Statistical modeling of Environmental data with non-detects

Jennifer Niemann, Senior Capstone Project

Dr. Mallick, Capstone Advisor

December 14, 2016

Abstract

In the environmental sciences, portions of collected data are often reported as non-detect, meaning that the actual data point is known only to be below the detection limit of a measuring device. In mainstream statistics, this type of data is known as left-censored data. In this project, the performance of a variety of substitution methods will be examined, as well as maximum likelihood estimation and the Kaplan-Meier method for estimating summary statistics (primarily the mean) of left-censored data with respect to certain statistical criteria like bias and mean squared error. The performance of these methods were also investigated in the context of construction of confidence intervals for mean and upper tolerance limits. After identifying the best method, the results were applied to a real life environmental data set provided by Neptune and Company, Inc.

1 Introduction

Environmental scientists are frequently interested in estimating the mean (or median) contaminant concentration in a particular area. For example, it may be necessary to estimate the amount of arsenic in ground water or the amount of sulfur dioxide in air. It is very common in these environmental data sets to have observations that are unable to be detected by the measuring device, and thus reported as non-detects, or below the detection limit. In other words, the amount of contaminant present is less than the detection limit of the measuring device being used. These non-detectable amounts are also known as censored observations, and they contain limited information about the actual contaminant concentration in these areas, i.e.

we only know that the actual concentration is below the detection limit or reporting limit.

The definition of detection limit varies. According to the US Environmental Protection Agency [2], the detection limit is defined as "the minimum concentration of a substance that can be measured and reported with 99% confidence that the analyte concentration is greater than zero." In other words, it is 'the lowest concentration of a chemical that can reliably be distinguished from a zero concentration'. When a data set includes non-detectable observations, analysis cannot be done in a traditional manner, since a number of observations are only known to be below a certain threshold. The purpose of this study is to investigate and discover the best way to estimate summary statistics using data containing censored observations.

There is currently no singular, preferred method for working with censored observations when analyzing data. Researchers have debated the best method for analysis, and have yet to agree on any one technique. Different studies have come to different conclusions depending on the sample size and percent of censored observations in the sample. She [9] compared several methods, and concluded that the Kaplan-Meier method performed best. Both Lubin et al. [7] and Hewitt and Ganser [4] found that substitution performed poorly, and that maximum likelihood estimation produced the best results. However, they did not find the Kaplan-Meier method to be a well performing method. Antweiler and Taylor [1] observed that the Kaplan-Meier method did perform well, along with two substitution methods also providing good results, but the maximum likelihood estimation method did not produce good estimates in their study. Helsel [3] gives recommendations of when to use both the Kaplan-Meier and maximum likelihood estimation methods, depending on the sample size and percent of censored observations, but discourages the use of substitution altogether. Therefore, it is clear that there is no universal approach when working with censored data. This research will try to establish which method is most appropriate in a variety of situations by testing various statistical techniques that can be used when dealing with such left-censored samples containing a range of percentages of censored observations.

We define the distributions used in this research in Section 2, different methods for handling left-censored are presented in Section 3, and various criteria for assessment or evaluation of efficiency of these methods are discussed in Section 4. In Section 5, we present the simulation results, followed by a real life example in Section 6. Finally, some concluding remarks are given in Section 7.

2 Distributions

The most well-known and frequently used probability distribution is the normal (or Gaussian) distribution. The approximate shape of a normal distribution is symmetrical, unimodal, and bell-shaped.

Definition 1. A random variable X is said to have a normal probability distribution if and only if, for $\sigma > 0$ and $-\infty < \mu < \infty$, the density function of X is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], -\infty < x < \infty. \tag{1}$$

The normal density function contains two parameters, μ and σ. For X, a normally distributed random variable with parameters μ and σ, the mean is $E(X) = \mu$ and the variance is $V(X) = \sigma^2$. The normal distribution is used widely because of its characteristic of supporting all real values $-\infty < x < \infty$ and the fact that the mean of a random sample of size n from any distribution converges to a normal distribution as $n \to \infty$, which makes it a suitable approximation for many distributions.

For this research, however, the lognormal distribution will be more appropriate, because it has a positive support (or range). This aligns with environmental science studies, where the values being measured will be positive, such as the level of contaminant in a water sample.

Definition 2. A random variable Y has a lognormal probability distribution with parameters μ and σ if $X = \ln Y$ has a normal distribution with mean μ and standard deviation σ. The probability density function $f(y)$ and cumulative distribution function $F(y)$ of the lognormal distribution are given by

$$\begin{aligned} f(y) &= \frac{1}{y\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln y - \mu)^2}{2\sigma^2}\right], y > 0 \\ F(y) &= P(Y \le y) = \int_{-\infty}^{y} f(t)\mathrm{d}t. \end{aligned} \tag{2}$$

The parameters for the lognormal distribution are the mean and standard deviation for the corresponding normally distributed random variable X, where $Y = e^X$. The mean of the lognormal distribution is $E(Y) = \exp\left(\mu + \frac{1}{2}\sigma^2\right)$ and the variance is $V(Y) = \exp(2\mu + \sigma^2) * \exp(\sigma^2 - 1)$.

3 Overview of Methods

Censored observations are not quantified, since we only know that they are below a detection limit. The problem is how to work with a data set that includes censored observations. We can not simply discard the censored observations, as this leads to loss of important information. So, in this section we explain some existing methods, such as substitution, maximum likelihood estimation, and the Kaplan-Meier method for analyzing such data sets.

3.1 Substitution

Multiple studies argue against using substitution because it could be considered fabrication of data. In other words, substitution is essentially creating false data, and can be very invasive to a study. However, substitution is a method that is still widely used in industry because of its simplicity, and is perhaps acceptable under certain conditions. In this study, six different substitution methods are investigated, which include replacing the non-detects with the detection limit (DL), $DL/2$, $DL/\sqrt{2}$, 0, values evenly spaced between 0 and DL, and random numbers between 0 and DL.

Example 1

To illustrate these substitution methods, we will use the following sample:

$$0.8, 1.2, 2.1, 2.7, 3.5, 4.6, 5.3, 6.9, 7.7, 8.2.$$

The mean of this sample is 4.3. Now suppose the detection limit(DL) is 3, i.e., all observations below this value were not detected. All that is known is that these values are less than 3. Therefore, our censored sample becomes

$$< 3, < 3, < 3, < 3, 3.5, 4.6, 5.3, 6.9, 7.7, 8.2.$$

The six substitution methods are demonstrated in Table 1.

Data	0	DL	$DL/2$	$DL/\sqrt{2}$	Evenly spaced	Random
< 3	0	3	1.5	2.12	0	0.4
< 3	0	3	1.5	2.12	1	0.9
< 3	0	3	1.5	2.12	2	1.0
< 3	0	3	1.5	2.12	3	2.9
3.5	3.5	3.5	3.5	3.5	3.5	3.5
4.6	4.6	4.6	4.6	4.6	4.6	4.6
5.3	5.3	5.3	5.3	5.3	5.3	5.3
6.9	6.9	6.9	6.9	6.9	6.9	6.9
7.7	7.7	7.7	7.7	7.7	7.7	7.7
8.2	8.2	8.2	8.2	8.2	8.2	8.2
Mean	3.62	4.82	4.22	4.47	4.22	4.14

Table 1: Demonstration of the six substitution methods.

3.2 Maximum Likelihood Estimation

A more generally accepted method is maximum likelihood estimation. Described by Helsel [3], this method assigns an assumed distribution to the data—both the detected values and the proportion of censored observations. Using the assumed distribution, i.e. the lognormal in this case, the values above the DL and the proportion of data below the DL can be used to estimate the parameters of interest that best match the data to the distribution.

Definition 3. Let $y_1, y_2, ..., y_n$ be sample observations taken on corresponding random variables $Y_1, Y_2, ..., Y_n$ whose distribution depends on an unknown parameter θ. Then, if $Y_1, Y_2, ..., Y_n$ are continuous random variables, the likelihood function $L(\theta) = L(\theta|y_1, y_2, ..., y_n)$ is defined to be the joint density evaluated at $y_1, y_2, ..., y_n$. This likelihood function is given as $L(\theta) = L(\theta|y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f(y_i|\theta)$. Then the maximum likelihood estimator (MLE) of θ is the the value that maximizes the likelihood function $L(\theta)$.

Definition 3 describes the likelihood of matching the observed data to the assumed distribution with parameter θ when all the observations are detected. However, we will need the likelihood function in such a way that it accounts for both observed and censored data. Observed data comes into the likelihood function through the probability density function $f(y)$ and the censored observations can be accounted for by the cumulative distribution function $F(y) = P(Y \leq y) = \int_{-\infty}^{y} f(t)dt$, as we know the non-detects are

known to be less than or equal to y. Therefore, the form of the likelihood function is as follows:

$$L(\theta \mid y_1, y_2, ..., y_n) = \prod_{y \in D} f(y) \cdot \prod_{y \in C} F(y) \tag{3}$$

where D is the set of all observed or detected values and C is the set of all left-censored or non-detected values. Instead of maximizing the above likelihood function, an easier and more common approach is to maximize the log-likelihood function, which is the natural logarithm of $L(\theta)$.

Example 2

Using the left-censored data set from Example 1, we obtain the likelihood function as follows:

$$L(\mu, \sigma) = F(3)^4 \cdot f(3.5) \cdot f(4.6) \cdot f(5.3) \cdot f(6.9) \cdot f(7.7) \cdot f(8.2) \tag{4}$$

and the log-likelihood function as

$$\begin{aligned} \ln L(\mu, \sigma) &= \ln[F(3)^4 \cdot f(3.5) \cdot f(4.6) \cdot f(5.3) \cdot f(6.9) \cdot f(7.7) \cdot f(8.2)] \\ &= 4 \cdot \ln F(3) + \ln f(3.5) + \ln f(4.6) + \ln f(5.3) + \ln f(6.9) \\ &\quad + \ln f(7.7) + \ln f(8.2). \end{aligned} \tag{5}$$

Maximizing the above equation, we obtain the the MLEs for this example as $\hat{\mu} = 1.35$ and $\hat{\sigma} = 0.57$. Hence, the MLE of the lognormal mean is $\exp(\hat{\mu} + \hat{\sigma}^2/2) = 4.54$

3.3 Kaplan-Meier Estimation

The Kaplan-Meier (KM) method is a nonparametric method for dealing with censored data. It is widely used in survival or lifetime data analysis to estimate the survival function, which is then used to estimate different summary statistics. This method requires the use of right-censored data. Therefore, left-censored data must be transformed into right-censored data before applying this method.

Definition 4. The Kaplan-Meier estimator of the survival function $S(t) = P(T \geq t)$ is:

$$\begin{aligned} \hat{S}(t) &= \prod_{j:t_j<t} \frac{r_j - d_j}{r_j} \\ &= \prod_{j:t_j<t} \left(1 - \frac{d_j}{r_j}\right), \end{aligned} \tag{6}$$

where t_j is the set of distinct death times observed in the sample; r_j is the number of individuals "at risk" right before the j^{th} death time t_j; and d_j is the number of deaths at t_j.

Then the mean is estimated by computing the following integral:

$$\hat{\mu}_{KM} = \int_0^{t_{max}} \hat{S}(t)\, dt, \tag{7}$$

where t_{max} is the largest observed death time. Since $\hat{S}(t)$ is a step function, we can estimate this integral using a summation, i.e.,

$$\hat{\mu}_{KM} = \int_0^{t_{max}} \hat{S}(t)\, dt \approx \Sigma_j[\hat{S}(t_{j-1})(t_j - t_{j-1})]. \tag{8}$$

In order to transform left-censored data into right-censored data, each observation value must be subtracted from a value greater than the maximum value in the data set. Once transformed, the Kaplan-Meier method can be performed on left-censored data.

Example 3

We consider the left-censored data set From Example 1,

$$< 3, < 3, < 3, < 3, 3.5, 4.6, 5.3, 6.9, 7.7, 8.2$$

The maximum value in this data set is 8.2. Therefore, the fixed value used for transformation must be > 8.2. We subtract the values from 9. This gives the transformed right-censored data set

$$> 6, > 6, > 6, > 6, 5.5, 4.4, 3.7, 2.1, 1.3, 0.8.$$

Table 2 summarizes the necessary values to compute the Kaplan-Meier estimates.

j	t_j	r_j	d_j	$1-\frac{d_j}{r_j}$	$\hat{S}(t_j)$	$t_j-(t_{j-1})$	$\hat{S}(t_{j-1})(t_j-t_{j-1})$
0	0	10	0	1	1	–	–
1	0.8	10	1	0.9	0.9	0.8	0.8
2	1.3	9	1	0.889	0.8	0.5	0.45
3	2.1	8	1	0.879	0.7	0.8	0.64
4	3.7	7	1	0.857	0.6	1.6	1.12
5	4.4	6	1	0.833	0.5	0.7	0.42
6	5.5	5	1	0.8	0.4	1.1	0.55

Table 2: Computation of the Kaplan-Meier estimates

Then the estimated mean computed from the transformed right-censored data set, using the Kaplan-Meier method, is

$$\begin{aligned}\hat{\mu}_{KM} &= \Sigma_j[\hat{S}(t_{j-1})(t_j - t_{j-1})] \\ &= 0.8 + 0.45 + 0.64 + 1.12 + 0.42 + 0.55 \\ &= 3.98\end{aligned}$$

Therefore, the estimated mean using our original left-censored data is $(9 - \hat{\mu}_{KM}) = 5.02$.

4 Evaluation Criterion

In this section, we define and discuss different criteria for assessing effectiveness of the methods presented in the previous section for estimating summary statistics. Performance of these methods will be assessed in context of bias, mean squared error, construction of confidence interval for mean, and upper tolerance limits.

4.1 Bias and Mean Squared Error

Bias is a measure of how far off the estimated parameter is from the true parameter.

Definition 5. Bias of an estimator $\hat{\theta}$, for estimating the parameter θ, is defined as $B(\hat{\theta}) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$. An estimator $\hat{\theta}$ is said to be unbiased if $B(\hat{\theta}) = 0$ i.e., if $E(\hat{\theta}) = \theta$.

Statistical bias of an estimator is the expected amount of over or underestimation done by the estimator while estimating a parameter. We want bias to be at a minimum, or nonexistent, indicating unbiasedness.

Mean squared error (MSE) of an estimator is the mean of the squared error.

Definition 6. The mean squared error (MSE) of an estimator $\hat{\theta}$ is defined as $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + [B(\hat{\theta})]^2$, where $V(\hat{\theta})$ is variance of the estimator $\hat{\theta}$.

In other words, mean squared error is a measurement of the amount of squared deviation, or squared error, the estimator displays on average. We want MSE to be as small as possible.

4.2 Confidence Intervals

Confidence intervals with a $(1-\alpha)100\%$ level of confidence give information as to where the parameter of interest could be in value. It is expected that $(1-\alpha)100\%$ of random samples will produce confidence intervals that enclose the desired parameter.

Result 1: Let $X_1, \ldots, X_n$ be a random sample from normal distribution with mean μ and standard deviation σ. Then

a) $\bar{X} = \dfrac{\Sigma_{i=1}^n X_i}{n} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$,
b) $\dfrac{(n-1)S_x^2}{\sigma^2} = \dfrac{\Sigma_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{(n-1)}$, and
c) $\bar{X}$ and S_x^2 are independently distributed.

Result 2: Let $X_1, \ldots, X_n$ be a random sample from normal distribution with mean μ and standard deviation σ. Then an approximate $(1-\alpha)100\%$ confidence interval for $\beta = \mu + \frac{\sigma^2}{2}$ is given by

$$\begin{aligned} \hat{\beta} &\pm t_{(n-1);1-\alpha/2} \cdot \sqrt{\hat{V}(\hat{\beta})} \\ \left(\bar{X} + \frac{S_x^2}{2}\right) &\pm t_{(n-1);1-\alpha/2} \cdot \sqrt{\frac{S_x^2}{n} + \frac{S_x^4}{2(n-1)}}, \end{aligned} \tag{9}$$

where $t_{m;\nu}$ is the ν^{th} percentile of a t-distribution with $(n-1)$ degrees of freedom.

Proof: From Result 1, we obtain $E(\bar{X}) = \mu$ and $E\left(\frac{(n-1)S_x^2}{\sigma^2}\right) = (n-1)$, i.e., $E(S_x^2) = \sigma^2$. Thus, $E(\hat{\beta}) = E\left(\bar{X} + \frac{S_x^2}{2}\right) = E(\bar{X}) + \frac{E(S_x^2)}{2} = \mu + \frac{\sigma^2}{2} = \beta$.

Hence, $\hat{\beta} = \bar{X} + \frac{S_x^2}{2}$ is an unbiased point estimator of β.

Also from Result 1, we obtain

$$\begin{aligned} V(\bar{X}) &= \frac{\sigma^2}{n}; \\ V\left(\frac{(n-1)S_x^2}{\sigma^2}\right) &= 2(n-1) \\ \Rightarrow \frac{(n-1)^2}{\sigma^4} V(S_x^2) &= 2(n-1) \\ \Rightarrow V(S_x^2) &= \frac{2\sigma^4}{(n-1)} \end{aligned} \tag{10}$$

Using expression (10) and noting that $\bar{X}$ and S_x^2 are independently distributed, we can easily derive variance of $\hat{\beta}$ as:

$$\begin{aligned} V(\hat{\beta}) &= V\left(\bar{X} + \frac{S_x^2}{2}\right) \\ &= V(\bar{X}) + \frac{1}{4} V(S_x^2) \\ &= \frac{\sigma^2}{n} + \frac{\sigma^4}{2(n-1)}. \end{aligned}$$

Then the estimated variance of $\hat{\beta}$ is given by $\hat{V}(\hat{\beta}) = \frac{S_x^2}{n} + \frac{S_x^4}{2(n-1)}$.

Result 3: Let $Y_1, \ldots, Y_n$ be a random sample from lognormal distribution with parameters μ and σ. Then $X_1 = \ln(Y_1), \ldots, X_n = \ln(Y_n)$ is a random sample from normal distribution with mean μ and standard deviation σ. Then an approximate $(1-\alpha)100\%$ confidence interval for mean of the lognormal distribution $E(Y) = \exp\left(\mu + \frac{\sigma^2}{2}\right) = \exp(\beta)$ is given by $(\exp(\hat{\beta}_L), \exp(\hat{\beta}_U))$, where

$$\begin{aligned} \hat{\beta}_L &= \left(\bar{X} + \frac{S_x^2}{2}\right) - t_{(n-1);1-\alpha/2} \cdot \sqrt{\frac{S_x^2}{n} + \frac{S_x^4}{2(n-1)}} \\ \hat{\beta}_U &= \left(\bar{X} + \frac{S_x^2}{2}\right) + t_{(n-1);1-\alpha/2} \cdot \sqrt{\frac{S_x^2}{n} + \frac{S_x^4}{2(n-1)}} \end{aligned}$$

Proof: From Result 2, we obtain an approximate $(1-\alpha)100\%$ confidence interval for β as $(\hat{\beta}_L, \hat{\beta}_U)$. Hence, by exponentiating we obtain an approximate $(1-\alpha)100\%$ confidence interval for $\exp(\beta)$ as $(\exp(\hat{\beta}_L), \exp(\hat{\beta}_U))$.

4.3 Upper Tolerance Limit

An Upper Tolerance Limit (UTL) provides more information about a population than confidence intervals. Confidence intervals provide information about just one population characteristic or parameter. UTLs indicate that p percent of the population lies below the UTL with confidence level $(1-\alpha)$.

For a normal distribution with mean μ and standard deviation σ, the $100p^{th}$ percentile (or quantile) is given by

$$q_p = \mu + z_p\sigma,$$

where z_p is $100p^{th}$ percentile of the standard normal distribution.

Definition 7. A $(1-\alpha)100\%$ upper confidence limit for q_p is defined as a $(p, 1-\alpha)$ one-sided upper tolerance limit (UTL) for the normal population, where p is the content and $(1-\alpha)$ is the coverage. In other words, with $(1-\alpha)100\%$ confidence we say that $100p\%$ of the normal population lies at or below the UTL [5].

Result 4: Let the random variable Z follow a standard normal distribution and be independent of X, which follows a Chi-square distribution with m degrees of freedom. Then

$$T = \frac{Z+\delta}{\sqrt{X/m}} \sim t_m(\delta),$$

i.e, T follows a non-central t distribution with m degrees of freedom and constant non-centrality parameter δ.

Result 5: Let $X_1, \ldots, X_n$ be a random sample from normal distribution with mean μ and standard deviation σ. Then $(p, 1-\alpha)$ one-sided upper tolerance limit (UTL) is given by

$$\bar{X} + \frac{1}{\sqrt{n}} t_{n-1;1-\alpha}(\sqrt{n}z_p) \cdot S_x, \tag{11}$$

where $t_{m;\nu}(\delta)$ is the ν^{th} percentile of a non-central t-distribution with non-centrality parameter δ and $(n-1)$ degrees of freedom.

Proof: Assume that $(p, 1-\alpha)$ upper tolerance limit is of the form $\bar{X} + k \cdot S_x$, where k is called the tolerance factor and $\bar{X}$ and S_x are the mean and standard deviation respectively of the random sample. Then the tolerance factor k needs to be determined in such a way that at least p proportion

of the population measurements are less than $\bar{X} + k \cdot S_x$ with confidence $(1 - \alpha)$. Therefore,

$$\begin{aligned} P_{\bar{X},S_x}\{P(X < \bar{X} + k \cdot S_x \mid \bar{X}, S_x) \geq p\} &= 1 - \alpha \\ \Rightarrow P_{\bar{X},S_x}\left\{P\left(\frac{X-\mu}{\sigma} < \frac{\bar{X}-\mu}{\sigma} + \frac{k \cdot S_x}{\sigma} \mid \bar{X}, S_x\right) \geq p\right\} &= 1 - \alpha \\ \Rightarrow P_{\bar{X},S_x}\left\{P\left(Z < \frac{\bar{X}-\mu}{\sigma} + \frac{k \cdot S_x}{\sigma} \mid \bar{X}, S_x\right) \geq p\right\} &= 1 - \alpha \\ \Rightarrow P_{\bar{X},S_x}\left\{\Phi\left(\frac{\bar{X}-\mu}{\sigma} + \frac{k \cdot S_x}{\sigma}\right) \geq p\right\} &= 1 - \alpha, \end{aligned} \tag{12}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Now,

$$\begin{aligned} \Phi\left(\frac{\bar{X}-\mu}{\sigma} + \frac{k \cdot S_x}{\sigma}\right) &\geq p \\ \Rightarrow \left(\frac{\bar{X}-\mu}{\sigma} + \frac{k \cdot S_x}{\sigma}\right) &\geq z_p \\ \Rightarrow \frac{\bar{X}-\mu}{\sigma} - z_p &\geq -\frac{k \cdot S_x}{\sigma} \\ \Rightarrow \frac{\frac{\bar{X}-\mu}{\sigma} - z_p}{S_x/\sigma} &\geq -k. \end{aligned} \tag{13}$$

So, from equations (12) and (13), we obtain

$$\begin{aligned} P_{\bar{X},S_x}\left\{\frac{\frac{\bar{X}-\mu}{\sigma} - Z_p}{S_x/\sigma} \geq -k\right\} &= 1 - \alpha \\ \Rightarrow P_{\bar{X},S_x}\left\{\frac{-\left(\frac{\bar{X}-\mu}{\sigma}\right) + Z_p}{S/\sigma} \leq k\right\} &= 1 - \alpha. \end{aligned} \tag{14}$$

Note: If $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, then $-\bar{X} \sim N(-\mu, \frac{\sigma}{\sqrt{n}})$. This implies that $\frac{\bar{X}-\mu}{\sigma} \sim N(0, \frac{1}{\sqrt{n}})$ and $-\left(\frac{\hat{X}-\mu}{\sigma}\right) \sim N(0, \frac{1}{\sqrt{n}})$, i.e., both $\frac{\bar{X}-\mu}{\sigma}$ and $-\left(\frac{\bar{X}-\mu}{\sigma}\right)$ are iden-

tically distributed. So continuing from equation (14), we get:

$$\begin{aligned} P_{\bar{X},S_x}\left\{\frac{\frac{\bar{X}-\mu}{\sigma}+z_p}{S_x/\sigma}\leq k\right\} &= 1-\alpha \\ \Rightarrow P_{\bar{X},S_x}\left\{\frac{\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right)+\sqrt{n}z_p}{S_x/\sigma}\leq \sqrt{n}k\right\} &= 1-\alpha \\ \Rightarrow P_{\bar{X},S_x}\left\{\frac{\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)+\sqrt{n}z_p}{S_x/\sigma}\leq \sqrt{n}k\right\} &= 1-\alpha. \end{aligned} \tag{15}$$

Note: From Result 1, we have that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $\frac{S_x^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n-1}$ and they are independently distributed. Hence, from Result 4, we can say that $\frac{\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)+\sqrt{n}z_p}{S_x/\sigma} \sim t_{n-1}(\sqrt{n}z_p)$, i.e., it follows a non-central t-distribution with $n-1$ degrees of freedom and non-centrality parameter $\sqrt{n}z_p$. Therefore, equation (15) becomes:

$$\begin{aligned} & P(t_{n-1}(\sqrt{n}z_p)\leq \sqrt{n}k) = 1-\alpha \\ \Rightarrow\ & \sqrt{n}k = t_{n-1;1-\alpha}(\sqrt{n}z_p) \\ \Rightarrow\ & k = \frac{1}{\sqrt{n}}t_{n-1;1-\alpha}(\sqrt{n}z_p). \end{aligned}$$

Hence, we get the $(p, 1-\alpha)$ upper tolerance limit of the normal population as

$$\bar{X} + \frac{1}{\sqrt{n}}t_{n-1;1-\alpha}(\sqrt{n}z_p)\cdot S_x. \tag{16}$$

Result 6: Let $Y_1, \ldots, Y_n$ be a random sample from lognormal distribution with parameters μ and σ. Then $X_1 = \ln(Y_1), \ldots, X_n = \ln(Y_n)$ is a random sample from normal distribution with mean μ and standard deviation σ. Then a $(p, 1-\alpha)$ upper tolerance limit of the lognormal distribution is obtained by simply exponentiating the $(p, 1-\alpha)$ upper tolerance limit for the corresponding normal distribution, i.e.,

$$UTL_{lognormal} : \exp[\bar{X} + \frac{1}{\sqrt{n}}t_{n-1;1-\alpha}(\sqrt{n}z_p)\cdot S_x]. \tag{17}$$

5 Simulation Results

In this section, we rank different methods for dealing with left-censored data using various assessment criterion discussed in the previous section. Ideally, we want the methods being studied to have a bias as close to 0 as possible, MSE as small as possible, and we want them to produce confidence intervals and tolerance limits with true/actual coverage very close to the pre-specified level of confidence.

Simulations for this project were performed in R using the NADA package [6] with the following algorithm:

1. Generate a random sample of size 50 from a lognormal distribution with known parameters.
2. Simulate censoring by using a particular theoretical percentile of the sample data as a detection limit.
3. Transform lognormal sample into normal sample by the natural logarithmic transformation.
4. Perform various methods in order to find point estimates or confidence interval or upper tolerance limits.
5. Then re transform this quantities by exponentiating them.
6. Repeat steps 1-5 a large number of times (say 10000 times).
7. Then using the known parameters, compute average error or squared error as estimates of bias and MSE, respectively. Also compute proportion of confidence intervals or upper tolerance limits containing true value of the parameters to estimate the actual coverage.

When simulating censoring, we used theoretical percentiles of our sample data. For example, to simulate a dataset with 40% non-detects, we considered any value falling below the theoretical 40^{th} percentile of the data as a non-detect. Our simulations cover multiple censored percentages (between 5% and 60%). These ranges were chosen in accordance with industry standards.

Figures 1 and 2 display the results of the eight different methods of working with censored data used on simulated lognormal samples containing varying percentages of censored observations. Figure 1 shows the estimated bias and MSE values. The method of substituting with 0 results in negative bias, while the methods of substituting with $\frac{DL}{\sqrt{2}}$ and DL, as well as

the Kaplan-Meier and Maximum Likelihood Estimation methods result in positive bias. The methods of substituting with $\frac{DL}{2}$, values evenly spaced between 0 and the DL, and random values result in approximately unbiased estimates for all percentages of censored values. The graph of the estimated MSE values display that only the Maximum Likelihood Estimation method produces undesirable values of MSE. The remaining seven methods all produced small MSE values up to about 50% of the data containing censored observations. After this percentage of censored observations, the methods of substituting with $\frac{DL}{2}$, values evenly spaced between 0 and the DL, and random values between 0 and the DL produce the smallest estimates of MSE. Figure 2 shows the actual coverage of 95% confidence intervals of the lognormal mean and $(0.9, 0.95)$ upper tolerance limits. The methods of substituting with $\frac{DL}{2}$ and $\frac{DL}{\sqrt{2}}$ produce the most efficient results in regards to actual coverage for up to about 30% of the data being made up of non-detects. These methods produce the most confidence intervals and UTLs with coverage close to the pre-specified 95% level of confidence. Maximum Likelihood Estimation and substituting with 0, evenly spaced values, and random values between 0 and the DL are too conservative, meaning that the confidence intervals are too wide, and the UTLs are too high to be of any practical value.

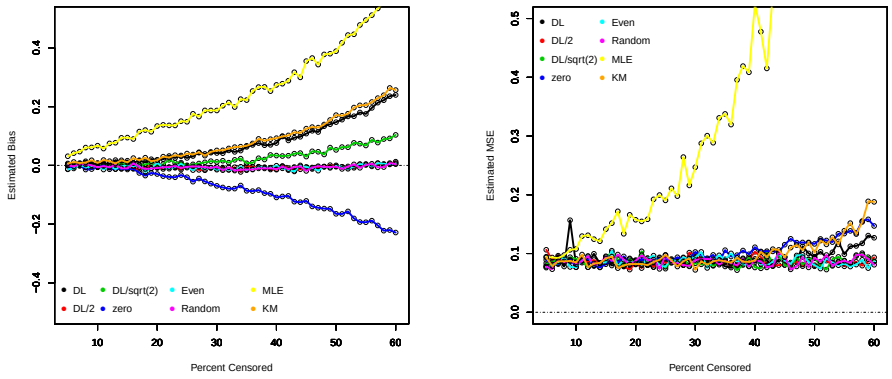

Figure 1: Estimated Bias and MSE of lognormal mean with varied Percentage of non-detects

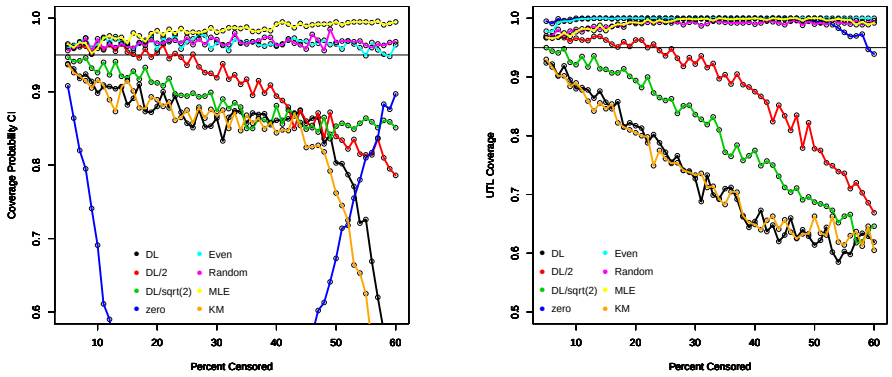

Figure 2: Actual coverage of 95% confidence interval of lognormal mean and $(0.9, 0.95)$ upper tolerance limits.

6 An Example

The following data was provided by Neptune and Company, Inc [8]. These are measurements for the concentrations of Arsenic in soil samples taken at different locations across a site that is undergoing clean up after being polluted. Concentrations are reported in mg/kg. The detection limit of the measuring device is 5 mg/kg, i.e., concentrations of Arsenic below this value are not detected. Table 3 displays the 66 observations.

< 5	< 5	5.6	< 5	< 5	6.7	5.5	< 5	9.9	5.3	5.2
6.5	< 5	7.3	5.2	5.5	5.8	< 5	6.7	< 5	< 5	6.4
< 5	5.3	< 5	< 5	5.5	< 5	5.6	5.1	5.2	< 5	< 5
5.4	8.5	< 5	6.3	5.8	5.1	< 5	6.9	5.4	6.1	5.9
5.3	< 5	6.5	8.8	6.8	7.8	6.3	8.8	7.3	7.4	8.5
6.4	6.0	6.7	7.9	< 5	6.4	< 5	< 5	6.3	9.4	8.0

Table 3: Arsenic concentrations in soil samples.

In this sample, about 32% of values are non-detectable, or left-censored. Now we need to verify if lognormal model is a good fit to the data. To do this, we use a probability (quantile-quantile) plot, where the sample quantiles are estimated by the Regression on Order Statistics (ROS) method. From the probability plot in figure 3, it looks like the data fit a lognormal distribution reasonably well.

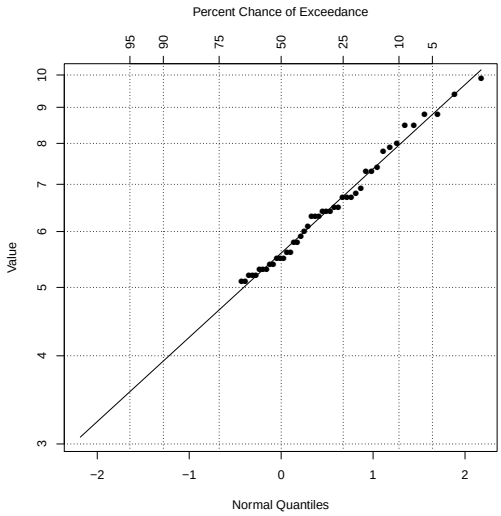

Figure 3: Probability plot to assess lognormality of the data using ROS method.

The eight methods were then performed on this data set. The mean, lower and upper confidence interval limits, and UTLs are displayed in table 4 for each method. In correspondence to the simulated samples, substituting with 0 results in the smallest mean value (giving negative bias) and substituting with the DL results in one of the largest mean values (giving positive bias). Also supporting the information found from the simulated samples, the UTL for the methods of substituting with 0, evenly spaced and random values between 0 and the DL are much too conservative, as the values are even larger than the maximum value in the data set. If this data set's results are analyzed using the method of substituting with one of the effective methods discovered in this project, $\frac{DL}{\sqrt{2}}$, the estimate of the mean arsenic concentration in soil is $5.584 mg/kg$. Also, 95% of arsenic concentrations in soil are estimated to be between $5.162 mg/kg$ and $6.0592 mg/kg$. In addition, with 95% confidence, 90% of arsenic concentrations in soil are estimated to be below $8.8184 mg/kg$.

Methods	Mean	Lower limit	Upper limit	UTL
zero	4.4591	4.0719	6.817	14.5769
DL	6.05	5.7667	6.3371	8.0344
$DL/2$	5.2545	4.7061	5.9974	10.0265
$DL/\sqrt{2}$	5.584	5.162	6.0592	8.8184
Even	5.2545	4.7199	7.0391	14.0225
Random	5.0973	4.7046	8.2888	18.0703
MLE	5.8179	5.4573	6.2023	8.4625
KM	6.0818	5.805	6.3632	8.0215

Table 4: Results from Arsenic data analysis.

7 Recommendations and Future Work

Based on the four criterion for selecting the most effective method for working with censored data, it is recommended that the methods of substituting the non-detects with the values $\frac{DL}{2}$ and $\frac{DL}{\sqrt{2}}$ be used for data with a percentage of censored observations up to 30%. These are not the expected results, as the Maximum Likelihood Estimation in general is the most reliable method for dealing with censored data. However, these results are consistent with what is often used in the industry or environmental sciences. Methods of substituting with varying fractions of the DL are commonly used in the industry, because they produce fairly good results without having to specify a certain distribution for the sample. Using a parametric method, such as Maximum Likelihood Estimation, depends on having a sample that is a good fit for that specified distribution, which is very difficult to do in practice.

This research will be continued to determine the most effective method when dealing with multiple detection limits, as opposed to a singular DL. Further explanation as to why the Maximum Likelihood Estimation method may not have produced the expected results will also be explored. The analysis of methods when dealing with multiple detection limits will be very applicable to environmental sciences, as having multiple detection limits is a common occurrence in the environmental sciences when different measuring devices are used.

References

[1] Ronald C Antweiler and Howard E Taylor. Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. summary statistics. *Environmental science & technology*, 42(10):3732–3738, 2008.

[2] U.S. EPA. Definition and procedure for the determination of the method detection limit, revision 1.11. *US Environmental Protection Agency. Code of Federal Regulations*, 40, 1984.

[3] Dennis R Helsel. *Statistics for censored environmental data using Minitab and R*, volume 77. John Wiley & Sons, 2011.

[4] Paul Hewett and Gary H Ganser. A comparison of several methods for analyzing censored data. *Annals of Occupational Hygiene*, 51(7):611–632, 2007.

[5] K. Krishnamoorthy and T. Mathew. *Statistical Tolerance Regions: Theory, Applications, and Computation*. John Wiley & Sons, 1 edition, 2009.

[6] Lopaka Lee. *NADA: Nondetects And Data Analysis for environmental data*, 2013. R package version 1.5-6.

[7] Jay H Lubin, Joanne S Colt, David Camann, Scott Davis, James R Cerhan, Richard K Severson, Leslie Bernstein, and Patricia Hartge. Epidemiologic evaluation of measurement data in the presence of detection limits. *Environmental health perspectives*, pages 1691–1696, 2004.

[8] Neptune and Company Inc. Private Communication. 2015.

[9] Nian She. Analyzing censored water quality data using a non-parametric approach1. *JAWRA Journal of the American Water Resources Association*, 33(3):615–624, 1997.