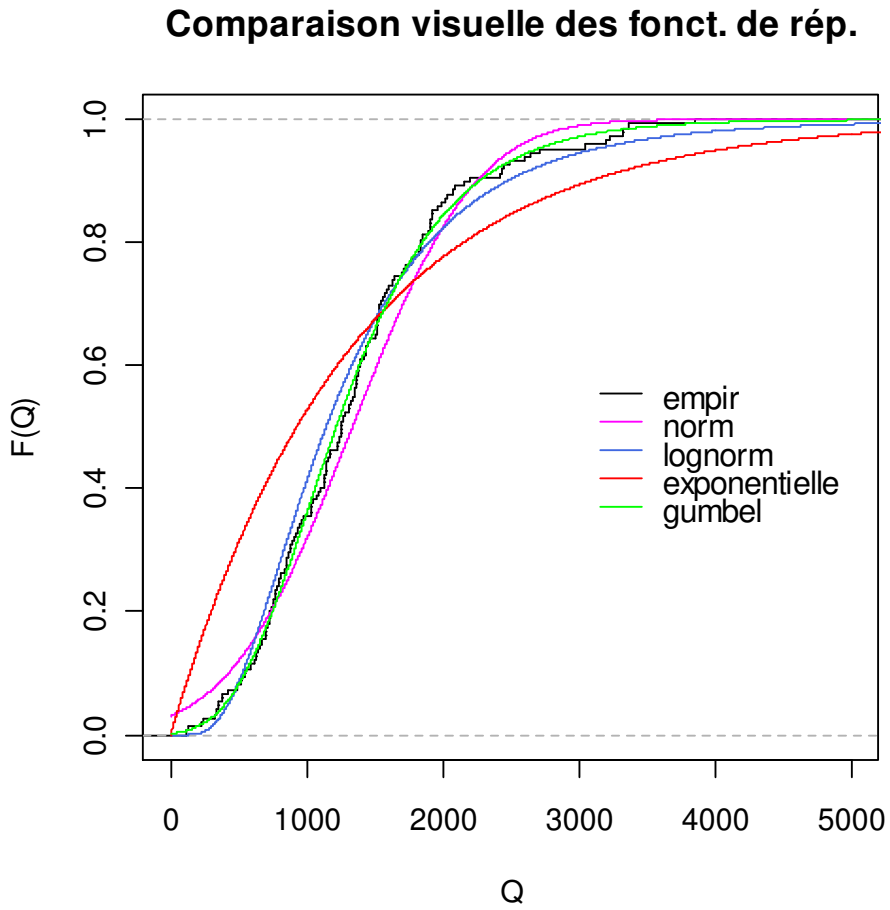


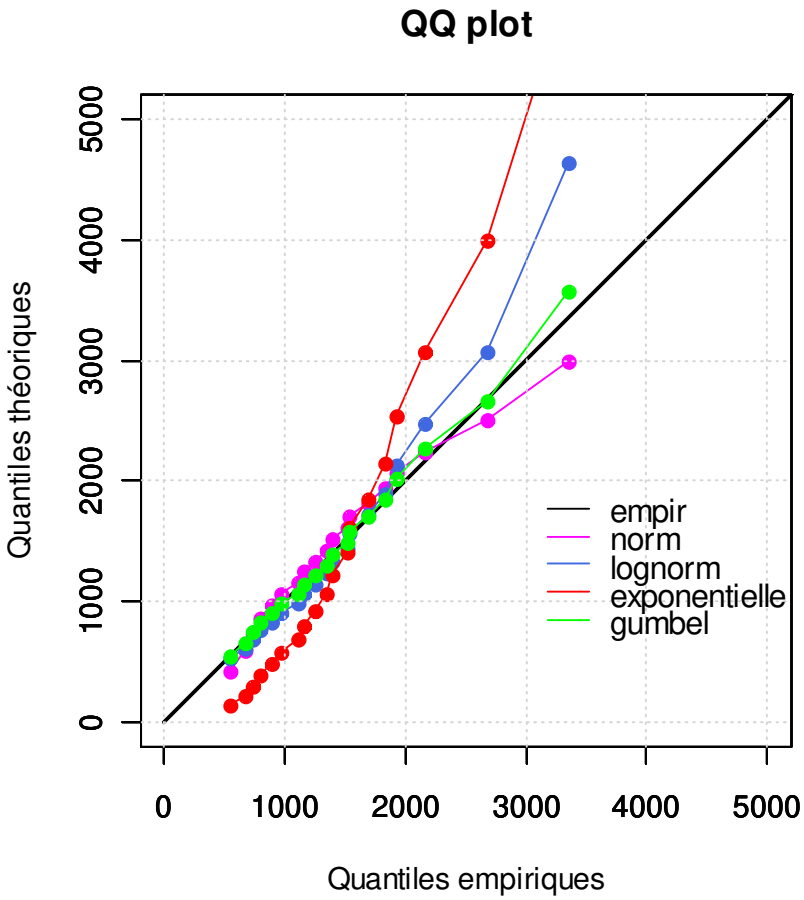
Exemple d'ajustement (2/2) - Estimation par maximum de vraisemblance

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 basent 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

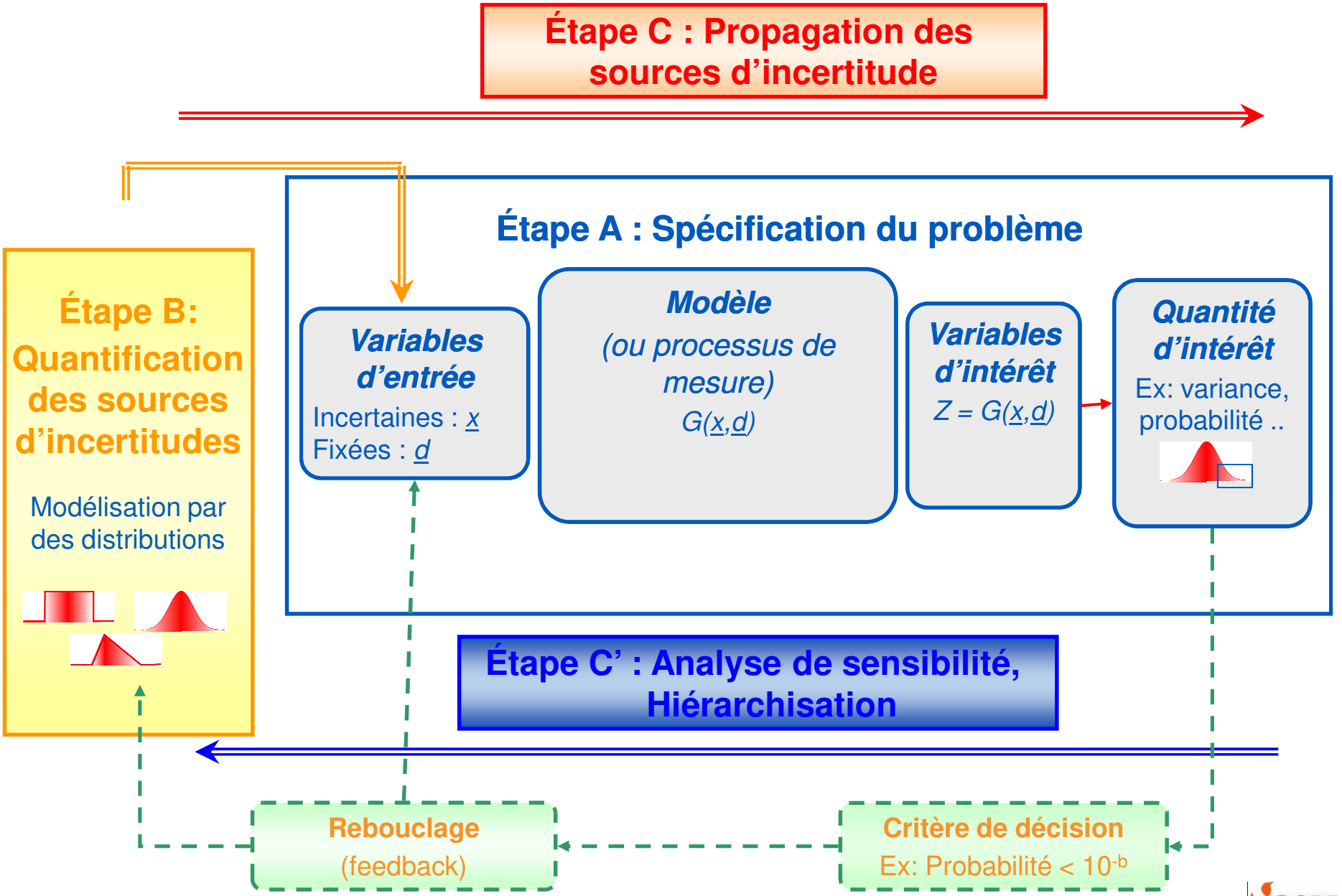
- Préférence pour la loi de Gumbel
- Attention au faible pouvoir discriminant des tests d'adéquation
- Autres critères de sélection (basés sur le rapport des vraisemblances) :

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

Plan du cours 1

1. Introduction
2. Modélisation des sources d'incertitudes
3. **Propagation des incertitudes**

Schéma générique



3.1 Cumul quadratique - Introduction

- « Cumul quadratique » : terme couramment employé par les praticiens pour désigner une méthode analytique, particulièrement simple
  - Fondement : deux résultats élémentaires de calcul des probabilités
    - $X_1, \ldots, X_N$: variables aléatoires
    - $a_1, \ldots, a_N$: réels

$$\mathbb{E}\left[\sum_{i=1}^{N} a_i X_i\right] = \sum_{i=1}^{N} a_i \mathbb{E}\left[X_i\right]$$

$$\mathbb{V}\left[\sum_{i=1}^{N} a_i X_i\right] = \sum_{i=1}^{N} a_i^2 \mathbb{V}\left[X_i\right] + 2 \sum_{1 \le i < j \le N} a_i a_j Cov\left[X_i, X_j\right]$$

    - Ces formules donnent la moyenne et la variance de Z=G(X) si G est un modèle linéaire
    - … d'où l'idée de « linéariser » localement le modèle G par un développement de Taylor

Cumul quadratique – Mise en œuvre

Données : les valeurs moyennes des X_i : $\mu_i = \mathbb{E}[X_i]$

la matrice de covariance ou la matrice de corrélation des X_i :

$$Cov[X_i, X_j] = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

$$\rho_{ij} = \mathbb{E}\left[\frac{X_i - \mu_i}{\sigma_i}\frac{X_i - \mu_i}{\sigma_j}\right]$$

Développement de Taylor de G(•) au voisinage de E(X) :

$$G(X) = G(\mu) + \sum_{i=1}^{N} \left.\frac{\partial G}{\partial X_i}\right|_{X=\mu} (X_i - \mu_i)$$
$$+ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \left.\frac{\partial^2 G}{\partial X_i \partial X_j}\right|_{X=\mu} (X_i - \mu_i)(X_j - \mu_j) + o\left(\|X - \mu\|^2\right)$$

En général, dans les applications le développement est d'ordre 1

Cumul quadratique – Développement d'ordre 1

Calcul de la moyenne de Z

$$\mathbb{E}[Z] = G(\mu)$$

La moyenne de la réponse est égale, au premier ordre,
à la réponse calculée aux valeurs moyennes des entrées

Calcul de la variance de Z

$$\mathbb{V}[Z] = \mathbb{E}\left[(Z - \mathbb{E}[Z])^2\right] = \mathbb{E}\left[\left(G(\mu) + \sum_{i=1}^{N} \frac{\partial G}{\partial X_i}\Big|_{X=\mu} (X_i - \mu_i) - G(\mu)\right)^2\right] =$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial G}{\partial X_i}\Big|_{X=\mu} \frac{\partial G}{\partial X_j}\Big|_{X=\mu} \mathbb{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

$$\mathbb{V}[Z] = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial G}{\partial X_i}\Big|_{X=\mu} \frac{\partial G}{\partial X_j}\Big|_{X=\mu} \rho_{ij}\sigma_i\sigma_j$$

<u>Remarques :</u>

++ Ne nécessite que moyenne et covariance de X

-- A ne pas utiliser pour les modèles G(.) fortement non linéaires

-- Ne restitue que moyenne et variance de Z => pas d'extrapolations sur la loi de Z

++ si X est gaussien et G(.) est linéaire, alors Z est gaussien

Cumul quadratique – Variables indépendantes

Calcul de la variance si les X_i sont indépendantes :

$$\mathbb{V}[Z] = \sum_{i=1}^{N} \left(\frac{\partial G}{\partial X_i}\Big|_{X=\mu} \right)^2 \sigma_i^2$$

Formule du « cumul quadratique »

Contribution de chaque variable d'entrée à l'incertitude sur la variable de sortie

- Termes « déterministes » → composants du gradient de G(•)
- Termes liés à l'incertitude de la variable X_i (variance)

$$\eta_i^2 = \frac{1}{\mathbb{V}[Z]} \left(\frac{\partial G}{\partial X_i}\Big|_{X=\mu} \right)^2 \sigma_i^2$$

Indices de sensibilité (normés)

L'analyse de sensibilité est réalisée de manière directe

3.2 Méthodes de simulation Monte Carlo

Méthodes Monte Carlo

- Méthodes générales pour l'évaluation d'une grandeur numérique, utilisant la simulation aléatoire

- Idée de base en propagation d'incertitudes : évaluer la quantité d'intérêt, sur la base d'un échantillon aléatoire de G(X)

Monte Carlo – fondements (1/4)

- Calcul de l'intégrale :

$$I = \int_{\mathcal{X}} h(x) f(x) dx$$

h(•) : fonction déterministe
X : v.a. de densité f(x)

$$\int_{\mathcal{X}} h(x) f(x) dx = \mathbb{E}[h(X)]$$

Formellement, c'est l'espérance de h(X).

$$x^{(1)}, x^{(2)}, \ldots, x^{(n)}$$

Échantillon aléatoire de X

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} h(x^{(i)}) \rightarrow \mathbb{E}[h(x)]$$

$$\hat{I} \rightarrow I$$

D'après la loi des grands nombres, l'estimateur Monte Carlo converge (p.s.) vers la grandeur recherchée

Estimateur Monte Carlo

Monte Carlo – fondements (2/4)

- Variance de l'estimateur Monte Carlo

$$\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} h(X^{(i)})\right] = \frac{1}{n^2}\, n\mathbb{V}\left[h(X)\right] = \frac{1}{n}\mathbb{V}\left[h(X)\right]$$

Variance d'une somme de n v.a. i.i.d.

  - La variance de h(X) est estimée par son estimateur :

$$\mathbb{V}\left[h(X)\right] \approx \frac{1}{n}\sum_{i=1}^{n}\left(h(x^{(i)}) - \hat{I}\right)^2$$

  - D'où l'expression générale pour la variance de l'estimateur MC

$$\mathbb{V}\left[\hat{I}\right] \approx \frac{1}{n^2}\sum_{i=1}^{n}\left(h(x^{(i)}) - \hat{I}\right)^2$$

  - Notons : $\sigma^2_{\hat{I}} = \mathbb{V}\left[\hat{I}\right]$

Monte Carlo – fondements (3/4)

- Loi asymptotique de l'estimateur
  - D'après le Théorème Central Limite :

$$\frac{\sqrt{n}}{\sigma_{h(X)}}\left(\hat{I}-I\right) \sim \mathcal{N}(0,1)$$

Convergence asymptotique de la loi de l'estimateur vers une loi normale

  - avec $\sigma_{h(X)} = \sqrt{\mathbb{V}[h(X)]}$

  - D'où les intervalles de confiance pour l'erreur Monte Carlo :

$$\epsilon_n = \hat{I} - I$$

Erreur Monte Carlo

$$\epsilon_n \in \left[-q_{(1-\alpha/2)} \cdot \sigma_{\hat{I}},\ q_{(1-\alpha/2)} \cdot \sigma_{\hat{I}}\right]$$

Intervalle de confiance de probabilité α

Quantiles de la loi norm. standard

$$\sigma_{\hat{I}} = \frac{\sigma_{h(X)}}{\sqrt{n}}$$

Monte Carlo – fondements (4/4)

- La vitesse de convergence est de l'ordre de $1/\sqrt{n}$
  - par ex. multiplier par 100 la nombre n de tirages permet de diviser par 10 l'écart type de l'erreur
  - convergence relativement lente mais
    - Indépendante de la dimension de X
    - Indépendante de la forme de la fonction h(•), sous des conditions de régularité assez larges
    - Estimateur non biaisé
    - La précision dépend uniquement de n (et donc du temps de calcul)

- La vitesse de convergence peut être pénalisante dans certain cas (notamment pour estimer des quantiles ou des probabilités de défaillance)

Monte Carlo et propagation d'incertitudes

- Revenons à la propagation d'incertitudes de X à Z=G(X)

$$x^{(1)}, x^{(2)}, \ldots, x^{(n)} \longleftarrow$$ n-échantillon de X

- Estimation Monte Carlo de moyenne et variance de Z :

$$\mathbb{E}[G(X)] \approx \frac{1}{n}\sum_{i=1}^{n} G(x^{(i)})$$

$$\mathbb{V}[G(X)] \approx \frac{1}{n}\sum_{i=1}^{n}\left[G(x^{(i)}) - \frac{1}{n}\sum_{i=1}^{n} G(x^{(i)})\right]^2$$

- Les moments de Z sont estimés par les moments empiriques

3.3 Estimation de quantiles (1/3)

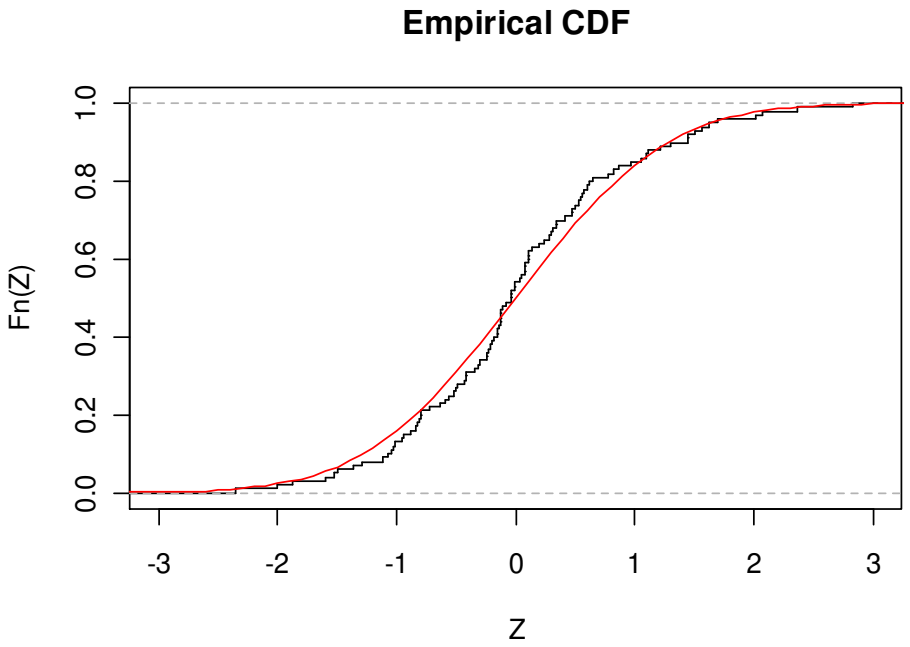
- Fonction de répartition empirique et estimateur du quantile

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

← Définition

$$\hat{F}_n(z) \to F(z)$$

← Théorème de Glivenko – Cantelli : convergence vers la fonction de rép. F(z)



Estimateur Monte Carlo d'un quantile de probabilité p : quantile empirique

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

Estimation de quantiles (2/3)

- Méthode pratique
  - Construire un échantillon ordonné à partir de $G\left(x^{(1)}\right), G\left(x^{(2)}\right) \ldots, G\left(x^{(n)}\right)$
  - Appelons-le $z^{(1)}, z^{(2)}, \ldots, z^{(n)} \quad z^{(1)} \leq z^{(2)} \leq \ldots \leq z^{(n)}$
  - $\hat{z}_p = z^{(\lceil np \rceil)}$ ← Plus petit entier strict. sup. à np
    - Définition équivalente à la précédente
    - Par exemple, si n=100 et p=0.95, alors il faut prendre la 96ème valeur dans l 'échantillon ordonné
    - Cette méthode n'a réellement de sens que si $\frac{1}{N} < p < 1 - \frac{1}{N}$
- Loi asymptotique de l'estimateur

$$\frac{\sqrt{n}}{\tau}\left(\hat{z}_p - z_p\right) \sim \mathcal{N}(0,1) \qquad \tau^2 = \frac{p(1-p)}{\left(f\left(z_p\right)\right)^2}$$

  - Expression peu pratique, car fait appel à la densité (inconnue) de z
    - Utilisée pour comparer (théoriquement) la variance de différents estimateurs

Estimation de quantiles (3/3)

- Intervalles de confiance pour le quantile
  - Asymptotiquement, l'erreur d'estimation tend vers une loi normale de moyenne nulle. Donc : $\mathbb{P}\left(\hat{z}_p \geq z_p\right) = 1/2$
    - Cet estimateur à une chance sur deux de surestimer ou sous-estimer la vraie valeur … ce qui peut poser problème (ex. études de sûreté)
    - Besoin d'estimateurs plus conservatifs du quantile. Idée :
      - Remplacer $z^{(\lceil np \rceil)}$ par $z^{(np+r)}$ pour avoir un nouveau $\hat{z}_p$ tel que :

$$\mathbb{P}\left(\hat{z}_p \geq z_p\right) \geq \beta \qquad \beta > 1/2$$

- Proposition :

Soit $j \in \mathbb{N}_0, j \leq n$

$$\mathbb{P}\left(j \text{ parmi les } z^{(i)} \text{ sont } > z*\right) = \binom{n}{j} \left(1 - F(z*)\right)^j F(z*)^{n-j} \Rightarrow$$

$$\mathbb{P}\left(j \text{ parmi les } z^{(i)} \text{ sont } > \hat{z}_p\right) = \binom{n}{j} \left(1 - p\right)^j p^{n-j}$$

Le nb de dépassements d'un seuil z_p par une suite de n v.a. i.i.d suit une loi binomiale B (n,F(z_p)) où F est la fct de répartition des z_i

Estimation de quantiles - Wilks

- On peut montrer que :

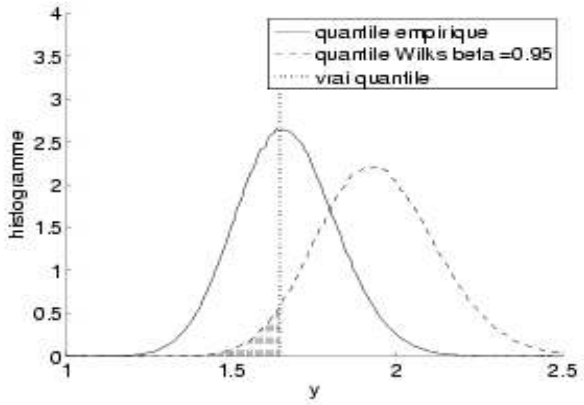
$$\mathbb{P}\left(z^{(np+r)} > z_p\right) = \sum_{j=n(1-p)-r+1}^{n} \mathbb{P}\left(j \text{ parmi les } z^{(i)} \text{ sont } > z_p\right) = 1 - C_p(n,r)$$

$$C_p(n,r) = \sum_{j=0}^{n(1-p)-r} \binom{n}{j} (1-p)^j \, p^{n-j}$$

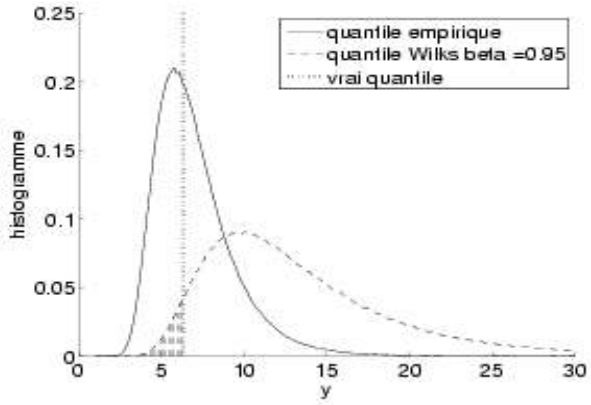
  - Donc, si r est le plus petit entier tel que : $C_p(n,r) \leq 1-\beta \implies 1 - C_p(n,r) \geq \beta$
  - alors, $\mathbb{P}\left(z^{(np+r)} > z_p\right) \geq \beta$

- C'est la méthode de Wilks
  - Estimateurs conservatifs des quantiles :
    - à n fixé, trouver le niveau de confiance β du quantile
    - à β fixé, déterminer le nombre n de tirages nécessaires

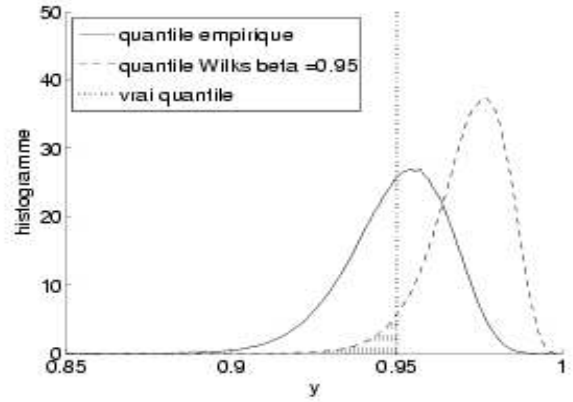
Comparaisons estimateur empirique / Wilks (n=200)



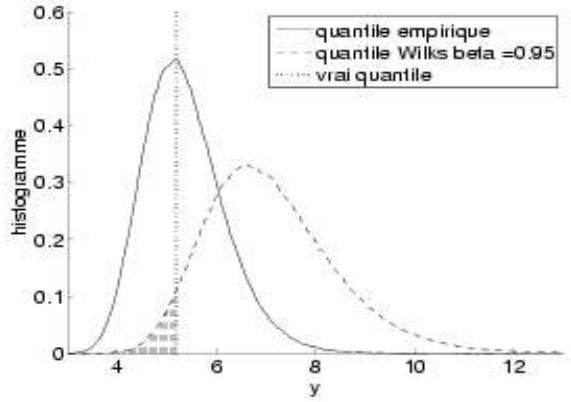
X N (0,1)



X Cauchy



X U (0,1)



X log-normal

Echantillonnage par la formule de Wilks

Commentaire :

- Méthode permettant de calculer :
  - ✓*N*, la taille de l'échantillon minimal nécessaire
  - ✓la valeur du quantile
- Méthode robuste
- S'applique à tout type de distribution même multimodale ou continue par morceaux

Contrainte :

L'échantillon doit être « purement » aléatoire (i.i.d.)

Exemple : tailles d'échantillons minimales pour un α-fractile unilatéral au niveau de confiance β en utilisant Wilks à l'ordre un

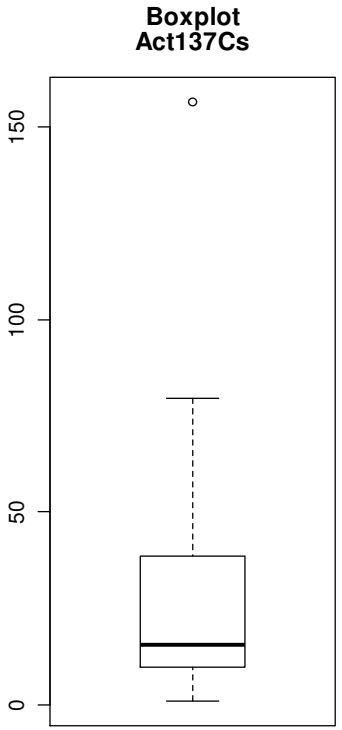
Z_{max} est la valeur maximale d'un N-échantillon i.i.d de Z

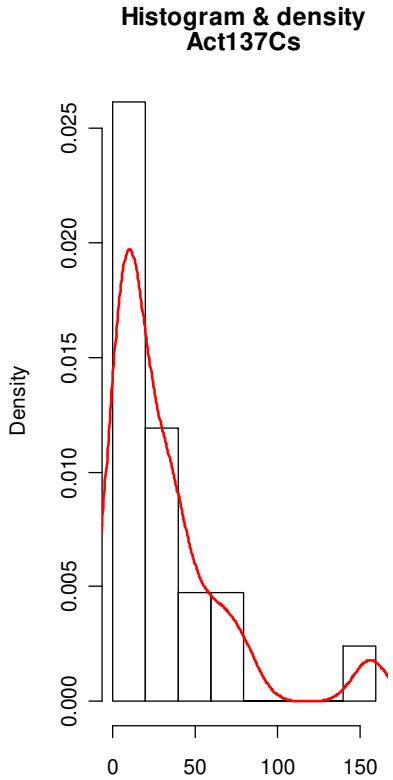
$$P\left[P\left(Z \le Z_{\max}\right) \ge \alpha\right] \ge \beta, \qquad N \text{ solution de } 1-\alpha^N \ge \beta$$

α	0.50	0.90	0.90	0.95
β	0.95	0.90	0.95	0.90
n	5	22	29	45

Exemple d'utilisation de la formule de Wilks en dehors du champ de la simulation numérique

Exemple : 21 mesures (concentrations d'un polluant) sur les murs d'une cellule
=> **Problème des échantillons de faible taille (*n* < 30)**





Statistiques élémentaires

- Moyenne = 31.45
  (mesure sensible aux valeurs élevées)
- Médiane = 15.4

- Ecart type = 36.11
- Min = 0.83 – Max = 156.67

- Skewness = 2.02
- Kurtosis = 4.19

Question de sûreté : comment garantir (avec un certain niveau de confiance) que la contamination n'excède pas un seuil intolérable en certains endroits ?

Exemples de questions économiques importantes : prédire la proportion totale d'activités < 50 , la proportion totale d'activités > 100 ?

Utilisation d'inégalités probabilistes

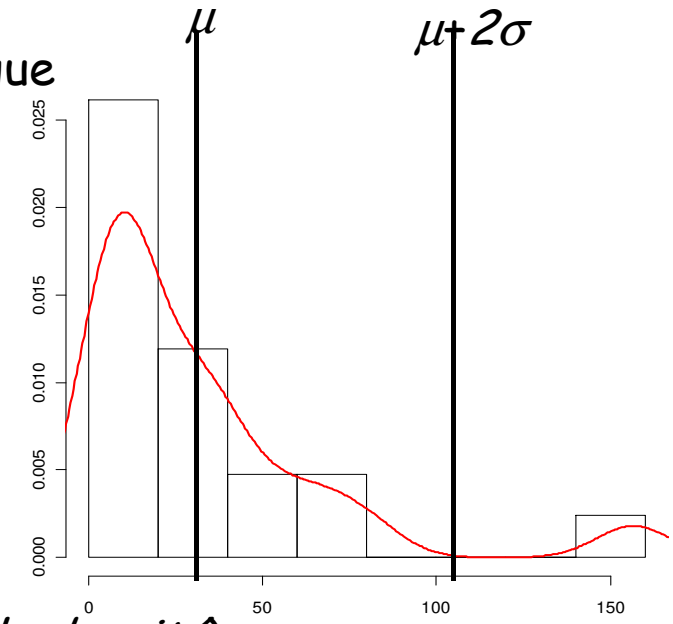
Pour une variable aléatoire X de moyenne μ and variance σ^2 , pour $X > \mu$, on a :

- L'inégalité de Bienaymé-Tchebytcheff : $P(X \leq \mu + k\sigma) > \frac{k^2}{1+k^2}$

Plus de 72% de la surface < 100

Borne très pessimiste

μ and σ^2 sont remplacés par leur estimateur empirique



- L'inégalité de Guttman :

$$P(X \leq \mu + k\sigma) > \frac{q^2}{1+q^2} \text{ with } q^2 = \frac{(k^2-1)^2}{\gamma_2 - 1}$$

Plus de 82% de la surface < 100

Nécessite la connaissance du kurtosis

- L'inégalité de Meidell (*hypothèse d'unimodalité de la densité*) :

$$P(X \leq \mu + k\sigma) > \frac{(3k/2)^2}{1+(3k/2)^2}$$

Plus de 89% de la surface < 100

Ces outils nous donnent des estimateurs sans IC

Utilisation de la formule de Wilks

Pour un échantillon i.i.d. $(X_1,\ldots,X_n)$, si n est solution de $1-\alpha^n \geq \beta$ et $X_{max} = \max\{X_1,\ldots,X_n\}$, on a $$P\left[P\left(X \leq X_{\max} \middle| (X_1,\ldots,X_n)\right) \geq \alpha\right] \geq \beta$$

Cela donne :

1. La taille minimale n pour α et β
2. Pour un échantillon donné, la valeur du α-quantile avec un degré de confiance β

α	0.50	0.90	0.95	0.95
β	0.95	0.95	0.90	0.95
n	5	29	45	59

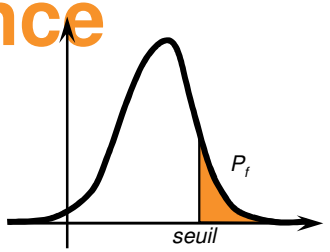
Pas d'hypothèse sur la distribution et pas besoin d'estimer des moments
(seule hypothèse : échantillon i.i.d.)

Application aux mesures de concentration de polluant :

- Wilks (n=21, r=2, β=0.9) -> plus de 83% de la surface < 80 (avec 90% de degré de confiance)

- Meidell (hypothèse d'unimodalité, estimation de σ) -> plus de 80% de la surface < 80

3.4 Estimation de probabilités de défaillance



- Défaillance du système : événement $Z \leq 0$
  - écriture classique (sans perte de généralité) où le seuil est nul et le système défaille quand la variable d'état est négative
  - Défaillance si R-S $\leq$ 0 (Résistance – Sollicitation)

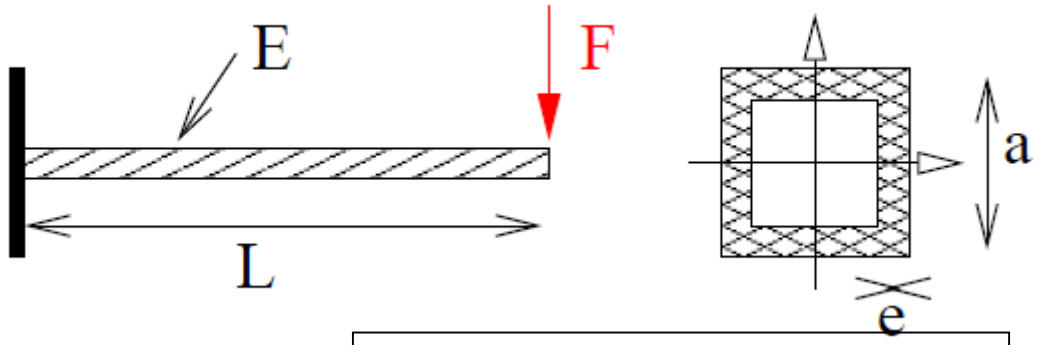
- Domaine de défaillance : $\mathcal{D}_f = \left\{ x \in \mathcal{X} : G(x) = z \leq 0 \right\}$

- Probabilité de défaillance : $p_f = \int_{\mathcal{D}_f} f(x)dx = \int_{\mathcal{X}} I_{\mathcal{D}_f}(x)\ f(x)dx = \mathbb{E}\left[I_{\mathcal{D}_f}(X) \right]$

  - Problème : calcul de l'espérance de la v.a. $I_{\mathcal{D}_f}(x)$

- Indicateur de défaillance : $I_{\mathcal{D}_f}(x) = \mathbb{1}_{\{G(x) \leq 0\}}$

Exemple classique en fiabilité des structures : poutre en flexion



Flèche : $y(E,F,L,I) = \frac{FL^3}{3EI}$

(déplacement vertical du bout)

F : force appliquée

E : module d'Young de la poutre

L : longueur de la poutre

I : moment quadratique

$$I = \frac{a^4 - (a-e)^4}{12}$$

Variable	Loi	Paramètres
F	A déterminer à partir des données mesurées	
E	Beta	r = 0.93 ; t = 3.2 ; a = 2.8e7 , b = 4.8e7
L	Uniforme	a = 250 ; b = 260
I	Beta	r = 2.5 ; t = 4.0 ; a = 3.1e2 , b = 4.8e7

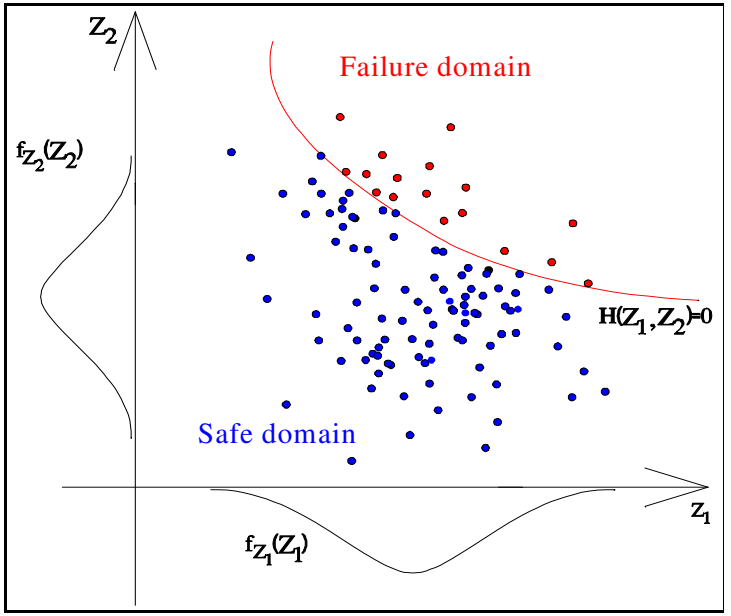
NB : les incertitudes proviennent des défauts/imprécisions dans les procédés de fabrication, dans les mesures, …

On s'intéresse à p_f = P(y > 30 cm) => probabilité de rupture de la poutre

Estimateur Monte Carlo de p_f (1/3)

- Estimateur Monte Carlo (naïf) :

$$\hat{p}_f = \frac{1}{n}\sum_{i=1}^{n} I_{\mathcal{D}_f}\left(x^{(i)}\right)$$



- Variance de l'estimateur :

$$\mathbb{V}[\hat{p}_f] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} I_{\mathcal{D}_f}\left(x^{(i)}\right)\right] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} I_{\mathcal{D}_f}\left(x^{(i)}\right)\right]$$

- Puisque : $I_{\mathcal{D}_f}\left(X^{(1)}\right),\ I_{\mathcal{D}_f}\left(X^{(2)}\right),\ldots I_{\mathcal{D}_f}\left(X^{(n)}\right) \sim \mathcal{B}(p_f)$ Bernouilli $i.i.d.$

- Alors : $\mathbb{V}[\hat{p}_f] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left[I_{\mathcal{D}_f}(x)\right] = \frac{1}{n^2}\, n\, p_f(1-p_f)$

$$\mathbb{V}[\hat{p}_f] = \frac{1}{n}p_f(1-p_f)$$

Estimée par :

$$\mathbb{V}[\hat{p}_f] \approx \frac{1}{n}\hat{p}_f(1-\hat{p}_f)$$

- On retrouve la convergence asymptotique vers une loi normale (propriété loi binom) … et toutes les propriétés des estimateurs MC

Estimateur Monte Carlo de p_f (2/3)

- Décroissance en racine de n : $\sigma_{\hat{p}_f} = \frac{1}{\sqrt{n}} \sqrt{p_f(1-p_f)}$
- Coefficient de variation : $cv = \frac{\sigma_{\hat{p}_f}}{\mathbb{E}[\hat{p}_f]} = \sqrt{\frac{p_f(1-p_f)}{n} \frac{1}{p_f^2}} = \sqrt{\frac{1-p_f}{n\ p_f}}$
- Pour des valeurs faibles de p_f : $p_f \to 0 \implies \frac{1-p_f}{p_f} \to \frac{1}{p_f}$

$$cv \approx \sqrt{\frac{1}{n\ p_f}}$$

« Erreur relative », précision de l'estimation

- Par exemple, si on veut estimer une proba $p_f = 10^{-r}$ avec un cv = 10%,

$$\sqrt{\frac{1}{n\ 10^{-r}}} = 10^{-1} \implies n = 10^{r+2}$$

10^{r+2} valeurs de G(X), donc 10^{r+2} appels au code G !

  - « règle du pouce » de l'ingénieur

Estimateur Monte Carlo de p_f (3/3)

- Estimateur « naïf » car coûteux !
  - Les temps de calcul deviennent vite prohibitifs dans la réalité industrielle
  - Par ex. pour des p_f de l'ordre de 10^{-4} → 10^6 appels à G(•)
  - Ce qui est coûteux est l'appel à G(•) !
    - Dans certains cas, un appel à G(•) peut demander des heures de temps CPU
- Plusieurs « parades »
  - Utiliser des techniques MC « accélérées » (à n égal, réduction de la var.)
  - Utiliser des techniques approximées (hypothèses supplémentaires) de type FORM/SORM pour une estimation rapide de p_f

Echantillonnage préférentiel (1/3)

- Idée : modifier la densité de tirage des X pour concentrer les tirages dans des régions plus intéressantes en termes de contribution au calcul de l'espérance de h(X)

$$I = \int_{\mathcal{X}} h(x) f(x) dx = \int_{\mathcal{X}} h(x) \frac{f(x)}{\varphi(x)} \varphi(x) dx = \int_{\mathcal{X}} h(x) w(x) \varphi(x) dx$$

- C'est l'espérance de la fonction $h(x)w(x), \quad X \sim \varphi(x)dx$

1) Générer un échantillon ($x^{(i)}$) à partir de la densité $\varphi(x)dx$

2) Puis, évaluer :

$$\hat{I}^{is} = \frac{1}{n} \sum_{i=1}^{n} h(x^{(i)}) w(x^{(i)}) \qquad \mathbb{V}\left[\hat{I}^{is}\right] = \frac{1}{n} \mathbb{V}\left[h(X) \frac{f(x)}{\varphi(x)}\right]$$

  - Estimateur sans biais de I, à condition que le support de φ(x) contienne celui de f(x)

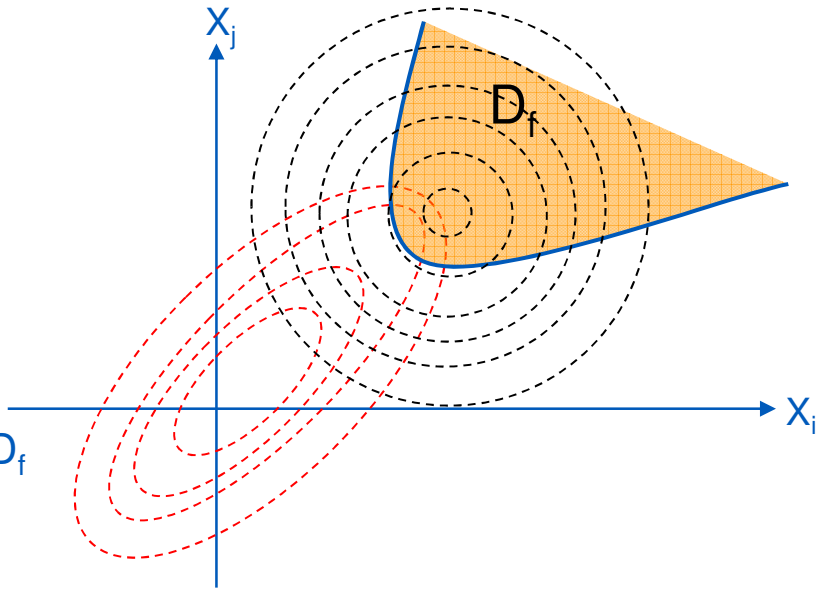
Echantillonnage préférentiel (2/3)

- Cette méthode ne garantit pas une réduction de la variance $\forall$ $\varphi(x)$
- Le choix de la « loi instrumentale » $\varphi(x)$ est crucial
  - Théoriquement : densité optimale $$\varphi^*(x) = \frac{|h(x)| f(x)}{\int_{\mathcal{X}} |h(x)| f(x) dx}$$
  - La constante de normalisation est aussi difficile à évaluer que I !
  - Néanmoins, ce résultat a un intérêt pratique …
- Estimation d'une probabilité de défaillance p_f par échantillonnage préférentiel
  - Ici : $h(x) = I_{\mathcal{D}_f}(x) = \mathbb{1}_{\{G(x) \leq 0\}}$
  - Densité optimale : $$\varphi^*(x) = \frac{I_{\mathcal{D}_f}(x) f(x)}{\int_{\mathcal{X}} I_{\mathcal{D}_f}(x) f(x) dx} = \frac{I_{\mathcal{D}_f}(x) f(x)}{p_f}$$

Echantillonnage préférentiel (3/3)

- La densité optimale est la loi conditionnelle de X sachant que $X \in D_f$
- C'est assez intuitif → La méthode est d'autant plus efficace qu'elle génère des points dans le domaine de défaillance

- Plusieurs manières de procéder …
  - Méthode courante : Avoir une première idée de la configuration de D_f (par exemple avec une méthode de type FORM/SORM)
  - Centrer la loi instrumentale sur un point de D_f (par exemple, le « point de conception » P*)
- Synonymes : Échantillonnage pondéré, tirage d'importance, importance sampling



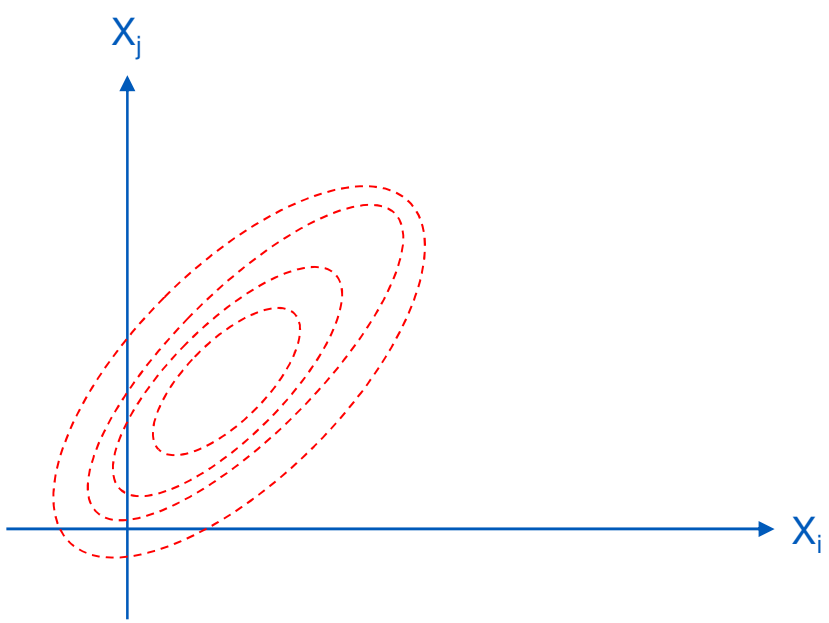
Méthode FORM (1/6)

- FORM: First Order Reliability Method

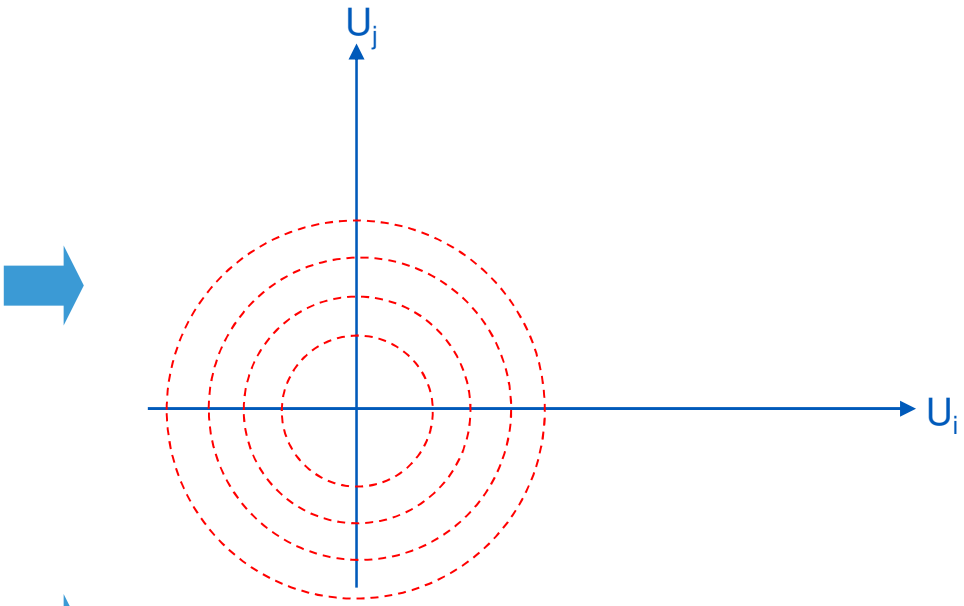
- Méthode typiquement fiabiliste (AFS)

- 3 étapes
  1. Transformation des variables X_i en d'autres variables dont la distribution de probabilité présente des « bonnes propriétés »
     Transformation isoprobabiliste → espace gaussien standard
  2. Recherche des conditions de défaillance les plus probables
  3. Évaluation de la probabilité de défaillance

Méthode FORM (2/6)
Transformation isoprobrobabiliste $U = \mathcal{T}(X)$



Espace physique



Espace « standard »

- Chaque composant de U suit une loi normale centrée-réduite
- Les composants de U sont indép.
- Les surfaces iso-prob. sont des sphères

$$\phi_d(\mathbf{u}) = \frac{1}{\left(\sqrt{2\pi}\right)^d} \exp\left(-\frac{1}{2}\sum_{i=1}^{d} u_i^2\right)$$

Les points qui contribuent le plus à p_f sont les plus proches de l'origine dans l'espace standard

FORM : transformation isoprobabiliste

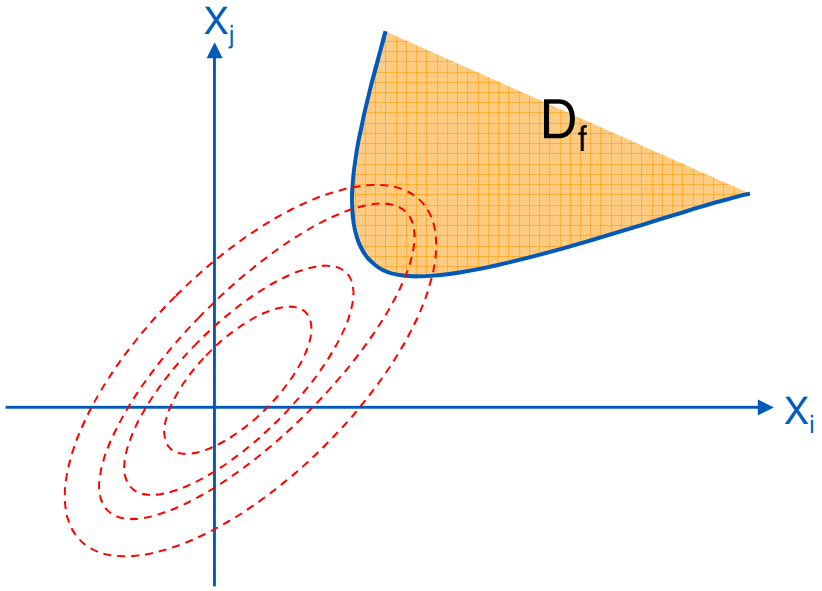
- Transformation de Rosenblatt

$$
\begin{aligned}
T \quad : \quad & u_1 = \Phi^{-1}\left(F_1(z_1)\right) \\
& u_2 = \Phi^{-1}\left(F_2(z_2|z_1)\right) \\
& \vdots \\
& u_N = \Phi^{-1}\left(F_N(z_N|z_{N-1},\cdots,z_2,z_1)\right)
\end{aligned}
$$

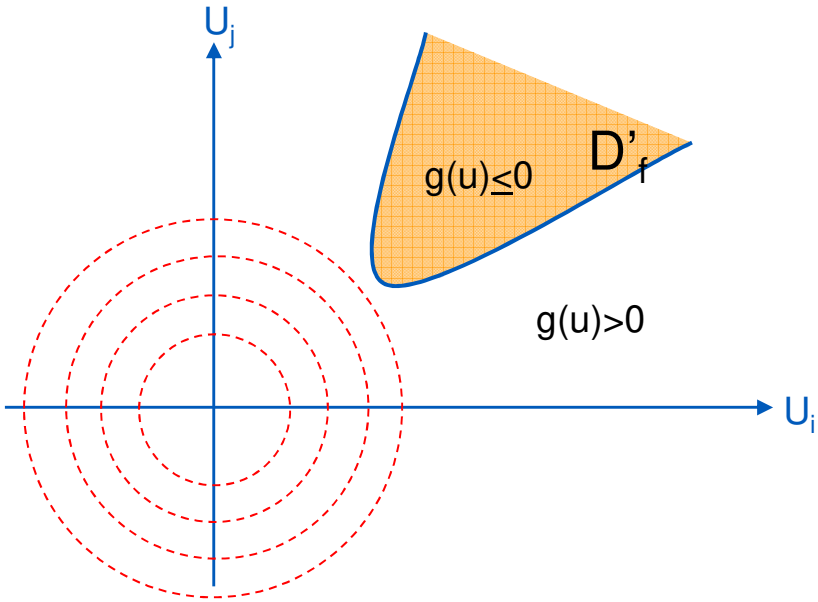
Méthode FORM (3/6) – Transformation isoprob.

- Nouvelle expression de la probabilité de défaillance

$$\mathbb{P}\left(G(X) \leq 0\right) = \mathbb{P}\left(G(\mathcal{T}^{-1}(U) \leq 0\right) = \mathbb{P}\left(g(U) \leq 0\right)$$



Espace physique



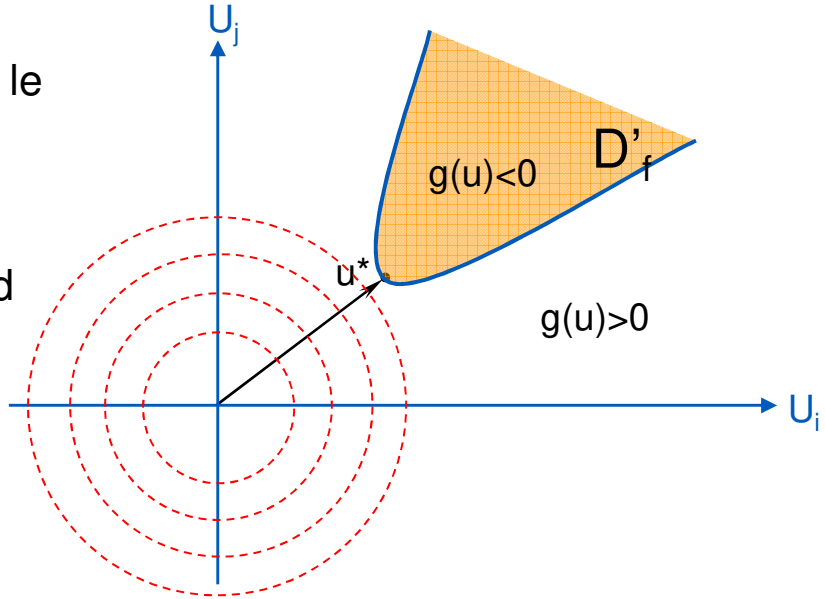
Espace « standard »

– Expression de la prob. de défaillance :

$$p_f = \int_{\mathcal{T}(\mathcal{X})} \mathbb{1}_{g(u)\leq 0} f_u(u)du$$

Méthode FORM (4/6) – Recherche des conditions de défaillance « les plus probables »

- A chaque point de l'espace standard, on peut associer des conditions de fonctionnement ou de défaillance du système.
  - Le point de « défaillance » le plus probable est le plus proche de l'origine (où la densité est max, puisque la moyenne de U est le vecteur nul)
  - Rappel : la valeur de la densité $f_U(u)$ ne dépend que de ||u|| (distance de u de l'origine)
  - Appelons-le P* : point de conception
  - Appelons u* le vecteur OP*
  - La recherche de u*, qu'on suppose unique, est un problème d'optimisation sous contraintes



$$u^* = \min_{g(u)=0} \beta(u) = \min_{g(u)=0} \sqrt{u^t u}$$

Méthode FORM (5/6) – Evaluation de p_f

- Hypothèse :
  - Remplacement de la surface limite g(u)=0 par l'hyperplan passant par P* et orthogonal au vecteur u*, d'équation :

$$\sum_{i=1}^{N} \alpha_i u_i + \beta = 0$$

β : norme de u*
α_i cos. direct de u*

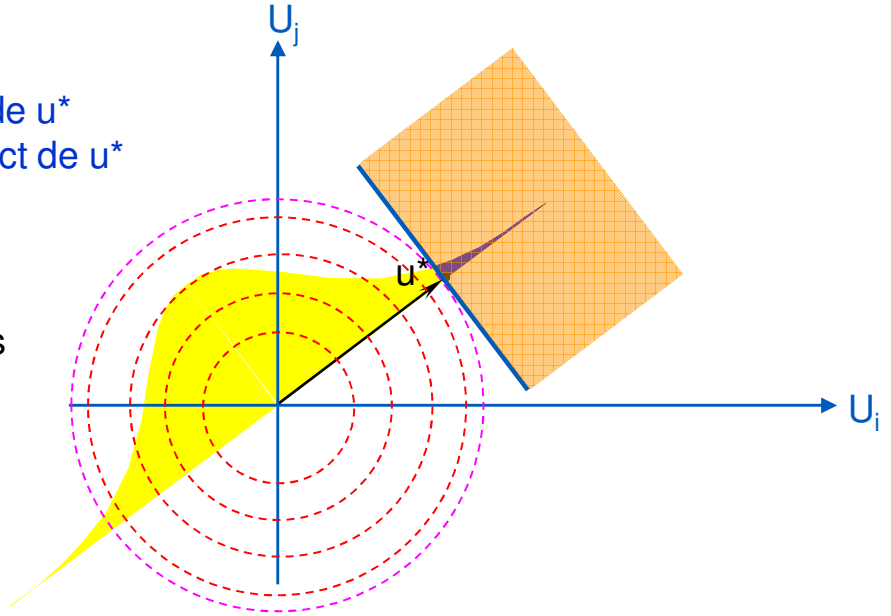
  - Approximation basée sur l'hypothèse que les points éloignés de P* (où l'approximation est mauvaise) contribuent peu au calcul de p_f → leur proba est très faible

$$p_f \approx \mathbb{P}\left(\sum_{i=1}^{N} \alpha_i u_i + \beta \leq 0\right) =$$

$$\mathbb{P}\left(\sum_{i=1}^{N} \alpha_i u_i \leq -\beta\right) = \Phi(-\beta)$$

Combinaison linéaire de v.a. N(0,1) avec coeff. αi normés, donc c'est une N(0,1)

fonction de rép. de la loi normale centrée-réduite



β : « Indice de fiabilité »
α_i : « Facteurs d'importance » FORM
→ analyse de sensibilité de pf aux variables Ui.

Méthode FORM (6/6) – Commentaires

- La linéarité de la surface de défaillance dans l'espace standard est vérifiée en théorie si (cond. suffisantes) :
  - Elle est linéaire dans l'espace physique et les variables X sont normales
    - Plus on s'éloigne de ces hypothèses, moins l'approximation FORM est bonne
  - Si la surface de défaillance a toujours la même courbure dans l'espace standard on déduit des formules d'encadrement de l'erreur liée à l'approximation FORM

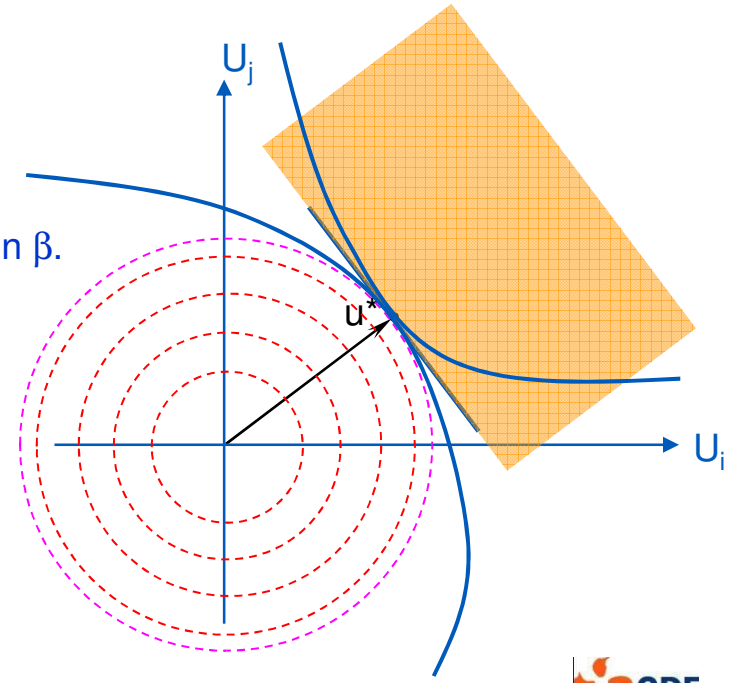
- Cas concave (concavité tournée vers l'origine) :

$$\Phi(-\beta) \leq p_f \leq 1 - F_{\chi^2_N}(\beta^2)$$

Approx. FORM — Prob. de l'espace ext. à la sphère de rayon β.

- Cas convexe (trivial !) : $0 \leq p_f \leq \Phi(-\beta)$

- Ces formules donnent un encadrement très grossier de l'approximation FORM



Méthode SORM

- SORM – Second Order Reliability Method
  - On remplace l'hyperplan de FORM par une surface d'ordre 2
  - Formule de Breitung : $$p_f \approx \Phi(-\beta) \prod_{i=1}^{N} \frac{1}{\sqrt{1+\beta\kappa_i}}$$

    N-1 courbures principales dans le point de conception NB $\kappa_i=(1/R_i)$ → 0 si la surface est un hyperplan → on retrouve le résultat FORM)
  - D'autres formules sont disponibles (Hohembichler, Tvedt) …

FORM/SORM : Avantages et limites

- Avantages :
  - **Temps de calcul réduit** par rapport aux méthodes de simulation
  - Pas de dépendance du temps de calcul au niveau de probabilité à calculer
  - **Même durée pour des probabilités de l'ordre de 10^{-1} ou de 10^{-9} !**
    - Calcul de facteurs d'importance et d'un point de conception
  - Etudes de sensibilité simples

- Limites :
  - Approximation pas toujours correcte
  - Pas d'erreur de mesure :
    - Seulement une borne supérieure peu adaptée
  - G doit être *différentiable*
  - Hypothèse d'un minimum global contrasté

Conclusions sur la propagation d'incertitudes

- **Enjeu :** Arbitrer entre précision de l'estimateur et coût des calculs

- Si possible**, Monte Carlo** est à privilégier : indépendant de la dimension des entrées, estimation non biaisée, fournit un intervalle de confiance sur l'estimation

  Mais : coût important en nombre d'évaluations du modèle

- Si le code de calcul est trop coûteux en CPU, il existe des méthodes alternatives :
  - Méthode Monte Carlo accélérées (tirage d'importance, etc.)
  - Méthodes quasi-Monte Carlo (cf. cours 2) - Mais : fléau de la dimension

  - Méthodes approchées :
    - Cumul quadratique (développement de Taylor) - Mais : hypothèse linéaires
    - Méthodes FORM/SORM : estimation rapide de p_f . Cette première estimation peut être utilisée pour construire un tirage d'importance

  - Utilisation d'un modèle de substitution du code de calcul (cf. cours 3) ayant un coût pratiquement nul (métamodèle)
    - Attention : un nouveau terme d'erreur apparaît
    - Le calage du métamodèle demande aussi un certain nombre d'appels au vrai modèle G

Crédits & Bibliographie

- Tutoriel « Incertitudes », JdS 2011, A. Pasanisi (EDF R&D)
- Formation « Démarche Incertitudes », IMdR-LNE
- Cours « Uncertainty analysis for engineering activities », 2011, F. Mangeant (EADS IW)
- Cours « Introduction à la statistique », Ecole d'été CEA-EDF-INRIA, A. Marrel (CEA)
- De Rocquigny, Devictor & Tarantola (eds), Uncertainty in industrial practice, Wiley, 2008
- Robert & Casella, Monte Carlo Statistical Methods, 2nd Edition, Springer, 2004
- Rubinstein, Simulation and the Monte Carlo method, Wiley, 1981
- Guide to the expression of Uncertainty in Measurements (GUM), ISO publication
- Lemaire, Fiabilité des structures. Lavoisier, 2005

Ce cours est disponible sur : http://www.gdr-mascotnum.fr/doku.php?id=iooss1#academic