Chemosphere 138 (2015) 599–608

Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

Estimating the mean and standard deviation of environmental data with below detection limit observations: Considering highly skewed data and model misspecification

Niloofar Shoari *, Jean-Sébastien Dubé, Shoja'eddin Chenouri

Department of Construction Engineering, École de Technologie Supérieure, Montreal, Canada
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada

HIGHLIGHTS

- Conflicting opinions exist concerning the suitability of estimation methods to handle left-censored environmental data.
- The performance of parametric estimators is greatly affected by data skewness.
- The robustness of parametric estimators to model misspecification is evaluated.
- MLE based on lognormality performs poorly for highly skewed data and under model misspecification.
- MLE based on gamma is more robust to variations in data skewness and to model misspecification.

ARTICLE INFO

Article history:
Received 18 February 2015
Received in revised form 2 July 2015
Accepted 7 July 2015
Available online 25 July 2015

Keywords:
Left-censored
Maximum likelihood
Kaplan–Meier
Regression on order statistics

ABSTRACT

In environmental studies, concentration measurements frequently fall below detection limits of measuring instruments, resulting in left-censored data. Some studies employ parametric methods such as the maximum likelihood estimator (MLE), robust regression on order statistic (rROS), and gamma regression on order statistic (GROS), while others suggest a non-parametric approach, the Kaplan–Meier method (KM). Using examples of real data from a soil characterization study in Montreal, we highlight the need for additional investigations that aim at unifying the existing literature. A number of studies have examined this issue; however, those considering data skewness and model misspecification are rare. These aspects are investigated in this paper through simulations. Among other findings, results show that for low skewed data, the performance of different statistical methods is comparable, regardless of the censoring percentage and sample size. For highly skewed data, the performance of the MLE method under lognormal and Weibull distributions is questionable; particularly, when the sample size is small or censoring percentage is high. In such conditions, MLE under gamma distribution, rROS, GROS, and KM are less sensitive to skewness. Related to model misspecification, MLE based on lognormal and Weibull distributions provides poor estimates when the true distribution of data is misspecified. However, the methods of rROS, GROS, and MLE under gamma distribution are generally robust to model misspecifications regardless of skewness, sample size, and censoring percentage. Since the characteristics of environmental data (e.g., type of distribution and skewness) are unknown a priori, we suggest using MLE based on gamma distribution, rROS and GROS.

1. Introduction

It is often necessary to estimate statistical parameters of contaminant concentration distributions. For example, in contaminated site characterization, this helps us to determine the average level of contamination of a remediation unit or to make statistical inferences to differentiate contaminated soil layers. Complications occur when the contaminant concentrations can not be quantified because the precision of the laboratory instrument is not sufficient to distinguish the presence of the contaminant from the background noise. As a result, a qualitative information is obtained since all we know is that the concentration lies between zero and the detection limit (DL) of measuring instruments (El-Shaarawi and Piegorsch, 2012; Ofungwu, 2014). A measurement that is less than the DL is called a left-censored data point.

* Corresponding author.
E-mail address: niloofar.shoari.1@ens.etsmtl.ca (N. Shoari).

http://dx.doi.org/10.1016/j.chemosphere.2015.07.009
0045-6535/

Furthermore, the concentration data might contain multiple DLs due to the use of different measuring instruments, analytical methods, or combining data sets with different DLs (Jin et al., 2011 and He, 2013).

In survival analysis, there are several statistical methods to accommodate right-censored data that can be adapted to address the problem of left-censoring in environmental studies. The most common methods to handle left-censored data include (i) the Maximum Likelihood estimator (MLE), (ii) methods based on Regression on Order Statistics (ROS), and (iii) Kaplan–Meier (KM) procedure. The MLE and ROS-based methods are parametric approaches that assume a predetermined distribution for the data whereas, the KM method is a non-parametric approach and does not require any distributional assumption. The two common versions of ROS are the robust ROS (rROS) and gamma ROS (GROS) methods that rely on lognormal and gamma assumptions, respectively.

Although several studies try to offer guidelines about how to deal with left-censored data through Monte-Carlo simulations (Singh et al., 2006; Helsel, 2010; Helsel, 2012), there has been no general agreement on an appropriate strategy. Literature review reveals that, in addition to sample size (Gardner, 2012) and percentage of censoring (Kroll and Stedinger, 1996; Huynh et al., 2014), skewness of the underlying distribution influences the performance of the methods (EPA, 2006). To our knowledge, only a few studies consider skewness when assessing the performance of the statistical methods in estimating the distributional parameters. For example, EPA (2006) guidelines state that conclusions derived for low skewed distributions cannot be generalized to moderately and highly skewed ones. We believe that the reason for which the conclusions of previous studies are not in general agreement is the fact that the impact of skewness was overlooked. In fact, the comparative simulations that were based on low to moderately skewed distributions or the simulations in which the results were averaged over a wide range of distributions generally argue in favor of the MLE method under lognormal assumption (Shumway et al., 2002; European Food Safety Authority, 2010; Hewett and Ganser, 2007; Lynn, 2001; Jain et al., 2008). On the other hand, studies that include more skewed distributions report poor performance of MLE under lognormal assumption (Gilliom and Helsel, 1986 and Helsel and Cohn, 1988).

In addition to the issue of skewness mentioned earlier, there is an issue regarding the performance of the parametric methods in the case of misspecified distributions. The common practice in environmental literature is to assume that data are lognormally distributed and to use the MLE and rROS methods based on this assumption (El-Shaarawi, 1989; Huybrechts et al., 2002; Baccarelli et al., 2005; Caudill et al., 2007; Leith et al., 2010). It is crucial to know how these methods behave if the underlying parametric model is misspecified. This occurs because

(a) There is no evidence that all environmental data are actually lognormal.
(b) There is not any straightforward extension of goodness-of-fit tests to establish the true underlying distribution of a given environmental data set due to the presence of left-censored observations.

Unfortunately, comprehensive studies that examine the robustness of the parametric estimators in the case of model misspecification are rather rare. Although the MLE method under lognormal assumption has been widely studied (for example, Gilliom and Helsel, 1986; She, 1997; Shumway et al., 2002; Hewett and Ganser, 2007, among others), only a few environmental studies have attempted to investigate the performance of MLE under Weibull and gamma assumptions (Schmoyeri et al., 1996 and European Food Safety Authority, 2010).

This paper aims at unifying the existing literature on environmental data analysis in the presence of left-censored data by addressing the above mentioned issues. To infer conclusions applicable to more realistic scenarios, we investigate the robustness of the methods under study to variations in data skewness and departures from a distributional assumption. This is key in the analysis of concentration data as neither the underlying distribution nor the skewness is exactly known a priori. We employ an extensive simulation exercise to evaluate the performance of the MLE, rROS, GROS, and KM methods in estimating distributional parameters in simulation scenarios based on different levels of skewness and data generating distributions. The particular objective of this work is to address the issue of the robustness of the parametric methods (i.e., MLE, rROS and GROS). This is achieved by:

(a) Investigating the robustness of MLE and rROS based on lognormal assumption when the data are generated from Weibull, gamma, and some mixture distributions.
(b) Investigating the robustness of MLE under Weibull assumption when the data are generated from lognormal, gamma, and some mixture distributions.
(c) Investigating the robustness of MLE and GROS based on gamma assumption when the data are generated from lognormal, Weibull, and some mixture distributions.

Careful collection and chemical analysis of environmental samples leads to obtaining concentration data sets that are representative of the actual contamination level of the sampling location. However, extracting correct information contained in the data and estimating the contamination level at the scale of a remediation unit or the site is possible using adequate statistical methods. Decisions made upon appropriate statistical methods protect human health and environment, optimize the allocation of financial resources and save time and effort. The conclusions of this study are applicable to any process that include contaminant quantification such as environmental monitoring and risk assessment.

2. Estimation techniques

In this section, we briefly describe the most common statistical methods for analyzing left-censored data. These are maximum likelihood estimation, methods based on regression on order statistics, and Kaplan–Meier methods.

Maximum Likelihood estimation (MLE) utilizes a likelihood function to estimate the distributional parameters. The likelihood function describes the likelihood of observed data, given any member of an assumed parametric family of distributions. In this method, the distributional parameter θ (e.g., the mean and standard deviation) is estimated by maximizing the likelihood function with respect to these parameters. Let $y_1, y_2, \ldots, y_n$ be some observations (i.e., contaminant concentrations) and let $\mathbf{DL} = (DL_1, \ldots, DL_n)$ denote the vector of censoring points (detection limits). The observed concentration data consist of pairs (x_i, δ_i) where $x_i = \max\{y_i, DL_i\}$ and $\delta_i = I(y_i \geqslant DL_i)$, meaning that $\delta_i = 1$ if $y_i \geqslant DL_i$ (in that case $x_i = y_i$) and $\delta_i = 0$ if $y_i < DL_i$ (in that case $x_i = DL_i$) for any $i = 1, \ldots, n$. For a random sample of size n, the likelihood contribution from the i^{th} observation is expressed as the probability density function $f(x_i;\theta)$, if the observation is not censored, and as the cumulative density function $F(x_i;\theta)$ if it is left-censored. For a full sample of n observations, the likelihood function is given by