Propagation of uncertainties through numerical models

Fabrice Gamboa
Bertrand Iooss
A few words about my company: EDF = Électricité de France.

220 hydraulic dams - 58 nuclear power plants - Solar energy - Wind power - …
Uncertainty problems in energy production systems

Many uncertainties for the energy production and the safety due to:
- hazards (demand, weather, …),
- incomplete system knowledge (ageing, physics, …),
- internal aggressions (failures, …)
- external aggressions (earthquake, …)

In order to better understand, prove the safety and optimize its industrial processes, EDF R&D develops some physical numerical simulation codes.
Large database of component failures

CLASSICAL STATISTICS

Small database of system failures

Expert Judgment

BAYESIAN STATISTICS

• No observed failure
• Physical model

SIMULATION UNCERTAINTY

Examples:
… Chemical accident scenario
… Dam failure
… Waste repository
Introduction to uncertainty management
Starting point: Uncertainties everywhere in a modeling chain!

Main problem: credibility of predictions

- Simplifications
  - Model uncertainties
- Numerical approximation
  - Numerical uncertainties
- Coding Errors
- Parameters
  - Epsitemic uncertainties
- Variables of interest
  - Stochastic uncertainties
- Input data
- Algorithm

Real observations

- Variables of interest
  - statistician
- Numerical model
  - Computer scientist

- Physical phenomenon
  - physicist

- Numerical model
  - mathematician

- Physicist

 physicien

 Statistician

 Mathematician

 Computer Scientist
Example 1: particle dispersion in atmosphere (1/3)

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)

Topography and location of sources

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

Computation of wind field (direction and amplitude)

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

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

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

Results are strongly sensitive to meteorological data
Example 2: Seismic fragility curve (industrial facilities protection to earthquakes)

Study on a mock-up (SMART)

Building a numerical model (finite-element based structural mechanics)
Validation on experiments

Computing the seismic fragility curves:
Proba (failure | PGA)

PGA = Peak ground Acceleration

- F. Gamboa & B. Iooss
Example 3: Optimization of photovoltaic energy production

Time consuming computer code (electrical model) with:

- 6 hours per model run
- 3 « controllable » input variables:
  - elevation (0.4m to 1m),
  - inclination (0° to 50°),
  - distance inter-stands (2.5m to 10m)
- 2 uncertain variables: meteo and albedo
- In output: the production (kWh) of a stand
Similar safety and uncertainty issues in Computational Science & Engineering and Nature sciences

Car and plane: Conception

Climate Modeling: Prediction

Nuclear industry: Conception, Maintenance, risks

Oil, gas, CO2: Production optimization

Astrophysics: Understanding
Summary: 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 (topics of rare events inference)
  – **Conception studies**: **optimize** system performances and robustness
Uncertainty management - The generic methodology

Step A: Problem specification
- Input variables
  Uncertain: x
  Fixed: u
- Model
  (or measurement process)
  \( G(x, u) \)
- Variables of interest
  \( Z = G(x, u) \)
- Quantity of interest
  Ex: variance, probability...

Step B: Quantification of uncertainty sources
- Modeling with probability distributions
- Direct methods, statistics, expertise

Step B': Quantification of sources
- Inverse methods, calibration, assimilation

Step C': Sensitivity analysis, Prioritization

Step C: Propagation of uncertainty sources

Feedback process

Decision criterion
Ex: Probability < 10\(^{-b}\)
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

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

Measure of the “vagueness” of the information on X provided by f(x)

<table>
<thead>
<tr>
<th>Information</th>
<th>Maximum Entropy pdf</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X \in [a, b] )</td>
<td>Uniform</td>
</tr>
<tr>
<td>( X \sim U(a, b) )</td>
<td></td>
</tr>
<tr>
<td>( \mathbb{E}(X) = \mu ) \quad X \in [0, \infty] )</td>
<td>Exponential</td>
</tr>
<tr>
<td>( X \sim \mathcal{E}(1/\mu) )</td>
<td></td>
</tr>
<tr>
<td>( \mathbb{E}(X) = \mu ) \quad \sigma^2 )</td>
<td>Normal</td>
</tr>
<tr>
<td>( X \sim \mathcal{N}(\mu, \sigma^2) )</td>
<td></td>
</tr>
</tbody>
</table>
Some mathematical methods for uncertainty propagation
Step A – Focus on the quantity of interest

**Inputs:** $X$

**Output:** $Z$

What is really interesting in our study?

- Mean, median, variance, (moments) of $Z$
  - Ex: in the design stage
- Quantiles (extremes), probability of threshold exceedance
  - Ex: in the certification stage

Formally, the quantity of interest is a particular feature of the pdf of $Z$
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
  => 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 »
  \to justification of a safety criterion

[Analytical methods sometimes applicable]

[Numerical methods (optimization, Monte Carlo sampling)]
Quadratic combination method - Introduction

- Analytical and simple practical method
- Based on 2 elementary results of probability theory

- $X_1, \ldots, X_p$ : random variables
- $a_1, \ldots, a_p$ : real numbers

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

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

- These formula give the mean and variance of $Z=G(X)$ if $G$ is a linear model

- ... it gives the idea to locally linearize the model $G$ by a Taylor decomposition
Quadratic combination method

Data :

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

variance-covariance matrix of $X_i :$

\[
C_{ov} [X_i, X_j] = \mathbb{E} [(X_i - \mu_i) (X_j - \mu_j)]
\]
\[
\rho_{ij} = \mathbb{E} \left[ \frac{X_i - \mu_i}{\sigma_i} \frac{X_j - \mu_j}{\sigma_j} \right]
\]

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

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

In general, we use the 1st order development in the applications
Quadratic combination method – First order

Mean of $Z$

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

Variance of $Z$

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

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

$$\text{Var}(Z) = \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial G}{\partial X_i} \bigg|_{X=\mu} \frac{\partial G}{\partial X_j} \bigg|_{X=\mu} \rho_{ij} \sigma_i \sigma_j$$

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 pdf of $Z$

++ if $X$ is Gaussian and $G(.)$ is linear, then pdf of $Z$ is Gaussian
Quadratic combination method – Independent case

If the $X_i$ s are independent:

$$\text{Var}(Z) = \sum_{i=1}^{p} \left( \frac{\partial G}{\partial X_j} \right)_{X=\mu}^2 \sigma_i^2$$

Quadratic summation formula

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

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

Sensitivity indices (normed)

Sensitivity analysis is directly obtained
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:

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

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

\[ x^{(1)}, x^{(2)}, ..., x^{(n)} \]

Random sample of X

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

Monte Carlo estimator

\[ \hat{I} \to I \]

h(•) : 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 estimator
Monte Carlo (2/3)

Variance of the Monte Carlo (MC) estimator

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

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

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

- General expression for the variance of the MC estimator:

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

- We note:

\[
\sigma_{\hat{I}}^2 = \operatorname{Var} [\hat{I}]
\]
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:

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

Quantiles of the standard normal law

Monte Carlo error

Confidence interval with probability \(\alpha\)

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

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

$$x^{(1)}, x^{(2)}, ..., x^{(n)} \quad \text{n-sample of } X$$

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

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

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

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

- System failure: event  $Z \leq 0$

- 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:

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

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

- We have:

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

Estimated by:

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

- Asymptotical convergence to a normal law and other MC estimator properties
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:
  \[
  cv = \frac{\sigma_{\hat{p}_f}}{E[\hat{p}_f]} = \sqrt{\frac{p_f(1 - p_f)}{n}} \frac{1}{p_f^2} = \sqrt{\frac{1 - p_f}{n \ p_f}}
  \]

- For small values of $p_f$:
  \[
  p_f \rightarrow 0 \quad \Rightarrow \quad \frac{1 - p_f}{p_f} \rightarrow \frac{1}{p_f}
  \]
  \[
  cv \approx \sqrt{\frac{1}{n \ p_f}}
  \]
  « Relative error », precision of the estimate

- For example, if we estimate a proba $p_f = 10^{-r}$ with $cv = 10\%$,
  \[
  \sqrt{\frac{1}{n \ 10^{-r}}} = 10^{-1} \quad \Rightarrow \quad n = 10^{r+2}
  \]
  \[
  10^{r+2} \text{ values of } G(X), \text{ then } 10^{r+2} \text{ calls to the code } G
  \]

⇒ prohibitive in terms of required cpu times
⇒ use of improved methods: approximate methods (FORM/SORM), accelerated Monte Carlo methods, metamodel-based methods, …
Importance sampling (1/3)

• Idea: modifying the sampling prob. distribution of $X$ in order to concentrate the samples in most interesting regions (in terms of contribution to computation of expectation of $h(X)$)

$$I = \int_X h(x) f(x) \, dx = \int_X h(x) \frac{f(x)}{\varphi(x)} \varphi(x) \, dx = \int_X h(x) w(x) \varphi(x) \, dx$$

• It is the expectation of the function $h(x) w(x)$, $X \sim \varphi(x) \, dx$

1) Produce a sample $(x^{(i)})$ from density $\varphi(x) \, dx$

2) Then, compute:

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

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

• Unbiased estimator of $I$, in condition that the support of $\varphi(x)$ contains the one of $f(x)$
Importance sampling (2/3)

• This method does not guarantee any variance reduction $\forall \varphi(x)$
• The choice of the « instrumental law » $\varphi(x)$ is crucial
  
  – theoretically: optimal density: $\varphi^*(x) = \frac{|h(x)| f(x)}{\int_X |h(x)| f(x) dx}$

  – The normalization constant is as difficult to evaluate as evaluating $I !$
  – However, practical result …

• Estimation of a failure probability $p_f$ by importance sampling

  – Here: $h(x) = I_{D_f}(x) = 1_{\{G(x) \leq 0\}}$

  – Optimal density: $\varphi^*(x) = \frac{I_{D_f}(x) f(x)}{\int_X I_{D_f}(x) f(x) dx} = \frac{I_{D_f}(x) f(x)}{p_f}$
Importance sampling (3/3)

- The optimal density is the conditional law of $X$ knowing that $X \in D_f$
- Intuitive result $\rightarrow$ the method is mostly efficient if it produces samples in the failure domain

- Some practical algorithms:
  - Obtain a first idea of the configuration of $D_f$ (first Monte Carlo runs)
  - Center the instrumental law on a point of $D_f$ (for example on the design point $P^*$ obtained with the FORM method)
FORM method (1/5)

**FORM**: First Order Reliability Method

3 steps:

1. Transformation of inputs $X_i$ to other inputs whose probability distributions have « good properties »
   Isoprobabilistic transformation $\rightarrow$ standard Gaussian space

2. Search of the most probable failure conditions

3. Estimation of the failure probability
FORM (2/5)
Isoprobabilistic transformation

\[ U = \mathcal{T}(X) \]

Points which mostly contribute to \( p_f \) are the nearest to the origin in the standard space.

Each component of \( U \) follows a \( N(0,1) \)

Component of \( U \) are independent

The iso-prob. surfaces are spheres

\[
\phi_d (\mathbf{u}) = \frac{1}{(\sqrt{2\pi})^d} \exp \left( -\frac{1}{2} \sum_{i=1}^{d} u_i^2 \right)
\]
FORM : isoprobabilistic transformation

• Rosenblatt transformation

\[ T : \quad u_1 = \Phi^{-1}(F_1(z_1)) \]
\[ u_2 = \Phi^{-1}(F_2(z_2|z_1)) \]
\[ \vdots \]
\[ u_N = \Phi^{-1}(F_N(z_N|z_{N-1}, \ldots, z_2, z_1)) \]
FORM (3/5) – Isoprobabilistic transformation

• New expression of the failure probability

\[ \mathbb{P}(G(X) \leq 0) = \mathbb{P}(G(T^{-1}(U) \leq 0) = \mathbb{P}(g(U) \leq 0) \]

\[ D_f \]

\[ D'_f \]

Physical space

« Standard » space

- Expression of the failure probability:

\[ p_f = \int_{T(X)} 1_{g(u) \leq 0} f_u(u) \, du \]
FORM (4/5) – Search of the most probable failure conditions

• To each point of the standard space, some operating or failure conditions are associated

  – The most probable failure point is the nearest from the origin (where the density is max because the mean of \( U \) is the null vector)

  – Reminder : the value of density \( f_U(u) \) depends only on \( ||u|| \) (distance of \( u \) from origin)

  – Call it \( P^* \) : design point
  – Call \( u^* \) the vector \( OP^* \)

  – The search of \( u^* \), with an uniqueness hypothesis, is an optimization problem under constraints

\[
 u^* = \min_{g(u)=0} \beta(u) = \min_{g(u)=0} \sqrt{u^t u}
\]
FORM (5/5) – Evaluation of $p_f$

- **Hypothesis**: 
  - Replacement of the limit surface $g(u)=0$ by the hyperplan intersecting $P^*$ and orthogonal to $u^*$, with equation:

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

  $\cdot \alpha_i$ : projection of $u^*$ on $U_i$

  $\beta$ : norm of $u^*$

- Approximation based on the hyp. that points far away from $P^*$ have small contributions to $p_f \rightarrow$ their proba is very small

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

$\beta$ : « Reliability index »

$\alpha_i$ : « Importance factors » FORM $\rightarrow$ sensitivity indices of variables $U_i$ to $p_f$

Linear combination of r.v. $N(0,1)$ with normed coeff. $\alpha_i$  
$\Rightarrow N(0,1)$

Distribution function of $N(0,1)$
FORM/SORM : Pros and cons

**Pros:**
- **Reduced computing times** with respect to other methods
- No dependency between computing times and value of $p_f$
- Getting the importance factors and a design point

**Cons:**
- Approximation not always valid
- No measure of the error which is made
- G has to be *differentiable*
- Hypothesis of a unique design point
Quantile estimation (1/2)

$p$-quantile definition:  \( q_p = \inf \{ z : P(Z \leq z) \geq p \} \)

Probability distribution function and quantile estimators

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

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

Glivenko – Cantelli theorem: convergence to \( F(z) \)

Erpirical CDF

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 \( G(x^{(1)}), G(x^{(2)}), ..., G(x^{(n)}) \)
  – Call it: \( \hat{z}^{(1)}, \hat{z}^{(2)}, ..., \hat{z}^{(n)} \) \( \hat{z}^{(1)} \leq \hat{z}^{(2)} \leq ... \leq \hat{z}^{(n)} \)
  – \( \hat{z}_p = \hat{z}([np]) \)
    • For example, if \( n=100 \) and \( p=0.95 \), then we have to take 96\(^{th}\) value in the ordered sample
    • Of course, we need \( \frac{1}{N} < p < 1 - \frac{1}{N} \)

• Asymptotic law of the estimator:

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

$$P[P(Z \leq Z_{\text{max}}) \geq \alpha] \geq \beta,$$  \( N \) 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 $^{137}$Cs (Bq/cm$^2$)

- 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-Tchebycheff 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:

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

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 \( X_{\text{max}} = \max\{X_1, \ldots, X_n\} \) we have

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

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 \( (n=21, \ r=2, \ \beta=0.9) \) \( \rightarrow \) more than 83\% of the surface < 80 Bq/cm\(^2\) (with a 90\% degree of confidence)

- Meidell (unimod., \( \sigma \) estimate) \( \rightarrow \) more than 80\% of the surface < 80 Bq/cm\(^2\)
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


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