ELSEVIER

Analytica Chimica Acta 446 (2001) 429–438

ANALYTICA
CHIMICA
ACTA

www.elsevier.com/locate/aca

Determining environmental standards using bootstrapping, bayesian and maximum likelihood techniques: a comparative study

Frederik A.M. Verdonck[a,*], Joanna Jaworska[b], Olivier Thas[a], Peter A. Vanrolleghem[a]

[a] *Department of Applied Mathematics, Biometrics & Process Control (BIOMATH), Ghent University,*
Coupure Links 653, B-9000 Gent, Belgium
[b] *Procter & Gamble, ETC, Temselaan 100, B-1853 Strombeek-Bever, Belgium*

Received 7 November 2000; received in revised form 14 February 2001; accepted 19 February 2001

Abstract

Environmental standards must be set in ways which give full recognition to all sources of uncertainty and variability of the toxicity data used to derive these standards. Toxicity data such as NOECs form a variability distribution describing species sensitivity distribution (SSD). In EU environmental regulations the 5th-percentile of SSD is used to set the quality criteria. In this paper, a comparison is made between the application of techniques characterising uncertainty and variability (bootstrap, maximum likelihood estimation (MLE) and Bayesian approaches) using small toxicity data sets to calculate the 5th-percentile. Estimating lower and upper uncertainty bounds of a specific percentile gives different results when different methods are used. Bayesian and MLE methods were found to be superior to parametric bootstrapping because they are easier to use and not so computationally intensive. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Uncertainty; Variability; Bootstrapping; Environmental standards; Risk assessment; Small sample size

1. Introduction

The goal of a comprehensive risk assessment is to estimate the probability and the extent of adverse effects occurring to man, animals or ecological systems due to possible exposure(s) to chemicals. The assessment of whether a chemical presents a risk to organisms in the environment is based on the comparison of an environmental concentration with a predicted no effect concentration to ecosystems. The predicted no effect concentration is determined based on no observable effect concentration (NOEC) toxicity data testing the sensitivity of an organism towards a chemical. Various species sensitivities towards a chemical can be captured in a variability distribution, called species sensitivity distribution (SSD). From this distribution of species sensitivities, a hazardous concentration is identified at which a certain percentage *p* of all species is assumed to be affected. The hazardous concentration is also used in quality standard setting.

In the deterministic framework of risk assessments, the predicted no effect concentration is a single value. In the probabilistic framework, the predicted no effect concentration is determined from the 5th-percentile of the SSD. One uses the lower 95% confidence bound of the estimated percentage to ensure that the specified level of protection is achieved.

A distinction ought to be made between variability and uncertainty (or confidence level) of the SSD

[*] Corresponding author. Tel.: +32-92645937;
fax: +32-92646220.
E-mail address: frederik.verdonck@biomath.rug.ac.be
(F.A.M. Verdonck).

PII: S0003-2670(01)00938-2

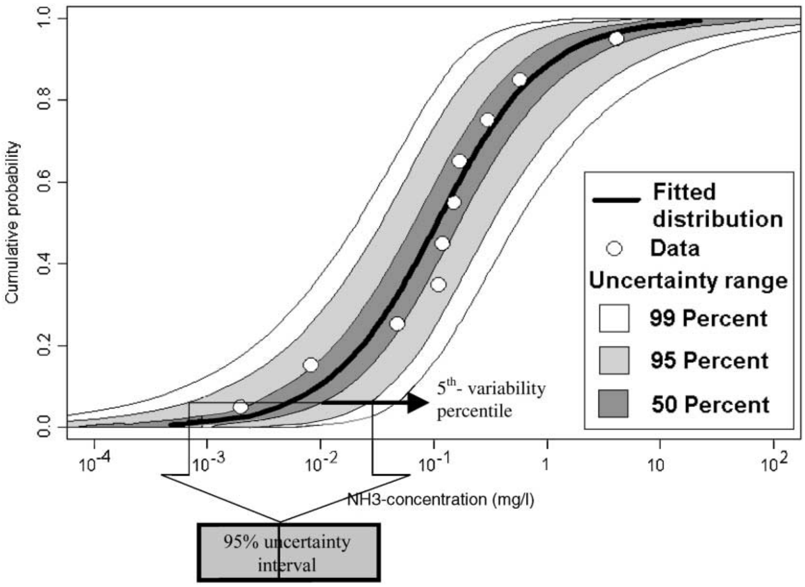


Fig. 1. Example of an uncertainty or confidence band around a cumulative variability distribution function (number of samples = 10, log–logistic distribution).

(Fig. 1). Variability represents heterogeneity or diversity in a well characterised population. Fundamentally a property of nature, variability is usually not reducible through further measurement or study (e.g. variation of chemical concentrations throughout the year due to river flow variability, e.g. species sensitivity towards a chemical). Uncertainty represents partial ignorance or lack of perfect information about poorly characterised phenomena or models, which is sometimes reducible through further measurement or study (e.g. measurement error) [1]. In case of a SSD it is uncertainty of the true shape of a distribution not limited by the sample size.

Several techniques can be used to characterise variability and uncertainty: bootstrapping, the maximum likelihood estimation method (MLE) and Bayesian approaches. Fig. 1 gives an example of the construction of an uncertainty band on a cumulative distribution function based on a limited data set ($n = 10$). The cumulative distribution function well illustrates the fact that of increasing concentrations on a community of species have increasing effect on organisms. For each percentile of the variability distribution, a confidence or uncertainty interval can be calculated (e.g. 95% uncertainty interval for the 5th-variability percentile in Fig. 1).

This area of quantitative risk analysis is currently an active area of research, but mainly methods from classical statistics, such as bootstrap [2,3] or maximum likelihood approach [4] have been applied so far, with an emphasis on parametric analyses. Parametric bootstrapping and maximum likelihood methods were found to produce similar results (for sample sizes 5, 10 and 20) [3]. Jagoe and Newman [5] compared the non-parametric bootstrapping (resampling) with the maximum likelihood method (assuming lognormal distributed data). The parametric method was found to be superior to the resampling, only in the case of lognormally distributed data. Newman et al. [6] proposed non-parametric bootstrapping as the best technique (for sample sizes larger than 20) because no assumptions have to be made on underlying distributions. But, so far all these techniques together have not been compared for small data sets (e.g. sample size = 20 or less).

Aldenberg and Jaworska [7] compared Bayesian and MLE approaches for the Gaussian (normal) model (for several sample sizes). Despite vastly different numerical schemes both approaches lead to identical answers.

In practice, data sets on toxicity tests are scarce and if available often only at small sample sizes. As a consequence, this raises the question: “Given small sample sizes, which techniques are most suitable and which parametric or non-parametric distribution should be used?” To try to answer this question, several methods to estimate the 5th-percentile of an estimated distribution and its confidence interval are compared.

2. Methods

After first giving some terminology, further explanation is given on the determination of the non-parametric and parametric percentiles. Next, a technical overview is given of the statistical methods for characterising variability and uncertainty.

The terminology used in Section 4 is visualised in Fig. 2 as an uncertainty/confidence bar of the 5th-variability percentile. This bar is the 90° left rotation of the horizontal bar in Fig. 1. Cullen and Frey [1] summarise several possible methods for computing non-parametrically the percentile of an observed data set. These methods are referred to as “plotting positions”. The plotting position is an estimate of the cumulative probability of a data point. First, rank ordering the data is needed. Then, the mean plotting system calculates the cumulative probabilities of a point x_i as follows: $F_x(x_i) = (i)/(n+1)$, or alternatively $F_x(x_i) = (i-0.5)/(n)$ which is known as Hazen plotting. In the formulae, *i* stands for the rank order and *n* stands for the total number of samples. Cullen and Frey [1] described that mean plotting system gives erratic results for small sample sizes. Once the observed data set is plotted, percentiles can be calculated taking the inverse (interpolated) empirical distribution function.

In parametric methods, a percentile can be calculated depending on the parametric distribution according to the following equation:

$$\alpha^{\text{th}}\text{-percentile} = \mu - K(s)$$

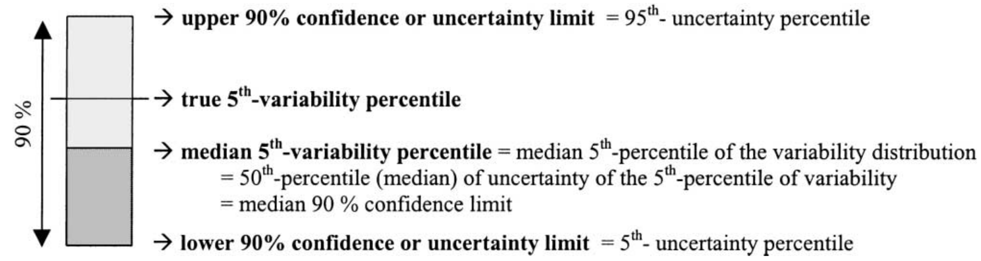
where μ denotes the mean of the data set; σ the standard deviation of the data set; *K* denotes a tabulated extrapolation factor; *K*-values for lognormal distribution and 5th-percentile can be found in [7]; *K*-values for the log–logistic distribution and 5th-percentile can be found in [8].

The use of parametric distributions is one way to model that smaller or higher values than those present in the data set may occur in the real system. Frey and Rhodes [3] showed that the uncertainty in the tails of frequency distributions could be large especially with small data sets. This statement appropriately reflects the limitations of extrapolating from a small data set to the tails of a parametric distribution.

Several methods can determine uncertainty bands on a variability distribution. The difference between these methods and the importance of the effects caused by choosing the wrong distribution will be investigated. The techniques are introduced below.

2.1. Bootstrapping

A detailed description of the bootstrapping method can be found in literature [1,2,9]. Given a data set of sample size *n*, the general approach in bootstrap simulation is to assume a non-parametric or parametric (e.g. lognormal, triangular, etc.) distribution which



→ upper 90% confidence or uncertainty limit = 95th- uncertainty percentile
→ true 5th-variability percentile
90 %
→ median 5th-variability percentile = median 5th-percentile of the variability distribution
= 50th-percentile (median) of uncertainty of the 5th-percentile of variability
= median 90 % confidence limit
→ lower 90% confidence or uncertainty limit = 5th- uncertainty percentile

Fig. 2. Terminology of uncertainty and variability results.

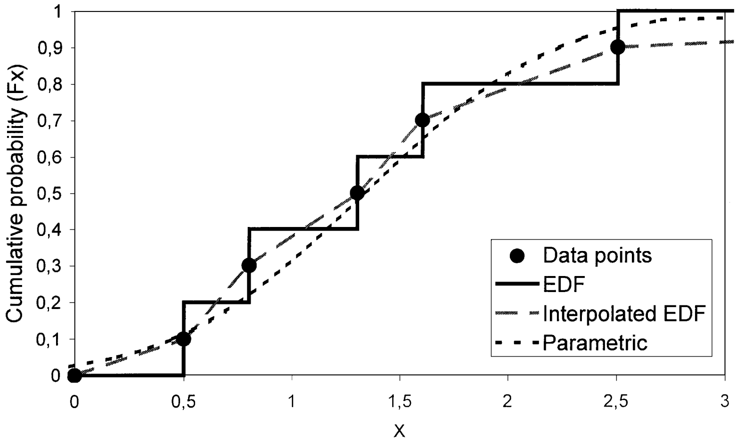


Fig. 3. EDF, interpolated EDF and a parametric fit for a given data set.

describes the quantity of interest, to perform r replications (e.g. $r = 5000$) of the original data set by randomly drawing, with replacement, n values, and then calculate r values of the statistic of interest.

Different types of bootstrapping were studied: two non-parametric techniques, each with two different plotting systems for constructing an empirical cumulative distribution function, and one parametric technique (assuming the lognormal distribution). More details can be found below.

2.1.1. Non-parametric bootstrapping

One approach is to use the actual data set itself and to randomly select, with replacement, the actual values of the data set. This is sometimes referred to as resampling. The data can be represented via an empirical distribution function (EDF). The solid dark line in Fig. 3 gives an empirical distribution function for a given data set.

A second approach is to fit an interpolated empirical cumulative distribution function (interpolated EDF $\neq$ EDF) to the data. The broken, grey line in Fig. 3 gives an interpolated EDF. Such a distribution has minimum and maximum values, which can be constrained by the minimum, and maximum values in the data set, or have to be determined explicitly. In the application of standard setting, zero can be considered as a minimum.

As introduced earlier, there are several “plotting positions” for constructing an EDF and an interpolated EDF: mean and Hazen plotting.

2.1.2. Parametric bootstrapping

A third approach is to assume a parametric distribution rather than an empirical distribution. This approach is called parametric bootstrapping. Efron and Tibshirani [9] discussed this method in detail. The broken, black line in Fig. 3 represents a fitted exponential distribution for a given data set.

Each approach will lead to a different estimate of the confidence interval. Non-parametric or distribution-free approaches do not require assumptions regarding the probability model for the underlying population distribution. However, they also tend to yield wider confidence intervals than parametric methods do.

2.2. Maximum likelihood estimation method (MLE)

The general idea of MLE is to choose an estimator for the parameter(s) in a distribution (e.g. mean, 5th-percentile, etc.) so as to maximise the likelihood of the sample data. An ML-estimator can be thought of as an estimate for which the observed data are most ‘likely’. From a statistical point of view, the method of maximum likelihood is considered to be more robust (with some exceptions) and yields estimators with good statistical properties. In addition, they provide efficient methods for quantifying uncertainty through confidence bounds. Although the methodology for maximum likelihood estimation is simple, the implementation is mathematically intense. More information on the strengths of the MLE can be found in [1].

2.3. Bayesian approaches

The Bayesian statistical method reverses the role of sample and model, the sample is fixed and unique, and the model itself is uncertain. This statistical viewpoint corresponds better to the practical situation the individual researcher is facing; there is only one sample and there are doubts what model to use, or, if the model is chosen, what values the parameters will take. The uncertainty of the model is modelled by assuming that the parameters of the model are distributed [7].

If one assumes parameter values to be distributed, one has to presuppose a so-called (in this case a non-informative) prior distribution for the parameters, to specify the initial state of knowledge about them, before the data are used. The prior distribution is transformed into the so-called posterior distribution by multiplication with the classical likelihood function, by which the information in the data is introduced. This is essentially Bayes' theorem. The posterior distribution summarises our increase in knowledge about the parameters due to observing the data. A Bayesian simulation focuses on the evaluation of the joint posterior distribution of the parameters. For further technical details the reader is referred to [10].

3. Data sets

Two different kinds of data sets were considered.

The first, a synthetic one, contains 20 positive values (see Table 1) drawn randomly from a lognormal distribution ($\exp(N(\mu, \sigma))$ with $\mu = 2$ and $\sigma = 1$). The arithmetic mean of the parent distribution equals $\exp(\mu + 0.5\sigma^2) = 12.2$, approximately, and the arithmetic mean of this sample equals 14, exactly.

The second series of data sets come from literature laboratory and field measurements. They consist of toxicity database of Cu (Copper — 20 data points) and linear alkyl benzene sulfonate (LAS) — 17 data points) which can be found in [11].

Table 1
Data set 1: a hypothetical data set of 20 samples, data values drawn from a lognormal distribution of the form $\exp(N(\mu, \sigma))$ with $\mu = 2$ and $\sigma = 1$

0.832858	2.573425	3.724999	9.227466	14.99063
0.903766	2.602635	4.258860	10.80821	15.05903
1.821690	2.659332	6.221531	10.85469	18.03431
2.463967	3.689074	8.331888	14.64650	24.97989

To explore lognormality, normal *Q–Q* charts of the log-transformed data were plotted. In a normal *Q–Q* chart, observed values of a single numeric variable are plotted against the expected values if the log-transformed sample were from a normal distribution. If the log-transformed sample is from a normal distribution, points will cluster around a straight line (see Fig. 4).

Both data sets are lognormal distributed according to the Kolmogorov Smirnov statistic for lognormality. The *Q–Q* plots indicate that the data are lognormal distributed around the mean, but tend to deviate at the tails, especially the upper tail for the Cu data set (Fig. 4a). These two possible outliers may influence the entire parametric fit, i.e. the mean and standard deviation will be overestimated. As a consequence, it can be expected that the 5th-variability percentile will be underestimated.

4. Results and discussion

All the previously discussed methods were assessed for their performance in calculating the 5th-variability percentile and its uncertainty/confidence estimates.

An illustration of the results for LAS is shown in Figs. 5 and 6. One can conclude that there is a distinct difference in shape between the parametric and non-parametric distributions estimated from the data. The parametric methods tend to produce smoother and smaller uncertainty or confidence bands compared to the non-parametric methods. Non-parametric techniques are more efficiently fit to data. This clearly demonstrates that the choice of distribution is an important, general problem.

In order to select a lower bound, it is necessary to specify both the desired percentiles of variability, and uncertainty. For example, one point estimate would be the 5th-percentile of uncertainty for the 5th-percentile of variability.

Initially, the hypothetical lognormal data set was studied. Fig. 7 gives the 90% uncertainty intervals of the median 5th-variability percentile obtained with all methods tested on the 20 data points of the hypothetical lognormal distribution. Let us compare the

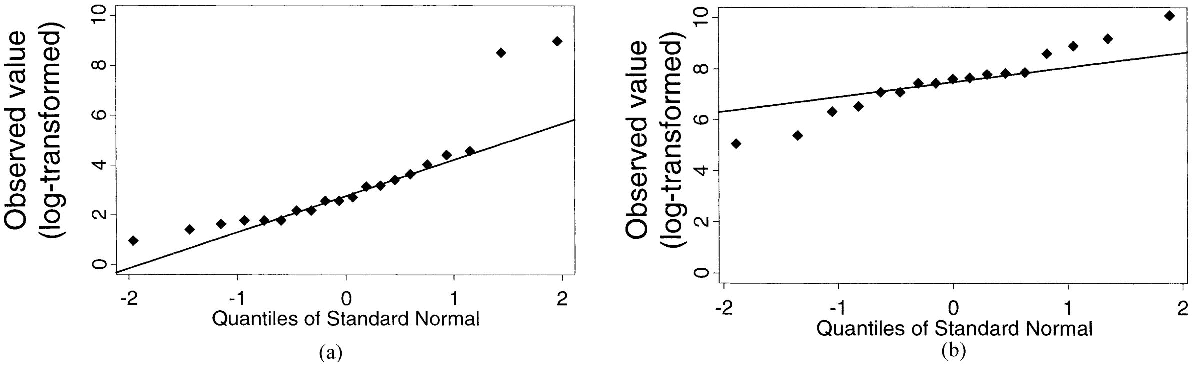


Fig. 4. Normal *Q–Q* plots for (a) log Cu and (b) log LAS.

parametric methods. The maximum likelihood method and the Bayesian approach lead to the same results, as Aldenberg and Jaworska [7] already concluded. The parametric bootstrap results are similar to the MLE and Bayesian analysis results, although with wider confidence bands. Because MLE and Bayesian analysis are easier to use and not so computationally intensive, they should be preferred over the parametric bootstrap.

The true 5th-variability percentile lies within the 90% uncertainty interval of all methods. This is to be expected, as data set 1 is lognormally distributed. The methods are good, given the correct assumption of the distribution.

The lower 90% uncertainty limit and the median 5th-variability percentile are (almost) equal for the re-sampling procedure (non-par bootstrap EDF in Fig. 7). In case of small data sets, the 5th-percentile

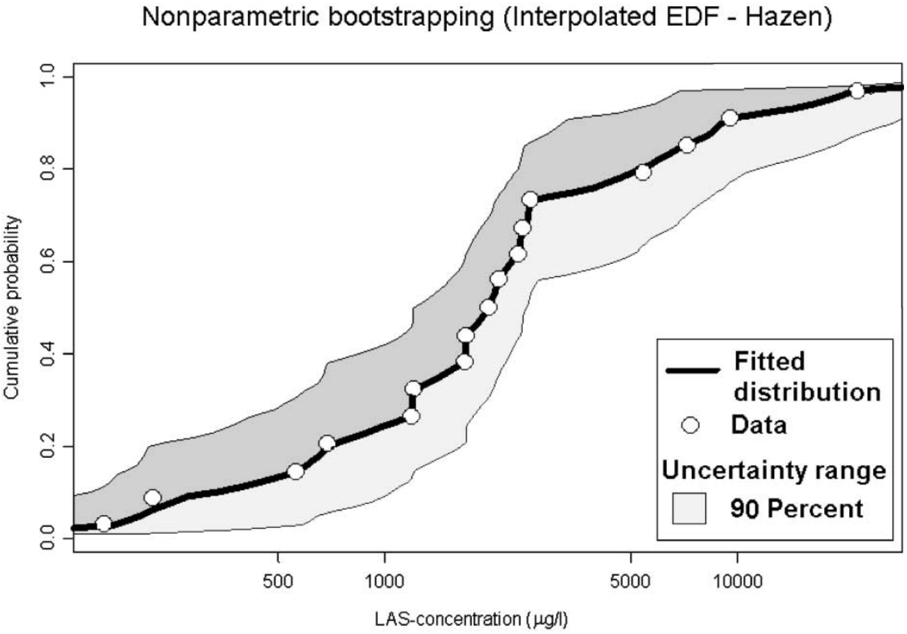


Fig. 5. Cumulative distribution function for the LAS data set based on non-parametric bootstrapping (interpolated EDF and Hazen plotting method).

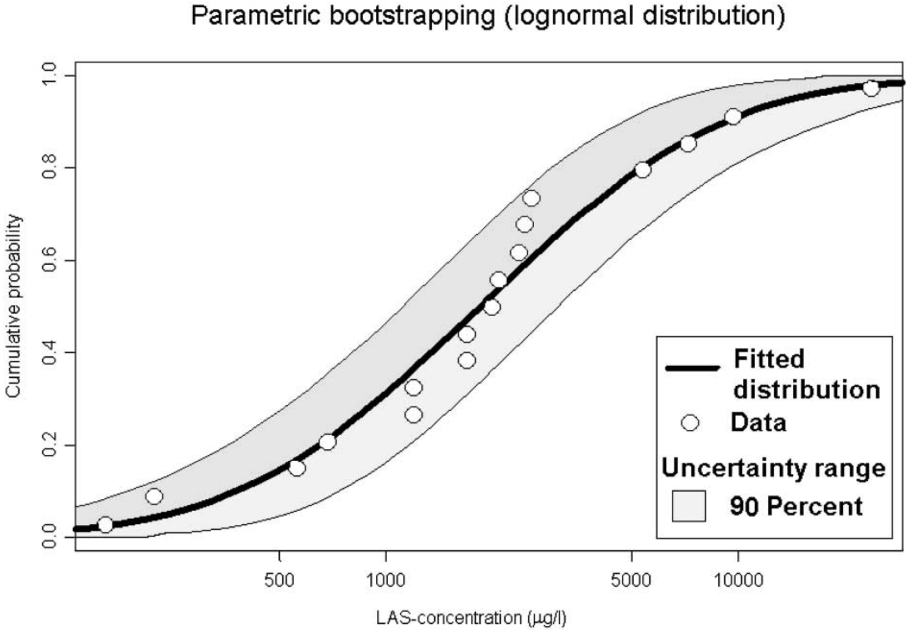


Fig. 6. Cumulative distribution function for the LAS data set based on parametric bootstrapping (lognormal distribution).

has to be estimated between zero and the first point. If the empirical distribution function (see Fig. 3) is used, the 5th-percentile is the first point itself. In other words, the uncertainty interval is bounded by the first (and smallest) data point (namely 0.832858). As a result, the first data point is selected many times. When linear interpolation between zero and the first point is used (as in the interpolated empirical distribution function, see Fig. 3), the lower 90% uncertainty limit is not bounded by the first point and as a result

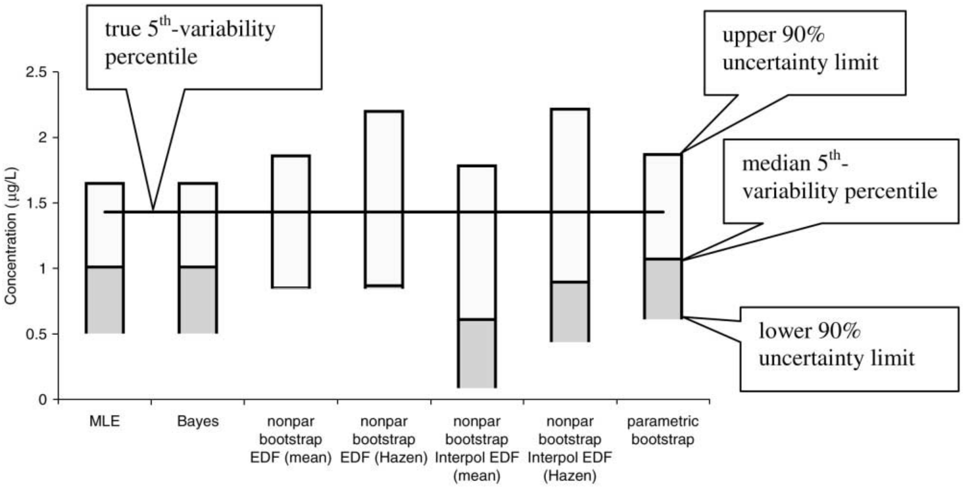


Fig. 7. Uncertainty (90%) or confidence intervals of the 5th-variability percentile following various methods for 20 data points of the hypothetical lognormal data set (thick line: true 5th-variability percentile).

Table 2
Coverage, or percentage (%) of the samples that the actual 5th-percentile value is included in the 90% confidence interval, calculated for different methods and distributions

Method	Distribution	Coverage (%)
Maximum likelihood method	Lognormal	90.6
Bayesian statistics	Lognormal	90.6
Bootstraps		
Parametric	Lognormal	88.9
Non-parametric (resampling)	EDF	58.6[a]
	EDF	63.6[b]
Non-parametric	Interpolated EDF	93.7[a]
	Interpolated EDF	94.5[b]

[a] Values obtained from mean plotting.

[b] Values obtained from Hazen plotting.

linear interpolation accounts for the possibility that the lower 90% uncertainty limit can be smaller than the first data point.

The mean and Hazen plotting system (in case of non-parametric bootstrapping with interpolated EDF) show significant differences. A factor of 4 was observed between the minimum and the maximum of the estimated lower 90% uncertainty limit.

Note that Fig. 7 is only one possible realisation of confidence intervals. Therefore, as a validation exercise, 20 new data points from the same hypothetical lognormal distribution were considered. The uncertainty interval of the median 5th-variability percentile was again estimated. The coverage of the true 5th-variability percentile over the uncertainty interval was checked for every method. If this process is repeated 1000 times, the uncertainty interval should cover the true 5th-variability percentile 900 times, i.e. the method with coverage closest to 90% should be considered as the most suitable method. The results are displayed in Table 2.

Differences between methods are mostly determined by the choice of the probability model. All parametric methods assuming lognormal distribution give similar results. The results show that all parametric methods also give the best results. This is to be expected as the hypothetical data set is lognormally distributed. The non-parametric resampling procedure clearly underestimates the uncertainty. This method is often used in literature. On the other hand, the non-parametric bootstrapping, based on an interpolated EDF, overestimates the uncertainty interval, which is normal as non-parametric techniques tend to have larger uncertainty estimates.

Finally, two true toxicity data sets (LAS and Cu) were studied (see Figs. 7 and 8). For LAS, a possible

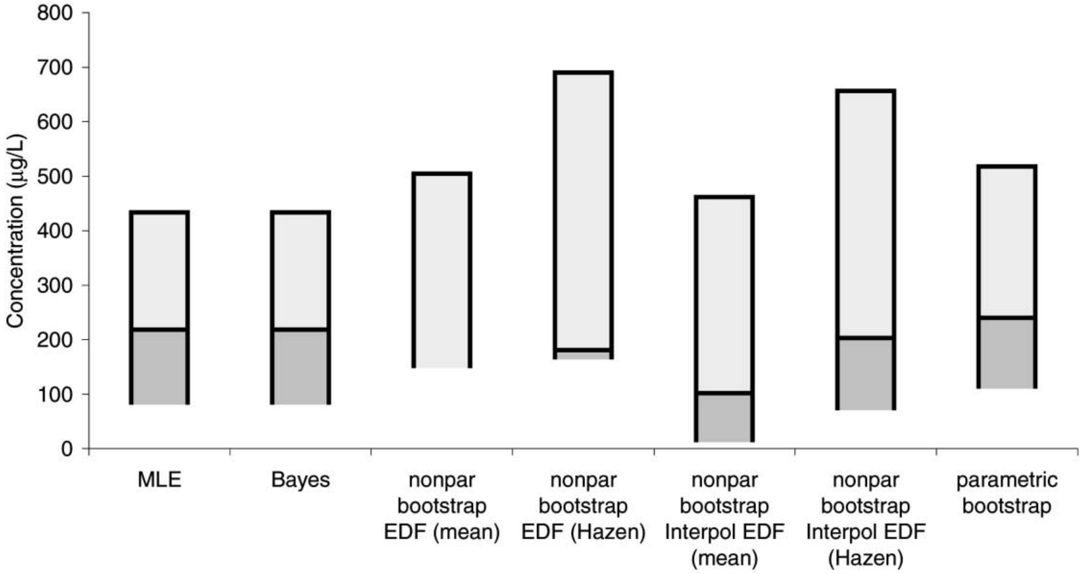


Fig. 8. Uncertainty (90%) or confidence intervals of the 5th-variability percentile following various methods for 17 data points of the LAS data set (concentration in μg/l).

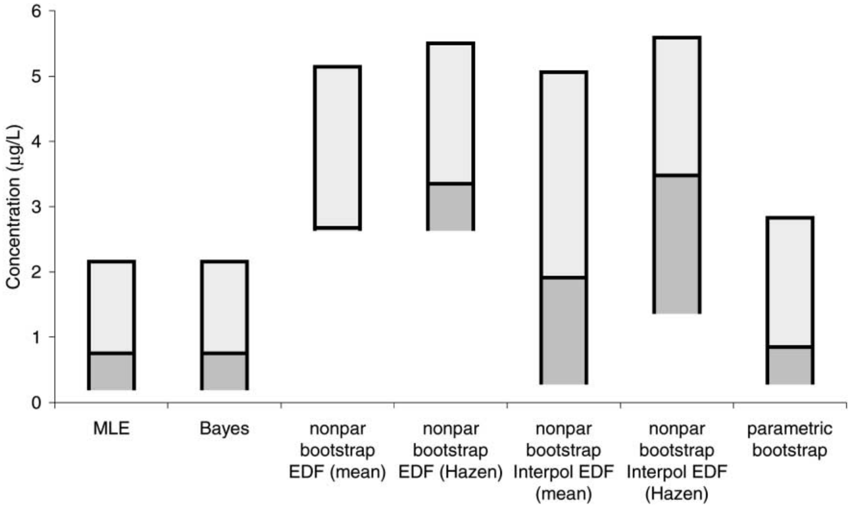


Fig. 9. Uncertainty (90%) or confidence intervals of the 5th-variabiltiy percentile following various methods for 20 data points of the Cu data set (concentration in μg/l).

median 5th-variability percentile could be identified (median of all methods is around 200 mg/l) that could be situated within the 90% uncertainty bands of all methods. Fig. 8 shows large similarities to Fig. 7. However, for Cu, no possible median 5th-variability percentile could be identified that would lie within the 90% uncertainty bands of all methods, since these do not overlap. For LAS, a factor of 2.4 and for Cu, a factor of almost 5 was found between the minimum and the maximum of the estimated median 5th-variability percentile of the various methods.

From Fig. 9, it can be seen that the results are very sensitive to the choice of the assumed distribution (parametric as in MLE, Bayesian approach and parametric bootstrapping) or not (as in non-parametric bootstrapping). Furthermore, as already outlined in Section 3, the influence of potential outliers may not be underestimated. A detailed outlier study should be performed.

5. Conclusions

Possible statistical framework for characterising uncertainty and variability in quality standard setting were examined.

The considered methods display varying robustness and accuracy in determining lower confidence limits of the 5th-variability percentile. The considered methods display varying degree of robustness when sample size decreases. The most suitable methods to estimate lower end percentiles such as 5th-percentile were found to be the maximum likelihood estimation method, Bayesian analysis and non-parametric bootstrapping (using interpolated EDF and the Hazen plotting system).

At this stage, there is no direct reason to prefer parametric or non-parametric methods. However, the results are very sensitive to the choice of the method (a factor of 5 difference was observed when results from different methods were compared). Differences between methods are for a large part determined by the choice of the probability model.

Some non-parametric methods should not be used for estimating low percentiles given a small sample size. All resampling techniques showed they were rather arbitrary and inaccurate because they are bounded by the smallest data point.

For estimating 5th-percentiles of small sample sizes, the Hazen plotting and the mean plotting system are used in literature but one should be aware that both systems give different results (a factor of 2 was observed here) at low sample sizes.

Further research on the influence of outliers may reveal more information.

Acknowledgements

This research has been funded by a scholarship from the Flemish Institute for the Improvement of Scientific–Technological Research in the Industry (IWT). The authors also like to thank Colin Janssen (Ghent University, Laboratory for Environmental Toxicology and Aquatic Ecology) and two anonymous reviewers for their useful suggestions.

References

[1] A.C. Cullen, H.C. Frey, Probabilistic techniques in exposure assessment. A handbook for dealing with variability and uncertainty in models and inputs, 1999.

[2] A.C. Davison, D.V. Hinkley, Bootstrap Methods and their Application, Cambridge University Press, Cambridge, 1997.

[3] H.C. Frey, D.S. Rhodes, Hum. Ecol. Risk Assessment 4 (1998) 423–468.

[4] D.E. Burnmaster, A.M. Wilson, Risk Assessment for Chemicals in the Environment, Wiley, New York, 1998.

[5] R.H. Jagoe, M.C. Newman, Ecotoxicology 6 (1997) 293–306.

[6] M.C. Newman, D.R. Ownby, L.C.A. Mézin, D.C. Powell, T.R.L. Christensen, S.B. Lerberg, B.-A. Anderson, Environ. Toxicol. Chem. 19 (2000) 508–515.

[7] T. Aldenberg, J.S. Jaworska, Ecotoxicol. Environ. Safety 46 (2000) 1–18.

[8] T. Aldenberg, W. Slob, Ecotoxicol. Environ. Safety 25 (1993) 48–63.

[9] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.

[10] G.E.P. Box, G.C. Tiao, Bayesian Inference in Statistical Analysis, Addison-Wesley, New York, 1973.

[11] D.J. Versteeg, S.E. Belanger, G.J. Carr, Environ. Toxicol. Chem. 18 (1999) 1329–1346.