Analyses statistiques des expériences numériques

Cours 2 : Propagation des incertitudes et inférence d'événements rares

Bertrand Iooss

Polytech Nice Sophia

Décembre 2025

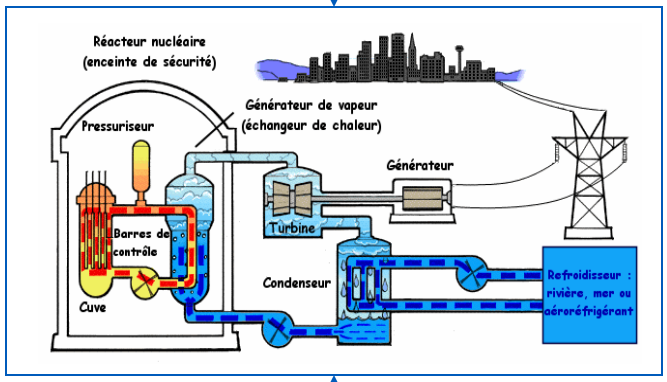
CHANGER L'ÉNERGIE ENSEMBLE

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

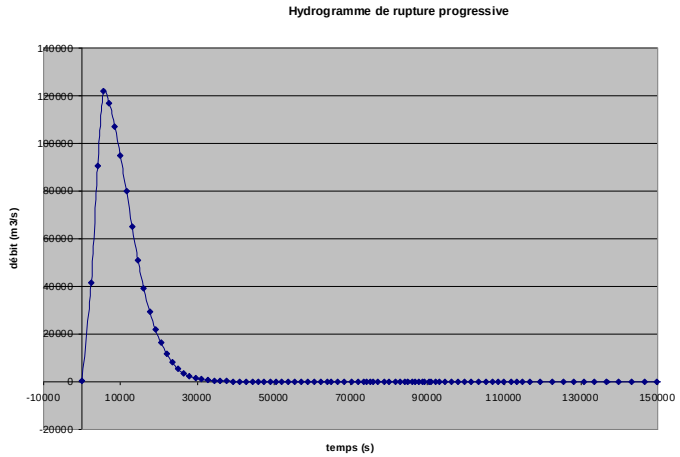
Pas de REX de défaillance mais un modèle physique

INCERTITUDES EN SIMULATION NUMERIQUE

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

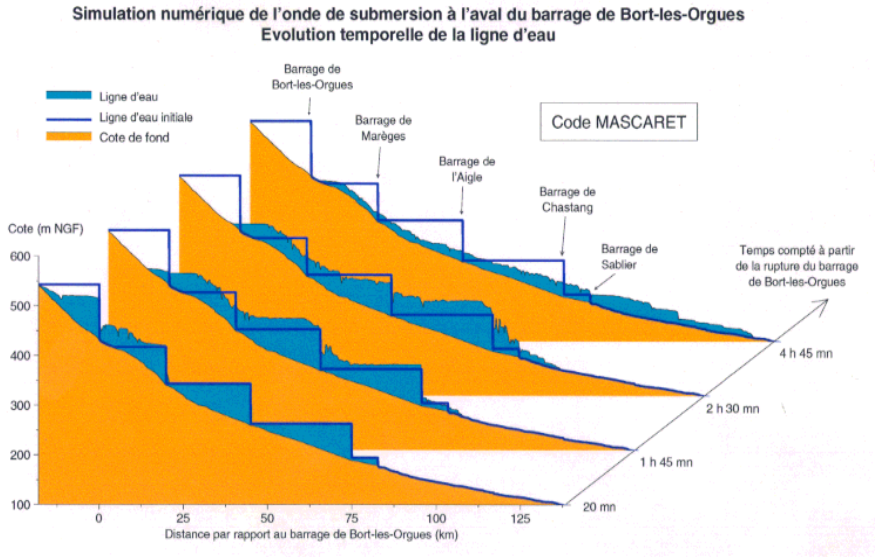
- L'objectif : évaluer la côte maximale de l'eau et le temps d'arrivée de l'onde de submersion

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

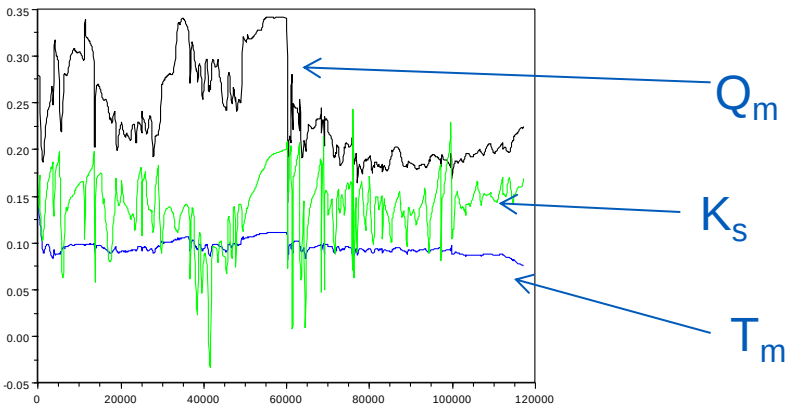


Exemple : simulation de rupture d'un barrage (2/2)

- 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



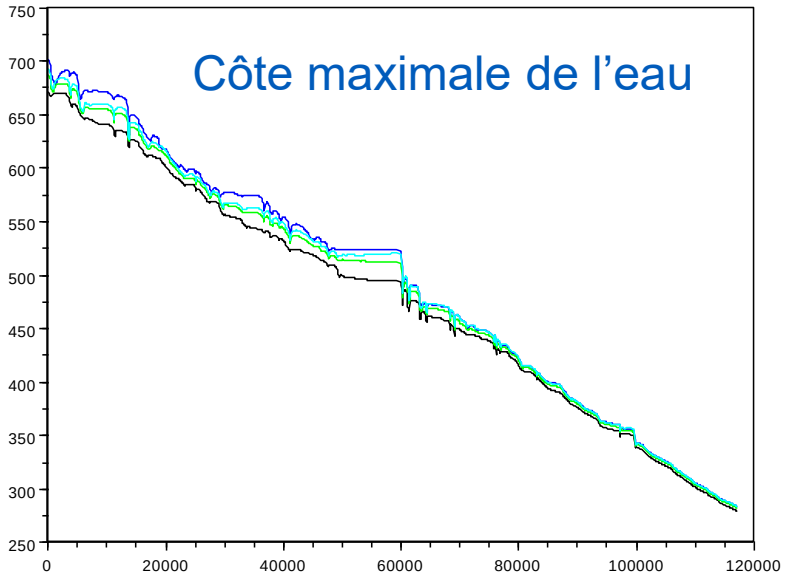
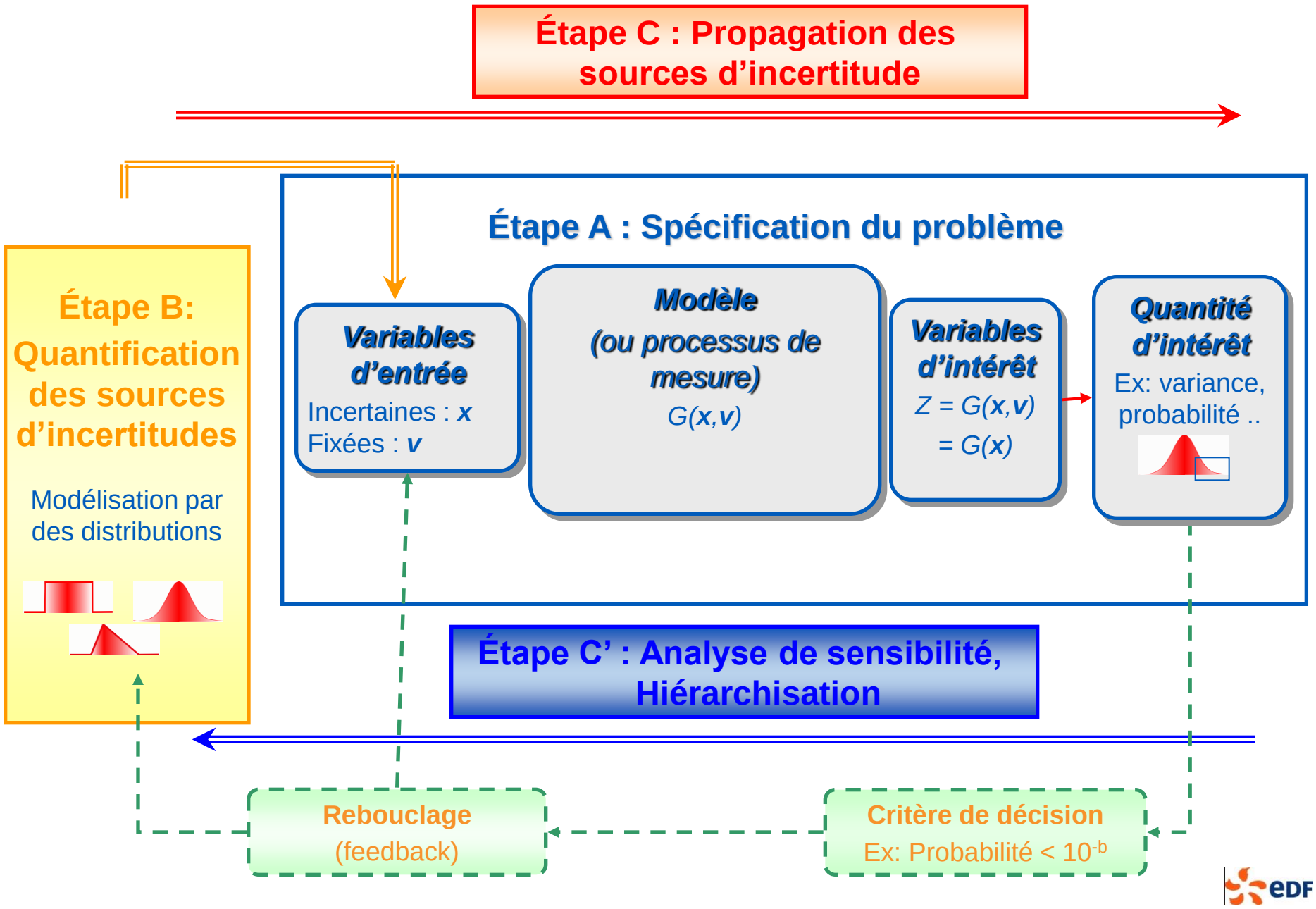
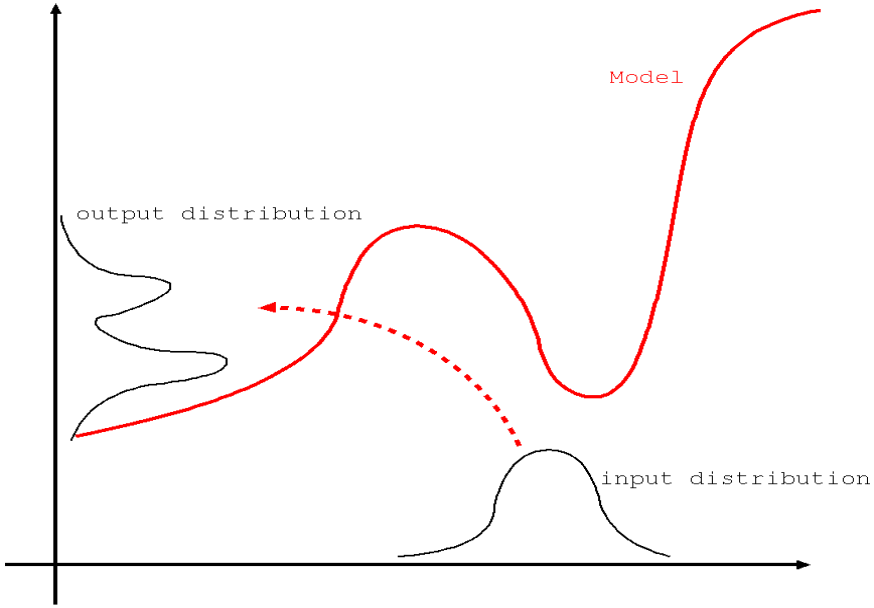


Schéma générique



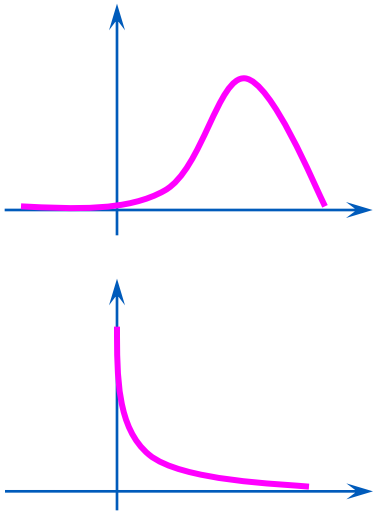
Propagation d'incertitudes

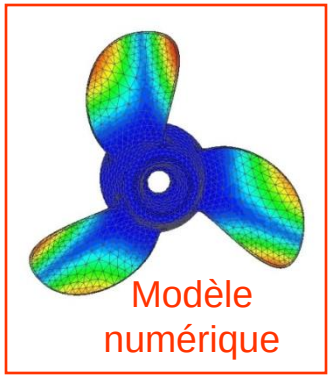
- Transfert des incertitudes de **X** ϵ $\mathfrak{R}^d$ vers Z ϵ $\mathfrak{R}$, via la fonction déterministe $G(\bullet)$
- **X** (noté aussi $\underline{X}$ ou X) est un vecteur aléatoire, avec une certaine mesure de proba
- $Z = G$ (**X**) devient un vecteur aléatoire, avec une mesure de proba à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

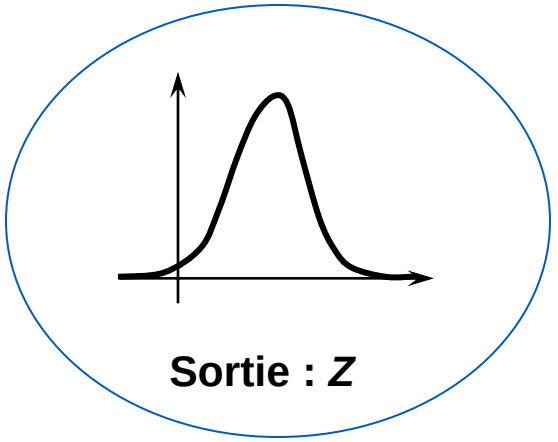


- Problème conceptuellement « simple » mais (parfois) de mise en œuvre complexe
- Le choix de la méthode dépend très fortement de la « quantité d'intérêt » de l'étude
- … d'où l'importance de l'étape A, peu mathématique mais essentielle en pratique

Etape A : la quantité d'intérêt (1/2)

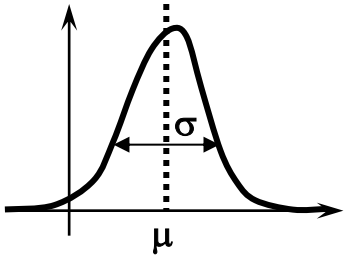




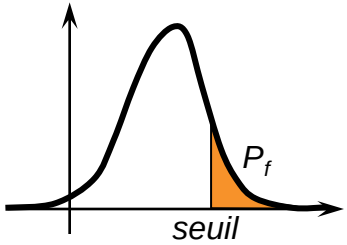


Entrées : X

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

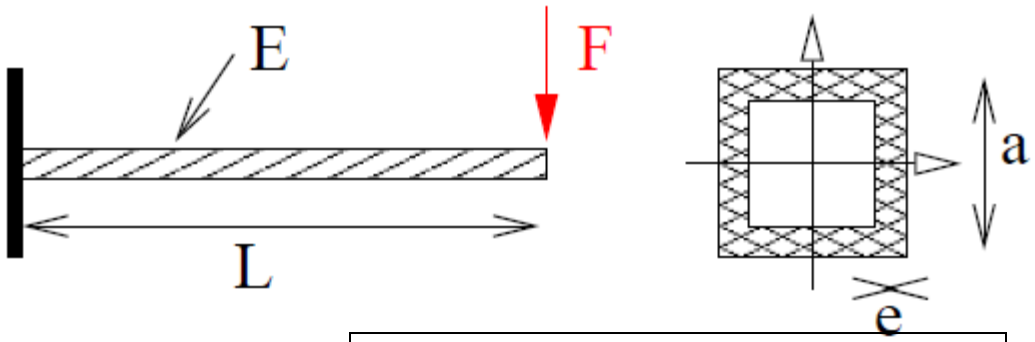
→ Méthodes numériques (optimisation, échantillonnage Monte Carlo)

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 $P_f = \mathrm{P}(Z \geq z^*)$

TP : 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	Normale	moy = 30000 ; écart-type 9000
E	Triangulaire	a = 2.8e7 ; b = 4.8e7 ; c = 3.0e7
L	Uniforme	a = 250 ; b = 260
I	Triangulaire	a = 310 ; b = 450 ; c = 400

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

Tendance centrale : estimation de moyenne et variance

Etape C - 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
    - $X_1, \ldots, X_d$: variables aléatoires
    - $a_1, \ldots, a_d$: réels

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

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

    - Ces formules donnent la moyenne et la variance de $Z=G(\mathbf{X})$ 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 :

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

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

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

$$G(\mathbf{X}) = G(\mu) + \sum_{i=1}^{d} \left.\frac{\partial G}{\partial X_i}\right|_{X=\mu} (X_i - \mu_i)$$

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

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

Cumul quadratique – Développement d'ordre 1

Calcul de la moyenne de Z

$$\mathrm{E}(Z) = G(\mu)$$

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

$$\mathrm{Var}(Z) = \mathrm{E}(Z - \mathrm{E}(Z))^2 = \mathrm{E}\left(G(\mu) + \sum_{i=1}^{d} \left.\frac{\partial G}{\partial X_i}\right|_{X=\mu} (X_i - \mu_i) - G(\mu) \right)^2$$

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

$$\mathrm{Var}(Z) = \sum_{i=1}^{d}\sum_{j=1}^{d} \left.\frac{\partial G}{\partial X_i}\right|_{X=\mu} \left.\frac{\partial G}{\partial X_j}\right|_{X=\mu} \rho_{ij}\sigma_i\sigma_j$$

<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 :

$$\mathrm{Var}(Z) = \sum_{i=1}^{d} \left(\left. \frac{\partial G}{\partial X_i} \right|_{X=\mu} \right)^2 \sigma_i^2$$

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)

$$\eta_i^2 = \frac{1}{\mathrm{Var}(Z)} \left(\left. \frac{\partial G}{\partial X_i} \right|_{X=\mu} \right)^2 \sigma_i^2$$

Indices de sensibilité (normés)

L'analyse de sensibilité est réalisée de manière directe

Etape C - 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 i.i.d. 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 niveau 1-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 le 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 $Z=G(\mathbf{X})$

$x^{(1)}, x^{(2)}, ..., x^{(n)}$ ← n-échantillon i.i.d. 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

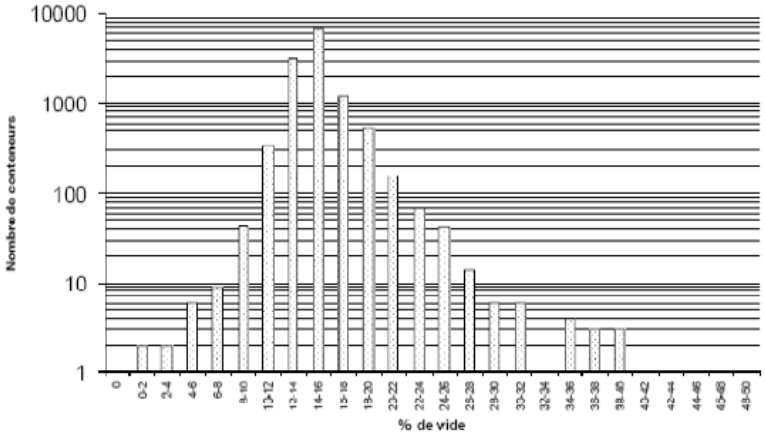
Exemple : propagation d'incertitudes dans un calcul thermique du stockage de déchets radioactifs (1/3)

[EDF R&D/SINETICS – Lefebvre, Barate, Leroyer]

- L'Andra impose une contrainte de température maximale de 100°C sur l'argile, diminuée de 10°C pour cause « d'incertitudes »
- Problème : ces 10°C de marges sont-ils pertinents ?
- Solution : réaliser des calculs de propagation d'incertitudes dans un calcul de thermique du stockage des déchets de Haute-Activité et à Vie Longue
  - Variable d'intérêt :
    - Température maximale en Barrière Géologique
  - Critères d'étude :
    - **Tendance centrale** : moyenne et écart-type des variables d'intérêt et hiérarchisation des sources d'incertitudes
  - Temps cpu pour 1 calcul de thermique :
    - 10 mn sur 8 processeurs

Exemple : propagation d'incertitudes dans un calcul thermique du stockage de déchets radioactifs (2/3)

- Trois familles d'incertitudes étudiées :
  - Taux de vide des colis standard de déchets vitrifiés
    - Incertitude stochastique réductible
    - Source : REX de production des colis à La Hague



  - Propriétés physiques du stockage
    - Incertitude stochastique irréductible
    - Source : dossier Andra 2005

Unité thermique	Profondeurs [m] (forage EST205)	Conductivité thermique à 20°C (W/m/K)	Chaleur spécifique à 20°C [J/kg/K]
		λ_0 (écart-type)	C_p^b (écart-type)
Sous zone 2	*–473,5 à –516*	1,3 ⊥ *(0,17)* 1,9 // *(0,48)*	1005 *(70)* [c]

  - Puissance thermique résiduelle des colis standard de déchets vitrifiés
    - Incertitude épistémique
    - Source : calculs d'incertitude sur la puissance thermique résiduelle (projet PRETING, département SINETICS)

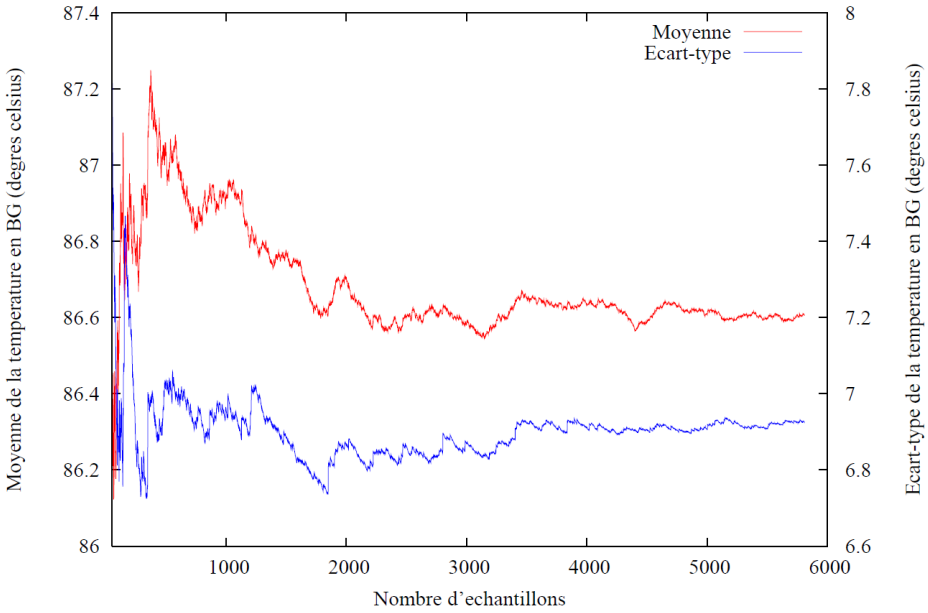
Exemple : propagation d'incertitudes dans un calcul thermique du stockage de déchets radioactifs (3/3)

Résultats :

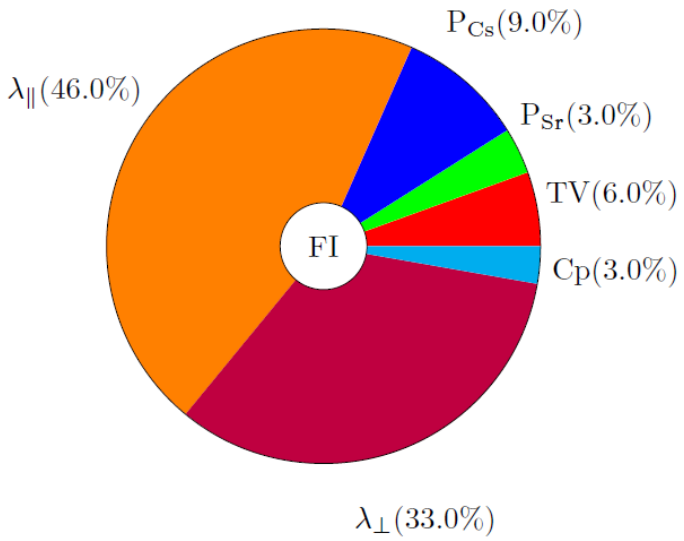
- Moyenne et écart type de la température maximale sur la barrière géologique

Nombre de calculs	Moyenne	Écart-type
5812	86.60°C	6.93°C

- On retrouve l'ordre de grandeur de 10°C de marge prise par l'Andra sur le critère thermique
- Prédominance des incertitudes sur les paramètres physiques de l'argile

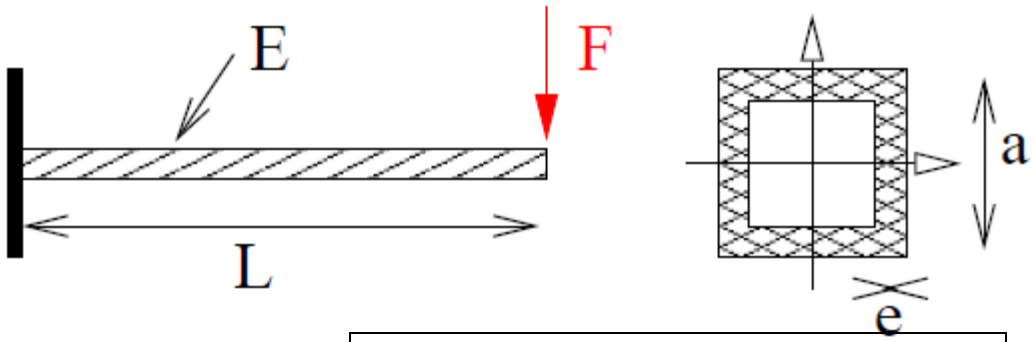


- Convergence des estimateurs Moyenne et Écart-Type



- Facteurs d'importance des incertitudes d'entrée

TP : 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	Normale	moy = 30000 ; écart-type 9000
E	Triangulaire	a = 2.8e7 ; b = 4.8e7 ; c = 3.0e7
L	Uniforme	a = 250 ; b = 260
I	Triangulaire	a = 310 ; b = 450 ; c = 400

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

On s'intéresse à la moyenne et à la variance du déplacement de la poutre

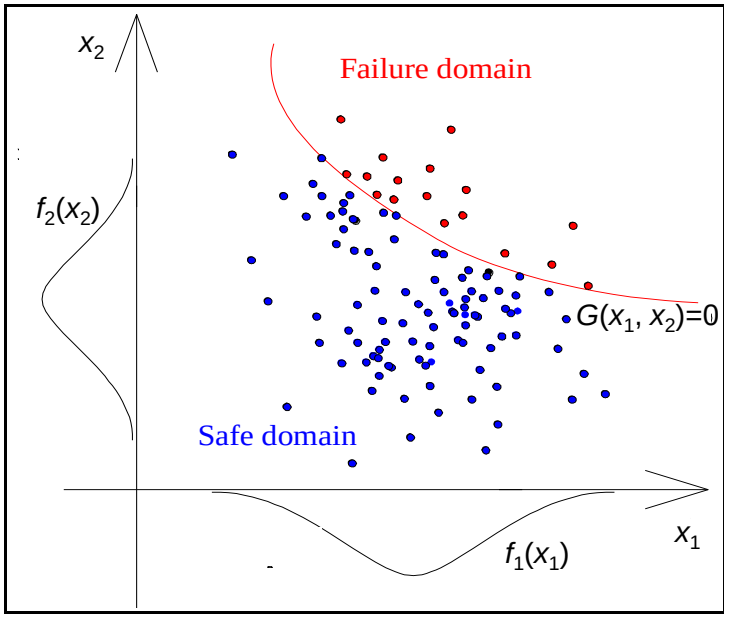
Inférence d’évenements rares

Inférence d'événements rares

- 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 (« failure ») :

$$\mathcal{D}_f = \{x \in \mathcal{X} : G(x) = z \leq 0\}$$

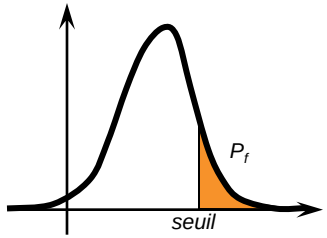


- Probabilité de défaillance :

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

- Quantile : z_α tel que $\mathrm{P}(Z < z_\alpha) = \alpha$

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)

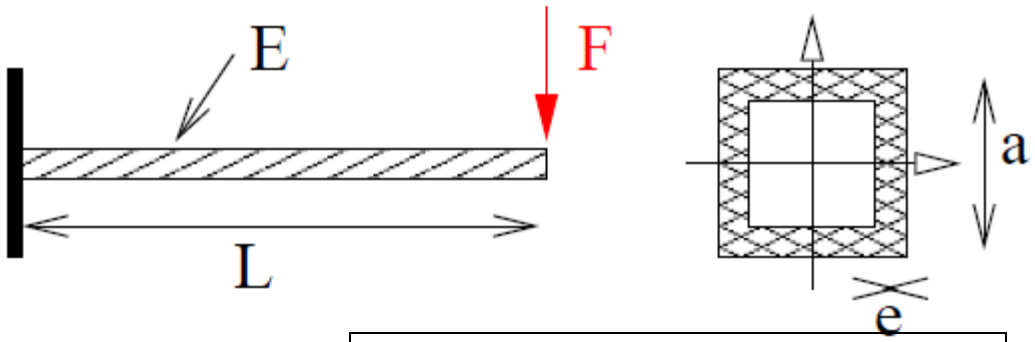
- Domaine de défaillance : $\mathcal{D}_f = \left\{ x \in \mathcal{X} : G(x) = z \leq 0 \right\}$

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

TP : 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	Normale	moy = 30000 ; écart-type 9000
E	Triangulaire	a = 2.8e7 ; b = 4.8e7 ; c = 3.0e7
L	Uniforme	a = 250 ; b = 260
I	Triangulaire	a = 310 ; b = 450 ; c = 400

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

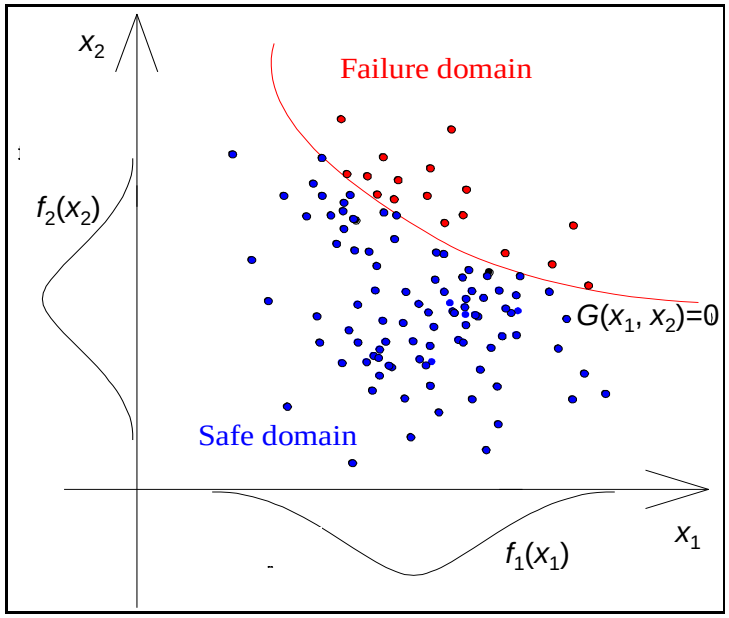
Estimateur Monte Carlo de p_f (1/3)

- 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 $G(\bullet)$
  - Ce qui est coûteux est l'appel à $G(\bullet)$!
    - Dans certains cas, un appel à $G(\bullet)$ 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

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 3) - 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

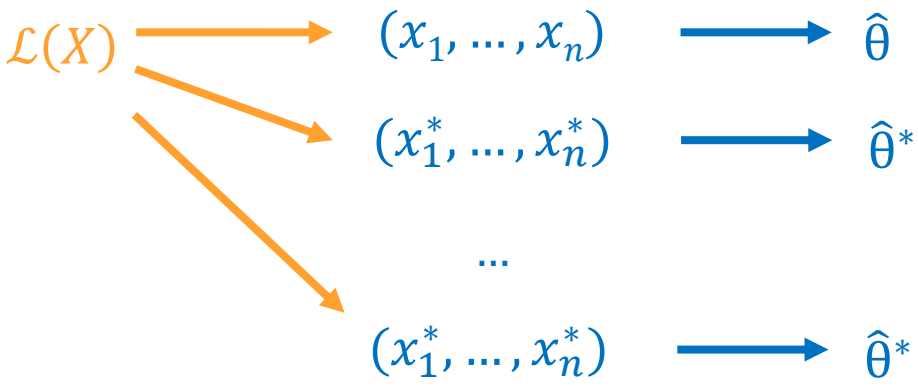
  - Utilisation d'un modèle de substitution du code de calcul (cf. cours Michaël Binois) ayant un coût pratiquement nul (métamodèle)
    - Attention : un nouveau terme d'erreur apparaît
    - Le calage du métamodèle demande aussi un certain nombre d'appels au vrai modèle G

Annexe : Introduction au bootstrap

Incertitude EN ESTIMATION

- $(x_1, \dots, x_n)$ échantillon d'une loi inconnue, de paramètre d'intérêt θ
- Soit $\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$ un estimateur de θ (Monte-Carlo par exemple)
- But : Quantifier l'incertitude sur l'erreur d'estimation $(\hat{\theta} - \theta)$
  - Biais $b = E[(\hat{\theta} - \theta)]$
  - Variance : $v = E[(\hat{\theta} - \theta)^2]$
  - Int. de confiance : $IC_{1-\alpha} = \hat{\theta} + \left[q_{\frac{\alpha}{2}}(\theta - \hat{\theta}); q_{1-\frac{\alpha}{2}}(\theta - \hat{\theta})\right]$
- Ces trois indicateurs dépendent de la loi $\mathcal{L}(\hat{\theta} - \theta)$, inconnue

Principe du bootstrap (1/3)

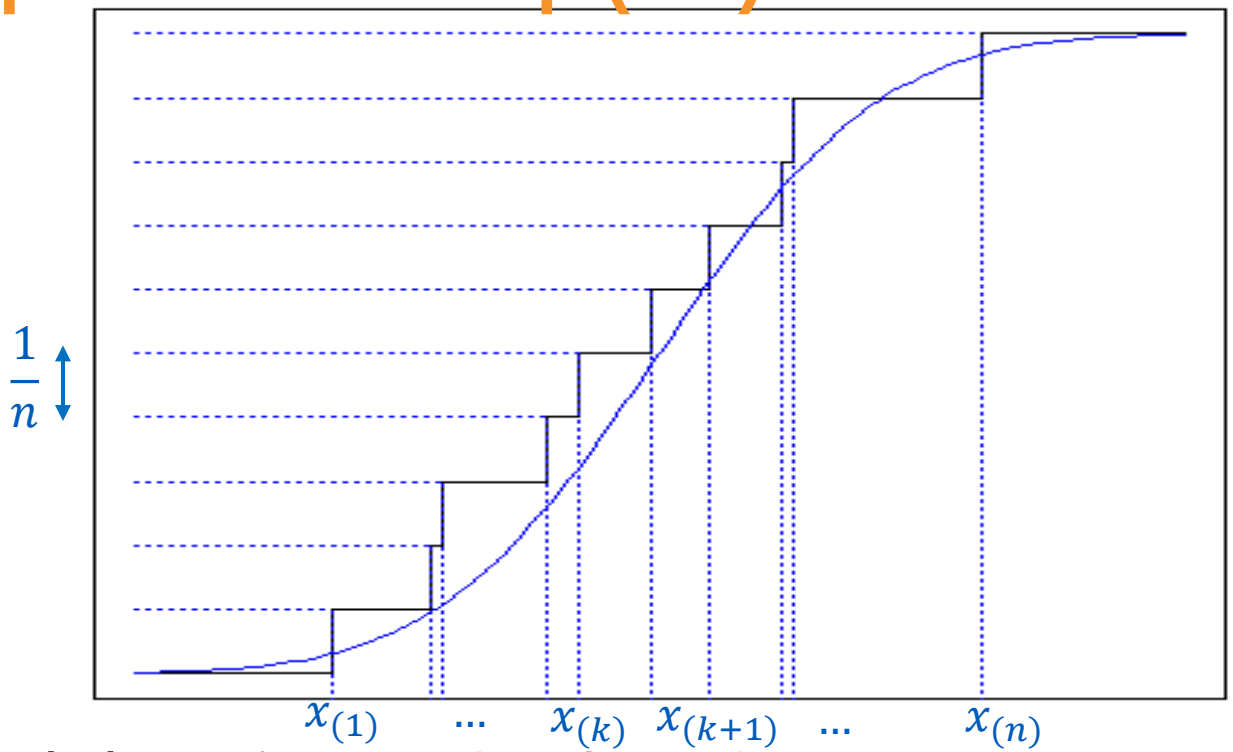


- Idée : Estimer les caractéristiques de la loi $\mathcal{L}(\hat{\theta} - \theta)$ par Monte-Carlo, à l'aide d'un grand nombre de tirages dans la loi de $\hat{\theta}$
- Nécessite de simuler de nouvelles valeurs possibles de $\hat{\theta}$, à l'aide des étapes suivantes :

1. Générer un nouvel échantillon $(x_1^*, \dots, x_n^*)$, de même loi que $(x_1, \dots, x_n)$
2. Ré-estimer le paramètre à l'aide du nouvel échantillon : $\hat{\theta}^* = \hat{\theta}(x_1^*, \dots, x_n^*)$

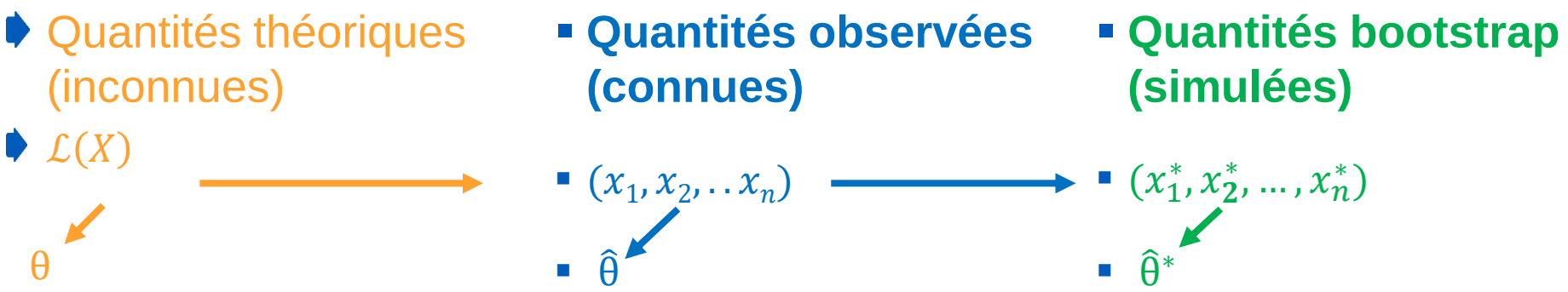
- Question : Comment générer $(x_1^*, \dots, x_n^*)$?

Principe du bootstrap (2/3)



- Objectif : Générer $(x_1^*, \ldots, x_n^*)$ de même loi que $(x_1, \ldots, x_n)$
  - **Inversion générique :** Requiert la fonction de répartition $F(x) = P[X \leq x]$, **inconnue**
  - **Solution :** Utiliser la **fonction de répartition empirique** $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \leq x\}}$
    - Revient à **tirer** les x_i^* **uniformément avec remise** parmi $x_1, x_2, .. x_n$

Principe du bootstrap (3/3)



- L'approche bootstrap consiste à :

1. Approcher la valeur du paramètre θ par son estimateur $\hat{\theta}$
2. Approcher la loi de l'estimateur $\hat{\theta}$ par celle de l'estimateur bootstrap $\hat{\theta}^*$

- Les grandeurs caractéristiques de $\mathcal{L}(\hat{\theta} - \theta)$ (moyenne, variance, etc) sont alors estimées par Monte-Carlo à l'aide de l'échantillon $(\hat{\theta}_b^* - \hat{\theta})_{1 \leq b \leq B}$

MISE EN OEUVRE DU BOOTSTRAP

1. Générer un échantillon bootstrap $(x_1^*, \dots, x_n^*)$, en tirant uniformément avec remise dans l'échantillon d'origine $(x_1, x_2, .. x_n)$:
   - Pour tout i, x_i^* est tiré au hasard parmi $x_1, x_2, .. x_n$
   - En général, $(x_1^*, \dots, x_n^*)$ contient des doublons
2. Ré-estimer θ à partir des $(x_1^*, \dots, x_n^*)$: on obtient une valeur bootstrap $\hat{\theta}^*$
3. On répète les étapes 1. et 2. un grand nombre B de fois ; on obtient ainsi un **échantillon bootstrap** $(\hat{\theta}_1^*, \dots, \hat{\theta}_B^*)$

➧ On peut alors estimer l'incertitude sur l'erreur d'estimation par :

- Le **biais bootstrap** : $\hat{b}^* = E\left[\left(\hat{\theta}^* - \hat{\theta}\right)\right] = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta})$
- La **variance bootstrap** : $\hat{v}^* = E\left[\left(\hat{\theta}^* - \hat{\theta}\right)^2\right] = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta})^2$
- L'**int. de confiance bootstrap** : $\hat{\theta} + \left[q_{\frac{\alpha}{2}}(\hat{\theta} - \hat{\theta}^*); q_{1-\frac{\alpha}{2}}(\hat{\theta} - \hat{\theta}^*)\right] = 2\hat{\theta} - \left[\hat{\theta}^*_{\left(\left[B\left(1-\frac{\alpha}{2}\right)\right]\right)}; \hat{\theta}^*_{\left(\left[B\frac{\alpha}{2}\right]\right)}\right]$

Bootstrap : avantages et limites

- **Avantages du bootstrap :**
  - Méthode **simple** et **générique**
  - S'applique à **tous les paramètres**, **et toutes les méthodes d'estimation**
  - Ne nécessite aucune hypothèse sur la forme de la loi : méthode **non paramétrique**
  - **Théoriquement fondée** : les estimateurs bootstrap sont asymptotiquement exacts

- Inconvénients :
  - Peut être très **coûteux** (dépend de la méthode d'estimation)
  - **Validité sujette à caution** pour les petits échantillons
  - **Peu adapté aux valeurs extrêmes** : limité aux bornes de l'échantillon observé
  - Nécessite un **échantillon i.i.d.**
    - Il existe des extensions dans le cas non i.i.d. (bootstrap paramétrique par ex.)

Crédits & Bibliographie

- Tutoriel « Incertitudes », JdS 2011, A. Pasanisi (EDF R&D)
- 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
- Hahn & Meeker, Statistical intervals. A guide for practitionners, Wiley-Interscience, 1991
- Morio and Balesdent, Estimation of rare event probabilities in complex aerospace and other systems, WP, 2016