Implications of Left-Censored Environmental Data Due to Labeling "Non-Detects"

LAURIE RUGEMER

Department of Mathematical Sciences
Montana State University

May 3, 2019

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Laurie Rugemer

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date	Megan Higgs, PhD Writing Project Co-Advisor
Date	Andrew Hoegh, PhD Writing Project Co-Advisor
Date	Mark C. Greenwood Writing Projects Coordinator

Contents

1 Abstract

Estimating mean concentrations of harmful analytes in the environment is an important and often completed task taken on by risk assessors, researchers, and statisticians working in the environmental field. Often these data include very low concentrations of analytes that are considered to be below the level that a laboratory instrument can reliably detect and the raw instrument reading is censored at a specified threshold (or detection limit). This paper focuses on understanding how detection limits arise, along with common methods used to deal with these censored readings in a data analysis. There is no "best" way to deal with detection limits and censored observations of this nature, and methods that may be reasonable in one situation may not be in another. The takeaway is that there should not be a default method for approaching any analysis with left-censored observations due to laboratory decision-making and instrument calibration. An appropriate, reasonable method depends on many factors, including the proportion of censored values in the data and their location in relation to the observed values. We, as statisticians, should continue to educate others about the loss of information when observations are censored at a detection limit and, when possible, build relationships that might allow us to be involved pre-processing to minimize the amount of information that is lost.

2 Introduction

2.1 Background

Many processes of the modern world like mining, oil extraction, and manufacturing leave behind chemical waste that, even in very small quantities, has potentially very harmful impacts on the health of the environment. As a necessary consequence of these processes, samples are being collected to measure the concentrations of these concerning chemicals and make vital decisions regarding their remediation. Risk assessments are undertaken that evaluate if it is necessary to clean-up an area where their has been, for example, a man-made disaster that left concentrations of harmful chemicals in surface water, ground water and/or sediment.

One example is the Love Canal disaster where a neighborhood outside of Niagara Falls, NY was built on top of a chemical dump site in the 1950s and for many years chemicals from the drums seeped into the soil and ground water, causing major health problems for the residents. Eventually this area was declared a superfund site and Environmental Protection Agency (EPA) and New York State Department of Environmental Conservation (NYSDEC) remediation efforts ensued (Beck 1979).

Another example, and one I will examine more later, is the Deep Water Horizon oil disaster of 2010 where an

oil drilling rig in the Gulf of Mexico exploded, killing 11 workers and dumping 4 billion gallons of oil into the ocean over 87 days (EPA, n.d.). Various organizations (private, federal, state and academic) have been monitoring the water and sediment in surrounding areas for chemicals that are at potentially high levels for human and environmental health (NOAA, n.d.).

These are just two examples of chemical processes and waste gone wrong that require sample collection and analysis in order to understand the concentrations of these chemicals and make decisions that will impact human and ecological health. I start with these because the small decisions we make as statisticians can have far-reaching consequences down the line, and it is often the case in environmental analyses such as these that we are dealing with chemicals that can be harmful in very small amounts.

2.2 Motivation

It is often the case in environmental studies that researchers, risk assessors, statisticians, etc... are interested in understanding the concentrations of certain compounds in an area. Samples are collected and sent to a laboratory in order to get the concentration values. Often in laboratory analyzed samples, where there are potentially very low concentrations of a compound of interest, there may be concentrations that are so low that they are considered to be below the level that a laboratory instrument can reliably detect. The instrument sometimes gives a value and sometimes doesn't, but laboratories lack confidence that these instrument readings correctly reflect a non-zero concentration. As a result, the raw value associated with these observations is often not recorded or provided in results from the lab, but instead the detection limit (DL) is reported or provided (e.g. provided in the data set as "< DL"). Assuming the DL is greater than the actual concentration in the sample, this creates left censored data (censored at the left side, or lower values) coming from the lab, and the researcher, risk assessor, etc... will receive the data with a combination of observed (detected) readings and censored readings. The process for specifying a detection limit differs greatly depending on the guidelines set by the laboratory (D. R. Helsel 2005). Analysts and researchers often refer to these censored observations as "non-detects". This paper focuses on understanding left-censored data in scenarios like these and how they generally arise.

It is first important to reframe how we discuss these instrument readings and why labeling them as "non-detects" may make it easier for users of the data to disregard them as unimportant pieces of information, when in reality they were possibly detected but were below a detection limit, and the data were then censored after that. This censoring at the start makes subsequent data analyses more complicated and great thought must be put into how one deals with the censored values relative to the specific goal of analyzing the entire

dataset. Instead of calling them "non-detects", it would be more valuable to discuss them as "readings censored at a detection limit" or "censored instrument readings".

This process of censoring instrument readings at a detection limit means that information is intentionally lost before any data analysis can be performed, and we assume that the measured concentration is between 0 and the DL and hence left-censored. How this lost information is dealt with and incorporated into the analysis (as we will see later) can drastically alter the results.

Many statisticians argue that statistical tools have been developed to deal with uncertainty and that it would be more beneficial to give the researcher all of the available information, including the instrument readings below the detection limit, than to categorize instrument readings into "non-detects" and "detects" using a DL definition. As Millard, Najara and Dixon state in their yet unpublished book with a chapter dedicated to censored data of this type, "as an aside, it seems very silly to us as statisticians to be making up values that an analytic chemist has measured but not reported" (Millard Unpublished, 16). The field of analytical chemistry does not trust the measurement error at low concentrations of a chemical, and the system of reporting is set up to protect against the error of providing a non-zero concentration when in fact the analyte is not present in the sample. For many years statisticians have been arguing that not providing the full dataset is detrimental to analyses and eventual decision making and that it would be better to give all of the instrument measurements and a measurement precision (Porter 1988; Millard Unpublished; Gilbert 1987). There can still be measurement error included in the variability quantified in the statistical analysis. The problem may be in the magnitude of measurement error relative to the concentrations. How well can we dichotomize these observations into detect or non-detect given the error around the DLs *and* error in dichotomizing based on the DLs?

2.3 Project Goal

If an agency or risk assessor is investigating environmental remediation at a site, their protocol may be to estimate a mean concentration of a chemical, calculate the upper confidence limit (UCL) of the mean concentration, and ultimately compare that UCL to a specified cutoff value in order to make remediation decisions. I acknowledge this is a simplification of the process leading to the point that a 95% UCL is compared to a cutoff value, but for the purposes of this paper I am not going to explore the steps in-depth, nor will I spend time discussing the potential issues with using a UCL in this way. I focus on estimating a mean with a 95% UCL because of its common use in practice and because this provides a tangible example tied to decision making that can be greatly sensitive to how censored data are treated in the analysis. I am

not, however, recommending this approach for mean and UCL estimation.

Calculation of a UCL involves not only the estimated mean concentration of a compound, but also the standard deviation from the sample observations. If observations censored at a detection limit are taken out or replaced with a number (like the detection limit), the estimated mean, standard deviation, and therefore the UCL, may not be trustworthy (not that these are always trustworthy with completely uncensored data). For example, even if the estimated mean concentration is greater than it should be, the standard deviation may be smaller than if all information was available and therefore the UCL calculated may be smaller than it would be with all information. If this is compared to a threshold value for human and ecological health and is erroneously below this cutoff, then the remediation may not move forward when it should (Wendelberger 1994).

Because risk assessment decision making can largely be based on one numerical summary (a 95% UCL), it is important that data analysts and researchers think very carefully about how this number is calculated, what it represents and how sensitive it is to decisions and assumptions (and generally about how statistical analyses are done). This translates to thinking carefully about how instrument readings censored at a detection limit (labeled as "non-detects") are treated. My goal is to examine how and why detection limits occur, as well as potential sensitivity in 95% UCL results based on common decisions. I will motivate this through a brief comparison of common methods for dealing with data of this type through examination of a Deep Water Horizon sediment dataset to understand how choices regarding these censored readings can drastically change results.

3 Detection Limits

3.1 Laboratory Instrument Readings

It is an accepted part of laboratory analytics that low instrument readings of an analyte in a sample may actually be a zero concentration that the laboratory instrument is incorrectly reading as an observed concentration. Generally the laboratory analyzing the samples has specified a "detection limit" under which instrument readings of a chemical are considered to be too small to confidently conclude a non-zero instrument reading as present in the sample and are therefore censored in the dataset given to researchers and generally labeled as "non-detect" (Lampert 1991). The dataset will usually have "< 5", for example, in a cell. The laboratory wants to guard against reporting reporting an analyte is present in a sample when in fact it is not. These data are not being used for presence/absence conclusions; they are being used as continuous

measurements.

Detection limits can have many definitions and take many forms (e.g. quantification limit, reporting limit, method detection limit), but the focus of this report is not the various definitions or nomenclature around them, but rather the implications in using them to censor instrument readings. Generally, the detection limit controls for false positive but also allows for false negatives. It is generally assumed that any concentration not picked up by an instrument is below the detection limit, which is often the case (D. R. Helsel 2005).

3.2 Defining and Estimating Detection Limits

The specification of a detection limit takes many forms. The common thread is that the laboratory asserts that "values measured above this threshold are unlikely to result from a true concentration of zero" while values below it are "not considered significantly different from a blank signal at a specified level of probability" (D. R. Helsel 2005, 22).

Daniel Helsel provides an explanation for one method of setting the detection limit in his book *Nondetects and Data Analysis* (2005, p.22). A laboratory may use solutions with known low concentrations of a chemical and take repeated measurements (5 is common) to find the approximate variation around a "true" concentration, the idea being that this variation will be the same for true concentrations of 0. The laboratory must assume, then, that "the measurement error at zero concentration is the same as the low standard concentration - the standard deviation is constant between zero and the concentration standard" (p.22). Once the standard deviation is approximated, the laboratory assumes the concentration is distributed normally. If this was done for a known low concentration (of 2, for example), then the normal distribution is simply shifted from a center of 2 to a center of 0 and the detection limit is set near the "upper end" of the distribution centered at 0. "The choice of a distance to represent the detection limit is made so that no more than a small percentage of the measured values truly originating from a zero concentration will fall above the limit" (D. R. Helsel 2005, 23).

Lambert, Peterson and Terpenning (1991) investigate the Love Canal study (mentioned earlier), a study in 1988 investigating pollutants in the soil using (at the time) "state of the art" mass spectroscopy (p.267) (Lampert 1991). In order to estimate the detection limit for the samples, the analysts put known levels of a pollutant into "clean soil" (soil assumed to be free of contaminants) and analyzed them to see what value the instrument would read. Then they regressed the concentrations from the instrument on the "true" concentrations (which they assume they know accurately because they put known amounts into what they believe to be soil free of the pollutant) and plotted the fitted regression line with 95% confidence interval bands assuming normality. They then took the 95% upper confidence limit at the 0 concentration and this

became the detection limit (Lampert 1991).

The methods for specifying or estimating detection limits vary depending on the laboratory and guiding body. The EPA, for example, published revised guidelines in 2016 regarding the setting of method detection limits (MDLs), which they define as the "minimum measured concentration of a substance that can be reported with 99% confidence that the measured concentration is distinguishable from method blank results" (EPA 2016, 1). Essentially they are referring to calculating the upper confidence limit of a one-sided 99% confidence interval using a multiplier from the t-distribution and the sample standard deviation from the "replicate method blank sample analyses" (EPA 2016, 3).

3.2.1 No rigid threshold

Lambert, Peterson, and Terpenning (1991) inform us that non-detectable concentrations can occur at higher levels as well. In the case of the Love Canal study (1988), and in this case when researchers used "spiked" samples to calibrate the machine, even with very high concentrations at times the machine did not always pick up the chemical. This is an instance where the instrument reading would have been censored at the detection limit and assumed to be below this value, but the actual concentration was greater than the detection limit. Conversely, they also found that it can be the case that because of interfering compounds in the sample, some measurements below the detection limit were measured as detects. "Because soil constituents and interfering compounds vary among field samples, there can be no strict threshold that separates non detects and detects" (Lampert 1991, 267). The detection limit, therefore, "is not an appropriate summary of the limitations of environmental data and the practice of assuming that all non detects are below the detection limit, which is common, is unwise" (Lampert 1991, 275).

These are just a few examples of how detection limits are specified or estimated; the key is that the procedure depends very much on the laboratory analyzing the samples and it is important understand how the detection limits were specified and the ambiguity in them. Ideally, statisticians would be able to communicate with the laboratory and advocate for the release of the DLs *and* information (meta-data) for how they were specified/estimated, along with the instrument readings. The takeaway is that it is important to educate researchers and analysts about this issue and continue to build relationships with the laboratory analysts if possible, ultimately advocating for the retention of as much information as possible and the documentation of any decisions made along the way that impact the data that are ultimately given to researchers.

3.3 Multiply Censored

It is also important to note that there may be more than one detection limit for a dataset, as "the level of censoring is determined by the analytical chemist and is based on the confidence with which the analytical signal can be discerned from noise. Samples taken over time may be censored at different levels as changes in analytical technology alter the precision of a method" (Porter 1988, 857). Helsel also mentions this phenomenon of multiply censored data because "detection limit thresholds change over time or with varying sample characteristics or among different laboratories" (D. R. Helsel 2005, 16).

4 Common Mean & 95% UCL Estimation Methods

Many articles in the literature investigating analysis of data with some observations censored using detection limits focus on comparisons of methods that differ in how they incorporate the censored readings into the analysis. Generally, the literature ranks possible methods by using simulated data and comparing the estimates of summary statistics for several different methods and often looking for the "best" approach (D. Helsel 2012). My goal here is not to find the "best" approach to dealing with censored observations of this nature, but to see how different the results can be depending on the method chosen. In general it isn't realistic to think there is a "best" approach for all data sets and analytes. The course I advocate is for researchers, statisticians, data analysts, and risk assessors to ask for the raw values before any are censored so that the possibility exists for using all available information in the analysis. Certainly this is no simple feat and it may be easier to do for professionals who are working directly with a small laboratory where communication with analysts is possible. Ultimately, sensitivity of the results to different choices and methods should be assessed.

There are varying methods presented in the literature for dealing with values labeled as "non-detects". These include, but are not limited to, substitution, deletion, maximum likelihood estimation, non-parametric Kaplan Meier method, and regression on order statistics (ROS), both parametric and robust. Beyond frequentist methods, there are many Bayesian approaches that can be used, but the focus of this paper is on the more common frequentist and likelihood-based approaches. In the second edition of *Statistics for Censored Environmental Data*, Helsel includes a section reviewing much of the literature comparing varying methods, and ultimately recommends differing methods based on the number of samples, the proportion of observations labeled as "non-detects", and where those values are in relation to the uncensored values (D. Helsel 2012).

My initial goal with this project was to obtain a dataset that had both the detection limits and the censored instrument readings so that I might compare results using the data set containing all raw readings (and all

information) with results using the data set with censored readings. Antweiler and Taylor (2008) attempted a similar objective by comparing results from two datasets - both from the same samples, that were processed by two instruments differing in their sensitivity. They then presented the raw values from the more sensitive instrument as the "truth" and treated that dataset as a reference to compare the results from the dataset with censored readings based on the detection limits from the less sensitive reading. An issue here is that there is error associated with all of the instrument readings and the observations from the more sensitive analysis are not, in that sense, the "truth". Their assumption is that "even though there is analytical variability between the analyses, each was measuring the same real quantity" (Antweiler 2008, 3732). Trying to make conclusions based on differences between the two datasets isn't necessarily a good comparison to make.

4.1 Motivating Example: Deep Water Horizon Disaster

In the end, I was not able to make comparisons between results using the raw censored values because the dataset I could obtain did not adequately fit my goal of estimating a mean concentration and 95% UCL with and without the censoring. Instead, I decided to focus on the potential differences in results based on several common methods for dealing with data where some observations are censored at the detection limit and there are multiple detection limits associated with the data set. I use a real dataset of sediment samples taken at one location over time near the Deep Water Horizon disaster in the months that followed the explosion.

In April 2010, a massive explosion at the offshore oil rig killed 11 workers and sent more than 4 billion gallons of oil into the ocean before it was contained 87 days later. It's not hard to see why this would be a major concern for marine life as well as general ecological and human health. As a result, data were collected by various federal, state and academic organizations to measure the concentrations of chemicals of concern in the water and sediment. These data were obtained from the NOAA website and consist of sediment samples taken over a six month period at a site near the disaster (NOAA, n.d.). I use these data to estimate a mean concentration over the six months for that location, and associated 95% UCL, but I acknowledge that these data are time series data taken at one site over a period of time and not specifically a characterization of that site. They characterize one location over time, rather than a larger site with multiple sampling locations at one point in time. I chose them because this scenario is well known and part of my motivation for this paper is for readers to understand the real world implications of choices made in dealing with censored values, and because it is an interesting dataset in that there are multiple detection limits. I decided to concentrate on the chemical pyrene, which is included in the class of polycyclic aromatic hydrocarbons (PAHs), which are particularly carcinogenic and have generally been thought of as potentially cancer-inducing chemicals

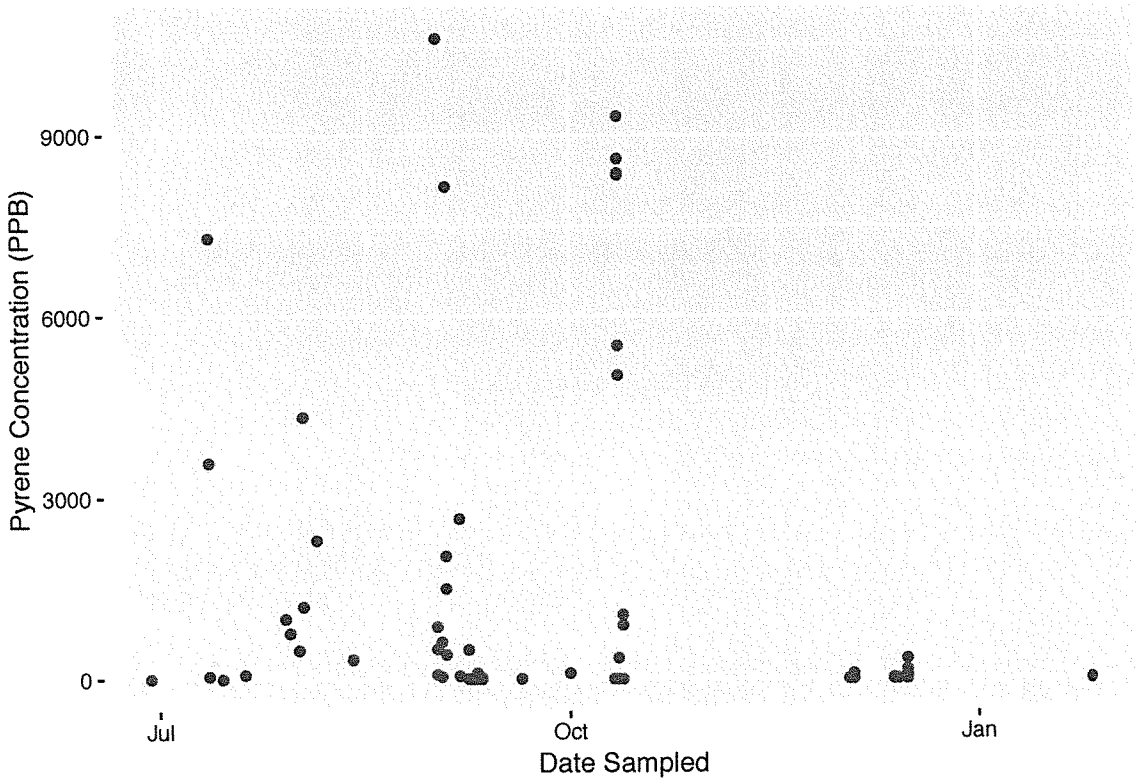

Figure 1: Pyrene concentrations from sediment at a site near the Deep Water Horizon disaster against the dates they were sampled. Samples were taken over a 6 month period from June 2010 - January 2011.

(ToxTown, n.d.). (Public Health, n.d.)

4.2 Data Exploration

These concentrations were taken at one site over a six month period (Figure 1), for a total of 131 observations. Even though these are time series data, my main goal was to use them to estimate a mean concentration and 95% UCL. Although these data should really be treated differently in a real analysis, my intention was to look at differences in these estimates based on common methods used for left-censored data of this type and not to ultimately use these estimates as results in an analysis. None of the methods compared will account for any serial autocorrelation because my goal is simply to assess the sensitivity of results to the methods.

The data are extremely right-skewed with many small pyrene concentrations along with some concentrations at much higher levels (Figure 2). As a result, a natural log transformation of the concentrations was considered moving forward, and the data are much more symmetric after transforming them using a natural log transformation (Figure 3).

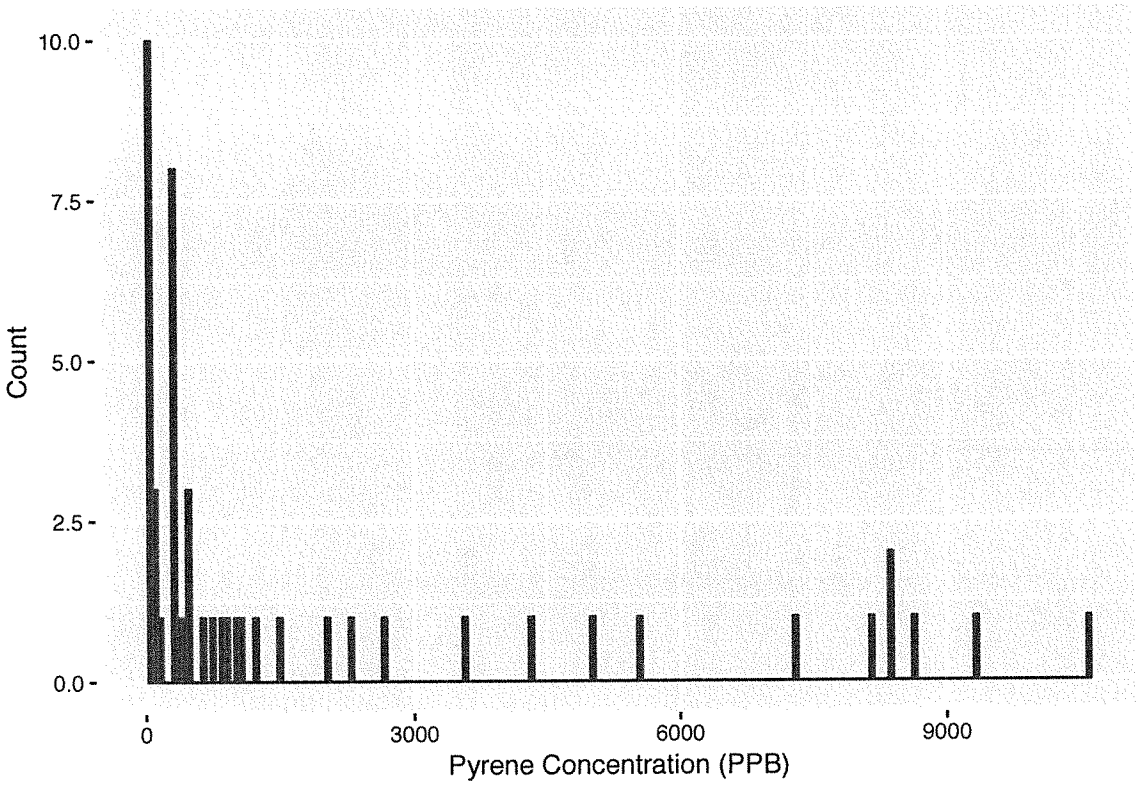

Figure 2: The distribution of pyrene concentrations from one site near the Deep Water Horizon over a six month period. Due to the right skewness of the data, a natural log transformation of the concentrations was used moving forward.

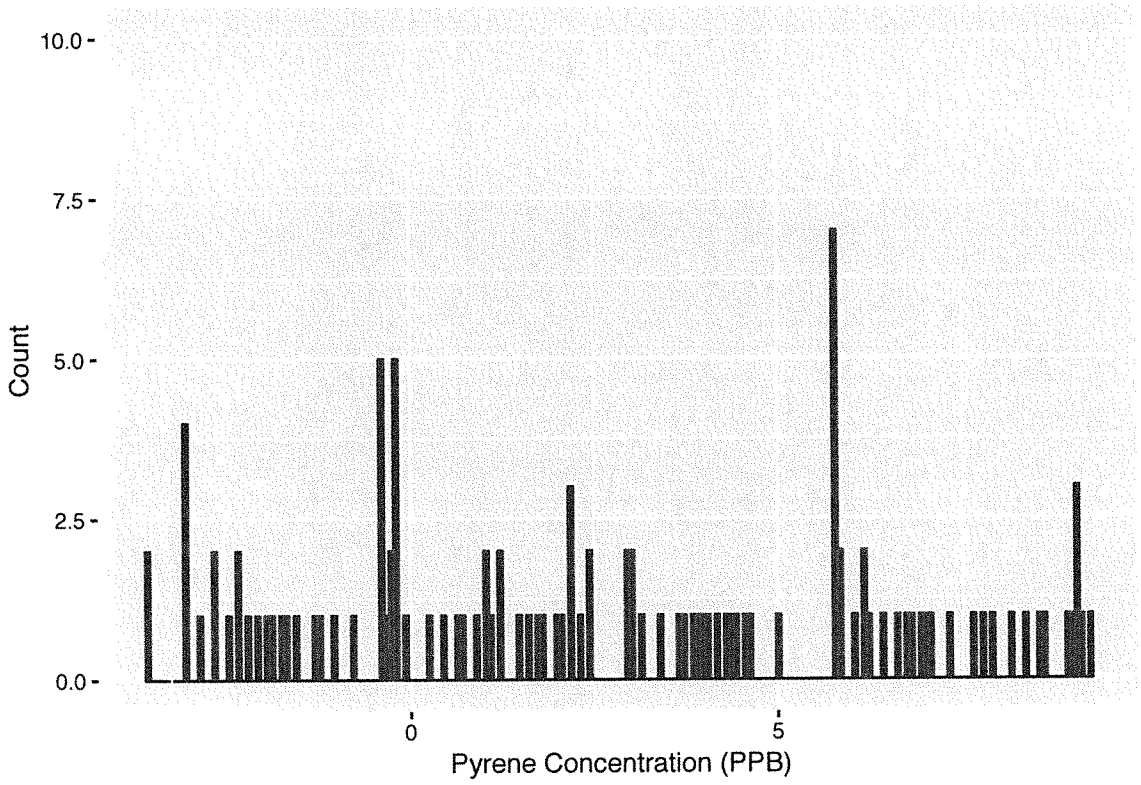

Figure 3: The distribution of ln(pyrene concentrations) from one site near the Deep Water Horizon taken over a six month period.

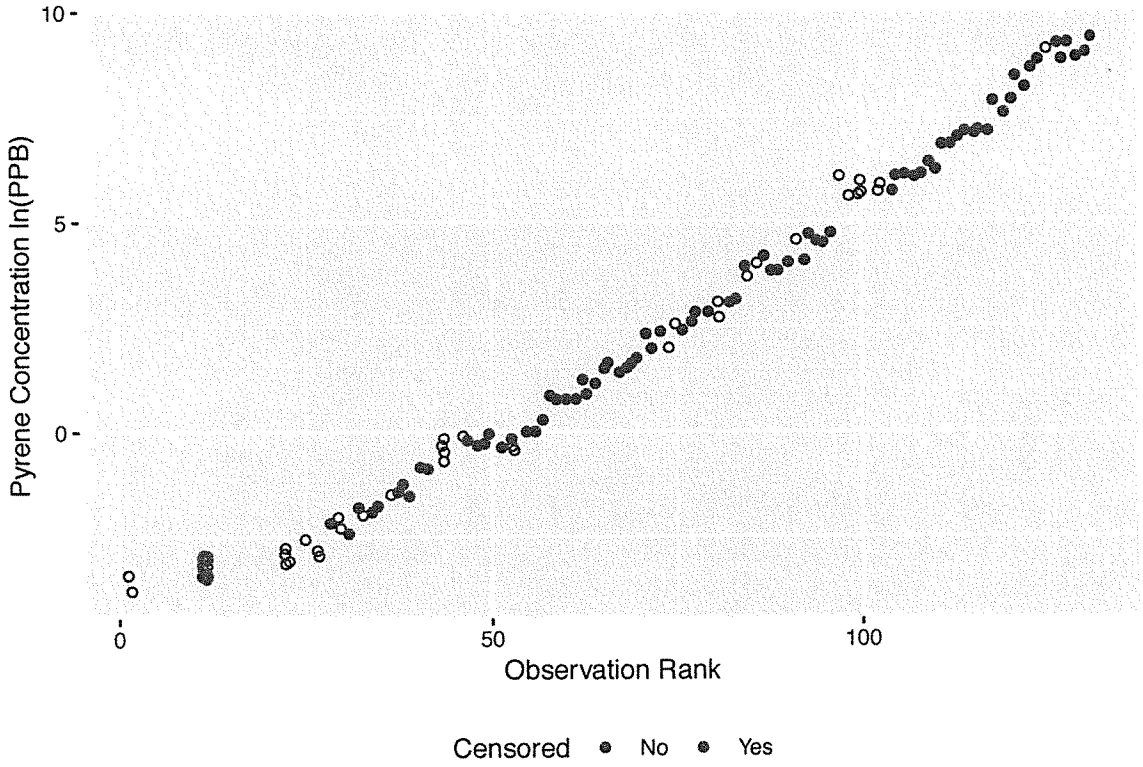

Figure 4: Pyrene concentrations in sediment near the Deep Water Horizon oil spill disaster were sampled from 7/13/10 to 1/27/11. The blue points indicate readings over the associated detection limit ('detects') while the concentrations for the red points are plotted at the stated detection limits associated with that sample ('non-detects').

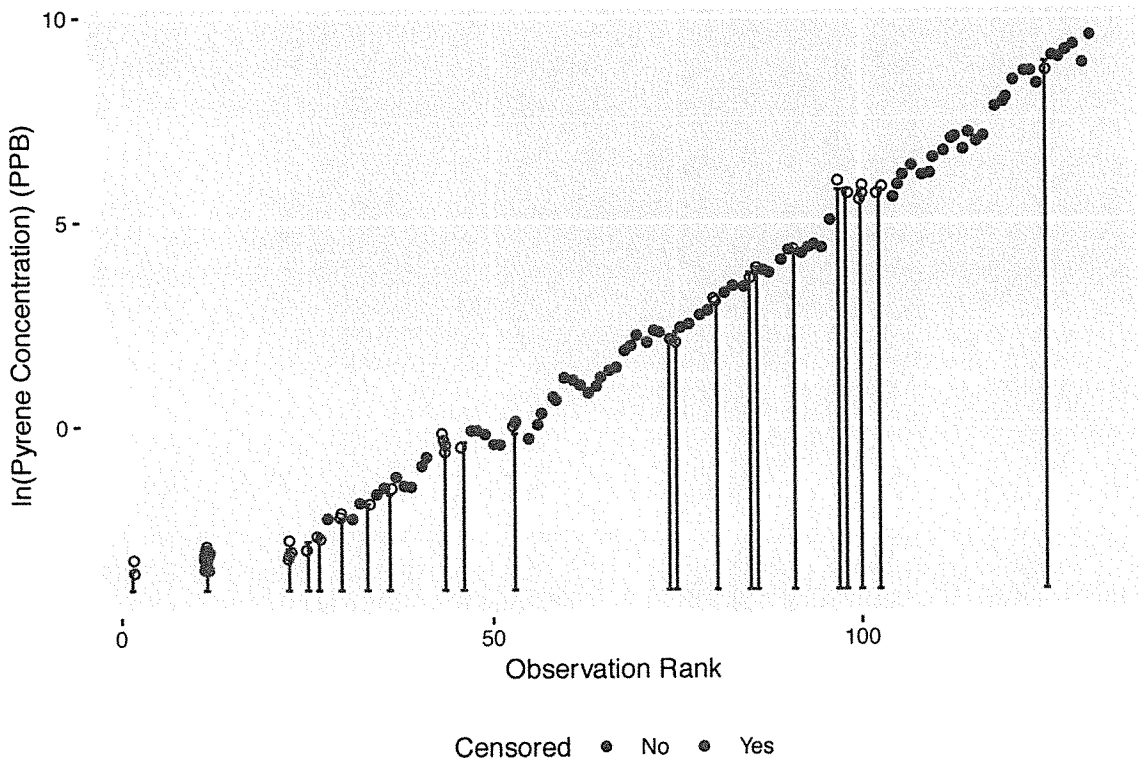

Figure 5: Pyrene concentrations in sediment near the Deep Water Horizon oil spill disaster were sampled from 7/13/10 to 1/27/11. The blue points indicate readings over the associated detection limit ('detects') while the concentrations for the red points are plotted at the stated detection limits associated with that sample ('non-detects'). The red lines indicate the interval along which the censored readings could fall.

It is easy to see that the censored values, as one might expect, are not all simply associated with a single detection limit and the detection limits associated with some observations are greater than concentrations for other uncensored values (Figure 4). There are multiple detection limits with these data and almost half (53 out of 131) of the observations are censored. A more accurate depiction of the censored data would show them as intervals, as they really could fall anywhere between 0 and the DL (Figure 5). Part of the understanding around "non-detects" is understanding that when we think of them as a point value, we diminish the fact that we really don't know enough about where they fall to assign them one value.

4.3 Common Method Comparison

Here, I investigate how the estimate of the mean of the natural log of pyrene concentrations and associated 95% UCL changes with the implementation of a few different common methods. Each method is briefly summarized, followed by a table presenting the results from all methods together to facilitate comparisons.

As with any statistical analysis, the data analyst should always understand the software they are using, any assumptions being made, and generally justify decisions made along the way. I discuss decisions in the estimation of the mean concentration and 95% UCL for each method in order to compare across results and see the differences. However, assumptions should not be taken lightly but should be given adequate thought before moving forward. My goal is not to dive too deeply into any one method, but to give an overview and use the statistical package (NADA) created by Dennis Helsel (Lee 2017) to obtain estimates and confidence intervals. My focus has been on understanding how these data come about and use this motivating example to demonstrate that common methods used can potentially give results with different practical consequences and should be considered carefully.

Sensitivity to methods is expected to depend on where the detection limits fall relative to instrument readings above the detection limits, as we see here with multiply censored data. Other factors include the proportion of the data that are censored and the number of samples (D. Helsel 2012). Because every situation is different in these and other ways, sensitivity of results depends greatly on the characteristics of an individual data set and making generalizations about which approach to use is often not useful. It is recommended to always investigate the sensitivity of results from different methods with left-censored data like these before making a decision about which method to use. And again, having all raw instrument readings would aid in this process and allow for the possibility of using all information in an analysis.

4.3.1 Substitution

There are several suggestions for values to use in place of the instrument readings for censored values (i.e., different substitutions), including inserting the detection limit (Figure 1), half the detection limit or even inserting 0 for the censored readings. These seemingly innocuous changes may have practical consequences in terms of decision using the results.

Helsel (2005) strongly advocates against any kind of substitution, as the analyst is essentially "fabricating" point values that can really fall anywhere within an interval and not incorporating the uncertainty in location of the value. Many methods can create a pattern that would not have been found in the original uncensored

data and "add artificial invasive data that create a pattern alien to the original observations" (D. R. Helsel 2005). In their most recent guide for the ProUCL software that is generally used for statistical analyses, the EPA recommends moving away from the use of substituting 1/2 the reported detection limit for the censored value (EPA 2015). However, it is worth noting this method because many researchers still view it as an easy way to incorporate values under the detection limit into the analysis and it may be the case, in some situations, that it is not worse to do this in comparison to other methods. Substitution with DL, for example, would just use the concentrations as plotted in Figure 4 without acknowledging the interval censoring. I investigated substitution with the detection limit and half the detection limit.

4.3.2 Deletion of Censored Values

A story that resonates with me, and that showed me the power of the decisions that we make as researchers, statisticians, and data analysts, is the infamous data story from the Challenger explosion in 1986. As is fairly well known, the Challenger exploded as the result of an O-ring failure due to the cold temperatures on the day of the launch. As Helsel (2005) aptly depicts in *Nondetects and Data Analysis*, engineers deleted values that were considered to be "below a damage detection threshold" (p.2) and decided that damage results at low temperatures were inconclusive and went forward with the launch. However, if they had not deleted those values they would have seen a pattern such that all of the "non-detects" occurred at temperatures above 65 degrees and they may have thought twice about launching in such cold temperatures. This story encapsulates some of the danger with our decisions regarding the treatment of values below a detection threshold set by the analyst, with deletion being a sure way to lose even more information and potentially miss important patterns in our data.

4.3.3 Maximum Likelihood Estimation

In Maximum Likelihood Estimation (MLE), it is assumed that the data (both censored and uncensored) follow a parametric distribution (such as the normal or lognormal distributions). This is a stringent assumption and generally this method will work better with a larger number of observations where the adequateness of distributional assumptions are easier to assess. This method uses the proportion of censored observations that are below each detection limit, the uncensored readings, and the probability density function of an assumed parametric distribution. The likelihood function assuming a parametric distribution is used, which is the likelihood of the parameters (in this case the mean and standard deviation) given the observed data. The goal is to find values of the parameters that maximize this likelihood (the maximum likelihood estimates).

In the case of left-censored data like these, the likelihood function has two pieces; one for the censored observations, which is actually a survival function (S(x)), which is 1 - F(x), where F(x) is the cumulative distribution function of the assumed parametric distribution. The part of the likelihood function representing the uncensored data is the probability distribution of the assumed distribution.

The uncensored values and detection limits are represented by x. An indicator variable (γ) is 0 for a censored observation and 1 for an uncensored observation. $L = \Pi p[x_i]^{\gamma_i} * (F[x_i])^{1-\gamma_i}$ (D. Helsel 2012, 15). For uncensored observations, the second part of the equation becomes a 1 and for censored observations it stays in the equation. "The estimates of mean and standard deviation will be the parameters for the assumed distributional shape that had the highest likelihood of producing the observed values for the uncensored observations and the observed proportion of data given below each of the reporting limits." (D. Helsel 2012, 16). The NADA package in R, from Helsel, will estimate the mean and variance assuming a normal or lognormal distribution using the method of maximum likelihood (Lee 2017). For this example, I performed MLE using the ln(pyrene) concentrations assuming a normal distribution (instead of keeping the concentrations on the original scale and assuming they follow a lognormal distribution).

It is important when using this method to check the distributional assumption being made. In this case I am assuming the the ln(pyrene concentrations) follow a normal distribution, which is not necessarily a reasonable assumption based on the normal probability plot, as the points largely deviate from the QQ line. (Figure 6).

4.3.4 Non-Parametric Survival Analysis: Kaplan-Meier method

Unlike ML estimation, with non-parametric methods we do not need to make the assumption that the data follow a parametric distribution like the normal or lognormal distribution. Instead, the relative positions or ranks of the concentrations and detection limits are used. This can be helpful when censored data are present because if we assume that censored values are lower than their detection limit, then they will be ranked lower in position and there is no assumption of the distances between censored and uncensored observations (D. Helsel 2012).

Survival analysis is being used more in environmental research because of the presence of censored data. Typically survival analysis has been used in right-censored data situations; for example if a study is examining survival outcomes of patients with a certain disease and some of the patients are still alive at the end of the study, then their survival times are right-censored because it is unknown how much longer those people will live. The Kaplan-Meier (KM) method is often used to estimate summary statistics and confidence intervals using survival analysis and was developed for right-censored data of this type. However, if left-censored data

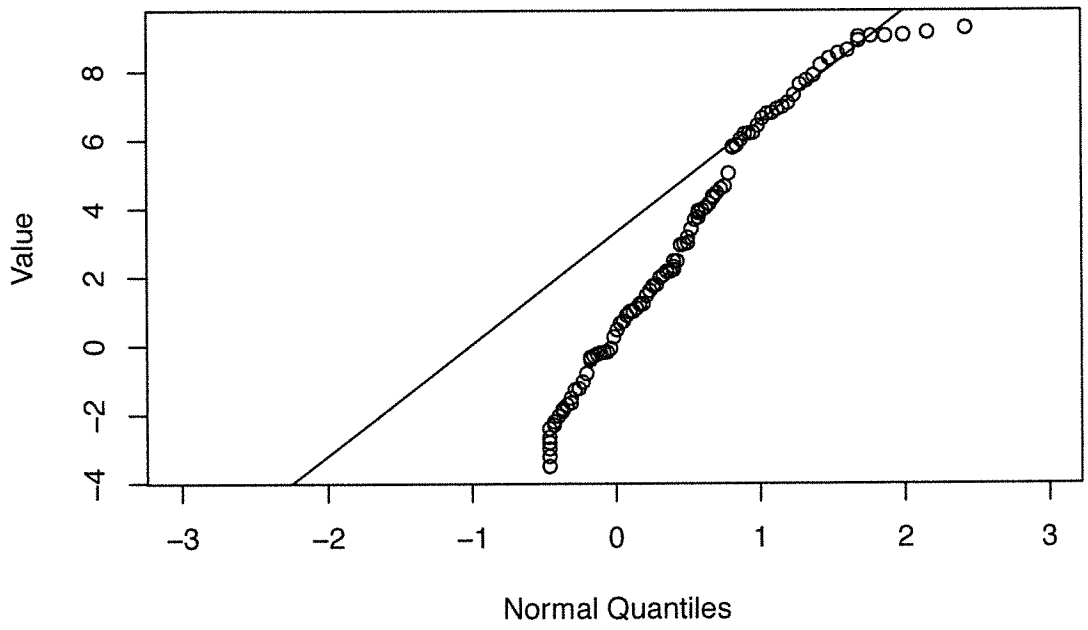

Figure 6: Normal plot of ln(pyrene concentrations). The points greatly deviate from the line and the assumption of normality needed to use MLE does not appear reasonable.

are flipped to resemble right-censored data, this method can also be used (which the NADA package does automatically when using KM) (Lee 2017).

Essentially, the KM method gives estimates of the survival probability function (S). By flipping the data to make it fit the right-censored scenario needed for KM, the uncensored observations become the "deaths" and are ranked from smallest to largest and the censored observations are accounted for in the rankings of the data (given their associated detection limit). As described by Helsel in *Statistics for Censored Environmental Data*:

"The number at risk (b) equals the number of observations, both detected and censored, at and below each detected concentration. The number of uncensored observations at that concentration is d. The incremental survival probability is the probability of"surviving" to the next lowest uncensored concentration, given the number of data at and below that concentration, or $(b - d)/d$. *The survival function probability is the product of the incremental probabilities to the point. The mean can be estimated by integrating the area under the KM survival curve." (p.74)*

Generally, the detection limit is used to represent the threshold associated with the censored values when computing the survival curve. As a result, the estimate of the mean can be higher than it should be, particularly if the large detection limits are used. So, in essence if there is only one detection limit, it gives

the same results as substituting the detection limit. With multiple detection limits (as is the case in this example), the substitution is at the smallest detection limit. Estimates of the standard error using KM generally assume that the data follow a normal distribution (D. Helsel 2012, 74–76). This means the 95% UCL still depends on the assumption of normality, which was already shown to be suspect in the previous section.

4.3.5 "robust" ROS (regression on order statistics)

Regression on order statistics (ROS) can be either fully parametric or not ("robust"). I focus on the robust version of ROS and explain it more fully, as this is the version used in the NADA for R package (Lee 2017) and is common based on recommendations to use. The robust ROS (rROS) assumes a parametric distribution (e.g. normal or lognormal) for the censored readings and otherwise uses the uncensored observations to directly compute summary statistics.

The basic idea of rROS is to regress the uncensored readings against a quantity called a "normal score" and then the collection of values of the censored observations are predicted based on the regression model fit (with their normal scores used as the explanatory variable). Normal scores are calculated by estimating the probability of going above each detection limit (in the case of multiple detection limits) and then computing plotting positions for each observation (both censored and uncensored). One detection limit is associated with one normal score. Then, the normal scores for censored readings are used as the explanatory variables in the regression equation and used to predict concentrations for the collection of censored observations (the predictions are not meant to be used as imputed values for specific concentrations, only as a group to estimate summary statistics). These predicted values and the uncensored values are combined to proceed with the estimation of means, confidence intervals, and other summary statistics (D. Helsel 2012). The 95% confidence interval for the mean using rROS is calculated using a t-based interval.

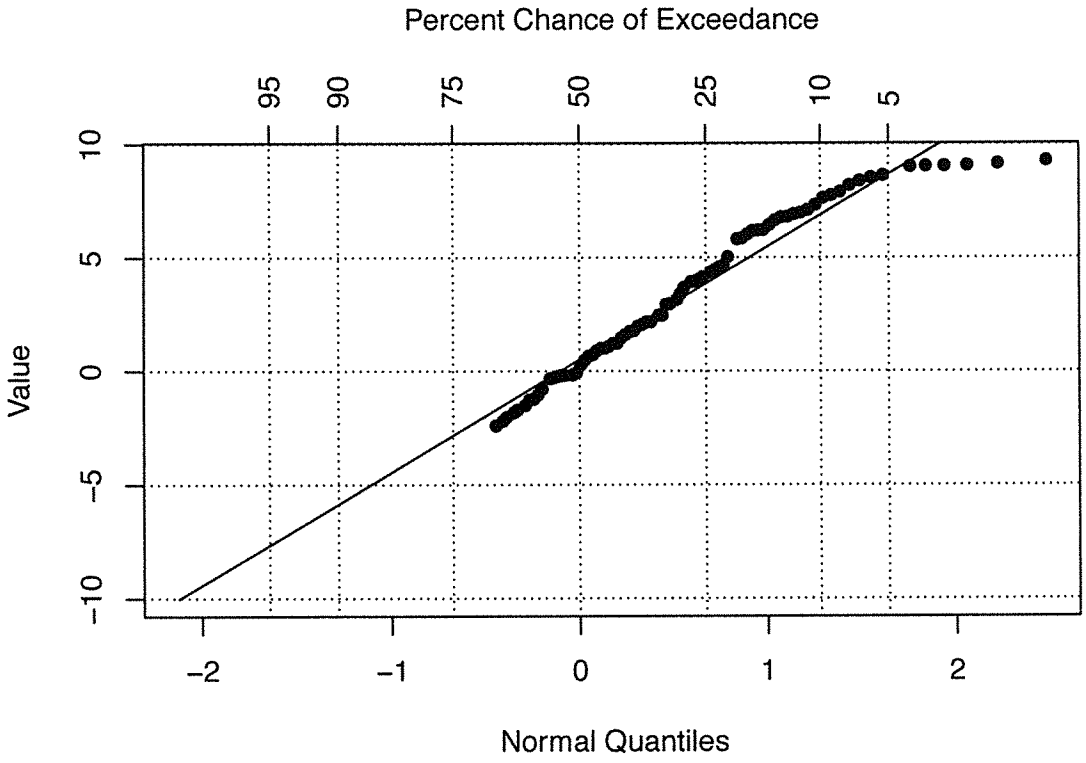

Figure 7: Probability plot and rROS line for ln(pyrene concentrations) using the cenros function in the NADA package from R.

5 Comparison of results

Table 1. Estimates of means, standard deviations, and 95% confidence intervals for the Deep Water Horizon dataset based on differing common methods for left-censored data of this type, on the natural log scale.

Method	Mean	SD	Lower.CI	Upper.CI
DL	1.87	3.996	-6.033	9.774
1/2DL	1.59	4.17	-6.659	9.839
Deletion	3.445	3.478	-3.435	10.325
MLE	3.286	0.342	2.616	3.956
KM	3.717	0.345	3.041	4.393
rROS	0.629	4.598	-8.465	9.723

It is clear that the different approaches will result in practically different confidence intervals. For example, as is the case between substituting the detection limit or half the detection limit and robust ROS versus using Kaplan Meier or Maximum Likelihood Estimation. Robust ROS results in the smallest estimated mean concentration at 0.629 ln(PPB), while using the detection limit or half the detection limit gives mean concentrations that are twice as much at 1.87 and 1.59 ln(PPB) respectively. KM, MLE and deletion of censored readings all have similar and much higher estimated mean concentrations above 3 ln(PPB). MLE and KM have the smallest estimate of the standard deviation around 0.34 ln(PPB). In terms of the 95% upper confidence limits, MLE and KM have the narrowest intervals overall and the upper confidence limit is around 4 ln(ppb) for both of them. However, deleting the censored observations yields the highest 95% upper confidence limit at 10.325, while substitution with the detection limit or half the detection limit and robust ROS gave 95% upper confidence limits just under that at around 9.7-9.8 ln(PPB). These differences on the log scale are much greater on the original concentration scale (which is the scale remediation decisions would be made on). Just looking at a simple backtransformation of the smallest and largest estimated 95% UCL's, the smallest would be 52.25 PPB and the largest would be 30485.3 PPB. This is a huge difference in results.

It is interesting that substitution yields the lowest estimates for mean concentration, but has (along with deletion and rROS) the largest estimates for 95% UCL, much higher than from MLE or KM. This is due to the fact that all intervals were calculated assuming a normal distribution and t-multiplier. Because the estimated standard deviation is much lower for MLE and KM methods, the intervals are much narrower than any of the other methods. Estimating the standard deviation is often harder than estimating a mean and is

usually a forgotten part of calculating a confidence interval.

Imagine comparing the upper 95% confidence limit with one threshold value and using that to largely make a decision about whether or not a potentially harmful chemical is present in high enough concentrations to warrant remediation. The method chosen could easily alter this decision based on how each method computes summary statistics and confidence intervals, as discussed earlier.

6 Conclusion

There is no "best" way to deal with detection limits and censored observations of this nature, and methods that may be reasonable in one situation may not be in another. The takeaway from my research into censored values of this kind is that there should not be a default method for approaching any analysis with left-censored observations due to laboratory decision-making and instrument calibration. An appropriate, reasonable method depends on many factors, including the proportion of censored values in the data and their location in relation to the observed values. We, as statisticians, should continue to educate others about the loss of information when observations are censored at a detection limit and, when possible, build relationships that might allow us to be involved pre-processing to minimize the amount of information that is lost.

Overall, it is important to advocate for the release of all instrument readings, including those usually censored at the detection limit, in order to have the possibility of using all available information to conduct an appropriate analysis. If the raw values are not available, it is equally as important to conduct a sensitivity analysis comparing the results based on several methods being considered to see how drastically (or not) results differ in terms of practical consequences and use this in the decision making process.

References:

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

42 (10): 3732–8.

Beck, E. C. 1979. "The Love Canal Tragedy." https://archive.epa.gov/epa/aboutepa/love-canal-tragedy.html.

EPA. 2015. "ProUCL Version 5.1 User Guide."

———. 2016. "Definition and Procedure for the Determination of the Method Detection Limit, Revision 2."

———. n.d. "Deep Water Horizon - Bp Gulf of Mexico Oil Spill." https://www.epa.gov/enforcement/deepwater-horizon-bp-gulf-mexico-oil-spill.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitering.* New York, New York: Wiley.

Helsel, D. 2012. *Statistics for Censored Environmental Data Using Minitab and R: Second Edition.* Hoboken, NJ: John Wiley & Sons.

Helsel, Dennis R. 2005. *Nondetects and Data Analysis: Statistics for Censored Environmental Data.* Hoboken, NJ: John Wiley & Sons.

Lampert, Peterson, D. 1991. "Nondetects, Detection Limits, and the Probability of Detection." *Journal of the American Statistical Association* 86 (414): 266–77.

Lee, Lopaka. 2017. *NADA: Nondetects and Data Analysis for Environmental Data.* https://CRAN.R-project.org/package=NADA.

Millard, Dixon, S.P. Unpublished. "Chapter 11: Censored Data."

NOAA. n.d. "Ship Data: Deep Water Horizon Support." https://www.nodc.noaa.gov/deepwaterhorizon/specialcollections.html.

Porter, Ward, S. 1988. "The Detection Limit." *Environmental Science and Technology* 22 (8): 856–61.

Public Health, Illinois Department of. n.d. "POLYCYCLIC Aromatic Hydrocarbons (Pahs)." http://www.idph.state.il.us/cancer/factsheets/polycyclicaromatichydrocarbons.htm.

ToxTown. n.d. "Polycyclic Aromatic Hydrocarbons (Pahs)." https://toxtown.nlm.nih.gov/chemicals-and-contaminants/polycyclic-aromatic-hydrocarbons-pahs.

Wendelberger, K., J & Campbell. 1994. "Non-Detect Data in Environmental Investigations." American Statistical Association.