# Quantificazione delle incertezze nella simulazione Monte Carlo

P. Saracco, M. Batic, M.G. Pia

# Uncertainty quantification in generic Monte Carlo Simulation: a mathematical framework

## How to do it?

**Abstract:**
Uncertainty Quantification (UQ) is the capability of predicting the uncertainty of experimental observables produced by Monte Carlo particle transport, deriving from uncertainties in the physics modeling components (such as cross sections, atomic and nuclear parameters, geometrical description of the experimental apparatus and so on) used in the simulation.

We establish a general mathematical framework for UQ: in the case of a single input uncertainty the problem is analytically soluble, while the case of many uncertainties requires the additional hypothesis of statistical independence and involves some predictable approximations.

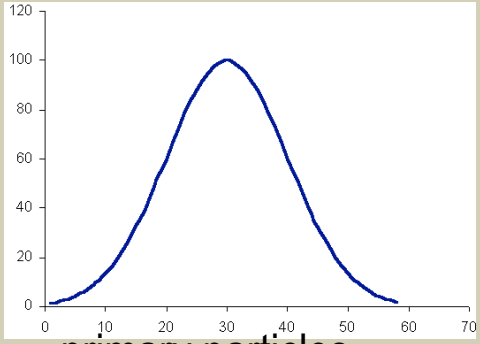
# Monte Carlo in HEP

cross sections,
branching ratios,
physics models,
physics parameters..

input

**MC**
**engine**

observable




**Event generator**
*(Pythia, Herwig…)*
**Particle transport**
*(Geant4, MCNP, MARS…)*

primary particles,
deposited energy…

How much can we trust the observables produced by MC?

A simulation engine is needed when a given physical model is too much complicated to be analytically solved

Physical model

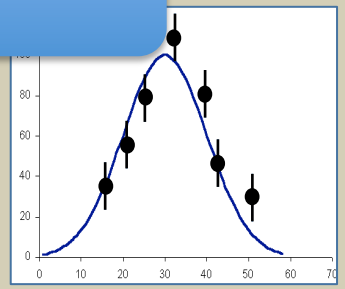
(Transport equation)

Geometry
Composition
…

User's data

Physical data

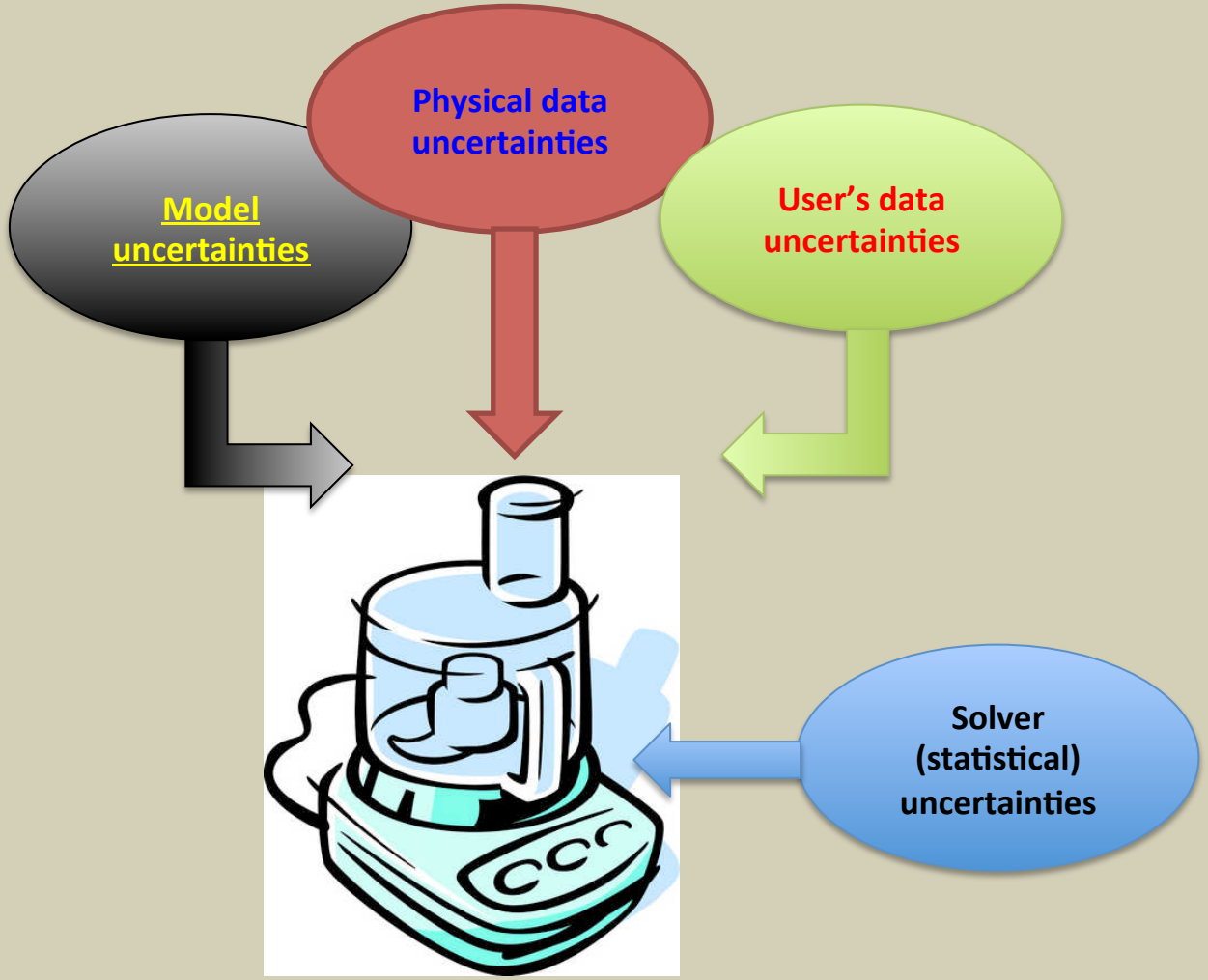
Cross sections
Models
…




SOLVER
Simulation
Engine
(MC)

Output for some observable + solver errors (statistical for MC)

… but in a non idealized situation there are many sources of uncertainties




**Result of the simulation are affected by a complex mix of various sources of uncertainties**

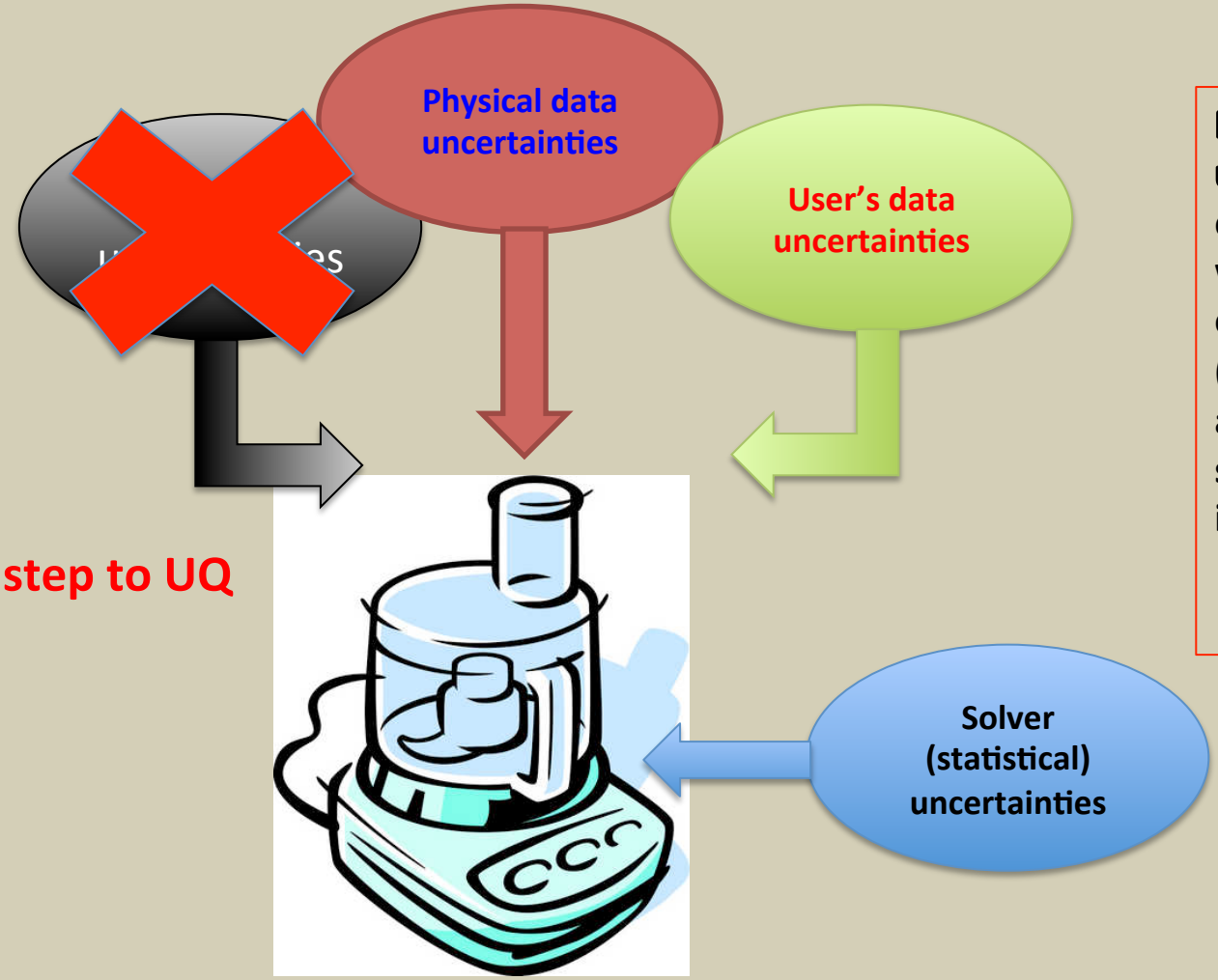
## Tentative classification of uncertainties (possibly non exhaustive)

- **PARAMETER UNCERTAINTY**, when some of the computer code inputs are unknown, or known with errors
- **MODEL INADEQUACY**, which may derive from **STRUCTURAL UNCERTAINTY** (for example approximations in the physical model) or **ALGORITHMIC UNCERTAINTY** (deriving from the numerical methods employed to solve the model)
- **RESIDUAL VARIABILITY**, when the process itself is inherently unpredictable or stochastic
- **PARAMETRIC VARIABILITY**, when some of the inputs are **INTENTIONALLY** uncontrolled or left unspecified (backward problems, robust design, …)

**Parameter uncertainty plays a major role in MC simulations of particle transport because in practice all the physical input to the simulation is affected by experimental or theoretical uncertainties**.

**Algorithmic uncertainty**, in this acceptation, is (mainly) of statistical origin for Monte Carlo simulations.

… but in a non idealized situation there are many sources of uncertainties

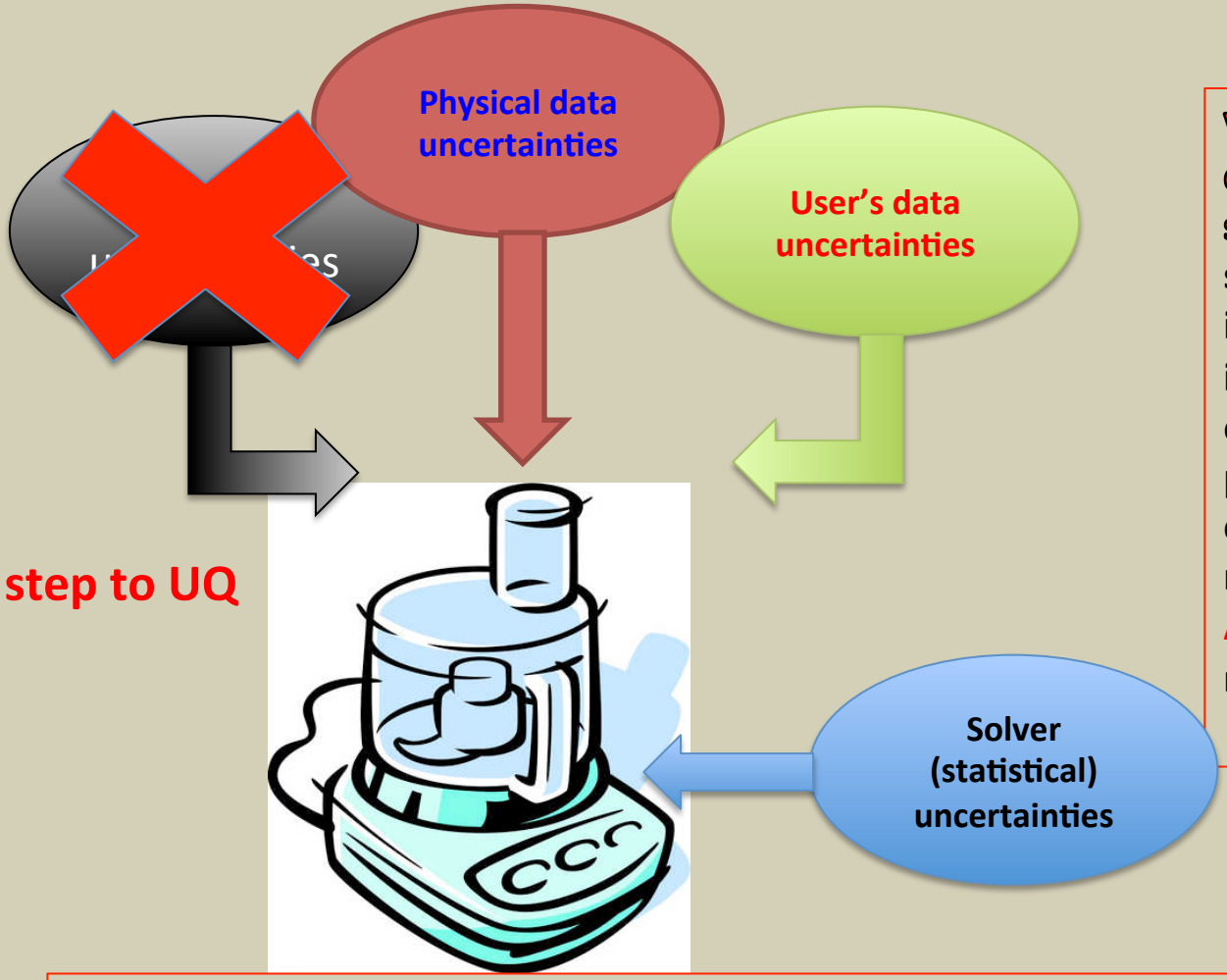



**1st step to UQ**

**Neglect model uncertainties: transport equation is the "true world" and there are no other approximations (like condensed path approximation) in its solution other than those induced by MC sampling**

**Result of the simulation are affected by a complex mix of various sources of uncertainties**

… but in a non idealized situation there are many sources of uncertainties

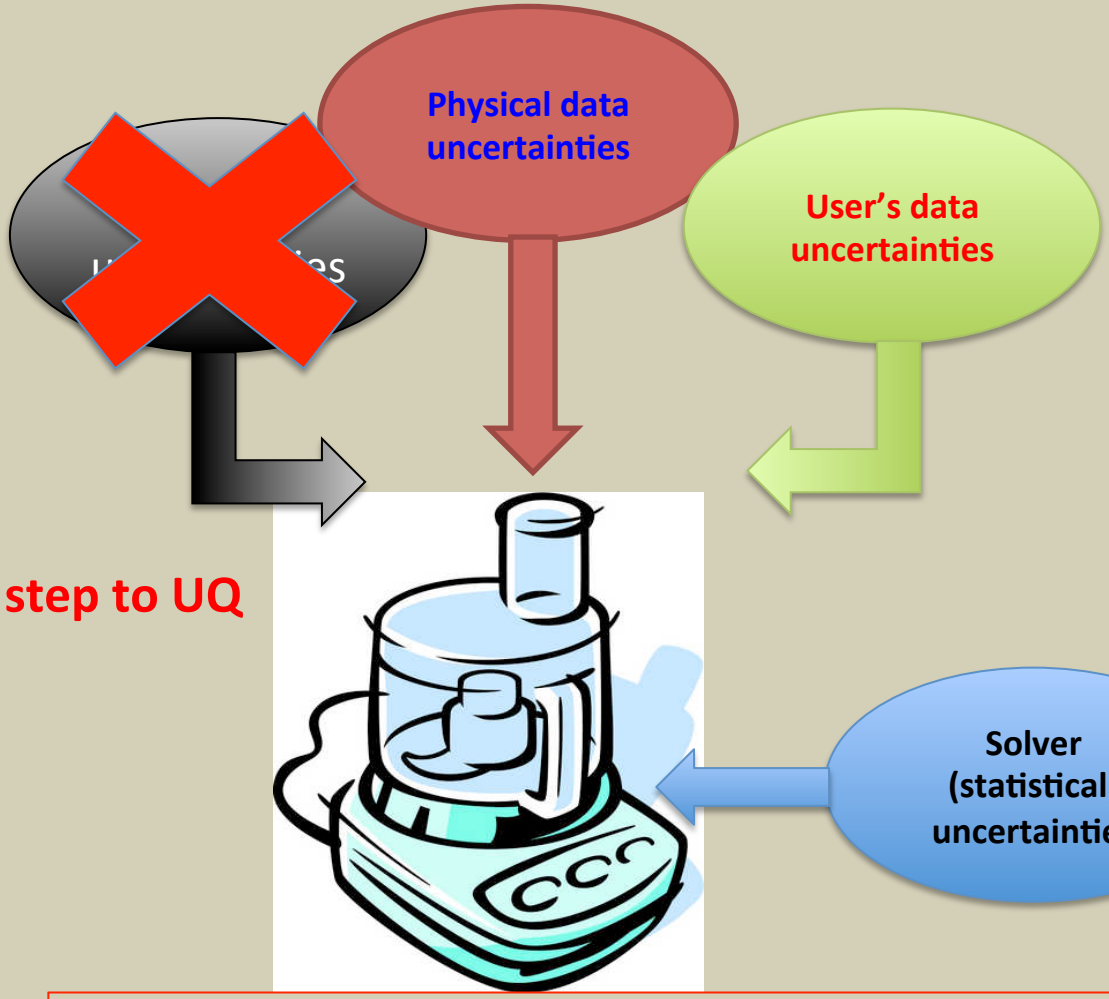



**2nd step to UQ**

**Validation of physical data needed for simulation: MC users should have accurate infos about confidence intervals (better PDFs) for each of the physical parameters used by the code (cross sections, models employed, …) AND the possibility to make variations**

**Because depending on the experimental configuration small variations in some parameter may result in large variations of the output or large variations in some other parameter may have no practical effect at all**

… but in a non idealized situation there are many sources of uncertainties



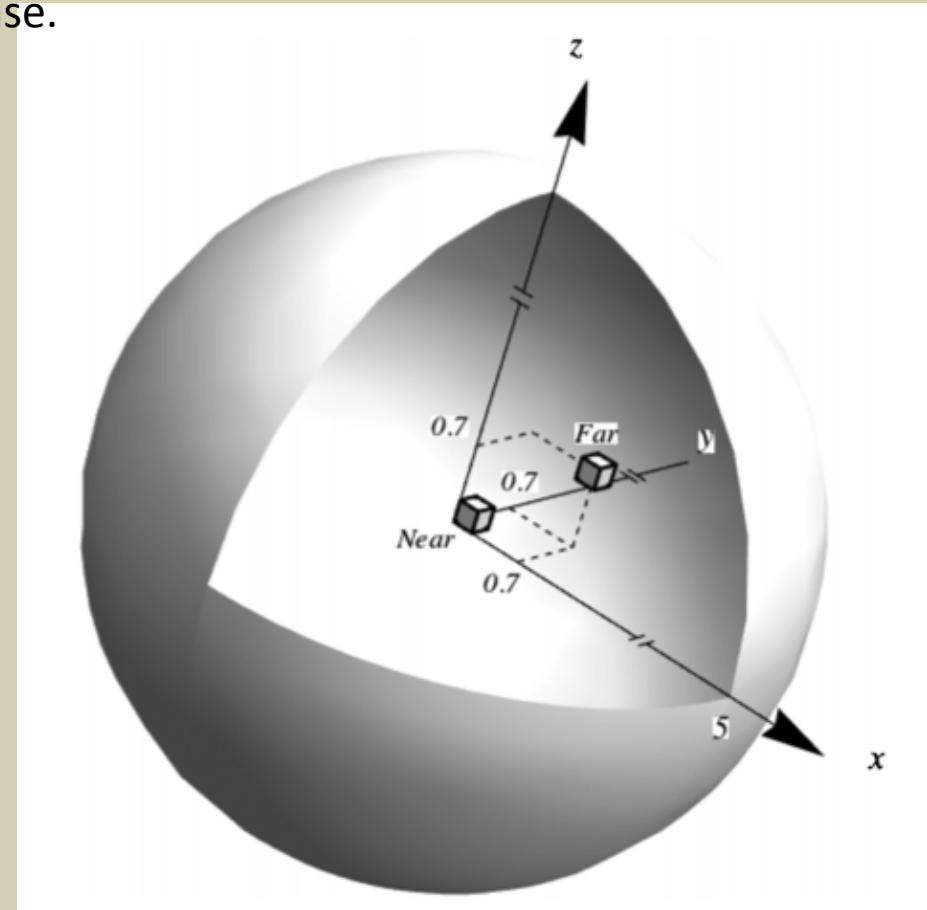

**3nd step to UQ**

**Accurate study of the experimental configuration: MC users should select a proper restricted set of parameters (physical and/ or user controlled) on which perform an UQ attempt**

**Because as we shall see it is out of human possibilities (and possibly meaningless) to perform UQ over possibly hundreds of different parameters**

**We expect that result of the simulation depends both on the value of the input unknown(s), on the position of the detector and on the sample dimension**

**"TOY MC"**: a (very) simplified "transport code", a random path generator ruled by two constant parameters describing the relative probability of absorption ($\Sigma_A$) and scattering processes ($\Sigma_S$), sampling an observable – track-length in this case.



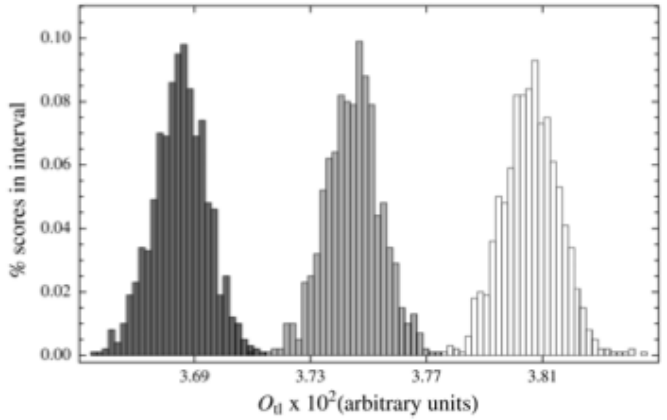




Fig. 4. Results for 1000 of Monte Carlo simulations for the observable $O_{tl}$, each encompassing $10^6$ events, for an observable scored close to the primary particle source (see text), produced with different values of the $\Sigma_S$ input physical parameter: $\Sigma_S = 1$ (white histogram), $\Sigma_S = 1.1$ (grey histogram) and $\Sigma_S = 1.2$ (black histogram).

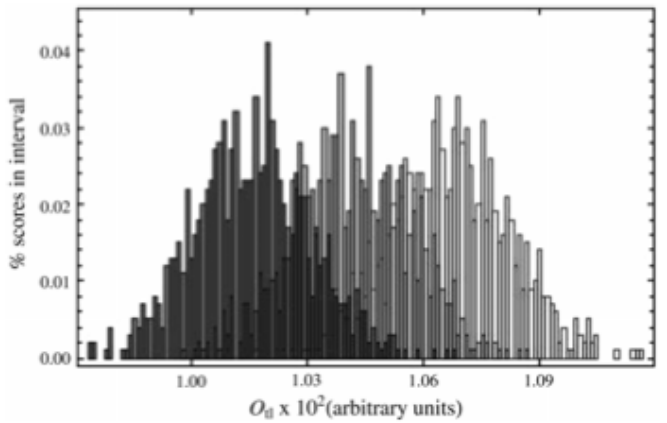
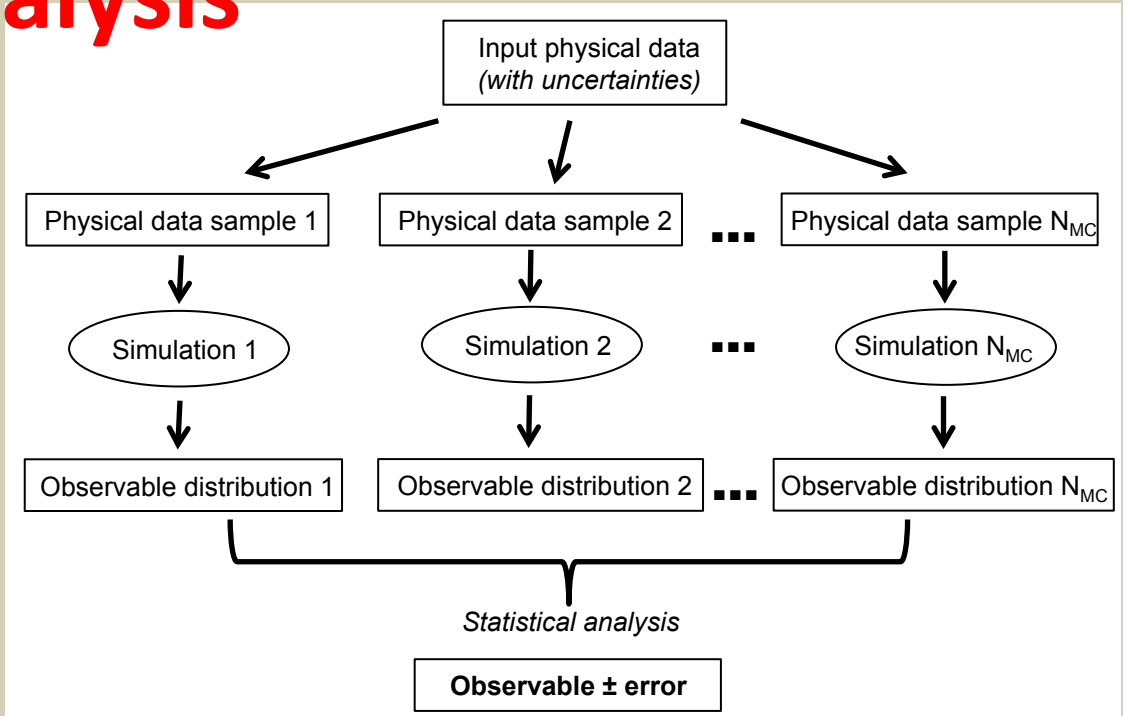



Fig. 5. Results for 1000 of Monte Carlo simulations for the observable $O_{tl}$, each encompassing $10^6$ events, for an observable scored far away from the primary particle source (see text), produced with different values of the $\Sigma_S$ input physical parameter: $\Sigma_S = 1$ (white histogram), $\Sigma_S = 1.1$ (grey histogram) and $\Sigma_S = 1.2$ (black histogram).

# Sensitivity analysis

**Computational cost**
**(brute force implies that thousands of MC runs are needed)**




To reduce computational costs, one can transform a full sensitivity analysis into the search for the most probable output

➔ this means giving up a full statistical characterization of the output

- Mathematical methods
- Software toolkits *(DAKOTA, PSUADE..)*

**Then (4rd step to UQ) we assume to have N independent source of uncertainty $x_1$, …,$x_N$ with their associate known PDFs (may be flat, normal or …) and we want to derive the PDF for an observable Y in the range (y,y+dy) from a MC simulation encompassing $N_E$ events:**

**A not trivial assumption, by the way**

$$G_{MC}(y) = \int_{-\infty}^{+\infty} d\vec{x}\, f_1(x_1)\cdots f_N(x_N) \exp\left[-\frac{(y - y_0(\vec{x}))^2}{\sigma_{y_0}^2 / N_E}\right] \sqrt{\frac{N_E}{2\pi\sigma_{y_0}^2}}$$

**from the Central Limit Theorem.**

**This result trivially derives from probability composition**

**In the limit $N_E \longrightarrow \infty$:**

$$G(y) = \int_{-\infty}^{+\infty} d\vec{x}\, f_1(x_1)\cdots f_N(x_N)\, \delta(y - y_0(\vec{x}))$$

**The task is to obtain G(x), not $G_{MC}$(x), but we need a general method to handle these expressions**

**COMMENTS:**

- This expression is exact and it has no (more) reference to MC: it relies only on the assumption of independence of the (input) uncertainties (in principle not necessary) and on probability composition rules. **We could establish this relation without any reference to simulation**
- There are however 2 **seemingly obvious, but absolutely non trivial assumptions**: (a) the existence of an (underlying) deterministic physical model (in this case some form of the transport equation) and (b) the ability to assign probabilities to the input unknowns

**How (b) can be done in common cases like this?**

**(ionization cross section study from Pia, Seo, Batic, Begalli, Kim, Quintieri, Saracco - IEEE Trans. Nucl. Sci. 58(2011)3246)**

**Cannot be done by MC users: validation studies are needed AND a methodology must be assessed (evidence theory???)**

**FUTURE RESEARCH TASK**

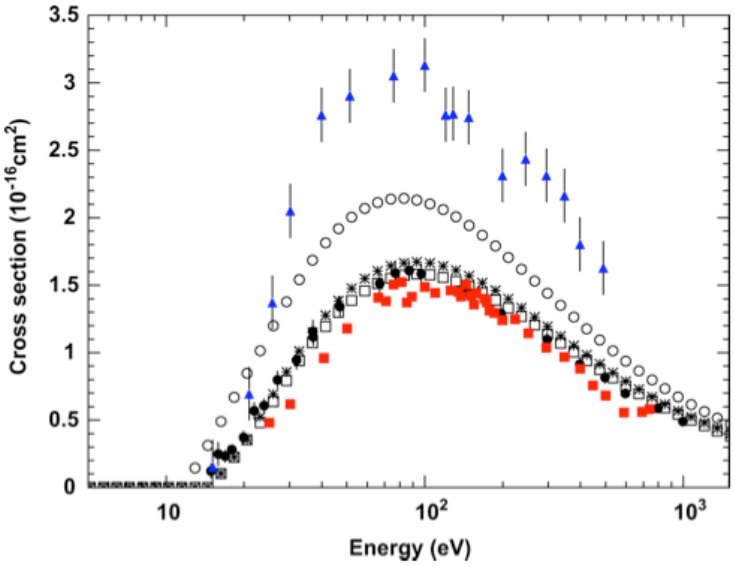



Fig. 26. BEB electron impact ionization cross section, $Z = 7$: BEB model with EADL binding energies for all shells (empty circles), BEB model with EADL binding energies except for ionization energies replaced by NIST values (empty squares), BEB model with Lotz binding energies (asterisks) and experimental data from [103] (black circles), [104] (red squares) and [105] (blue triangles).

# How it works (1 – dimension)

$$G(y) = \int_{-\infty}^{+\infty} dx\, f(x)\,\delta(y - y_0(x)) = \left|\frac{dx(y_0)}{dy_0}\right|_{y_0=y} f(x(y))$$

**We must know (and invert) the "susceptivity" $y_0(x)$!!!**
**We can use MC to study this (see later).**
**Lower computational cost**

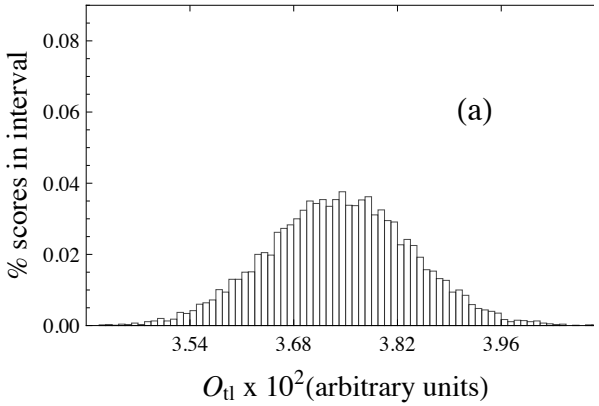
EXAMPLE: a (very) simplified "transport code", a random path generator ruled by two constant parameters describing the relative probability of absorption ($\Sigma_A$) and scattering processes ($\Sigma_S$), sampling an observable – track-length in this case.

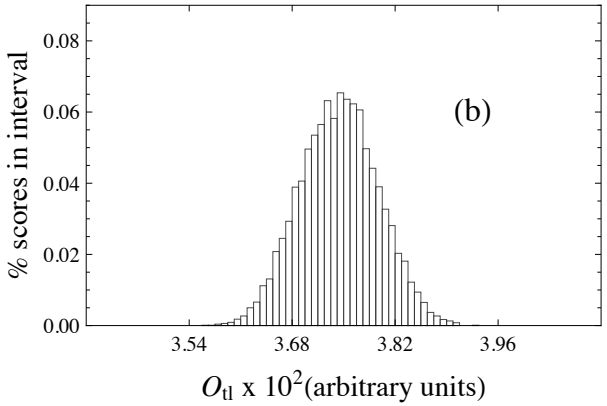
*(by the way, this simulates the propagation of neutral particles in an uniform medium with constant scattering and absorption cross-sections and isotropic scattering)*

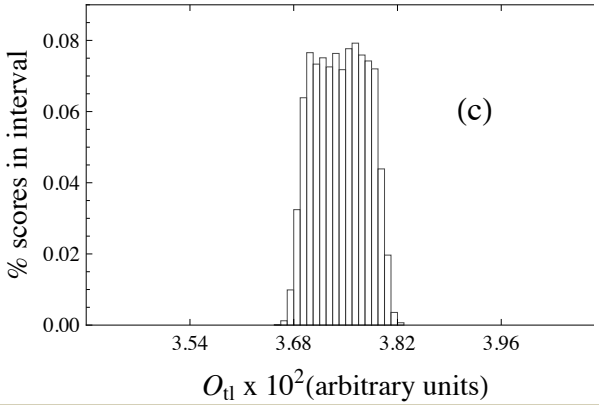
If $\Sigma_S$ is affected by some uncertainty, say $\Sigma_{S,\,min} < \Sigma_S < \Sigma_{S,\,max}$ we run many simulations varying its value with some known probability (for instance flat)

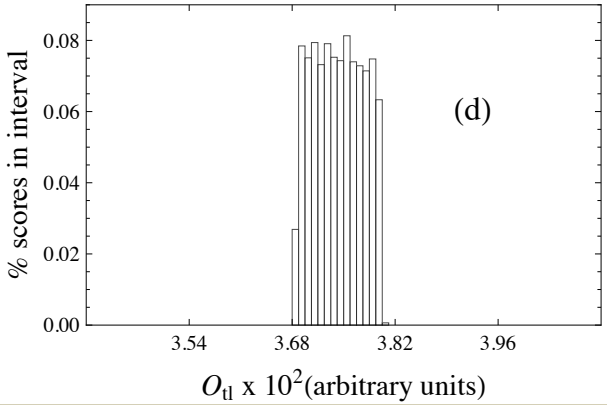
# Disentangling uncertainties

Algorithmic (statistical) and parameter uncertainties




Results for a track length observable scored in a volume near the source

15000 simulations

(a) $10^5$ events

(b) $5\ 10^5$ events

(c) $10^6$ events

(d) $10^8$ events

**linear susceptivity (?)**

As statistical errors decrease, the distribution of the observable **is dominated by parameter uncertainties only**

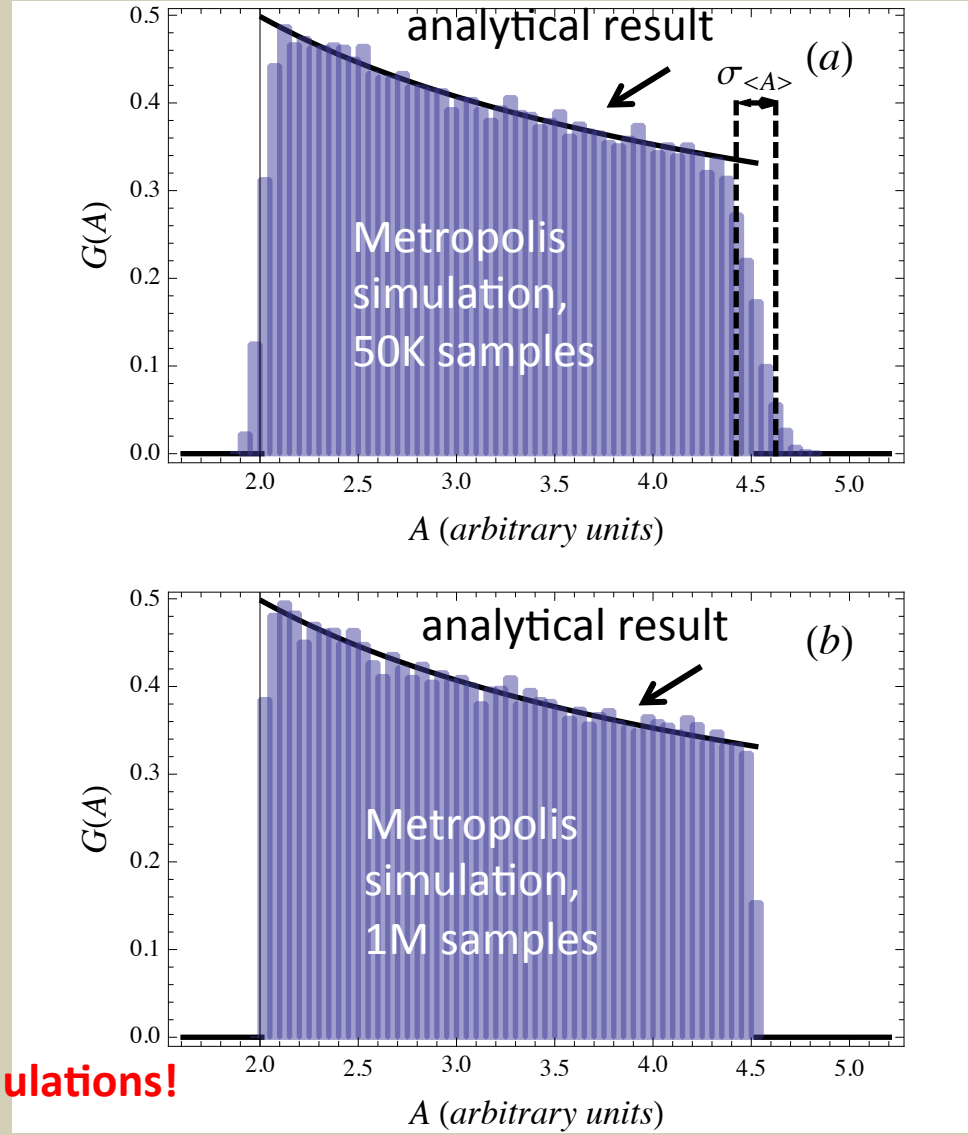
# How all it works (II) : verification
## (when all is under control …)

Example: evaluate the **area of the circle** with some **flat uncertainty** on the measure of its **radius**

$f(R)=\theta(R-R_{min})\theta(R_{max}-R)/(R_{max}-R_{min})$

$A(R) = \pi R^2$ the exact solution (susceptivity) is **known**

$$G(A) = \frac{\vartheta(A - A_{min})\vartheta(A_{max} - A)}{2\left(\sqrt{A_{max}} - \sqrt{A_{min}}\right) \times \sqrt{A}}$$




**100K simulations!**

# How all it works (III) : verification

If susceptivity $y_0(x)$ is linear and input PDF is flat the expression for $G_{MC}(x;N)$ is

$$G_{\rm emp}(x) \simeq \frac{1}{2(b-a)}\left[\mathrm{erf}\left(\frac{\sqrt{N}(x-a)}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{\sqrt{N}(x-b)}{\sqrt{2}\sigma}\right)\right]$$

Which explains the observed behavior and gives hints on the precision needed in the simulations.

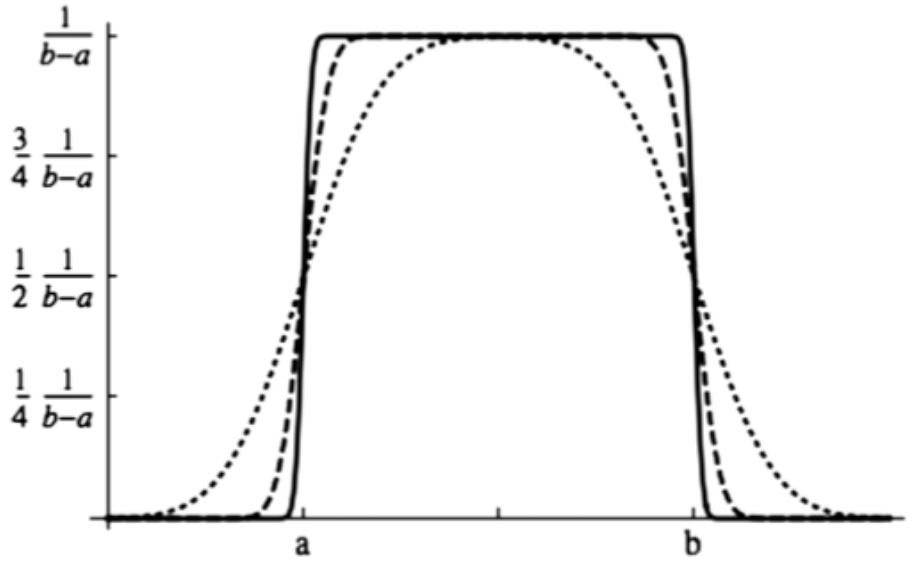



Fig. 8. Form of $G_{\rm emp}(x)$ for different values of $\sigma^2/N$: $10^{-3}$ dotted line, $10^{-4}$ dashed line and $10^{-5}$ solid line.

# A direct sampling of $G_{MC}(y)$ requires thousands of MC simulations

Knowledge of G(x) (not of $G_{MC}$) is

**the ultimate goal of UQ**

**Knowledge of $y_0(x_1,\ldots,x_N)$ required**

*Caution: statistical errors un $y_0(x_1,\ldots,x_N)$*

**USE MC to sample $y_0(x_1,\ldots,x_N)$**

**Fixed values of $x_1,\ldots,x_N$**

*…with high statistics on $y_0(x_1,\ldots,x_N)$*

**Reduced number of simulations needed**

# The task of UQ

The feasibility of UQ requires:

1) To **know input uncertainties** and their probability distributions - Validation of MC modeling ingredients needed

2) To be able to **solve explicitly for G(x)**:
   An exact mathematical context needed

3) To use MC simulations to **determine parameters in G(x)**:
   *(in the previous example, to find* $A_{max/min}$ *from simulation and to determine the proper behavior* $A^{-1/2}$*)*
   Possible with few simulations with predetermined accuracy

2) is independent from the features of the specific problem and can be solved under wide assumptions

**this is an exact mathematical framework for UQ**

# Many parameters problem

In the generic case we have many input parameter unknowns:

$$G(y) = \int_{-\infty}^{+\infty} d\vec{x}\, f(\vec{x})\delta(y - y_0(\vec{x}))$$

We make two "reasonable" assumptions:

- The $x_k$ are independent: $f(x_1,\ldots,x_N)=f_1(x_1)\ldots f_N(x_N)$
- $y_0(\vec{x})$ is linear (if necessary subdivide the domain of variability of the unknowns in such a way to fulfill the condition)

Under these hypothesis the evaluation of G(x) reduces to a well known problem in probability theory: the determination of the weighted sum of a certain number of independent stochastic variables.

$$G(y) = \int_{-\infty}^{+\infty} \prod_k dx_k f_k(x_k)\, \delta\left( y - \bar{y}_0 - \sum a_k (x_k - \bar{x}_k) \right)$$

Not soluble in general, but can we **approximately** solve it, with a prefixed accuracy?

# Some remarks

Under these assumptions $\sigma_y^2 = \sum_k \left( \frac{\partial y_0}{\partial x_k} \right)^2 \sigma_k^2$ with $\sigma_k^2$ the variances of the individual input unknowns.

Is this quantity a good measure of the uncertainty for y? **In general the answer is negative**

For M input unknowns we need *a priori* M+1 simulations to

determine the values $\frac{\partial y_0}{\partial x_k}$, a task that can be pursued **reasonably if the number of input unknowns is not so large.**

So detailed physical knowledge of the problem at hand is required to select a proper set of physical parameters on which is meaningful to attempt a full Uncertainty Quantification.

**We then emphasize that a full UQ is PROBLEM SPECIFIC**

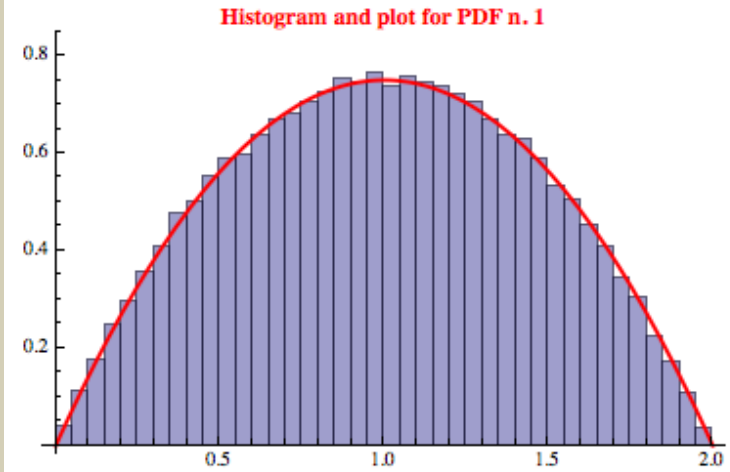
**BUT we are not sure that $\sigma_x^2$ is a proper measure of the output uncertainty, since we do not know the exact form of G(x).**

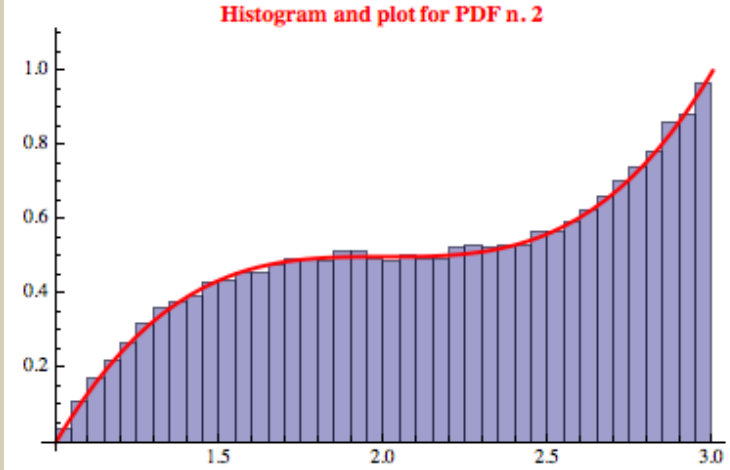
In some **useful selected cases** the form of G(x) is known:

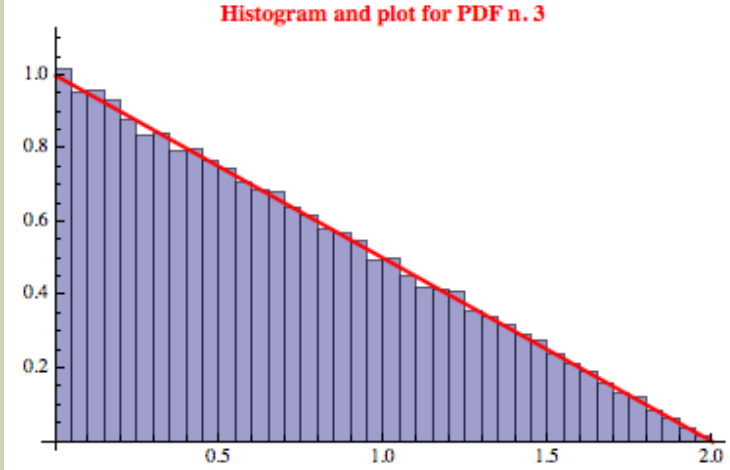
- all the input unknowns are normally distributed: in such case G(x) is normal with the quoted variance
- all the input unknown are uniformly distributed: in such case a generalization of the Irwine-Hall distribution holds
- all the input unknowns have $\alpha$-stable distributions with the same $\alpha$ value: in such case G(x) is again a stable distribution with the same $\alpha$ value (e.g. the Lorentz distribution)

**We recently proved that a general (polynomial) form exists for the weighted sum of generic polynomial distributions over different intervals: this result can be used directly to find an approximate form of G(x) with arbitrary predetermined accuracy in the general case (currently working on some technicalities: which is the best way to obtain a polynomial approximation to a given PDF, from a computational point of view).**

**It is a proper generalization of the Irwine-Hall distribution.**



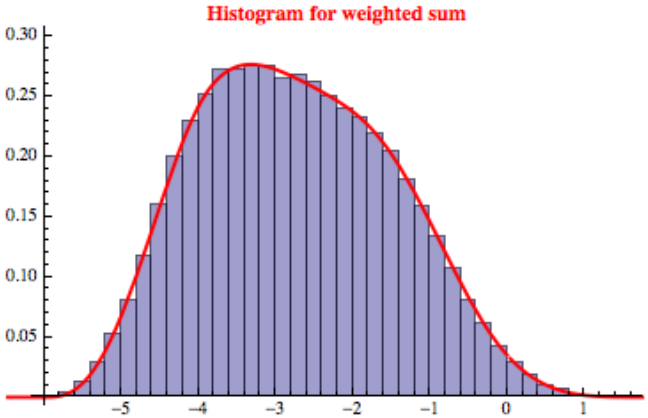




$F_1(x) = (x-1)^2-1, \quad 0<x<2, \quad$ weight 1

$F_2(x) = 1+(x-2)^3, \quad 1<x<3, \quad$ weight -2

$F_3(x) = 2-x, \quad 0<x<2, \quad$ weight 1




Nodes are $\{-6, -4, -2, 0, 2\}$

Form of PDF is

$$-6 < x < -4 \quad -\frac{(6+x)^3\left(8800+1488x-656x^2-40x^3-10x^4+x^5\right)}{286\,720}$$

$$-4 < x < -2 \quad \frac{-147\,712-321\,536x-191\,744x^2-56\,448x^3-7840x^4-448x^5+8x^7+x^8}{286\,720}$$

$$-2 < x < 0 \quad \frac{9984-20\,480x+14\,336x^2-2688x^3-7840x^4-2240x^5-112x^6+8x^7+x^8}{286\,720}$$

$$0 < x < 2 \quad -\frac{(-2+x)^5\left(312+140x+18x^2+x^3\right)}{286\,720}$$

In red the theoretical result, the histograms are random samples

# Current scope of applicability

Single parameter uncertainty (see [1]):

- complete analysis of uncertainty propagation available
- simulation is used solely to determine the values of the parameters defining the output probability density function
- confidence intervals for the output are known with a statistical error that can be predetermined

Many parameter uncertainty (see also [2]):

- a complete UQ is possible only for independent input uncertainties.
- calculability issues may exist, in practice, if the number of parameters considered is high and/or if linearity of $x_0(\Sigma_k)$ is questionable
- confidence interval for the output are affected by the statistical errors in the determination of the required parameters AND by errors in the polynomial approximations required
- in principle a predefined accuracy can be obtained
- calculation issues must be studied

[1] - P. Saracco, M. Batic, G. Hoff, M.G. Pia – "Theoretical ground for the propagation ofuncertainties in Monte Carlo particle transport", submitted to IEEE Trans. Nucl. Phys., 2013.
[2] – P. Saracco, M.G. Pia – "Uncertainty Quantification and the problem of determining the distribution of the sum of N independent stochastic variables: an exact solution for arbitrary polynomial distributions on different intervals", submitted *to Journ. Math. Phys.*, 2013.

## A (hidden) hypothesis

Input unknowns should not be modified by simulation …

(a) Energy deposited in the system by transported particles may modify locally the temperature and, then, the cross sections averaged over the motion the atoms of the materials. General purpose MC does not handle this.
(b) Activation processes

In these cases one should couple the MC simulation with an appropriate solver to keep track of the evolution of the (whole) system: thermo-hydraulics, Bateman equations, …

**The only true hypotheses on which the approach is grounded are:**

**(a) independence of the input unknowns**
**(b) Abibility to assign PDF**

# Conclusion and outlook

We have established:

- A **novel conceptual approach**
- A **mathematical framework**
- **Calculation methods** for single and many parameter uncertainties

to determine the **intrinsic uncertainty**
of the results of Monte Carlo simulation
*(beyond statistical uncertainty)*

These developments are applicable to Monte Carlo simulation in general
Particle transport
Event generators

## Outlook

- Verification in a realistic experimental scenarios
- Application software system
- **Enlargement of dynamical models and solution methods (other than MC)**