Treatment of uncertainties in numerical simulation

Bertrand Iooss
Introduction
Starting point: uncertainties everywhere in a modeling chain!

Main problem: credibility of predictions

- Physical model
- Simplifications
- Numerical approximation
- Coding Errors
- Algorithm

- Observations
- Input data
- Variables of interest
- Parameters
- Coding Errors
- Stochastic uncertainties
- Epistemic uncertainties
- Mathematician
- Statistician
- Physicist
- Computer scientist
Similar safety and uncertainty issues in CS&E and Nature sciences

**CS&E : Computational Science & Engineering**

- **Climate Modeling : Prediction**
- **Nuclear industry : Conception, Maintenance, risks**
- **Oil, gas, CO2: Production optimization**
- **Car and plane: Conception**
- **Astrophysics: Understanding**
Accidental scenario of pollutant release

Domain of study: 10 km around an industrial site

2 arbitrary sources (at ground level):
   * source 1: tracer (gas)
   * source 2: iodine (particles)

Projection for 4 days

Meteorological data: wind, temperature, humidity, rain

Rugosity of the ground (vegetation)

[Source: CEA]
Exemple 1: particle dispersion in atmosphere (2/3)

Computation of wind field (direction and amplitude)

Visualization of the wind with flux lines
Exemple 1: particle dispersion in atmosphere (3/3)

Use of a **computer code** of lagrangian particle dispersion (solving the Euler equations of fluid mechanics)

Visualization of gas concentrations en gaz after a 5 hours’ release

Plume under the particular form

Concentration plume (at 10 m level)

Results are strongly sensitive to meteorological data
Exemple 2: Models in hydrology

Uncertainties in model parameters that govern surface and ground water transport, …
Exemple 3: Uncertainties in oil reservoir characterization

• Scalar uncertain parameters:
  – Reservoir Geometry: limits, thickness, faults, etc...
  – Petrophysical properties: porosity, permeability, ...
  – Fluid properties water/Oil/Gaz: contacts between fluids, viscosity, ...
  – Rock/Fluid interactions, Well Data, etc...

• Spatial uncertain parameters:
  – several realizations of a unique geological structure
  – geostatistical parameter ⇒ represented as a "seed variable"
  – Exemples:
    • geostatistical seed
    • Structure maps
    • Stochastic fracture networks

[Source : IFP EN]
Effect of geostatistical uncertainties?

- How to characterize this effect of geostatistics?
Main stakes of uncertainty management

- **Modeling phase:**
  - Improve the model
  - Explore the best as possible different input combinations
  - Identify the predominant inputs and phenomena in order to prioritize R&D

- **Validation phase:**
  - Reduce prediction uncertainties
  - Calibrate the model parameters

- **Practical use of a model:**
  - Safety studies: assess a risk of failure (rare events)
  - Conception studies: optimize system performances and robustness
Uncertainties in simulation experiments

\[ Y = a_1 x_1 + a_2 x_2 \]

**Ancient way**
\[ \Delta Y = a_1 \Delta x_1 + a_2 \Delta x_2 \]

**Pre-modern way**
\[ \text{`s identified to R.V.} \]
\[ \text{... but same algebra} \]
\[ \sigma_Y = \sqrt{a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2} \]

**Really Modern way**
\[ \text{`s fully treated as R.V.} \]

Can give moments, quantiles, and even pdf of \( Y \) ...
...if fair waiting time

Venez’ course 2012 – Treatment of uncertainties

- B. Iloos 12
Which (parametric) uncertainty sources?

• Epistemic uncertainty
  
  – It is related to the lack of knowledge or precision about a parameter which is deterministic in itself (or can be considered deterministic under some accepted hypotheses). E.g. a characteristic of a material.

• Stochastic (or aleatory) uncertainty
  
  – It is related to the real variability of a parameter, which cannot be reduced (e.g. the discharge of a river in flood risk assessment of a riverside area). The parameter is stochastic in itself.

• Reducible vs non-reducible uncertainties
  
  – Epistemic uncertainties are (at least theoretically) reducible
  – Instead, stochastic uncertainties are (in general) irreducible (the discharge of a river will never be predicted with certainty)
A (very) simplified example: flood water level calculation

- $Z_c$: Flood level (variable of interest)
- $Z_m$ et $Z_v$: level of the riverbed, upstream and downstream (random)
- $Q$: river discharge (random)
- $K_s$: Strickler’s roughness coefficient (random)
- $B, L$: Width and length of the river cross section (deterministic)

Strickler’s Formula

\[ Z_c = Z_v + \left( \frac{Q}{K_s \cdot \sqrt{(Z_m - Z_v)/L \cdot B}} \right)^{3/5} \]

General framework

**Input Variables**
- Uncertain: $X$
- Fixed: $d$

**Model**

$G(X,d)$

**Output variables of interest**

$Z = G(X, d)$
Which output variable of interest?

- Formally, we can link the output variable of interest \( Z \) to a number of continuous or discrete uncertain inputs \( X \) through the function \( G \):

\[
Z = G(X, d)
\]

- \( d \) denotes the “fixed” variables of the study, representing, for instance a given scenario. In the following we will simply note:

\[
Z = G(X)
\]

- The output variable of interest can be of dimension 1 or >1
- The function \( G \) can present itself as:
  - an analytical formula or a complex finite element code,
  - with high / low computational costs (measured by its CPU time),

- The uncertain inputs are modeled thanks to a random vector \( X \), composed of \( p \) univariate random variables \((X_1, X_2, ..., X_p)\) linked by a dependence structure.
Methodology
**Step B**: Quantification of uncertainty sources

Modeling with probability distributions

**Input variables**
- Uncertain: x
- Fixed: u

**Model (or measurement process)**

\[ G(x,u) \]

**Variables of interest**

\[ Z = G(x,u) \]

**Step A**: Problem specification

**Quantity of interest**

Ex: variance, probability..

**Step C**: Propagation of uncertainty sources

**Step C’**: Sensitivity analysis, Prioritization

**Step B’**: Quantification of sources

Inverse methods, calibration, assimilation

**Observed variables**

\[ Z_{obs} \]

**Decision criterion**

Ex: Probability < 10^{-b}

**Feedback process**
Step A – Focus on the quantity of interest

What is really interesting in our study?

Ex: in the prior stage of a new product design

Mean, median, variance, (moments) of $Z$

Ex: in the design stage

Quantiles (extrems), probability of treshold exceedence

Ex: in the certification stage

Formally, the quantity of interest is a particular feature of the pdf of $Z$
A particular quantity of interest: the “probability of failure”

- G models a system (or a part of it) in operating conditions
  - Variable of interest $Z$ → a given state variable of the system (e.g. a temperature, a deformation, a water level etc.)

- Following an « operator » point of view
  - The system is in safe operating condition if $Z$ is above (or below) a given “safety” threshold

- System “failure” event: $Z \leq 0$
  - Classical formulation (no loss of generality) in which the threshold is 0 and the system fails when $Z$ is negative
  - Structural Reliability Analysis (SRA) “vision”: Failure if $C - L \leq 0$ (Capacity – Load)

- Failure domain:
  - Problem: estimating the mean of the random variable $\mathcal{D}_f = \{ x \in \mathcal{X} : G(x) \leq 0 \}$

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

$$p_f = \int_{\mathcal{D}_f} f(x)dx = \int_{\mathcal{X}} I_{\mathcal{D}_f}(x) f(x)dx = \mathbb{E} [I_{\mathcal{D}_f}(X)]$$
Step B - Quantification of uncertainty sources

Different cases with respect to available information

1. A lot of data
   - Fitting of probability distributions
   - Statistical hypothesis test (often parametric tests)

2. Few data (n < 10)
   - Hypothesis on parametric probability distribution
   - Non-parametric tests: less powerful, wide bounds
   - Expert judgement, then Bayesian inference

3. No data
   - Expert judgment techniques
   - Maximum entropy principle

\[ H(X) = -\int_X f(x) \log(f(x)) \, dx \]

<table>
<thead>
<tr>
<th>Information</th>
<th>Maximum Entropy pdf</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X \in [a, b] )</td>
<td>Uniform ( X \sim U(a, b) )</td>
</tr>
<tr>
<td>( \mathbb{E}(X) = \mu )</td>
<td>Exponential ( X \sim \mathcal{E}(1/\mu) )</td>
</tr>
<tr>
<td>( X \in [0, \infty[ )</td>
<td>Normal ( X \sim \mathcal{N}(\mu, \sigma) )</td>
</tr>
<tr>
<td>( \mathbb{E}(X) = \mu )</td>
<td></td>
</tr>
<tr>
<td>( \mathbb{V}(X) = \sigma^2 )</td>
<td></td>
</tr>
</tbody>
</table>

Measure of the “vagueness” of the information on \( X \) provided by \( f(x) \)
Some comments (Step B). Dependency

• Taking into account the dependency between inputs is a crucial issue in uncertainty analysis
  – Using copulas structure → CDF of the vector $X$ as a function of the marginal CDF of $X_1 \ldots X_n$:

    $$F(x_1, x_2, \ldots, x_n) = C(F(x_1), F(x_2), \ldots, F(x_n))$$

Example: All bivariate densities here have the same marginal pdf’s (standard Normal) and the same Spearman rank coeff. (0.5)
Step C - Uncertainty propagation: main principles

Propagate uncertainties from $X$ to $Z$, via the deterministic function $G(\cdot)$:

- Conceptually simple problem, but with sometimes a complex implementation
- Choice of method strongly depends on the quantity of interest
  $\Rightarrow$ importance of step A

This quantity of interest is linked to decisional issues

Two kinds of problems:

- Central tendency (ex. mean) or dispersion (variance)
  - Metrology
- High quantile, « probability of failure »
  $\Rightarrow$ justification of a safety criterion

\[ \text{Analytical methods sometimes applicable} \]
\[ \text{Numerical methods (optimization, Monte Carlo sampling)} \]
Step C’ - Sensitivity analysis: main objectives

• Reduction of the uncertainty of the model outputs by prioritization of the sources
  • Variables to be fixed in order to obtain the largest reduction (or a fixed reduction) of the output uncertainty
    A purely mathematical variable ordering
  • Most influent variables in a given output domain
    - if reducibles, then R&D prioritization
    - else, modification of the system
    The individual cost of the reduction may change the previous variable ordering

• Simplification of a model
  • determination of the non-influent variables, that can be fixed without consequences on the output uncertainty
  • building a simplified model, a metamodel
Uncertainties management for cpu time consuming models

A useful solution: the metamodel (model of the numerical model)

\[ p \text{ input variables} \]
\[ X = (X_1, ..., X_p) \]

Physical phenomena
observed experiences

Computer code
simulated experiences
\[ y(X) \]
Time consuming

Metamodell
\[ \hat{Y}(X) \]
Negligible cost
Predicted experiences

Use of the metamodel:

- **C′**: Sensitivity analysis
  - \[ \text{Variance de } \hat{Y} \]
  - Distribution of the inputs
  - Metamodell
  - Distribution of the output
  - \[ Y_{sr} = f_{sr}(X) \]

- **C**: Uncertainty propagation (via Monte Carlo methods)
  - Identification of input parameters values
  - Adequation between observed and simulated experiences

- **B′**: Calibration

Venez' course 2012 – Treatment of uncertainties
- B. Ilooss

24
V&V process: Verification and Validation

To sum up:

**Verification**: do I solve the equations right?

**Validation**: do I solve the right equations?

*(at least for the intended application)*

Two levels for Verification:

1. **Code Verification**: some kind of "internal" correctness of the code may be assessed by formal methods from Software Engineering

2. **Calculation Verification**: concerns the calculations themselves

   Convergence, grid adaptation, solution algorithms, ...

   Is the solution closed to the exact one?

We’ll talk later on subtlties between Code and Model(s)
Some mathematical methods for uncertainty propagation
Quadratic combination method

Data: mean values of $X_i$: $\mu_i = \mathbb{E} [X_i]$ 

variance-covariance matrix of $X_i$:

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

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

Taylor expansion of $G(\bullet)$ around $E(X)$:

$$G(X) = G(\mu) + \sum_{i=1}^{N} \frac{\partial G}{\partial X_i} \bigg|_{X=\mu} (X_i - \mu_i)$$

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

En général, dans les applications le développement est d’ordre 1
Quadratic combination method – First order

Mean of Z

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

Variance of Z

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

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

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

Remarks:
++ Needs only mean and covariance of X
-- Do not use if G(.) is strongly non linear
-- Provides only mean and variance of Z => no extrapolation for the distrib. law of Z
++ if X is gaussian and G(.) is linear, then Z is gaussian
Quadratic combination method – Independent case

If the $X_i$'s are independent:

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

Quadratic summation formula

Contribution of each input variable to the uncertainty of the output variable

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

Sensitivity indices (normed)
Methods of Monte Carlo simulation

- General methods to evaluate a numerical quantity, using some random simulations

- In uncertainty propagation: use a random sample of \( G(X) \) to evaluate the quantity of interest

- We suppose that we know how to simulate an i.i.d (independent and identically distributed) sample of \( X_i \) following its probability distribution \( f_i \)
Monte Carlo in general (1/3)

• Computation of the integral:

\[ I = \int_{\mathcal{X}} h(x) f(x) \, dx \]

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

Random sample of X

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

\[ \hat{I} \rightarrow I \]

Monte Carlo estimator

h(\bullet) : deterministic function
X : r.v. with density f(x)

From the law of large numbers, the Monte Carlo estimator converges (a.s.) to the true quantity
Monte Carlo (2/3)

- Variance of the Monte Carlo estimator

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

- Variance of \( h(X) \) is given via its estimator:

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

- General expression for the variance of the MC estimator

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

- We note:

\[
\sigma_i^2 = \mathbb{V} \left[ \hat{I} \right]
\]
Monte Carlo (3/3)

- Asymptotic law of the estimator from Central Limit Theorem:

\[
\frac{\sqrt{n}}{\sigma_{h(X)}} \left( \hat{I} - I \right) \sim \mathcal{N}(0, 1) \quad \text{with} \quad \sigma_{h(X)} = \sqrt{\mathbb{V}[h(X)]}
\]

Low convergence speed (in \(1/\sqrt{n}\)) but:

Independence with respect to the dimension of \(X\) and to the form of \(h(\cdot)\)

Unbiased estimator

Precision only depends on \(n\) (then on the cpu time of \(h(.)\))

- Confidence intervals for the Monte Carlo error:

\[
\epsilon_n = \hat{I} - I
\]

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

Monte Carlo error

Confidence interval with probability \(\alpha\)

Quantiles of the standard normal law

\[
\sigma_{\hat{I}} = \frac{\sigma_{h(X)}}{\sqrt{n}}
\]
Monte Carlo and uncertainty propagation

- Propagation of the uncertainties of $X$ to $Z = G(X)$

$n$-sample of $X$

- Monte Carlo estimator of mean and variance of $Z$:

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

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

- Moments of $Z$ are estimated by the empirical moments
Estimation of a probability of failure

• System failure : event \( Z < 0 \)

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

• Failure probability: \( p_f = \int_{\mathcal{D}_f} f(x)dx = \int_{\mathcal{X}} I_{\mathcal{D}_f}(x) f(x)dx = \mathbb{E}[I_{\mathcal{D}_f}(X)] \)

  - Problem : computation of the mean of the random variable \( I_{\mathcal{D}_f}(x) \)

• Failure indicator : \( I_{\mathcal{D}_f}(x) = 1_{\{G(x) \leq 0\}} \)
Monte Carlo estimation of $p_f$ (1/3)

- Naive Monte Carlo estimator:

$$\hat{p}_f = \frac{1}{n} \sum_{i=1}^{n} I_{D_f}(x^{(i)})$$

- Variance of the estimator:

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

- As $I_{D_f}(X^{(1)})$, $I_{D_f}(X^{(2)})$, \ldots $I_{D_f}(X^{(n)}) \sim \mathcal{B}(p_f)$ Bernouilli i.i.d.

- We have:

$$\mathbb{V}[\hat{p}_f] = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{V}[I_{D_f}(x)] = \frac{1}{n^2} n \ p_f(1-p_f)$$

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

- Asymptotical convergence to a normal law and other MC estimator properties

Venez’ course  2012 – Treatment of uncertainties

- B. Iooss

36
Monte Carlo estimation of $p_f$ (2/3)

- Decrease in square root of $n$:
  \[ \sigma_{\hat{p}_f} = \frac{1}{\sqrt{n}} \sqrt{p_f(1-p_f)} \]

- Variation coefficient:
  \[ cv = \frac{\sigma_{\hat{p}_f}}{\mathbb{E}[\hat{p}_f]} = \sqrt{\frac{p_f(1-p_f)}{n}} \frac{1}{p_f^2} = \sqrt{\frac{1-p_f}{n \ p_f}} \]

- For small values of $p_f$:
  \[ p_f \to 0 \implies \frac{1-p_f}{p_f} \to \frac{1}{p_f} \]
  \[ cv \approx \sqrt{\frac{1}{n \ p_f}} \]

- For example, if we estimate a proba $p_f = 10^{-r}$ with $cv = 10\%$, \[ \sqrt{n \cdot 10^{-r}} = 10^{-1} \implies n = 10^{r+2} \]

$\Rightarrow$ prohibitive required cpu times

$\Rightarrow$ Use of improved methods: approximate methods (FORM/SORM), accelerated Monte Carlo methods, metamodel-based methods, …
**Importance sampling (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 :

\[
\hat{I}_{is} = \frac{1}{n} \sum_{i=1}^{n} h(x^{(i)})w(x^{(i)})
\]

\[
\mathbb{V} \left[ \hat{I}_{is} \right] = \frac{1}{n} \mathbb{V} \left[ h(X)\frac{f(x)}{\varphi(x)} \right]
\]

- Estimateur sans biais de $I$, à condition que le support de $\varphi(x)$ contienne celui de $f(x)$
Importance sampling (2/3)

• Cette méthode ne garantit pas une réduction de la variance \( \forall \phi(x) \)
• Le choix de la « loi instrumentale » \( \phi(x) \) est crucial
  - Théoriquement : densité optimale \( \varphi^*(x) = \frac{|h(x)| f(x)}{\int_X |h(x)| f(x) dx} \)
  - 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 : \( h(x) = I_{D_f}(x) = 1_{\{G(x) \leq 0\}} \)
  - Densité optimale : \( \varphi^*(x) = \frac{I_{D_f}(x) f(x)}{\int_X I_{D_f}(x) f(x) dx} = \frac{I_{D_f}(x) f(x)}{p_f} \)
Importance sampling (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
Quantile estimation (1/2)

- Probability distribution function and quantile estimator

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

\[ \hat{F}_n(z) \to F(z) \]

Monte Carlo estimator of a p-quantile: empirical quantile

\[ \hat{z}_p = \inf \left( z : \hat{F}_n(z) \geq p \right) \]
Quantile estimation (2/2)

• In practice:
  – Build an ordered sample from \( G(x^{(1)}) , G(x^{(2)}) , \ldots , G(x^{(n)}) \)
  – Call it: \( z^{(1)} , z^{(2)} , \ldots , z^{(n)} \quad z^{(1)} \leq z^{(2)} \leq \ldots \leq z^{(n)} \)
  – \( \hat{z}_p = z^\lceil np \rceil \)
    • For example, if \( n=100 \) and \( p=0.95 \), then we have to take 96\textsuperscript{th} value in the ordered sample
    • Of course, we need \( \frac{1}{N} < p < 1 - \frac{1}{N} \)

• Asymptotic law of the estimator:
  \[
  \frac{\sqrt{n}}{\tau} (\hat{z}_p - z_p) \sim \mathcal{N}(0, 1) \quad \tau^2 = \frac{p(1-p)}{(f(z_p))^2}
  \]
Quantile estimation – Wilks formula

• We can show that:

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

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

– Then, if \( r \) is the smallest integer such that \( C_p(n, r) \leq 1 - \beta \) \( \implies \) \( 1 - C_p(n, r) \geq \beta \)

– then, \( \mathbb{P} \left( z^{(np+r)} > z_p \right) \geq \beta \)

• We obtain the Wilks method

– Conservative estimator of quantiles:
  • With a fixed \( n \), find \( \beta \) (the confidence level of the quantile)
  • With a fixed \( \beta \) fixé, find \( n \) (required number of code runs)
Sampling via Wilks formula

Comments:
- Robust method
- No hypothesis on the distribution function

Constraint:
- Can only be applied to pure random sample (i.i.d.)

Example with Wilks at first order and unilateral quantile
$Z_{\text{max}}$ is the maximal value of the N-sample (i.i.d) of $Z$

$$P[P(Z \leq Z_{\text{max}}) \geq \alpha] \geq \beta, \quad N \text{ solution of } 1 - \alpha^N \geq \beta$$

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>0.50</th>
<th>0.90</th>
<th>0.90</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>0.95</td>
<td>0.90</td>
<td>0.95</td>
<td>0.90</td>
</tr>
<tr>
<td>$n$</td>
<td>5</td>
<td>22</td>
<td>29</td>
<td>45</td>
</tr>
</tbody>
</table>
Application example
Example: pollution measures on a few points of a room wall

21 activity measures in $^{137}\text{Cs} \text{ (Bq/cm}^2\text{)}$

- Mean = 31.45
- Median = 15.4
- Standard-deviation = 36.11
- Min = 0.83 – Max = 156.67
- Skewness = 2.02
- Kurtosis = 4.19

**Safety issue:** guarantee (with a certain confidence level) that the contamination does not exceed a threshold over all the room wall

**Examples:** prediction of the amount of different category of wastes
(proportion of activities < 50 Bq / cm$^2$, > 100 Bq / cm$^2$, ...)

- waste quantities in different types of storage (deep geologic, subsurface, no storage)
- different costs
Useful probabilistic tools: universal inequalities

For a random variable $X$ with mean $\mu$ and variance $\sigma^2$, we can use for $X > \mu$:

- the Bienaymé-Tchebytcheff inequality: $P(X \leq \mu + k\sigma) > \frac{k^2}{1 + k^2}$

More than 72% of the surface < 100 Bq/cm²

Pessimistic bound

$\mu$ and $\sigma^2$ are replaced by their empirical estimates

- the Guttman inequality:

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

More than 82% of the surface < 100 Bq/cm²

Needs the knowledge of the kurtosis

- the Meidell inequality (unimodality hypothesis):

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

More than 89% of the surface < 100 Bq/cm²

All these tools give unsafe estimates
Using the Wilks formula

For an i.i.d. sample \(\{X_1, \ldots, X_n\}\) of a random variable \(X\), if \(n\) is solution of \(1-\alpha^n \geq \beta\) and \(X_{\max} = \max\{X_1, \ldots, X_n\}\) we have

\[
P\left[P(X \leq X_{\max} | (X_1, \ldots, X_n)) \geq \alpha\right] \geq \beta
\]

It gives:
1. the minimal sample size \(n\) for \(\alpha\) and \(\beta\)
2. for a given sample, the \(\alpha\)-quantile value, with a \(\beta\) confidence degree

<table>
<thead>
<tr>
<th>(\alpha)</th>
<th>0.50</th>
<th>0.90</th>
<th>0.95</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta)</td>
<td>0.95</td>
<td>0.95</td>
<td>0.90</td>
<td>0.95</td>
</tr>
<tr>
<td>(n)</td>
<td>5</td>
<td>29</td>
<td>45</td>
<td>59</td>
</tr>
</tbody>
</table>

No hypothesis on distribution function and no needs of parameter estimates

More general result linking \(n\) and order \(r\) (rank in the ordered sample \(\{X_1, \ldots, X_n\}\))

Application (measures in \(^{137}\text{Cs}\)):

- Wilks \((n=21, r=2, \beta=0.9)\) -> more than 83% of the surface < 80 Bq/cm² (with a 90% degree of confidence)

- Meidell (unimod., \(\sigma\) estimate) -> more than 80% of the surface < 80 Bq/cm²
Conclusions on step C (uncertainty propagation)

• **Challenge:** balance between precision of the estimate and cpu time cost

• Use **Monte Carlo** if possible: independent of input dimension, unbiased estimation, gives a confidence interval
  **BUT:** needs large number of model runs to obtain convergence

• If this cost is unreachable, alternative methods exist:
  – Accelerated Monte Carlo method (importance sampling, etc.)
  – Méthodes quasi-Monte Carlo (cf. cours 2) - **But:** curse of dimensionality
  – **Approximate methods:**
    • Quadratique summation - **But:** linear hypothesis
    • FORM/SORM : fast estimation of $p_f$. Can be used to initialize another method (importance sampling)

  – Using a surrogate model of the computer code (metamodel)
Bibliography


- De Rocquigny, Devictor & Tarantola (eds), Uncertainty in industrial practice, Wiley, 2008

- M. Lemaire, A. Chateauneuf, and J. Mitteau, Structural reliability, Wiley Online Library, 2009


- Guide to the expression of Uncertainty in Measurements (GUM), ISO publication