Comparison of Statistical Methods In Handling Minimum Detection Limits

Pi-Erh Lin and Xu-Feng Niu

Department of Statistics

Florida State University

Tallahassee, Florida 32306–4330

July 1998

Technical Report (Draft) Submitted to

the Florida Department of Environmental Protection

Contents

Comparison of Statistical Methods In Handling Minimum Detection Limits

1 Introduction

Monitoring and analyzing phosphorus levels of ecosystems is an important topic in Florida environmental study. The Florida Department of Environmental Protection (FDEP) is currently gathering data from different agencies for the purpose of evaluating the health of aquatic systems around the state.

For samples of low-level phosphorus concentrations, a testing laboratory would reported its results as "below the detection limit" (BDL) in the cases where the concentrations were not detected at or below the minimum detection limit (MDL). For example, the South Florida Water Management District has been collecting atmospheric deposition data from 19 monitoring sites in weekly intervals since the early 1980's. The MDL for total phosphorus (TP) concentration used by the agency is 3.5 $\mu g/L$. In this study, measurements below the detection limit in a data set will be called the BDL portion and those above the detection limit (ADL) will be called the ADL portion of the sample.

Statistical methods for computing summary statistics of data with BDL values have been proposed by many researchers. They include simple substitution methods, maximum likelihood methods based on distribution, and robust regression methods. Helsel and Gilliom (1986) and Gleit (1985) compared the performance of several estimating methods based on thousands of simulated data sets. Helsel and Gilliom (1986) also applied some of these methods to analyze water-quality data. The simple substitution methods, maximum likelihood methods, and robust regression methods are summarized in Helsel and Hirsch (1992, Chapter 13).

Most recently, Ahn (1998) studied and compared several maximum likelihood methods and a regression method for estimating population parameters based on two wet total phosphorus (TP) data sets with BDL values. He concluded that the one-step restricted maximum likelihood method gave more accurate estimates for the wet TP data than other methods he studied. Ahn (1998) also proposed a method to estimate population parameters by combining the estimates obtained separately from both BDL and ADL portions. He showed that the proposed method improved over the conventional maximum likelihood estimates for the two data sets under his consideration.

In this study, four statistical methods of handling the MDLs will be compared based on simulations. The simulation study is essentially similar to the study performed by Helsel and Gilliom (1986). The main purpose of this research is to assess the performance of two uniform-distribution substitution methods, the one-step restricted maximum likelihood method, and the regression method in terms of their accuracy in estimating population parameters based on data sets containing BDL values. The families of distribution under study will be log-normal and Gamma. A robust and simple method will be recommended to the FDEP for the analysis of water-quality data.

2 Statistical Methods

In this section, we list four basic statistical methods which will be applied to handle the BDL values in environmental data analyses. Our attention will be restricted to analyzing a data set containing only one MDL, i,e., the data set is left-censored at the MDL. Methods for handling multiple MDLs in a data set will be discussed in later studies.

Throughout this section, let $\{x_1, \ldots, x_n\}$ be an ordered random sample from a population, i.e., x_1 is the smallest value in the sample and x_n is the largest. Suppose that $m(< n)$ values in the sample are smaller than a given MDL (=c, say). In other words, $\{x_1, \ldots, x_m\}$ is the BDL portion and $\{x_{m+1}, \ldots, x_n\}$ is the ADL portion of the sample.

In practice, the BDL values, $\{x_1, \ldots, x_m\}$, are not available and need to be estimated. In a simulation study, a full sample with both BDL and ADL portions will be generated from a population and statistics such as sample mean and sample variance based on the full sample can be calculated. Four different methods of handling the BDL values will be applied to calculate the statistics using only the ADL portion of the sample. The calculated statistics will then be compared to the statistics calculated from the full sample for the purpose of assessing the performance of the estimation methods under consideration.

2.1 Uniform distribution method based on the original scale

One substitution method is to fill in the BDL values in a sample based on a uniform distribution. More specifically, assume that the BDL values are independent and uniformly distributed on the interval [0, MDL]. This method can be described in two steps:

- generate m values $\{x_1^*, \ldots, x_m^*\}$ from the uniform distribution on the interval [0, MDL] and treat $\{x_1^*, \ldots, x_m^*\}$ as the real values for the BDL portion.
- combine $\{x_1^*, \ldots, x_m^*\}$ with the ADL portion $\{x_{m+1}, \ldots, x_n\}$ of the sample and calculate the sample statistics.

Notice that the expected value for the uniform distribution on [0, MDL] is MDL/2 and the sample mean $\hat{\mu}_B = \sum_{i=1}^m x_i^*/m$ for the BDL portion should be close to MDL/2. It is easy to see that the sample mean based on $\{x_1^*, \dots, x_m^*, x_{m+1}, \dots, x_n\}$ is

$$\begin{aligned}\hat{\mu} &= \left[\sum_{i=1}^m x_i^* + \sum_{i=m+1}^n x_i\right] \Big/ n = \left[m\left(\sum_{i=1}^m x_i^*/m\right) + \sum_{i=m+1}^n x_i\right] \Big/ n \\ &= \left[m\hat{\mu}_B + \sum_{i=m+1}^n x_i\right] \Big/ n \approx \left[m(MDL/2) + \sum_{i=m+1}^n x_i\right] \Big/ n.\end{aligned}$$

Therefore, filling-in the BDL values based on the uniform distribution on [0, MDL] is in fact approximately equivalent to substituting the BDL values by MDL/2 for the estimation of the sample mean.

When the sampled population is log-normal, the left tail of the underlying distribution cannot be well approximated by any uniform distribution. The substitution method based on a uniform distribution on [0, MDL] has no theoretical basis on which any accurate estimates for the population parameters may be expected. This method suffers the same drawbacks as any simple substitution methods (e.g., substituting the BDL values simply by 0, MDL/2, or MDL).

2.2 Uniform distribution method based on the logarithm scale

Another substitution method is to fill in the BDL values based on the logarithm scale. In this method the m log(BDL) values are assumed to be independent and uniformly distributed on the interval [0, log(MDL)]. The procedure and calculations of this method are similar to those for Method 1.

- generate m values $\{y_1, \dots, y_m\}$ from the uniform distribution on [0, log(MDL)] and treat $\{x_1^* = e^{y_1}, \dots, x_m^* = e^{y_m}\}$ as the real values for the BDL portion.

- combine $\{x_1^*, \dots, x_m^*\}$ with the ADL portion $\{x_{m+1}, \dots, x_n\}$ of the sample and calculate the sample statistics.

Intuitively, when a random variable Y has a log-normal distribution, the distribution of $\log(Y)$ is normal that has a flatter left tail than the log-normal distribution. One may expect that simple substitutions based on the logarithm scale will give more accurate estimates for population parameters than substitutions based on the original scale. But similar to Method 1, uniform-distribution substitution based on the original scale, this method has very little theoretical basis.

2.3 One-step restricted maximum likelihood method

The one-step restricted maximum likelihood method was proposed by Persson and Rootzen (1977) for the estimation of mean and variance based on a censored normal sample. Suppose that $\{y_1, \ldots, y_n\}$ is a random sample from a normal distribution with mean μ_y and variance σ_y^2. Then the probability of each observation falling below the MDL ($= c$) is

$$P(y_i < c) = P\left((y_i - \mu_y)/\sigma_y < (c - \mu_y)/\sigma_y\right) = \Phi(\theta),$$

where $\theta = (c - \mu_y)/\sigma_y$ and $\Phi(\cdot)$ is the standard normal distribution. Let K be the number of observations in the sample whose values are below $MDL = c$. Then K is a random variable with a binomial distribution $Binomial(n, \Phi(\theta))$. For a sample with m BDL values, i.e., $K = m$, Persson and Rootzen (1977) pointed out that a natural estimate for $\Phi(\theta)$ is m/n, which implies that a good estimate for θ is

$$\theta^* = \Phi^{-1}(m/n).$$

If the sample observations or measurements $\{x_1, \ldots, x_n\}$ are from a log-normal distribution, the one-step restricted maximum likelihood method can be applied to $y_i = \log(x_i)$. Using the notation in Ahn (1998), let $d = \log(MDL) = \log(c)$ and let $\hat{\mu}_{yA}$ and $\hat{\sigma}_{yA}^2$ denote the sample mean and sample variance of $\{y_i : y_i \geq d\}$, respectively. Then the one-step restricted maximum likelihood estimates for the mean and variance of $y = \log(x)$ are

$$\begin{aligned} \hat{\mu}_y &= \hat{\mu}_{yA} - a\sigma^*, \\ \hat{\sigma}_y^2 &= \hat{\sigma}_{yA}^2 - (a\theta^* - a^2)(\sigma^*)^2, \end{aligned} \tag{2.1}$$

where $a = nf(\theta^*)/(n - m)$, with $f(\cdot)$ being the density function of the standard normal distribution, and where

$$\sigma^* = (1/2)\{C + [C^2 + 4\hat{\sigma}_{yA}^2 + 4(\hat{\mu}_{yA} - d)^2]^{1/2}\}$$

with $C = \theta^*(\hat{\mu}_{yA} - d)$.

Note that in (2.1) θ^* is an estimate of the parameter $\theta = (d - \mu_y)/\sigma_y$. Ahn (1998) suggested to estimate θ by $\epsilon = (d - \hat{\mu}_{yA})/\hat{\sigma}_{yA}$. But it is obvious that $\hat{\mu}_{yA}$ is an over-estimate of μ_y and σ_{yA} is an under-estimate of σ_y, which implies that $|\epsilon|$ is an over-estimate of $|\theta|$. Alternately, we recommend the use of $\theta^* = \Phi^{-1}(m/n)$, originally proposed by Persson and Rootzen (1977), as a natural estimate for θ.

After the estimates $\hat{\mu}_y$ and $\hat{\sigma}_y^2$ are obtained, the mean value and variance of the original random variable x can be estimated by

$$\begin{aligned} \hat{\mu} &= \exp(\hat{\mu}_y + \hat{\sigma}_y^2/2), \\ \hat{\sigma}^2 &= \hat{\mu}^2[\exp(\hat{\sigma}_y^2) - 1]. \end{aligned} \tag{2.2}$$

It should be pointed out that the one-step restricted maximum likelihood method discussed in this subsection is specially designed for normal distributions or log-normal distributions. When the distribution assumption is violated, i.e., for example, when the sampled population is a Gamma distribution, the method may give poor estimates for the population parameters.

2.4 Regression method

The regression method is studied by Helsel and Gilliom (1986) and summarized in Helsel and Hirsch (1992, Chapter 13). Define

$$p_i = \frac{i - \omega}{n + 1 - \omega},$$

where ω is used to correct bias in the extreme observations. Following the recommendation by Newman et al. (1989, 1995), $\omega = 3/8$ will be used in this study.

Assume that $\{x_1, \ldots, x_n\}$ is a random sample from a log-normal distribution. Let $z_i = \Phi^{-1}(p_i)$ be the p_ith quantile (or the 100 p_ith percentile) of the standard normal distribution. A linear regression model can be fitted to $\{(y_i, z_i) : i = m+1, \ldots, n\}$ with $y_i = \log(x_i)$ as the response variable. The model has the form:

$$y_i = \alpha + \beta z_i + \epsilon_i, \quad i = m+1, \ldots, n, \tag{2.3}$$

where the ϵ_i's are independent and normally distributed random errors.

After fitting the model, the estimated equation $\hat{y}_i = \hat{\alpha} + \hat{\beta} z_i$ can be used to extrapolate the BDL values $\{x_1^* = e^{y_1}, \ldots, x_m^* = e^{\hat{y}_1}\}$. The summary statistics of the data such as the sample mean and variance can then be calculated based on the $\{x_{m+1}, \ldots, x_n\}$ and the fill-in values $\{x_1^*, \ldots, x_m^*\}$.

The regression method is essentially designed for data sets from log-normal distributions. Helsel and Gilliom (1986) performed an intensive simulation study and showed that this method produces consistently small errors for various summary statistics such as the sample mean, variance, skewness, and kurtosis.

3 Simulation Study

A simulation study is performed to compare the performance of the four statistical methods in terms of estimating the mean and standard deviation of a distribution based on samples containing BDL values. Log-normal and Gamma distributions will be used as the target distributions for different water-quality parameters. The MDLs used in this study are 2, 2.5, 3, 3.5, and 4 $\mu g/L$. The four methods described in Section 2 will be called Methods 1-4.

For a chosen distribution, 100 independent samples, each with 1000 replicates, will be generated. The four methods presented in Section 2 will be used to estimate the mean value and standard deviation of each sample after removing the BDL values. The final estimates of mean and standard deviation for each distribution will be the average of the 100 estimated values.

For example, consider the estimation based on Method 1, the uniform distribution method based on the original scale. For each simulated sample, the BDL values $\{x_1, \ldots, x_m\}$, where the number m varies for different samples, are removed from the sample and $\{x_1^*, \ldots, x_m^*\}$ are generated from the uniform distribution on [0, MDL]. The the simulated BDL values $\{x_1^*, \ldots, x_m^*\}$ are combined with the ADL portion $\{x_{m+1}, \ldots, x_n\}$ to calculate the sample mean and standard deviation. For the ith sample, the sample mean and standard deviation based on $\{x_1, \ldots, x_n\}$ are denoted by $\bar{x}_i$ and s_i, respectively, while the estimated mean and standard deviation based on the pseudo-sample $\{x_1^*, \ldots, x_m^*, x_{m+1}, \ldots, x_n\}$ are denoted by $\bar{x}_i^*$ and s_i^*, respectively. The final estimates for the population parameters μ and σ based on the pseudo samples are $\bar{x}^* = \sum_{i=1}^{100} \bar{x}_i^*/100$ and $\bar{s}^* = \sum_{i=1}^{100} s_i^*/100$, respectively.

3.1 Log-normal Distributions

Table 1. Estimated mean and standard deviation by the four methods based on 100 samples generated from the log-normal distribution with mean 12.1825 and standard variation 15.9692 ($\log(x) \sim N(2,1)$).

Sample mean and S.D. (Average of 100) based on full samples:
(12.2561, 16.2425)

Estimated Mean Values By the Four Methods

MDL($\mu g/L$)	Method 1	Method 2	Method 3	Method 4	BDL%
2.0	12.2224	12.2657	12.2434	12.2556	9.6
2.5	12.2052	12.2582	12.2431	12.2549	13.9
3.0	12.1854	12.2429	12.2435	12.2537	18.3
3.5	12.1630	12.2198	12.2461	12.2522	22.7
4.0	12.1482	12.1906	12.2463	12.2506	26.9

Estimated S.D. Values By the Four Methods

MDL($\mu g/L$)	Method 1	Method 2	Method 3	Method 4	BDL%
2.0	16.2662	16.2356	16.1273	16.2427	9.6
2.5	16.2778	16.2405	16.1233	16.2430	13.9
3.0	16.2905	16.2501	16.1224	16.2436	18.3
3.5	16.3046	16.2644	16.1533	16.2443	22.7
4.0	16.3132	16.2817	16.1513	16.2451	26.9

Comments:

0). Estimates by the four methods should be compared to the averaged sample mean 12.2561 and the averaged sample standard deviation 16.2425, not to the population mean and standard deviation.

1). Method 1 under-estimates the mean and over-estimates the standard deviation. The estimates become worse when MDL increases.

2). Method 2 gives moderately good estimates.

3). Method 3 under-estimates both the mean and standard deviation.

4). Method 4 gives very good estimates.

5). Ranked order from the best to the worst: (Method 4, Method 2, Method 3, Method 1)

Table 2. Estimated mean and standard deviation by the four methods based on 100 samples generated from the log-normal distribution with mean 15.1803 and standard variation 27.2425 ($\log(x) \sim N(2, 1.2)$).

Sample mean and S.D. (Average of 100) based on full samples:
(15.3106, 27.0934)

Estimated Mean Values By the Four Methods

MDL($\mu g/L$)	Method 1	Method 2	Method 3	Method 4	BDL%
2.0	15.2812	15.3413	15.3013	15.3095	13.7
2.5	15.2701	15.3401	15.3010	15.3080	18.2
3.0	15.2626	15.3353	15.3071	15.3062	22.5
3.5	15.2559	15.3203	15.3124	15.3045	26.6
4.0	15.2584	15.3080	15.3121	15.3023	30.3

Estimated S.D. Values By the Four Methods

MDL($\mu g/L$)	Method 1	Method 2	Method 3	Method 4	BDL%
2.0	27.1096	27.0767	27.6374	27.0940	13.7
2.5	27.1155	27.0774	27.6304	27.0947	18.2
3.0	27.1196	27.0800	27.6854	27.0955	22.5
3.5	27.1228	27.0873	27.7401	27.0961	26.6
4.0	27.1215	27.0932	27.7237	27.0969	30.3

Comments:

0). Estimates by the four methods should be compared to the averaged sample mean 15.3106 and the averaged sample standard deviation 27.0934, not to the population mean and standard deviation.

1). Method 1 under-estimates the mean. The estimates become worse when MDL increases.

2). Method 2 gives moderately good estimates.

3). Method 3 over-estimates the standard deviation.

4). Method 4 gives very good estimates.

5). Ranked order from the best to the worst: (Method 4, Method 2, Method 3, Method 1)

Table 3. Estimated mean and standard deviation by the four methods based on 100 samples generated from the log-normal distribution with mean 10.1758 and standard variation 9.6347 ($\log(x) \sim N(2, 0.8)$).

Sample mean and S.D. (Average of 100) based on full samples:
(10.2258, 9.6055)

Estimated Mean Values By the Four Methods

MDL($\mu g/L$)	Method 1	Method 2	Method 3	Method 4	BDL%
2.0	10.1995	10.2234	10.2395	10.2259	5.1
2.5	10.1752	10.2102	10.2389	10.2262	8.8
3.0	10.1453	10.1873	10.2377	10.2268	12.8
3.5	10.1124	10.1535	10.2371	10.2271	17.5
4.0	10.0767	10.1104	10.2368	10.2278	22.0

Estimated S.D. Values By the Four Methods

MDL($\mu g/L$)	Method 1	Method 2	Method 3	Method 4	BDL%
2.0	9.6306	9.6075	9.7241	9.6053	5.1
2.5	9.6536	9.6191	9.7226	9.6050	8.8
3.0	9.6785	9.6384	9.7144	9.6045	12.8
3.5	9.7053	9.6656	9.7089	9.6040	17.5
4.0	9.7335	9.6988	9.7002	9.6035	22.0

Comments:

0). Estimates by the four methods should be compared to the averaged sample mean 10.2258 and the averaged sample standard deviation 9.6055, not to the population mean and standard deviation.

1). Method 1 under-estimates the mean and over-estimates the sample standard deviation. The estimates become worse when MDL increases.

2). Method 2 under-estimates the mean and over-estimates the sample standard deviation for large MDLs.

3). Method 3 over-estimates the mean and standard deviation.

4). Method 4 gives very good estimates.

5). Ranked order from the best to the worst: (Method 4, Method 3, Method 2, Method 1)

3.2 Gamma distributions

Table 4. Estimated mean and standard deviation by the four methods based on 100 samples generated from the Gamma distribution with mean 12.1825 and standard variation 15.9692.

Sample mean and S.D. (Average of 100) based on full samples:
(12.0433, 15.7643)

Estimated Mean Values By the Four Methods

MDL($\mu g/L$)	Method 1	Method 2	Method 3	Method 4	BDL%
2.0	12.1213	12.2454	14.2220	12.2638	27.9
2.5	12.1559	12.2772	13.9075	12.3263	31.4
3.0	12.1946	12.3047	13.6841	12.3915	34.7
3.5	12.2348	12.3310	13.5180	12.4591	37.7
4.0	12.2868	12.3517	13.3908	12.5252	40.4

Estimated S.D. Values By the Four Methods

MDL($\mu g/L$)	Method 1	Method 2	Method 3	Method 4	BDL%
2.0	15.7083	15.6191	34.7518	15.6103	27.8
2.5	15.6849	15.5978	31.9243	15.5700	31.4
3.0	15.6593	15.5798	29.8339	15.5293	34.7
3.5	15.6338	15.5631	28.1835	15.4884	37.7
4.0	15.6008	15.5503	26.7867	15.4494	40.4

Comments:

0). Estimates by the four methods should be compared to the averaged sample mean 12.0433 and the averaged sample standard deviation 15.7643, not to the population mean and standard deviation.

1). Method 1 over-estimates the mean and under-estimates the sample standard deviation. The estimates become worse when MDL increases.

2). Method 2 over-estimates the mean and under-estimates the sample standard deviation. The estimates are worse than those given by Method 1.

3). Method 3 over-estimates the mean and the estimated standard deviations are the worst.

4). Method 4 over-estimates the mean and under-estimates the sample standard deviation. The estimates are worse than those given by Methods 1 and 2.

5). Ranked order from the best to the worst: (Method 1, Method 2, Method 4, Method 3)

Table 5. Estimated mean and standard deviation by the four methods based on 100 samples generated from the Gamma distribution with mean 15.1803 and standard variation 27.2425.

Sample mean and S.D. (Average of 100) based on full samples: (15.1845, 27.0047)

Estimated Mean Values By the Four Methods

MDL($\mu g/L$)	Method 1	Method 2	Method 3	Method 4	BDL%
2.0	15.4048	15.5832	22.4989	15.5230	40.8
2.5	15.4765	15.6439	21.4255	15.6071	43.7
3.0	15.5538	15.7028	20.6547	15.6947	46.3
3.5	15.6344	15.7541	20.0375	15.7784	48.5
4.0	15.7270	15.8089	19.5448	15.8586	50.3

Estimated S.D. Values By the Four Methods

MDL($\mu g/L$)	Method 1	Method 2	Method 3	Method 4	BDL%
2.0	26.8857	26.7887	135.6812	26.8275	40.8
2.5	26.8484	26.7573	116.6942	26.7836	43.7
3.0	26.8090	26.7273	103.6845	26.7410	46.3
3.5	26.7686	26.7018	93.3882	26.7012	48.5
4.0	26.8089	26.6750	85.2448	26.6638	50.3

Comments:

0). Estimates by the four methods should be compared to the averaged sample mean 15.1845 and the averaged sample standard deviation 27.0047, not to the population mean and standard deviation.

1). Method 1 over-estimates the mean and under-estimates the sample standard deviation. The estimates become worse when MDL increases.

2). Method 2 over-estimates the mean and under-estimates the sample standard deviation. The estimates are worse than those given by Method 1.

3). Estimates using Method 3 are extremely poor.

4). Method 4 over-estimates the mean and under-estimates the sample standard deviation. The estimates are worse than those given by Method 1 but better than those given by Method 2.

5). Ranked order from the best to the worst: (Method 1, Method 4, Method 2, Method 3)

Table 6. Estimated mean and standard deviation by the four methods based on 100 samples generated from the Gamma distribution with mean 10.1758 and standard variation 9.6347.

Sample mean and S.D. (Average of 100) based on full samples: (10.1837, 9.7096)

Estimated Mean Values By the Four Methods

MDL($\mu g/L$)	Method 1	Method 2	Method 3	Method 4	BDL%
2.0	10.1831	10.2535	10.9029	10.2977	15.8
2.5	10.1854	10.2619	10.8038	10.3387	19.6
3.0	10.1884	10.2619	10.7366	10.3823	23.3
3.5	10.1957	10.2582	10.6926	10.4265	26.8
4.0	10.2034	10.2523	10.6651	10.4649	30.3

Estimated S.D. Values By the Four Methods

MDL($\mu g/L$)	Method 1	Method 2	Method 3	Method 4	BDL%
2.0	9.7128	9.6453	14.5084	9.6073	15.8
2.5	9.7109	9.6379	13.8076	9.5732	19.6
3.0	9.7087	9.6379	13.2430	9.5383	23.3
3.5	9.7021	9.6409	12.7687	9.5043	26.8
4.0	9.6959	9.6454	12.3913	9.4687	30.3

Comments:

0). Estimates by the four methods should be compared to the averaged sample mean 10.1837 and the averaged sample standard deviation 9.7096, not to the population mean and standard deviation.

1). Method 1 give very good estimates for the sample mean and standard deviation.

2). Method 2 over-estimates the mean and under-estimates the sample standard deviation.

3). Method 3 over-estimates the mean and the sample standard deviation. The estimates are worse than those given by Methods 1, 2, and 4.

4). Method 4 over-estimates the mean and under-estimates the sample standard deviation. The estimates are worse than those given by Method 2 but better than those given by Method 3.

5). Ranked order from the best to the worst: (Method 1, Method 2, Method 4, Method 3)

4 Summary

A simulation study is carried out to evaluate the performance of four statistical methods for the estimation of sample statistics based on a data set containing values below a detection limit. The four methods are (a) uniform-distribution method based on the original scale, (b) uniform-distribution method based on the logarithm scale, (c) the one-step restricted maximum likelihood method, and (d) the regression method.

Log-normal and Gamma distributions have many similarities. They often can be mistaken from each other, and or mis-spesified. In this simulation study, these two families of distributions are chosen as the underlying populations. The log-normal distributions chosen for this study are $\log(x) \sim N(2, 1)$, $\log(x) \sim N(2, 1.2)$, and $\log(x) \sim N(2, 0.8)$, where $N(\mu, \sigma)$ denotes the normal distribution with mean μ and standard deviation σ. The three Gamma distributions chosen have the same means and standard deviations as the three log-normal distributions. Other distributions, such as $\log(x) \sim N(1.8, 1)$, $\log(x) \sim N(1.8, 1.2)$, $\log(x) \sim N(1.8, 0.8)$, $\log(x) \sim N(1.6, 1)$, $\log(x) \sim N(1.6, 1.2)$, and $\log(x) \sim N(1.6, 0.8)$ have also been simulated. The simulation results give similar conclusions to those of the first three distributions; hence they are not presented in this report.

One hundred samples, each with 1000 observations, were generated from a chosen distribution. The MDLs used in this study are 2, 2.5, 3.0, 3.5, and 4 $\mu g/L$. For each sample, the BDL values are removed and the sample mean and standard deviation are estimated by the four methods. The final estimates given by each method for a chosen distribution are the averages of the 100 estimates.

Here are some initial findings based on the simulation:

- A Gamma distribution has a higher percentage of BDL values than a log-normal distribution with the same mean and standard deviation.

- For log-normal distributions, the ranked order for the four methods from the best to the worst is (Method 4, Method 2, Method 3, Method 1), i.e., the regression method gives the most accurate estimates and the uniform-distribution method based on the original scale gives the worst estimates.

- For Gamma distributions, the ranked order for the four methods from the best to the worst is (Method 1, Method 2, Method 4, Method 3), i.e., the the uniform-distribution method based on the original scale gives the most accurate estimates and the one-step restricted maximum likelihood method gives the worst estimates.

The results indicate that when the underlying population is a Gamma distribution but mis-specified as a log-normal distribution, the regression method and the one-step restricted maximum likelihood method perform worse than the uniform-distribution methods. In particular, the one-step restricted maximum likelihood method gives very poor estimates for the sample statistics. This is not surprising since the regression method and the one-step restricted maximum likelihood method both were designed specifically for samples from normal (or log-normal) distributions.

The conclusions from this study are the follows:

- For data sets with BDL values, the performance of different estimation methods depend on the distribution family of the underlying sampled-population. Before analyzing a data set with BDL values, the population family of the data set should be carefully studied and determined.

- If a water-quality variable has a normal or log-normal distribution, the regression method is recommended to handle the BDL values for the purpose of estimating the population parameters.

- If a water-quality variable does not belong to the normal family or the log-normal family, the regression method and the one-step restricted maximum likelihood method need to be modified to provide accurate estimates.

5 References

Ahn, H. (1998), "Estimating the Mean and Variance of Censored Phosphorus Concentrations in Florida Rainfall", *Journal of the American Water Resources Association*, Vol.34, No.3, 583-593.

Gleit, A. (1985), "Estimation for Small Normal Data Sets with Detection Limits", *Environmental Science Technol*, 19(12): 1201-1206.

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

Helsel, D. R. and Hirsch, R. M. (1992), *Statistical Methods in Water Resources*, Elsevier, New York.

Kite, G. W. (1988), *Frequency and Risk Analysis in Hydrology*, Water Resources Publications, Fort Collins, Colorado.

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

Newman , M. C., Greene, K. D., and Dixon, P. M. (1995), *UNCENSOR Version 4.0*, Savannah River Ecology Laboratory, Aiken, South Carolina.

Pearson, E. S. and Rootzen, H. (1977), "Simple and Highly Efficient Estimators for a Type-I Censored Normal Sample", *Biometrika*, 64(1): 123-128.