Probabilistic risk bounds for the characterization of radiological contamination

Géraud Blatman, Bertrand Iooss, Nadia Pérot

Probabilistic risk bounds for the characterization of radiological contamination

Géraud Blatman*, Bertrand Iooss† and Nadia Pérot‡
* EDF Lab Les Renardières, 77818 Moret-sur-Loing, France
† EDF Lab Chatou, 78401 Chatou, France
‡ CEA Nuclear Energy Division, 13108 Saint-Paul-lès-Durance, France

December 10, 2016

Abstract

The radiological characterization of contaminated elements (walls, grounds, objects) from nuclear facilities often suffers from a too small number of measurements. In order to determine risk prediction bounds on the level of contamination, some classic statistical methods may then reveal unsuited as they rely upon strong assumptions (e.g. that the underlying distribution is Gaussian) which cannot be checked. Considering that a set of measurements or their average value arise from a Gaussian distribution can sometimes lead to erroneous conclusion, possibly underconservative. This paper presents several alternative statistical approaches which are based on much weaker hypotheses than Gaussianity. They result from general probabilistic inequalities and order-statistics based formula. Given a data sample, these inequalities make it possible to derive prediction intervals for a random variable, which can be directly interpreted as probabilistic risk bounds. For the sake of validation, they are first applied to synthetic data samples generated from several known theoretical distributions. In a second time, the proposed methods are applied to two data sets obtained from real radiological contamination measurements.

1 Introduction

In nuclear engineering, as for most of industrial domains, one often faces up with difficult decision making processes, especially when safety issues are involved. In order to consider in a rigorous and consistent way all the environment uncertainties in a decision process, the probabilistic framework offers an invaluable help. For example, a non exhaustive sampling of a process or an object induces some uncertainties that have to be known in order to control their effects.

In particular, the estimation of risk prediction bounds is an important element of a comprehensive probabilistic risk assessment of radioactive elements (e.g. walls, grounds, objects) coming from nuclear industry. The

radiological characterization of contaminated elements in a nuclear facility may be difficult because of practical and/or of strong exploitation constraints, often limiting the number of possible measurements. Nevertheless the estimation of the radioactivity levels is essential to assess the risk of exposure of the nuclear dismantling operators as well as the risk of environment contamination [1].

Drawing up a radiological inventory based on a small number of measurements (say in the order of 10) is a particularly difficult statistical problem. The shortage of data can lead either to a coarse over-estimation, which has large impact on economic costs, or to a coarse under-estimation, which has unacceptable impact in terms of public health and environment protection. In the past, a few works have been realized and presented in the literature. [2] focused on the problem of defining a sampling strategy and assessing the representativeness of the small samples at hand. In the context of irradiated graphite waste, [3] developed a method to assess the radionuclide inventory as precisely as possible with a 2.5% risk of under-assessment. In a recent work, [4] described several sampling methods to estimate the concentration of radionuclides in radioactive waste, by using correlation between different radionuclides activities. When the characterized contamination exhibits some spatial continuity and when the spatial localization of the measurement can be chosen, geostatistical tools can be used, as shown in [5, 6, 7].

In this work, we focus on the uneasy task of radiological characterization based on a small number of data which are assumed to be statistically independent (and non-spatially localized). This task belongs to a quite general class of problems: the statistical analysis of small-size data sample (see for example [8, 9, 10]). In this case, the classical statistical tools turn out to be unsuited. For example, assuming that a set of measurements or their average value arise from a Gaussian distribution can lead to erroneous and sometimes unconservative conclusions. Indeed, if the estimation of the mean value is of interest, normal distribution-based bounds may only be used in the asymptotic limit of a very large sample, and the convergence to this asymptotic regime may be very slow in the presence of a noticeably skewed actual data-generating distribution. Even if some solutions exist to correct this large-size sample requirement, the Gaussian distribution hypothesis can be either non valid, either unjustifiable.

Alternative statistical tools, called concentration inequalities (but also denoted as universal inequalities or robust inequalities), are applicable without knowing the probability distribution of the studied variable. In general, from a sample of data, statistical intervals allow the derivation of [11]:

1. Confidence intervals for the estimation of the mean (or other distribution parameters) of a random variable. For example, we can determine the size of the measurement sample to collect in order to reach a given precision on the average of contamination measures. Such a process allows one to optimize the sampling strategy and offers invaluable economic gains.

2. Prediction intervals for a random variable. For example, we can compute the probability that the value of a point contamination is larger

than a given critical value. In practice, regulatory threshold values are set for different waste categories. Determining the probability that the contaminant value is smaller than a given threshold can be used to predefine the volumes of waste by category. Prediction intervals

3. Tolerance intervals which extend prediction intervals to take into account the uncertainty on the parameters of the variable distribution. A tolerance interval gives the statistical interval within which, with some confidence level, a specified proportion of a sampled population falls.

In this paper, we focus on the second and third intervals (note that confidence intervals are also addressed in [12]), restricting their names to "prediction intervals" in the following for simplicity. Easy to state and easy to use, the Bienaymé-Chebyshev inequality [13] is the most famous probabilistic inequality. Unfortunately, it comes at the expense of extremely large bounds which make this inequality unsuited to practical situations and then not often used. From these considerations, [14] has proposed to use the more efficient (but little-known) Guttman inequality [15]. Even if it requires no additional assumption, the Guttman inequality has however the drawback to require an estimate of the kurtosis value (*i.e.* the fourth-order statistical moment) of the studied variable. In a context of a small amount of data (around ten), a precise estimation of the kurtosis seems to be unrealistic.

In another context, the quality control domain, [16] has developed narrower bounds than the Bienaymé-Chebyshev ones, showing at the same time how the three-sigma rule can be justified (based on a unimodality assumption proven for example in [17]). Starting from this statistical literature as well as from old results about unimodal and convex distributions (see [18] for a more recent reference), several useful inequalities have been identified in [12]. While the latter work also focused on the validity range of each inequality, its results were preliminary; the present paper extends this work to estimate risk prediction bounds in a robust way, with some applications to real radiological characterization problems. Furthermore, we make a connection between this risk bound estimation problem and the problem of computing conservative estimates of a quantile, classically addressed in nuclear thermal-hydraulic safety using the so-called Wilks formula [19]. Comparisons are then performed between the various approaches.

The following section provides all the probabilistic inequalities that we can use to solve our problem. For the sake of validation, all these inequalities are applied in section 3 to synthetic data samples generated from several known theoretical distributions. More precisely, the accuracy of the resulting prediction intervals are compared to those obtained from standard methods such as the Gaussian approximation. Section 4 shows how the probabilistic inequalities can be used in practice, more precisely to analyze radiological contamination measures. A conclusion synthesizes the results of this work.

2 Probabilistic inequalities for prediction intervals

We are interested in the determination of a unilateral prediction interval. It allows one to define a limit value that a variable cannot exceed (or reach, depending on the context) with a given probability level. In a radiological practical context, it can then be used to estimate, on the basis of a few contaminant measures, the quantity of contaminant which does not exceed a safety threshold value.

Mathematically, a unilateral prediction interval for a random variable $X \in \mathbb{R}$ reads:

$$\mathbb{P}(X \geq S) \leq \alpha \tag{1}$$

where $S \in \mathbb{R}$ is the threshold value and $\alpha \in [0, 1]$ is the risk probability. For an absolutely continuous random variable, it is equivalent to the following inequality

$$\mathbb{P}(X \leq S) \geq 1 - \alpha = \gamma \ . \tag{2}$$

In other words, S is a quantile of X of order greater than γ.

In the two following subsections, we introduce some theoretical inequalities which require the knowledge of the mean μ and the standard deviation σ of X, that are therefore supposed to exist. Such inequalities are of the following form:

$$\mathbb{P}(X \geq \mu + t) \leq \left(1 + \frac{t^2}{k\sigma^2}\right)^{-1} \tag{3}$$

where $t \geq 0$ and k is a positive constant.

General hypothesis of all the inequalities: X is absolutely continuous with finite mean and variance.

Hypothesis for their applicability: The sample of X is i.i.d. (independent identically distributed)

2.1 The Gaussian approximation

Provided that the random variable X is normally distributed, the derivation of an unilateral prediction interval related to a given risk probability α is straightforward. Denoting by z_u the quantile of level u of the standard Gaussian distribution $\mathcal{N}(0, 1)$ (whose value can be easily found in standard normal tables or basic statistical software), one gets

$$\mathbb{P}\left(\frac{X - \mu}{\sigma} \ \geq \ z_{1-\alpha}\right) = \alpha \ , \tag{4}$$

which is equivalent to

$$\mathbb{P}\left(X \ \geq \ \mu + \sigma z_{1-\alpha}\right) = \alpha \ . \tag{5}$$

This relationship is a special case of Eq. (3) in which the right hand side is no longer an upper bound but the actual risk probability α and

where $t = \sigma z_{1-\alpha}$. It is easy to show that parameter k is then equal to $z_{1-\alpha}^2 \alpha/(1-\alpha)$.

Note that the so-called Owen's method develop corrected formulas in order to take into account the ignorance of the mean and standard deviation of the variable [20, 10]. However, it is also based on the assumption of normality and strong care must be taken when applied to distributions other than normal. In this case, the application of a Gaussian approximation may provide non conservative bounds though in the presence of a small sample, especially in the case of a significantly skewed random variable X. So-called concentration inequalities are presented in the sequel in order to determine conservative intervals in such situations.

2.2 Concentration inequalities

In probability theory, the concentration inequalities relate the tail probabilities of a random variable to its statistical moments[1]. Therefore, they provide bounds about the deviation of a random variable away from a given value (for example its mean value). The various inequalities come from the information one can give about the random variable (mean, variance, bounds, positiveness, ...). This is a very old research topic in the field of statistics and probability. For example, [21] has reviewed thirteen such classic inequalities. New results were obtained in the past decades based on numerous mathematical works focused on concentration of measure principles (see for example [22]). We restrict our work to three classical inequalities that seem to be the most useful for the radiological characterization problems with small samples.

2.2.1 Bienaymé-Chebychev inequality

The Bienaymé-Chebychev inequality writes [13]:

$$\forall\, t \;\geq\; 0, \quad \mathbb{P}(X \geq \mu + t) \leq \left(1 + \frac{t^2}{\sigma^2}\right)^{-1} \tag{6}$$

which corresponds to Eq. (3) with $k = 1$. As μ and σ are unknown in practical applications, they are replaced with their empirical counterpart $\hat{\mu}$ and $\hat{\sigma}$ (*i.e.* their estimates from the sample values). This inequality does not require any hypothesis on the probability distribution of X.

In fact, Eq. (6) is also known under the name of Cantelli inequality or Bienaymé-Chebychev-Cantelli inequality. This is an extension of the classical Bienaymé-Chebychev inequality (see [22]) where an absolute deviation is considered inside the probability term of Eq. (6). For the two following inequalities, the same rearrangement is made.

Hypothesis of Bienaymé-Chebychev (BC) inequality: None

[1] The first moment of a random variable X is the mean $\mu = E(X)$; the second moment is the variance $\sigma^2 = E[(X-\mu)^2]$; the third moment is the skewness coefficient $\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$; the fourth moment is the kurtosis $\gamma_2 = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$.

2.2.2 Camp-Meidell inequality

The Camp-Meidell inequality writes [23, 16]:

$$\forall\, t \;\geq\; 0, \quad \mathbb{P}(X \geq \mu + t) \leq \left(1 + \frac{9}{4}\frac{t^2}{\sigma^2}\right)^{-1} \tag{7}$$

which corresponds to Eq. (3) with $k = 4/9$. As μ and σ are unknown in practical applications, they are replaced with their empirical counterpart $\hat{\mu}$ and $\hat{\sigma}$.

It is interesting to note that this inequality, in its two-sided version, justifies the so-called "three-sigma rule". This rule is classically used in manufacturing processes, as it states that 95% of a scalar-valued product result X ranges in the interval $[\mu - 3\sigma, \mu + 3\sigma]$. In fact, it has been shown in [17] that this rule is only valid for a result X following a unimodal distribution. Indeed, one proof of the expression (7) is based on the bounding of the distribution function by a linear function [24]. Then, this inequality requires the hypothesis of derivability on the probability distribution of X and unimodality on the probability density function (pdf) of X. It can then be applied to all the unimodal continuous probability laws used in practice (e.g. uniform, Gaussian, triangular, lognormal, Weibull, etc.).

Hypothesis of Camp-Meidell (CM) inequality: Unimodality of the pdf

2.2.3 Van Dantzig inequality

The Van Dantzig inequality reads [24]:

$$\forall\, t \;\geq\; 0, \quad \mathbb{P}(X \geq \mu + t) \leq \left(1 + \frac{8}{3}\frac{t^2}{\sigma^2}\right)^{-1} \tag{8}$$

which corresponds to Eq. (3) with $k = 3/8$. As μ and σ are unknown in practical applications, they are replaced with their empirical counterpart $\hat{\mu}$ and $\hat{\sigma}$.

One can note that this inequality is little known. This can be explained by the low gain obtained with respect to the CM inequality. One demonstration of the expression (8) is based on the bounding of the distribution function by a quadratic function [24]. This inequality requires the hypothesis of second-order derivability on the probability distribution of X and convexity on the density of X. In fact, it can be applied to all the unimodal continuous probability laws in their convex part. The tail of most of the classical pdfs is convex, as for example the exponential law (on all the domain), the triangular law (for $\mu + t \geq$ mode), the Gaussian law (for $t > \sigma$), the Weibull law, etc. However, it is not valid for uniform variables.

Hypothesis of Van Dantzig (VD) inequality: Convexity of the pdf tail

2.2.4 Conservative estimates based on bootstrap

Applying the three previous concentration inequalities require to know the mean μ and the standard deviation σ of the variable under consideration.

In most of practical situations, these quantities are unknown and they are directly estimated from their sample counterparts. However, poor confidence can be given to these estimates when dealing with small sample cases: substituting the actual moments by their sample estimates can indeed lead to overly optimistic results. To overcome this problem, we propose a penalized approach based on bootstrap, which is a common tool in the statistical practice [25].

The principles of the bootstrap variation which is used in this work are as follows: for a given sample of size n, we generate a large number B of resamples, *i.e.* samples made of n values selected randomly with replacement in the original sample. We then compute the empirical mean and standard deviation of each resample, and then apply the inequality that we study. Finally, we obtain B resulting values and can compute some statistics such as high quantiles (of order β which is the confidence value). We can take for example the 95%-quantile (i.e. $\beta = 0.95$) to derive a large and conservative value.

2.3 Using the Wilks formula

We consider the quantile estimation problem of a random variable as stated in Eq. (2), where $\gamma = 1 - \alpha$ is the order of the quantile. This problem is equivalent to the previous one of risk bound estimation (Eq. (1)). The classic (empirical) estimator is based on order statistics derivations [26] of a Monte-Carlo sample. With a small-size sample (typically with less than 100 observations), this estimator gives very imprecise quantile estimate (*i.e.* with large variance), especially for low (less than 5%) and large (more than 95%) orders β [27].

Another point of view consists in calculating a tolerance limit instead of a quantile, thanks to some order statistics theorems [26]. For an upper bound, this provides a majoring value of the desired quantile with a given confidence level (for example 95%). Based on this principle, the Wilks formula [28, 19]) allows us to precisely determine the required sample size in order to estimate, for a random variable, a quantile of order α with confidence level β. It was introduced in the nuclear engineering community by the German nuclear safety institute (GRS) at the beginning of the 1990s [29], and then used for various safety assessment problems (see for example [30], [31] and [2]).

We restrict our explanations below to the one-sided case. Suppose we have an i.i.d. n-sample $X_1, X_2, \ldots, X_n$ drawn from a random variable X. We note $M = \max_i(X_i)$. For M to be an upper bound for at least $100\gamma\%$ of possible values of X with given confidence level β, we require

$$\mathbb{P}[\mathbb{P}(X \leq M) \geq \gamma] \geq \beta \, . \tag{9}$$

The Wilks formula stands that the sample size n must therefore satisfy the following inequality:

$$1 - \gamma^n \geq \beta \, . \tag{10}$$

In Table 1, we present several consistent combinations of the sample size n, the quantile order γ and the confidence level β.

γ	0.9	0.9	0.9	0.95	0.95	0.95	0.95	0.99	0.99
β	0.5	0.9	0.95	0.4	0.5	0.78	0.95	0.95	0.99
n	7	22	29	10	14	30	59	299	459

Table 1: Examples of values given in the first-order case by Wilks formula (Eq. (10)).

Eq. (10) is a first-order equation because the upper bound is set equal to the maximum value of the sample. To extend Wilks formula to higher orders, we consider the n-sample of the random variable X sorted into increasing order: $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(r)} \leq \ldots \leq X_{(n)}$. For all $1 \leq r \leq n$, we set

$$G(\gamma) = \mathbb{P}[\mathbb{P}(X \leq X_{(r)}) \geq \gamma]. \tag{11}$$

According to the Wilks formula, the previous equation can be recast as

$$G(\gamma) = \sum_{i=0}^{r-1} C_n^i \gamma^i (1-\gamma)^{n-i} \,. \tag{12}$$

The value $X_{(r)}$ is an upper-bound of the γ-quantile with confidence level β if $1 - G(\gamma) \geq \beta$.

Increasing the order in the use of Wilks formula helps reduce the variance in the quantile estimator, the price being the requirement of a larger n (according to formula (12) with $\beta = 1 - G(\gamma)$ and fixed γ). The Wilks formula can be used in two ways:

- When the goal is to determine the sample size n, to be measured for a given $\gamma-$quantile with a given level of confidence β, the formula (10) can be used with the fixed order s (corresponding to the s^{th} greater value, $s = n - r + 1$): First order ($s = 1$) gives $r = n$ (maximal value for the quantile), second order ($s = 2$) gives $r = n - 1$ (second larger value for the quantile), etc.
- When a sample of size n is already available, then the formula (10) can be used to determine the couples (α, β) and the orders s for the estimation of the Wilks quantile.

Hypothesis of the Wilks formula: None.

3 Numerical tests

3.1 Introduction

The objective of this section is to assess the degree of conservatism of the various approaches presented above, namely the Gauss-approximation based inequality (Section 2.1), the variations of the concentration inequalities (Section 2.2) and the Wilks method (Section 2.3). To this purpose, we consider four probability distributions which are assumed to generate the data:

- one normal (or Gaussian) distribution with mean 210 and standard deviation 20,
- three lognormal distributions with standard deviations equal to 30, 50 and 70, and with means calculated in such a way that the density maxima (*i.e.* the modes) be all equal to 210.

It is recalled that a random variable X is said to be lognormal if $\log(X)$ is normal. Let us note that the lognormal distribution is classically used for modelling the environmental data, such as a pollutant concentration. Moreover, the hypotheses of all the concentration inequalities are valid for these distributions.

As shown in Figure 1, the distributions exhibit more or less significant skewness. The tests carried out for the most skewed distributions may challenge the robustness of the probabilistic inequalities.

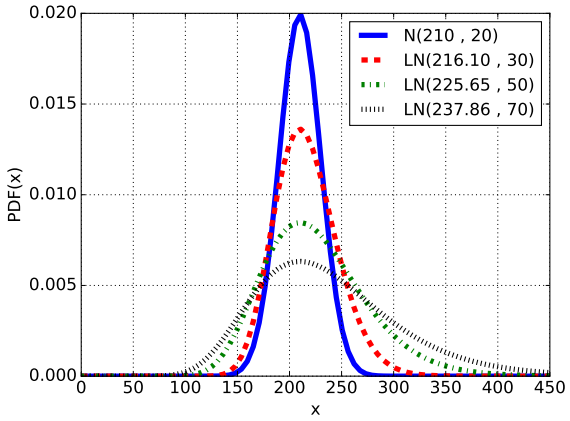


Figure 1: Four different theoretical pdfs for the random variable X.

For each theoretical distribution, we want to estimate the minimum probability that a value randomly drawn from this same distribution exceed a given threshold S. This probability corresponds to the variable α in Eq. (1), and the threshold S to the quantity $\mu + t$ in Eqs. (6), (7) and (8). In other words, the problem can be cast as follows:

$$\text{Estimate } \alpha \text{ such that } \mathbb{P}(X \geq S) \leq \alpha,$$

where X is a random variable following one of the four theoretical distributions.

The numerical analysis is organized into two parts:

1. *The distribution moments are assumed to be perfectly known.* Thus we can compute the exact values of the α-estimates given by the concentration inequalities (Eqs.(6), (7), (8)). We can also estimate α using the Gaussian assumption. However, it does not make sense to use the Wilks formula at this stage since the sample uncertainty is neglected.

2. *The distribution moments are considered to be unknown (realistic case).* The moments are estimated from the data sample at hand. Hence the α estimates are affected by the uncertainty of the sample. In other words, these estimates are random, and they can be characterized by their statistical distributions. In this context, it is relevant to quantify the probability that the estimates underpredict or overpredict the α value obtained theoretically for each method.

3.2 Analysis with known moments

The methods reviewed in this subsection are the three concentration inequalities and the Gaussian approximation. For the concentration inequalities, the risk α is estimated as follows:

$$\alpha \;=\; \left(1 + \frac{t^2}{k\sigma^2}\right)^{-1} \quad \text{with } t \;=\; S - \mu, \tag{13}$$

where μ and σ denote the *exact* values of the distribution mean and standard deviation. For the Gaussian approximation, α is estimated through the evaluation of the cumulative density function of the normal random variable of mean μ and standard deviation σ. The estimates are compared to the actual probabilities that any random realization exceed the threshold S. A different value for S is given for each distribution, so that the exceeding probability α is neither too low nor too high. S is chosen as the quantile of order 95% of the distribution under consideration. (Accordingly the actual value of α is 5%.)

The estimates of the risk α are reported in Table 2. It is observed that, according to the theory, the concentration inequalities are always conservative, constantly overestimating the actual risk level. As expected, the BC formula is the most conservative, followed by the CM one and then by the VD one. CM reveals particularly interesting because a factor two is gained over BC. The degree of conservatism decreases as the distribution skewness increases. The Gaussian approximation reveals of course exact when the distribution is itself Gaussian, but it provides too optimistic estimations of α when the distribution is not Gaussian.

Table 2: Estimates of the risk α obtained from the Gaussian approximation and the concentration inequalities. The true risk is equal to 5%.

Methods	Distributions for X			
	$\mathcal{N}(210, 20)$	$\mathcal{LN}(216.10, 30)$	$\mathcal{LN}(225.65, 50)$	$\mathcal{LN}(237.86, 70)$
Gauss	0.05	0.04	0.04	0.03
BC	0.27	0.25	0.23	0.23
CM	0.14	0.13	0.12	0.12
VD	0.12	0.11	0.10	0.10

3.3 Analysis with unknown moments

3.3.1 Sample moments based risk estimates

In practice, the distribution of the population from which were generated the data is unknown, and hence so are the distribution mean and standard deviation. Therefore those parameters have to be estimated from the data sample at hand in order to derive the risk if exceeding the threshold S. This induces some randomness in the estimates of the risk α since the sample is itself random. As a consequence, it may be possible to more or less noticeably underestimating α in practical situations.

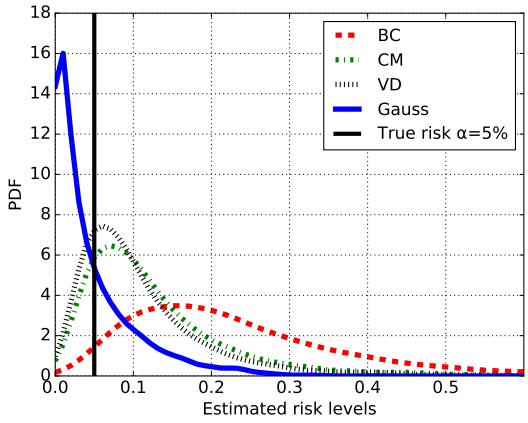
In order to study the level of conservatism of the various approaches when subjected to a sample randomness, we estimate the statistical distribution of the estimates and inspect the proportion of non-conservative estimates, *i.e.* estimates less than the actual value $\alpha = 5\%$. We only consider the most skewed distribution in this subsection, *i.e.* the lognormal one with mean 237.86 and standard deviation 70. From the theoretical distribution, we randomly draw a large number N of samples of a given size n. In this study, we choose $N = 5,000$ and $n \in \{10, 30\}$. For a given distribution and a given sample size n, N risk estimates are computed from the various methods, leading to a sample of N estimates of α. The empirical distribution of this sample is compared to the true risk 5%, and the proportion of non-conservative estimates out of the N values is computed.

The results obtained with $n = 10$ and $n = 30$ are represented in Figures 2 and 3, respectively. It appears that the Gauss approximation strongly underestimates the actual risk level, with about 70% of its results being smaller than the reference risk value 0.05. (It has to be noted that the method provided many negative values of risk levels that were truncated to zero.) The concentration inequalities reveal much more conservative, especially the BC one. Nonetheless, when dealing with samples of size $n = 10$, the CM and the VD formulas yield a non-negligible proportion of too optimistic estimates of α. The three types of inequalities are more conservative when more data are available, that is when the sample size n is set equal to 30.

These results are compared to the Wilks approach at first order. First order (which means that the threshold is the sample maximum) is taken because the sample sizes are extremely small in these tests. According to the formula (10), for an actual risk of level $\alpha = 0.05$ and a sample size $n = 10$, the proportion of non-conservative estimates is less than 0.60 (this value corresponds to the quantity 1-β). This value decreases up to 0.21 when the sample size is equal to $n = 30$. Thus the Wilks method lies between the Gauss approximation and the concentration inequalities for this test case in terms of conservatism. However, an advantage of the method is that it directly gives an upper bound 1-β of the risk of being non-conservative, in contrast to all the other strategies.

3.3.2 Penalized risk estimates based on bootstrap

It has been shown that applying the Gauss and the robust methods by simply substituting the actual moments by their sample estimates can



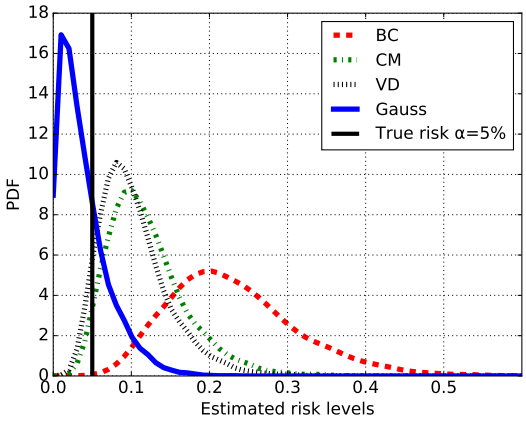
Proportions of non-conservative estimates			
Gauss	BC	CM	VD
0.68	0.02	0.16	0.22

Figure 2: Statistical distributions of the α estimates based on samples of size $n = 10$, for the lognormal distribution of mean 237.86 and standard deviation 70. The actual risk is equal to 5%.

lead to overly optimistic results. To overcome this problem, we propose a penalized approach based on bootstrap (see Section 2.2.4). The principles is as follows. For a given sample of size n, we generate a large number B of resamples (say, $B = 500$). We compute the empirical mean and standard deviation of each resample, and then derive in each case an estimate of α as shown in the previous subsection. This results in a bootstrap set of B estimates of α. In a conservative way, we compute a high quantile of a given order of this set, say 5%. The calculated value serves as an estimate of the risk α.

As in the previous subsection, a focus is given to the lognormal distribution with greatest skewness. The results related to sample sizes n equal to 10 and 30 are plotted in Figures 4 and 5, respectively. As expected, it is observed that the bootstrap-based estimators are significantly more conservative than their "moment-based" counterparts. In particular, for $n = 10$, the level of conservatism was roughly doubled. For $n = 30$, all the concentration inequalities led to very small proportions (less than 1%) of underprediction of the actual risk α. The Gaussian approximation remains quite not reliable though for the skewed distribution under consideration.

The drawback of the conservatism of the robust approaches is that they can yield grossly exaggerated estimations of the risk of exceeding the threshold value, especially the BC formula. On the other hand, the VD formula relies upon assumptions about the density function of the population which are not easy to check in practice. In the lack of further



Proportions of non-conservative estimates			
Gauss	BC	CM	VD
0.72	0.00	0.03	0.06

Figure 3: Statistical distributions of the α estimates based on samples of size $n = 30$, for the lognormal distribution of mean 237.86 and standard deviation 70. The actual risk is equal to 5%.

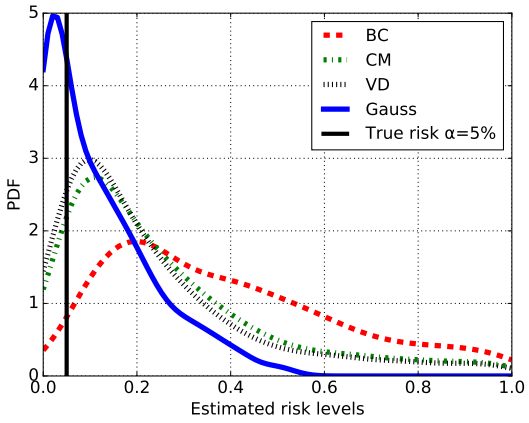
investigations, the combination of the CM inequality with a bootstrap penalization seems to be a reasonable approach.

In conclusion of all these tests based on small-size data samples, inadequacy of the Gaussian approximation has been outlined, while concentration inequalities, used in a conservative manner (by a boostrap technique), reveal strong robustness. The Wilks formula offers the advantage to directly give an upper bound of the risk of being non-conservative, but is not so advantageous when dealing with very small-size sample (low conservatism). If their assumptions can be defended, the Camp-Meidell and Van Dantzig inequalities must be preferred, as the Bienaymé-Chebychev inequality gives most often too conservative results. In the next section, real application cases try to illustrate the practical usefulness of all these tools.

4 Applications

4.1 Case 1: Contamination characterization

This case study concerns the radiological activity (denoted X) in Cesium 137 of a large-size population of waste objects. This characterization enables one to put each waste object in a suitable waste category, e.g. low-activity or high-activity. The reliability of this classification is all the more crucial as it directly affects the total cost of waste management. Indeed,



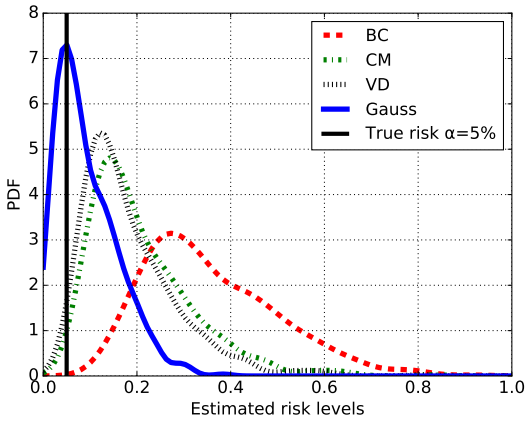
Proportions of non-conservative estimates			
Gauss	BC	CM	VD
0.41	0.01	0.08	0.11

Figure 4: Statistical distributions of the bootstrap α estimates based on samples of size $n = 10$, for the lognormal distribution of mean 237.86 and standard deviation 70. The actual risk is equal to 5%.

putting objects in high-activity waste category is much more expensive than in the low-activity one.

An exhaustive characterization of the population of the waste objects is impossible and only 21 measurements have been done. Reasoning in terms of statistics, it is assumed that this small-size sample ($n = 21$) has been randomly chosen from an unkwown infinite population associated with some probability distribution. Each object of this sample has been characterized by its ^{137}Cs activity measure (in Bq/cm^2). The summary statistics which are estimated with these data are the following: mean $\hat{\mu} = 31.45$, median = 15.4, standard deviation $\hat{\sigma} = 36.11$, Min = 0.83, Max = 156.67. Figure 6 shows the boxplot, the histogram and a smoothed-kernel density of these data. This distribution looks like a lognormal one, with a high asymmetry, a mean much larger than the median, a standard deviation larger than the mean, a lot of low values and a few high values. The extreme value at 156.67 seems to be isolated from the rest of the sample values, but we have no argument to consider it as an outlier. Moreover, the actual data density is considered as unimodal because there is no physical reason to believe that this high value comes from a second population with a different contamination type. Even if it can be subject to discussion, the hypothesis of the convexity of the density tail can also be proposed.

From 21 activity measures, one wants to estimate the proportion of the total population which has a radiological activity larger than a given threshold. Firstly, the quantity of waste objects whose activity does ex-



Proportions of non-conservative estimates			
Gauss	BC	CM	VD
0.32	0.00	0.00	0.00

Figure 5: Statistical distributions of the bootstrap α estimates based on samples of size $n = 30$, for the lognormal distribution of mean 237.86 and standard deviation 70. The actual risk is equal to 5%.

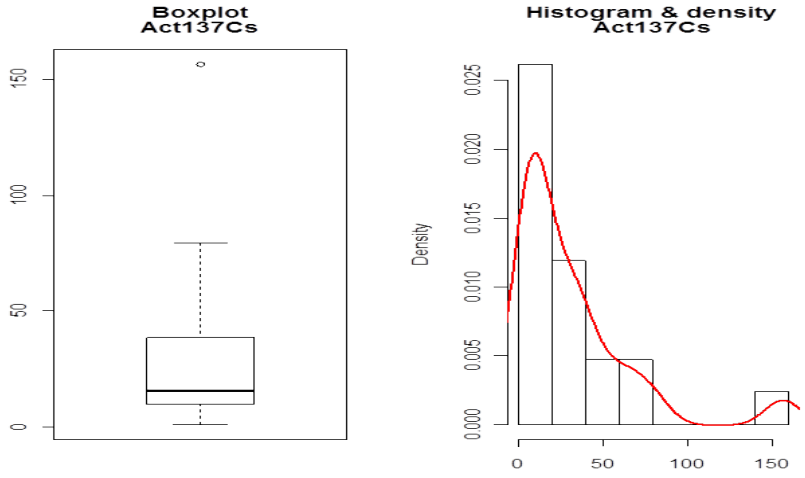


Figure 6: Boxplot (left), histogram and smoothed-kernel density function (right) of the 21 activity measures (Case 1).

ceed the threshold $S = 100$ Bq/cm^2 has to be determined. It could be an

important issue in order to predefine the volume of this waste category.

We did not succeed to fit (with a good confidence degree) a parametric statistical distribution on these data. Then, only distribution-free tools, as those discussed in this paper, can be used to build prediction intervals with a sufficient degree of confidence. The probabilistic inequalities of type (3) are applied, by replacing μ and σ by their estimates:

$$\mathbb{P}(X \geq \hat{\mu} + t) \leq \left(1 + \frac{t^2}{k\hat{\sigma}^2}\right)^{-1} \tag{14}$$

where $t = S - \hat{\mu}$ and the values of k depends on the inequality ($k = 1$ for BC, $k = 4/9$ for CM and $k = 3/8$ for VD). This equation can also be expressed by using S:

$$\mathbb{P}(X \geq S) \leq \left(1 + \frac{(S - \hat{\mu})^2}{k\hat{\sigma}^2}\right)^{-1} \tag{15}$$

The first row of Table 3 gives the risk bound results of the various inequalities for the threshold set equal to 100 Bq/cm^2 and using the empirical estimates of μ and σ. The second row provides bootstrap-based conservative estimates of the risk bound by taking the 95% quantile of $B = 10,000$ risk bounds estimated from B replicas of the data sample. The interpretation of these two results reveals:

- By BC inequality, we obtain from (15) $\left(1 + \frac{(S-\hat{\mu})^2}{\hat{\sigma}^2}\right)^{-1} = 0.2172$ then we coarsely estimate that less than 21.7% of the population has an activity larger than 100 Bq/cm^2, and we can guarantee (at a 95% confidence degree) that less than 44.8% of the population has an activity larger than 100 Bq/cm^2.
- By CM inequality, we obtain from (15) $\left(1 + \frac{(S-\hat{\mu})^2}{\frac{4}{9}\hat{\sigma}^2}\right)^{-1} = 0.1098$ then we coarsely estimate that less than 11% of the population has an activity larger than 100 Bq/cm^2, and we can guarantee (at a 95% confidence degree) that less than 26.5% of the population has an activity larger than 100 Bq/cm^2.
- By VD inequality, we obtain from (15) $\left(1 + \frac{(S-\hat{\mu})^2}{\frac{3}{8}\hat{\sigma}^2}\right)^{-1} = 0.0942$ then we coarsely estimate that less than 9.4% of the population has an activity larger than 100 Bq/cm^2, and we can guarantee (at a 95% confidence degree) that less than 23.3% of the population has an activity larger than 100 Bq/cm^2.

This simple application illustrates the gain we can obtain by using CM or VD inequalities instead of BC: Knowing from the bootstrap conservative estimates that 24% of the population, instead of 45%, of the waste objects can be classified in a high-activity waste category would allow to avoid a too much conservative estimation of the waste management cost.

The Wilks formula is now used to illustrate what kind of statistical information can be inferred for the given data sample. For a sample size $n = 21$, we can estimate two types of quantile:

Table 3: Case 1: Estimates of the risk α obtained from the concentration inequalities and Wilks formula, for different threshold values S.

S	BC	CM	VD	Wilks
100	0.217	0.110	0.094	-
100 $\beta = 0.95$	0.448	0.265	0.233	-
156.67 $\beta = 0.95$ $\beta = 0.78$	0.169 0.118	0.083 0.056	0.071 0.048	0.133 0.071
79.67 $\beta = 0.95$ $\beta = 0.999$	0.665 0.877	0.469 0.760	0.427 0.728	0.207 0.427

- An unilateral first order $\gamma-$quantile with a level of confidence β, and then we have to deduce $\alpha = 1-\gamma$ and β from the equation (10). We obtain the following solutions:
  $\mathbb{P}[\mathbb{P}(X \leq 156.67) \geq 0.896] \geq 0.9;\ (\alpha, \beta) = (10.4\%, 90\%)$,
  $\mathbb{P}[\mathbb{P}(X \leq 156.67) \geq 0.867] \geq 0.95;\ (\alpha, \beta) = (13.3\%, 95\%)$.

- An unilateral second order $\gamma-$quantile with a level of confidence β, and then we have to deduce $\alpha = 1-\gamma$ and β from the equation (12) with $s = 2$ and $r = n-1$. We obtain the following potential solutions:
  $\mathbb{P}[\mathbb{P}(X \leq 79.67) \geq 0.827] \geq 0.9;\ (\alpha, \beta) = (17.3\%, 90\%)$,
  $\mathbb{P}[\mathbb{P}(X \leq 79.67) \geq 0.793] \geq 0.95;\ (\alpha, \beta) = (20.7\%, 95\%)$.

In Table 3 (rows 3 and 4), we compare these results with those of the concentration inequalities by adjusting the corresponding thresholds S. Indeed, comparisons cannot be made with $S = 100$ because Wilks formula can only be applied with a quantile value coming from the data sample values. This is the major drawback of the Wilks formula. For the low-quantile case (79.67, in line 4), *i.e.* large risk bounds, Wilks formula shows its relevance because it gives always less conservative results than BC, CM and VD inequalities: The gain is a factor of two. However, in the high-quantile case (156.67, in line 3), *i.e.* small risk bounds, CM and VD inequalities provide less conservative results with 8.3% and 7.1% of the population that can have an activity higher than 156.67 Bq/cm^2 and 13.3% for Wilks method with a confidence β of 95% (resp. 5.6% and 4.8% for CM and VD, 7.1% for Wilks method with a confidence β of 78%): The gain with VD is a factor of two. Same results are also obtained by showing the values of β obtained with Wilks using the conservative α result obtained by VD inequality.

4.2 Case 2: H_2 flow rate characterization for drums of radiactive waste

Some categories of radioactive waste drums may produce hydrogen gas because of the radiolyse reaction on organic matter like PVC, Polyethylene or cellulose mixed with α-emitters in the waste. The evaluation of the hydrogen flow rate (denoted X in l/drum/year) produced by radioactive waste drum is required for their disposal in final waste repositories. However, considering the time required for the H_2 flow rate measurement of only one drum (more than one month) and the need to characterize a population of several thousands of drums, only a small-size ($n = 38$) randomly chosen sample have been measured.

The summary statistics which are estimated with the data of H_2 flow rate are the following: mean $\hat{\mu} = 2.18$, median $= 1.43$, standard deviation $\hat{\sigma} = 2.67$, Min $= 0.02$, Max $= 13.97$. Figure 7 shows the boxplot, the histogram and a smoothed-kernel density of these data. As for the Case 1, distribution looks like a lognormal one, with a high asymetry, a mean much larger than the median, a standard deviation larger than the mean, a lot of low values and a few high values. The extreme value at 13.97 seems to be isolated from the rest of the sample values, but we have no argument to consider it as an outlier. The fit of a lognormal distribution was not rejected by a Kolmogorov-Smirnov test (with the threshold $\alpha = 5\%$): $X = \mathcal{LN}(0.23, 1.16)$. Therefore, the density can be considered as unimodal and the hypothesis of the convexity of the density tail is also accepted. We can directly estimate the 95%-quantile from the theoretical distribution $X = \mathcal{LN}(0.23, 1.16)$:

$$q_{95\%} = 8.4827 \text{ l/drum/year}.$$

However, due to the small size of data that served to the pdf fit, poor confidence can be given to this value and its justification to safety authorities could be difficult.

The first row of Table 4 gives the risk bound results of the different inequalities for the threshold $S = 10$ l/drum/year and using the empirical estimates of μ and σ. The second row provides bootstrap-based conservative estimates of the risk bound by taking the 95% quantile of $B = 10,000$ risk bounds estimated from B replicas of the data sample. The interpretation of these two results reveals:

- By BC inequality, we coarsely estimate that less than 10.5% of the population has an activity larger than 10 l/drum/year, and we can guarantee (at a 95% confidence degree) that less than 21.2% of the population has a H_2 flow rate larger than 10 l/drum/year.

- By CM inequality, we coarsely estimate that less than 5% of the population has an activity larger than 10 l/drum/year, and we can guarantee (at a 95% confidence degree) that less than 10.7% of the population has a H_2 flow rate larger than 10 l/drum/year.

- By VD inequality, we coarsely estimate that less than 4.2% of the population has an activity larger than 10 l/drum/year, and we can

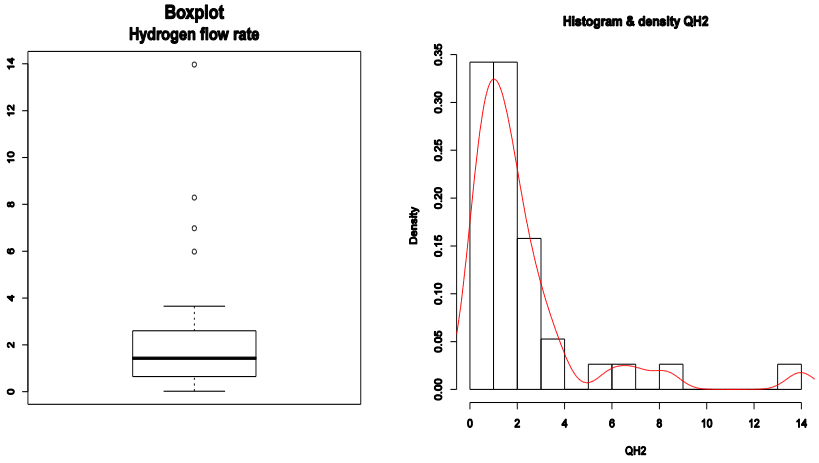


Figure 7: Boxplot (left), histogram and smoothed-kernel density function (right) of the 38 hydrogen flow rates (Case 2).

guarantee (at a 95% confidence degree) that less than 9.2% of the population has a H_2 flow rate larger than 10 l/drum/year.

As for case 1, the gain (here, about 12%) we can obtain by using CM or VD inequality instead of BC is relatively important for the estimation of the waste management cost.

Table 4: Case 2: Estimates of the risk α obtained from the concentration inequalities and Wilks formula, for different threshold values S.

S	BC	CM	VD	Wilks
10	0.105	0.050	0.042	-
10				
$\beta = 0.95$	0.212	0.107	0.092	-
13.97				
$\beta = 0.95$	0.099	0.047	0.040	0.076
$\beta = 0.78$	0.070	0.032	0.027	0.040
8.29				
$\beta = 0.95$	0.315	0.170	0.147	0.119
$\beta = 0.98$	0.360	0.200	0.174	0.147

The Wilks formula is at present used to illustrate for the given data sample what kind of statistical information can be inferred. For a sample size $n = 38$, we can estimate two types of quantile :

- An unilateral first order $\gamma-$quantile with a level of confidence β, and then we have to deduce $\alpha = 1-\gamma$ and β from the equation (10). We

obtain the following solutions:
$\mathbb{P}[\mathbb{P}(X \leq 13.97) \geq 0.941] \geq 0.9;\ (\alpha, \beta) = (5.9\%, 90\%)$,
$\mathbb{P}[\mathbb{P}(X \leq 13.97) \geq 0.924] \geq 0.95;\ (\alpha, \beta) = (7.6\%, 95\%)$.

- An unilateral second order $\gamma-$quantile with a level of confidence β, and then we have to deduce $\alpha = 1 - \gamma$ and β from the equation (12) with $s = 2$ and $r = n - 1$. We obtain the following potential solutions:
$\mathbb{P}[\mathbb{P}(X \leq 8.29) \geq 0.901] \geq 0.9;\ (\alpha, \beta) = (9.9\%, 90\%)$,
$\mathbb{P}[\mathbb{P}(X \leq 8.29) \geq 0.881] \geq 0.95;\ (\alpha, \beta) = (11.9\%, 95\%)$.

In Table 4 (lines 3 and 4), we compare these results with those of the concentration inequalities by adjusting the corresponding thresholds S. Indeed, comparisons cannot be made with $S = 10$ because Wilks formula can only be applied with a quantile value coming from the data sample values. This is the major drawback of the Wilks formula. For the low-quantile case (8.29, in line 4), *i.e.* large risk bounds, Wilks formula shows its relevance because it gives always less conservative results than BC, CM and VD inequalities. However, in the high-quantile case (13.97, in line 3), *i.e.* small risk bounds, CM and VD inequalities provide less conservative results: The gain with VD is a factor of two. Same results are also obtained by showing the values of β obtained with Wilks using the conservative α result obtained by VD inequality.

5 Conclusion

In this paper, we studied several statistical tools to derive risk prediction and tolerance bounds in the context of nuclear waste characterization. The main challenge was related to the small number of data which are usually available in real situations. The normal assumption is then generally unfounded, especially in the case of strongly asymmetrical distribution of the data which are frequently observed in real characterization studies. Much narrower bounds exist in the statistical literature and this paper has highlighted them. Moreover, these are distribution-free tools and no strong assumptions are needed, e.g. with respect to the normality of the distribution of the studied variable. These tools are distribution statistical devices which can provide practical confidence bounds for radiological probabilistic risk assessment.

A few concentration inequalities, used in a conservative way (by a boostrap technique), reveal strong robustness. However, the prediction bounds given by the standard Bienaymé-Chebyshef inequality are very broad: then, their use in risk assessment leads to unnecessary significant conservatism. If their assumptions (unimodality and tail convexity of the pdf) can be justified, the Camp-Meidell and Van Dantzig inequalities must be preferred. Without any assumption, the Wilks formula offers the advantage to directly give an upper bound of the risk of being non-conservative, but is not so advantageous when dealing with very small-size samples or low risk bounds. Indeed, the excess of conservatism can be greater than when using the concentration inequalities. Moreover, the Wilks formula could suffer from a lack of flexibility in practical situations.

As perspectives, more recent concentration inequalities [32, 22] can be studied and can potentially give much narrower intervals. Numerical comparisons with Owen's method [10] have also to be developed. For another issue, it is also shown in [12] how to use probabilistic inequalities to determine the precision on the estimation of the mean of the random variable from a measurement sample. With this kind of inequalities, we can find the minimal number of measurements to realize in order to reach a given confidence level on the mean estimate. In conclusion, applications of these tools could be numerous, in all safety issues based on expensive experimental processes. Further research works and applicative studies could lead to develop useful guides for practitioners, in particular in the nuclear dismantling context.

Acknowledgments

Earlier work that inspired this article were published in [12]. The authors wish to thank Hervé Lamotte, Alexandre Le Cocguen, Dominique Carré and Ingmar Pointeau from the CEA Department of Nuclear Services and Thierry Advocat, head of the CEA GFDM research program to permit the use of the H_2 flow rates data from drums of radioactive waste. We also thank Léandre Brault for many useful comments on a first version of this paper.

References

[1] J. Attiogbe, E. Aubonnet, L. De Maquille, P. De Moura, Y. Desnoyers, D. Dubot, B. Feret, P. Fichet, G. Granier, B. Iooss, J-G. Nokhamzon, C. Ollivier Dehaye, L. Pillette-Cousin, and A. Savary. *Soil radiological characterisation methodology.* CEA-R–6386. Commissariat à l'énergie atomique et aux énergies alternatives (CEA). CEA Marcoule Center, Nuclear Energy Division, Radiochemistry and Processes Department, Analytical Methods Committee (CETAMA), France, 2014.

[2] N. Pérot and B. Iooss. Quelques problématiques d'chantillonnage statistique pour le démantèlement d'installations nucléaires. In *Conférence* $\lambda\mu$16, Avignon, France, october 2008.

[3] B. Poncet and L. Petit. Method to assess the radionuclide inventory of irradiated graphite waste from gas-cooled reactors. *Journal of Radioanalytical and Nuclear Chemistry*, 298:941–953, 2013.

[4] B. Zaffora, M. Magistris, G. Saporta, and F. La Torre. Statistical sampling applied to the radiological characterization of historical waste. *EPJ Nuclear Sci. Technol.*, 2:11, 2016.

[5] N. Jeannée, Y. Desnoyers, F. Lamadie, and B. Iooss. Geostatistical sampling optimization of contaminated premises. In *DEM - Decommissionning challenges: an industrial reality?*, Avignon, France, 2008.

[6] Y. Desnoyers, J-P. Chilès, D. Dubot, N. Jeannée, and J-M. Idasiak. Geostatistics for radiological evaluation: study of structuring of extreme values. *Stochastic Environmental Research and Risk Assessment*, 25:1031–1037, 2011.

[7] A. Bechler, T. Romary, N. Jeannée, and Y. Desnoyers. Geostatistical sampling optimization of contaminated facilities. *Stochastic Environmental Research and Risk Assessment*, 27:19671974, 2013.

[8] A.R. Brazzale, A.C. Davison, and N. Reid. *Applied asymptotics - Case studies in small-sample statistics*. Cambridge University Press, 2007.

[9] P-C. Pupion and G. Pupion. *Méthodes statistiques applicables aux petits échantillons*. Hermann, 2010.

[10] E.G. Schilling and D.V. Neubauer. *Acceptance sampling in quality control*. CRC Press, second edition, 2009.

[11] G.J. Hahn and W.Q. Meeker. *Statistical intervals. A guide for practitioners*. Wiley-Interscience, 1991.

[12] G. Blatman and B. Iooss. Confidence bounds on risk assessments - application to radiological contamination. In *Proceedings of the PSAM11 ESREL 2012 Conference*, pages 1223–1232, Helsinki, Finland, june 2012.

[13] R. Nelson. *Probability, stochastic processes, and queueing theory: The mathematics of computer performance modeling*. Springer, 1995.

[14] G. Woo. Confidence bounds on risk assessments for underground nuclear waste repositories. *Terra Research*, 1:79–83, 1988.

[15] L. Guttman. A distribution-free confidence interval for the mean. *The Annals of Mathematical Statistics*, 19:410–413, 1948.

[16] F. Pukelsheim. The three sigma rule. *The American Statistician*, 48(2):88–91, 1994.

[17] D.F. Vysochanskii and Y.I. Petunin. Justification of the 3σ rule for unimodal distribution. *Theory of Probability and Mathematical Statistics*, 21:25–36, 1980.

[18] S. Dharmadhikari and K. Joag-dev. *Unimodality, convexity, and applications*. Academic Press, Inc, 1988.

[19] W.T. Nutt and G.B. Wallis. Evaluation of nuclear safety from the outputs of computer codes in the presence of uncertainties. *Reliability Engineering and System Safety*, 83:57–77, 2004.

[20] D.B. Owen. *Factors for one-sided tolerance limits and for variables sampling plans*. SCR-607. Sandia Corporation Monograph, 1963.

[21] I.R. Savage. Probability inequalities of the Tchebycheff type. *Journal of Research of the National Bureau of Standards-B. Mathematics and Mathematical Physics*, 65B:211–226, 1961.

[22] S. Boucheron, G. Lugosi, and S. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

[23] B. Meidell. Sur un problème du calcul des probabilités et les statistiques mathématiques. *Compte-Rendu de l'Académie des Sciences*, 175:806–808, 1922.

[24] D. Van Dantzig. Une nouvelle généralisation de l'inégalité de Bienaymé (extrait d'une lettre à M. M. Fréchet). *Annales de l'Institut Henri Poincaré*, 12:31–43, 1951. Available at: http://archive.numdam.org.

[25] B. Efron and R.J. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, 1993.

[26] H.A. David and H.N. Nagaraja. *Order statistics*. Wiley, New-York, third edition, 2003.

[27] C. Cannamela, J. Garnier, and B. Iooss. Controlled stratification for quantile estimation. *Annals of Apllied Statistics*, 2:1554–1580, 2008.

[28] S.S. Wilks. Determination of sample sizes for setting tolerance limits. *Annals of Mathematical Statistics*, 12:91–96, 1941.

[29] E. Hofer. Probabilistische unsicherheitsanalyse von ergebnissen umfangreicher rechenmodelle. *GRS-A-2002*, 1993.

[30] A. de Crécy, P. Bazin, H. Glaeser, T. Skorek, J. Joucla, P. Probst, K. Fujioka, B.D. Chung, D.Y. Oh, M. Kyncl, R. Pernica, J. Macek, R. Meca, R. Macian, F. D'Auria, A. Petruzzi, L. Batet, M. Perez, and F. Reventos. Uncertainty and sensitivity analysis of the LOFT L2-5 test: Results of the BEMUSE programme. *Nuclear Engineering and Design*, 12:3561–3578, 2008.

[31] E. Zio and F. Di Maio. Bootstrap and order statistics for quantifying thermal-hydraulic code uncertainties in the estimation of safety margins. *Science and Technology of Nuclear Installations*, 2008, Article ID 340164, 9 pages, 2008.

[32] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.