UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

Robust Estimation of Mean and Variance Using Environmental Data Sets with Below Detection Limit Observations

Journal Article Published May 2001
Clearance Number 01-062

Robust Estimation of Mean and Variance Using Environmental Data Sets with Below Detection Limit Observations

by

Anita Singh
Lockheed-Martin Environmental Systems & Technologies Company
980 Kelly Johnson Drive
Las Vegas, NV 89119

John Nocerino
United States Environmental Protection Agency
National Exposure Research Laboratory
Post Office Box 93478
Las Vegas, NV 89193-3478

Notice

The U.S. Environmental Protection Agency (EPA), through its Office of Research and Development (ORD), funded and collaborated in the research described here. It has been subjected to the Agency's peer review and has been approved as an EPA publication. The U.S. Government has a non-exclusive, royalty-free license in and to any copyright covering this article.

Summary

Scientists, especially environmental scientists often encounter trace level concentrations that are typically reported as less than a certain limit of detection, L. Type I, left-censored data arise when certain low values lying below L are ignored or unknown as they cannot be measured accurately. In many environmental quality assurance and quality control (QA/QC), and groundwater monitoring applications of the United States Environmental Protection Agency (U.S. EPA), values smaller than L are not required to be reported. However, practitioners still need to obtain reliable estimates of the population mean, μ, and the standard deviation (sd), σ. The problem gets complex when a small number of high concentrations are observed with a substantial number of concentrations below the detection limit. The high outlying values contaminate the underlying censored sample, leading to distorted estimates of μ and σ. The U.S. EPA, through the National Exposure Research Laboratory-Las Vegas (NERL-LV), under the Office of Research and Development (ORD), has research interests in developing statistically rigorous robust estimation procedures for contaminated left-censored data sets. Robust estimation procedures based upon a proposed (PROP) influence function are shown to result in reliable estimates of population parameters of mean and sd using contaminated left-censored samples. It is also observed that the robust estimates thus obtained with or without the outliers are in close agreement with the corresponding classical estimates after the removal of outliers. Several classical and robust methods for the estimation of μ and σ using left-censored (truncated) data sets with potential outliers have been reviewed and evaluated.

Key Words: Type I censoring, Type II censoring, left-censored (truncated) data, detection limit, robust statistics, Monte Carlo simulation, mean square error (MSE), PROP influence function, unbiased maximum likelihood estimation (UMLE), Cohen’s maximum likelihood estimation, Persson and Rootzen’s restricted maximum likelihood estimation (RMLE), expectation-maximization (EM) algorithm, regression methods.

Table of Contents

Section 1

Introduction

The processing of the analytical results of environmental samples containing potentially hazardous chemicals is often complicated by the fact that some of these pollutants are present at trace levels, which cannot be measured reliably and therefore are reported as results lying numerically below a certain limit of detection, L. This results in left-censored data sets. In many environmental monitoring applications, values smaller than L are not even required to be reported. However, since the presence of some of these toxic pollutants (*e.g.*, dioxin) in the various environmental media can pose a threat to human health and the environment even at trace level concentrations, these non-detects cannot be ignored or deleted (often done in practice) from subsequent analyses. For site characterization purposes such as to establish mean contamination levels at various parts of a polluted site, it is desirable to obtain reliable estimates of μ and σ using the left-censored data sets. The problem gets complicated when some outliers are also present in conjunction with the non-detects. Also, sometimes in environmental applications, non-detects (*e.g.*, due to matrix effects) exceed the observed values adding to the complexity of the estimation procedures. Improperly obtained estimates of these parameters can result in inaccurate estimates of cleanup standards, which in turn can lead to incorrect remediation decisions at a polluted site. In this article, emphasis is given to obtain robust estimates of population mean and sd using left-censored data sets with potential outliers in the right tail of a data set. In this study, it is assumed that all non-detects are smaller than the observed values.

In general, censoring means that observations at one or both extremes (tails) are not available. In Type I censoring, the point of censoring (*e.g.,* the detection limit, L) is "fixed" a priori for all observations and the number, $k(\geq 0)$, of the censored observations varies. In Type II censoring, the number of censored observations, k, is fixed a priori, and the point(s) of censoring vary. For example, Type II right-censoring (large values are not available) typically occurs in life testing and reliability applications. In a life testing application, n items (*e.g.*, electronic items) are subjected to a life testing experiment which terminates as soon as (n-k) of the n data values have been observed (failed). The lifetimes of the remaining k living objects are unavailable or being censored.

The estimation of the parameters of normal and lognormal populations from censored samples has been studied by several researchers, including *Cohen* [1950, 1959], *Persson and Rootzen* [1977], *Gleit* [1985], *Schneider* [1986], *Gilliom and Helsel* [1986]. A myriad of estimation procedures for Type I left-censored data exist in the literature, including simple substitution methods and several rigorous procedures such as Cohen's maximum likelihood estimation (MLE) procedure, Persson and Rootzen's restricted MLE (RMLE) method, and regression methods (*Gilliom and Helsel* [1986], *Newman, Dixon, and Pinder* [1989]). The commonly used substitution methods are: replacement of below detection limit data by zero, or by half of the detection limit, L/2, or by the detection limit, L itself.

Using Monte Carlo simulation experiments, several researchers, including *Gleit* [1985], *Gilliom and Helsel* [1986], and *Haas and Scheff* [1990], concluded that the data substitution methods resulted in a biased estimate of the population mean. In practice, probably due to computational ease, these data substitution methods are commonly used in many environmental applications. Depending upon the sample size, n, and the censoring intensity, k, substitution of the censored values by L/2 is one of the recommended methods in some U.S. EPA guidance documents, such as the *Guidance for Data Quality Assessment, 96*. None of the simulation studies conducted so far included the unbiased maximum likelihood estimation (UMLE) method. Also, the results and conclusions of the above-mentioned studies are not directly comparable due to reasons discussed in the following paragraphs.

Gleit [1985] performed simulation experiments for a "fixed" detection limit, L, for various censoring intensities. Based upon *Dempster, Laird, and Rubin's* [1977] expectation-maximization (EM) algorithm, *Gleit* used the conditional expected values of order statistics of the Gaussian distribution for censored observations. Based on the low mean square error (MSE) criterion, he recommended the use of the EM method which replaces all of the non-detects by the conditional expected value of the order statistics as given by equation (11) below. *Gleit's* simulation experiments did not include *Persson and Rootzen's* RMLE method or any of the regression methods.

Gilliom and Helsel [1986] performed simulation experiments for various distributions and several levels of censoring intensities. Their simulation experiments used the "computed" detection limit based on the distribution used. For example, for a normal distribution with mean, 5, and sd, 2, ~ N(5,2), and for censoring intensities of 30% L and 60% L, will be $5+2*z_{0.30} \sim 5-2*0.525=3.95,$ and $5+2*z_{0.60} \sim 5+2*0.255=5.51,$ respectively, where z_{α} represents a value of the standard normal deviate such that area to the left of z_{α} is $\alpha.$ Thus, the limit, L, changes with the censoring intensity. They concluded that the extrapolation regression approach on the log-transformed data results in an estimate of the population mean with the smallest root mean square error (RMSE). Their simulation study did not include the RMLE method and the EM algorithm.

Haas and Scheff [1990] and *Lechner* [1991] compared the performance of classical estimation methods for left-censored samples in terms of bias and MSE. They also used the "computed" detection limit based on the distribution used in their simulation experiments. They concluded that the bias-corrected RMLE procedure results in as good estimates as the Cohen's MLE method, and also the RMLE method possesses lower bias and MSE than the regression and substitution methods. They also suggested that the RMLE is less sensitive to the deviations from normality. Their study did not include the EM method.

The objective of the present article is to develop robust procedures which yield reliable estimates of population parameters from left-censored data sets in the presence of outliers, and also to compare the performances of the various estimation procedures. The authors of this article performed Monte Carlo simulation experiments for both the "fixed" and the "computed" detection limit cases to assess the performances of the various classical and robust procedures in terms of bias and MSE. Several methods, including the EM algorithm, MLE, UMLE, RMLE, and the regression method, have been considered. The results of a couple of simulation runs are presented here to demonstrate the differences in the performances of these methods for the two cases: 1) L stays fixed for all censoring levels, and 2) L is computed based on the distribution used and, therefore, varies with the censoring intensity. In environmental applications, the first case (1) of a fixed detection limit occurs quite frequently.

The occurrence of non-detects in combination with potential outliers is inevitable in data sets originating from environmental applications. The data set resulting from such a combination of non-detects in the left tail of the distribution, and high concentrations in the right tail of the distribution, typically do not follow a well-known statistical distribution. The problem gets complex when multiple detection limits (reporting limits) are present. In practice, such a data set may have been obtained from two or more populations with significantly different mean concentrations such as the one coming from the clean background part of the site and the other obtained from a contaminated part of the site. Unfortunately, many times such a data set can be modeled incorrectly by a lognormal distribution (which could pass the lognormality test). Also, a normally distributed data set with a few extreme (high) observations can be incorrectly modeled by a lognormal distribution with the lognormal assumption hiding the outliers [*Singh, Singh, and Engelhardt* (1997, 2000)]. An example is discussed next to elaborate on this point.

Example 1. A simulated data set with 15 observations has been obtained from a mixture of two normal populations. Ten observations (representing background) were generated from a normal distribution with a mean of 100 and a sd of 50, and five observations (representing contamination) were generated from a normal distribution with a mean of 1000 and a sd of 100. The mean of this mixture distribution is 400. The generated data are: 180.51, 2.33, 48.67, 187.07, 120.21, 87.96, 136.75, 24.47, 82.23, 128.38, 850.91, 1041.73, 901.92, 1027.18, and 1229.94. The data set failed the normality test based on several goodness-of-fit tests, such as the Shapiro-Wilk test, the W-test (W=0.7572), and the Kolmogorov-Smirnov (K-S = 0.35) test. However, when these tests were carried out on the log-transformed data, the test statistics are insignificant at the $\alpha = 0.05$ level of significance with W=0.8957 and K-S = 0.168, suggesting that a lognormal distribution provides a reasonable fit to the data. Based upon those tests, one might conclude that the observed data come from a single background lognormal population, a situation which occurs frequently in practice. It is, therefore, warranted to make sure that the data come from a single population before one would try to use a lognormal distribution on a data set.

Like full uncensored samples, the classical procedures used on left-censored data sets with potential high outliers result in distorted estimates of location and scale. A brief description of some of the above-mentioned procedures is given in Section 2 and some real and simulated data sets are discussed in Section 3. Results of a few simulation runs for the various classical methods are given in Section 4, and the conclusions are summarized in Section 5. The simulation results based on robust procedures are in agreement with the classical procedures without outliers. However, due to the length of the present article, results of the complete Monte Carlo study (using data sets with outliers) to assess the performances of the various classical and robust procedures are not included in this article, and will be submitted for publication at a later date.

Section 2

Mathematical Formulation

In the following, it has been implicitly assumed that the data set under consideration has been obtained from a "single" normal population, perhaps after a suitable *Box-Cox* [1964] type transformation (including the log-transformation) with unknown mean, μ, and sd, σ. *Cohen* [1950, 1959] derived the maximum likelihood (ML) equations for censored samples and prepared tables of the constants needed to obtain the MLEs of μ and σ. The ML equations are solved iteratively using a suitable numerical method such as the Newton-Raphson method. Some computer programs (*e.g.*, UNCENSOR by *Newman et al.* [1989]) are available to compute the MLEs and the RMLEs from left-censored samples obtained from normal and lognormal populations. Some of these classical and robust methods have been incorporated into a computer program, CENSOR (see *Scout: A Data Analysis Program*), for estimation of μ and σ from left-censored data sets with potential outliers.

Let $x_1, x_2, ..., x_n$ be a random sample from a normal, N(μ,σ), population with k of the non-detects, $x_1, x_2, ..., x_k$, lying numerically below the detection limit, L. Let ϕ and Φ be the probability density function (pdf) and cumulative distribution function (cdf) of the standard normal distribution (snd). The logarithm of the likelihood function is given as follows:

$$\ln L(\boldsymbol{x}, \mu, \sigma) = k \ln \Phi(Z) - n_0 \ln \sigma - \sum_{k+1}^{n} (x_i - \mu)^2 / 2\sigma^2 + constant, \tag{1}$$

where $n_0 = (n-k)$ and $Z = (L-\mu)/\sigma$, with $\Phi(Z)$ representing the probability that an observation is less than L. The mean $,\bar{x}_o,$ and variance $,s_o^2,$ using the (n-k) observed data values are:

$$\bar{x}_o = \sum_{i=k+1}^{n} x_i/(n-k), \text{ and } s_o^2 = \sum_{i=k+1}^{n} (x_i - \bar{x}_o)^2/(n-k). \tag{2}$$

A brief description of some of the robust procedures to estimate population parameters from contaminated left-censored samples is given as follows.

Robust Procedures

When dealing with data sets originated from environmental applications, one is faced with the dual problem of the occurrence of below detection limit concentrations (non-detects) in the left tail and possibly some extreme concentrations in the right tail of the distribution of the contaminant (*e.g.*, lead) under consideration. The presence of outliers leads to distorted estimates of the population mean, μ, and the sd, σ. It is, therefore, important that these unusual observations in both tails of the distribution be treated adequately. For full uncensored data sets, simple robust estimates such as the trimmed mean or Winsorized mean (*Hoaglin, Mosteller, and Tukey* [1983]) are sometimes used to estimate the population mean in the presence of outliers. For example, a *100p%* trimmed mean is obtained by using only the middle *n(1-2p)* data values and the *np* values are omitted from each of the two (left and right) tails of the data set. *Gilbert* [1987] suggested the use of the Winsorized and trimmed means for the estimation of μ and σ for left-censored data sets. Depending upon the censoring intensity, the use of the trimmed and Winsorized mean has also been recommended in some guidance documents, such as *Guidance for Data Quality Assessment*, 1996. *Helsel* [1990] discussed the use of non-parametric, distribution-free procedures, and of a simple robust pair, (median, MAD/0.6745), to estimate μ and σ, where MAD represents the median absolute deviation. *Gilliom and Helsel* [1986] suggested the use of the least-squares regression on the log-transformed data to obtain robust estimates of μ and σ from left-censored data sets.

In this article, we use robust M-estimation procedures based on the notion of the influence function (*Hampel* [1974]), which assigns reduced weights to the outlying observations. For full uncensored data sets, several robust procedures exist in the literature for the estimation of the population mean and the variance (*Huber* [1981], *Rousseeuw and Leroy* [1984], *Staudte and Sheather* [1990], *Singh and Nocerino* [1995]). For left-censored data sets, in order to identify and subsequently assign reduced weights to the outliers that may be present in the right tail of a data set, the robust sample mean, $\overline{x}_o^*,$ and the robust sd, $s_o^*,$ using the (n-k) observed values, need to be obtained first. These values are then used in the various estimation methods, such as MLE, UMLE, RMLE, and the EM method to obtain robust estimates of the population mean and sd. *Singh and Nocerino* [1995] showed that for full data sets, the PROP influence function works very well for 1) the identification of multiple multivariate outliers, and 2) the robust estimation of the population mean vector and the dispersion matrix. In this article, those techniques are extended to obtain the robust estimates of the population mean and the variance using left-censored data sets with outliers.

The PROP influence function (*Singh* [1993]) and the corresponding iteratively obtained sample mean and sd based on the (n-k) detected observations are given as follows.

$$\begin{aligned}\psi(d_i) &= d_i &&;\ d_i \le d_o\\ &= d_o\exp\left[-(d_i-d_o)\right] &&;\ d_i > d_o\end{aligned} \tag{3}$$

$$\overline{x}_o^* = \sum_{(k+1)}^{n} w_1(d_i)\,x_i \Big/ \sum w_1(d_i)\ ; \qquad s_o^{*2} = \sum_{(k+1)}^{n} w_2(d_i)\,(x_i - \overline{x}_o^*)^2/\nu\,. \tag{4}$$

Here, $d_i^2 = (x_i-\overline{x}_o^*)^2/s_o^{*2};\ i=k+1,k+2,...,n,$ and d_o^2 is the α*100% critical value from the scaled beta distribution, $(n-k-1)^2\ \beta(1/2,\ (n-k-2)/2)/(n-k),$ of the distances, $d_i^2.$

The weights are given by $w_1(d_i) = w_i = \psi(d_i)/d_i,$ and $w_2(d_i) = w_i^2 = w_1^2(d_i),$ with the degrees-of-freedom, $\nu = wsum2 - 1,$ and $wsum1 = \sum w_1(d_i), wsum2 = \sum w_2(d_i).$

Since the number of outliers present in a data set is usually unknown, more than one value of α should be tried on the same data set. The commonly used values of the level α are 0.01, 0.05. In the presence of multiple outliers, higher values of α (*e.g.*, 0.1) may be needed (*Singh and Nocerino* [1995]) to unmask outliers. This is especially true when the sample size is small (*e.g.*, 20 or less). It needs to be pointed out that the outlier identification procedures based on influence functions typically identify extreme observations in both tails of the underlying distribution. When dealing with left-censored data sets, one is concerned with the identification of outlying observations that might be present in the right tail of the distribution; therefore, reduced weights are to be assigned to those extreme observations found in the right tail only. Each of the observed detected values in the left-tail is assigned a unit weight.

Cohen's Method

Cohen's MLEs for the mean and the variance are obtained by solving the following equations:

$$\hat{\mu}_{MLE} = \bar{x}_o - (\bar{x}_o - L)\lambda(g, h), \text{ and } \hat{\sigma}^2_{MLE} = s_o^2 + (\bar{x}_o - L)^2\lambda(g, h), \tag{5}$$

where $g = s_o^2/(\bar{x}_o - L)^2$ and $h = k/n.$ The estimates of μ and σ given by equation (5) are biased. For a Type II censored data set from normal population, *Saw* [1961] tabulated the first-order bias correction terms, which were simplified by *Schneider* [1986] and are given as follows.

$$Bias_{\hat{\mu}} = -\exp[2.692 - 5.439(n-k)/(n+1)], \text{ and} \tag{6}$$

$$Bias_{\hat{\sigma}} = -[0.312 + 0.859(n-k)/(n+1)]^{-2}. \tag{7}$$

In practice, the bias corrections given by equations (6) and (7) are also used for Type I censored data. The bias-corrected MLE, denoted by UMLE, is given as follows.

$$\hat{\mu}_{UMLE} = \hat{\mu}_{MLE} - \frac{\hat{\sigma}_{MLE} Bias_{\hat{\mu}}}{(n+1)}, \text{and } \hat{\sigma}_{UMLE} = \hat{\sigma}_{MLE} - \frac{\hat{\sigma}_{MLE} Bias_{\hat{\sigma}}}{n+1}. \tag{8}$$

The corresponding robust ML and UML estimates of μ and σ are obtained by using the robust estimates, $\bar{x}_o^*$ and s_o^*, in place of $\bar{x}_o,$ and s_o in equations (5) and (8), respectively.

Expectation Maximization (EM) Algorithm

Dempster, Laird, and Rubin [1977] developed the EM algorithm to maximize the likelihood function based upon censored and missing data. The iterative EM algorithm works on the observed values assuming that no observations were censored. At the initial iteration, using the observed (n-k) data values, one could start with some convenient estimates for μ and σ, such as the sample mean and sd, or a simple one-step robust pair represented by the median and MAD/0.6745. The iterations are defined as successively maximizing the expectation of the conditional likelihood function of the complete data, given the type of censoring. *Gleit* [1985] used this procedure for left-censored samples and found it to possess a lower MSE than the various other substitution and likelihood procedures. For the single detection limit case, the estimates of μ and σ at the (j+1)th iteration are given as follows (*Shumway, Azari, and Johnson* [1989]).

$$\hat{\mu}_{j+1} = [\sum_{i=k+1}^{n} x_i + \sum_{i=1}^{k} E_j(X_i | X_i \le L)]/n, \tag{9}$$

$$\hat{\sigma}^2_{j+1} = [\sum_{i=k+1}^{n} (x_i - \hat{\mu}_j)^2 + \sum_{i=1}^{k} E_j((X_i - \mu_j)^2 | X_i \le L)]/(n-1), \quad \text{where} \tag{10}$$

$$E_j[X_i | X_i \le L] = \hat{\mu}_j - \hat{\sigma}_j[\phi(Z)/\Phi(Z)], \quad \text{with} \quad Z = (L - \hat{\mu}_j)/\hat{\sigma}_j, \tag{11}$$

$$E_j[(X_i - \mu_j)^2 | X_i \le L] = \hat{\sigma}^2_j (1 - Z[\phi(Z)/\Phi(Z)]). \tag{12}$$

Thus the EM method is an iterative substitution method in which at each iteration all of the non-detects are replaced by the same conditional expected value as given by equation (11). In the presence of outliers, the conditional expected value given by equation (11) gets distorted (*e.g.*, becomes negative), and results in inadequate estimates given by equations (9) and (10). Typically, contaminant concentrations are non-negative and substituting a negative value for non-detects will be inappropriate. In these cases, the non-detects have to be replaced by zero, or half of the detection limit, L/2 (see Example 4). In this article, whenever the conditional expected value became negative, it was replaced by L/2. As shown in the examples to follow, the robust EM estimation procedure takes care of this problem by assigning reduced weights to the outlying observations. The robust EM estimates at the (j+1)th iteration are given as follows:

$$\hat{\mu}_{j+1} = [\sum_{i=k+1}^{n} w_i x_i + \sum_{i=1}^{k} E_j(X_i | X_i \le L)]/(wsum1 + k), \quad \text{and} \tag{13}$$

$$\hat{\sigma}^2_{j+1} = [\sum_{i=k+1}^{n} w_i^2 (x_i - \hat{\mu}_j)^2 + \sum_{i=1}^{k} E_j((X_i - \mu_j)^2 | X_i \le L)]/(wsum2 + k - 1). \tag{14}$$

Restricted Maximum Likelihood (RMLE) Method

Persson and Rootzen [1977] obtained the restricted likelihood estimates by simplifying the ML equations. The likelihood function can be written as follows:

$$L(\boldsymbol{x},\mu,\sigma) = [\Phi(Z)]^{k}(2\pi\sigma^{2})^{-(n-k)/2}\exp-[\sum_{i=(k+1)}^{n}(y_i+Z\sigma)^{2}/2\sigma^{2}], \tag{15}$$

where $y_i=x_i-L$; i:=k+1, k+2,..., n. The random variable, (n-k), representing the number of observed values above L, can be expressed as a binomial random variable with the pdf given below.

$$\text{P(No. of observations lying above L = r)} = [n!/r!(n-r)!](1-\Phi(Z))^{r}\Phi^{n-r}(Z) \tag{16}$$

where r = 0,1,2,..., n. An estimate of $\Phi(Z)$, the probability that an observation lies below L, is k/n. Thus, for 0<k<n, an estimate, $\lambda_{k/n}$, of Z is given by $\hat{Z}=\lambda_{k/n}=\Phi^{-1}(k/n)$.
Substituting $\lambda_{k/n}$ for Z in equation (15) and then maximizing the resulting restricted likelihood function yields the following closed form estimates of μ and σ.

$$\hat{\sigma}_{RML}=\frac{1}{2}[c+(c^{2}+\frac{4}{(n-k)}\sum_{k+1}^{n}y_i^{2})^{1/2}], \text{ and } \hat{\mu}_{RML}=L-\lambda_{k/n}\hat{\sigma}_{RML}, \tag{17}$$

where $c=\lambda_{k/n}\sum_{k+1}^{n}y_i/(n-k)$. The estimates given by equation (17) are biased, which can be corrected as follows. For left-censored samples, $E[\bar{x}_o]=\mu+\sigma\alpha$, and $E[s_o^{2}]=\sigma^{2}[1+(\alpha Z-\alpha^{2})]$, where $\alpha=\phi(Z)/(1-\Phi(Z))$, and the bias-corrected RMLEs are given as follows:

$$\hat{\mu}_{BRML}=\bar{x}_o-\hat{\alpha}\hat{\sigma}_{BRML}, \text{ and } \hat{\sigma}_{BRML}=[s_o^{2}-(\hat{\alpha}\lambda_{k/n}-\hat{\alpha}^{2})\hat{\sigma}_{RML}^{2}]^{1/2}, \tag{18}$$

where $\hat{\alpha}=\phi(\lambda_{k/n})/(1-k/n)$.

The robust RMLEs are obtained by assigning reduced weights to each of the outlying observation present in the right tail of the data set. The bias corrected robust RMLEs are given by:

$$\hat{\mu}^{*}_{BRML}=(\sum_{i=k+1}^{n}w_i x_i/wsum1)-\hat{\alpha}\hat{\sigma}_{BRML}, \text{ and} \tag{19}$$

$$\hat{\sigma}^{*}_{BRML}=[(\sum_{i=k+1}^{n}w_i^{2}x_i^{2}/wsum2)-(\sum_{i=k+1}^{n}w_i x_i/wsum1)^{2}-(\hat{\alpha}\lambda_{k/n}-\hat{\alpha}^{2})\hat{\sigma}_{RML}^{2}]^{1/2}. \tag{20}$$

Regression Method

The ordinary least squares (OLS) regression line is obtained by fitting a model to the observed data (perhaps after a suitable transformation) and the hypothetical normal quantiles. In other words, it is assumed that the k censored observations, $x_1, x_2, ..., x_k,$ follow the zero-to-detection limit portion of a normal (transformed) distribution. A least squares regression line is obtained using the (n-k) pairs, $(q_{(i)}, x_{(i)});$ i=k+1, k+2,...,n, where $x_{(i)}$ are the observed values arranged in ascending order. The n quantiles, $q_{(i)},$ are obtained using an appropriate normal probability statement, such as $P[z \le q_{(i)}] = (i-3/8)/(n+1/4); i:=1,2,...,n$ (*Johnson and Wichern* [1988]). The fitted OLS regression line is given by:

$$x_{(i)} = a + bq_{(i)}, i:=k+1, k+2, ..., n. \tag{21}$$

The mean, μ, and sd, σ, can be estimated in two ways: 1) by using the intercept and slope of the fit given by equation (21), and 2) by the extrapolation of the non-detects obtained using the model given by equation (21). The extrapolation regression approach (labeled as the Regress method) estimates the population mean and sd using the (n-k) observed data values and the k extrapolated non-detects. For full data sets, *Barnett* [1975] used the intercept and the slope of the regression line to estimate the population mean and sd. *Newman et al.* [1989] followed a similar approach, and used the intercept and the slope of the OLS line given by equation (21) to estimate μ and σ from left-censored data sets.

Hashimoto and Trussell [1983], *Gilliom and Helsel* [1986], and *Helsel* [1990] used the OLS regression on the log-transformed data and extrapolated the non-detects using the regression model thus obtained. Their studies suggested that this method is fairly robust for the estimation of μ and σ using left-censored data sets with potential outliers. Let *Org* stand for the original units and *Ln* stand for the log-transformed data. Using equation (21) on the log-transformed data, the non-detects in transformed units are obtained by extrapolation corresponding to the first k normal quantiles. These non-detects can be back-transformed in the original units, and sample mean and sd then can be computed using the n data points in the original units. Alternatively, the mean, $\hat{\mu}_{Ln},$ and sd, $\hat{\sigma}_{Ln},$ are computed using the observed log-transformed data and the extrapolated non-detects (log-transformed). Assuming lognormality, *EL-Shaarawi* [1989] estimated μ and σ by back-transformation using the following equations. Note that these estimates suffer from transformation bias and neither are unbiased nor have the minimum variance [*Gilbert*, 1987].

$$\hat{\mu}_{Org} = \exp(\hat{\mu}_{Ln} + \hat{\sigma}^2_{Ln}/2), \text{ and } \hat{\sigma}^2_{Org} = \hat{\mu}^2_{Org}(\exp(\hat{\sigma}^2_{Ln}) - 1). \tag{22}$$

From the examples discussed below in Section 3, it is observed that the OLS regression approaches do not perform well in all cases. The OLS regression model on original or log-transformed data does get distorted by the outliers. This results in distorted estimates of intercept (population mean) and slope (sd), which give rise to infeasible extrapolated non-detects. For example, the estimated non-detects can become negative, larger than L, and even larger than some of the observed values (*e.g.*, $x_{(k)}$), which results in biased estimates of mean and sd (see Example 4). In these situations, typically subjective checks are provided: negative estimates can be replaced by L/2, and the estimated non-detects greater than L can be replaced by L itself. The mean and variance are then computed using the replacement values.

It is well known that the OLS estimates of intercept and slope (*Rousseeuw and Leroy* [1984]), and, hence, the mean and sd and the extrapolated non-detects, get distorted even by the presence of a single outlier. In the presence of outliers, the use of the log-transformation alone will not result in robust estimates of intercept and slope. Robust regression methods as given by *Rousseeuw and Leroy* [1984], and by *Singh and Nocerino* [1995], may be used to obtain robust estimates of slope and intercept. Some examples are considered next to illustrate the procedures described here. The discussion and use of the robust regression for censored data sets are beyond the scope of the present article. All computations are performed using the CENSOR program. In the following, all replacement values (when applicable) for non-detects are listed in parentheses. Since the substitution methods do result in biased estimates of μ and σ, their computations are omitted from most of the examples and simulation results discussed in the following sections.

Section 3

More Examples

Example 2. A simulated data set of size 15 was obtained from a normal population with mean, μ = 1.33, and sd, σ = 0.2, N(1.33, 0.2), with L=1.0, and k=2. The left-censored data are: <1.0, <1.0, 1.2883, 1.1612, 1.1560, 1.3251, 1.1568, 1.5638, 1.2914, 1.3253, 1.2884, 1.4688, 1.4581, 1.3641, 1.1342. The sample mean and sd obtained using the 13 observed data values were 1.306 and 0.134, respectively. Classical and robust procedures produced similar results and are given as follows in Table 1. For this simulated data set, observe that all of the methods, except the first two substitution methods, resulted in similar results.

Table 1. Classical/Robust results for N(1.33, 0.2) with n=15, k=2, and L=1.0.

Method	Zero	L/2	L	MLE	UMLE	RMLE	Regress*	EM**
Mean	1.13	1.20	1.27	1.25	1.26	1.26	1.27	1.25
sd	0.48	0.31	0.16	0.18	0.19	0.17	0.16	0.19
							*(0.979, 1.061)	**(0.91)

Note: Substitution values for non-detects are given in parentheses, identified by one (1) asterisk (*) for the Regress method and by two (2) asterisks (**) for the EM method.

Example 3. Next, the data set of Example 2 was contaminated with two outliers, 3.8561 and 6.2513, from a normal, N(5,2) population. Outliers distorted the classical estimates of the mean and the sd for all of the methods, and also distorted the intercept and slope of the OLS regression. The sample mean and sd for the 15 observed data points were 1.806 and 1.4, respectively. The results are given in Table 2. Notice that for the EM algorithm, the outliers distorted the conditional replacement value of 0.91 to 0.025. The robust MLE, RMLE, and the EM methods, on the other hand, resulted in fairly accurate estimates, and the robust results of Table 2 with the outliers, and the classical results of Table 1, without the outliers, are in close agreement.

Table 2. Results for N(1.33, 0.2), with outliers from N(5,2), n=17, k=2, and L=1.0.

Method	***Classical***		***Robust***	
	Mean	**sd**	**Mean**	**sd**
L	1.71	1.34	1.27	0.16
MLE	1.60	1.41	1.26	0.17
UMLE	1.61	1.49	1.26	0.18
RMLE	1.55	1.50	1.26	0.17
EM*	1.60	1.46 *(0.025)	1.25	0.18 *(0.91)

Note: Substitution values for non-detects are given in parentheses, identified by one asterisk (*) for the EM method.

Example 4. This left-censored data set is taken from the U.S. EPA RCRA guidance document [1992]. The detection limit is 1450. The data with 3 non-detects and 21 observed values are: <1450, <1450, <1450, 1850, 1760, 1710, 1575, 1475, 1780, 1790, 1780, 1790, 1800, 1800, 1840, 1820, 1860, 1780, 1760, 1800, 1900, 1770, 1790, 1780. The sample mean and sd obtained using 21 observed data are 1771.91 and 92.702, respectively. The classical and robust estimates are summarized in Table 3. Note that the substitution by L/2 method resulted in a biased estimate of mean with the highest variability, which is one of the most frequently used methods in environmental applications. All of the likelihood methods and the EM method resulted in fairly similar estimates. The regression method resulted in estimated non-detects larger than L. Figure 1 displays the classical fit obtained using the observed data, $(q_i, x_{(i)})$, i= 4, 5, ..., 21. This model is then used to estimate the non-detects by extrapolation. The graph with extrapolated non-detects is given in Figure 2. As can be seen in this figure, the three estimated non-detects (circled) are larger than the detection limit, L=1450, and even larger than the smallest observed value of 1475. The use of those extrapolated non-detects (higher than the reporting value), of course, will reduce the spread in data, but will also result in a highly biased estimate of the mean. This resulted in sd=103.21 and a high-biased estimate of the mean=1751.36. Using the slope and the intercept of the regression line, we have an estimate of sd=92.15, and an estimate of the mean=1751.36. This example alone suggests that the OLS regression approaches are not suitable for the estimation of the mean and the sd from censored samples.

Table 3. Classical/Robust results (without outliers), n=24, k=3, and L=1450.

Method	**L/2**	**L**	**MLE**	**UMLE**	**RMLE**	**Regress***	**EM****
Mean	1641.04	1731.67	1724.0	1724.94	1725.55	1751.36	1723.66
sd	364.09	138.92	153.65	159.39	144.37	103.21	157.80
						*(1571.91, 1613.25, 1637.46)	**(1385.97)

Note: Substitution values for non-detects are given in parentheses, identified by one (1) asterisk (*) for the Regress method and by two (2) asterisks (**) for the EM method.

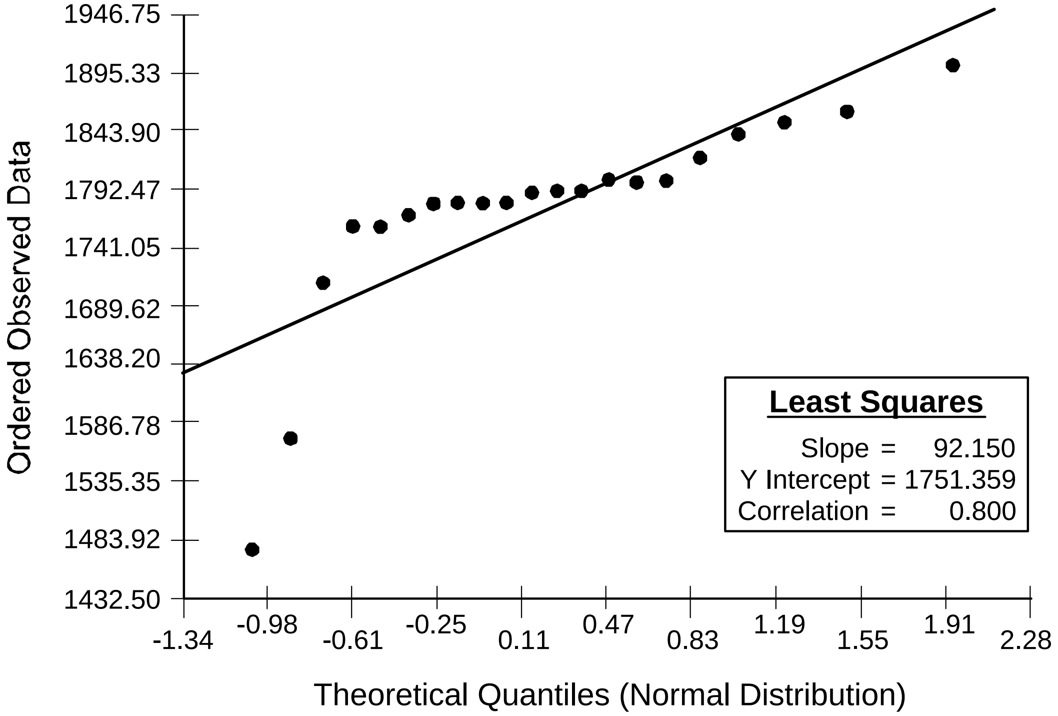

Figure 1. Classical Fit for Observed Data.

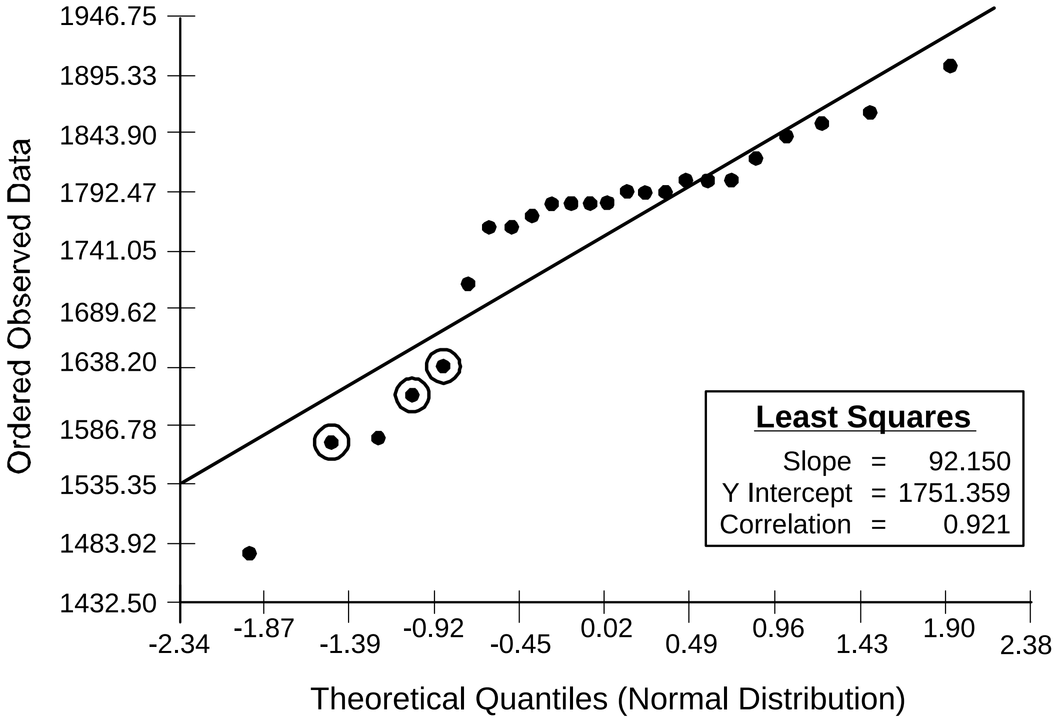

Figure 2. Non-Detects Obtained Using the Classical Fit.

In order to further illustrate how the presence of outliers distorts these estimates, three arbitrarily chosen outliers, 7000, 8000, and 11000 are added to the data set of this example. The relevant classical and robust statistics for the contaminated data set are summarized below in Table 4. The classical observed sample mean and sd for the data with outliers (24 values) are 2633.75 and 2410.35. Using equation (21), the intercept and slope of for the left-censored contaminated data set are 2216.51 and 2061.25. Use of this OLS fit resulted in distorted negative values for the extrapolated non-detects which are given in Table 4.

Table 4. Results with 3 discordant values, n=27, k=3, and L=1450.

	Classical		***Robust*** (α=0.01, 0.05)	
Method	**Mean**	**sd**	**Mean**	**sd**
MLE	2317.64	2437.60	1729.83	147.41
UMLE	2329.79	2516.74	1730.56	152.19
RMLE	2204.61	2609.06	1731.14	139.17
Regress*	2216.51	2574.10	*(-1899.83, -995.304, -469.177)	
EM**	2312.80	2491.23 **(-254.79)	1723.66	157.80 **(1385.97)

Note: Substitution values for non-detects are given in parentheses, identified by one asterisk (*) for the Regress method and by two asterisks (**) for the EM method.

For the classical EM procedure, the estimated non-detects became negative. Notice that the robust results for the MLE, the RMLE, and the EM-based algorithm are in close agreement with or without the outliers as can be seen by comparing Tables 3 and 4. Also note that the robust replacement value of 1385.97 for the EM algorithm is in agreement with the corresponding classical value without the outliers.

Next, the log-transformation of the data set with the outliers is considered. The corresponding estimates are given below in Table 5. The classical mean and sd for the observed log-transformed data are 7.675 and 0.537. In the following, all back-transformation results are obtained using equation (22). The outliers distorted the estimates of the mean and the sd for all of the methods, including the regression method.

The OLS fit on the log-transformed data is given in Figure 3. The intercept and slope are 7.577 and 0.481, respectively. Using this fit, the estimated non-detects are 6.62, 6.83, and 6.95, which, when converted back to the original units, are 749.95, 925.19, and 1043.15, respectively. The resulting estimates (using all 27 data points) of the mean and the sd in the original units are 2441.79 and 2333.93, which are obviously influenced by the three outliers. Thus, as mentioned above, the log-transformation alone cannot produce robust regression estimates. The only advantage of using the log-transformed data is that the replacement values for the non-detects did not become negative for the regression and the EM method.

Table 5. Classical Estimates for Log-Transformed Data with Outliers, n=27, and k=3.

	Log Transform		*Back Transform*	
Method	**Mean**	**sd**	**Mean**	**sd**
MLE	7.59	0.56	2314.35	1393.69
UMLE	7.59	0.57	2344.59	1456.60
RMLE	7.58	0.58	2315.67	1476.96
Regress*	7.58	0.58	2311.29	1461.96 *(6.62, 6.83, 6.95)
EM**	7.59	0.57	2327.96	1438.29 **(6.92)

Note: Substitution values for non-detects are given in parentheses, identified by one asterisk (*) for the Regress method and by two asterisks (**) for the EM method.

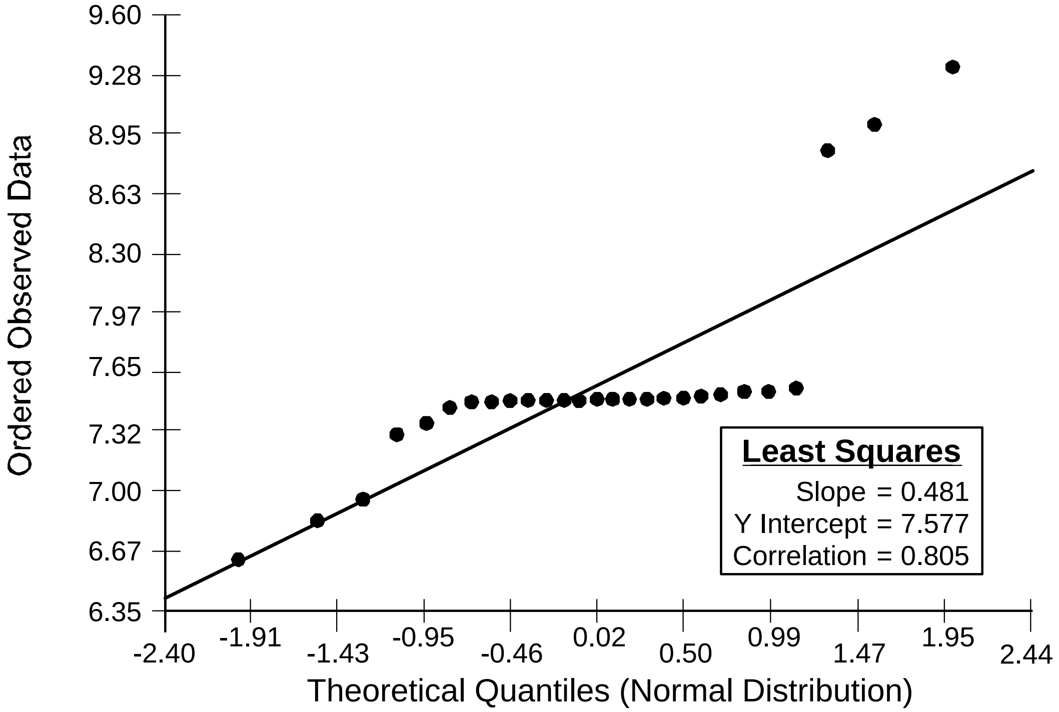

Figure 3. OLS Fit with Outliers: Log-Transformed Data.

The PROP estimates with α=0.05 on the log-transformed data are given in Table 6. The robust mean and sd for the observed data are 7.48 and 0.0086, respectively. The results summarized in Table 6 are in close agreement with the robust results (Table 4) obtained using the data in the original units and also with the estimates (Table 3) obtained using the classical procedure without the outliers.

Table 6. Robust Estimates for Log-Transformed Data with Outliers, n=27, and k=3.

	Log Transform		***Back Transform***	
Method	**Mean**	**sd**	**Mean**	**sd**
MLE	7.45	0.09	1731.10	155.89
UMLE	7.45	0.09	1732.34	161.09
RMLE	7.45	0.08	1731.72	146.78
EM*	7.45	0.10	1725.57	166.57 (7.24)

Note: Substitution values for non-detects are given in parentheses, identified by one asterisk (*) for the EM method.

Section 4

Simulation Experiments and Results

The performances of the various procedures are measured in terms of bias and MSE. Bias of an estimate, $\hat{\theta},$ of a parameter, θ , is defined as its departure, $(\hat{\theta}-\theta)$, from the parameter. For a simulation experiment with N iterations, these are given by bias $=\sum_{1}^{N} (\hat{\theta}_i-\theta)/N,$ and MSE $=\sum_{1}^{N} (\hat{\theta}_i-\theta)^2/N,$ where $\hat{\theta}_i$ is an estimate of $\theta,$ obtained from the sample generated at the ith iteration, i:=1,2, ..., N. Some results of the simulation experiments for left-censored samples without outliers obtained from the normal population, N(5,2), with 5000 iterations each, are discussed as follows. Several classical estimation methods for various values of L, sample sizes, and censoring intensities are considered. For the "fixed" detection limit case, L was set at 1.0, 2.0, 4.0, and 5.0 for each of the censoring levels. For the "computed" detection limit case, for α100% censoring intensity, the detection limit, L, is given by the equation, $L=\mu+\sigma*z_\alpha,$ where z_α , the critical value of the standard normal distribution, is given by $P(Z\le z_\alpha)=\alpha.$ Selected graphs of bias and MSE for some values of L (*e.g.*, 2 and 4), sample sizes (*e.g.*, 5 - 25), and censoring intensities (*e.g.*, 10% - 60%) are presented here. Since the various substitution methods do not perform well, therefore, most of the results and graphs discussed are for the MLE, UMLE, RMLE, EM, and the regression methods. From these graphs (Figures 4-27), as expected, it is observed that the MSE and bias for all of the methods decrease with the sample size. Also, it is noticed that the differences in MSEs of the various likelihood procedures decrease as the sample size increases (*e.g.*, Figures 4-6 and 9-11).

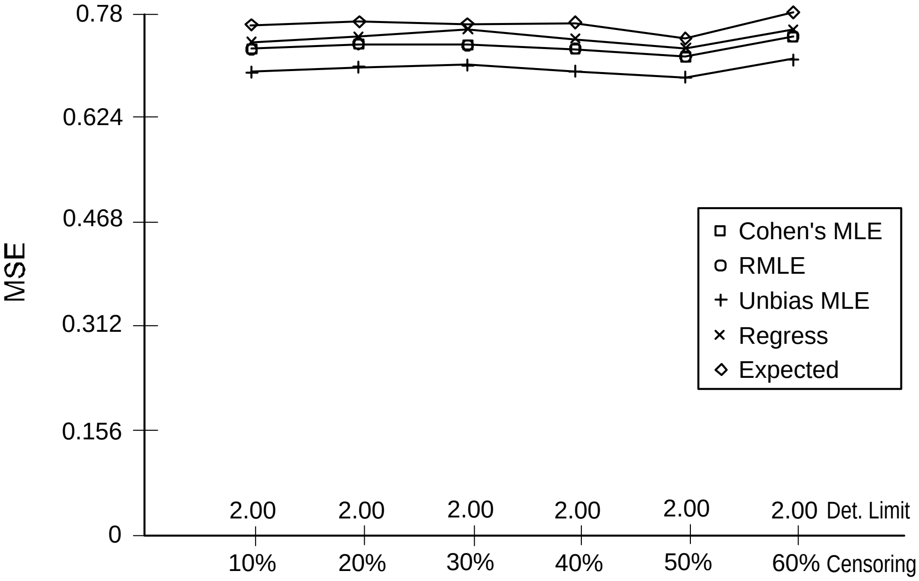

Figure 4. MSE: 5000 Simulation Runs For N(5,2), L = 2.0 (n = 5).

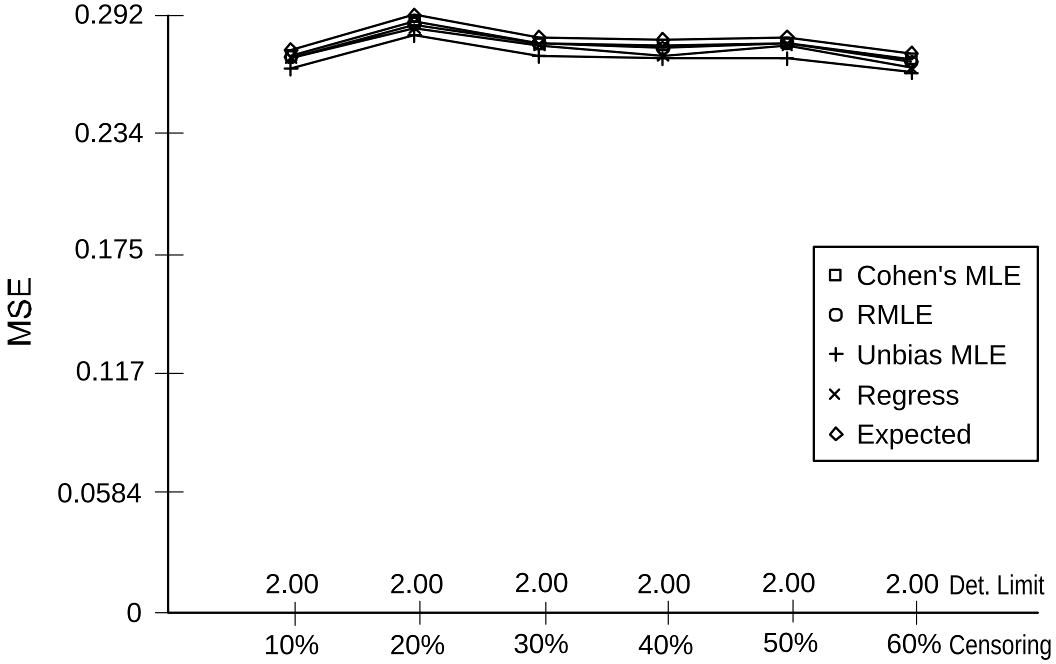

Figure 5. MSE: 5000 Simulation Runs For N(5,2), L = 2.0 (n = 15).

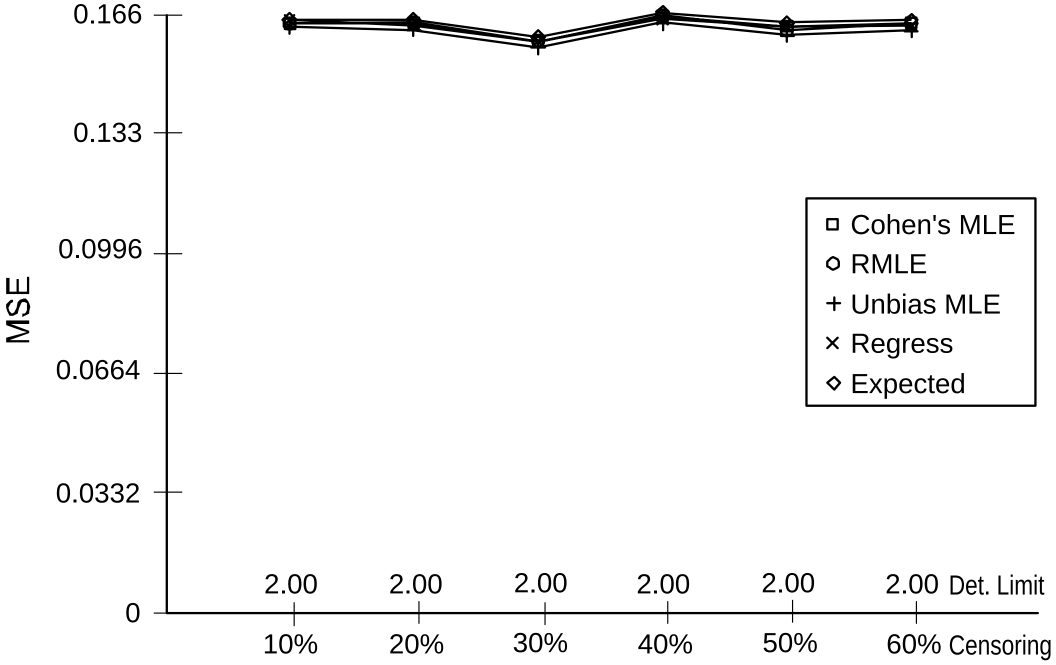

Figure 6. MSE: 5000 Simulation Runs For N(5,2), L = 2.0 (n = 25).

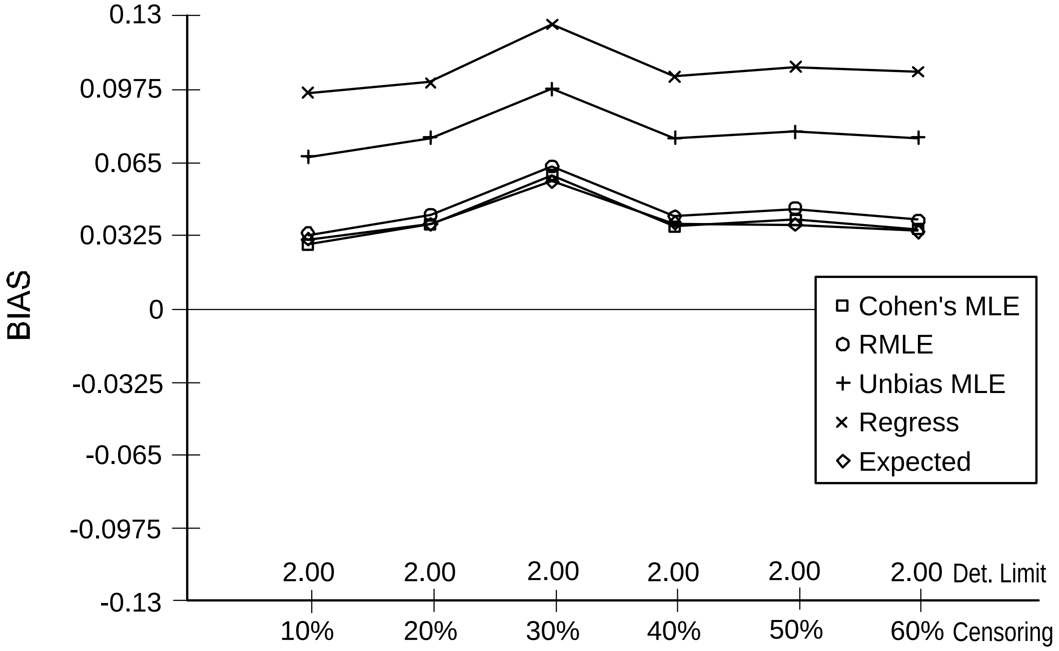

Figure 7. Bias: 5000 Simulation Runs For N(5,2), L = 2.0 (n = 5).

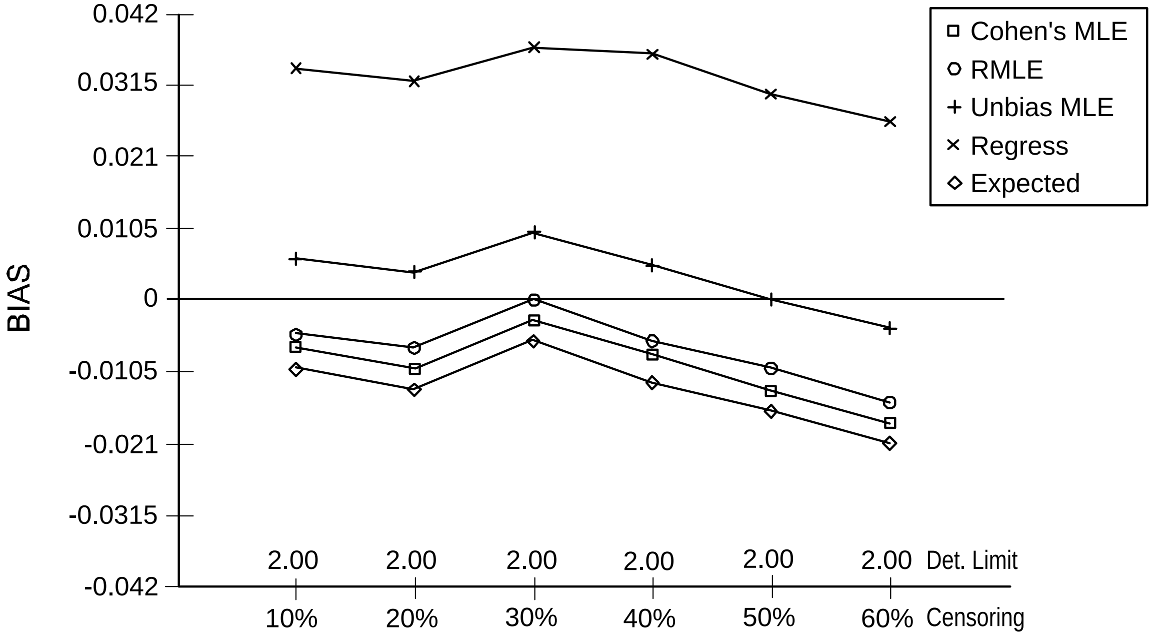

Figure 8. Bias: 5000 Simulation Runs For N(5,2), L = 2.0 (n = 15).

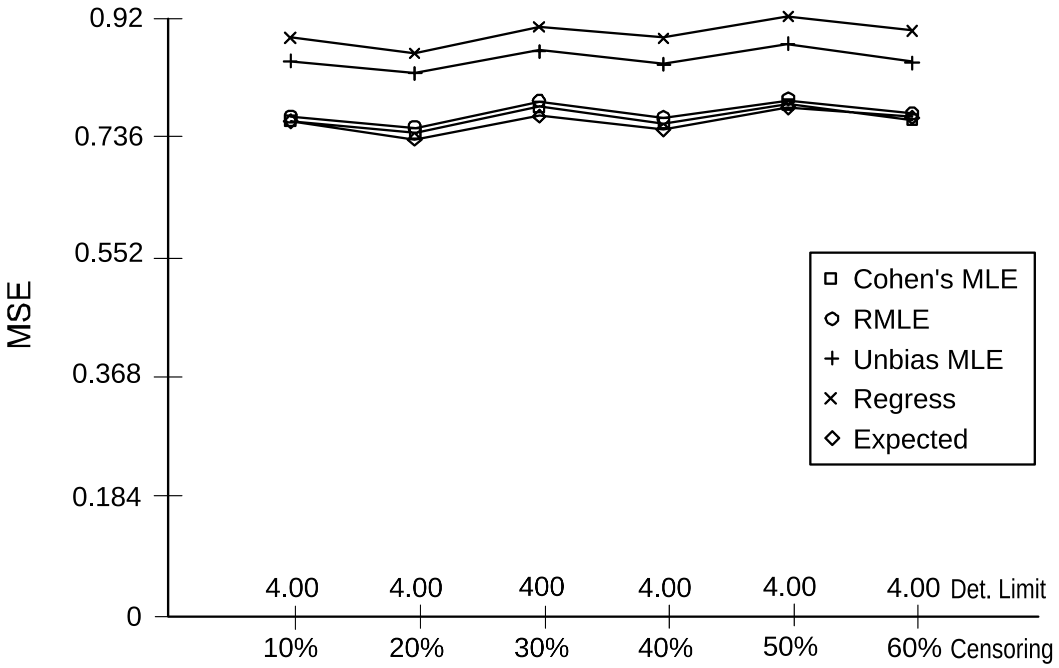

Figure 9. MSE: 5000 Simulation Runs For N(5,2), L = 4.0 (n = 5).

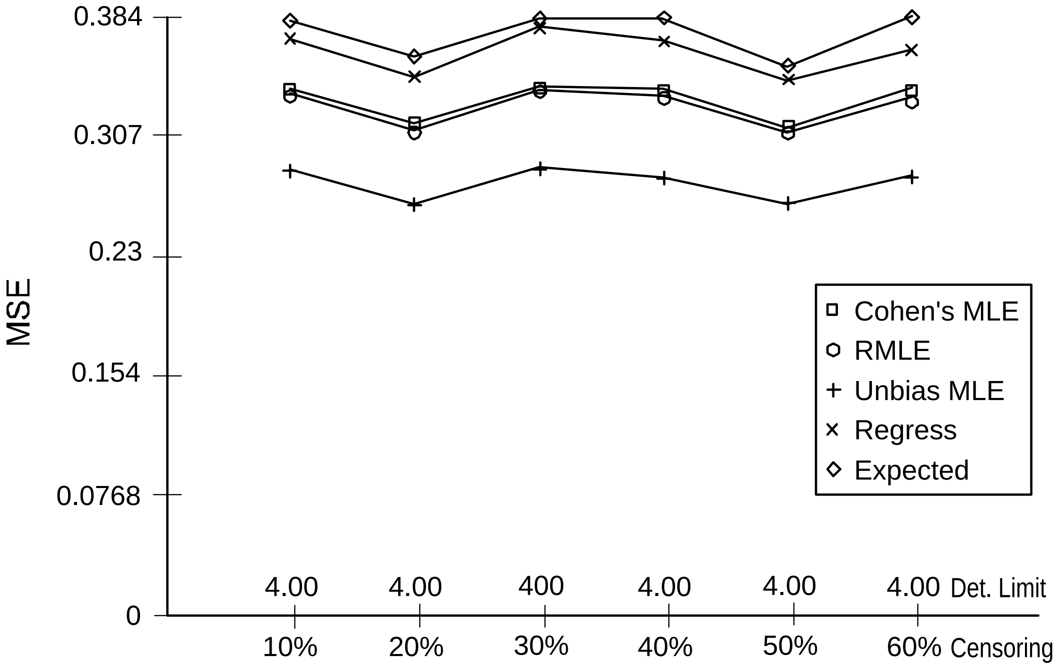

Figure 10. MSE: 5000 Simulation Runs For N(5,2), L = 4.0 (n = 15).

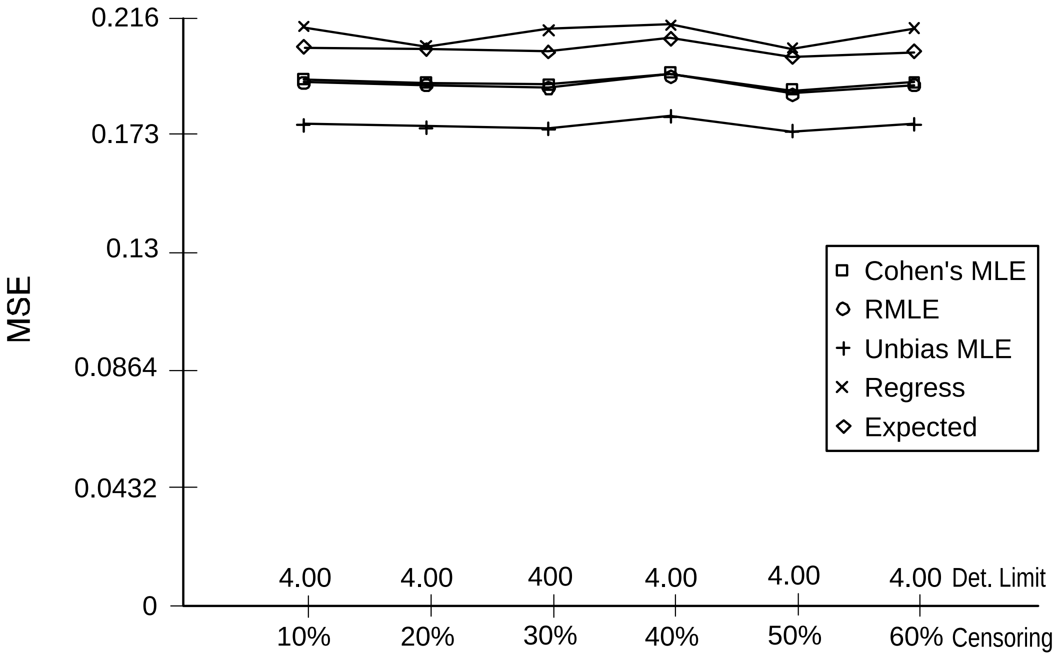

Figure 11. MSE: 5000 Simulation Runs For N(5,2), L = 4.0 (n = 25).

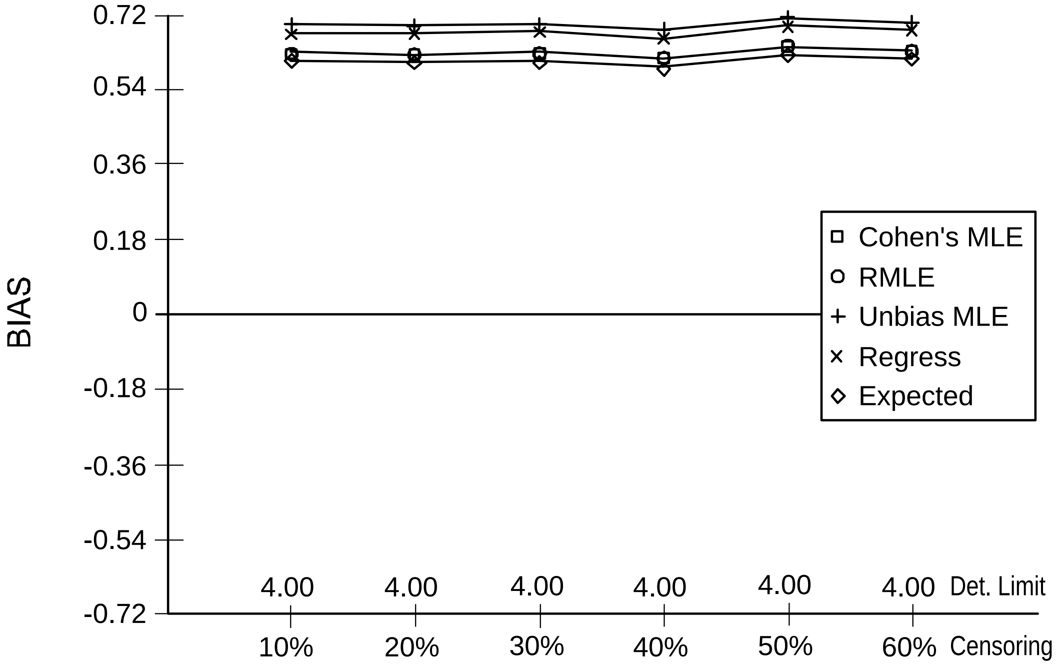

Figure 12. Bias: 5000 Simulation Runs For N(5,2), L = 4.0 (n = 5).

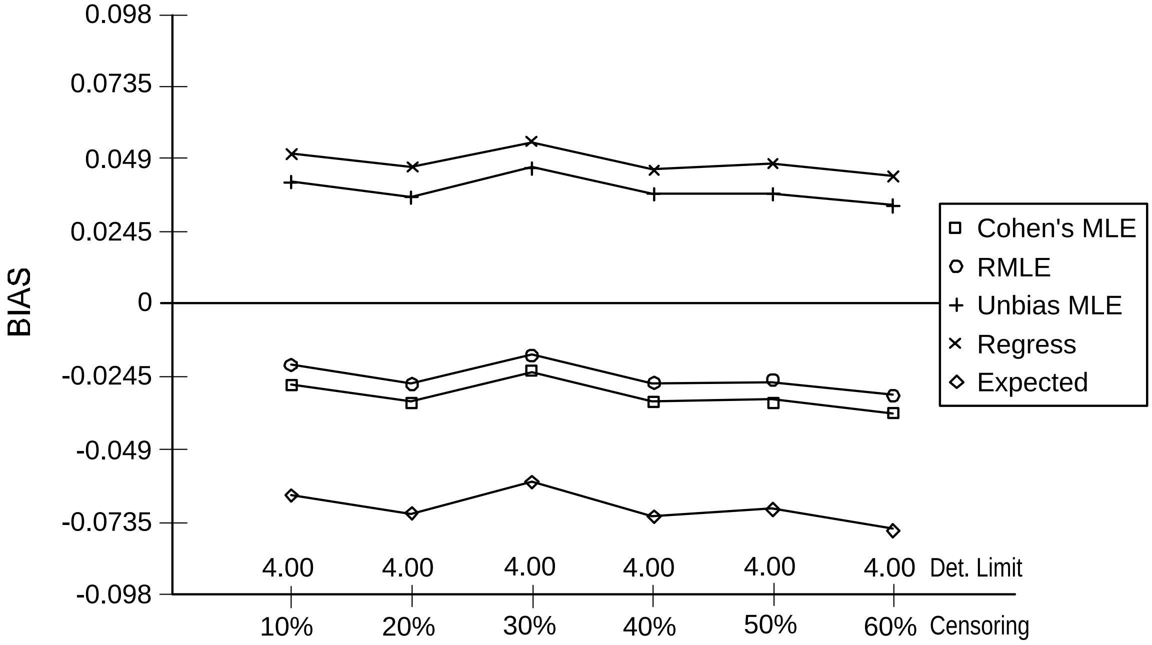

Figure 13. Bias: 5000 Simulation Runs For N(5,2), L = 4.0 (n = 15).

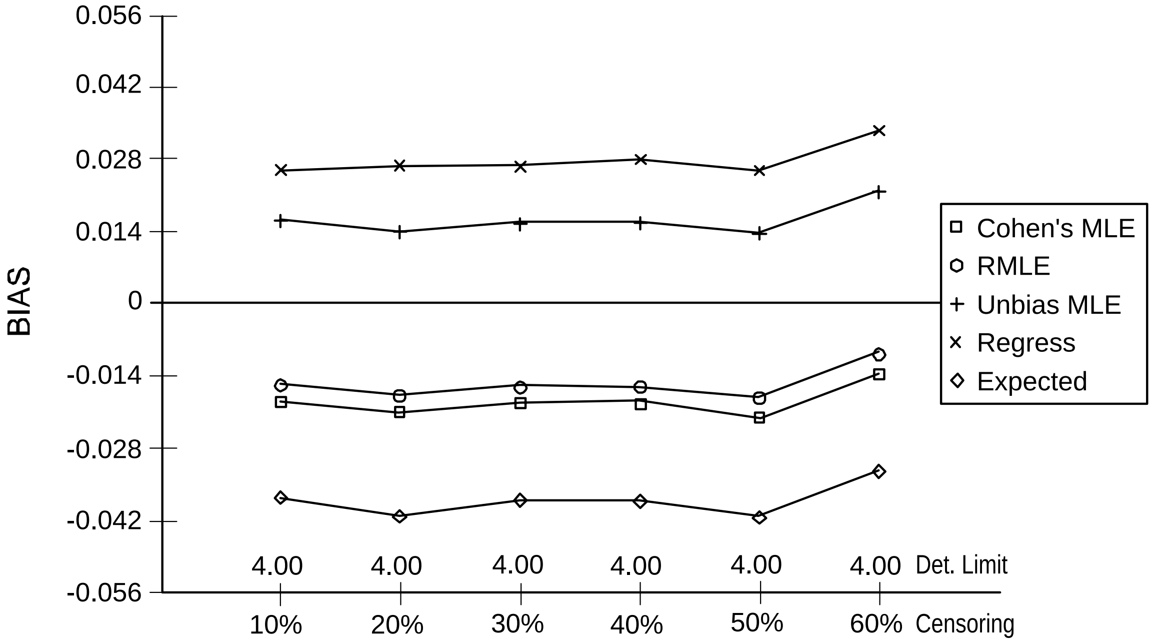

Figure 14. Bias: 5000 Simulation Runs For N(5,2), L = 4.0 (n = 25).

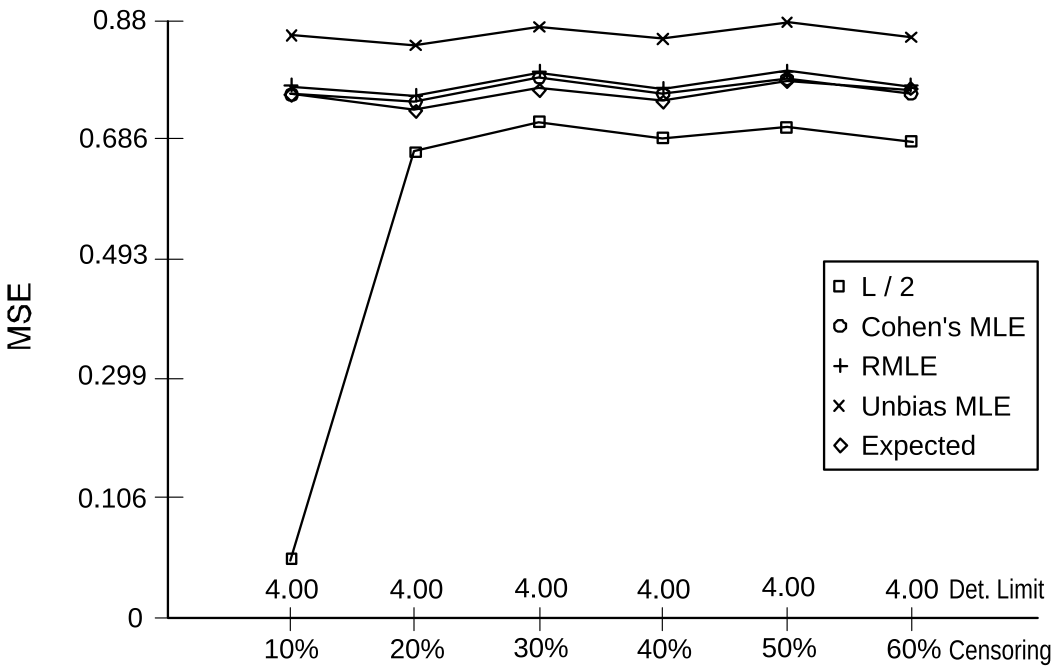

Figure 15. MSE: 5000 Simulation Runs For N(5,2), L = 4.0 (n = 5).

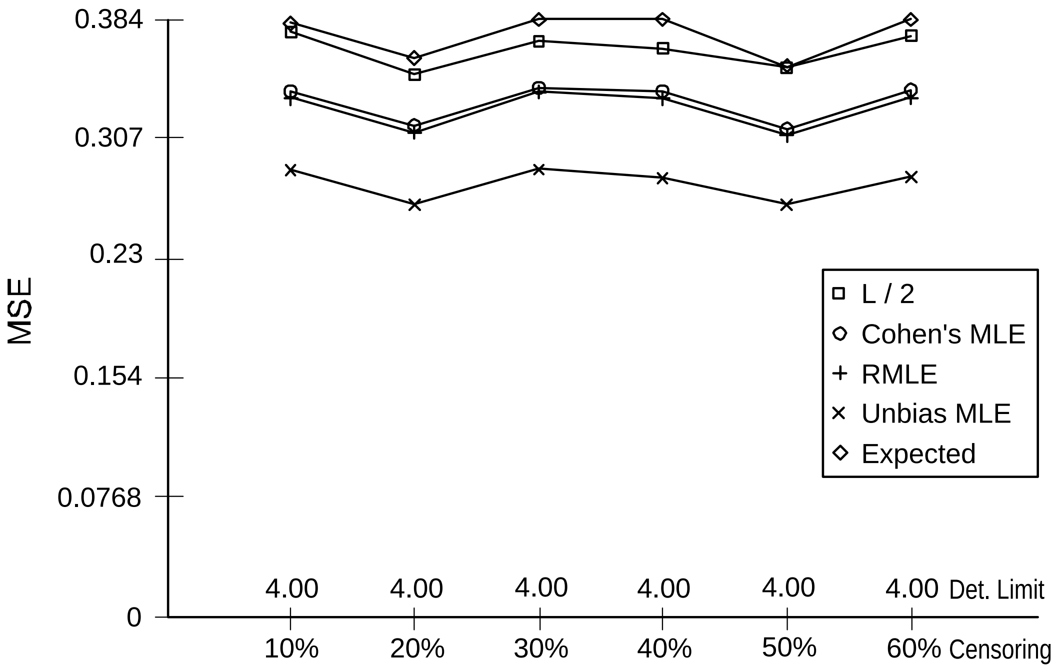

Figure 16. MSE: 5000 Simulation Runs For N(5,2), L = 4.0 (n = 15).

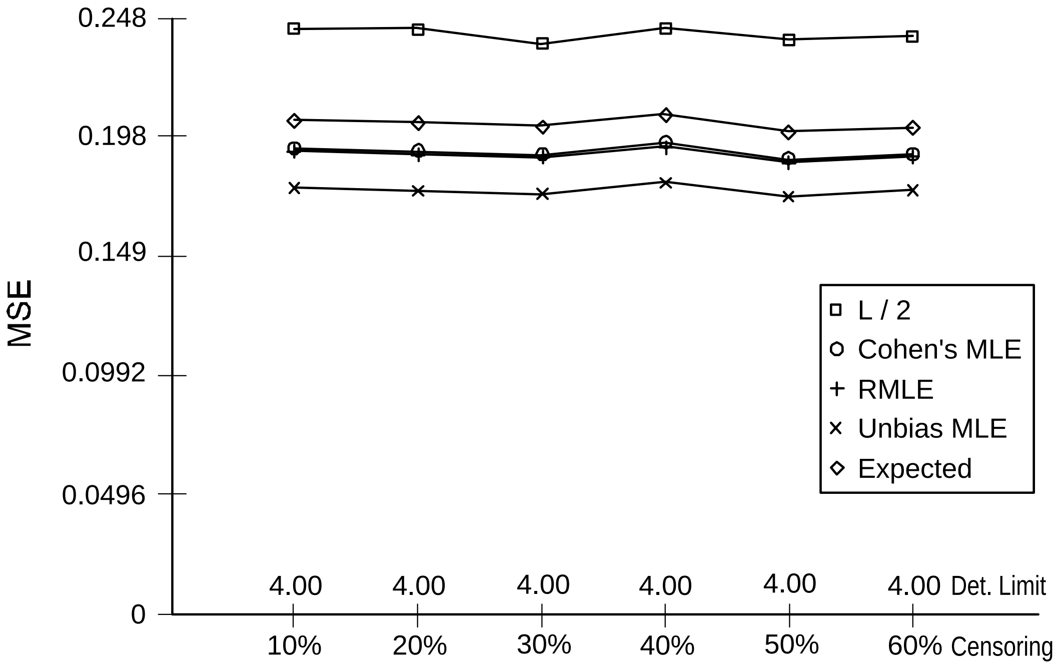

Figure 17. MSE: 5000 Simulation Runs For N(5,2), L = 4.0 (n = 25).

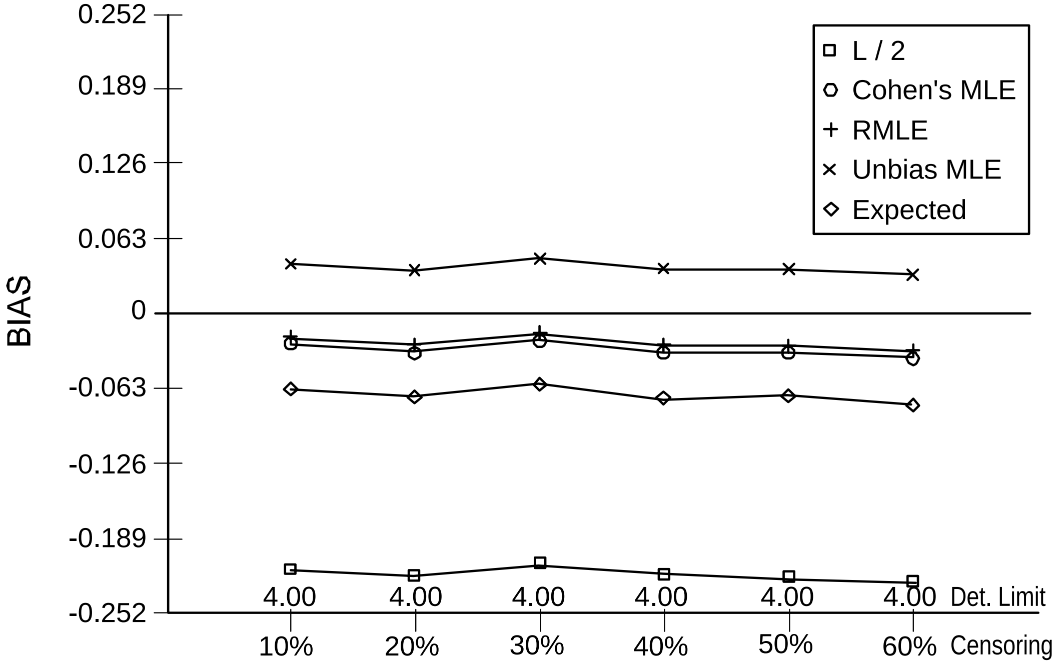

Figure 18. Bias: 5000 Simulation Runs For N(5,2), L = 4.0 (n = 15).

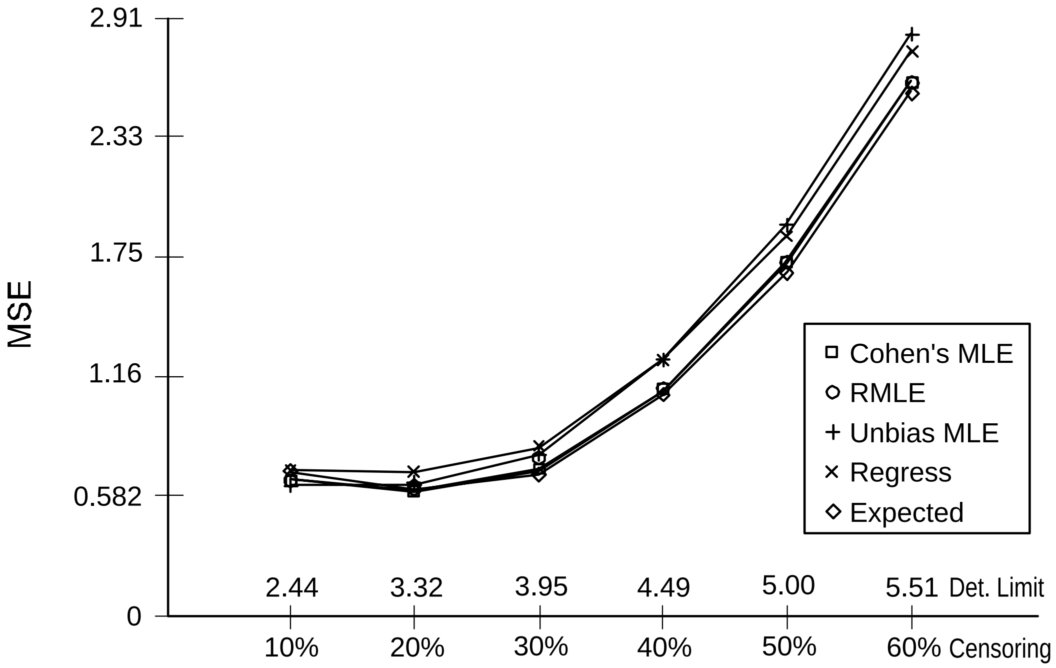

Figure 19. MSE: 5000 Simulation Runs For N(5,2), L Computed (n = 5).

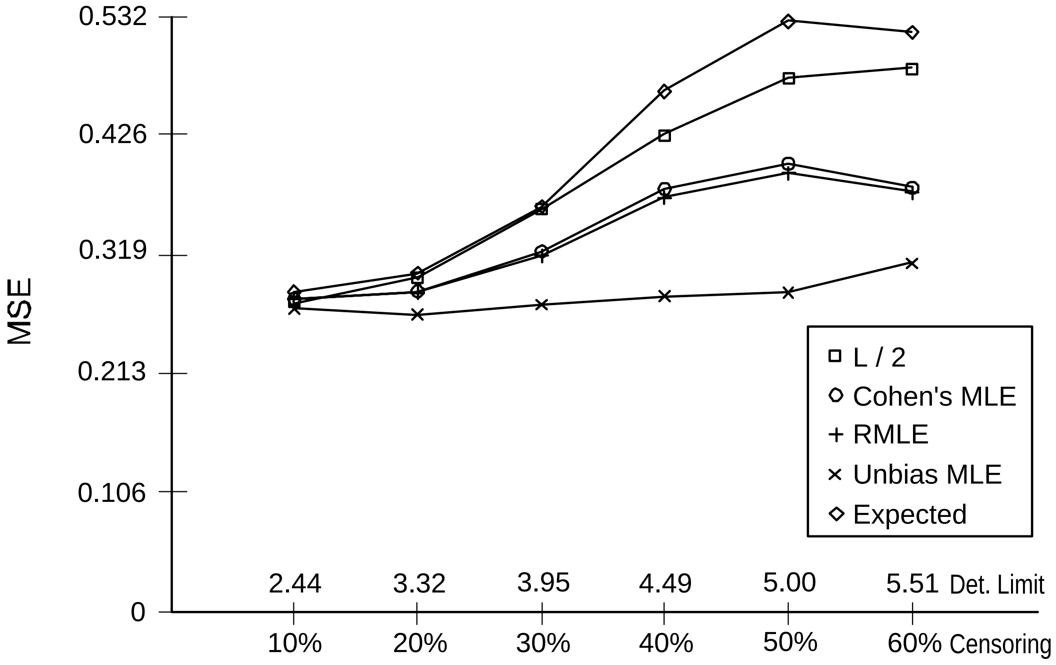

Figure 20. MSE: 5000 Simulation Runs For N(5,2), L Computed (n = 15).

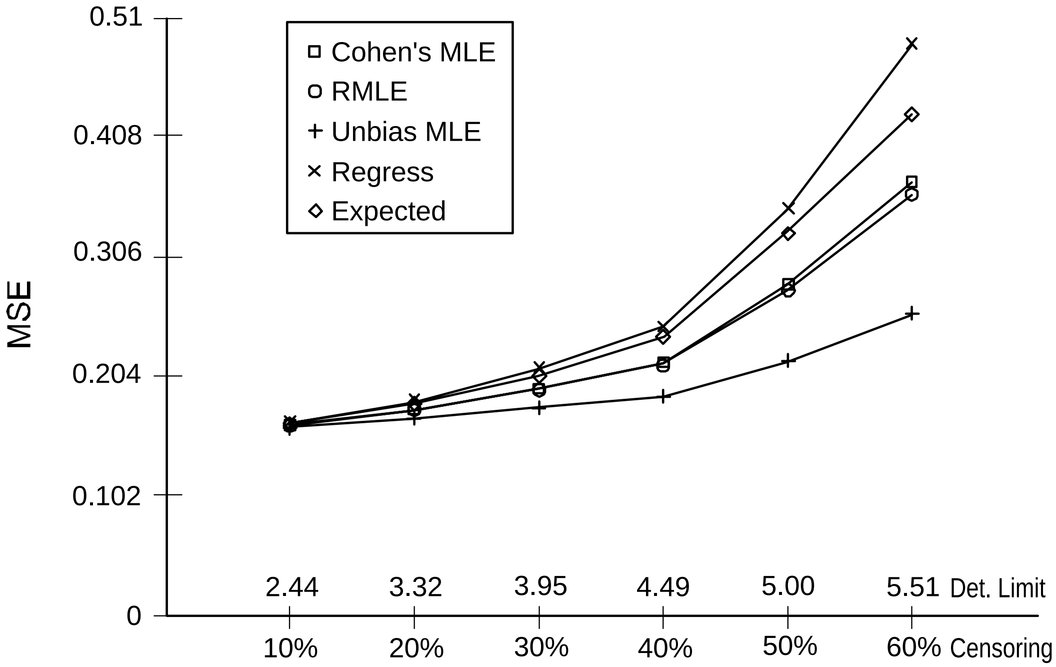

Figure 21. MSE: 5000 Simulation Runs For N(5,2), L Computed (n = 25).

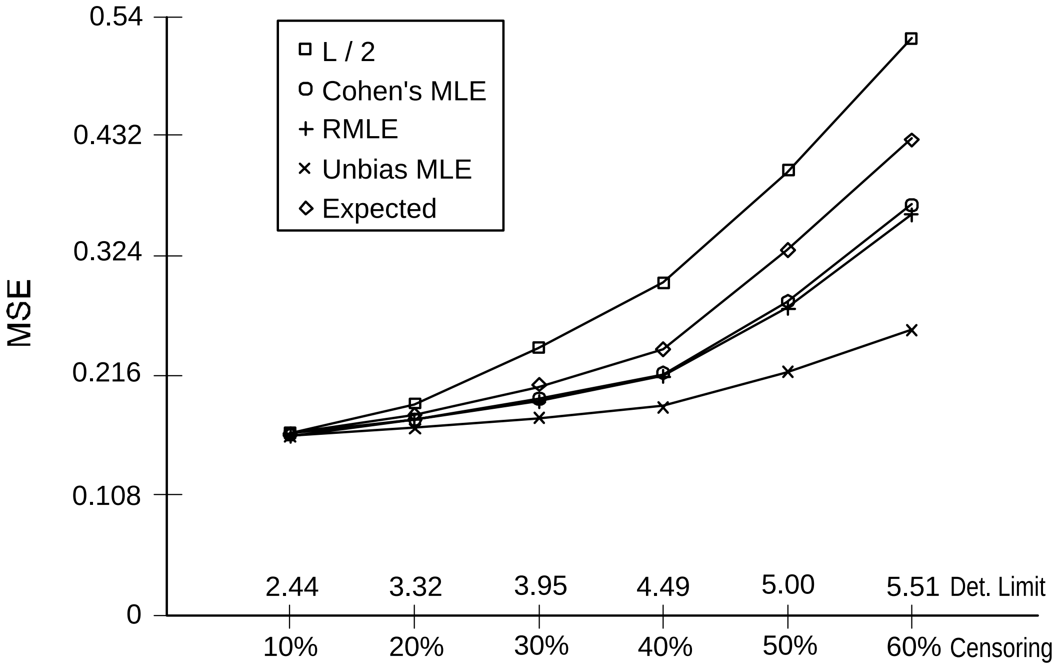

Figure 22. MSE: 5000 Simulation Runs For N(5,2), L Computed (n = 25).

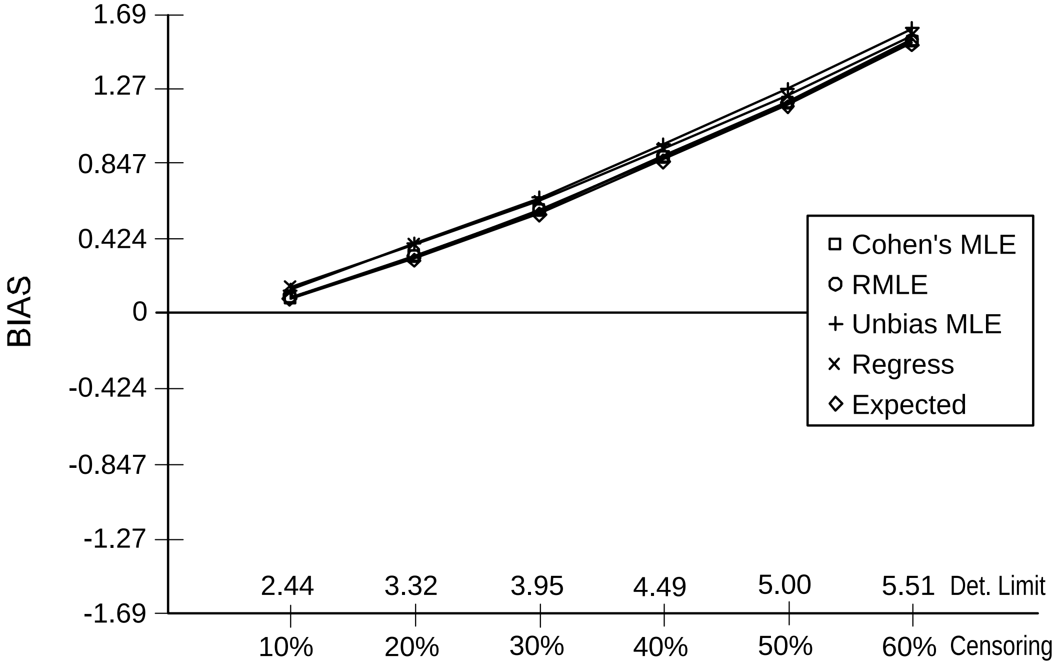

Figure 23. Bias: 5000 Simulation Runs For N(5,2), L Computed (n = 5).

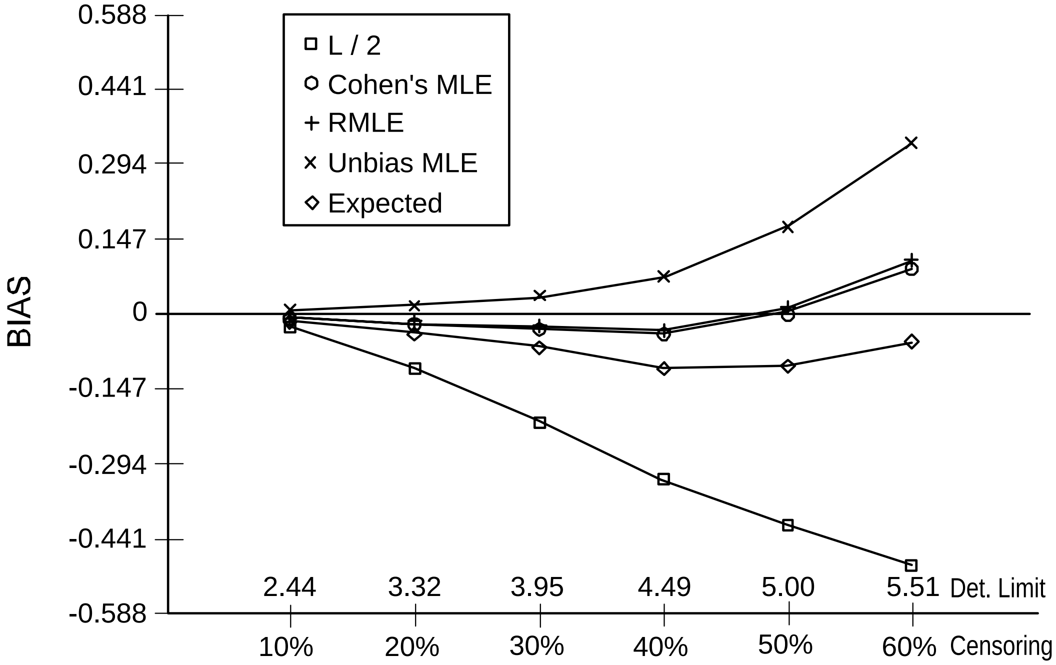

Figure 24. Bias: 5000 Simulation Runs For N(5,2), L Computed (n = 15).

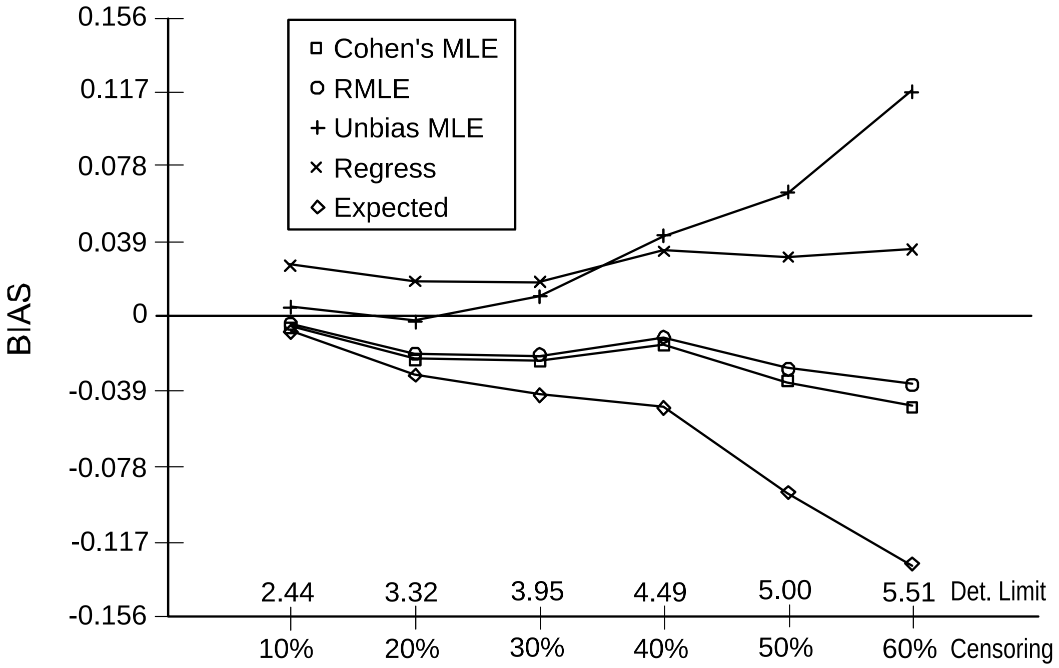

Figure 25. Bias: 5000 Simulation Runs For N(5,2), L Computed (n = 25).

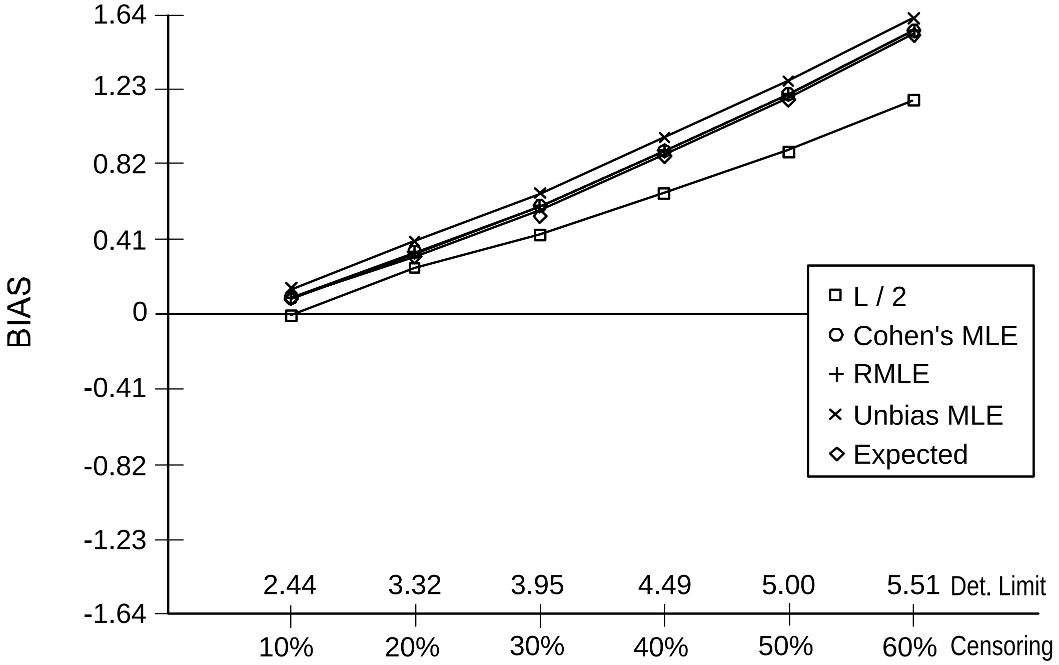

Figure 26. Bias: 5000 Simulation Runs For N(5,2), L Computed (n = 5).

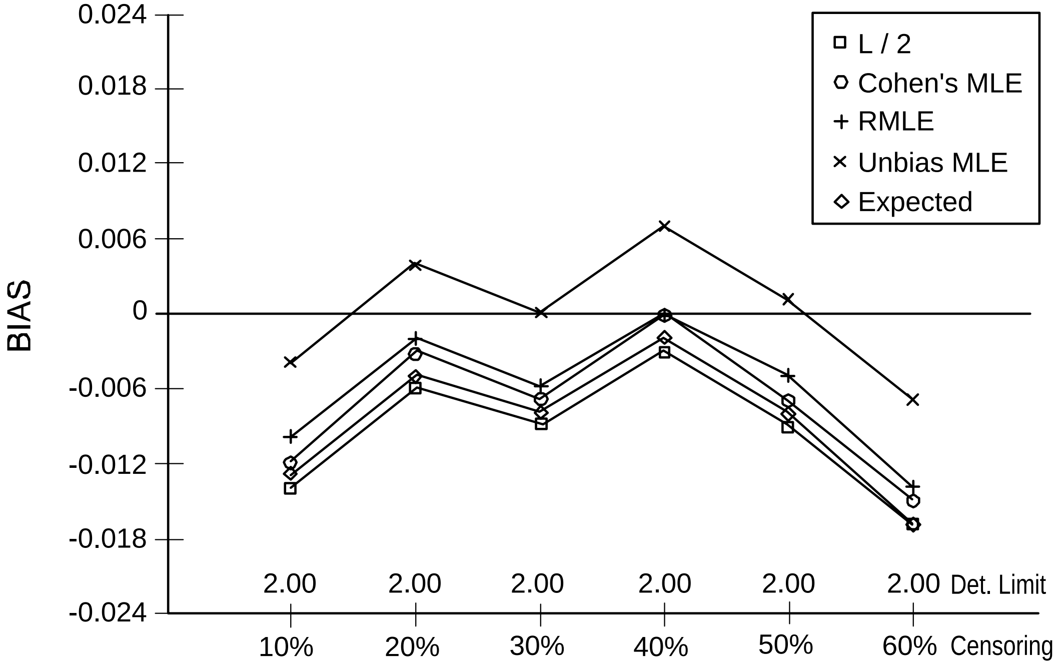

Figure 27. Bias: 5000 Simulation Runs For N(5,2), L = 2.0 (n = 25).

Fixed Detection Limit

Figures 4 through 6 have the MSEs and Figures 7 and 8 have the bias for various procedures when L=2.0. Figures 9 through 11 show the MSEs and Figures 12-14 display the bias for various estimation methods when L=4.0. It is observed that, for samples of smaller (*e.g.*, less than 10) sizes, the UMLE method yields a higher bias than the MLE, RMLE, and the EM methods (Figures 7 and 12). However, when L (*e.g.*, 1, 2) is much smaller than the mean, μ (*e.g.*, 5), the UMLE method has the smallest MSE and the EM method has the largest MSE for samples of all sizes. When L is closer (*e.g.*, 4 or 5) to mean, μ, for samples of smaller sizes (Figures 9 and 12), the MSE and bias for the UMLE become larger than those of the EM, MLE, and the RMLE methods, with the EM method having the smallest MSE and bias. This observation concurs with Gleit's (1985) findings for the EM method. But as the sample size increases (*e.g.*, becomes 15 or larger), as expected, the situation reverses, and the UMLE method results in the smallest MSE, while the EM and the regression methods yield larger MSEs (Figures 10 and 11). Note that, to some extent, this behavior of the MSE and bias of the EM method is similar to the substitution by L/2 or L methods, except that, as the sample size increases, the MSE and bias for the later two substitution methods become much greater than those of the EM method, as can be seen in Figures 15-18 and 24. The EM method, after all, is just a substitution method in which all of the non-detects are replaced by an optimally obtained conditional expected value. From Figures 13 and 14, it is noticed that the bias for the EM and the regression methods becomes fairly large as the sample size increases.

Also, it is observed that as the sample size increases, the bias of the MLE and RMLE methods starts becoming smaller (in magnitude) than that of the UMLE method for all censoring intensities (Figures 13-14). From all of these graphs, it is observed that the bias and MSEs obtained using the RMLE and Cohen's MLE methods are stable, always stay very close to each other, and lie in the middle of the respective bias and MSE of the other methods for all sample sizes and censoring levels. Actually, in most cases, the RMLE method even results in a smaller bias and MSE than the MLE method as can be seen from Figures 8, 10, 11, 13, 14, and 16-18. Also, note that the differences in the MSE of the three MLE methods (UMLE, Cohen's, and RMLE) decrease as the sample size increases.

Computed Detection Limit

Figures 19-22 are the graphs of the MSE and Figures 23-27 have the bias for various sample sizes and censoring levels. From these graphs, as expected, it is observed that the MSE for all of the methods increase with the detection limit, L, or the censoring intensity, for all sample sizes. It is observed that for small sample sizes, the EM method (the optimal replacement by the conditional expected value method) results in smaller MSE and bias (Figures 19 and 23), especially when the detection limit starts coming closer to the mean value. The bias for the EM and L/2 methods becomes unacceptably high with increased sample size, as can be seen in Figures 24 and 25. As observed earlier, note that, as the sample size increases, the UMLE method results in the smallest MSE, and the substitution by L/2 method, the EM method, and the regression method yield MSEs much larger than the three MLE methods. However, the UMLE method results in a bias which is larger than those of the MLE and RMLE methods. This increase in the bias of the UMLE becomes quite noticeable with increases in the sample size, the detection limit, and the censoring intensity. This is especially true when the censoring level starts exceeding 30%. Moreover, from all of these graphs, it is observed that both the bias and the MSEs obtained using the RMLE and Cohen's MLE methods are stable and always stay close together for all of the sample sizes and censoring levels. Also, as noticed earlier, the RMLE method does result in a smaller bias and MSE than the MLE method (Figures 20-25) most of the time. These observations concur with the conclusions derived by *Haas and Scheff* (1990).

Section 5

Summary and Conclusions

In this article, two questions which arise when dealing with left-censored data sets have been addressed. Those two questions are: 1) "Which method should be used for the estimation of the population mean and sd from left-censored data sets?" and 2) "What is an appropriate robust estimation procedure in the presence of potential outliers in the right tail of the distribution?"

The various substitution methods are simple, but do not perform well in most cases as they yield estimates with a larger bias and MSE than those obtained using the MLE methods. Also, it is observed that for larger sample sizes (*e.g.*, >=15), the EM method results in a bias and MSE larger than those of the MLE methods. The examples presented here lead to the conclusion that the OLS regression-based approaches cannot be recommended for routine use. The estimated non-detects obtained by extrapolating the fitted model many times result in infeasible estimates, which become negative or even greater than the detection limit, L. The examples and the simulation results presented in this article clearly establish that, in most cases, the three MLE methods (Cohen's MLE, UMLE, and RMLE) perform better than the various substitution and regression methods.

All of the classical estimation procedures, including the maximum likelihood and substitution methods, result in distorted estimates in the presence of outliers. In the presence of outliers, the EM method sometimes produces negative estimates of the non-detects, which in turn result in a biased estimate of the population mean. The OLS regression models get distorted by the outlying observations; therefore, regression estimates obtained using raw or log-transformed data are no longer reliable. Thus, the OLS regression method based on the log-transformed data is not a "true robust" method. It is observed that the robust estimation procedure based on the PROP influence function results in stable and reliable estimates of the population parameters. Moreover, the resulting robust estimates, with or without the outliers, and the classical estimates, without the outliers, stay in close agreement.

The performance of the various estimation methods described here depend upon several things, such as the sample size, the censoring intensity, and the value of the detection limit, L. The conclusions derived from the simulation results and graphs presented in this article are summarized as follows.

- When the detection limit, L, is closer to the population mean, it is observed that for samples of smaller sizes (*e.g.*, 5-10), the EM method and the other substitution methods such as the L/2 method result in a smaller bias and MSE than the three MLE (UMLE, Cohen's, and RMLE) procedures. However, as the sample size increases (*e.g.*, 15 or larger), the EM method, along with the other substitution methods, results in a higher bias and a larger MSE than the three likelihood procedures.

- For values of L much smaller than the mean, the UMLE method results in the smallest MSE for samples of all sizes.

- The differences in the MSE of the three MLE methods decrease as the sample size increases.

- The simulation results suggest that, although the UMLE method does result in the smallest MSE for samples of size 15 or larger, the bias of the UMLE becomes larger (in magnitude) than the MLE and the RMLE methods. This increase in the bias of the UMLE method becomes quite noticeable as the detection limit increases and the censoring intensity starts exceeding 30%. Thus, for higher censoring intensities, the MLE or the RMLE method may be used to obtain estimates of the population mean and sd from a left-censored data set.

- The RMLE method is simple and results in estimates which are in close agreement with the Cohen's ML estimates. It is observed that the bias and MSEs obtained using the RMLE and Cohen's MLE methods are stable and always stay close together for all sample sizes and censoring intensities. Actually, in most cases, the RMLE method results in a smaller bias and MSE than those obtained using the Cohen's MLE method. This is especially true as the sample size increases.

Using the examples and results described here, the following recommendations can be made:

- For data sets with potential outliers, the robust estimation procedures based on influence functions, such as the PROP influence function, should be used for the estimation of population parameters.

- For samples of small sizes (*e.g.*, 10 observations or less), the EM method or the substitution by L/2 (or L) method may be used, especially when L is closer to the mean.

- For samples of larger sizes (*e.g.*, 15 observations or larger), the UMLE method may be used for censoring levels of 30% or less.

- However, since the differences in the MSE of the three MLE methods (UMLE, Cohen's, and RMLE) decrease as the sample size increases, and in order to make things easier for a typical user, it is recommended that for larger sample sizes, or for samples with censoring levels exceeding 30%, the much simplified RMLE method may be used for the estimation of the population parameters.

- The results of our study clearly establish that one should stay away from the substitution methods, especially when the sample size is larger than 10 observations.

References

Box, G.E.P., and Cox, D.R., "An analysis of transformation," *Journal of Royal statistical Society*, Ser. B, 39, pp. 211-252, 1964.

Barnett, V., "Convenient probability plotting positions for the normal distribution," *Appl. Statist.*, 25, No. 1, pp. 47-50, 1976.

Cohen, A.C., Jr., "Estimating the mean and variance of normal populations from singly truncated and double truncated samples," *Ann. Math. Statist.*, Vol. 21, pp. 557-569, 1950.

Cohen, A.C., Jr., "Simplified estimators for the normal distribution when samples are singly censored or truncated," *Technometrics,* Vol. 1, No. 3, pp. 217-237, 1959.

Cohen, A.C., Jr., "Truncated and censored samples," 119, Marcel Dekker Inc., New York, NY 1991.

Dempster, A.P., Laird, N.M., and Rubin, D.B., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society,* Ser. B, 39, pp. 1-38, 1977.

El-Shaarawi, A.H., "Inferences about the mean from censored water quality data," *Water Resources Research*, 25, pp. 685-690, 1989.

Gilbert, R.O., "Statistical Methods for Environmental Pollution Monitoring," Van Nostrand Reinhold, New York, 1987.

Gilliom, R.J., and Helsel, D.R., "Estimation of distributional parameters for censored trace level water quality data. 1. Estimation Techniques," *Water Resources Research*, 22, pp. 135-146, 1986.

Gleit, A., "Estimation for small normal data sets with detection limits," *Environmental Science and Technology*, 19, pp. 1206-1213, 1985.

Haas, C.H., and Scheff, P.A., "Estimation of averages in truncated samples," *Environmental Science and Technology*, 24, pp. 912-919, 1990.

Hampel, F.R., "The Influence Curve and Its Role in Robust Estimation," *Journal of American Statistical Association*, 69, pp. 383-393, 1974.

Hashimoto, L.K., and Trussell, R.R., "Evaluating water quality data near the detection limit." Presented at the Advanced Technology Conference, *Am. Water Works Assoc.,* Las Vegas, NV, June 5-9, 1983.

Helsel, D.R., "Less than obvious, Statistical treatment of data below the detection limit," *ES&T Features Environmental Sci. Technol.,* Vol. 24, No. 12, pp. 1767-1774, 1990.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W., "Understanding Robust and Exploratory Data Analysis," John Wiley, New York, 1983.

Huber, P.J., "Robust Statistics," John Wiley, New York, 1981.

Johnson, R.A., and Wichern, D.W., "Applied multivariate statistical analysis," Prentice Hall, 1988.

Lechner, J.A., "Estimators for Type-II Censored (Log) Normal Samples," *IEEE Transactions on Reliability*, Vol. 40, No. 5, pp. 547-552, 1991.

Newman, M.C., Dixon, P.M., and Pinder, J.E., "Estimating mean and variance for environmental samples with below detection limit observations," *Water Resources Bulletin*, Vol. 25, No. 4, pp. 905-916, 1990.

Persson, T., and Rootzen, H., "Simple and highly efficient estimators for a Type I censored normal sample," *Biometrika*, 64, pp. 123-128, 1977.

Practical Methods for Data Analysis, Guidance for Data Quality Assessment, EPA QA/G-9, QA96 Version, 1996.

Saw, J.G., "The bias for the maximum likelihood estimates of location and scale parameters given a Type II censored normal sample," *Biometrika*, 48, pp. 448-451, 1961b.

Schneider, H., "Truncated and censored samples from normal populations," Vol. 70, Marcel Dekker Inc., New York, 1986.

“Scout: A Data Analysis Program” (revised March 1999), Technology Support Bulletin, U.S. EPA, ORD, NERL, ESD-LV, Las Vegas, NV 89193-3478. CENSOR was developed to be a module that will be incorporated into the next version of Scout.

Shumway, A.H., Azari, A.S., Johnson, P., "Estimating mean concentrations under transformation for environmental data with detection limits," *Technometrics*, Vol. 31, No. 3, pp. 347-356, 1989.

Singh, A., "Omnibus robust procedures for assessment of multivariate normality and detection of multivariate outliers," *Multivariate Environmental Statistics*, Patil, G.P. and Rao, C.R., Editors, pp. 445-488, Elsevier Science Publishers, 1993.

Singh, A. and Nocerino, J.M., "Robust procedures for the Identification of Multiple Outliers," Handbook of Environmental Chemistry, Statistical Methods, Vol. 2, Part G, 229-277, Springer, Germany, 1995.

Singh, A.K., Singh, A., and Engelhardt, M. (1997). The Lognormal Distribution in Environmental Applications. Technology Support Center Issue, 182CMB97. EPA/600/R-97/006.

Singh, A.K., Singh, A., and Engelhardt, M. (2000). Some Practical Aspects of Sample Size and Power Computations for Estimating the Mean of Positively Skewed Distributions in Environmental Applications. EPA/600/S-99/006.

Staudte, R.G. and Sheather, S.J., "Robust Estimation and Testing," John Wiley, 1990.

Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities, Addendum to Interim Final Guidance, Office of Solid Waste, Waste Management Division. U.S. EPA, 1992.