



EUROPEAN
COMMISSION

Community research

PAMINA

Performance Assessment Methodologies in Application to Guide the Development of the Safety Case

(Contract Number: **FP6-036404**)



DEVELOPMENT AND TESTING OF A TEMPLATE TO PRESENT PA RESULTS DELIVERABLE (D-N°:2.1.B.2)

Author(s):

R. Bolado & A. Badea (JRC)

Date of issue of this report: **19/08/09**

Start date of project : **01/10/2006**

Duration : **36 Months**

Project co-funded by the European Commission under the Euratom Research and Training Programme on Nuclear Energy within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
RE	Restricted to a group specified by the partners of the [PAMINA] project	
CO	Confidential, only for partners of the [PAMINA] project	

Foreword

The work presented in this report was developed within the Integrated Project PAMINA: **P**erformance **A**ssessment **M**ethodologies **I**N **A**pplication to Guide the Development of the Safety Case. This project is part of the Sixth Framework Programme of the European Commission. It brings together 25 organisations from ten European countries and one EC Joint Research Centre in order to improve and harmonise methodologies and tools for demonstrating the safety of deep geological disposal of long-lived radioactive waste for different waste types, repository designs and geological environments. The results will be of interest to national waste management organisations, regulators and lay stakeholders.

The work is organised in four Research and Technology Development Components (RTDCs) and one additional component dealing with knowledge management and dissemination of knowledge:

In RTDC 1 the aim is to evaluate the state of the art of methodologies and approaches needed for assessing the safety of deep geological disposal, on the basis of comprehensive review of international practice. This work includes the identification of any deficiencies in methods and tools.

In RTDC 2 the aim is to establish a framework and methodology for the treatment of uncertainty during PA and safety case development. Guidance on, and examples of, good practice will be provided on the communication and treatment of different types of uncertainty, spatial variability, the development of probabilistic safety assessment tools, and techniques for sensitivity and uncertainty analysis.

In RTDC 3 the aim is to develop methodologies and tools for integrated PA for various geological disposal concepts. This work includes the development of PA scenarios, of the PA approach to gas migration processes, of the PA approach to radionuclide source term modelling, and of safety and performance indicators.

In RTDC 4 the aim is to conduct several benchmark exercises on specific processes, in which quantitative comparisons are made between approaches that rely on simplifying assumptions and models, and those that rely on complex models that take into account a more complete process conceptualization in space and time.

The work presented in this report was performed in the scope of RTDC 2.

All PAMINA reports can be downloaded from <http://www.ip-pamina.eu>.



TABLE OF CONTENTS

1.	Introduction	4
2.	Notation	6
3.	Potentially targeted output variables	7
4.	Statistics to characterise output uncertainty	9
4.1	Numeric statistics	9
4.1.1	Central tendency statistics	9
4.1.2	Quantiles	13
4.1.3	Dispersion statistics	15
4.1.4	Shape statistics	17
4.2	Sample size selection	20
4.2.1	The empirical estimator	21
4.2.2	Wilks estimator	22
4.2.3	Tolerance intervals	23
4.3	Graphic tools	25
4.3.1	The ECDF and the ECCDF	25
4.3.2	The histogram	28
4.3.3	PDF estimation	29
4.3.4	Boxplots	33
5.	Tools for identifying most important input parameters	35
6.	Time dependent output variables	37
7.	A template to present PA results	39
7.1	Non time-dependent output variables	39
7.1.1	Uncertainty analysis	39
7.1.2	Sensitivity analysis	40
7.2	Time-dependent output variables	40
7.2.1	Uncertainty analysis	40
7.2.2	Sensitivity analysis	41
7.3	Remarks about suggested and optional statistical indicators	42
8.	Template application	45
8.1	Non time-dependent output variables (Peak annual dose rate in biosphere due to ^{129}I and time to the peak)	47
8.1.1	Uncertainty analysis results	47
8.1.2	Sensitivity analysis results	48
8.2	Time-dependent output variables (Annual dose rate in biosphere due to ^{129}I at all times and at six selected times)	50
8.2.1	Uncertainty analysis results	50
8.2.2	Sensitivity analysis results (all times)	52
9.	Conclusions	53
	References	54
	Annex A: Monte Carlo simulation	56
	Annex B: Classical inference methods	67
	Annex C: Properties of Quantile Estimators	76

1. Introduction

The process to develop a Performance Assessment of a nuclear High Level Waste repository (HLW) involves modelling the whole system, which classically is considered to be divided into three parts: i) The near field or engineered facilities including the disturbed part of the geosphere, ii) the far field or part of the geosphere that hosts the repository, and iii) the biosphere, eventual sink of radioactive pollutants. Modelling such a system means modelling the inventory of radionuclides, the processes that deteriorate the facility and that produce the release of radionuclides in the long term, their transport through the geosphere and their spread over the biosphere, which ultimately will produce doses on humans. All those models will be integrated as submodels of the system model.

Moreover, implementers should be able to foresee potential disruptive scenarios that could induce 'worse than expected' behaviour of the system. This involves addressing events and processes that, though unlikely, could reasonably happen, and would produce more adverse consequences than the expected normal evolution of the system. Two activities are triggered when alternative scenarios are identified: Likelihood estimation and adapting the system model to the specific physical and chemical conditions produced by the scenario.

Parameters such as coefficients, boundary and initial conditions of the differential equations used in the system model are usually affected by uncertainty. Characterising these uncertainties is an extremely time consuming task that includes laboratory and field experiments, collection of historical records, search in databases and use of expert judgment. Formally, as soon as scenarios are identified and their probabilities are estimated, the system model is available and parameter uncertainty is assessed, computations could be started to estimate the adverse consequences to humans and the environment in the future. Eventually, the sampled input data and the values of the output variables obtained via Monte Carlo simulation will be available. This enormous quantity of data may be used to characterize the uncertainty of output variables and to identify what input parameters contribute more to such uncertainty.

The target of this report is to show a systematic way to present results produced in a PA study, which are typical elements of an Uncertainty Analysis (UA) and of a Sensitivity Analysis (SA). Its structure is as follows. Chapter 2 is dedicated to show the mathematical notation used along the whole report, chapter 3 is dedicated to recall potential output variables that deserve to be included, typically safety and performance indicators and related variables. Chapter 4 provides an overview of most useful numeric and graphic statistics available to characterise output variable uncertainty. Each statistic considered is described, showing main advantages and shortcomings. Some attention is also dedicated to the issue of selecting an appropriate sample size. Chapter 5 is dedicated to show a selection of a minimum set of SA tools that, in the opinion of the authors, should be used to analyse input-output relations. Chapter 6 indicates the simple tasks needed to show the results of the statistics described in chapters 4 and 5 when applied to time-dependent output variables. Chapter 7 provides the actual template proposed, which is based on the previous chapters.

Chapter 8 is the actual application of the template to the dose rate due to ^{129}I in the Spanish reference concept in granite. The peak dose rate due to ^{129}I and the time to the peak are taken as non time-dependent output variables, while the dose rate due to ^{129}I at all times and at 6 specific times are taken as time-dependent output variables. Chapter 9 gives the main conclusions of this report. After the references, three annexes provide some extensions about subjects mentioned in the main text, which could be of help to some readers. It is important to stress that the whole text has been written keeping in mind the idea that the data used in the PA are generated via Monte Carlo simulation, under non-biasing sampling schemes.

Many examples of application of the different statistics described in this report are shown along the different sections. PrvÁková et al. (2008) has been the source of data for the examples. This reference (PAMINA milestone M4.3.2) reports the results of a benchmark study developed in a probabilistic framework to compare the results obtained with two models that simulate the behaviour of an Intermediate and Low-level waste (ILW) repository in indurated clay (argillite) in France. The output variables considered are the molar flows of three radionuclides (^{129}I , ^{79}Se and ^{94}Nb) at six different surfaces. 24 input random parameters describe release rates of each waste component and hydraulic and transfer properties of each porous medium (permeability, diffusion, porosity, adsorption, solubility limit). This data set will be called 'reference data set' along the whole text, and the study itself will be called the 'reference study' or 'reference problem'.

2. Notation

Upper-case letters will be used along the whole text to denote random variables (or *variates*), while their realizations will be denoted by the corresponding lowercase letters. The letter X (x) will be associated with the input parameters and the letter Y (y) with the output.

- rv : Random variable
- X, Y : Random variables;
- (X_1, X_2, \dots, X_n) : A random sample of random variable X ;
- (x_1, x_2, \dots, x_n) : The corresponding realization of the random sample;
- $X = (X_1, X_2, \dots, X_d)$: A random vector of size d ;
- $Y=Y(X)$: The output of the numerical model;
- F : The cumulative distribution function (CDF): $F(x) = P(X \leq x)$;
- f : The probability density function (PDF) : $F(x) = \int_{-\infty}^x f(t)dt$;
- \mathbb{R} : set of real numbers;
- x_α, q_α : the α -quantile of X , defined as $F(x_\alpha) = \alpha$;
- $\lfloor x \rfloor$: The largest integer smaller or equal to x ;
- $\lceil x \rceil$: The smallest integer larger or equal to x ;
- $X_{(k)}$: Order statistics (of order k) ;
- μ : Mean of a random variable;
- σ^2 : Variance of a random variable;
- \bar{x} : Sample mean;
- σ_x^2, s^2 : Sample variance; σ_x, s : sample standard deviation
- $E(.)$: mathematical expectation
- $Var(.)$: variance
- iid : independent, identically distributed
- $\#$: Cardinal of a set

3. Potentially targeted output variables

In a PA, a large variety of output variables is obtained as a result of the Monte Carlo simulation. Among them, the most important is the effective total dose rate over time, which is used in the regulation of many countries as the main measure of risk/safety associated to the repository. Consequently, safety limits are usually imposed on it. In addition to this fundamental output variable, other output variables are also relevant, either because they provide additional information about the system safety (alternative safety indicators), or because they provide measures of the system performance, or of the performance of some subsystem (performance indicators). The European Commission FP-5 research project SPIN (Becker et al., 2002) was dedicated to the identification and study of safety and performance indicators, in order to select the most appropriate set.

In addition to the effective total dose rate, SPIN partners identified as relevant safety indicators the following two output variables:

- The radiotoxicity concentration in biosphere water
- The radiotoxicity flow¹ from geosphere

The effective total dose rate over time was considered very relevant at early times (first several thousand years), while the radiotoxicity concentration in biosphere water was considered most relevant at intermediate times (between several thousand years and several tens of thousand years) and the radiotoxicity flow from the geosphere at late times (hundreds of thousand years and beyond).

The following types of output variables were identified as main performance indicators

- Inventories in compartments
- Inventories outside compartments
- Flows getting out of compartments
- Concentrations in compartment water
- Transport time through compartments

Among the three safety indicators and the five types of performance indicators selected, all but one type are time dependent variables; only the last one, the transport time through compartments, does not vary in each realization over time (only one value per realization). Statistics considered in chapter 4 are designed for non-time dependent variables; necessarily they will be adapted to show uncertainty evolution over time.

¹ Consistently, in many works related to the area of radioactive waste repository PA, flows are called fluxes. In this report, the authors prefer to call flow to any quantity whose units are $X \cdot t^{-1}$, where X may be mass, volume, etc. Becker et al. (2002) use the expression 'flux' to name this safety indicator, which we replace by 'flow'.



These output variables are the main candidates to be analysed; which of them will formally be analysed depends on the interest of the organisation developing the study. Nevertheless, other potential candidates can also be considered. For some of those outputs selected that are time dependent, it could be of interest to analyse also the following two output variables (non time-dependent) associated to them: their peaks and the times to the peaks. The peak of a variable takes one single value per realization (if more than one peak happens, the largest is taken as the peak) and provides an upper bound for that variable. For example, in the case of the effective total dose rate, if its peak fulfils the safety criteria, it will also fulfil them.

4. Statistics to characterise output uncertainty

Descriptive statistics are used to characterise the uncertainty associated to random variables. Since typical output variables obtained in a PA via Mont Carlo simulation are random variables, common descriptive statistics are appropriate to characterise their uncertainty. The realizations of the rvs are generically denoted by (x_1, x_2, \dots, x_n) . Along the whole text the word **statistic** will design ‘a function of a sample where the function itself is independent of the sample's distribution; the term is used both for the function and for the value of the function on a given sample’.

All the tools described in this section may be applied to results obtained via Monte Carlo simulation, under different possible sampling schemes with the only restriction of not biasing the sample. This includes Simple Random Sampling (SRS) and the following variance reduction techniques: Latin hypercube Sampling (LHS), Proportional Stratified Sampling, Control Variates and input space Dimension Reduction (either trivial or achieved via Dimensional Analysis). On the other hand, special biasing techniques such as importance sampling need special output data post-processing tasks which are not addressed in this report. Annex A provides a short summary about Monte Carlo simulation and the aforementioned variance reduction techniques.

4.1 Numeric statistics

Numeric statistics characterise different properties of a random variable. Four types are considered in the following subsections: statistics of central tendency, statistics that characterise the full distribution of the variable, statistics of spread or dispersion and statistics that characterise specific aspects of the distribution shape.

4.1.1 Central tendency statistics

Measures of central tendency, or of location, compute one single number that gives the best possible representation of the value around which the data are located. Four statistics are the most used: mean, median, geometric mean and mode.

The mean

The most important measure of central tendency is the arithmetic mean of the sample, defined by:

$$\bar{x} = 1/n \sum_{i=1}^n x_i . \quad (4.1.1.1)$$

Other notations: \bar{x}_n (whenever knowing the sample size is needed) and μ .

Important characteristics of the arithmetic mean:

- It is a linear statistic in the following sense: for two samples x and y of the same size $\overline{ax + by} = a\bar{x} + b\bar{y}$ ($a, b \in \mathbb{R}$).
- It is not a robust statistic; it is very sensitive to extreme values (see the example in page 12).
- It gives a very good measure of location for homogeneous symmetric sets of data. The variance (see section 4.1.3) is minimised when the mean is the measure of location used as a reference.
- If the sample is heterogeneous (existence of data obtained under different conditions), the mean can become completely useless as a measure of central tendency (this problem affects all measures of central tendency); it could even take a value outside the range of definition of the variable under study. For instance a sample made of two subsamples of equal size that do not overlap at all, the arithmetic mean can be in between, just where the variable takes no value. This could be the case when combining the results obtained for two scenarios whose output values do not overlap.

The arithmetic mean is the most important measure of central tendency because it is the best available estimator for the actual mean of any random variable. It is a statistic, thus it is a random variable itself, and it will take different values in different samples. For this, as for any of the other statistics described in this report, we know that the value obtained from a sample is not the same as the actual value that we want to estimate, but we expect that it will be close. Confidence intervals are the tools provided by classical inference methods to get an estimate of the error committed when estimating a given characteristic of a random variable with a given statistic (see annex B for a short review of classical inference methods and for a correct interpretation of a confidence interval).

Nevertheless, exact confidence intervals are available only for input uncertain parameters of a very limited number of types of random variables. The output random variables that appear in a PA do not frequently fit well those types and do usually show not so good behaviour (highly skewed, high kurtosis). In case confidence intervals be needed, which is the case of a PA, when a measure of uncertainty about the estimates given is needed, asymptotic confidence intervals may be estimated, see annex B. Unfortunately, the problem of asymptotic confidence intervals is that we never know if the asymptotic conditions have been either achieved or not, so the actual confidence level is never known (i.e. we could provide a 60% confidence interval when we actually think it is a 95% confidence interval). That is why the authors of this report are not in favour of providing confidence intervals for the means of output variables. When output variables fit reasonably well standard probability models, which is not frequent, classical inference methods may be used, as explained in annex B, to estimate parameters (i.e. means) and provide confidence intervals.

In case such an interval is asked for the mean, the confidence interval supported by classical inference theory is

$$\left[\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right] , \quad (4.1.1.2)$$

where \bar{x} plays the role of $\hat{\theta}_{ML}$, s/\sqrt{n} plays the role of $\sigma(\hat{\theta}_{ML})$ in the last row of table B.2 in annex B, and s is the sample standard deviation as defined in (4.1.3.4). This interval will have a confidence level close to $1-\alpha$ (the expected one) only if s is very close to the real standard deviation of the random variable under study, which is only true for asymptotic values of the sample size ($n \gg$). The authors of this report do strongly advice not to use this type of asymptotic confidence intervals when the sample kurtosis is high, see section 4.1.4.

The geometric mean

The geometric mean of the sample is defined as:

$$\tilde{x} = \left(\prod_{i=1}^n x_i \right)^{1/n} . \quad (4.1.1.3)$$

The geometric mean is only of interest when all sampled values are positive. It may also be computed when there are null values, but then it is also null. The geometric mean gives a measure of central tendency when a logarithmic scale is used. The usual way to compute the geometric mean (in order to avoid numeric problems) is computing the arithmetic mean of the logarithm of the actual sampled values and transforming the obtained result consequently, i.e.:

$$\tilde{x} = 10^{\left[(1/n) \sum_{i=1}^n \log_{10}(x_i) \right]} . \quad (4.1.1.4)$$

The geometric mean is always either equal to or smaller than the arithmetic mean. In cases when positive and null values are mixed in the same sample, it may be of interest to compute a geometric mean restricted to the m sampled positive values ($m < n$). The geometric mean does also answer the question “if all quantities had the same value, what would that value have to be in order to achieve the same product”.

As many rv used as outputs in PA studies are spread over several orders of magnitude, the geometric mean is useful to estimate the “center of mass” of the data in a logarithmic scale (the arithmetic mean estimates the “center of mass” in a linear scale).

The median

Another alternative to estimate a central value is the median. The median of a sample is the value that splits the sample in two equal parts. Suppose that the sample sorted in ascending order is $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, then

$$\text{med}(x) = \begin{cases} (x_{(n/2)} + x_{(n/2+1)})/2 & \text{if } n \text{ is even} \\ x_{((n+1)/2)} & \text{if } n \text{ is odd} \end{cases} \quad (4.1.1.5)$$

It is the value of x for which the CDF $F(x) = 1/2$.

The median is a robust indicator, but it is more difficult to perform algebraic computations using it than using the mean. For instance, the linearity property is no longer valid. On the other hand, the median is conserved when applying a strictly monotonic increasing transform to the sample (the transform of the median is the median of the transformed values), which is not the case for the mean.

Example:

The sample data ($n=51$) represents the release of ^{94}Nb getting out of the fractured zone after 5000 years computed in the reference problem; see Prváková et al. (2008). In figure 1, the point in the upper left corner is an extreme value.

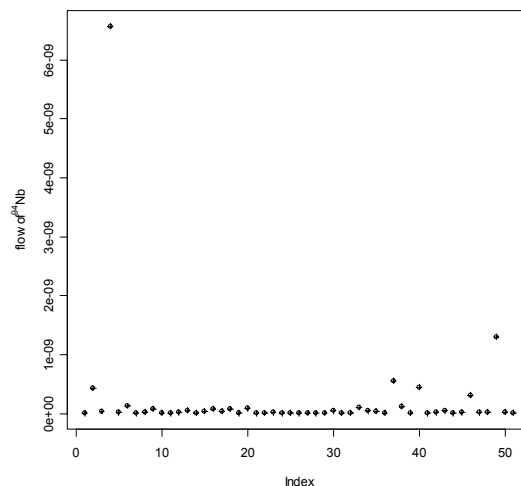


Figure 1.- A sample of ^{94}Nb getting out of the fractured zone after 5000 years.

The mean of the whole sample is $\bar{x}_{51} = 2.18E-10$, while if we exclude the extreme point we obtain $\bar{x}_{50} = 9.13E-11$. The effect on the mean of that single value is huge; excluding it from the sample produces a decrease of 58% in the mean. This is not the case with the median.

The original median (sample of size 51) was $med_{51}(x) = 2.22E-11$. After removing the extreme value, the new median is $med_{50}(x) = 2.21E-11$; the two values are quite similar, which is due to its robustness as a measure for the central tendency.

The mode

The mode is the location of a local maximum of the PDF. A PDF can be multimodal, which often means that we are dealing with heterogeneous populations. For discrete data, the mode is the most frequently observed value. However, the estimation of the mode using a sample depends entirely on the method used to estimate the PDF (see section 4.3.3).

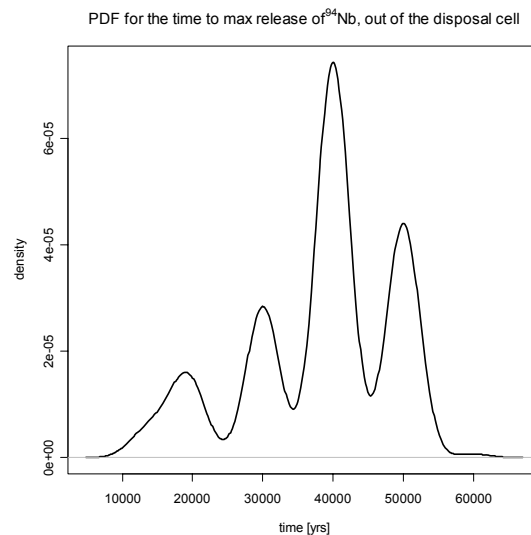


Figure 2.- Example of a multimodal PDF; there are 4 modes around 19000, 30000, 40000 and 50000 years.

4.1.2 Quantiles

Quantiles generalize the median for a probability α different from $\frac{1}{2}$, i.e. they are values that split the data in two parts, such as the proportion of data inferior or equal to this value is equal to α . The α -quantile q_α is defined by the equation

$$F(q_\alpha) = \alpha, \quad \forall \alpha \in [0,1]. \quad (4.1.2.1)$$

However, when the cumulative distribution is not a strictly increasing function, this equation might have either an infinite number of solutions or no solution at all, as can be seen in figure 3. The usual conventions to overcome this problem are based on the ordered observations $x_{(1)} \leq \dots \leq x_{(n)}$. The smallest observation corresponds to a probability of 0 and the largest

one to a probability of 1. The i^{th} observation corresponds to α -quantile q_α (i.e. $q_\alpha = x_{(i)}$), where α may be defined as follows:

$$\alpha = \begin{cases} (i-1)/(n-1) & \text{or} \\ (i-0.5)/n & \text{or} \\ i/(n+1) & \text{or} \\ i/n & \text{or} \\ (i-1/3)/(n+1/3) & \text{or} \\ (i-3/8)/(n+1/4). \end{cases} \quad (4.1.2.2)$$

In (4.1.2.2), the two emphasized expressions are the most used ones:

- the first one because it has a symmetry with respect to the CDF: the smallest observation corresponds to a probability of 0 and the largest one to a probability of 1; it is the one used by default by some statistical software packages such as R [<http://cran.r-project.org/>] and S [<http://www.insightful.com/>];
- the fourth one because it corresponds exactly to the definition of the empirical cumulative distribution function (see formula (4.2.1.1)).

Concerning the other expressions in (4.1.2.2):

- the second one is popular amongst hydrologists,
- the third one is used by other statistical software packages such as Minitab [<http://www.minitab.com/>] and SPSS [<http://www.spss.com/>],
- using the fifth expression, one obtains a quantile estimate that is approximately median-unbiased (i.e. the median of the estimator is approximately unbiased) regardless of the distribution of x ,
- using the last expression, one obtains a quantile estimate that is approximately unbiased for the expected order statistics if x is normally distributed.

More details about different alternatives to define quantiles may be found in Hyndman and Fan (1996).

If α is not exactly one of the values in (4.1.2.2), a linear interpolation may be used to estimate

$$q_\alpha, \text{ as for example } \alpha = (1-a)\frac{(i-1)}{n-1} + a\frac{i}{n-1} \Rightarrow q_\alpha = (1-a)x_{(i)} + ax_{(i+1)}, \quad 0 < a < 1.$$

Example:

Let us consider the following $n=6$ sample $\{0,1,2,3,4,5\}$. We assume that the i^{th} observation is the estimation of the α -quantile q_α , where $\alpha = (i-1)/(n-1)$. We want to estimate the 1/4 quantile, which is not of the form $\alpha = (i-1)/(n-1)$.

But, as $\frac{1}{5} < \frac{1}{4} < \frac{2}{5}$, there exists a value $a = 0.25$, ($0 < a < 1$), and a value $i = 2$, such that

$\frac{1}{4} = (1-a)\frac{i-1}{5} + a\frac{i}{5}$. We hence obtain $q_{1/4} = 0.75x_{(2)} + 0.25x_{(3)} = 0.75 \times 1 + 0.25 \times 2 = 1.25$.

The median is the $\frac{1}{2}$ quantile. Some other particular quantiles frequently used are:

- percentiles, the 1/100-quantiles
- deciles, the 1/10-quantiles
- quartiles, the 1/4-quantiles.

For more details concerning quantile estimation see section 4.2.

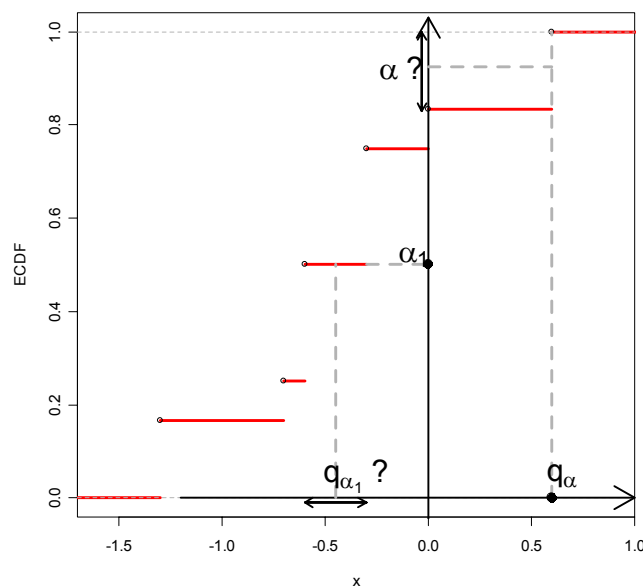


Figure 3.- Empirical cumulative distribution function and quantile signification

4.1.3 Dispersion statistics

The measures of dispersion are important for describing the spread of the data around a central value. Two distinct samples could have similar means or medians but completely different degrees of dispersion around them.

The range

The range is defined as the difference between the largest and smallest sample values:

$$range = x_{(n)} - x_{(1)} = \max(x) - \min(x). \quad (4.1.3.1)$$

It is one of the simplest measures of variability to calculate, but it depends only on extreme values (and hence it is a non robust indicator) and provides no information on the data distribution.

The interquartile range (interval)

The interquartile range is defined as the difference between the 3rd and the 1st quartiles, i.e. $q_{3/4} - q_{1/4}$. It is a robust indicator. The meaning of this indicator is that at least 50% of the “central” data are contained in this interval. It is also used for drawing the boxplots (see section 4.3.4).

The variance and the standard deviation

The variance was meant to measure the mean deviation from the mean value of the sample, by taking into account positive and negative deviations in the same manner. This is the reason for introducing the quadratic function sample variance as:

$$\text{var}(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (4.1.3.2)$$

As the variance does not have the same units as the sample (because of the squares), the standard deviation has been introduced:

$$\sigma_x = \sqrt{\text{var}(x)} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (4.1.3.3)$$

Alternative definitions of the variance and the associated standard deviation are

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2; \quad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (4.1.3.4)$$

Saporta (1990) provides further information about the different definitions of the variance.

If the sample is approximately normal, then

- The interval mean \pm one standard deviation contains approximately 68% of the measurements in the series.

- The interval mean \pm two standard deviations contains approximately 95% of the measurements in the series.
- The interval mean \pm three standard deviations contains approximately 99.7% of the measurements in the series.

When the distribution that generates the sample is unknown, similar rules, based on Chebyshev's inequality may be applied, see Jordaan (2005). However, the bounds that are computed are rather loose, but they are valid irrespective of the distribution that generates the data; knowing the sample mean and the sample standard deviation is enough to calculate them. Chebyshev's inequality states that $fr(|x_i - \bar{x}| \geq ks) \leq 1 - 1/k^2$, where k is any real number and fr stands for relative frequency. It provides useful information for $k > 1$:

- The interval mean \pm two standard deviations contains at least 75% of the measurements in the series.
- The interval mean \pm three standard deviations contains at least 89% of the measurements in the series.

The geometric standard deviation

The sample geometric standard deviation (gsd_x) is

$$gsd_x = 10^{\sigma_y}, \quad (4.1.3.5)$$

where σ_y is the standard deviation of the variable $Y = \log_{10}(X)$. The geometric standard deviation is helpful to assess the spread of values around the geometric mean. When applied to the geometric mean and the geometric standard deviation, Chebyshev's inequality states that $fr(gm \times gsd^{-k} \leq y_i \leq gm \times gsd^k) \geq 1 - 1/k^2$. When k is set to 2 and 3 respectively, the following two statements may be said

- The interval *geometric mean* \times (*geometric standard deviation*) $^{\pm 2}$ contains at least 75% of the measurements in the series;
- The interval *geometric mean* \times (*geometric standard deviation*) $^{\pm 3}$ contains at least 89% of the measurements in the series.

4.1.4 Shape statistics

The moments of a rv allow to characterize its probability distribution. Moments may be computed with respect to the origin (0) or with respect to a measure of central tendency, usually the mean. The first order moment with respect to the origin is the mean of the rv and the second order moment with respect to the mean is the variance. The third and the fourth moments define the shape of the distribution.

The skewness coefficient

The skewness coefficient is the third standardized moment with respect to the mean, i.e.

$$\gamma_1 = \frac{1}{n} \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{\sigma_x^3}, \quad (4.1.4.1)$$

where σ_x is computed as in expression (4.1.3.3). A positive coefficient means that the distribution has a long right tail, (the distribution is also known as right-skewed) while a negative coefficient means that the distribution has a long left tail (the distribution is also known as left-skewed), see figure 4. Any symmetric distribution has a skewness coefficient equal to 0, as for example the normal and the uniform distributions. It should be noted though that some non-symmetric distributions could also have a null skewness coefficient.

Other statistics may also be used to detect lack of symmetry, such as the difference between the mean and the median. The mean is larger than the median in a right-skewed set of data, while it is smaller for left-skewed set of data. Positive (all values larger than 0) right-skewed sets of data do also show large standard deviations compared to their means.

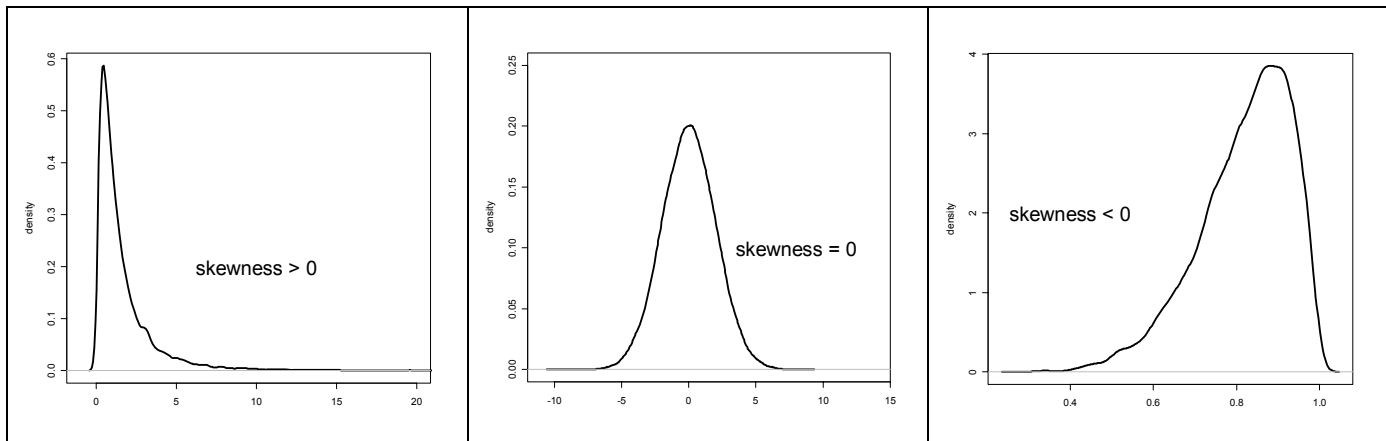


Figure 4.- PDFs for distributions with different skewness coefficients: > 0 (right-skewed, left); = 0 (symmetric, center); < 0 (left-skewed, right).

The kurtosis

The kurtosis coefficient is the fourth standardized moment with respect to the mean, i.e.

$$\gamma_2 = \frac{1}{n} \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{\sigma_x^4}. \quad (4.1.4.2)$$

It represents a measure of the “peakedness” of the distribution, see figure 5. A kurtosis coefficient larger than 3 means that the distribution has sharper “peaks” and flatter “tails” than a normal distribution (leptokurtic distribution). A kurtosis equal to 3 means that the

distribution is approximately normal (mesokurtic distribution). A kurtosis below 3 means that the distribution is flatter than the normal distribution (platykurtic distribution).

The kurtosis coefficient is sometimes defined as $\mu_4 / \sigma^4 - 3$, where μ_4 is the sample fourth moment around the mean (numerator –including factor 1/n- in expression (4.1.4.2)), in order to make the kurtosis of the normal distribution equal to zero instead of 3. The kurtosis is not an intuitive coefficient; it is quite difficult to say, by looking at a PDF if the distribution has a large or small kurtosis. What is important, in terms of shape, for a leptokurtic distribution is that there is a sharper “peak” around the mean (which means a higher probability than a normally distributed rv of values close to the mean) and “fat tails” (which means a higher probability than a normally distributed rv of extreme values), as it can be seen in figure 6.

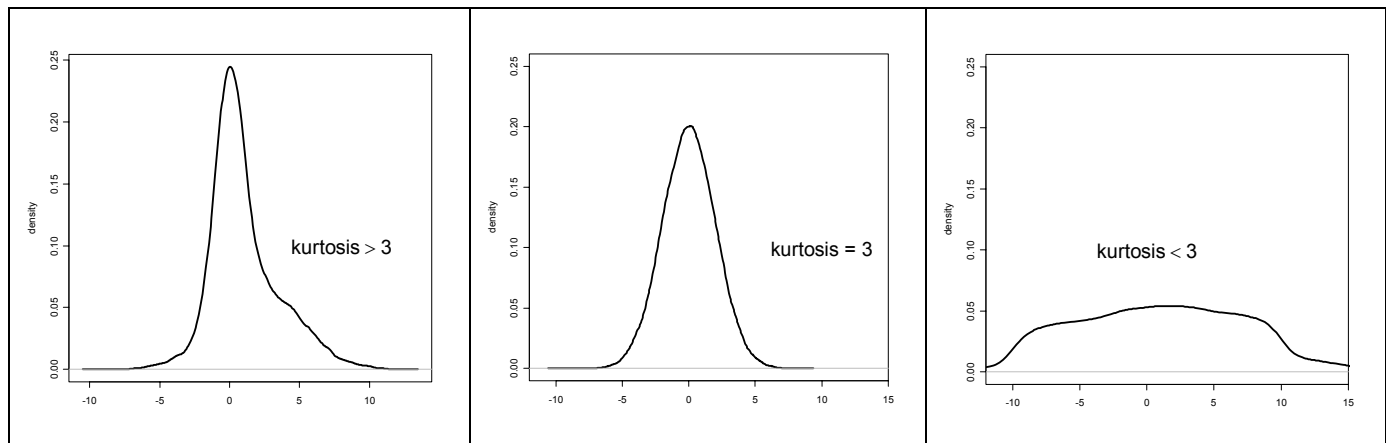


Figure 5.- PDFs for distributions with different kurtosis : >3 (left) ; =3 (center); <3 (right).

The kurtosis gives also an indication about the problems that can arise when trying to estimate the mean of a random variable. Expression (4.1.1.2) provides the way to estimate a confidence interval for the mean of a random variable that follows an unknown probability distribution. It includes the sample standard deviation (s) as the key data to know the width of the interval. The variance of the square of this statistic for a generic random variable is

$$Var(s^2) = \sigma^2 \left(2 / (n-1) + (K-3) / n \right)^{1/2}, \quad (4.1.4.3)$$

where K is the real kurtosis of the random variable. Only when this variance is very small the sample standard deviation is close to the real standard deviation of the random variable and the asymptotic confidence interval becomes accurate. Large values of K make that variance larger and makes necessary to take larger sample sizes to get accurate estimates of the standard deviation. When the sample kurtosis (which is the best available estimator for the real kurtosis) is large, the sample standard deviation may show a large variability. That is the reason not to use confidence intervals for the mean obtained under asymptotic assumptions

such as the one given in (4.1.1.2), unless the asymptotic conditions are known to be fulfilled, which is not frequent.

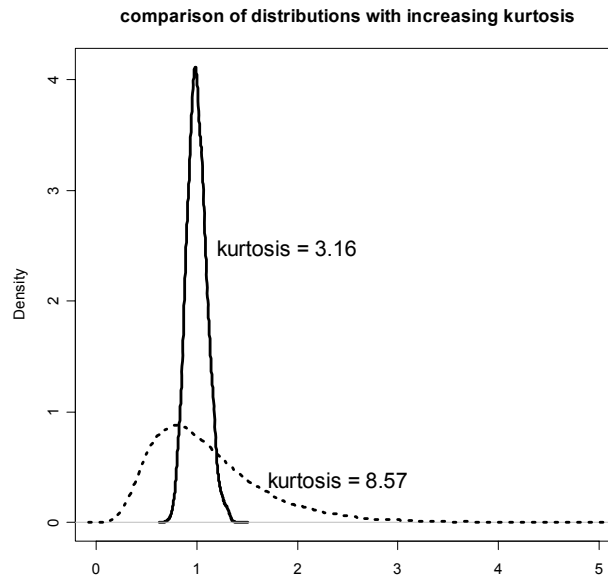


Figure 6.- Influence of increasing Kurtosis

4.2 Sample size selection

The larger the sample size, the more information the sample contains. Nevertheless, increasing the sample size means increasing the computational cost. So there is a balance between the information we can obtain and its cost. In order to set sample sizes we will consider two concepts: conservative estimates of quantiles and tolerance intervals.

Quantile estimation is a very important issue when dealing with outputs of PA codes since, in many occasions, safety limits are based on quantiles. This is the case, for example, when a safety limit is set in the following terms: In order to be acceptable, the repository should not produce, at any time in the future, an effective total dose rate higher than $D_L \text{ mSv}\cdot\text{y}^{-1}$ with a probability higher than 0.05. This means that the percentile 95% of the output variable 'effective total dose rate' should be lower than D_L .

In the case of a real-valued rv Y , quantile estimation means determining the value y such that the likelihood that Y takes a value lower than y is some prescribed value. Using the CDF of Y , $F(y) = P(Y \leq y)$, we seek an estimation of the α -quantile y_α defined by $F(y_\alpha) = \alpha$.

In the next sections is a description of how to set minimum sample sizes to achieve given quantities of information. Most of the ideas used come from the theory of order statistics. The references for this section are Cannamela et al. (2007), David and Nagaraja(2003), Guba et

al. (2003), Makai and Pal (2006), Nutt and Wallis (2004 and 2005), Orechwa (2005), Wallis (2003), Wallis (2006) and Wilks (1941).

4.2.1 The empirical estimator

Let $\hat{F}_{EE}(y) = \frac{1}{n} \sum_{i=1}^n 1_{Y_i \leq y}$ be the empirical estimator of the CDF, where $1_{Y_i \leq y}$ is the indicator function (one for $Y_i \leq y$ and zero for $Y_i > y$). This leads to the following estimator of the α -quantile

$$\hat{Y}_{\alpha,n} = \inf\{y, \hat{F}_{EE}(y) > \alpha\} = Y_{(\lceil \alpha n \rceil)}. \quad (4.2.1.1)$$

This means that, if we need to estimate the 95% percentile of Y and we have a sample of size 100, $\lceil \alpha n \rceil = \lceil 0.95 \cdot 100 \rceil = 95$. The estimate would be the sixth largest observation of Y in the sample (the 95th smallest observation). The properties of this estimator are given in annex C. The variance of this estimator is large. Moreover, from the asymptotical law, we obtain $P(\hat{Y}_{\alpha,n} \geq y_\alpha) \approx 0.5$, which comes from the fact that this is an asymptotically unbiased estimator of the α -quantile. So, for sufficiently large sample sizes (we never know if the sample size actually used is large enough), roughly 50% of the times our estimate will be larger and 50% of the times it will be smaller than the actual α -quantile.

Let us assume that we wish to estimate and provide a $100(1-\delta)\%$ confidence interval for the quantile α of a random variable X . Given a sample size n , the point estimate used to estimate that quantile will be the order statistic $Y_{(\lceil \alpha n \rceil)}$, as proposed in (4.2.1.1). The confidence interval sought will be limited by the values of the order statistics $Y_{(r)}$ and $Y_{(s)}$, where $r < \lceil \alpha n \rceil < s$, and the r and s selected are the closest integer numbers that meet the condition

$$\pi(r, s, n, \alpha) = \sum_{j=r}^{s-1} \binom{n}{j} \alpha^j (1-\alpha)^{n-j} \geq 1 - \delta. \quad (4.2.1.2)$$

The way to do it is to start with $r = \lceil \alpha n \rceil$ and $s = \lceil \alpha n \rceil + 1$, and decrease r by one unit and increase s by one unit alternatively until inequality (4.2.1.2) is fulfilled. In the case of $\alpha=0.95$, $\delta=0.95$ and $n=100$, we obtain $r=91$ and $s=100$ (the 95% confidence interval for the 95% percentile is the interval defined by the 10th largest observation and the largest observation). It could happen that the sample size is not large enough to produce the confidence interval with the desired confidence level. This problem may be solved enlarging the sample size or considering a confidence interval with a smaller confidence level (larger δ).

When working in the area of safety, we are usually interested in estimating extreme (high) quantiles, such as the 95%, the 99%, etc. In those cases, it could be advisable to be more

confident that our estimate is above the actual quantile. Then we would be interested in a conservative estimator of the α -quantile like Wilks' introduced below.

4.2.2 Wilks estimator

In order to set a conservative estimator for the quantile of interest (y_α), we should decide how confident we want to be that our estimator exceeds the real quantile. The confidence level β would typically be either 0.95 or 0.99. The estimator considered is the one that fulfils

$$\hat{Y}_{\alpha,n} = Y_{(\lceil \alpha n \rceil + s)}, \text{ such that } P(Y_{(\lceil \alpha n \rceil + s)} \geq y_\alpha) = \beta, \quad (4.2.2.1)$$

where $s \geq 1$. This estimator, based on order statistics, is referred to as Wilks estimator, and is based on the probabilistic distribution of the number of times a sample of the rv exceeds a certain threshold. Let us rename it as $Y_{(n-r+1)}$. For each couple (n,r) we will get a given value $P(Y_{(n-r+1)} > y_\alpha)$, which is (see annex C for more details)

$$\sum_{j=0}^{n-r} \binom{n}{j} \alpha^j (1-\alpha)^{n-j} \quad (4.2.2.2)$$

For a fixed r , we may (numerically) compute the smallest value of n needed to make this expression (4.2.2.2) larger than or equal to β (for fixed values of β and α).

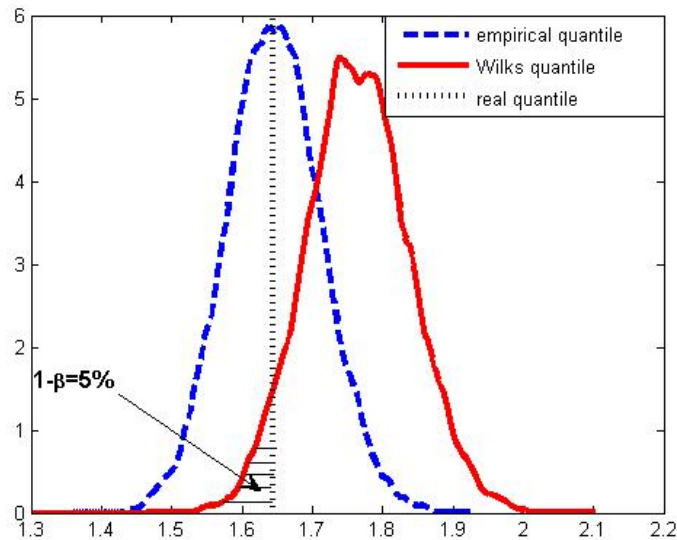


Figure 7.- Comparison between distributions of the empirical and the Wilks estimators, for samples from a normal distribution (the real 95% quantile for the normal distribution is 1.6449). For this example, the variance of the empirical estimator is 0.0045, while for the Wilks estimator is 0.0053.

Interpretation: n is the number of simulations required for the r^{th} largest value of the ordered sequence of outputs to exceed the α -quantile with a prescribed confidence level β .

Example: For $\alpha = \beta = 95\%$, we have the following couples :

$(r = 1, n = 59) ; (r = 2, n = 93) ; (r = 3, n = 124) \dots (r = 39, n = 991)$.

For $r=1$, the previous formula becomes

$$\beta = 1 - \alpha^n. \quad (4.2.2.3)$$

Remark: The variance of this estimator is even larger than the one of the empirical estimator (see figure 7).

In conclusion, the problem of quantile estimation can be solved by Monte Carlo techniques, but the estimates are imprecise (i.e. with large variance) if the number of runs is often considered “reasonable” (100 – 1000 runs).

4.2.3 Tolerance intervals

An alternative solution to set a sample size is to estimate a tolerance interval rather than a percentile, see Guba et al. (2003) and David and Nagaraja (2003). A tolerance interval has random bounds, denoted by L (lower) and U (upper) and the requirement for this interval is that it should contain at least a proportion γ of the population, with probability β (with prescribed γ and β). Hence we seek L and U such that

$$P \left[\int_L^U f(y) dy \geq \gamma \right] = \beta, \quad (4.2.3.1)$$

where f is the (unknown) underlying PDF.

It has been shown (see for instance David and Nagaraja (2003)) that the left hand side of equation (4.2.3.1) is independent of f if and only if the bounds L and U are order statistics (i.e. $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$). To see the necessary condition, let $L = Y_{(r)}$ and $U = Y_{(s)}$, $r < s$ (where $Y_{(0)} = -\infty$ and $Y_{(n+1)} = +\infty$), and then the equation (5.18) may be written as $P[F(Y_{(s)}) - F(Y_{(r)}) \geq \gamma] = \beta$. The quantities $F(Y_{(s)}), F(Y_{(r)})$ are order statistics for a uniform distribution in $[0,1]$. The distribution of the range (here the range is $F(Y_{(s)}) - F(Y_{(r)})$) of order statistics is known for uniform distributions and is given by David and Nagaraja (2003):

$$P[F(Y_{(s)}) - F(Y_{(r)}) \geq \gamma] = 1 - I_\gamma(s-r, n-s+r+1) = \beta, \quad (4.2.3.2)$$

where $I_\gamma(j, k)$ is the incomplete beta function². Equation (4.2.3.1) and (4.2.3.2) are not satisfied exactly, but some values of r and s may be chosen such that

$$P \left[\int_L^U f(y) dy \geq \gamma \right] \geq \beta. \quad (4.2.3.3)$$

Table 1.- Minimum sample sizes (n) for one and two sided tolerance intervals for different values of β and γ .

	Two sided tolerance intervals			One-sided tolerance intervals		
$\beta \setminus \gamma$	0.90	0.95	0.99	0.90	0.95	0.99
0.90	38	77	388	22	45	230
0.95	46	93	473	29	59	299
0.99	64	130	662	44	90	459

Table 2.- Values of γ for the standard tolerance interval for different values of n and of β (from Guba et al. (2003)).

n	γ values		
	$\beta=0.90$	$\beta=0.95$	$\beta=0.99$
10	0,66315	0,60584	0,49565
20	0,81904	0,78389	0,71127
30	0,87643	0,85141	0,79845
40	0,9062	0,88682	0,84528
50	0,92443	0,9086	0,87448
60	0,93671	0,92336	0,89442
70	0,94557	0,93402	0,9089
80	0,95225	0,94207	0,91989
90	0,95747	0,94837	0,92851
100	0,96166	0,95344	0,93554
125	0,96924	0,96262	0,94813
150	0,97432	0,96877	0,95658
175	0,97796	0,97318	0,96268
200	0,98069	0,9765	0,96736
225	0,98282	0,97909	0,97087
250	0,98453	0,98118	0,97375
275	0,98593	0,98287	0,97618
300	0,9871	0,98429	0,97809

² Which is defined by $I_\gamma(j, k) = \int_0^\gamma \frac{u^{j-1}(1-u)^{k-1}}{B(j, k)} du$, $B(j, k) = \frac{(j-1)!(k-1)!}{(j+k-2)!}$.

Application 1: When a minimum sample size is sought, we are focusing our attention on what happens when the sample maximum and minimum are selected to estimate a tolerance interval (standard **two-sided tolerance interval**: $L = Y_{(1)}$, $U = Y_{(n)}$), we obtain from (4.2.3.2) the following value for β :

$$\beta = 1 - \gamma^n - (n-1)(1-\gamma)\gamma^{n-1} . \quad (4.2.3.4)$$

This formula may be used in different manners. The standard way is to set the fraction of the random variable (γ) that we want to capture between the minimum and the maximum values in the sample and the confidence to get it (β); then it is solved numerically for n and the result rounded to the next highest integer. Table 1 provides the solution (minimum sample size) for several values of β and γ . For example, if we wish to get at least 99% of the random variable contained between the sample minimum and the sample maximum with 95% confidence, the sample size must be 473 at least. An alternative way is to fix n and β or n and γ and solve the equation in the third variable. Table 2 provides the least fraction of the random variable contained between the sample minimum and the sample maximum when the sample size and the confidence level are fixed (taken from Guba et al. (2003)). For $\beta=0.95$ and $n=100$, $\gamma=0.953$ is obtained.

Application 2: For the standard **one-sided tolerance interval** case, when $L = Y_{(0)}$, $U = Y_{(n)}$, we obtain the following value for β : $\beta = 1 - \gamma^n$, which is exactly the same result obtained in the case of the Wilks quantile estimator (and again we get $n=59$, for $\beta = 95\%$, $\gamma = 95\%$). Table 2 provides minimum sample sizes for one-sided tolerance intervals given different values of β and γ .

4.3 Graphic tools

Graphic tools are used to show the spread of the values of random variables along their support. The tools considered as most appropriate and consequently included in this report are the empirical cumulative distribution function (ECDF), its complementary curve (ECCDF), the histogram, the estimated probability density function (PDF) and the boxplots.

4.3.1 The ECDF and the ECCDF

The cumulative distribution function (CDF) is defined by $F(x) = P(X \leq x)$ and for a continuous rv it is also equal to $F(x) = \int_{-\infty}^x f(t)dt$, where $f(.)$ is the probability density function. The most important properties of the CDF are:

- It is non-decreasing monotonic function,
- $0 \leq F(x) \leq 1$,

- $P(a \leq X \leq b) = F(b) - F(a)$,
- $F'(x) = f(x)$.

Alternatively, the complementary cumulative distribution function (CCDF), which is equal to $1-F(x)$, may also be used. The use of the CCDF is widespread in the area of nuclear safety in general and specifically in the area of PA since many safety limits and safety criteria are given in terms of exceeding probabilities, which is the kind of information included in CCDFs.

In figure 8 we present the CDFs and the CCDFs of some of the most frequently used distributions, without specifying their parameters, in order to give an idea of their aspects. Whenever different sets of parameters are used the position and the spread of these curves will be different.

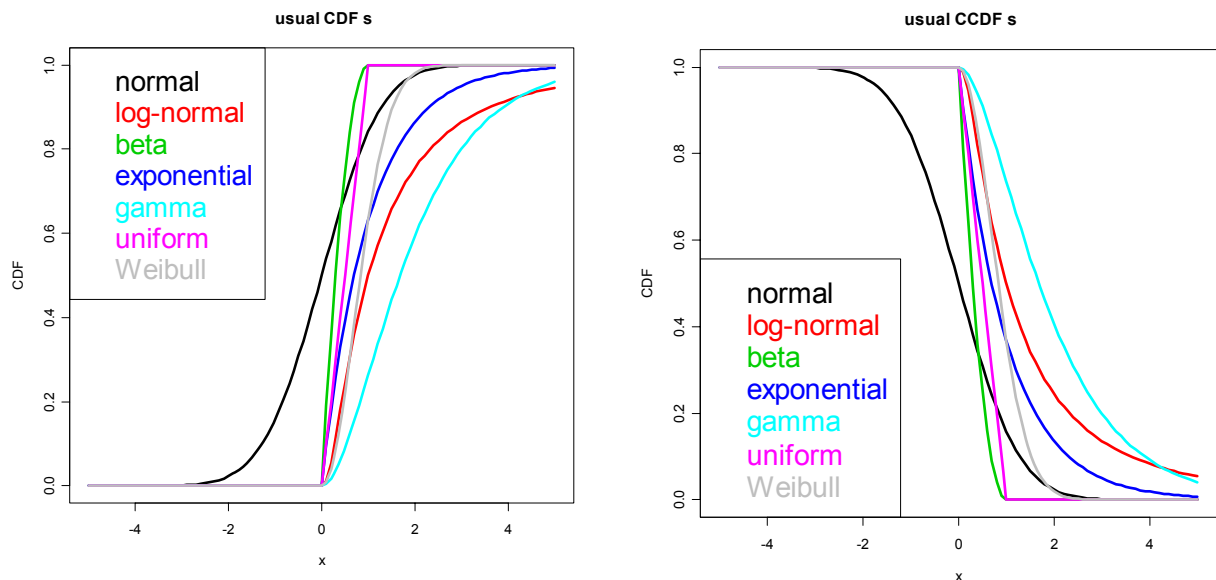


Figure 8.- Some of the most usual CDFs and the corresponding CCDFs.

The empirical cumulative distribution function (ECDF – $F_n(x)$) of a sample is the available tool to estimate the CDF of the corresponding rv, i.e.:

$$x \in IR, F_n(x) = \frac{1}{n} \# \{i, x_i \leq x\} , \quad (4.3.1.1)$$

where the symbol $\#$ denotes the cardinal of a set. The empirical complementary cumulative distribution function (ECCDF) is equal to $1 - F_n(x)$.

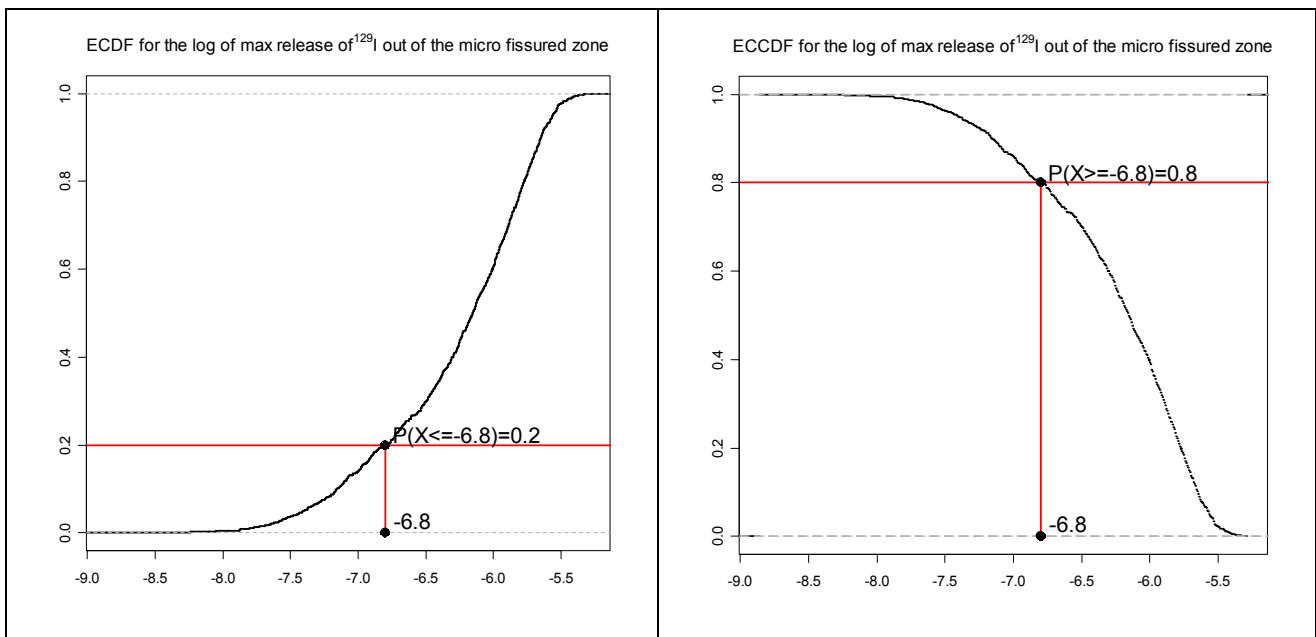


Figure 9.-Example of Empirical CDF and the corresponding ECDF

In figure 9 we represent the ECDF (left) and the ECDF (right) for some data from the benchmark in Prvakova et al. (2008). The sample (of size $n=1000$) represents the decimal logarithms of the peaks of the release of ^{129}I coming out of the disposal cell. It is easy to read directly on this representation that, for example, the percentage of the sample such that the $\log_{10} (^{129}\text{I})$ is less than or equal to -6.8 is around 20%. The same information can be read on the right panel of figure 9: the percentage of the sample such that the $\log_{10} (^{129}\text{I})$ is greater than or equal to -6.8 is around 80%.

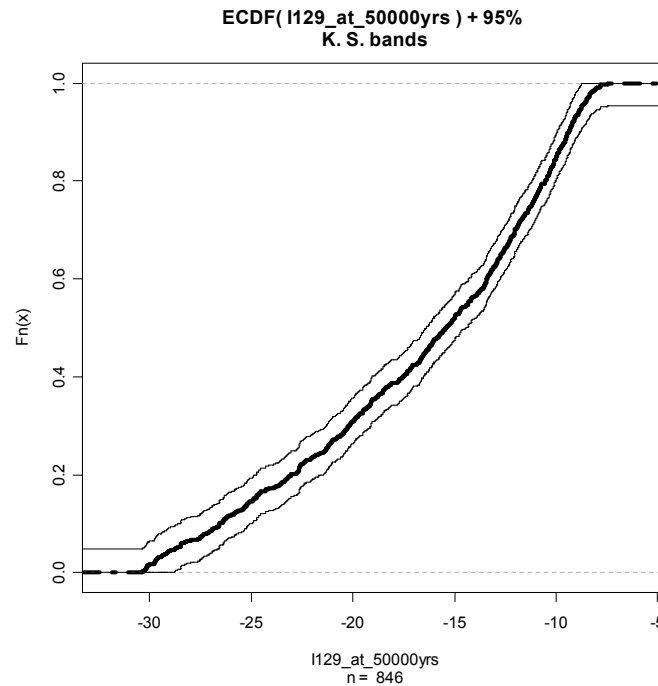


Figure 10.-ECDF for the log10 of the ^{129}I release at 50000 years, at the top and bottom of the repository, together with its 95% confidence bands.

Moreover, Kolmogorov-Smirnov confidence bands may be computed for any ECDF and any ECCDF (for details on Kolmogorov-Smirnov confidence bands see Owen (2001) and Conover (1980)). In figure 10 we present an example of ECDF together with its 95% confidence bands. See Annex B for a correct interpretation of this graphic representation.

4.3.2 The histogram

The histogram graphically summarizes and displays the distribution of a data set. The histogram is constructed by regrouping the data into k bins $C_1 = [a_0, a_1[$, $C_2 = [a_1, a_2[$, ..., $C_k = [a_{k-1}, a_k]$ and then defining the (relative) frequency of each bin

as $f_j = \frac{1}{n} \# \{i, x_i \in C_j\}$. A density is then inferred by a step function whose value for the

C_j bin is the associated frequency per unit length, i.e. $f_j / (a_j - a_{j-1})$. The surface below this step function is equal to 1. However, even if it is possible to define variable bin widths, the use of constant bin width is most popular. In the case of discrete variates two options are available: either using the cardinal of each bin (absolute frequency, see figure 11) or using the relative frequencies. Discrete variates can also be represented as bars. Figure 12 illustrates the importance of the choice of the number of bins (or equivalently of the bins widths): the left side picture is very “noisy”, too many bins have been displayed; on the contrary, the right hand side picture has not enough bins, and much of the information is therefore lost. The only reasonable histogram is the one in the middle, where the corresponding estimated PDF (see section 4.3.3) has been added. The number of segments

should be sufficient to represent the shape of the distribution but not so small so that noise becomes dominant.

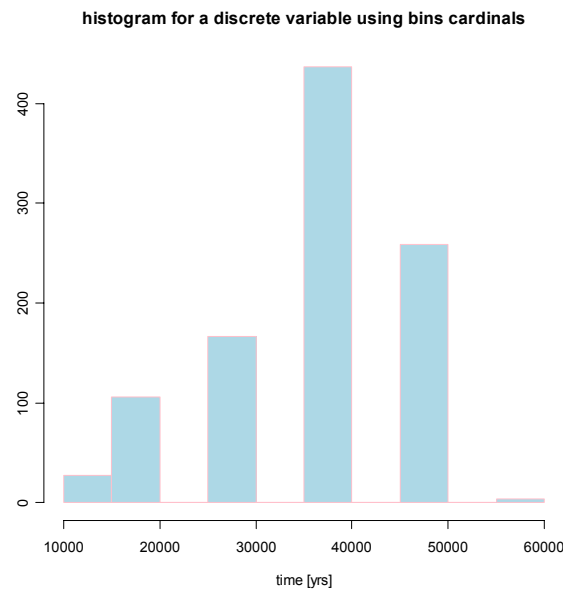


Figure 11.-Histograms for a discrete variable using as ordinate the bins cardinals.

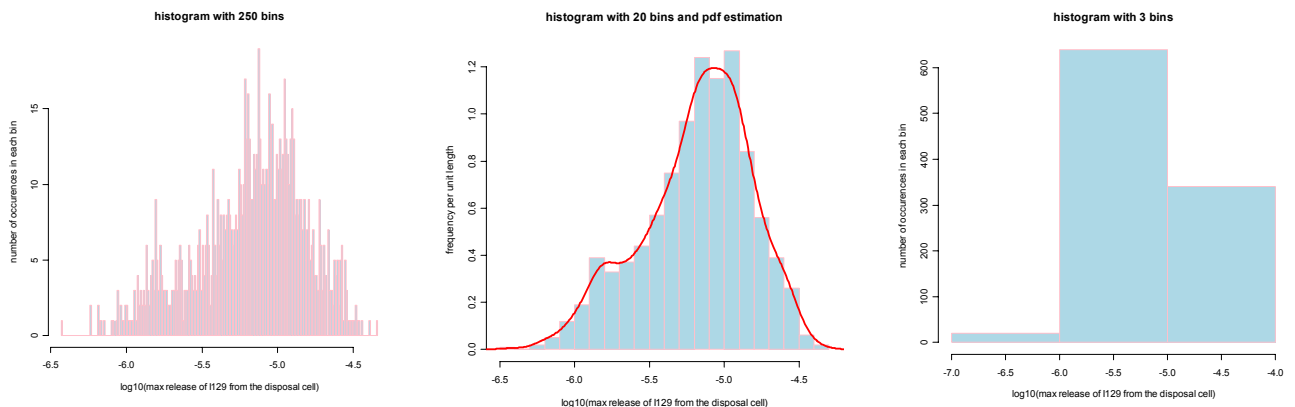


Figure 12.-Histograms for the same data, using different number of bins.

4.3.3 PDF estimation

The PDF gives the probability density of a random variable X at each value x . The integral of the PDF in the interval $[a,b]$ provides the probability that X takes values between a and b :

$$P(a \leq X \leq b) = \int_a^b f(x)dx . \quad (4.3.3.1)$$

If we consider a sample of size n ($x_i, i = 1, \dots, n$) from an unknown continuous probability distribution of density f , the histogram represents an approximation of the PDF f . The main deficiencies of the histogram are its discontinuity and the appropriate choice of the number of bins (or bin widths).

The best available method to estimate probability density function (PDF) is the Kernel method. This is a non-parametric method (because it does not assume a certain probability distribution) generalizing the histogram. The kernel estimator of f , denoted by \hat{f} is a sum of “bumps” of width h placed at the observations x_i :

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad (4.3.3.2)$$

K denotes the kernel. There are several desirable properties:

- positivity : $K \geq 0$,
- regularity : K has to be smooth enough,
- normalization : $\int_{-\infty}^{+\infty} K(x)dx = 1$,
- symmetry : $K(x) = K(-x)$,
- fast decreasing at infinity.

The most used kernels (see figure 13) are:

- Gaussian : $K(x) = 1/\sqrt{2\pi} \exp(-x^2/2)$
- Epanechnikov : $K(x) = \begin{cases} \frac{3}{4}(1-x^2), & \text{if } x \in [-1,1] \\ 0 & \text{otherwise} \end{cases}$
- Rectangular : $K(x) = \begin{cases} \frac{1}{2} & \text{if } x \in [-1,1] \\ 0 & \text{otherwise} \end{cases}$
- others : Triangular, Biweight, Cosine, Optcosine.

It can be seen from figure 13 that density curves are similar for the different Kernels. Thus the kernel is not as important as the choice of bandwidth, h . This scaling parameter (which has the same physical dimension as the sample) controls:

- the width of the probability mass spread around a point
- the smoothness or roughness of a density estimate.

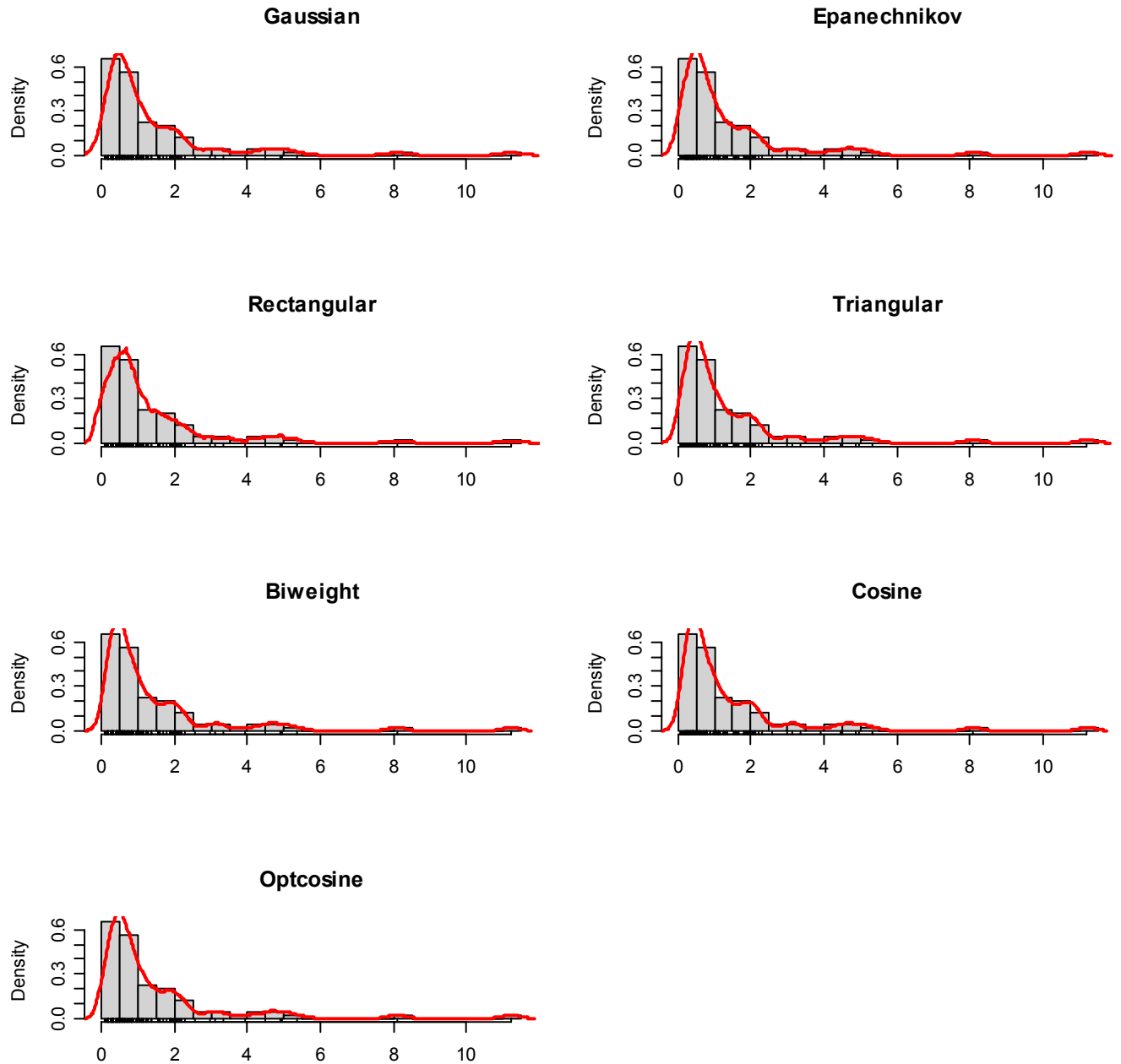


Figure13.-Influence of the kernel (h=1)

If the bandwidth is too small, the estimated density will be under-smoothed; a large value of h , on the other hand, would lead to an over-smoothed estimated density (see figure 14).

An **optimal bandwidth** may be computed for each kernel. The criterion to be minimized is either the Mean Integrated Squared Error (MISE) or the Asymptotic Mean Integrated Squared Error (AMISE). For instance, the optimal bandwidth for the Gaussian Kernel and MISE criterion is:

$$h_{opt} = 1.06 \hat{\sigma} n^{-1/5} \quad (4.3.3.3)$$

where $\hat{\sigma}$ is the empirical standard deviation of the sample, i.e. $\hat{\sigma} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

Unfortunately, the optimal bandwidth is over-smoothing if f is multimodal or somehow “not normal”.

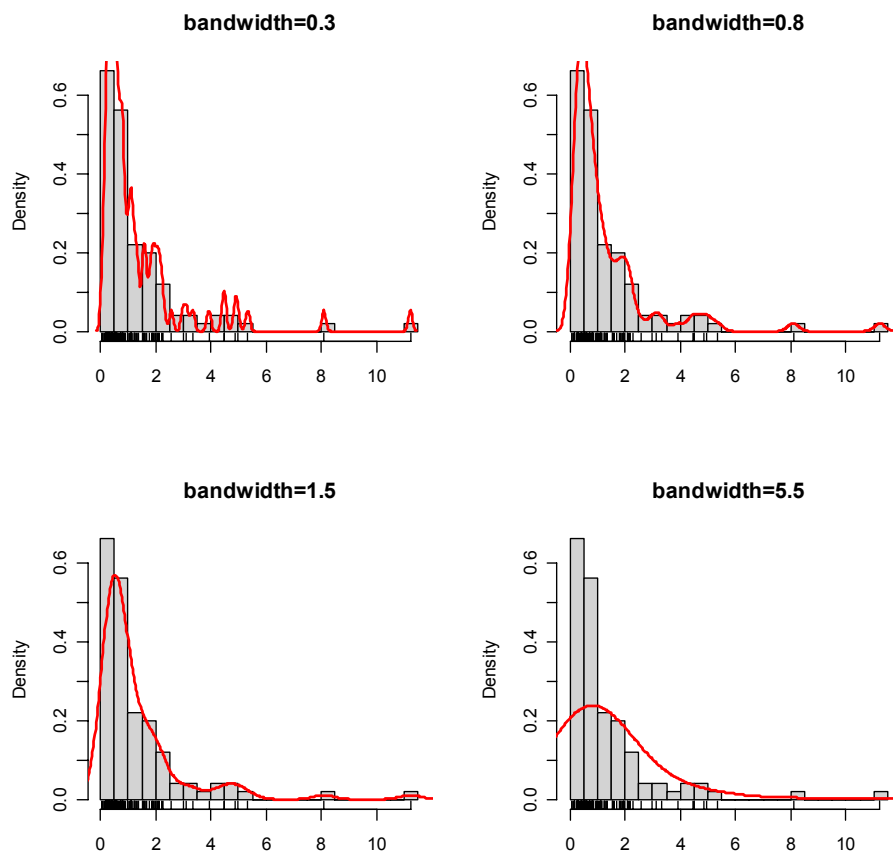


Figure 14.-Bandwidth influence, from under to over-smoothing (Gaussian kernel).

Another option is to use the adaptive kernel method, which consists of varying h with x_i , in order to have a small h where we have a high density of data and a large h where the data is sparse. The algorithm is outlined below.

- define a pilot estimation $\tilde{f}(x)$ (an optimal bandwidth kernel estimation with optimal bandwidth denoted by h_0), such that $\tilde{f}(x_i) > 0$
- compute $\lambda_i = \{\tilde{f}(x_i) / g\}^{-\alpha}$, where $\log(g) = (1/n) \sum \log(\tilde{f}(x_i))$ and $0 \leq \alpha \leq 1$ is a sensitivity parameter (a good choice is $\alpha = 1/2$)
- $$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_0 \lambda_i} K\left(\frac{x - x_i}{h_0 \lambda_i}\right).$$

Silverman (1986) gives further details concerning density estimation.

4.3.4 Boxplots

A boxplot (also known as a box-and-whisker plot) is a way to describe graphically groups of numerical data using five of their summaries (the smallest observation, lower quartile (Q_1), median, upper quartile (Q_3), and largest observation). It also indicates if there are some observations which might be considered outliers. The length of the “box” is the interquartile range ($IQR = Q_3 - Q_1$) and the line inside the box stands for the median. An outlier is any data that lies outside the interval $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$. The bounds of this interval are indicated by some tic marks and are connected to the box by a line.

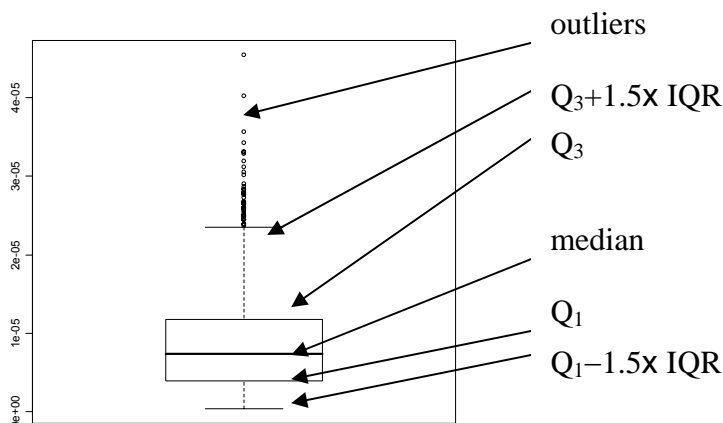


Figure 15.-Example of boxplot for data representing the peak of the release of ^{129}I coming out of the disposal cell; Prváková et al. (2008).

As we can see from figure 16, the boxplots, even if they show less information than histograms or PDFs, are very useful for making comparisons between different distributions; they may even suggest the existence of a second subpopulation instead of outliers (as it is the case for the left boxplot in figure 16).

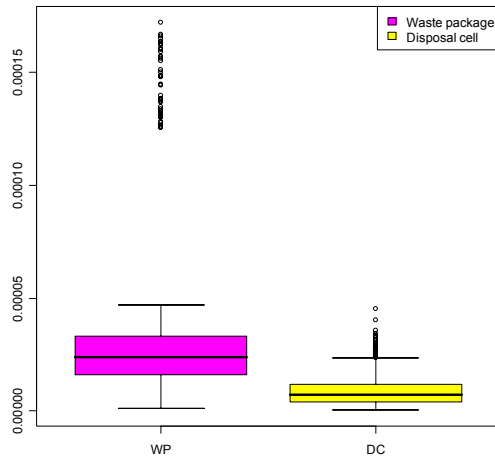


Figure 16.-Comparison of the distributions for the peaks of the release of ^{129}I coming out of the waste package and of the disposal cell by using boxplots; Prváková et al. (2008).

5. Tools for identifying most important input parameters

An important part of the PA is the SA. SA analysis methods may be classified as global, local and screening methods. In the context of a PA, global methods are preferred because they provide valuable information about the relation input-output taking into account the whole input space and the input distributions. In the following paragraphs we make a selection of a minimum set of techniques (numeric and graphic) that should be applied to the output targeted variables. We will discuss why they are selected but we will not describe them since they have already been extensively described in other reports generated within PAMINA, as for example milestones Badea and Bolado (2008) and Plischke et al. (2009) and deliverable Becker et al. (2009).

When analysing the relation between inputs and outputs in a probabilistic framework, several approaches may be adopted, among them the two most frequent are: to identify the (functional) relation between them, or to identify how the different inputs contribute to the variability (variance) of each output. The main numeric techniques available for the first approach are the regression/correlation based techniques and the Monte Carlo filtering techniques. The regression/correlation based techniques allow identifying linear and monotonic relations between inputs and outputs. Monte Carlo filtering techniques allow identifying relations between different regions of inputs and outputs, which are not necessarily either linear or monotonic. Several techniques are available to find out how much each input parameter contributes to the variance of an output variable (Sobol sensitivity indices or just sensitivity indices). Unfortunately, many of them are very expensive in computational terms and need sophisticated input designs that cannot be used simultaneously to perform uncertainty analysis. These facts bring us to consider correlation ratios (CR), also known as 'cheap methods' after Plischke et al. (2009) and Becker et al. (2009), as the most interesting ones to report about contributions to the variance. Among them, the correlation ratios are strongly supported by the authors of this report due to the simplicity of implementation and the good results provided. Additionally, these techniques may be supported by ancillary graphic techniques such as the scatterplots, the cobweb plots and the contribution to the sample mean plots (CSM plots); see Badea and Bolado (2008) for a description of all these techniques. The following set of techniques is proposed to be used to identify important input uncertain parameters:

- Regression/correlation based statistics: Standardised Regression Coefficients (SRCs) and Standardised Rank Regression Coefficients (SRRCs) and their respective coefficients of determination (R^2). It is always convenient transforming also input parameters and output variables appropriately (i.e. logarithmic transformation) for finding out the best possible regression/correlation based model.
- Use of Monte Carlo filtering techniques: Smirnov test and Mann-Whitney test (this one also known as Wilcoxon test). The use of the t test is not advised due to the difficulties to fulfil test hypotheses. Two rules to divide the output sample are proposed:

- Null/non-null observations, and
- 10%/90% rule (10% largest observations and the rest of the observations – 90% smallest observations), though other sensible rules (i.e. 5%/95%) are not excluded.
- Correlation ratios (cheap methods): Available cheap methods based on the computation of correlation ratios to estimate contributions to the variance of the output variable (sensitivity indices) are the Variance of the Conditional Expectation (CR-VCE), the Expectation of the Conditional Variance (CR-ECV), the Polynomial Fit (CR-FIT) and the Conditional Linear Model (CR-CLM). Any of these sensitivity techniques can be used to compute at least first order sensitivity indices. Though the study of the contribution to the variance of second and higher order interactions is convenient, restrictions imposed by the sample size used can make this either impossible or inaccurate. Moreover, correlation ratios may also be used to group contributions to the variance according to different criteria, as for example radionuclide, barrier, etc.
- Graphic tools: The different capability to represent information about several input parameters in the same plot encourages us to use more frequently CSM plots than any other type of plots. Cobweb plots are very useful to support numeric information obtained via Monte Carlo filtering techniques. Scatterplots should only be used to stress some remarkable effect detected via numeric SA techniques.

The best option to present results from numeric SA techniques is the table of results. It is also strongly advised to write together with the value of the statistic the rank of importance of each input parameter (1 for the most important, 2 for the second most important and so on), specially in the case of Monte Carlo filtering techniques, whose results are p-values from statistical tests, whose interpretation is not so obvious for persons lacking a good statistical background. Nevertheless, in the case of sensitivity indices, the fact that they are fractions of a total quantity (the variance), graphic representations such as pie charts may also be used.

6. Time dependent output variables

Most of the output variables considered in a PA are time-dependent. This obliges us to provide tools that allow showing the evolution of their uncertainty and the evolution of the different sensitivity indices over time. The procedure to do this consists in computing the same statistics used for non-time dependent output variables at each time considered in the simulation and to draw the results versus time in a picture. This may be done for all numeric statistics considered in sections 4 and 5. The use of linear or logarithmic scales in any of both axes depends on the effects that the analyst wants to stress.

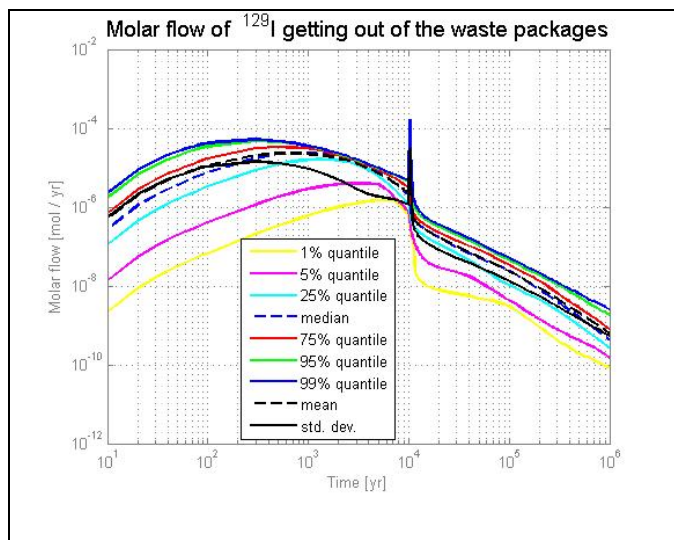


Figure 17.-Evolution over time of the main statistical indicators for the flow of ^{129}I getting out of the waste packages.

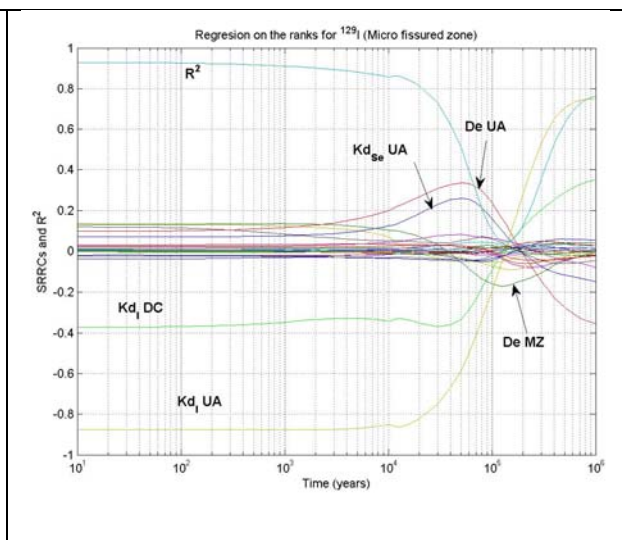


Figure 18.-Evolution over time of the SRRCs and R^2 for the flow of ^{129}I getting out of the micro-fissured zone.

Figures 17 and 18 show the results obtained for ^{129}I in the reference problem in two different compartments (the waste packages and the micro-fissured zone respectively) in Prváková et al. (2008). Figure 17 shows the most relevant statistics characterising the distribution of the ^{129}I molar flow getting out of the waste packages over time. We can see at 10^4 y the effect of the sudden degradation of concrete properties. Moreover, it can also be seen the decrease in the uncertainty (orders of magnitude) over time; at early times the results spread over more than three orders of magnitude, while at late times they are spread over less than two orders of magnitude. It could be convenient to put together with this picture the corresponding numeric statistics at some selected times. Figure 18 shows the results of the regression-based analysis for the ^{129}I molar flow getting out of the micro-fissured zone over time. It can be seen the change of sign of the SRRC associated to the $K_{d,i}UA$ (most important parameter all over time according to this SA technique), which produces a large drop in the



value of R^2 around 10^5 y. The effect of the concrete degradation on these SA statistics can also be seen. A lot of useful information about the system behaviour over time may be obtained from these types of graphic representations, mostly if UA and SA are used in a combined manner.

7. A template to present PA results

The objective of creating a template to present PA results is to show, in a systematic way, main results concerning the uncertainty about output variables and identify most important input parameters or sets of input parameters. In order to create this template we will classify output variables as either time-dependent or as non time-dependent. For each output variable a minimum set of statistics will be suggested, while other statistics will be considered as optional, and this will be done both for the uncertainty analysis and for the sensitivity analysis. Though candidate variables were mentioned in chapter 3, the selection of the output variables to study is not within the scope of this report; this is left completely to the election of the organisation interested in developing the study. The study of 'derived input parameters' (combinations of input parameters that have a stronger influence on the output than the original sampled parameters, see Cormenzana et al. (2009) and Bolado et al. (2009)) is also left to be decided by the organisation interested in the study.

7.1 Non time-dependent output variables

Among these output variables we will consider any variable that takes only one value per realisation. This includes peaks and time to peaks among others. Firstly we will consider the statistics needed to do the uncertainty analysis, then the ones related to the sensitivity analysis.

7.1.1 Uncertainty analysis

Quantitative indicators (presented in tables):

- Suggested: Mean, standard deviation, selection of quantiles (at least 1% and/or 5%, median, 95% and/or 99%), skewness coefficient and kurtosis.
- Optional: Variance, mode, other quantiles and derived quantities (i.e. minimum, 25%, 75%, other alternative quantiles, maximum, interquartile range and range), geometric mean, geometric standard deviation, and skewness coefficient and kurtosis for the logarithm of the variable. Confidence intervals for specific quantiles as required.

Graphic indicators:

- Suggested:
 - ECDF.
 - In case the peak of a time-dependent output variable and the time to the peak are considered in the study, it is also suggested to provide the scatterplot of both output variables.
- Optional: ECCDF, PDF and boxplots.

7.1.2 Sensitivity analysis

Quantitative indicators (presented in tables):

- Suggested:
 - SRCs and R^2 (both for raw values and for the logarithm of the values), SRRCs and corresponding R^2 .
 - First order sensitivity indices calculated via correlation ratios (presented as tables or as pie charts).
 - Smirnov and/or Mann-Whitney statistics based on any meaningful rule to divide the output sample (null/non-null observations, 10%/90%, etc.). Provide preferably the p-value of the associated test instead of the statistic itself. Use the value of the statistic only if the sample size is so large that several p_values become 0.0.
- Optional:
 - Second or higher order sensitivity indices calculated via correlation ratios (presented as tables or as pie charts), if needed and if the sample size is large enough.
 - Statistic associated to the CSM plot (maximum distance to the diagonal), as potential support to first order sensitivity indices.
 - Any other SA technique applicable to a sample obtained via a non-biasing sampling technique.

Graphic indicators:

- Suggested:
 - Contribution to the sample mean plot.
 - Cobweb plot as support to Monte Carlo filtering statistics.
- Optional:
 - Scatterplots. Only to highlight or support specific findings of numeric indicators and selecting conveniently the scales (linear, log, ranks).

7.2 Time-dependent output variables

Among these variables we also consider time dependent variables at selected times.

7.2.1 Uncertainty analysis

Quantitative indicators (presented in tables) for selected times:

- Quantitative indicators for selected times are optional:

- Mean, standard deviation, selection of quantiles (at least 1% and/or 5%, median, 95% and/or 99%), skewness coefficient and kurtosis.
- Variance, mode, other quantiles and derived quantities (i.e. minimum, 25%, 75%, other alternative quantiles, maximum, interquartile range and range), geometric mean, geometric standard deviation, and skewness coefficient and kurtosis for the logarithm of the variable. Confidence intervals for specific quantiles as required.

Graphic indicators at selected times:

- Graphic indicators at selected times are optional:
 - ECDF, ECCDF, PDF and boxplots.

Graphic indicators (presented in pictures: evolution of quantitative indicators over time):

- Suggested:
 - Mean, standard deviation, selection of quantiles (at least 1% and/or 5%, median, 95% and/or 99%), skewness coefficient and kurtosis.
 - Set of all individual runs.
- Optional:
 - Variance, mode, other quantiles and derived quantities (i.e. minimum, 25%, 75%, other alternative quantiles, maximum, interquartile range and range), geometric mean, geometric standard deviation, and skewness coefficient and kurtosis for the logarithm of the variable. Confidence intervals for specific quantiles as required.

7.2.2 Sensitivity analysis

Quantitative indicators (presented in tables) for selected times:

- Quantitative indicators for selected times are optional:
 - SRCs and R^2 (both for raw values and for the logarithm of the values), SRRCs and corresponding R^2 .
 - First order sensitivity indices calculated via correlation ratios (presented as tables or as pie charts).
 - Smirnov and Mann-Whitney statistics based on any meaningful rule to divide the output sample (null/non-null observations, 10%/90%, etc.). Provide preferably the P-value of the associated test instead of the statistic itself. Use the value of the statistic only if the sample size is so large that several p_values become 0.0.

- Second or higher order sensitivity indices calculated via correlation ratios (presented as tables or as pie charts), if needed and if the sample size is large enough.
- Statistic associated to the CSM plot (maximum distance to the diagonal), as potential support to first order sensitivity indices.
- Any other SA technique applicable to a sample obtained via a non-biasing sampling technique.

Graphic indicators at selected times:

- Graphic indicators at selected times are optional:
 - Contribution to the sample mean plots.
 - Cobweb plots as support to Monte Carlo filtering statistics.
 - Scatterplots. Only to highlight or support specific findings of numeric indicators and selecting conveniently the scales (linear, log, ranks).

Graphic indicators (presented in pictures: evolution of quantitative indicators over time):

- Suggested:
 - SRCs and R^2 (both for raw values and for the logarithm of the values), SRRCs and corresponding R^2 .
 - First order sensitivity indices calculated via correlation ratios.
- Optional:
 - Smirnov and Mann-Whitney statistics based on any meaningful rule to divide the output sample (null/non-null observations, 10%/90%, etc.). Provide preferably the P-value of the associated test instead of the statistic itself. Use the value of the statistic only if the sample size is so large that several P_values become 0.0.
 - Second or higher order sensitivity indices calculated via correlation ratios, if needed and if the sample size is large enough.
 - Statistic associated to the CSM plot (maximum distance to the diagonal), as potential support to first order sensitivity indices.
 - Any other SA technique applicable to a sample obtained via a non-biasing sampling technique.

7.3 Remarks about suggested and optional statistical indicators

Along this section, the reader could get the impression that the selection of a statistical indicator as suggested or optional is a little arbitrary. Certainly this is not the intention of the

authors of this document. The general rule to select a statistical indicator as suggested is providing essential information, while optional indicators are those that provide complementary information. In some cases, the reason for considering some indicators as optional is to avoid providing too much either overlapped or redundant information, in order to keep a moderate report size. In the opinion of the authors of this document, suggested statistics should always be provided, while the analysts have to decide what optional statistics have to be included in order to get some additional relevant pieces of information about the output variables and the system under study.

For example, if the analyst wishes to show the main characteristics of a non time-dependent output variable, the mean, the standard deviation, the five basic suggested quantiles (1%, 5%, median, 95% and 99%), together with the skewness coefficient and the kurtosis, are enough. The basic central tendency, dispersion and shape characteristics are summarised with these statistics, keeping the quantity of data shown still moderate. Providing additional statistics will not provide significant additional information at the expenses of creating too crowded tables. A graphic representation is always desirable. All the information contained in the sample is given by the empirical cumulative distribution function (ECDF). The empirical complementary cumulative distribution function (ECCDF) provides the same information. The estimated probability density function (PDF) includes some data modelling, and the result of the estimated curve depends on the smoothing model parameters. The boxplot is useful to identify data that differ from the data set bulk, but hides many distribution details. This is why the ECDF is suggested, while the others are considered as optional and should be used only to highlight some important feature.

In the case of time dependent output variables, uncertainty and sensitivity statistics can be reported as plots (uncertainty or sensitivity measure vs. time), but they can also be reported as tables at selected times. Plots are always suggested (evolution of all runs vs. time, evolution of some specific uncertainty statistics –mean, standard deviation and specific quantiles- vs. time, and evolution of some sensitivity indices –regression statistics and Sobol sensitivity indices- vs. time), because they summarize information in an optimal way. Nevertheless, when the natural scale of the ordinate axis (y axis) is logarithmic, providing the information in tables for specific times may also be convenient, due to the difficulty of estimating a value given in such kind of scale. This is the case, for example, of means, standard deviations and quantiles evolving over time. Providing those tables for regression statistics or Sobol sensitivity indices does not help so much because the values associated to important parameters may easily be read from in plots (a linear scale in the y-axis is adequate).

In general, regarding plots for sensitivity analysis of non time-dependent output variables, the use of cobweb plots and contribution to the sample mean plots (CSM plots) is encouraged, while the systematic use of scatterplots is discouraged always (not only for non time-dependent output variables). The reason for this selection is the possibility of including in one single plot the information corresponding to several input parameters in the case of the cobweb plots and the CSM plots, which cannot be done with scatterplots. The use of



sensitivity plots for time-dependent output variables is always considered optional due to the large space that this would occupy in a report (one plot per time considered). In such a case a careful selection of times should be done in order to generate a moderate size report. Scatterplots should only be used to highlight specific relevant selected effects.

8. Template application

The template designed in section 7 has been applied to the dose rates due to ^{129}I for the Spanish reference concept in granite; see ENRESA (2000). This reference concept is based on the disposal of spent fuel in carbon steel canisters in 500 m long galleries at a depth of 500m. Canisters are surrounded by high-density bentonite.

The expected behaviour of this type of repository is as follows. Water reaches the disposal drifts via small fractures, and saturates bentonite in a few decades. Minimum container lifetime due to anaerobic general corrosion is 20,000 years, although in the evaluation it is assumed that a few (up to 10) canisters fail much earlier due to a fabrication defect. After canister failure, and since no credit is given to the cladding as a barrier, there is an instantaneous release of some volatile radionuclides, such as ^{129}I , ^{36}Cl and ^{135}Cs . The gradual release of the radionuclides in the UO_2 matrix starts when groundwater reaches the waste.

The radionuclides released from the waste dissolve or precipitate in the water in the canister cavity, depending on their solubility limits. Dissolved radionuclides are transported through the bentonite buffer by diffusion and pass to the groundwater flow in the near field close to the disposal drift. Both sorption and anion exclusion in the bentonite are considered. Radionuclide transport in the granite is controlled by fractures, and in the calculations the geosphere is represented by a single one-dimensional planar fracture. Longitudinal dispersion and matrix diffusion into the wall rock are modelled, including sorption onto the rock matrix and anion exclusion.

The radionuclides that cross the geosphere are discharged to a river used by the critical group to produce most of its aliments. The dose to an average member of this critical group is used as the main indicator of the safety of the repository.

The model has 135 random independent input parameters and considers 29 chemical elements. No radionuclide is affected by more than 21 random parameters: 3 related to the canister failure, 3 to the release from the waste, 6 to the transport in the near field and 9 to the transport in the far field. Biosphere parameters are considered constant. As mentioned at the beginning of this section, this application is done only for the dose rate in biosphere due to ^{129}I . Only 18 input parameters affect the results obtained for this radionuclide. Their names and short descriptions are reported in table 3.

The peak of the dose rate due to ^{129}I and the time to the peak dose rate will be considered non time-dependent output variables in this application, while the dose rate due to ^{129}I at all times and at 6 specific times (3.0E+4, 1.0E+5, 3.0E+5, 1.0E+6, 3.0E+6 and 1.0E+7 a) will be time-dependent output variables. Sections 8.1 and 8.2 contain the implementation of the template for these output variables. All statistics (numeric and graphic) classified as 'suggested' in chapter 7 have been included in this implementation, together with a few that

were classified as 'optional'. The design of tables of results and the location of tables and pictures in pages has been done with the aim of optimising the space used.

Table 3- List of input random parameters (18) affecting Iodine (I) in the application case.

Input parameter	Short description
Waste (2)	
R1000	UO ₂ matrix alteration rate after 1000 years of cooling
GAPI	Fraction of the inventory of Iodine released immediately when the canister fails – Instant Release Fraction (IRF)
Canisters (3)	
FAIL	Number of canisters with initial defects that fail in a century approximately
MEAN	Mean duration of the canisters due to generalised corrosion
SLOPE	Parameter for the Weibull distribution of canister failures
Near field (5)	
VC1	Volume of water in the canister cavity
POR ANIONS BUFFER	Fraction of bentonite porosity accessible to anions
WATER NF	Groundwater flow in the granite of the near field
KDBEI	Distribution coefficient bentonite-water for Iodine
DIFFI	Diffusion coefficient in bentonite porewater (Dp) for Iodine
Water flow (4)	
WATER TT0	Water travel time for a kinematic porosity of 10 ⁻⁴
KINEM POR	Kinematic porosity of the granite
PECLET	Peclet number (longitudinal dispersion)
PATH LENGTH	Length of the geosphere pathway, from repository to biosphere
Granite matrix (4)	
FWA	Flow wetted area
THICKM	Thickness of the granite matrix
PORMAT	Porosity of the granite matrix
POR ANIONS GRANITE	Fraction of granite matrix porosity accessible to anions

8.1 Non time-dependent output variables (Peak annual dose rate in biosphere due to ^{129}I and time to the peak)

8.1.1 Uncertainty analysis results.

Table 4.-Uncertainty statistics for the Peak annual dose rate in biosphere due to ^{129}I and for the time to the peak.

	Mean	Std. dev.	Quantile 1%	Quantile 5%	Median	Quantile 95%	Quantile 99%	Skewness coefficient	Kurtosis
Peak	1.09E-06	7.01E-07	2.03E-07	2.90E-07	9.23E-07	2.56E-06	3.37E-06	1.22E+00	4.32E+00
Time to peak	2.34E+05	1.32E+05	8.04E+04	9.89E+04	1.98E+05	4.82E+05	7.30E+05	2.04E+00	1.00E+01

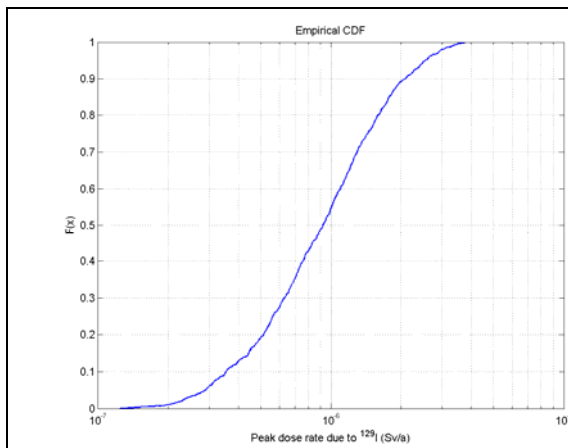


Figure 19.-ECDF for the Peak annual dose rate in biosphere due to ^{129}I .

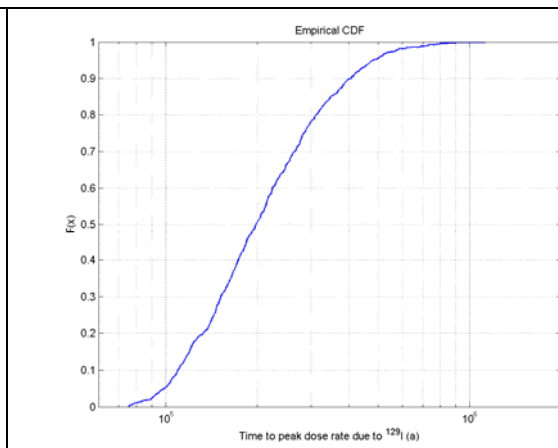


Figure 20.-ECDF for the Time to the peak.

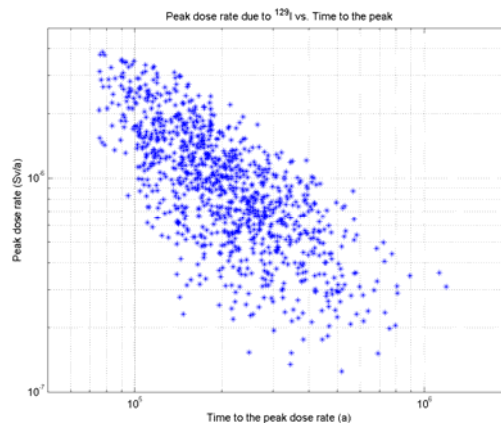


Figure 21.-Scatterplot. Peak annual dose rate in biosphere due to ^{129}I vs. Time to the peak.

8.1.2 Sensitivity analysis results

8.1.2.1 Sensitivity analysis results for the Peak annual dose rate in biosphere due to ^{129}I

Table 5.-Regression analysis, Sobol sensitivity indices computed via CR (CR-VCE) and non-parametric statistics results for the Peak annual dose rate in biosphere due to ^{129}I .

Input parameter	Regression analysis			Correlation ratios	Non-parametric statistics Rule 10%/90%	
	Raw values	Logarithmic transform.	Ranks			
	SRC	SRC	SRRC	1 st order S. indices	Mann-Whitney*	Smirnov*
DIFFI	-2.22E-02	-1.90E-02	-2.56E-02	3.11E-03		
PORMAT	-3.37E-02	5.87E-03	-8.86E-03	3.84E-03		
KDBEI	-2.17E-01	-2.36E-01	-2.49E-01	5.16E-02	4.35E-04 ⁽⁶⁾	8.88E-03 ⁽⁶⁾
POR_AN_BUF	-1.57E-03	-1.18E-02	2.21E-03	1.91E-03		
POR_AN_GRA	1.76E-02	-1.50E-02	-2.61E-06	4.02E-03		
R1000	1.11E-01	1.70E-01	1.46E-01	4.92E-03		
VC1	6.25E-03	-5.92E-04	-5.85E-03	2.38E-03		
WATER_NF	4.43E-01	5.91E-01	5.75E-01	2.61E-01	0.0 ⁽¹⁾	1.29E-16 ⁽¹⁾
FAIL	-2.61E-02	-1.27E-02	-1.45E-02	4.61E-03		
SLOPE	2.93E-02	6.84E-03	1.52E-02	2.44E-03		
MEAN	-6.14E-02	-4.25E-02	-3.88E-02	5.86E-03		
GAPI	4.09E-01	4.16E-01	4.02E-01	1.47E-01	0.0 ⁽²⁾	9.46E-15 ⁽²⁾
WATER_TT0	-3.35E-01	-2.94E-01	-2.97E-01	1.12E-01	1.77E-15 ⁽³⁾	5.35E-13 ⁽³⁾
FWA	-4.06E-02	-4.06E-02	-3.25E-02	3.36E-03		
KINEM_POR	-3.37E-01	-3.63E-01	-3.80E-01	1.42E-01	2.83E-10 ⁽⁴⁾	7.13E-08 ⁽⁴⁾
PATH LENG	3.55E-02	1.67E-02	4.03E-02	4.66E-03		
PECLET	1.36E-01	1.49E-01	1.65E-01	4.74E-02	3.10E-04 ⁽⁵⁾	6.21E-03 ⁽⁵⁾
THICKM	-6.12E-02	-3.20E-02	-4.91E-02	2.12E-03		
R ² / σ^2 fraction	6.35E-01	8.00E-01	7.87E-01	8.05E-01	NA	NA

The number between brackets stands for the importance order. 1 stands for the most important input. NA stands for Not Applicable.

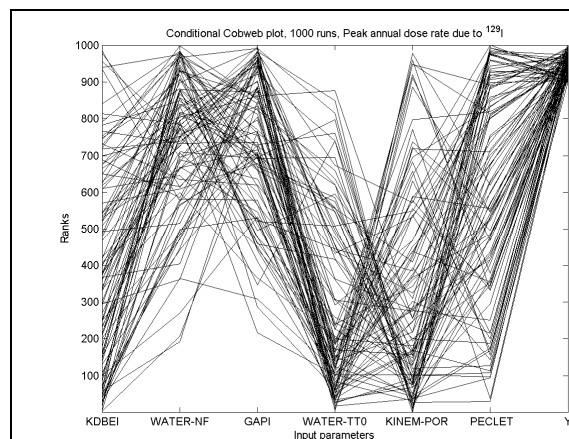


Figure 22.-Cobweb plot for the Peak annual dose rate in biosphere due to ^{129}I .

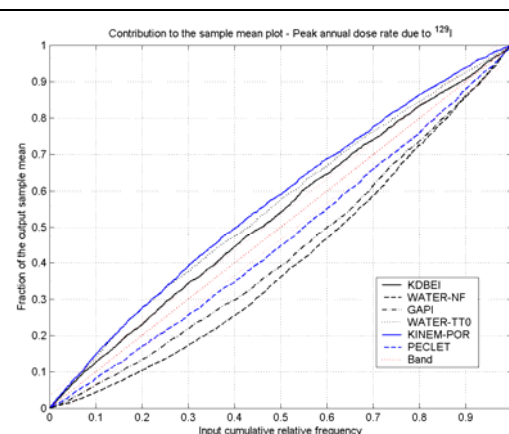


Figure 23.-CSM plot for the Peak annual dose rate in biosphere due to ^{129}I .

8.1.2.2 Sensitivity analysis results for the Time to the peak annual dose rate in biosphere due to ^{129}I

Table 6.-Regression analysis results, Sobol sensitivity indices computed via CR (CR-VCE) and non-parametric statistics results for the time to the peak.

Input parameter	Regression analysis			Correlation ratios	Non-parametric statistics Rule 10%/90%	
	Raw values	Logarithmic transform.	Ranks			
	SRC	SRC	SRRC	1 st order sens. indices	Mann-Whitney	Smirnov
DIFFI	1.69E-02	7.65E-03	1.42E-02	8.29E-04		
PORMAT	3.43E-02	8.72E-03	2.61E-02	6.52E-03		
KDBEI	1.65E-01	1.61E-01	1.60E-01	1.90E-02	8.66E-04 ⁽⁴⁾	3.98E-03 ⁽⁴⁾
POR AN BUF	-1.03E-02	-3.76E-03	-1.53E-02	1.61E-03		
POR AN GRA	1.53E-02	2.94E-02	7.42E-03	1.06E-03		
R1000	8.57E-02	7.11E-02	6.49E-02	1.37E-02	9.09E-03 ⁽⁵⁾	
VC1	5.20E-03	-9.33E-04	6.17E-03	2.44E-04		
WATER_NF	-2.50E-01	-3.77E-01	-3.55E-01	1.42E-01	2.45E-09 ⁽³⁾	8.93E-10 ⁽³⁾
FAIL	1.59E-02	2.92E-03	1.92E-02	3.86E-03		
SLOPE	-2.77E-03	-1.91E-03	-1.18E-02	1.04E-03		
MEAN	9.05E-02	9.44E-02	1.04E-01	9.51E-03		
GAPI	-6.70E-02	-5.98E-02	-4.96E-02	1.56E-03		
WATER_TT0	4.91E-01	6.09E-01	5.92E-01	2.31E-01	0.0 ⁽²⁾	4.54E-13 ⁽²⁾
FWA	1.05E-01	8.94E-02	7.52E-02	1.05E-02		
KINEM POR	6.09E-01	5.68E-01	5.64E-01	3.49E-01	0.0 ⁽¹⁾	9.23E-24 ⁽¹⁾
PATH LENG	1.33E-02	5.82E-03	-3.12E-02	2.98E-03		
PECLET	6.62E-02	7.25E-02	5.77E-02	1.18E-03		
THICKM	5.82E-02	4.41E-02	6.08E-02	3.65E-03		
R ² / σ^2 fraction	7.31E-01	8.82E-01	8.40E-01	7.99E-01	NA	NA

The number between brackets stands for the importance order. 1 stands for the most important input. NA stands for Not Applicable.

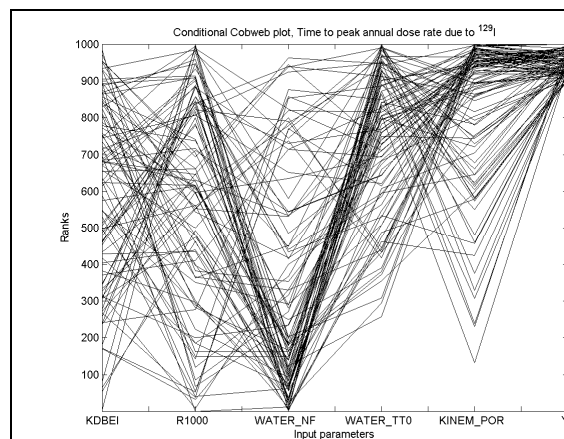


Figure 24.-Cobweb plot for the time to the Peak annual dose rate in biosphere due to ^{129}I .

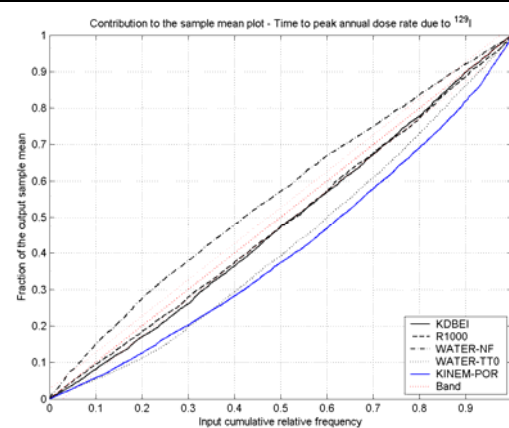


Figure 25.-CSM plot for the Time to the Peak annual dose rate in biosphere due to ^{129}I .

8.2 Time-dependent output variables (Annual dose rate in biosphere due to ^{129}I at all times and at six selected times)

8.2.1 Uncertainty analysis results

8.2.1.1 Uncertainty analysis results (all times)

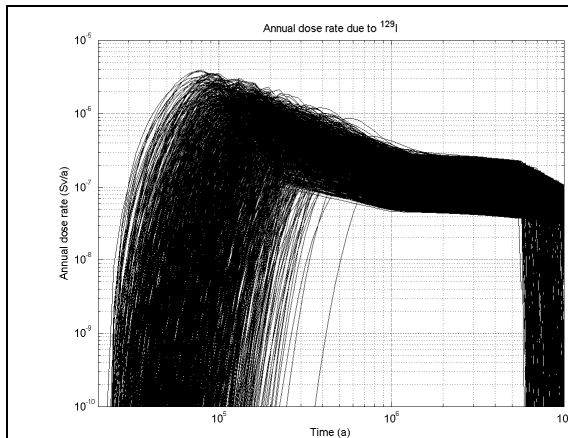


Figure 26.-Evolution of 1000 runs of the Annual dose rate in biosphere due to ^{129}I .

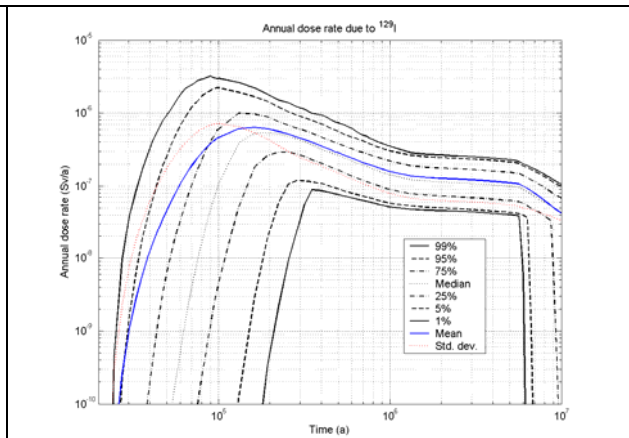


Figure 27.-Evolution of main uncertainty statistics for the Annual dose rate in biosphere due to ^{129}I .

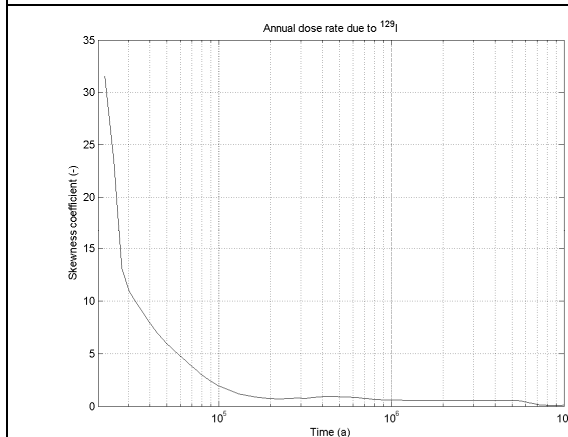


Figure 28.-Evolution of the Annual dose rate in biosphere due to ^{129}I Skewness coefficient.

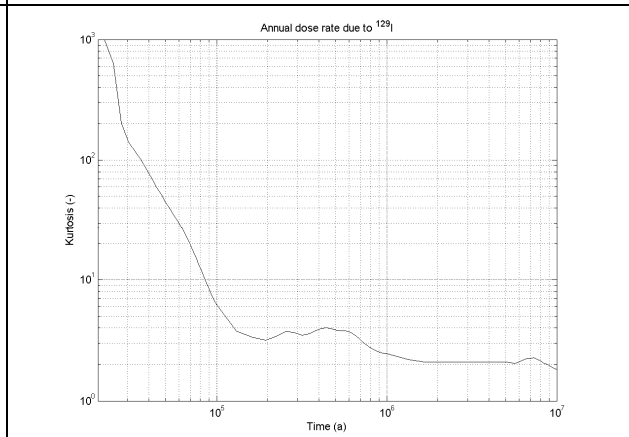


Figure 29.-Evolution of the Annual dose rate in biosphere due to ^{129}I kurtosis.

8.2.1.2 Uncertainty analysis results (6 selected times: 3.0E+4, 1.0E+5, 3.0E+5, 1.0E+6, 3.0E+6 and 1.0E+7 a)

Table 7.-Uncertainty statistics for the Annual dose rate in biosphere due to ^{129}I at six selected times.

Time (a)	Mean	Std. dev.	Quantile 1%	Quantile 5%	Median	Quantile 95%	Quantile 99%	Skewness coefficient	Kurtosis
3·10 ⁴	1.08E-09	7.78E-09	2.63E-27	1.43E-24	2.49E-20	1.62E-09	3.89E-08	1.09E+01	1.40E+02
10 ⁵	4.62E-07	7.17E-07	3.72E-17	3.77E-13	1.05E-07	2.22E-06	3.00E-06	1.94E+00	6.21E+00
3·10 ⁵	4.30E-07	2.32E-07	5.63E-08	1.18E-07	3.98E-07	8.64E-07	1.12E-06	7.63E-01	3.49E+00
10 ⁶	1.58E-07	7.92E-08	5.15E-08	5.79E-08	1.42E-07	3.03E-07	3.53E-07	6.02E-01	2.45E+00
3·10 ⁶	1.21E-07	6.03E-08	4.41E-08	4.74E-08	1.08E-07	2.34E-07	2.52E-07	5.41E-01	2.09E+00
10 ⁷	4.20E-08	3.37E-08	0.00E+00	0.00E+00	4.33E-08	9.84E-08	1.05E-07	1.01E-01	1.82E+00

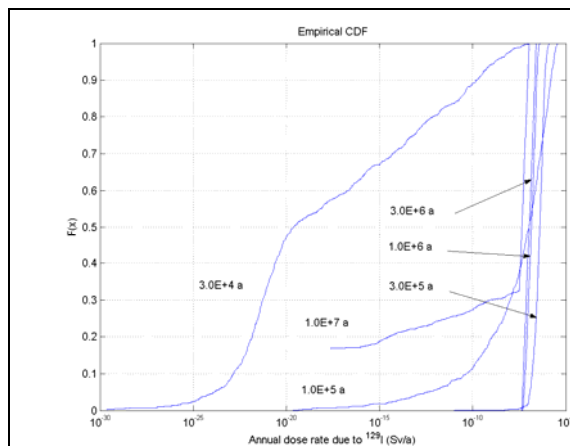


Figure 30.-ECDF for the Annual dose rate in biosphere due to ^{129}I at six different times.

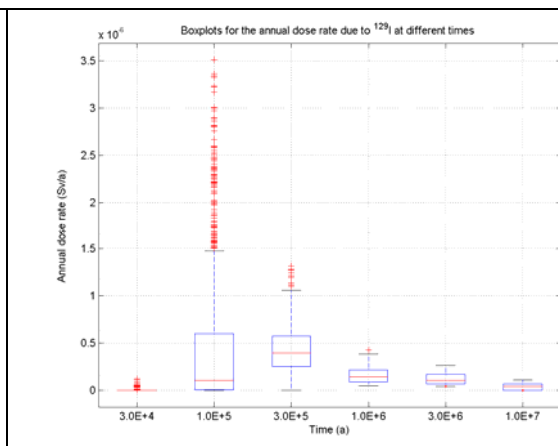


Figure 31.-Boxplots for the Annual dose rate in biosphere due to ^{129}I at six different times.

8.2.2 Sensitivity analysis results (all times)

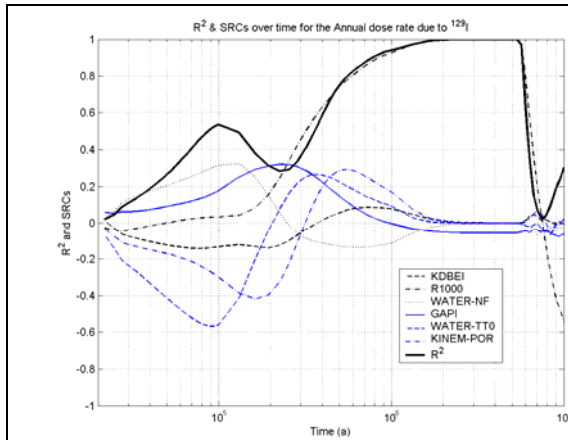


Figure 32.-Regression analysis results* (raw values) for the Annual dose rate in biosphere due to ^{129}I .

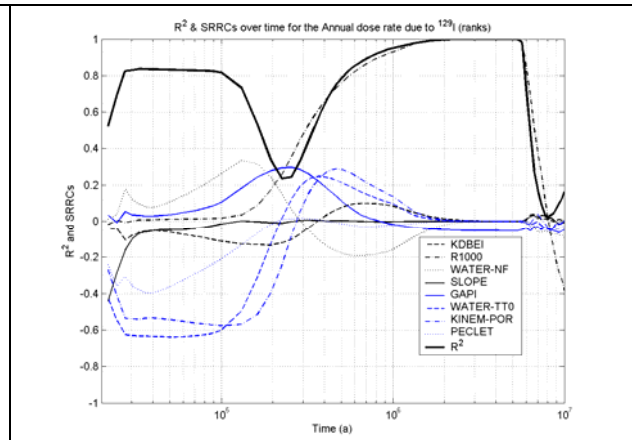


Figure 33.-Regression analysis results* (ranks) for the Annual dose rate in biosphere due to ^{129}I .

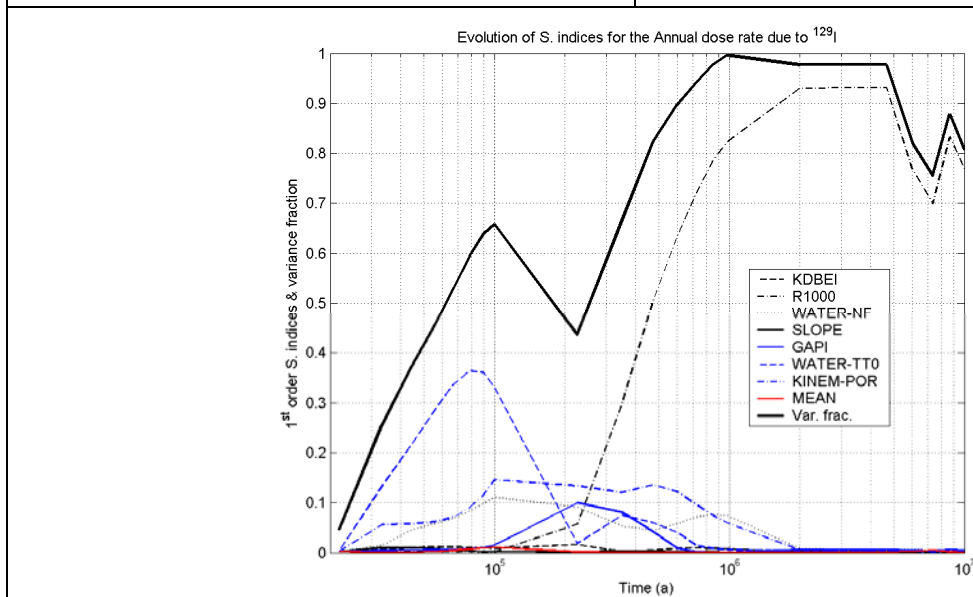


Figure 34.-First order sensitivity indices[†] and output variance fraction due to first order sensitivity indices for the Annual dose rate in biosphere due to ^{129}I .

*Only input parameters whose associated SRC or SRRC exceeds 0.1 at least at one time point are represented on this plot. [†]Only input parameters whose first order effect contribution to the output variable exceeds 1% at least at one time point are represented on this plot.



9. Conclusions

In this report we have described a set of statistics and techniques to perform uncertainty and sensitivity analysis in the framework of a PA. We have stressed their properties and also their deficiencies. We have also provided a template to use them efficiently, dividing them into either suggested or optional depending on the type of output variable under study. Finally, the template developed has been implemented using as a test data set the results obtained for the Biosphere annual dose rate due to ^{129}I in the Spanish reference concept in granite.

ACKNOWLEDGEMENTS

The authors of this document wish to explicitly acknowledge the support obtained from ENRESA, and specifically from José Luis Cormenzana and Miguel Ángel Cuñado, who provided us the data set used to test the template developed. We are also indebted to Daniel Galson who provided interesting comments and suggestions about the first draft of this report.

References

- [1] A. Badea and R. Bolado (2008). Review of Sensitivity Analysis Methods and Experience. PAMINA milestone
- [2] D.A. Becker, D. Buhmann, R. Storck, A. Alonso, J.L. Cormenzana, M. Hugl, F. van Genmert, P. O'Sullivan, A. Laciok, J. Marivoet, X. Sillen, H. Nordman, T. Vieno and M. Niemeyer (2002). Testing of Safety and Performance indicators (SPIN). Final report. EUR 19965 EN.
- [3] D.A. Becker, S. Spießl, K. R. Röhligh, E. Plischke, R. Bolado, A. Badea, J.L. Cormenzana, M.A. Cuñado, T.J. Schröder, J. Hart and R. Ávila (2009). Evaluation of approaches to sensitivity analysis. Deliverable D2.1.D.1, PAMINA IP.
- [4] R. Bolado, A. Badea, D.A. Becker, S. Spießl, K. Fischer-Appelt, J.L. Cormenzana, M.A. Cuñado, T.J. Schröder, J. Hart, J. Grupa, E. Rosca-Bocancea, R. Ávila, G. Pepin, F. Plas, K.-J. Röhligh and E. Plischke (2009). Lessons learnt from studies on sensitivity analysis techniques in the EU project PAMINA: Sensitivity analysis applied to different HLW PA models. In: Proceedings of ESREL 2009 conference. In press.
- [5] C. Cannamela, J. Garnier, and B. Iooss (2007). Controlled stratification for quantile estimation, submitted to J. Roy. Stat. Soc. B.
- [6] W.J. Conover (1980). Practical Nonparametric Statistics. Second edition. Ed. John Wiley & Sons.
- [7] J.L. Cormenzana, M.A. Cuñado and R. Bolado (2009). Sensitivity/uncertainty analyses. Application to a repository in granite. Milestone M2.1.D.8, PAMINA IP. Draft version.
- [8] H.A. David and H.N. Nagaraja (2003). Order Statistics, Third ed., J. Wiley & Sons, Wiley Series in Probability and Mathematical statistics.
- [9] ENRESA (2000). Evaluación del comportamiento y de la seguridad de un almacenamiento de combustible gastado en una formación granítica. 49-1PP-M-15-01 Rev.0. December 2001 (in Spanish).
- [10] A. Guba, M. Mihakly and P. Lenard (2003). Statistical aspects of best estimate method-1, Reliability Engineering and System Safety 80, 217–232.
- [11] R.J. Hyndman and Y. Fan (1996). Sample quantiles in statistical packages, *American Statistician*, 50, 361–365.
- [12] I.J. Jordaan (2005). Decisions under Uncertainty, Probabilistic Analysis for Engineering decisions, Cambridge University Press.
- [13] M. Makai and L. Pal (2006). Best estimate method and safety analysis II, Reliability Engineering and System Safety 91, 222–232.
- [14] W.T. Nutt and G.B. Wallis (2004). Evaluation of nuclear safety from the outputs of computer codes in the presence of uncertainties, [Reliability Engineering and System Safety](#) 83, 57–77.
- [15] W.T. Nutt and G.B. Wallis (2005). Reply to “Comments on ‘Evaluation of nuclear safety from the outputs of computer codes in the presence of uncertainties’ by W.T. Nutt and G.B. Wallis,” by Y. Orechwa, [Reliability Engineering and System Safety](#) 87, 137-145.
- [16] Y. Orechwa (2005). Comments on ‘Evaluation of nuclear safety from the outputs of computer codes in the presence of uncertainties’ by W.T. Nutt and G.B. Wallis, [Reliability Engineering and System Safety](#) 87, 133-135.
- [17] A. B. Owen (2001). Empirical Likelihood. Chapman and Hall/CRC, Boca Raton.
- [18] E. Plischke, K. R. Röhligh, A. Badea, R. Bolado, E. Ekströmand S. Hotzel (2009). Sensitivity analysis benchmark based on the use of analytic and synthetic PA cases. Milestone M2.1.D.11, PAMINA IP.

- [19] S. Prváková, R. Bolado-Lavín, A. Badea and K-F Nilsson (JRC); G. Pepin, E. Treille, F. Plas (ANDRA) (2008). PAMINA WP4.3 Benchmark 1 French Clay Repository: Application of Performance Assessment Methodologies for Clay Repository: Uncertainty and Sensitivity Analysis. PAMINA milestone M4.3.2.
- [20] G. Saporta (1990). Probabilités analyses des données et Statistiques. Editions Technip, Paris.
- [21] G.B. Wallis (2003). Contribution to the paper 'Statistical aspects of best estimate method-1' by A. Guba, M. Makai and L. Pal, [Reliability Engineering and System Safety](#) 80, 309-311.
- [22] G.B. Wallis (2006). Evaluating the probability that the outputs of a computer code with random inputs will meet a set of evaluation criteria, [Reliability Engineering and System Safety](#) 91, 820-827.
- [23] S. Wilks (1941). Determination of sample sizes for setting tolerance limits; Ann. Math. Statist., 12, pp. 91-96.

Annex A: Monte Carlo simulation

Whenever the system model is available and the distributions of the input parameters have been derived, the next step in the PA study is to propagate uncertainties in order to get information about the distribution of the output variables. Analytical uncertainty propagation methods can only be used for very simple systems with very few parameters. In more complex cases other methods need to be adopted. The most suited method, and in fact the most used one is the Monte Carlo method.

THE MONTE CARLO METHOD

The Monte Carlo method consists in sampling at random the vector of input parameters, running the system model computer code for each sample of that vector and getting a sample of the vector of output variables. Later on, the characteristics of the output variables may be estimated using the output samples obtained. One of the advantages of using the Monte Carlo method is that all statistical standard methods we need to estimate the output variables distributions and to test any hypothesis may be used. This makes it the most straightforward and powerful method available in the scientific literature to deal with uncertainty propagation in complex models. This method is valid for models that have static and also dynamic outputs. It is adequate for working with discrete and continuous inputs and outputs, and the implementation of computational algorithms required has no fundamental complexity.

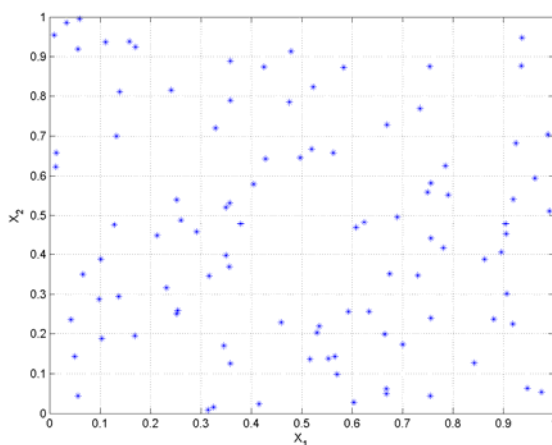


Figure A.1.- Simple random sample of size 100 of two random variables uniformly distributed in the region $[0,1] \times [0,1]$.

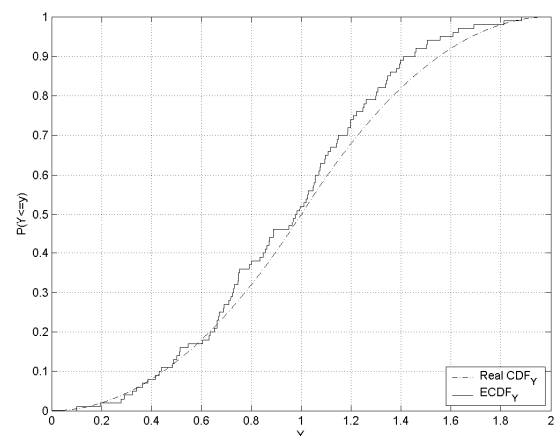


Figure A.2.- ECDF obtained from a simple random sample of size 100 of $Y = X_1 + X_2$ and its theoretical CDF.

Monte Carlo maps the input space into the output space point by point. In order to see this, let us consider a very simple model: $Y = X_1 + X_2$. Suppose X_1 and X_2 follow independent uniform distributions both of them defined in the interval $[0,1]$. For this simple model an

analytical propagation of uncertainties is feasible and the output Y follows a triangular distribution defined in the interval $[0, 2]$ and whose mode, mean and median are 1. This propagation may be done via Monte Carlo. First, a sample of size 100 is taken in the input space (see figure A.1). For each point shown in figure A.1, the value of the output is then computed. An empirical cumulative distribution function is built from the 100 values obtained (see figure A.2). For the sake of comparison the actual CDF of Y has also been drawn.

Monte Carlo may also be seen as a numerical integration method. In the same example, let us consider that we are primarily interested in the estimation of the mean of Y . This means that we are trying to estimate

$$\mu_Y = \int_{[0,1] \times [0,1]} (X_1 + X_2) dx_1 dx_2 . \quad (\text{A.1})$$

One of the possible approximations to compute this integral is to take the sample considered in figure A.1 and figure A.2 and to calculate the arithmetic mean

$$\hat{\mu}_Y = \frac{1}{100} \sum_{i=1}^{100} (x_{1i} + x_{2i}) . \quad (\text{A.2})$$

It is important to remark that the standard deviation of this estimator is

$$\sigma_{\hat{\mu}_Y} = \sigma_Y / \sqrt{n} , \quad (\text{A.3})$$

where σ_Y is the standard deviation of the output Y .

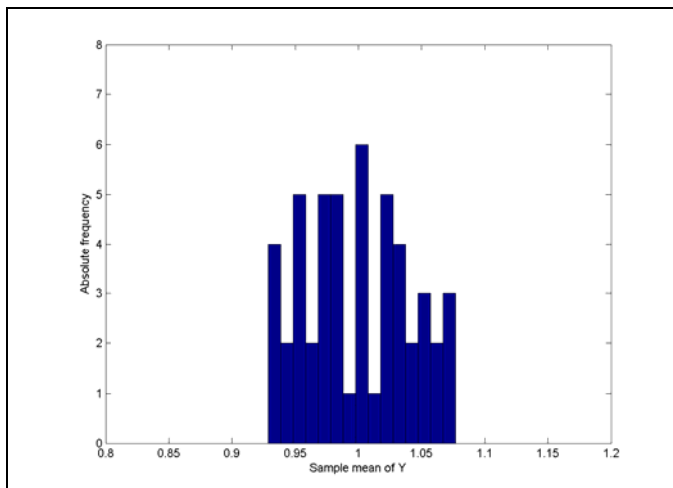


Figure A.3.- Histogram of the sample means obtained from 50 simple random samples of size 100 obtained via Monte Carlo simulation.

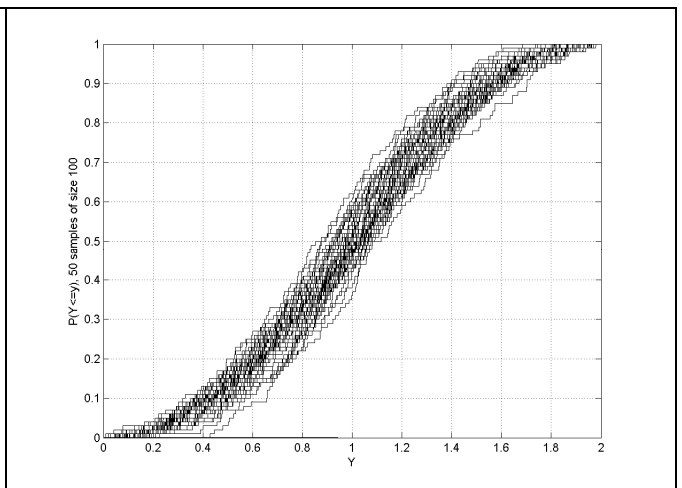


Figure A.4.- ECDFs obtained from 50 simple random samples of size 100 obtained via Monte Carlo simulation.

Figure A.3, which shows the histogram obtained from 50 simple random samples of size 100, similar to the one shown on figure A.1 and figure A.2. In this plot we can see that the range of $\hat{\mu}_Y$ is roughly 0.2, which means it represents one tenth of the range of Y (the range of this triangular distribution is 2). Figure A.4 shows the corresponding ECDFs.

It is important to remark that the standard error of $\hat{\mu}_Y$ does not depend on the dimension p of the space where the integral is computed, and that consequently the Monte Carlo method does not suffer from the curse of dimensionality. Metropolis and Ulam (1949) is the seminal paper about Monte Carlo, where many interesting suggestions are made about its applicability.

VARIANCE REDUCTION TECHNIQUES

The computational time to perform a Monte Carlo analysis depends on the number of simulations and cost per simulation. The computational time for complex problems with a large number of simulations often become prohibitive. The cost for each simulation can be reduced by simplifying the mathematical description of the problem. A second alternative is to reduce the number of simulations compared to standard random sampling without sacrificing the precision and confidence intervals of the outputs. Such techniques are referred to as Variance Reduction Techniques. Main techniques are Latin Hypercube sampling (LHS), stratified sampling, control variates, importance sampling and antithetic variates. In the following pages we discuss about the first three. Readers interested in getting further details about these techniques are suggested to see Hammersley and Handscomb (1964), Rubinstein (1981) and Robert and Casella (2004).

Stratified sampling

Input parameters may vary considerably. By stratification the population is sub grouped into relatively homogenous subgroups. The sampling is then performed for each of the strata. The strata must be mutually exclusive and collectively exhaustive. Stratified sampling is based on the fact that the variance of any random variable, once it has been divided in strata, may be decomposed into two contributions: the variability within each stratum and the variability between different strata, which means

$$\sigma^2 = \sum_{i=1}^h \omega_i \cdot \sigma_i^2 + \sum_{i=1}^h \omega_i \cdot (\mu_i - \mu)^2, \quad (\text{A.4})$$

where the first summand represents the variability within the h considered strata and the second one represents the variability between different strata. ω_i stands for the probability of stratum i , and μ_i and σ_i stand for the mean and the standard deviation of the (output) variable Y also in stratum i . If the sampling of each observation is restricted to a given stratum, its variability will be the variability of that stratum (σ_i) rather than the whole variability of Y (σ). The estimate for the mean of Y under stratified sampling is

$$\hat{\mu}_S = \sum_{i=1}^h \omega_i \hat{\mu}_i = \sum_{i=1}^h \omega_i \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad , \quad (\text{A.5})$$

where $\hat{\mu}_i$ is the estimate of the mean of Y in stratum i , which normally is computed as the average of the values of Y obtained in that stratum (y_{ij}), as shown in (A.5). n_i is the sample size in stratum i and the whole sample size is $n=n_1+\dots+n_h$. This estimator is an unbiased estimator of Y 's mean. It may be easily demonstrated that the variance of this estimator for a given number of samples is reduced with respect to the estimator provided by simple random sampling according to

$$Var(\hat{\mu}_S) = Var(\hat{\mu}) - \frac{1}{n} \sum_{i=1}^h \omega_i (\mu_i - \mu)^2 \quad , \quad (\text{A.6})$$

which means that the larger the differences between the means of the different strata the larger the decrease in the variance of the stratified sampling estimator.

The main problem affecting stratified sampling is that ideally what should be stratified is the output space, so that the second term on the right hand side of (A.6) would be large and so it would be profitable to stratify. Unfortunately, what can be easily stratified is the input sample space, which doesn't mean that the corresponding stratification in the output space will be so good. Under those circumstances, when large overlaps between different strata happen, the benefit from stratifying would not be so important, though some benefit will always be obtained according to (A.6).

Once the sample size has been chosen, there are two problems to be solved: 1) how to create the strata and 2) the sample size within each stratum. There is no clear rule to partition the input sample space. When no additional information is available about the system model, the most common strategy is to build a net of hypercubes via Cartesian product of the stratification performed in each input variable. When some information is available, it can be used for creating the stratification. In general, there are two ways to get information about the model: Studying the equations of the model and getting a small size sample. The study of the equations of the model may provide information on the relation between inputs and outputs and on the importance of combinations of specific sets of inputs, see 'Input Space Dimension Reduction' within this annex. A small size sample could be obtained via simple random sampling and it could be used to perform Sensitivity Analysis (SA). The use of SA techniques could help identifying the most relevant input variables; stratification could be performed only on these relevant input variables

Regarding the sample size per stratum, there are several options. The first option is to take proportional sampling, which means that the sample size in each stratum is proportional to the probability of the stratum: $n_i=n\omega_i$. Further improvement may always be achieved (McKay et al. (1979)) if the sample space is further stratified to getting as many strata as samples (one observation per stratum). In that case the reduction in the variance of the estimator with respect to simple random sampling is

$$Var(\hat{\mu}_s) = Var(\hat{\mu}_R) - \frac{1}{n^2} \sum_{i=1}^n (\mu_i - \mu)^2 \quad . \quad (A.7)$$

Not only the mean of the output is more accurately estimated when stratified sampling is applied, but its distribution is better estimated due to the evenness in the sampling all over the sampling space, no region is either over-sampled or under-sampled. Figure A.5 provides an idea about the way to get a stratified sample of size 9 with one observation per stratum (each stratum has probability $1/9$) in a 2-D input sample space. Figure A.6 provides the same information when the stratification is done on only one of the input variables (X).

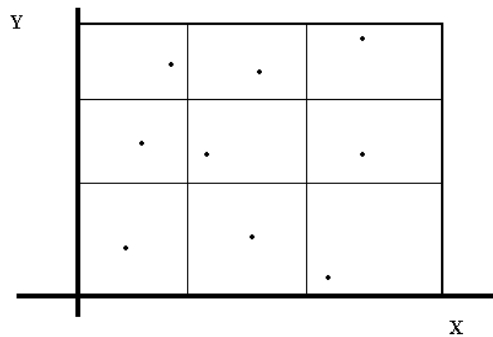


Figure A.5.- Stratified sample with nine observations for two variables; one observation per stratum, probability of each stratum $1/9$.

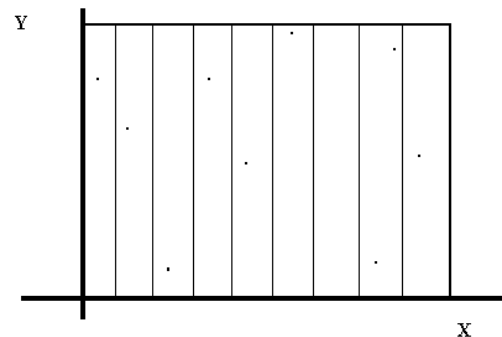


Figure A.6.- Stratified sample with nine observations for two variables; one observation per stratum, probability of each stratum $1/9$.

Latin Hypercube Sampling (LHS)

Latin Hypercube Sampling (LHS) is a cost-effective and reliable extension of stratified sampling, designed to generate collections of parameter values from multivariate distributions. In order to get a sample of size n , the procedure is the following one:

- a stratified sample is obtained for each input variable (n strata with probability $1/n$ each one and a sample of size 1 per stratum),
- get a permutation of each one of the samples of each input variable
- combine the first observations of all the variables (after permutation) to get the first observation of the input vector, combine the second observations of all the variables (after permutation) to get the second observation of the input vector and so on.

The procedure above is generally complemented by techniques to fill the input space in an optimal way, for instance by maximizing the minimum distance between the samples points. McKay et al. (1979) shows LHS produces unbiased estimators for the mean and the CDF of the output. They also demonstrate that a sufficient condition to get an estimation error for the sample mean and the CDF smaller than in the case of random sample is that the model has to be monotonic in all its input variables. Stein (1987) proved some asymptotic properties of LHS under general conditions: the variance of the estimators provided for the mean and the CDF are smaller (asymptotically) than the ones obtained under simple random sampling, with the degree of variance reduction depending on the additivity of the model. The estimates do also follow, asymptotically, a normal distribution. Iman and Conover (1982) developed a method to induce rank correlation between input variables sampled under this scheme and Stein (1987) introduced a method to induce correlations between input variables. Figure A.7 shows the way to generate a sample of size 5 through this method for a bivariate random vector.

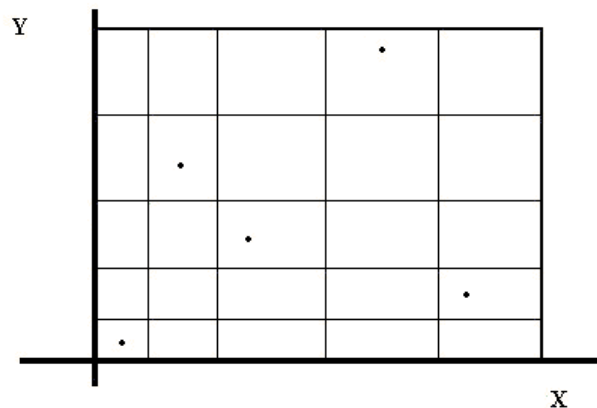


Figure A.7.- LHS sample of size 5 for two variables. Each stratum has the same probability

Control variates

This technique is based on decomposing the output random variable Y as a sum of two ancillary random variables Y' and Y'' in such a way that Y' should have a mean analytically and easily computable, or at least with a well known dependence on the vector of input parameters X (so that its mean could be computed with the needed accuracy at a low cost), and Y'' should have a small variance. Under these conditions, the mean may be split up as

$$\mu_y = \int_{S_x} y(\mathbf{x}) \cdot f(\mathbf{x}) \cdot d\mathbf{x} = \int_{S_x} y'(\mathbf{x}) \cdot f(\mathbf{x}) \cdot d\mathbf{x} + \int_{S_x} (y(\mathbf{x}) - y'(\mathbf{x})) \cdot f(\mathbf{x}) \cdot d\mathbf{x} , \quad (\text{A.8})$$

where $y''(x) = y(x) - y'(x)$ (figure A.8). Again, as in the case of importance sampling and stratified sampling, we need additional information to find Y' . If no such theoretical information is available, the most straightforward way to get it is using a previous small size sample. That sample may be used, for instance, to build a response surface (see Myers and Montgomery (2002)) that captures the main characteristics of the functional dependence of Y over X . The response surface obtained would be Y' (also represented as $y'(x)$ in this text). On one side Y' will usually be a polynomial that may be used to propagate uncertainty analytically or computationally using huge sample sizes, estimating the first integral on the right hand side of (A.8) with no or negligible error. On the other side, if the quality of this response surface is good, $y(x) - y'(x)$ would have small values for all values of input vector x , so that the last integral in (A.8) would be the only one introducing relevant error in the estimation of the mean, but much smaller than the one introduced by normal random sampling.

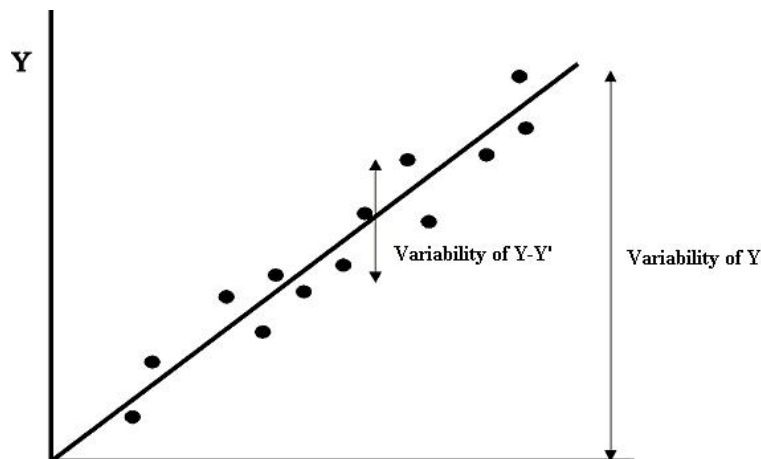


Figure A.8.- Intuitive idea behind control variates sampling technique.

INPUT SPACE DIMENSION REDUCTION

Let us consider a system of functional equations where $\mathbf{Y} = (Y_1, \dots, Y_n)$ are the dependent or output variables and $\mathbf{X} = (X_1, \dots, X_m)$ are the independent variables (e.g., space coordinates and time). Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ be the parameters of the system, that is, coefficients of the differential equations and of the initial and boundary conditions. The solutions of the system are $Y_j = I_j(\mathbf{X}; \boldsymbol{\theta})$.

In physics, one speaks of similarity between two problems when one can transform one problem into the other by a change of scale in the variables. It is shown that this is possible when a set of dimensionless numbers (in mathematical terms, we shall speak instead of

invariant functions), which are functions of the parameters θ ; coincide in both problems. A classical example is the Reynolds number in fluid mechanics. The dimension of the parameter space, originally p , can thus be reduced to the number of dimensionless quantities that define the system of functional equations. This problem is referred to in the literature as dimensional analysis, and though in many physics and engineering works it is formulated in terms of physical magnitudes and dimensions (Buckingham (1914); Langhaar (1951); Palacios (1964); Szirtes (1998)). A more abstract, mathematical, and hence physics independent language, is preferable when dealing with propagation of uncertainties, such as in Moran (1971).

Moran and Marshek (1972) generalized dimensional analysis consists of finding a set of linear transformations:

$$\begin{aligned} Y'_j &= K_j Y_j (j=1, \dots, n \geq 1) \\ X'_k &= K_{n+k} X_k (k=1, \dots, m \geq 1) \\ \theta'_e &= K_{n+m+e} \theta_e (e=1, \dots, p \geq 1) \end{aligned} \quad (\text{A.9})$$

of the \mathbf{Y} , \mathbf{X} , θ , where the K_j , $j = 1, \dots, n + m + p$ are constants, such that the system of functional equations is invariant under the transformations, that is, $Y_j = I_j(\mathbf{X}; \theta)$ transforms to $Y'_j = I'_j(\mathbf{X}'; \theta')$; where $\mathbf{X}' = X'_1, \dots, X'_m$ and $\theta' = (\theta'_1, \dots, \theta'_p)$. We note that the prime symbol stands for variable transformation and not for array transposition. A more general class of transformations could have been used, but we are restricted here to linear transformations (scale changes) because they have proved useful in many physical problems, while maintaining mathematical simplicity and a clear physical interpretation.

After introducing the transformations or scale changes into the system equations and boundary and initial conditions, and imposing the condition of invariance (the system equations maintain the same form before and after the transformation; $Y_j = I_j(\mathbf{X}; \theta) \Leftrightarrow Y'_j = I'_j(\mathbf{X}'; \theta')$), there appear restrictions linking the values of the K_i , $i = 1, \dots, n + m + p$. In most cases, the restrictions will reduce their degrees of freedom. So if initially there are $n+m+p$ transformation constants K_i and q restrictions, there will finally be $r = n+m+p-q$ degrees of freedom for the K_i . Then, the transformations can be defined in terms of a reduced set of constants, which are called A_j , $j=1, \dots, r$, and the set of transformations may be rewritten as

$$\begin{aligned} Y'_j &= A_1^{a_{j1}} \dots A_r^{a_{jr}} Y_j (j=1, \dots, n \geq 1) \\ X'_k &= A_1^{b_{k1}} \dots A_r^{b_{kr}} X_k (k=1, \dots, m \geq 1) \\ \theta'_e &= A_1^{c_{e1}} \dots A_r^{c_{er}} \theta_e (e=1, \dots, p \geq 1) \end{aligned} \quad (\text{A.10})$$

where the a_{ji} , b_{ki} and c_{ei} are exponents. In fact each restriction defines an invariant function or dimensionless number

$$\pi = Y_1^{\alpha_1} \dots Y_n^{\alpha_n} X_1^{\beta_1} \dots X_m^{\beta_m} \theta_1^{\gamma_1} \dots \theta_p^{\gamma_p} \quad (\text{A.11})$$

where the α_i , β_j , γ_k are also exponents, in such a way that (see Moran and Marshek (1972)) the system of functional equations can be expressed in terms of these invariant functions, instead of in terms of the original and larger set formed by \mathbf{Y} , \mathbf{X} , $\boldsymbol{\theta}$. The calculation of the invariants and of the expression of the system model in terms of the invariants is formalized in the theorems of Moran and Marshek (1972); see also appendix A of Mira et al. (2004) for details.

Usually, the reduction of dimension is in the space of input parameters and input variables (\mathbf{X} , $\boldsymbol{\theta}$), only very infrequently is the reduction performed in the space of output variables. Even when a reduction of dimension is obtained in the space of input parameters, it does not necessarily mean that this produces a benefit in the propagation of uncertainties. It is possible that the reduction of dimension happens in the part of the space of input parameters that is not affected by uncertainty (known constants); in that case no improvement is obtained. Moreover, in order to get some benefit, variance reduction techniques have to be applied in combination with dimension reduction. If an effective dimension reduction is obtained, using simple random sampling on this space doesn't lead to a net decrease in the variance of the estimators of the outputs; a simple random sample of the input space produce a simple random sample of the output space independently of the dimension of the equivalent input sampled space.

Mira et al. (2004) describe an application of dimension reduction obtained via dimensional analysis for the propagation of uncertainties of a simplified HLW repository. In this application, the original space of input parameters and input variables has dimension 7 and the transformed one 4. Nevertheless, the real reduction obtained is from 3 to 2 since only two input parameters and one input variable are affected by uncertainty and these inputs are concentrated in only 2 invariants in the transformed input space. Mira and his colleagues compare in their work four sampling techniques: simple random sampling, LHS and stratified sampling in the original 3-D input space and stratified sampling in the 2-D transformed input space. For this comparison 60 samples of size 64 were used. Figure A.9 shows the means of the means for the flow of ^{129}I getting into the biosphere at different times. All techniques produce unbiased results. Figure A.10 shows the standard deviations of the means for the same case and illustrates the improvement that is obtained when combining dimension reduction and stratified sampling with respect to the other techniques applied on the original input space. The results for LHS in the 2-D space shouldn't be taken into account since it shouldn't be called LHS the way this sampling scheme was actually applied in this test case.

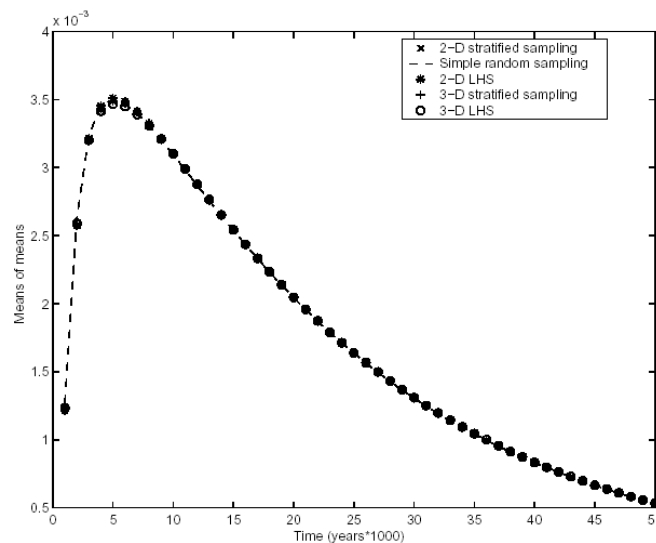


Figure A.9.- Evolution over time of the mean of the means for different sampling schemes with (dimensions reduction (2-D) curves) and without (no dimension reductions (3-D) or simple random curves) input space dimension reduction obtained via Dimensional Analysis.

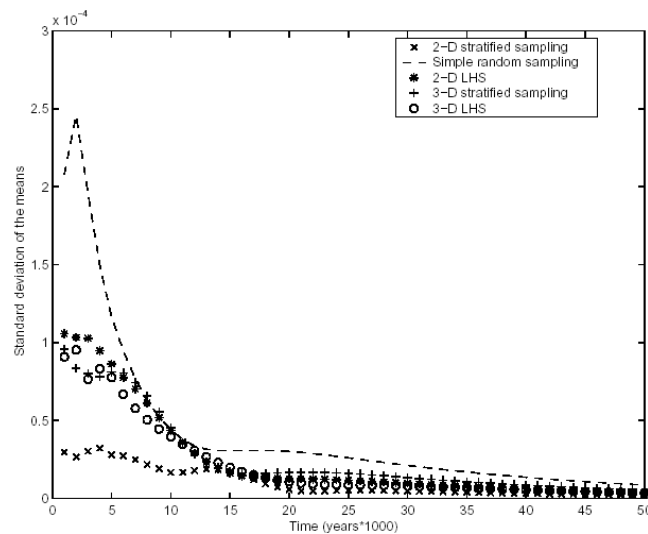


Figure A.10.- Evolution over time of the standard deviation of the means for different sampling schemes (2-D curves) and without (3-D or simple random curves) input space dimension reduction obtained via Dimensional Analysis.

Dimension reduction in the input space may also be obtained in a more immediate, less sophisticated way, referred to as trivial reductions of dimension. It frequently occurs that in the differential equations which describe the behavior of the system, some coefficients appear as elementary functions of a number of, let us say, original coefficients, such that, either due to physical reasons or because of reasons related the way experts address the problem, the uncertainty has been expressed in terms of probability distributions for the

original coefficients. Bolado and Mira (2004) showed that, as in the case of dimension reduction obtained via dimensional analysis, smaller errors are made in the estimation of the outputs when trivial reductions of dimension are combined with variance reduction techniques such as stratified sampling.

REFERENCES

- [1] R. Bolado and J. Mira (2004). Trivial reductions of dimensionality in the propagation of uncertainties: a physical example. *Environmetrics*, 15, 57-66.
- [2] E. Buckingham (1914). On physically similar systems. *Physics Review*, 4, 345 – 376.
- [3] J.M. Hammersley and D.C. Handscomb (1964). *Monte Carlo Methods*. Chapman and Hall.
- [4] R.L. Iman and W.J. Conover (1982). A distribution free approach to inducing rank correlations among input variables. *Communications in Statistics, Part B – Simulation and Computation*, 11, 311-334.
- [5] H.L. Langhaar (1951). *Dimensional Analysis and Theory of Models*. Wiley.
- [6] M.D. McKay, W.J. Conover and R.J. Beckman (1979). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code. *Technometrics*, 21, 239-245.
- [7] N. Metropolis and S. Ulam (1949). The Monte Carlo Method. *Journal of the American Statistical association*. Vol. 44, N. 247, 335-341.
- [8] J. Mira, R. Bolado and P. Solana (2004). The use of dimensional and similarity analysis in the propagation of uncertainties: a physical example. *Journal of computational and Graphical statistics*, 13, N. 2, 467-486.
- [9] M. Moran (1971). A generalisation of dimensional analysis. *Journal of the Franklin Institute*, 292, 423-432.
- [10] M. Moran and K.M. Marshek (1972). Some matrix aspects of generalised dimensional analysis. *Journal of Engineering Mathematics*, 6, 291-303.
- [11] R.H. Myers and D.C. Montgomery (2002). *Response Surface Methodology*, J. Wiley & Sons, Wiley Series in Probability and Statistics.
- [12] J. Palacios (1964). *Dimensional Analysis*. MacMillan.
- [13] C.P. Robert and G. Casella (2004). *Monte Carlo Statistical Methods*. Springer.
- [14] R.Y. Rubinstein. *Simulations and Monte-Carlo Method*. Wiley Series in Probability and Mathematical Statistics, J. Wiley & Sons, 1981
- [15] M. Stein (1987). Large sample properties of simulations using Latin Hypercube Sampling. *Technometrics*, vol. 29, n°2, 143-151.
- [16] T. Szirtes (1998). *Applied dimensional analysis and modeling*. McGraw-Hill.

Annex B: Classical inference methods

Classical inference methods are based on the assumption of having a random sample. The target is to determine the PDF that generated the random sample. This process may be divided in three steps:

- Model identification
- Parameter estimation, which is divided in two parts
 - Point estimation
 - Interval estimation
- Diagnosis of the model

Model identification consists in finding the most appropriate probability model (uniform, normal, log-normal, exponential, Weibull, etc.) for the sampled data. This task needs the use of graphic tools such as histograms, in addition to the experience in the field under study. Furthermore experts in the field will often have an idea of the distributions that could best represent the data. This part of the process certainly involves subjective elements.

Once the probability model has been identified, the parameters need to be determined. Most probability models are characterised by a set of parameters (parametric models), as for example the mean, μ , and the standard deviation, σ , in a normal (Gaussian) probability model. Estimation is done via techniques of point estimation. These techniques allow identifying a best choice for those parameters. Identifying best choices does not mean that those are the only acceptable ones; other similar values could also be acceptable. A measure of error or of likely alternatives is also needed. This is provided by *interval estimates*.

The last step consists in checking that the hypotheses considered in the whole process were correct. Three hypotheses are normally used: the type of probability model, the independence between the different observations and the homogeneity of the sample.

In the following pages special attention will be dedicated to both types of parameter estimation (step 2) and to checking that the assumed probability model is good enough (first hypothesis tested in step 3).

POINT ESTIMATION

The best-known and most widely used methods are the Maximum Likelihood Method and the Method of Moments. The main shortcoming of all these methods is their requirement of sample sizes to get good quality estimates. In practical situations with real engineering facilities it may be quite difficult to get the required sample size.

Method of moments

Method of Moments is probably the oldest inferential method to estimate the parameters of a PDF. K. Pearson developed the method of moments by the end of 19th century. The idea is quite simple. It consists in taking as an estimator of a parameter its equivalent sample quantity. So, the sample mean is the estimator for the mean, the sample variance is the estimator for the variance and so on.

Maximum Likelihood method

The Maximum Likelihood Method is the most widely used and most powerful estimation method in the classical context. Let us assume that we wish to study a random variable X (representing a parameter affected by uncertainty) of a known distribution function type $f(\mathbf{X}|\theta)$, but of unknown parameter θ . In order to estimate θ we take a random sample $\mathbf{X} = (X_1, X_2, \dots, X_n)$, which is assumed to be a random vector, whose components are independent and identically distributed (iid), so that its joint probability density function is

$$f(\mathbf{X}|\theta) = f(X_1, \dots, X_n|\theta) = \prod_{i=1}^n f(X_i|\theta). \quad (\text{B.1})$$

It is important to notice that, in this expression, under the classical view, before sampling, θ is unknown, but has an assigned value that determines what regions of \mathbf{X} are more likely and what regions are less likely. So, this is a function whose unknowns are \mathbf{X} . This is the meaning before sampling. As soon as the sample is available, \mathbf{X} is known, while θ remains unknown. The objective is to determine what value, among all the possible values of θ , makes the sample actually obtained the most likely one. The problem is hence to find the value of θ for which the function defined in (B.1) attains its maximum value. As it is convenient to look at the problem after getting the sample, expression (B.1) is usually written as

$$L(\theta|\mathbf{X}) = f(\theta|\mathbf{X}) = f(\theta|X_1, \dots, X_n) = \prod_{i=1}^n f(X_i|\theta), \quad (\text{B.2})$$

which means that, after sampling, the probability density function of the sample vector is changed into a function of the unknown parameter θ . 'L' stands for 'Likelihood'. From a practical point of view, the function whose maximum is actually computed is not L , but its logarithm $l(\theta|\mathbf{X})$. Both functions reach a maximum at the same point since the transformation to get one from the other one is a monotonic transformation.

As an example, let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be a sample of size n of a Gaussian random variable whose variance σ^2 is known. We wish to estimate the mean μ of the random variable under study. Under these circumstances, the likelihood function is

$$L(\mu|\mathbf{X}) = f(\mu|X_1, \dots, X_n) = \prod_{i=1}^n f(\mu|X_i) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} e^{-\frac{1}{2} \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2}, \quad (\text{B.3})$$

whose logarithm is

$$l(\mu|\mathbf{X}) = \ln(L(\mu|\mathbf{X})) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2} \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2. \quad (\text{B.4})$$

In order to compute the value of μ for which this expression reaches a maximum, we compute its first derivative with respect to μ

$$\frac{\partial l(\mu|\mathbf{X})}{\partial \mu} = -\sum_{i=1}^n \left(-\frac{1}{\sigma} \right) \left(\frac{X_i - \mu}{\sigma} \right). \quad (\text{B.5})$$

The maximum is obtained when this expression equals zero, which happens for the value

$$\hat{\mu} = \overline{X_n} = \frac{1}{n} \sum_{i=1}^n X_i. \quad (\text{B.6})$$

The reader may check, by computing the second derivative that, indeed, the likelihood function reaches a maximum when $\mu = \hat{\mu}$ (second derivative is less than zero when $\mu = \hat{\mu}$). The method may also be applied when a PDF is defined through a vector of parameters; in that case the usual rules for maximizing a multi-parameter function must be applied (to equal first partial derivatives to zero and to check conditions imposed on the Hessian matrix evaluated at the point where first partial derivatives are zero). The method provides a single value as an estimate. If needed, a confidence interval with the desired degree of confidence, may be obtained using interval estimation theory.

The maximum likelihood method has several properties that makes it the most widely used estimation method Mood et al. (1974):

- The estimators obtained through this method are asymptotically unbiased (the limits of their expected values when the sample size tends to infinite are the true values of the parameters).
- They are asymptotically normal since their distributions become normal when the sample size tends to infinite.
- They are asymptotically efficient; for large sample sizes, they are the most accurate estimators.
- They are sufficient since they summarise all the relevant information contained in the sample.
- They are invariant; if $\hat{\theta}$ is the maximum likelihood estimator of θ , and $\theta' = f(\theta)$, then $f(\hat{\theta})$ is the maximum likelihood estimator of θ' .

Table B.1.- The most useful probability distributions functions, their parameters and their maximum likelihood estimators. *The solutions of this system of equations, where ψ stands for the digamma function, are the maximum likelihood estimators. \dagger c is estimated recursively from the second equation, later on its estimate is substituted in the first one in order to get the estimator of α .

Distribution	PDF	Parameters	Estimators
Uniform	$\frac{1}{b-a}; a \leq x \leq b$ 0; otherwise	a: Minimum b: Maximum	$\hat{a} = \min\{x_1, \dots, x_n\}$ $\hat{b} = \max\{x_1, \dots, x_n\}$
Log-uniform	$\frac{1}{x \ln(b/a)}; a \leq x \leq b$ 0; otherwise	a: Minimum b: Maximum	$\hat{a} = \min\{x_1, \dots, x_n\}$ $\hat{b} = \max\{x_1, \dots, x_n\}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right];$ $\sigma > 0$	μ : mean σ^2 : variance	$\hat{\mu} = \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$ $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$
Log-normal	$\frac{1}{x\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right];$ $x > 0, \sigma > 0$	μ : mean of $\ln(x)$ σ^2 : variance of $\ln(x)$.	$\hat{\mu} = \bar{x}_n = \frac{1}{n} \sum_{i=1}^n \ln x_i$ $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (\ln x_i - \bar{x}_n)^2$
Exponential	$\lambda e^{-\lambda t}; t > 0$	λ : inverse of the mean	$\hat{\lambda}^{-1} = \frac{1}{n} \sum_{i=1}^n t_i$
Gamma	$\frac{t^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} e^{-t/\beta};$ $\alpha > 0, \beta > 0, x > 0$	α : shape param. β : scale param.	$\hat{\alpha}\hat{\beta} = \bar{x}_n$ $\frac{1}{n} \sum_{i=1}^n \ln x_i = \ln \hat{\beta} + \psi(\hat{\alpha})$ *
Weibull	$\frac{cx^{c-1}}{\alpha^c} \exp(-x/\alpha)^c$	α : scale param. c : shape param.	$\hat{\alpha} = \left(1/n \sum_{i=1}^n x_i^c\right)^{1/c}$ $\hat{c}^{-1} = \left(\sum_{i=1}^n x_i^c \ln x_i\right) \left(\sum_{i=1}^n x_i^c\right)^{-1} - \frac{1}{n} \sum_{i=1}^n \ln x_i$ †
Binomial	$\binom{n}{i} p^i (1-p)^{n-i}$	p: prob. of event	$\hat{p} = r/n$ r=number of times event happens out of n trials
Poisson	$\frac{\lambda^k}{k!} e^{-\lambda}$	λ : Mean n. of events per unit of time, length, surface, etc.	$\hat{\lambda} = r/n$ r= number of events n= sample size (s, m, m ² , etc.)

INTERVAL ESTIMATION

The purpose of point estimation is to give some single “best” value of each unknown parameter, based on sample data. Nevertheless, any point estimate cannot completely describe the distribution. Due to the way the estimation process is conducted, the estimate and the actual value of the parameters are close, but they are usually different. Scientists and engineers try to provide, at least, a measure of the error made when a point estimate is given. Interval estimation was created to solve this problem.

Confidence intervals are the main tool to estimate intervals for a given parameter in a probability model. The theory of confidence intervals is based on the study of the distribution of the sample mean, the sample variance and other statistics and on the concept of pivotal quantity. If we take a sample of size n from a Gaussian variable and we compute the sample mean we will get a given value, usually close to the mean μ of that variable. If we get a new sample of size n , we can compute a new sample mean. We may repeat the same process k times and we will get a sample of size k of the sample mean based on n observations. By plotting these k values as a histogram, we will get an idea of the distribution with the associated sample mean. Any standard statistics book (see Mood et al. (1974) or Casella and Berger (1990)) shows that, for a Gaussian variable, the sample mean follows a Gaussian distribution with mean μ and standard deviation σ/\sqrt{n} . The sample mean as a random variable has the same mean as the variable itself but its standard deviation is smaller. In fact, the larger n , the smaller its standard deviation. Additionally, its distribution is also normal. Taking into account the properties of normal distributions, this means that the quantity $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ follows a standard Gaussian distribution with mean=0, standard deviation=1. This quantity is referred to as ‘pivotal quantity’; it is a function of the sample values and the parameter studied but whose distribution does not depend on the actual value of the parameter. Knowing the distribution of this pivotal quantity, we obtain

$$P\left[-z_{\alpha/2} \leq (\bar{X} - \mu)/(\sigma/\sqrt{n}) \leq z_{\alpha/2}\right] = 1 - \alpha \Leftrightarrow P\left[\mu \in \bar{X} \pm z_{\alpha/2}(\sigma/\sqrt{n})\right] = 1 - \alpha, \quad (\text{B.7})$$

where $z_{\alpha/2}$ stands for the $100(1-\alpha/2)\%$ percentile of the standard Gaussian distribution. Expression (4.7) means that the interval

$$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] \quad (\text{B.8})$$

is a $100(1-\alpha)\%$ confidence interval for the mean of that normal distribution whose standard deviation is known. Typically α is set to 0.05 and then $z_{\alpha/2}=1.96$. In this case the interval obtained is a 95% confidence interval.

Table B.2: Confidence intervals for normal, exponential and generic probability distributions. $\dagger \chi^2_{\alpha/2}$ stands for the $100(1-\alpha/2)\%$ percentile of the corresponding χ^2 distribution (i.e. with as many degrees of freedom as indicated in the fourth column of the table). * Stands for asymptotic results, which means that they are valid for large sample sizes; all the others are exact results.

Distribution	Parameter	Pivotal quantity	Distribution of the pivotal quantity	Confidence interval
Normal	μ (σ^2 : known)	$(\bar{X} - \mu) / (\sigma / \sqrt{n})$	Standard Gaussian: $N(0,1)$	$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$
Normal	μ (σ^2 : unknown)	$(\bar{X} - \mu) / (\hat{\sigma} / \sqrt{n})$	t_{n-1}	$\left[\bar{x} - t_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$
Normal	σ^2	$(n-1)\hat{S}^2 / \sigma^2$	χ^2_{n-1}	$\left[(n-1)\hat{S}^2 / \chi^2_{\alpha/2}, (n-1)\hat{S}^2 / \chi^2_{1-\alpha/2} \right] \dagger$
Exponential	λ	$2n\lambda\bar{X}$	χ^2_{2n}	$\left[\chi^2_{1-\alpha/2} / (2n\bar{x}), \chi^2_{\alpha/2} / (2n\bar{x}) \right] \dagger$
Generic	θ	$\frac{\theta - \hat{\theta}_{ML}}{\sigma(\hat{\theta}_{ML})}$	*Standard Gaussian: $N(0,1)$	$\left[\hat{\theta}_{ML} - z_{\alpha/2} \sigma(\hat{\theta}_{ML}), \hat{\theta}_{ML} + z_{\alpha/2} \sigma(\hat{\theta}_{ML}) \right]$

Interpretation of confidence intervals

Suppose that a pivotal quantity is used to estimate a $100(1-\alpha)\%$ confidence interval $[\theta_1, \theta_2]$ for a given parameter θ of a probability model according to the procedure above described. A priori, the probability that the interval $[\theta_1, \theta_2]$ contains θ is $100(1-\alpha)\%$. The values θ_1 and θ_2 are computed on a sample; once they are computed, the true value of the unknown parameter is either in the interval $[\theta_1, \theta_2]$ or outside it, hence we cannot speak about probability any more. By repeating the experiment (i.e. by taking different samples and by computing the interval $[\theta_1, \theta_2]$) a certain number of times, in average $100(1-\alpha)\%$ of the cases, the true parameter will be in the confidence interval. But we don't know in which cases this will happen. This is the reason why the well-known expression "with confidence $100(1-\alpha)\%$ the parameter lies in the confidence interval" is used. Figure B.1 shows the results of generating via sampling 48 95% confidence interval. Only three of them do not contain the real value of the parameter (dashed line), which is close to what would be expected, between 2 and 3 intervals should not contain the real value (5% of 48).

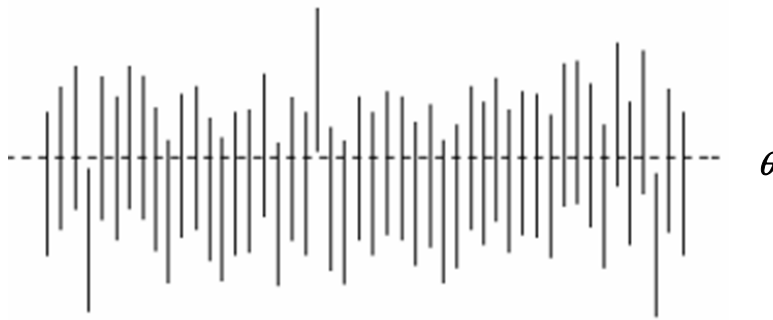


Figure B.1.- Repeated confidence interval (vertical lines) together with the true value of the parameter (horizontal line).

The main problem related to the use of confidence intervals is that exact confidence intervals are available only for the parameters of a few distributions such as normal, log-normal and exponential distributions. For any other distribution, only approximate confidence intervals are available, which are based on the asymptotic normality and lack of bias of maximum likelihood estimators. Table B.2 shows the most frequently used confidence intervals. Exact interval estimates are available for quantiles of any distribution, provided that large enough samples are available (see section 4.2.1 in the main report).

GOODNESS OF FIT TESTS

The last step of the inferential process is to check if the hypotheses under which it has been developed are true. The main hypothesis is the selected probability model. After selecting the model, the point estimation gives the best choices for the values of the parameters subject to some criteria (maximisation of the likelihood function or some other one). Both sets of information define completely the law that supposedly generated the data under study. Nevertheless, the best choice could be 'not good enough'. This is what we try to find out using goodness of fit tests. The main tests are the χ^2 (chi-square) test and Kolmogorov's test.

χ^2 (chi-square) test

The χ^2 test is based on the comparison of the histogram of the data with the estimated PDF. It consists of the following steps:

- group the data in k sets as done when drawing a histogram and count the number of data in each set (O_i),
- compute the probability of each set (p_i) under the assumed probability law. Compute the expected number of data in each set under the assumed probability distribution using the formula $E_i = np_i$,
- compute the discrepancy between what is expected under the assumed model and what has been obtained in the sample according to

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}.$$

- compare this value with the $1-\alpha$ quantile of the χ^2_{k-r-1} distribution (χ^2_α). Typically α is set to 0.05 or 0.01.
 - if $\chi^2 > \chi^2_\alpha$, reject the null hypothesis, which means that the PDF obtained through the estimation process and the data differ so much that it is very unlikely (probability $< \alpha$) that the data could have been generated under the estimated distribution.
 - if $\chi^2 \leq \chi^2_\alpha$, accept the null hypothesis. In this case the agreement between the estimated PDF and the data is good enough to consider that the PDF could have likely generated the data.

Here $k - r - 1$ is the number of degrees of freedom of the χ^2 distribution taken as a reference in the test; r is the number of parameters of the PDF that were estimated from the data to determine the PDF. So, if we consider that a given set of data could follow a normal distribution whose mean is unknown but whose variance is known. To define the PDF completely we estimate only the mean from the data. In this case $r=1$. If we estimate both the mean and the variance from the sample, r would be 2. The χ^2 test is an asymptotic test, it works well with large sample sizes, but it is not recommended to apply it to small data sets (in fact many authors discourage its use when the sample size is below 25 or 30).

Kolmogorov's test

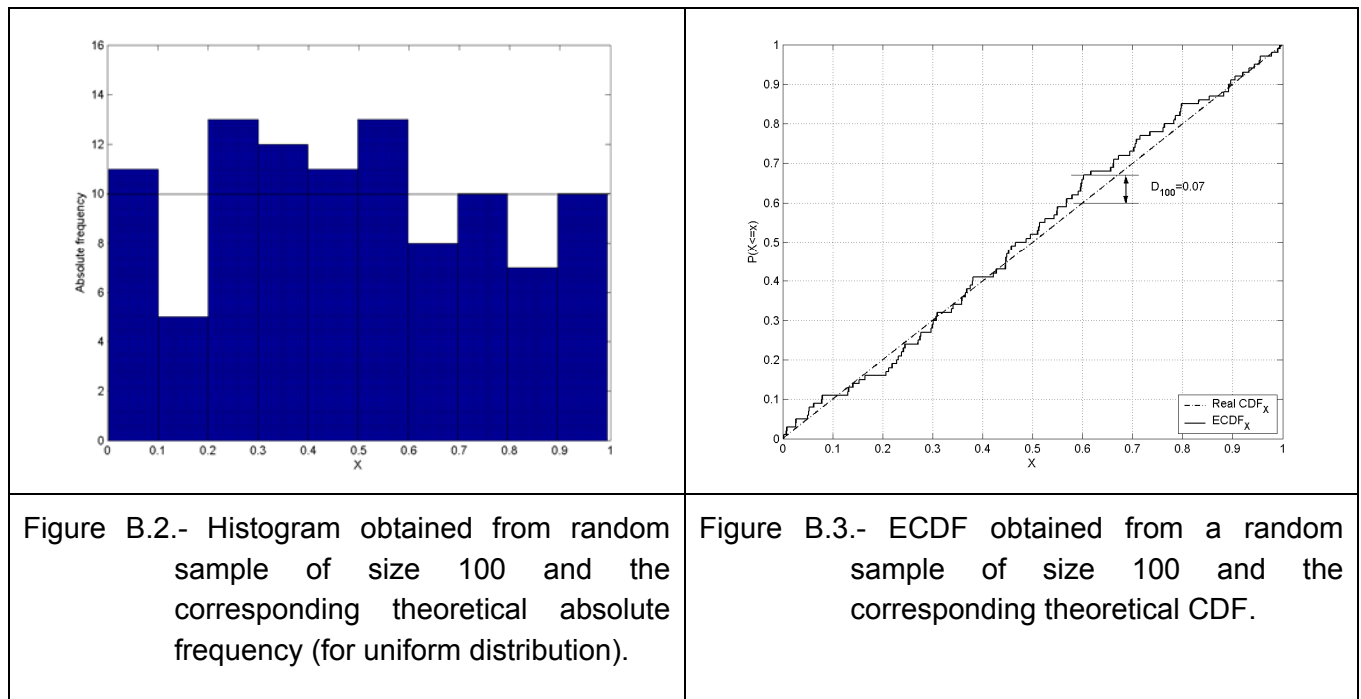
Kolmogorov's test is based on the comparison of the ECDF obtained from the data and the estimated CDF. The steps to perform the test are:

- draw the ECDF based on the data,
- draw the CDF according to the model selected and the estimated parameters,
- compute the maximum vertical distance (D_n) between the ECDF and the CDF,
- compare this value with the $1-\alpha$ quantile ($D(n)_\alpha$) of Kolmogorov's statistic ($D(n)$) distribution for a sample of size n . As usual, α is set to 0.05 or 0.01.
 - if $D_n > D(n)_\alpha$, reject the null hypothesis, which means that the CDF obtained through the estimation process and the data differ so much as to consider very unlikely (probability $< \alpha$) that the data could have been generated under the estimated distribution.
 - if $D_n \leq D(n)_\alpha$, accept the null hypothesis. In this case the agreement between the CDF and the data is good enough as to consider that the CDF could have likely generated the data.

Kolmogorov's test is an exact test that can be applied to any random sample, whatever its size is, though its capability to detect departures from the null hypothesis is quite limited for small sample sizes.

Example:

A random sample of size 100 has been obtained. We assume that it comes from a uniform distribution defined in the interval $[0, 1]$. In order to test this hypothesis we perform the χ^2 test and Kolmogorov's test. In order to perform the former, we plot a histogram of the data set, which is shown in figure B.2, and compare it with what would be expectable from the theoretical PDF (see horizontal line at height 10). Then we compute the quantity $\chi^2 = [(11-10)^2/10 + (5-10)^2/10 + (13-10)^2/10 + \dots + (10-10)^2/10] = 5.4$. Since $\chi^2 = 5.4 \leq 16.9 = \chi_{0.05}^2$ (the value of the statistic chi-square does not exceed the 95% percentile of the χ_9^2 - chi-square distribution with 9 degrees of freedom), the null hypothesis (the data set comes from a uniform distribution defined in the range $[0, 1]$) is accepted.



In order to apply Kolmogorov's test to the same data set, we draw the ECDF and the CDF, and we compute the maximum vertical distance between both curves, see figure B.3. We then compare the value D_{100} obtained with the 95% percentile of Kolmogorov's statistic for sample size 100. Since $D_{100} = 0.070 \leq 0.0136 = D(100)_{0.05}$, we accept the null hypothesis.

REFERENCES

- [1] G. Casella and R.L. Berger (1990). Statistical inference. Duxbury Press.
- [2] A.M. Mood, F.A. Graybill and D.C. Boes (1974). Introduction to the theory of statistics. Third edition. McGraw-Hill. Statistics series.

Annex C: Properties of Quantile Estimators

PROPERTIES OF THE EMPIRICAL ESTIMATOR

The estimator of the α -quantile $\hat{Y}_{\alpha,n} = \inf\{y, \hat{F}_{EE}(y) > \alpha\} = Y_{(\lceil \alpha n \rceil)}$ is biased with the following first two moments

$$E(\hat{Y}_{\alpha,n}) = y_\alpha - \frac{\alpha(1-\alpha)p'(y_\alpha)}{2(n+2)p^3(y_\alpha)} + O\left(\frac{1}{n^2}\right), \quad (C.1)$$

$$\text{Var}(\hat{Y}_{\alpha,n}) = \frac{\alpha(1-\alpha)}{(n+2)p^2(y_\alpha)} + O\left(\frac{1}{n^2}\right), \quad (C.2)$$

and it is asymptotically normal, which means that

$$\sqrt{n}(\hat{Y}_{\alpha,n} - y_\alpha) \xrightarrow{n \rightarrow \infty} N\left(0, \frac{\alpha(1-\alpha)}{p^2(y_\alpha)}\right). \quad (C.3)$$

The variance of this estimator is large, and it increases for extreme quantiles, for which the value of the PDF $f(y_\alpha)$ is small. Moreover, from the asymptotical law, one can see that $P(\hat{Y}_{\alpha,n} \geq y_\alpha) \approx 0.5$.

PROPERTIES OF WILKS ESTIMATOR

The Wilks estimator is the order statistic $Y_{(n-r+1)}$. The following proposition helps establishing the Wilks formula, which connects n (the size of the sample) and r, α, β from $P(Y_{(n-r+1)} > y_\alpha) \geq \beta$.

Proposition:

The number of times n iid rvs (Y_1, \dots, Y_n) exceed a certain threshold y follows a Binomial distribution $B(n, F(y))$, where F is the CDF of the rv Y_i .

This proposition is used as follows: let $(Y_{(1)}, \dots, Y_{(n)})$ be the ordered sample, $Y_{(1)} \leq \dots \leq Y_{(n)}$. The probability of the event $\{j \text{ of the } Y_i \text{ are } > y\}$, $\forall j$ is computed using the binomial distribution:

$$P(j \text{ of the } Y_i \text{ are } > y) = \binom{n}{j} (1 - F(y))^j F(y)^{n-j}. \quad (C.4)$$

For $y = y_\alpha$, we get

$$P(j \text{ of the } Y_i \text{ are } > y_\alpha) = \binom{n}{j} (1-\alpha)^j \alpha^{n-j}. \quad (\text{C.5})$$

On the other hand, the event $\{Y_{(n-r+1)} > y_\alpha\}$ occurs if and only if at least r of the Y_i are $> y_\alpha$.

This leads to

$$\begin{aligned} P(Y_{(n-r+1)} > y_\alpha) &= \sum_{j=r}^n P(j \text{ of the } Y_i \text{ are } > y_\alpha) = \sum_{j=r}^n \binom{n}{j} (1-\alpha)^j \alpha^{n-j} \\ &= \sum_{j=0}^{n-r} \binom{n}{j} \alpha^j (1-\alpha)^{n-j} = 1 - \sum_{j=n-r+1}^n \binom{n}{j} \alpha^j (1-\alpha)^{n-j} \end{aligned} \quad (\text{C.6})$$

As we want $Y_{(n-r+1)}$ to be sure to the level β , i.e. $P(Y_{(n-r+1)} > y_\alpha) \geq \beta$, we obtain the « Wilks formula »

$$\beta = 1 - \sum_{j=n-r+1}^n \binom{n}{j} \alpha^j (1-\alpha)^{n-j} = \sum_{j=0}^{n-r} \binom{n}{j} \alpha^j (1-\alpha)^{n-j} \quad (\text{C.7})$$

It is then possible to compute, for α, β fixed (for instance 95%, 95%), the couples r, n .