Tales from the Tail: Robust Estimation of Moments of Environmental Data with One-Sided Detection Limits

Ping-Hung Hsieh

College of Business
Oregon State University
Corvallis, OR 97331-2603

Ping-Hung Hsieh is Associate Professor of Quantitative Methods, College of Business, Oregon State University, Corvallis, OR 97331-2603. (Phone) 541-737-6060 (Fax) 541-737-4890 (Email) Ping-Hung.Hsieh@bus.oregonstate.edu.

Tales from the Tail: Robust Estimation of Moments of Environmental Data with One-Sided Detection Limits

Abstract

Estimating the means and standard deviations of environmental data remains a great challenge because a substantial percentage of observations lies below or above detection limits. The inadequacy of several common, ad hoc estimation procedures is clear; this study instead proposes a robust moment estimation procedure for environmental data with a one-sided detection limit. The procedure assumes that the tails of the underlying distribution of the (transformed) data are symmetric, and censoring only occurs on one side. Through an application of the Rényi representation theorem, it is possible to use observations from the other side to learn the shape of the distribution below the detection limit, without specifying any particular parametric model, and consequently, derive the moment estimates of the distribution. A simulation provides a comparison of estimation performance between the proposed procedure and several existing estimators, and several real-life samples offer a good illustration.

Keywords: Extreme Value Theory, Generalized Pareto Distribution, Peaks over Threshold, Rényi Representation Theorem, Tail Index.

Tales from the Tail: Robust Estimation of Moments of Environmental Data with One-Sided Detection Limits

1 Introduction

Environmental data typically are censored because the measuring devices or procedures used to collect such data cannot reliably quantify trace levels of concentration below a certain quantitation limit (d). When concentrations exist lower than d, they are reported as "less-than" values. However, with this portion of sample censored, even the simple task of estimating the mean concentration presents a great challenge, and the use of standard statistical techniques, developed with a complete sample in mind, generally is not applicable. To complicate matters even further, several studies suggest that the distribution of environmental data frequently exhibits multi-modal and long-tailed traits, even after log transformations. Thus, the development of a moment estimation procedure that deals with singly censored environmental data constitutes a research topic of significant interest in both statistics and environmetrics.

Techniques for estimating the mean and variance of environmental data with censoring can be loosely classified into three categories: substitution, distributional, or some combination of the two (cf. Helsel 1990, 2005a, b, c). A substitution technique aims to fill in the censored observations with certain numerical values to create a fabricated sample, to which standard complete sample statistical methods can be applied. For convenience, a fraction of d typically is used as the fill-in value, where 0, $d/2$, or d are three commonly used numbers. Although simple, this method has little theoretical basis and it performs poorly in comparison with other more complex methods (Gilliom and Helsel 1986; Gleit 1985; Helsel 2006; Helsel and Cohn 1988; Helsel and Gilliom 1986).

The distributional technique assumes that the underlying distribution that generates the data is known, a strong but convenient assumption. Commonly assumed distributions for environmental data include normal, lognormal, and delta distributions described by Aitchison (1955). Moulton and Halsey (1995) establish a mixture model that consists of a censored lognormal distribution and a point distribution located below the detection limit for a sample of antibody concentration values. The parameters of these distributions are typically estimated by a maximum likelihood (ML) estimation approach. With a normality assumption, Cohen (1959, 1961) has developed ML estimators adjusted for censored data to estimate the parameters, and Persson and Rootzén (1977) propose a restricted version to offer computationally simpler estimators. Saw (1961), Schmee, Gladstein and Nelson (1985), Schneider (1986), and Haas and Scheff (1990) suggest further extensions and refinements, and Shumway, Azari, and Johnson (1989) and Shumway, Azari, and Kayhanian (2002) propose exact maximum likelihood estimators (MLE) in conjunction with a Box-Cox transformation that leads to the best approximate normal likelihood. Cohn (2005) also offers an adjusted MLE that fits Type I censored data.

The third technique combines the two approaches and plays an important role in environmental literature. In particular, regression on order statistics (ROS), as proposed by Shumway, Azari, and Kayhanian (2002), assumes that observations x_i, for $i = 1, 2, \dots, n$, or the log-transformed observations, are independently and normally distributed. By regressing the uncensored observations on $\Phi^{-1}(\hat{p}_i)$ using weighted (Gupta 1952) or ordinary (Helsel and Gilliom 1986) least squares, where Φ^{-1} denotes the inverse of the standard normal distribution function and $\hat{p}_i$ is the observation rank, we can fill the censored observations with the predicted values from the regression model. A variation of this procedure, as summarized by Travis and Land (1990), considers the lognormal distribution rather than a normal distribution. The ROS performance, in both normal and lognormal cases also has been examined closely by Gilliom and Helsel (1986).

These distributional and ROS techniques make a rather strong assumption regarding the knowledge of the underlying distribution, which is rarely possible in empirical studies. Several simulation studies reveal that the resulting estimates are sensitive to a deviation from the assumed distribution, such as Gilliom and Helsel (1986) and Helsel and Cohn (1988). Shumway, Azari, and Johnson (1989) and Shumway, Azari, and Kayhanian (2002) also propose "normalizing" the data using the Box-Cox transformation to eliminate the distributional bias. Korn and Tyler (2001) suggest the use of a Student t distribution with between two to four degrees of freedom to accommodate outlying data, as is frequently observed in environmental samples even after a log transformation. Yet even their considerations still assume a specific global distribution.

The lack of theoretical justification for the fill-ins and the strong distributional assumptions in several existing techniques indicate the need to develop an estimation procedure that provides theoretically defensible fill-in values and is robust across a wide range of distributions commonly considered for modeling environmental data. For a left-censored sample, for example, the proposed estimation procedure uses the reliably measured observations above the censoring point d to learn about the behavior of the right tail of the underlying distribution. By assuming tail symmetry (as defined in the next section), this procedure then applies the properties of the right tail to the left tail, where the censored data are located, to calculate the expected value and the quantiles below d, which then serve as fill-ins. Other than the assumption of tail symmetry, no global distribution is assumed.

This article proceeds as follows: Section 2 lays out the theoretical foundation of the proposed method, followed by the derivation of the moment estimators in Section 3. A simulation study reveals the strengths and weaknesses of the estimators in Section 4, followed by an application to several real-life examples in Section 5. Section 6 summarizes and concludes our study and offers possible extensions to this work. Certainly, the problem of moment estimation from data with detection limits confronts researchers in many other fields, including industrial experimenters (Liu and Sun 2000) and exposure assessments (Finkelstein 2008; Flynn 2010); the proposed estimators can be applied in those circumstances as well. Other issues related to detection limits, such as estimating the probability of detection, have been discussed by Lambert, Peterson and Terpenning (1991) and are not the focus of this paper. Baccarelli, *et al.* (2005) also provide an overview of estimation methods with a focus on substitution procedures.

2 Theory

2.1 Assumptions

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution F and $X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}$ be the corresponding order statistics. Given a lower quantitation limit d, we assume that the values of the smallest r order statistics can not be reliably quantified, i.e., $X_{(n-r+1)} \leq d \leq X_{(n-r)}$. We define the lower tail of the distribution to be the set $\{x : x \leq d\}$ and the upper tail as $\{x : x \geq D\}$, where D will be defined next.

The class $\mathcal{F}$ of distributions under consideration consists of the distributions with symmetric lower and upper tail behaviors. Without loss of generality, consider a class of distribution F defined on a real line with symmetric tail behavior at the origin. A distribution

F belongs to $\mathcal{F}$ if $F(x) = 1 - F(-x)$ and the first derivative of $F(x)$ is monotone for $x \leq d < 0$ and $x \geq D = -d$. Symmetric distributions, such as normal, Student t, and Cauchy distributions, belong to this class $\mathcal{F}$. Certain mixtures of normal distributions, such as $0.5\Phi(1,1)+0.5\Phi(-1,1)$ and $0.8\Phi(0,1)+0.2\Phi(0,5)$ (as discussed in Haas and Scheff 1990), are also members of $\mathcal{F}$, where $\Phi(\mu,\sigma)$ is a normal distribution with mean μ and standard deviation σ. The shape of the distribution between d and D is not assumed.

The core idea of the proposed estimation procedure thus can be defined. It proposes to use the observations above D to reveal the shape of the upper tail, and then apply a tail symmetry assumption to obtain the shape of the lower tail, where censoring occurs. With knowledge of the distributional form below d, it becomes possible to calculate estimates of several theoretical quantities, such as the expected value below d and quantiles, and use them as fill-ins for censored observations. Finally, standard complete sample statistical methods, such as the sample mean and standard deviation formulas, are applied to the fabricated sample to obtain the moment estimates. Although the derivation with the proposed method assumes the data are left censored, the proposed method can be easily modified for right-censored samples as well.

2.2 Learning the Tail Behavior

The development of the proposed method relies on the following theorem:

Theorem (Rényi 1953) *Let $X_1, \ldots, X_n$ be a sample from a continuous, strictly increasing distribution F, and let $X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}$ be the order statistics. Let e_i be independent exponentially distributed random variables with expectation* 1.0. *Then*

$$X_{(i)} = F^{-1}\left(\exp\left[-\left(\frac{e_1}{n} + \frac{e_2}{n-1} + \cdots + \frac{e_i}{n-i+1}\right)\right]\right), \tag{1}$$

for $i = 1, 2, \ldots, n$.

Suppose that $X_{(r)} \geq D$ and $1 - F(x) = \bar{G}(x;\theta) = 1 - G(x;\theta)$ for $x \geq D$, where $G(\cdot;\theta)$ is a specified function with parameter θ. Then from Equation (1), it follows that for $X_{(r)} \geq D$,

$$e_i = (n - i + 1)\left[\log \bar{G}(X_{(i-1)};\theta) - \log \bar{G}(X_{(i)};\theta)\right], \tag{2}$$

for $i = 1, 2, \ldots, r$. By definition, $G(X_{(0)};\theta) = 1$. If the observed values of the order statistics above D are $X_{(i)} = x_{(i)}$, for $i = 1, 2, \ldots, r$, then the conditional likelihood function for θ is

$$L(\theta) \propto |J| \times \exp\left[n \log \bar{G}(x_{(r)};\theta) + \sum_{i=1}^{r} \log \frac{\bar{G}(x_{(i)};\theta)}{\bar{G}(x_{(r)};\theta)}\right], \tag{3}$$

where the Jacobian J is proportional to $\Pi_{i=1}^{r}(d \log \bar{G}(x_{(i)};\theta))/dx_{(i)}$. Let $g(\cdot;\theta)$ be the first derivative of $\bar{G}(\cdot;\theta)$, in which case the log-likelihood function $l(\theta)$ can be simplified as

$$l(\theta) \propto (n - r) \log \bar{G}(x_{(r)};\theta) + \sum_{i=1}^{r} \log g(x_{(i)};\theta), \tag{4}$$

and the MLE of the parameter θ can be obtained by maximizing the log-likelihood function in Equation (4).

2.3 Alternative Functional Form $G(\cdot;\theta)$

Without specifying a global distribution, the task at hand is to find an appropriate function $G(\cdot : \theta)$ that approximates the tail area of the distributions commonly used for environmental data. Three alternative functions of $G(\cdot;\theta)$ are considered: (1) power function with $1 - G(x;\alpha, C) = Cx^{-\alpha}$, (2) exponential function with $1 - G(x;\alpha, C) = C\exp(-\alpha x)$, and (3) Weibull-type function with $1 - G(x;\alpha, C) = \exp(-Cx^{\alpha})$, where $\alpha > 0$ and $C > 0$.

These functional forms encompass a large class of tail behavior. The power function describes a distributional tail that decays algebraically, as in the case of the long-tailed Student t distribution that is often considered an alternative to the normal distribution when modeling environmental data. The exponential function includes a class of distribution whose tails are longer than those of a normal distribution but decay exponentially. This class includes distributions such as the double exponential distribution. Finally, the Weibull-type function captures the tail behavior that decays faster than the exponential function and includes the flexible Weibull distribution.

3 The Proposed Moment Estimators

Because the global distribution is not specified, the proposed strategy is to fill in the censored data with certain numerical values and then apply standard sample moment formulas on the fabricated sample. Two approaches are used to determine the fill-in values.

Fill in with mean excess value. Define the mean excess function as $e(D) = E\left[X - D | X > D\right]$. The idea is to fill in all values below d with the estimated $d - e(D)$, where $e(D)$ is estimated by replacing the parameters with the corresponding MLE. For power and exponential functions, $e(D)$ are $D/(\alpha - 1)$ and $1/\alpha$, respectively. The advantage of using the power and exponential functions is that the MLE of the parameters have simple closed-form solutions. For the power function, the respective MLE of α and C are (Hill 1975)

$$\hat{\alpha} = \frac{r}{\sum_{i=1}^{r}(\log x_{(i)} - \log x_{(r)})} \quad \text{and} \quad \hat{C} = \frac{r}{n} x_{(r)}^{\hat{\alpha}}, \tag{5}$$

and for the exponential function, the respective MLE of α and C are

$$\hat{\alpha} = \frac{r}{\sum_{i=1}^{r}(x_{(i)} - x_{(r)})} \quad \text{and} \quad \hat{C} = \frac{r}{n} \exp(\hat{\alpha} x_{(r)}). \tag{6}$$

For the Weibull-type function,

$$e(D) = C^{-1/\alpha}\Gamma(1 + 1/\alpha)\left[1 - F_G(D^{\alpha})\right]\exp(CD^{\alpha}), \tag{7}$$

where $\Gamma(\cdot)$ is the gamma function, and F_G is a gamma distribution with the scale parameter C and the shape parameter $1/\alpha$. An optimization algorithm is needed to obtain the MLE of α and C, whereas a numerical algorithm is required to calculate $\Gamma(\cdot)$ and integrate F_G in this case. To facilitate the following discussion, the estimation procedures that use the mean excess fill-in values derived from the power, exponential, and Weibull-type functions take the designations **PwME**, **EwME**, and **WwME**, respectively.

Fill in with sample quantiles. Similar to the **ROS** approach, it is possible to fill in the censored data with the sample quantiles. Consider the largest order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(r)}$ located above the cutoff D. Define $H(x) = [1 - F(x)][1 - F(D)]^{-1}$, and $H(X_{(i)}) = \hat{p}_i$, where $i = 1, 2, \ldots, r$. Let $H^{-1}(\cdot)$ be the inverse function of $H(\cdot)$. Thus, the r smallest order statistics below the quantitation limit d, $X_{(n)} \leq X_{(n-1)} \leq \ldots \leq X_{(n-r+1)} \leq d$, can be replaced with $\hat{x}_{(n-i+1)} = d - H^{-1}(\hat{p}_i) + D$ to create a fabricated sample, where the observation rank $\hat{p}_i = (i - 0.5)/r$ for $i = 1, 2, \ldots, r$, see, for example, Hazen (1914) and Hyndman and Fan (1996). Finally, the fill-in values $\hat{x}_{(n-i+1)}$ for the censored $X_{(n-i+1)}$ are, for $i = 1, 2, \ldots, r$,

$$\hat{x}_{(n-i+1)} = d - D\hat{p}_i^{-1/\alpha} + D \quad \text{for the power function,} \tag{8}$$

$$\hat{x}_{(n-i+1)} = d - \left[D - \frac{1}{\alpha}\log\hat{p}_i\right] + D \quad \text{for the exponential function, and} \tag{9}$$

$$\hat{x}_{(n-i+1)} = d - \left[D^\alpha - \frac{1}{C}\log\hat{p}_i\right]^{1/\alpha} + D \quad \text{for the Weibull-type function.} \tag{10}$$

Cohen and Ryan (1989) replace the censored data with random deviates generated from an assumed distribution. The resulting moment estimates depend on the generated random deviates and can not provide unique estimates. The estimates also rely on a strong distributional assumption that goes against the core idea of the current article, so their approach is not considered further. For similar reasons, multiple imputation methods (e.g., Rubin 1987, Schafer 1997) that replace each censored value with a set of plausible values are also excluded. For the following discussion, the terms **PwSQ**, **EwSQ**, and **WwSQ** represent the estimation procedures that use the sample quantiles fill-in values derived from the power, exponential, and Weibull-type functions, respectively.

4 Performance Evaluation

The simulation study in this section evaluates the performance of the proposed estimation procedures in comparison with several extant approaches under various distributional assumptions and censoring percentages.

4.1 Estimation Procedures

In addition to the proposed **PwME**, **EwME**, **WwME**, **PwSQ**, **EwSQ**, **WwSQ**, the simulation considers the following estimation procedures: **FULL**, **DL/2**, **ROS**, **EM**, and **MEst**. The estimation procedure **FULL** uses the standard sample mean and standard deviation formulas on the entire simulated data without censoring. Using the sample moment formula offers a convenient approach in practice when all sample values are measured and available. The resulting estimates provide a basis for comparison. In contrast, the **DL/2** procedure replaces all censored values by half of the detection limit, then applies the usual sample moment formula to this fabricated sample to obtain the mean and standard deviation estimates. Although Helsel (2010) concludes against the use of **DL/2** and substitution methods in general, it is included as a performance benchmark for the other estimation procedures.

The regression on order statistics (**ROS**) approach (Newman, Dixon, Looney and Pinder, 1989) fits a regression model to the observed data, likely after a suitable transformation, and the hypothetical normal quantiles. Specifically, if the transformed observations $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n-1)} \leq y_{(n)}$ are assumed to be independently normally distributed with a common mean μ_y and variance σ_y^2, then each $y_{(i)}, i = 1, \ldots, n$, satisfies

$$y_{(i)} = \mu_y + \sigma_y \Phi^{-1}(P_i), \tag{11}$$

where $P_i = Prob\{Y_{(i)} \leq y_{(i)}\}$, and $\Phi^{-1}(\cdot)$ denotes the inverse of the cumulative normal distribution function. Suppose that the data are left censored and the smallest n_0 data values are below the detection limit d, i.e. $y_{(n_0)} < d \leq y_{(n_0+1)}$. In this case, **ROS** regresses the remaining largest $n - n_0$ uncensored data values on their adjusted ranks (Blom 1958) to obtain the mean and variance estimates. In other words, Equation 11 becomes

$$y_{(i)} = \mu_y + \sigma_y \Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right) + \epsilon_i, \tag{12}$$

where $i = n_0 + 1, \ldots, n$; the residual errors ϵ_i are i.i.d. normal with mean 0 and standard deviation σ_ϵ; and the probability P_i is replaced by the adjusted ranks as an accepted procedure in practice. For this simulation, the intercept and the slope parameters, μ_y and σ_y, respectively, are estimated by ordinary least squares. Extensions and variations of **ROS** appear in Gupta (1952) and Helsel and Gilliom (1986).

The expectation maximization (**EM**) algorithm developed by Dempster, Laird and Rubin (1977) attempts to deal with censored and missing data. Starting with some convenient estimates of μ_y and σ_y, **EM** defines an iterative sequence that involves re-estimating the mean and variance at each iteration, with all censored data replaced by the same conditional expected value, then re-evaluating the log-likelihood function with the new mean and variance estimates. The EM algorithm is guaranteed to increase the log-likelihood function and converge to a unique maximum, if it exists. Technical details are available in work by Dempster, Laird, and Rubin (1977), Gleit (1985), Shumway, Azari, and Johnson (1989), and Shumway, Azari, and Kayhanian (2002).

Finally, M−estimates (Huber 1981) offer a generalization of maximum likelihood estimates. Given a sample $y_i, i = 1, \ldots n$, and some function ρ, an M−estimate is defined as a solution $\hat{\theta}$ that minimizes

$$\sum_{i=1}^{n} \rho(y_i; \theta).$$

For the purposes of this study, the objective function is the negative log-likelihood function, defined as $\sum_{i=1}^{n} \rho(y_i; \delta_i, \theta)$, where $\theta = \{\mu_y, \sigma_y^2\}$, $\rho(y_i; \delta_i, \theta) = -\delta_i \log(f_\theta(y_i)) - (1 - \delta_i) \log(F_\theta(y_i))$; f and F are the density and distribution functions, respectively; and $\delta_i = 1$ if the ith observation y_i is uncensored and 0 if censored. In addition to the detection limit issue, Korn and Tyler (2001) note that the distribution of environmental data often exhibit longer tails than does a normal distribution, even after a log transformation. As a result, they focused on the use of the Student t density f and distribution F with v degrees of freedom, then applied an EM algorithm (Pettitt 1985) to obtain the solution $\hat{\theta}$. Their simulation showed that a choice of $v = 2$, 3, or 4 is a reasonable compromise between robustness and efficiency. Therefore, the present simulation uses $v = 3$, denoted **MEst**.

4.2 Simulation Design

Several prior studies have provided good starting points for designing the simulation (e.g., Haas and Scheff 1990; Helsel and Gilliom 1986; Hewett and Ganser 2007; Singh and Nocerino 2001). Distributions commonly assumed for (transformed) environmental data, such as normal, Student t, and mixtures of normals, appear in the simulation, and experiments with different parameter values indicated the properties of the proposed estimation procedures.

These distributions, along with their abbreviations in brackets, include the normal distribution $\Phi(10; 1)$ [**Norm**]; double exponential distribution with the mean 10 and standard deviation 1 [**DExp**]; Student t distribution with the mode at 10, and 2 and 5 degrees of freedom, denoted by [**Tdf**v], where $v = 2$, and 5; and the following mixtures of normal distributions: $0.5 \times \Phi(9; 1) + 0.5 \times \Phi(11; 1)$ [**M2Sa**], $0.5 \times \Phi(8; 1) + 0.5 \times \Phi(12; 1)$ [**M2Sb**], $0.3 \times \Phi(8; 1) + 0.4 \times \Phi(10; 1) + 0.3 \times \Phi(12; 1)$ [**M3Sa**], $0.45 \times \Phi(8; 0.25) + 0.10 \times \Phi(10; 0.25) + 0.45 \times \Phi(12; 0.25)$ [**M3Sb**], $0.8 \times \Phi(10; 1) + 0.2 \times \Phi(13; 1)$ [**MNSa**], and $0.1 \times \Phi(8; 1) + 0.9 \times \Phi(12; 1)$ [**MNSb**]. Because the standard deviation is location invariant, to avoid taking log-transformations on the negative values of already log-transformed data, the location of all distributions is shifted to 10. The density plots of the aforementioned distributions appear in Figure 1.

The **Norm** distribution serves to approximate the distribution of transformed environmental data, considering its well-studied properties and convenience. However, several empirical studies also suggest that the normal distribution may not be a good fit to the data, even after proper transformation. Therefore, the simulation considers the cases of **DExp**, whose tails decay exponentially, and **Tdf**v, which has longer tails than does a normal distribution, in particular, **Tdf2** is a heavy-tailed distribution with an infinite second moment. Following the simulation design in Haas and Scheff (1990), the present study uses the mixtures of normal distributions to explore different tail behavior and modality and includes the cases of the unimodal distributions **M2Sa** and **M3Sa**, a bimodal distribution **M2Sb**, and a tri-modal distribution **M3Sb**, all of which satisfy the tail symmetry assumption. To examine the impact of the tail symmetry assumption, the simulation also contains two distributions, **MNSa** and **MNSb**, that violate the assumption.

For each distribution with the true parameter $\theta = \{\mu, \sigma\}$, it is possible to generate $r = 1000$ samples of size $n = 100$ and 200, respectively, and take the smallest $p\%$ of observations as censored values, where p = 5, 10, 15, 20, 25, and 30 in the simulation. With the resulting estimate $\hat{\theta}$ for each of the estimation procedures under consideration, the quantification of its accuracy relies on calculating the root mean square error (RMSE), defined as $\sqrt{\sum_{i=1}^{r}(\hat{\theta}_i - \theta)^2/r}$.

4.3 Simulation Results

The RMSE values of the μ and σ estimates are summarized in Figures 2 and 3. Because the results from the two proposed fill-in procedures, namely, with mean excess value and with sample quantiles, are nearly identical, this section only details the results from **PwME**, **EwME**, and **WwME**. To enable the presentation of all estimation procedures in the same figures, the abbreviated designations in the figures use **p** for **PwME**, **e** for **EwME**, **w** for **WwME**, **F** for **FULL**, **D** for **DL/2**, **R** for **ROS**, **E** for **EM**, and **M** for **MEst**. A quick glance shows that, as expected, the RMSEs derived from a larger sample size $n = 200$ (denoted by $\times$ in the figures) are smaller than those from a smaller sample size $n = 100$ (denoted by $\circ$) for all combinations of censoring percentages and underlying distributions. The RMSEs of $\hat{\sigma}$ are generally higher than the corresponding RMSEs of $\hat{\mu}$, which implies that it is more difficult to get a good estimate of σ. Furthermore, when the true distribution deviates from normality with either longer tails and/or multiple modes, the RMSEs also increase. Finally, as the censoring percentage p increases, the performance of several estimation procedures, **DL/2** and **EM** in particular, deteriorates with an increasing RMSE, whereas the RMSEs of the proposed **PwME**, **EwME**, and **WwME** stay approximately the same.

In both figures, if **Norm** is the true underlying distribution, all estimation procedures perform equally well for low percentages of censoring. Except for **DL/2** and **EM**, the

RMSEs of the other estimation procedures remain similar to that of **FULL** as censoring intensity increases. The increase in RMSE for both **DL/2** and **EM** is in line with the findings in Singh and Nocerino (2001), who recommend against the use of **DL/2** and **EM** for samples with a size greater than 15 and/or larger censoring levels.

The poor performance of **DL/2** and **EM** as censoring intensity increases emerges again when the true underlying distribution is **DExp** or **Tdf5**. The **MEst** and **ROS** procedures are two front runners that outperform even **FULL** in the case of the μ estimation. However, their performances are less than ideal in the case of the σ estimation. In contrast, the performance of the proposed **PwME**, **EwME**, and **WwME** methods is consistent in both cases, with RMSEs close to the values for the front runners.

Both **M2Sa** and **M3Sa** are symmetric and unimodal, and the proposed estimation procedures continue to perform consistently. Even in the case of the μ estimation, **MEst** and **ROS** are no longer the best procedures, as in the **DExp** and **Tdf5** cases. This finding is most evident in the σ estimation case for **MEst**, which suggests that its performance is rather sensitive to departures from normality.

The proposed **PwME**, **EwME**, and **WwME** methods outperform the rest of the estimation procedures that use censored data when the true underlying distribution is bimodal (**M2Sb**) or tri-modal (**M3Sb**). The results highlight the strengths of the proposed estimation procedures, in that **EM**, **ROS**, and **MEst** were developed with normality in mind. Once the true distribution differs notably from **Norm**, their performances suffers, as clearly depicted in the results from the σ estimation in Figure 3. However, the proposed estimation procedures accommodate distributions such as **M2Sb** and **M3Sb**, and their RMSEs are almost as good as that of **FULL**, especially in the low censoring cases.

If the tail symmetry assumption is violated, the performance of an estimation procedure depends on the extent of the violation, as well as the censoring percentage. For the two special cases under consideration (**MNSa** and **MNSb**), the proposed estimation procedures compete equally well with the other procedures and even do better in general. It is premature to draw any conclusion about the performance of the proposed **PwME**, **EwME**, and **WwME** methods when the tail symmetry assumption is violated, but the results from both the **MNSa** and **MNSb** cases seem to suggest that they may not do worse than the other estimation procedures.

Suppose that two scientifically reputable analytical techniques are used to measure an environmental element. The more sensitive technique gives rise to a complete sample, whereas the other technique produces a sample with censored values. Antweiler and Taylor (2008) propose replacing the censored data with their uncensored counterparts to create a "complete" uncensored data set for the less sensitive technique. The mean and standard deviation of the fabricated data set are treated as the "true" mean and standard deviation and used as a comparison basis for various estimation procedures. Their "true" mean and standard deviation are essentially the same as the estimates derived from **FULL** in the simulation herein and should outperform all estimation procedures across all distributions and censoring percentages considered, as shown in the simulation. In other words, one would typically do better with a complete sample. However, both sample mean and standard deviation formulas are sensitive to extreme observations, and the performance of **FULL** is not without its own problems, as Table 1 shows. Recall that **Tdf2** is a Student t distribution with a true mean $\mu = 10$ and 2 degrees of freedom. It is a heavy-tailed distribution with an infinite variance. For 1000 generated samples of size 100 and 200, Table 1 summarizes the estimation results with various censoring percentages. On average, the proposed estimation procedures and **FULL** hit the target $\mu = 10$ across all censoring percentages. The difference is that the standard error of **FULL** is quite large in comparison. For example, for $n = 100$ and 5% censoring, the standard error of **FULL** is 0.366, whereas the standard error of the proposed procedures is approximately half that amount, ranging between 0.180 to 0.187. Needless to say, a large standard error often leads to less power in subsequent statistical analysis.

5 Real Data Examples

Three data sets from Helsel (2005c) serve to illustrate the use of the proposed estimation procedures. The first data set contains measurements of metals concentrations in stream sediments at 82 sites in New Mexico. The two variables, $y1996$ and $y2000$, represent measurements taken in 1996 and then in 2000 after wildfires. The metals concentrations below the detection limit 4 μg/L are recorded as censored observations. The second data set contains 423 measurements of Atrazine concentration, referred to as *AtraConc* in the following analysis, collected in streams across the Midwestern United States. There is one detection limit at 0.05 μg/L. Finally, the third data set contains measurements of dieldrin contamination in fish, denoted *Dieldrin*, collected at the Swindon, Burford, Northmoor, and Hannington Bridge sites near the Thames River, in the United Kingdom. The data set has a detection limit at 0.09. The histograms and normal probability plots of all log-transformed variables appear in Figure 4. The dashed line in the normal probability plot indicates a reference line without censored values. A quick glance at the figure suggests that a unimodal and symmetric distribution may be appropriate to model the $\log(y1996)$ and $\log(y2000)$ data, but a bimodal distribution may be more appropriate for the $\log(AtraConc)$ data. The distribution of $\log(Dieldrin)$ is more difficult to ascertain, due to its high percentage of censoring.

Table 2 summarizes the estimates of μ and σ derived from the estimation procedures discussed herein. As the simulation demonstrated, when the censoring percentage is low, the proposed estimation procedures, along with **ROS** and **EM**, should provide estimates comparable to those of **FULL**, were a complete sample available. However, **DL/2** is not expected to perform as well, even in the low censoring percentage case, and the performance of **MEst** should vary, depending on the underlying distribution. The censoring percentages of $\log(y1996)$ and $\log(y2000)$ are low at 4.88% and 1.22%, respectively. Except for **DL/2** and **MEst**, the estimates from the rest of the estimation procedures are similar, with the mean μ estimated around 2.49 and 2.53, and the standard deviation σ estimated around 0.53 and 0.48 for $\log(y1996)$ and $\log(y2000)$, respectively.

With a censoring percentage of approximately 10% and an underlying distribution that exhibits multiple modes, the proposed estimation procedures should perform better than the other competing procedures, and in most cases, their estimates are closest to those of **FULL**. The censoring percentage of $\log(AtraConc)$ is 11.11%, and its histogram suggests a bimodal distribution. The proposed procedures estimate μ as approximately -0.347 and σ to be 2.04, whereas the estimates from the other procedures vary dramatically, ranging from -0.122 to -0.529 for the mean estimates and from 1.773 to 2.190 for the standard deviation estimates.

It is difficult to assess the characteristics of the underlying distribution of $\log(Dieldrin)$, given its high censoring percentage and small sample size. But with a censoring percentage around 25%, the simulation again suggests that the use of the proposed estimation procedures is appropriate, considering their consistency across various distributional shapes. Table 2 shows that the mean and standard deviation estimates are -1.11 and 1.03 for the proposed procedures, respectively; for the rest of the estimation procedures, the estimates vary from -0.725 to -1.191 for the mean and from 0.478 to 1.167 for the standard deviation.

6 Summary and Conclusions

This article has proposed a moment estimation procedure for singly censored environmental data. With its weaker distributional assumption, the proposed procedure uses uncensored observations to learn about the tail behavior of the distribution where censored data reside. The censored observations are imputed with mean excess values or sample quantiles to create a fabricated sample, and the traditional sample moment estimators apply to the "complete" sample for the estimates. The simulation has demonstrated that the proposed procedure is robust according to the RMSE criterion, to censoring percentages below 30%, and a large class of distributions commonly used in modeling environmental data. This class includes the typical normal distribution, long-tailed double exponential distributions, heavy-tailed Student t distributions, and mixtures of normal distributions with multiple modes. Several real life data sets also were used to illustrate the proposed estimation procedures.

Yet several issues also require further exploration. In particular, a comparison between several imputation methods is a logical extension to the current study. Most notably, Rubin's multiple imputation procedure (Rubin 1987) that accounts for uncertainty in statistical inference due to missing values will add to the strength of the proposed method. Furthermore, regarding the choice of functional form $G(\cdot)$, this study has considered the power, exponential, and Weibull-type functional forms, but the larger question is whether there is a more flexible functional form that can capture a larger range of tail behavior. In addition, how can the tail symmetry assumption be tested with a censored sample, and how can the proposed tail-learning process be adjusted to account for samples with a larger percentage of censoring, as well as with multiple reporting limits? Finally, it still is necessary to explore different data transformation techniques and examine the possibility of a transformation bias.

Acknowledgements

The author thanks Dr. Leo R. Korn for providing the R codes of the **MEst** estimation procedure used in the simulation.

References

Aitchison, J. (1955), "On the Distribution of a Positive Random Variable Having a Discrete Probability Mass at the Origin," *Journal of the American Statistical Association,* 50, 901–908.

Antweiler, R. C. and Taylor, H. E. (2008), "Evaluation of Statistical Treatments of Left-Censored Environmental Data Using Coincident Uncensored Data Sets: I. Summary Statistics," *Environmental Science & Technology,* 42, 3732–3738.

Baccarelli, A., Pfeiffer, R., Consonni, D., Pesatori, A. C., Bonzini, M., Patterson Jr., D. G., Bertazzi, P. A., Landi, M. T. (2005), "Handling of Dioxin Measurement Data in the Presence of Non-detectable Values: Overview of Available Methods and Their Application in the Seveso Chloracne Study," *Chemosphere,* 60(7), 898–906.

Blom, G. (1958), *Statistical Estimates and Transformed Beta-Variables.* John Wiley & Sons, New York.

Cohen Jr., A. C. (1959), "Simplified Estimators for the Normal Distribution When Samples Are Singly Censored or Truncated," *Technometrics,* 1(3), 217–237.

—— (1961), "Table for Maximum Likelihood Estimates: Singly Truncated and Singly Censored Samples," *Technometrics,* 3, 535–554.

Cohen, M. A. and Ryan, P. B. (1989), "Observations Less than the Analytical Limit of Detection: A New Approach," *Journal of the Air Pollution Control Association,* 39(3), 328–329.

Cohn, T. A. (2005), "Estimating Contaminant Loads in Rivers: An Application of Adjusted Maximum Likelihood to Type I Censored Data," *Water Resources Research,* 41(7), 1–13.

Dempster, A. P., Laird, N. M. and Rubin, D. B., (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Finkelstein, M. M. (2008), "Asbestos Fibre Concentrations in the Lungs of Brake Workers: Another Look," *Annals of Occupational Hygiene,* 52, 455–461.

Flynn, M. R. (2010), "Analysis of Censor Exposure Data by Constrained Maximization of the Shapiro-Wilk *W*istic," *Annals of Occupational Hygiene,* 54, 263–271.

Gilliom, R. J. and Helsel, D. R. (1986), "Estimation of Distributional Parameters for Censored Trace Level Water Quality Data: 1. Estimation Techniques," *Water Resources Research,* 22(2), 135–146.

Gleit, A. (1985), "Estimation for Small Normal Data Sets with Detection Limits," *Environmental Science & Technology,* 19(2), 1201–1206.

Gupta, A. K. (1952), "Estimation of the Mean and Standard Deviation of a Normal Population from a Censored Sample," *Biometrika,* 39, 260–273.

Haas, C. N. and Scheff, P. A. (1990), "Estimation of Average in Truncated Samples," *Environmental Science & Technology,* 24(6), 912–919.

Hazen, A. (1914), "Storage to be Provided in Impounding Reservoirs for Municipal Water Supply (with discussion)," *Transactions of the American Society of Civil Engineers,* 77, 1539–1669.

Helsel, D. R. (1990), "Less Than Obvious: Statistical Treatment of Data Below the Detection Limit," *Environmental Science & Technology,* 24(12), 1766–1774.

—— (2005a), "Insider Censoring: Distortion of Data with Nondetects," *Human and Ecological Risk Assessment,* 11, 1127–1137.

—— (2005b), "More Than Obvious: Better Methods for Interpreting Nondetect Data," *Environmental Science & Technology,* 39 (20), 419A - 423A.

—— (2005c), *Nondetects and Data Analysis: Statistics for Censored Environmental Data*, John Wiley & Sons, Hoboken, NJ, 250 p.

—— (2006), "Fabricating Data: How Substituting Values for Nondetects Can Ruin Results, and What can be Done About it," *Chemosphere,* 65(11), 2434–2439.

—— (2010), "Much Ado About Next to Nothing: Incorporating Nondetects in Science," *Annals of Occupational Hygiene,* 54(3), 257–262.

—— and Cohn, T. A. (1988), " Estimation of Descriptive Statistics for Multiply Censored Water Quality Data," *Water Resources Research,* 24(12), 1997–2004.

—— and Gilliom, R. J. (1986), "Estimation of Distributional Parameters for Censored Trace Level Water Quality Data: 2. Verification and Applications," *Water Resources Research,* 22(2), 147–155.

Hewett, P. and Ganser, G. H. (2007), "A Comparison of Several Methods for Analyzing Censored Data," *Annals of Occupational Hygiene,* 51(7), 611–632.

Hill, B. M. (1975), "A Simple General Approach to Inference About the Tail of a Distribution," *Annals of Statistics,* 3(5), 1163–1174.

Huber, P. J. (1981), *Robust Statistics*, John Wiley & Sons, New York.

Hyndman, R. J. and Fan, Y. (1996), "Sample Quantiles in Statistical Packages," *The American Statistician,* 50(4), 361–365, November.

Korn, L. R. and Tyler, D. E. (2001), "Robust Estimation for Chemical Concentration Data Subject to Detection Limits," in *Statistics in Genetics and in the Environmental Sciences,* Fernholz, L. T., Morgenthaler, S. and Stahel, W. (editors), Trends in Mathematics, Birkhäuser Verlag Basel/Switzerland, 41–63.

Lambert, D., Peterson, B., and Terpenning, I. (1991), "Nondetects, Detection Limits, and the Probability of Detection," *Journal of the American Statistical Association,* 86, 266–277.

Liu, C. and Sun, D. X. (2000), "Analysis of Interval-Censored Data from Fractionated Experiments Using Covariance Adjustment," *Technometrics,* 42(4), 353–365.

Moulton, L. H. and Halsey, N. A. (1995), "A Mixture Model With Detection Limits for Regression Analyses of Antibody Response to Vaccine," *Biometrics,* 51, 1570–1578.

Newman, M. C., Dixon, P. M., Looney, B. B., and Pinder III, J. E., (1989), "Estimating Mean and Variance for Environmental Samples with Below Detection Limit Observations," *Water Resources Bulletin,* 25(4), 905–916.

Persson, T. and Rootzén, H. (1977), "Simple and Highly Efficient Estimators for a Type I Censored Normal Sample," *Biometrika,* 64(1), 123–128.

Pettitt, A. N. (1985), "Re-weighted Least Squares Estimation with Censored and Grouped Data: An Application of the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 47, 253–260.

Rényi, A. (1953), "On the Theory of Order Statistics," *Acta Mathematica Academiae Scientiarum Hungaricae*, 4, 191–231.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons, New York.

Saw, J. G. (1961), "Estimation of the Normal Population Parameters Given a Type I Censored Sample," *Biometrika*, 48(3/4), 367–377.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data.* Chapman and Hall, New York.

Schmee, J., Gladstein, D., and Nelson, W. (1985), "Confidence Limits for Parameters of a Normal Distribution From Singly Censored Samples, Using Maximum Likelihood," *Technometrics,* 27(2), 119–128, May.

Schneider, H. (1986), *Truncated and Censored Samples from Normal Populations*, Marcel Dekker, New York.

Shumway, R. H., Azari, A. S. and Johnson, P. (1989), "Estimating Mean Concentrations Under Transformation for Environmental Data with Detection Limits," *Technometrics,* 31(3), 347–356, August.

——, ——, and Kayhanian, M. (2002), "Statistical Approaches to Estimating Mean Water Quality Concentrations With Detection Limits," *Environmental Science and Technology,* 36(15), 3345–3353.

Singh, A. and Nocerino, J. (2001), "Robust Estimation of Mean and Variance Using Environmental Data Sets with Below Detection Limit Observations," U.S. Environmental Protection Agency, National Exposure Research Laboratory, Las Vegas, NV, Clearance Number 01062.

Travis, C. C. and Land, M. L. (1990), "Estimating the Mean of Data Sets with Nondetectable Values," *Environmental Science and Technology,* 24(7), 961–962.

n	Censoring %	**FULL**	**PwEM**	**EwEM**	**WwEM**	**DL/2**	**ROS**	**EM**	**MEst**
	5%	10.000	10.018	10.014	10.010	9.977	10.114	9.939	9.991
		(0.366)	(0.187)	(0.180)	(0.186)	(0.290)	(0.239)	(0.360)	(0.128)
	10%	9.999	10.027	10.012	9.984	9.823	10.096	9.849	9.979
		(0.316)	(0.183)	(0.168)	(0.169)	(0.247)	(0.182)	(0.347)	(0.134)
	15%	10.013	10.035	10.011	9.955	9.649	10.053	9.755	9.976
100		(0.348)	(0.209)	(0.155)	(0.170)	(0.265)	(0.149)	(0.416)	(0.136)
	20%	9.994	10.033	10.007	9.917	9.452	9.988	9.603	9.974
		(0.338)	(0.188)	(0.150)	(0.171)	(0.246)	(0.153)	(0.474)	(0.139)
	25%	9.986	10.028	10.002	9.871	9.238	9.907	9.466	9.972
		(0.295)	(0.168)	(0.144)	(0.184)	(0.223)	(0.175)	(0.459)	(0.137)
	30%	10.003	10.043	10.009	9.837	9.049	9.810	9.284	9.977
		(0.346)	(0.206)	(0.143)	(0.208)	(0.268)	(0.326)	(0.579)	(0.138)
	5%	9.999	10.025	10.013	10.004	9.984	10.118	9.933	9.992
		(0.279)	(0.145)	(0.134)	(0.136)	(0.181)	(0.154)	(0.330)	(0.097)
	10%	9.997	10.024	10.005	9.967	9.822	10.090	9.843	9.977
		(0.230)	(0.137)	(0.116)	(0.119)	(0.182)	(0.124)	(0.250)	(0.093)
	15%	10.006	10.034	10.009	9.929	9.650	10.048	9.732	9.979
200		(0.229)	(0.129)	(0.110)	(0.126)	(0.170)	(0.106)	(0.344)	(0.096)
	20%	9.996	10.033	10.002	9.883	9.457	9.973	9.563	9.972
		(0.262)	(0.144)	(0.099)	(0.138)	(0.191)	(0.116)	(0.458)	(0.090)
	25%	9.989	10.029	9.998	9.829	9.247	9.887	9.412	9.971
		(0.240)	(0.129)	(0.097)	(0.162)	(0.169)	(0.157)	(0.580)	(0.092)
	30%	9.955	10.039	10.000	9.779	9.052	9.770	8.985	9.970
		(1.083)	(0.174)	(0.096)	(0.195)	(0.221)	(0.329)	(5.988)	(0.093)

Table 1: Averages of 1000 estimates of the true mean μ. The samples were generated from a Student t distribution with 2 degrees of freedom. The true mean μ is 10.0, and the true standard deviation σ does not exist. Standard errors are given in parentheses.

Variable	n	Censoring %	Parameter	**PwEM**	**EwEM**	**WwEM**	**PwSQ**	**EwSQ**	**WwSQ**	**DL/2**	**ROS**	**EM**	**MEst**
log(y1996)	82	4.88%	μ	2.488	2.488	2.488	2.488	2.488	2.488	2.457	2.495	2.481	2.584
			σ	0.533	0.533	0.533	0.533	0.533	0.533	0.613	0.520	0.548	0.571
log(y2000)	82	1.22%	μ	2.534	2.536	2.533	2.535	2.536	2.534	2.527	2.536	2.534	2.583
			σ	0.482	0.478	0.485	0.481	0.478	0.484	0.504	0.478	0.480	0.475
log($AtraConc$)	423	11.11%	μ	-0.347	-0.346	-0.347	-0.346	-0.346	-0.347	-0.122	-0.404	-0.405	-0.529
			σ	2.042	2.041	2.043	2.049	2.047	2.048	1.773	2.155	2.156	2.190
log($Dieldrin$)	31	25.8%	μ	-1.108	-1.108	-1.115	-1.104	-1.105	-1.114	-0.725	-0.821	-1.191	-0.860
			σ	1.022	1.022	1.034	1.024	1.024	1.038	0.478	0.609	1.167	1.066

Table 2: Mean μ and standard deviation σ estimates of four real-life samples.

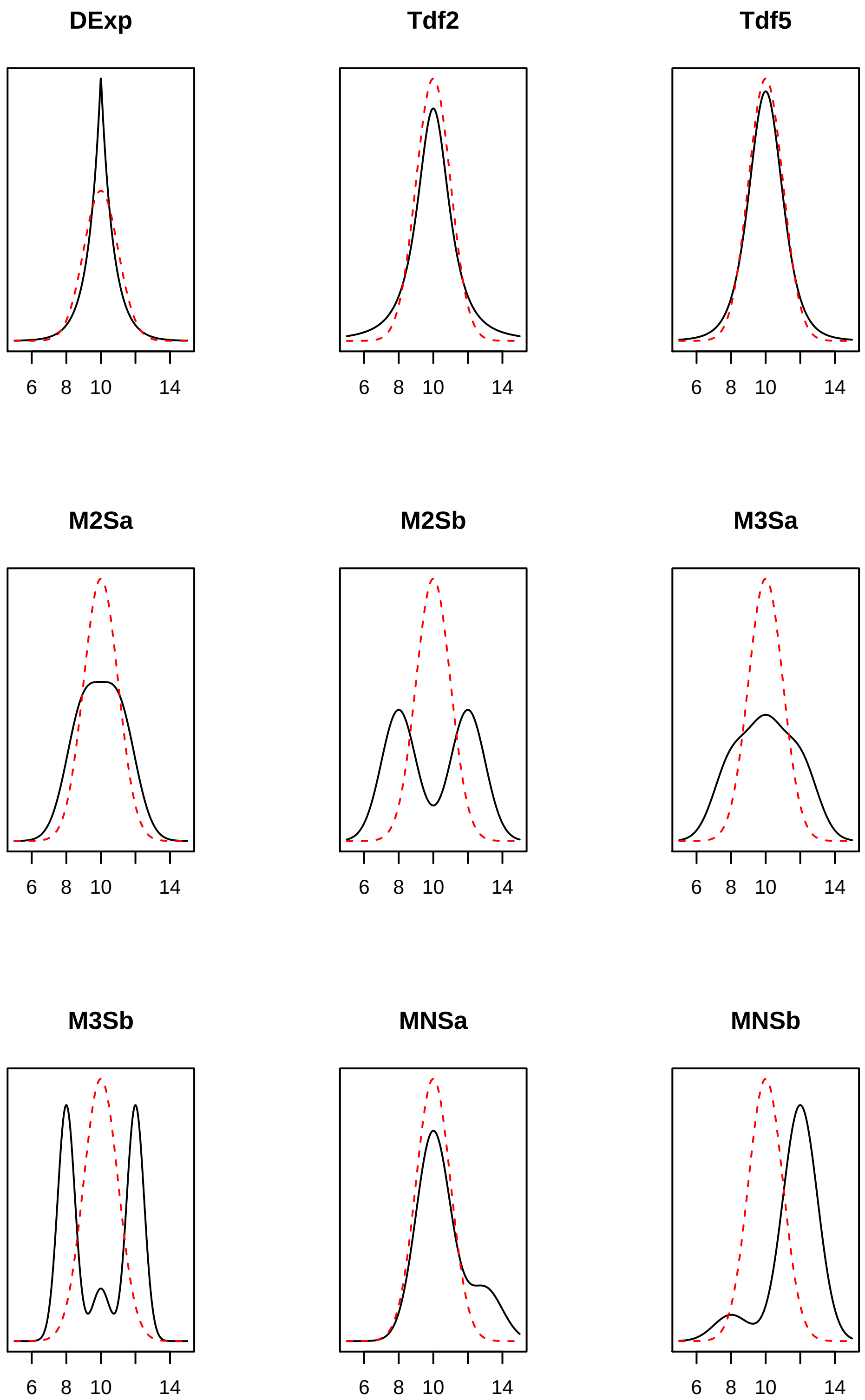

Figure 1: Densities of distributions considered in the simulation. The normal distribution is denoted by dashed lines.

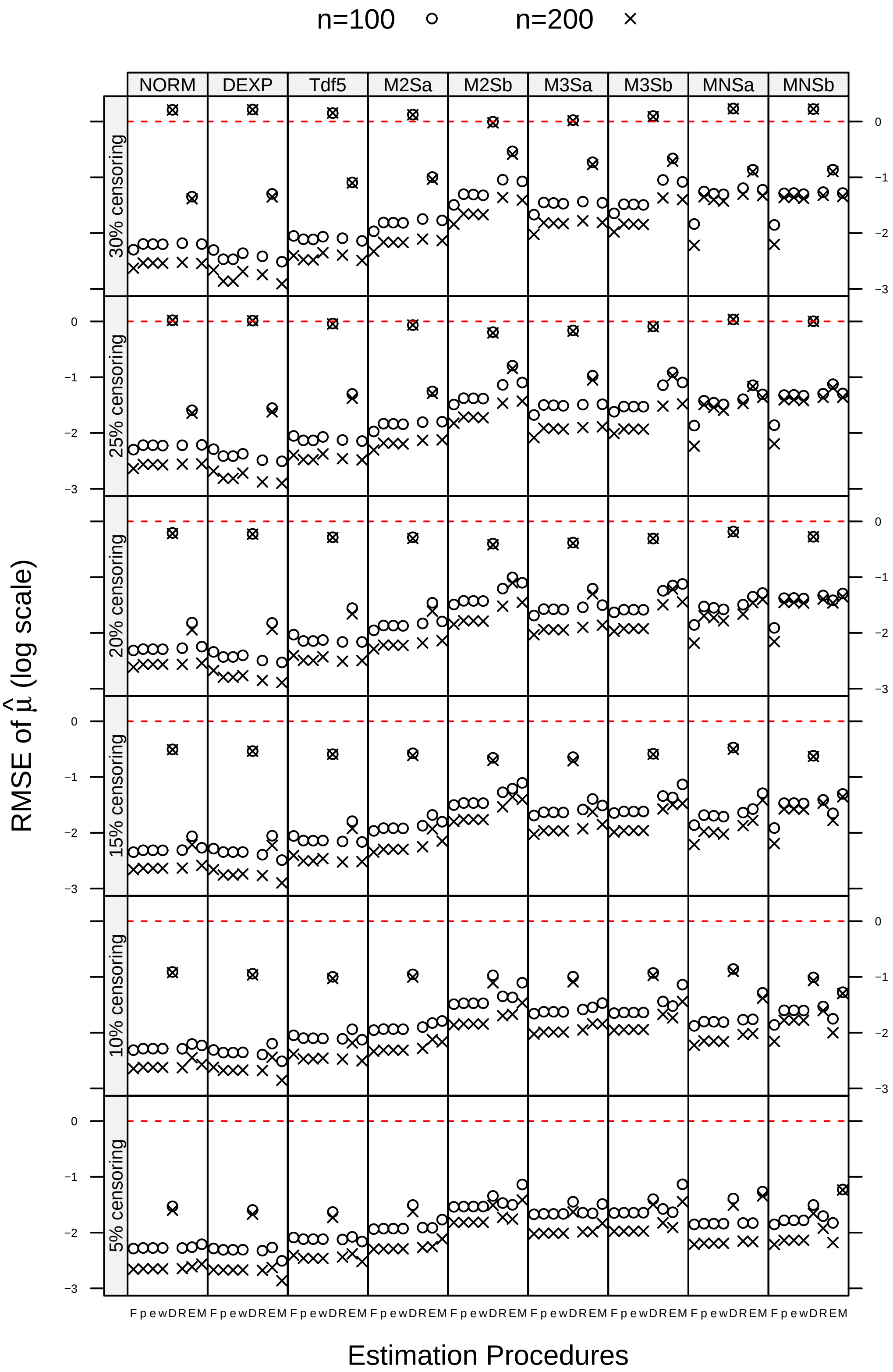

Figure 2: Performance of the mean μ estimation procedures. The root mean square errors on the log scale are reported.

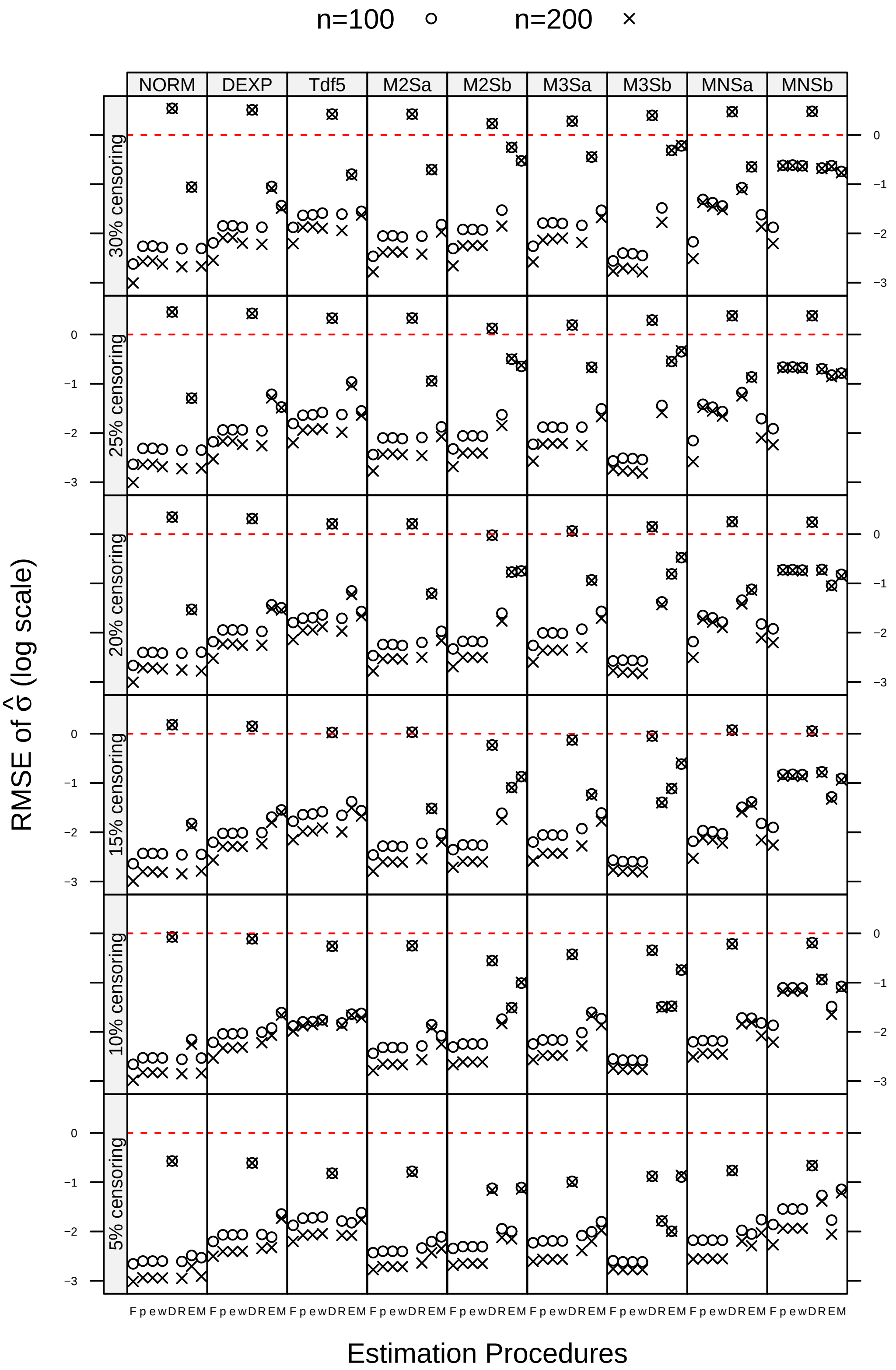

Figure 3: Performance of the mean σ estimation procedures. The root mean square errors on the log scale are reported.

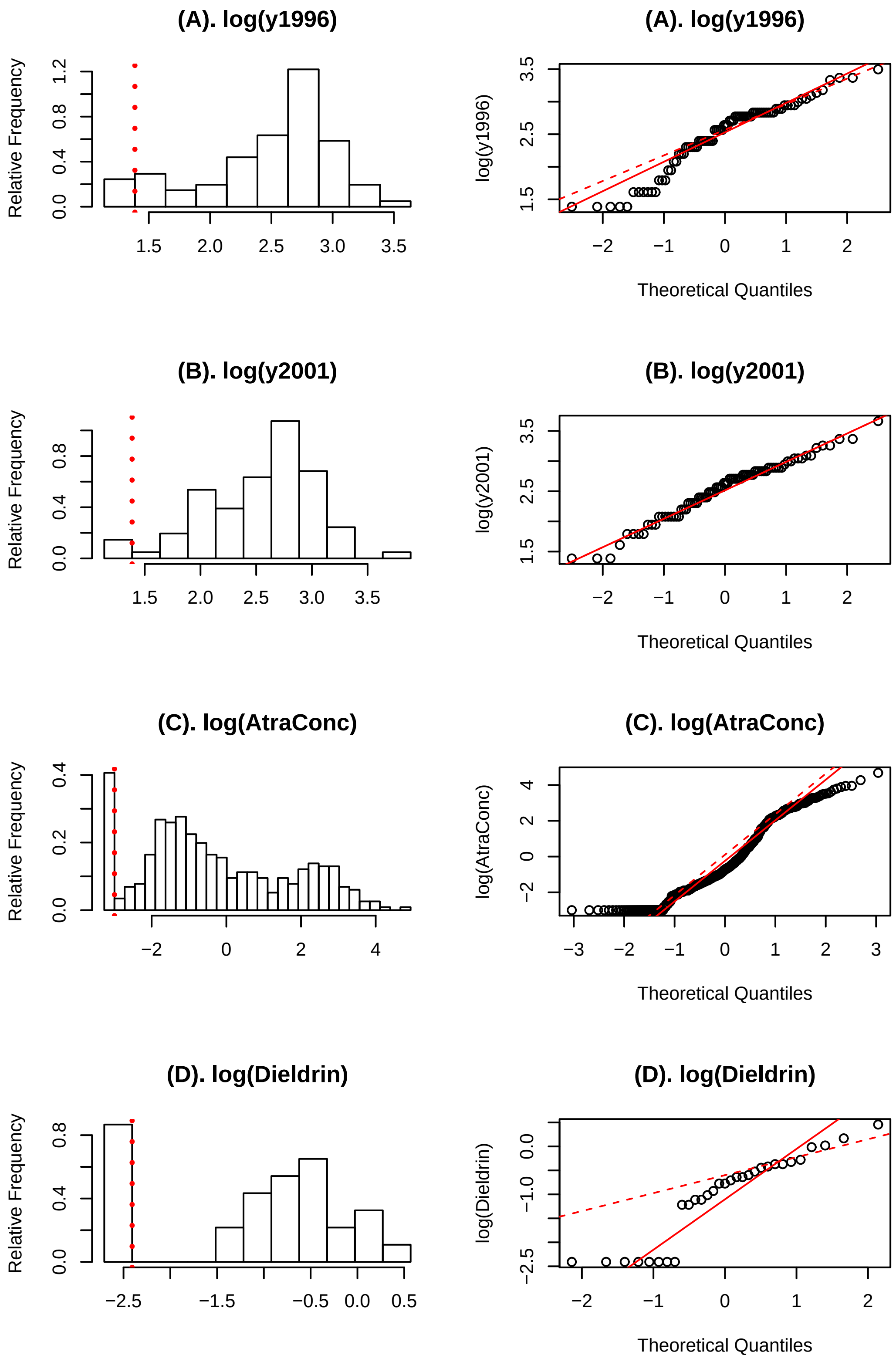

Figure 4: Histograms and normal probability plots for the log-transformed variables: (a) y1996, (b) y2001, (c) AtraConc, and (d) Dieldrin.