My amazing title

Tony Ni
April DD, 20YY

Submitted to the Department of
Mathematics and Statistics
of Amherst College in partial fulfillment
of the requirements for the degree of
Bachelor of Arts with honors.

Advisor:
Brittney Bailey

Abstract

The abstract should be a short summary of your thesis work. A paragraph is usually sufficient here.

Acknowledgments

Use this space to thank those who have helped you in the thesis process (professors, staff, friends, family, etc.). If you had special funding to conduct your thesis work, that should be acknowledged here as well.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Background

Coal is one of the most dangerous combustible fuels which is being burned in all across the world as one of the largest methods of obtaining energy. Yet, although it is a fossil fuel which is naturally abundant and easy to utilize, it is comprised of a long list of dangerous chemicals including – but not limited to: arsenic, radium, boron, and a large list of other chemicals which prove to be dangerous to humans and animals alike. (Kelderman et al., 2019)

Power plants produce electricity by burning this coal, and as a result of how prevalent it is within the US - over 100 million tons of coal ash are produced every year. This side-product as a result of the coal combustion is often disposed by directly being dumped into landfills and waste ponds. (Kelderman et al., 2019)

Only recently have these complaints and lawsuits regarding the disposing practices made by non-profit environmental organizations been heard. Due to the onslaught of pressure put on the Environmental Protection Agency – the Coal Ash Rule was bor in 2015. (Kelderman et al., 2019)

This rule has forced over 265 coal power plants – about 3/4 of all coal power plants in the US - to make data regarding chemical concentrations publicly available to the general population. (Kelderman et al., 2019)

In their analysis using this data, the Environmental Integrity Project – a non-profit

organization dedicated to issues involving environmental justice have concluded that essentially all groundwater under coal plants are contaminated. (Kelderman et al., 2019)

However, is this really the case? There are many naturally occurring chemicals existing in groundwater as as such, perhaps their claims are overstated.

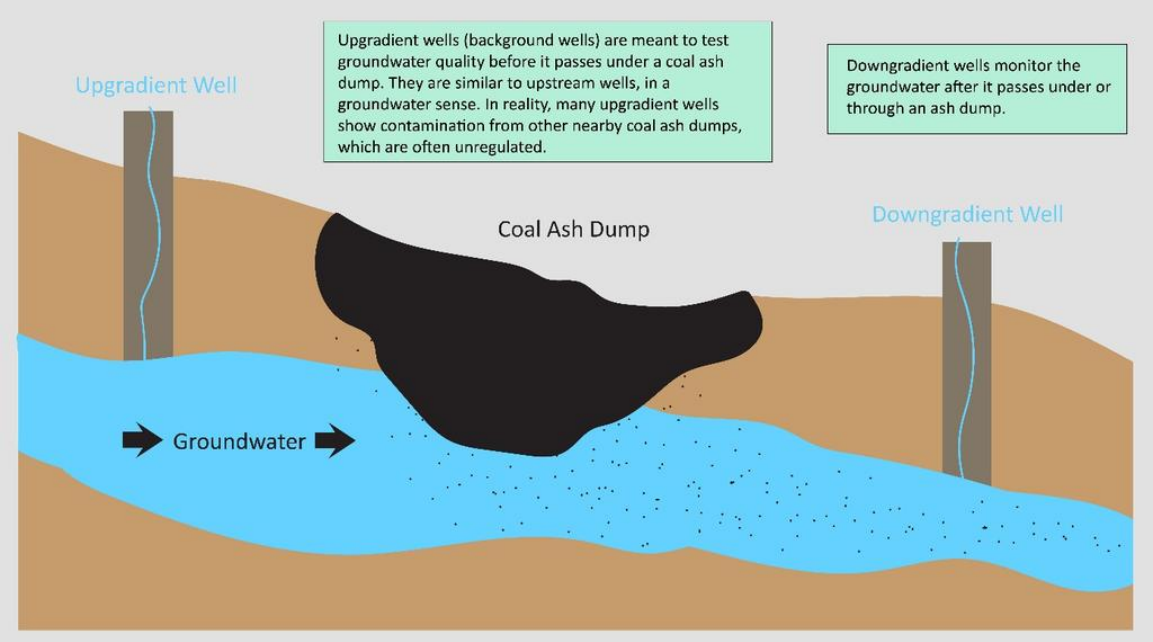

Figure 1.1: Difference Between Upgradient and Downgradient Wells

Typically in a coal ash plant, there exists two types of wells: upgradient wells and downgradient wells. These wells are essential to measure the amount of contamination being caused by coal ash. Upgradient wells, also known as background wells, measures the concentrations of chemicals in groundwater before it passes through an coal ash dump. Conversely, downgradient wells measure the concentrations of chemicals in groundwater after it passes through a coal ash dump.

- "80% of the US population is served by 14% of the utilities," so if something were to get into the water distribution system, it can easily spread amongst the

US population which is why contamination in water services is so important. (Byer & Carlson, 2019)

With this information, typically – one estimates the amount of chemical contamination caused by a coal as dump by subtracting the upgradient concentration from the downgradient concentration of a chemical (downgradient concentration - upgradient concentration).

However, due to the lack of proper reporting guidelines prior to the enactment of the Coal Ash Rule, we believe that there may be retired or even unregulated upgradient wells which can cause the concentrations of chemicals being recorded from these upgradient wells to be inaccurate or even completely wrong.

Our end goal remains the same as the EIP: to identify contaminated groundwater in coal plants – but to attempt to find a way to effectively correct the improper/inaccurate values resulting from LOD errors and other factors which the EIP may not have considered.

The limit of detection problem stems from the measuring devices' inability to obtain chemical concentrations smaller than a certain threshold amount, thus affecting the measurements recorded.

Our plan is to utilize methods designed to calculate descriptive statistics of interest when in the presence of left censored data. Specifically, we would like to calculate average concentrations of certain coal-risk heavy contaminants with regards to downgradient and upgradient wells – and seeing if our values differ if we had instead simply worked with the dataset without accounting for the left censored values.

1.2 Data

1.2.1 Coal Ash Rule

A large coal ash spill at the Tennessee Valley Authority (TVA) which occured on December 22, 2008 in Kingston, TN – prompted the Environmental Protection Agency (EPA) to propose a set of standardized regulations and procedures to address the concerns regarding coal ash plants nationwide in the US. (Environmental Protection Agency, 2020)

This was known as the Coal Ash Rule, passed on December 19, 2014. (Environmental Protection Agency, 2020)

Changes were made to the Coal Ash Rule over the years in the form of 'amendments,' one of which made required facility information and data to be made publicaly available to the public (April 15, 2015 rule change) (Environmental Protection Agency, 2020)

1.2.2 Source of Data

The data used in the study are from the results published in "Annual Groundwater Monitoring and Corrective Action Reports" which were made available to the public in March 2018 as a result of the Coal Ash Rule. (Environmental Integrity Project, 2020)

These reports are in PDF format and are thousands of pages long, which makes it difficult for individuals to look through the data in a meaningful way. (Environmental Integrity Project, 2020)

The EIP obtained the data from an online, publicly available database containing groundwater monitoring results from the first "Annual Groundwater Monitoring and Corrective Action Reports" in 2018 which was collected from coal plants and coal ash dumps under the Coal Ash Rule (Environmental Integrity Project, 2020)

They wrangled the data into a more accessible machine-readable format which contains information from over 443 annual groundwater monitoring reports posted by 265 coal ash plants, which is downloadable from the EIP's website. (Environmental Integrity Project, 2020)

1.2.3 Variables

The dataset contains information regarding chemical concentrations at coal plants. A coal plant consists of multiple disposal areas for the coal ash that it produces. At each disposal area, there are specific locations that groundwater is being measured, known as wells which represent an observation in the dataset. There are two types of wells – upgradient and downgradient wells. The variables consist of information regarding the specific chemical concentrations of each well.

1.2.4 Plan of Action

Within the report, the EIP mentions certain restrictions within the data that have caused their data to potentially be inaccurate (specifically, with limit of detection problems, and a large amount of missing chemical data). The limit of detection problem comes when measuring devices used to measure chemical concentrations are unable to detect below a certain threshold, causing large numbers of observations to have duplicate, wrong values – which can cause for misguided analysis.The other issue is less guided/formed, but for brevity, we think that a lot of the issues in the data comes from the potential possibility of contamination during data collection from investigators from non coal-ash sources. This may include things like: retired/unregulated wells which are old and have chemicals leaking into the groundwater, mismanagement in measuring, etc.

My project hopes to work with methods on handling this missing data – alongside investing potential uses of bootstrapping and other resampling methods (potentially?) in order to try to come up with a more statistically accurate and sound result by looking to assuage the problems that the EIP faced in their analysis. Specifically, to find a way to split up the data into ”uncontaminated” and ”contaminated” wells in

order to find the natural distribution of chemicals in each – and doing to so in the face of data corrupted by LOD problems and inaccuracies. I'm hoping to apply and compare different ways of altering the data to account for these myriad of issues in order to look for more salient findings that the EIP might have missed or if not, to see if improvements can be made regarding the way that contaminated coal ash wells are being identified.

Chapter 2 Methodology

The concept of missing data is ubiquitous within academic disciplines and often complicates innumerable types of real-world studies. Missing data will be defined within this thesis as *occurences within a dataset where there is no value stored for a variable in the observation of interest.* As most studies often utilize data collected through mediums such as surveys, questionnaires, or field research, missing data is an unavoidable problem. Missing data hinders one's ability to work with and analyze the phenomena at hand, as data is often the basis of all studies.

Barnard and Meng (1999) outline three significant issues when conducting analysis on missing data. Firstly, it can introduce estimation bias within an analysis. Estimation bias is defined as *the difference between and estimator's expected value and the true value of the parameter being estimated.* Another issue is that missing data can often lead to a reduction in statistical power, which can affect the conclusions one makes during studies involving hypothesis testing. Finally, missing data can introduce complications with statistical software and lead to functions not working as intended, if they have not accounted for the possibility of the data containing missingness.

This thesis will go into a more specific instance of missing data known as censoring, which is *the condition when one has only partial information regarding the values of a measurement within a dataset.* We will introduce and define the three types of censored data, discuss the challenges with the reporting of censored data, and explore common statistical approaches to handling censored data.

2.1 Censored Data

As discussed previously, censored data is a specific type of missingness where one has only partial information regarding the values of a measurement in a dataset. There are many types of censoring which can occur, but three main ones which are the most common: right censoring, interval censoring, and left censoring. ### Left Censoring {#left}

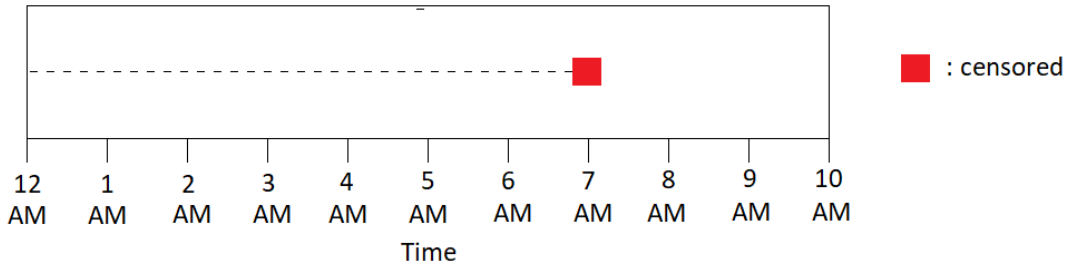

Figure 2.1: Left Censoring Example

Left censoring is a specific instance of censoring in which we only know that the true value of a data point falls below a certain threshold which we call the *limit of detection* (LOD).

To understand this concept better, consider the following example. Imagine a scenario in which you are attempting to estimate the time at which the sun rises each morning. You plan to wake up every morning far before the sun rises, but on the first day of the study, you oversleep and wake up at 7:00 A.M. with the sun already out. We now have an instance of left-censored data. We want to know the time at which the sun rose, but all we have is an upper limit (7:00 A.M.).

Left censoring is commonly found in environmental, water quality, and chemical-related research where the focus is on the concentration of an analyte. Due to limitations on measuring instruments, left censored data are commonly found in these types of studies. The most pressing issue of left-censored data mostly lie in the difficulty of distinguishing between extremely low values and statistical noise (Hall, Perry, & Anderson, 2020).

2.2 Challenges of Reporting Censored Data

There is no universal reporting practice for values below the LOD which can lead to confusion amongst researchers. The lack of standardization makes it difficult to distinguish LOD values with uncensored values. This can lead to LOD values unintentionally being overlooked, causing faulty analysis or conclusions which are heavily flawed.

In a study involving the precision of lead measurements near concentrations of the limit of detection, Berthouex (1993) discusses the disparity in practices within chemists on ways to record LOD values. He enumerates in a following list, common reporting practices in this field:

1. Reporting the trace, a chemical whose average concentration is less than 100 μg
2. Reporting the letters ND, which stand for "not detected"
3. Reporting the numerical LOD value itself
4. Reporting "<", followed by the numerical LOD value
5. Reporting some value between 0 and the LOD value, such as one-half the LOD value

6. Reporting the actual measured concentration, even if it falls below the LOD

7. Reporting the actual measured concentration, followed by "(LOD)"

8. Reporting the actual measured concentration with a precision (±) statement

The latter three methods are the best procedures to follow, especially from a practical and statistical point of view according to Gilbert (1987). He argues that assuming the small concentration values are not from some sort of measurement error during data collection, then the value holds value. As such, recording a measurement as "below LOD," without any sort of accompanying value would be discarding useful information which could have been used in practice and analysis.

Berthouex (1993) discusses the prevalence in regards to the practice of censoring data by reporting only values which are above the detection limit and discarding those which fail to yield quantifiable results. He discourages this practice and instead suggests the reporting of measurements, even when those values are below the limit of detection.

Further supporting the stance of keeping all concentration values rather than only those above the detection limit, Monte-Carlo experiements conducted by Gillom (1984) show that linear trends in water-quality data were far more easily able to be detected with uncensored data as compared to censored data. The methods they used to handle censored water quality data were found to produce wild and erratic estimates for the mean and standard deviation of datasets with higher censoring levels than those without. They found a general trend of decreasing classification success with increased censoring levels, attributing it to the limited availability of information in censored data.

2.3 Approaches

It is important to note that the values below the LOD still contain information, specifically that the values is between the lower bound value (if it exists) and the LOD (Chen et al., 2011). As such, there are a variety of statistical treatments to handle censored data which have been popularized in the statistical literature which will be discussed within this section.

Omission involves the deletion of data points which are deemed to be invalid as a result of left-censoring or any other deficiencies in the data. This is also more commonly known as *available-case analysis*, in which statistical analysis is conducted while only considering the observations which have no missing data on the variables of interest, and excluding the observations with missing values (May, 2012). May argues against this approach and claims that the loss of information from discarding data and the inflation of standard errors of estimates (when discussing missingness in a regression context) will invariably be inflated as a result of the decreased sample size. The advantages of omission lies in its ease of implementation.

Apart from available-case analysis, over the past century, a myriad of methods to deal with censoring have been developed to counter this issue – some more statistically sound than others. We will review some of the most common methods to estimate descriptive statistics involving censored data, which include: substitution, maximum likelihood estimation, Kaplan-Meier, and regression on order statistics (Lafleur et al., 2011).

2.3.1 Substitution Method

Often condemned in papers as a statistically unsound method to handle censored data, substitution methods are ubiquitous in the chemical and environmental sciences

as an appropriate and recommended method to work with left-censored chemical concentration data (Canales, 2018).

The substitution method simply involves imputing in a replacement value in lieu of the censored data point. The lack of a global, standardized replacement value to substitute is one of the most pronounced downside of this method. The replacement value used may differ between studies but common values include: $\frac{LOD}{2}$, $\frac{LOD}{\sqrt{2}}$, or LOD (Lee & Helsel, 2005). Different disciplines have their own suggested "best" replacement value to use, an example being $\frac{3}{4}$ times the LOD being a common replacement value in geochemistry (Crovelli, 1993). However, it must be recognized that the substitution method is a statistically unsound technique which is often used in non-rigorous statistical settings due to them being quite easy to implement (Chen et al., 2011). As such, there have been several studies in order to investigate the effectiveness of the method.

Proponents of the substitution method claim that the replacement value $\frac{LOD}{2}$ is useful for data sets in which the majority of the data are below the LOD or when the distribution of the data is highly skewed; the definition of "highly skewed" being any distribution with a geometric standard deviation (a measure of spread commonly used in tandem with log-normal distributions) of 3 or more (Hornung & Reed, 1989). They also suggest using $\frac{LOD}{\sqrt{2}}$ when there are only a few data points below the LOD or when the data is not highly skewed.

Substitution methods are flawed as they can often introduce a "signal" which was not originally present within the data, or even obstruct an actual signal which was present in the original data (Lee & Helsel, 2005). Numerous authors have advised against the usage of substitution methods for being statistically inappropriate to use. Glass and Gray (2001) found that both introduce large errors and biases in descriptive statistics of interest. Thompson and Nelson (2001) conducted a study in

which they found similar results, in that it often led to biased parameter estimates and "artificially small standard error estimates." Hewett and Ganser (2007) also found in their simulation study that the substitution method yielded the lowest average bias and root mean squared error values (comparison metrics to measure accuracy) in their estimation of the mean. Overall, the overall consensus seems to advise against the practice of these substitution techniques.

[paragraph talking about proponents of the substitution methods??]

2.3.2 Maximum Likelihood Estimation Method

Maximum likelihood (ML) estimation is a parametric technique which allows us to estimate the parameters of a distribution or model when the data is from a multivariate normal distribution.

To give a brief introduction to the mechanisms of ML estimation, let $f(x|\theta)$ denote the probability density function (PDF) which specifies the probability of observing the random variable x given the parameter θ.

Given a random, independently and identically distributed (*i.i.d.*) set of random variables $X_1, X_2, ..., X_n$ from $f(x|\theta)$, we know that each individual observation x_i's are statistically independent from one another, which allows us to express the PDF as the product of all individual densities. For every observed random sample $x_1, ..., x_n$, we can define the joint density function to be:

$$f(x_1, ..., x_n|\theta) = f(x_1|\theta)...f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

In most real-life scenarios, the actual (observed) data is already given, and our goal is to find the PDF which is most likely to generate our observed values. In order to solve this inverse problem, we introduce the likelihood function, which is defined as

the joint density of the observed data as a function of the parameter (with the data held as a fixed constant).

In mathematical notation, upon observing the given data, $f(x_1, ..., x_n|\Theta)$ becomes a function of θ alone, so we obtain a likelihood of:

$$lik(\theta) = f(x_1, ..., x_n|\theta)$$

It is important to recognize the difference which separates the likelihood function and the PDF. The PDF is a function of the observed data given a parameter(s). It gives information regarding the probability of a particular data value for a fixed parameter.

On the other hand, the likelihood function is a function of the parameter, given a set of observed data. It tells us the likelihood of observing a particular parameter value for a fixed set of data.

Our goal is to obtain the ML estimate of our parameter which maximizes the likelihood function, $lik(\theta)$, in other words, to obtain a θ which makes our observed data the most probable.

As we previously declared our random variables $X_1, X_2, ..., X_n$ to be i.i.d, we can rewrite the likelihood to be a product of the marginal densities:

$$lik(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

in which we can then maximize the likelihood to find the best mle of θ to best capture our observed data.

Yavuz et al. (2017) discuss the usage of MLE method, when missing data is present, and note that it is only appropriate to use for non-negative probability distributions such as: exponential, log-normal, normal, and Weibull.

When left censoring is present, the likelihood function changes in order to account

for both the censored observations and the uncensored observations and becomes:

$$lik(\theta) = \prod_{i=1}^{n} f(x_i|\theta)^{\delta_i} \times F(x_i|\theta)^{1-\delta_i}$$

in which δ_i is an indicator, representing whether or not if the ith observation is censored or not:

$$\delta_i = \begin{cases} 0, & \text{if censored} \\ 1, & \text{if uncensored} \end{cases}$$

From this updated definition of the likelihood function which must be used in the presence of left censored data, it is then possible to follow typical procedures to find the estimator, θ, which maximizes the likelihood, also known as the *maximum likelihood estimator*. With this knowledge, the descriptive statistics of interest (mean, variance, etc.) relating to the specified distribution can be calculated.

Canales (2018) outlines a imputation technique which involves replacing censored observations with values from the estimated parameterized distribution. However, is not mentioned explicitly how this imputation method is conducted.

The code for the MLE method will be handled with the `cenmle` function in the `NADA` package, which allows the user to specify censored and uncensored data, and uses the LOD as the placeholder. As this method is not an imputation technique, values are not replaced. This method allows us to calculate the summary statistics for the entire data set – including the censored values. (remove and write own code??)

As a technique which heavily relies upon knowing a distribution which best models the data, MLE is one of the most well-known parametric approaches to handling LOD values. Many studies use the MLE as a sort of baseline method of handling censored values, to which they compare their new techniques upon (Ganser & Hewett,

2010). However, it must be known that regardless of the prevalence of the MLE method, it is not free from its own downfalls. Canales (2018) found that the MLE method seems to underperform when the data in question was highly skewed, in which overinflated mean squared errors were often obtained. Being a technique which is so heavily dependent upon distributional assumptions, an incorrect specification of the distribution of the censored data will inevitably lead to misleading results (Bolks, DeWire, & Harcum, 2014).

2.3.3 Kaplan-Meier Method

As a phenomenon, censoring is most often discussed in survival analysis, which concerns itself with techniques to analyze a time to an *event variable*. As its name suggests, these variables measure the time which passes until some sort of event occurs. This can be as innocuous as the time until device breaks, time until birds migrate away from their homes, time until a person passes away, etc. Regardless of which, all these scenarios share a common problem in terms of the possibility of the data being "censored."

The Kaplan-Meier (KM) method is a common nonparametric technique used to deal with censored data. Nonparametric methods do not utilize any information regarding the parameters for a specified distribution, like the mean and standard deviation for the normal distribution. The KM method was originally developed to handle right-censored survival analysis data. The advantages of the KM method lie in its robustness as a nonparametric method, it performs well without having to depend upon distributional assumptions. Many recommend its usage for when there are cases of severe censoring, instances where $> 90\%$ of the data is censored (Canales, 2018).

To introduce the concept of the KM-estimator, it is helpful to take a look into its usages in survival analysis studies where the focus is often on a type of data known as

"time to event" data. These types of studies often involve events such time to death, time to failure, and so forth.

The KM-estimator is a statistic used to estimate the survival curve from the empirical data while accounting for the possibilities of certain values being censored. It does this by assuming that censoring is independent from the event of interest and that survival probabilities remain the same in observations found early in the study and those recruited later in the study.

The KM-estimator when performing an empirical estimation of the survival curve at time t can be represented by the following equation:

$$\hat{S}(t) = \prod_{t_i \leq\ t} \left(1 - \frac{d_i}{n_i}\right)$$

where t_i is the distinct event time, d_i is the number of event occurrences at time t_i, and n_i is the number of followup times (t_i) that are $\geq t_i$ (how many observations in sample survived at least/or past the time t_i) (Klein & Moeschberge, 2003).

Typically, the KM-estimator can only be used to estimate the distribution function of right-censored data, in which a data point is above a certain threshold, but it is unknown by how much. A simple tweak to the typical KM-method, allows for the estimation of the survival curve with left-censored values.

Helsel (2005, as cited in Yavuz et al., 2017) provides a detailed explanation on how to apply the KM method when left censoring is present. Firstly, it is essential to reverse the left-censored data through a transformation algorithm before using the KM method to change them into right-censored data.

Let $x_i \ldots x_n$ be the values for the observations $i = 1, 2, ..., n$. Arrange all the left-censored values in descending order and then subtract them by M, a constant bigger than the biggest value of the dataset, in order to get the transformed, right-censored

value, $M - x_i$. All values are then arranged in ascending order to be used to estimate the survival function through the Kaplan-Meier estimator.

It must be known that the KM-method is not an imputation procedure, but instead an estimation technique that allows for the calculation of descriptive statistics for left-censored datasets. Nian (1997 gives the expressions to calculate the estimated mean, median, and variance below:

$\hat{\mu} = \int_0^\infty \hat{S}(t)\ dt\ \hat{M} = \hat{S}^{-1}\left(\frac{1}{2}\right)\ Var(\hat{\mu}) = \sum_{i=1}^{r}\left(\int_{t_i}^\infty \hat{S}(t)\ dt\right)^2 \frac{d_i}{n_i(n_i-d_i)}$

2.3.4 Regression on Order Statistics

Lastly, regression on order statistics (ROS) combines both the parametric nature of the MLE approach and nonparametric nature of the KM method. ROS is a semi-parametric method which assumes an underlying normal or lognormal distribution for the censored measurements but makes no assumption towards the distribution of uncensored measurements.

(Environmental Protection Agency, 2009) provides a a more detailed explanation to the methodology of ROS, but the basic procedures will be outlined in this thesis.

ROS begins with the estimation of the cumulative probability associated with each distinct LOD. This cumulative probability is distributed equally between the censored values with a common LOD (see (Environmental Protection Agency, 2009), for more details). A regression model is fit between the uncensored values and the distributional quantiles. The slope and intercept of the regression line from this model is then used to estimate the mean and standard deviation of the distributional model which are then used to generate imputed values for the censored observations.

In order for ROS to be utilized, there needs to be at least 5 known values and more than half the values within the censored variables must be known. As regression is utilized in this method, the response variable must also be a linear function of

the explanatory variable (quantiles). Additionally, the errors should have constant variance (Lee & Helsel, 2005).

The `NADA` package contains the function `ros` which provides an implementation of regression on order statistics which allows us to calculate descriptive statistics for left censored values.

[INSERT PARAGRAPH TO TRANSITION TO CHAPTER 3 (?)]

Chapter 3 Simulations

Having discussed the various methods to handle left-censored data in the previous chapter, we now turn to a simulation study in order to evaluate the strengths and weaknesses of each method in cases of differing censoring rate, sample size, and distribution. We will also discuss the implementation of the methods, data generating mechanisms, and specific evaluation metrics to assess the performance of each method.

3.1 Aims

The question of which method is the best to use is frequently discussed topic within the field. Many studies have been conducted over the years to evaluate the performance of these methods to handle left-censored data, with the results being widely varied and largely inconclusive. First and foremost, a large issue comes in that every conducted study widely differs in the methods being investigated and the scope of the study. As an example of the broad differences between studies which can make comparisons difficult, (Antweiler, 2015) evaluates the effectiveness of 11 different methods with several censoring rates and distributional assumptions, using the median absolute deviation (MAD) as their performance metric of choice. Meanwhile, (Hall, Perry, & Anderson, 2020) focuses instead on the applications of such methods from a water-quality focused context and investigates the performance of the four methods used in this thesis – except with water stream concentration data, and no focus on

distributional assumptions nor censoring rates. As each of these studies are concerned with their own goals – the reasoning and conclusion that they reach will inevitably be different. Studies that are more focused on a general, broad audience, with no assumptions as to what sort of data the individual is working with – may find more use with the conclusion and results that investigators like Antweiler come up with. There may also be individuals who are more focused on the performance of such methods in a specific context, as in the study conducted by Hall. There is no common ground between statisticians on the optimality of methods, prompting our own foray into this topic. I wish to incorporate the detailed specifications of a simulation study, while keeping it applicable towards the coal contamination water quality data that I am hoping to apply the methods to.

Through our simulation study, we wish to identify settings where a method can be effective but also those in which the methods may not be able to perform quite as well. Several investigators in this field have found issues with certain methods underperforming under certain conditions, and brings up the possibility of particular methods being more equipped than others to deal with different rates of censoring. Specifically, a study on methodologies to handling left-censored microbial risk assessment conducted by (Canales, 2018) found that the substitution method seemed to work much better than expected while other methods, such as the MLE method, seemed to have trouble when applied to highly skewed data. Results from other works (Antweiler, 2015) suggest that regardless of the method being utilized, obtaining reliable estimates from datasets where censoring was greater than 40% was unfeasible. This particular study also suggests that the size of the data had no influence on the result of the estimate in question.

Claims regarding the effectiveness of methods with regards to censoring rates, distribution of data, and sample size are all highly contentious. In order to get a better

idea of sense of how these claims hold up, our goal is to evaluate the validity of those claims by conducting a simulation study of our own which will put those claims into practice. We must note that we don't expect one method to be significantly above the others in terms of performance for all settings, but we do hope to see which method might be best for certain settings, in terms of distributional assumptions, censoring rates, and sample sizes. This approach and aim of our simulation study will then largely depend on how we will generate the data to be used in order to achieve our goal.

3.2 Data-Generating Mechanisms

The data generated for use in our study will be obtained by using parametric draws from user-specified distributions (log-normal, exponential, and Weibull), as the methods utilized can only be used with non-negative distributions (Yavuz, Tekindal, & Dog, 2017). Our data-generating mechanism also alters criterion such as the sample size, $n_{obs} = \{10, 100, 1000\}$ and censoring rate $R = \{0.10, 0.20, 0.30, 0.40, 0.50\}$.

Censored values are generated and determined by first arranging the uncensored observations in ascending order, and then from the censoring rate. For instance, if our censoring rate were $R = 0.15$, the lowest 0.15 of the observations will be marked as censored while the rest remain uncensored.

3.3 Estimands

Each of the four methods discussed in the previous chapter are designed for usage in obtaining summary statistics for left censored data (Shoari, 2018). in our simulation study, we want to evaluate just how well these methods are able to estimate a population quantity. We will be using the sample mean as an estimator for our

estimand, the population mean, μ.

3.4 Performance Measures

Morris et al. (2019) define performance measures as numeric metrics used to assess the performance of the method in question. The criterions we will use to assess the performance of each of our four methods will consist of: bias, variance, and mean squared error (MSE).

3.4.1 Variance

Before defining variance, it is important to have a good grasp of the concept of precision. Precision simply refers to how far away estimates from different samples are from one another. Low precision indicates that the estimates from each sample are close to one another in value and vice versa.

Knowing this, variance is a metric which informs us on the precision of an estimator. It is defined as simply the average squared deviation of the estimator from its average, which in our case is defined as:

$Variance = E[(\hat{\mu} - E(\hat{\mu}))^2]$

Estimators with low variances generally remain close in value throughout all samples, while those with high variance may wildly differ between samples. As such, it is generally preferable to have an estimator with low variance. However, it is important to note that precision measurements, such as the variance, are not a sole indicator of an estimator's performance (Walther & Moore, 2005). While precise estimators are ideal, it is also important to assess the estimator's bias, how close it is to the true value.

3.4.2 Bias

Bias is defined as the difference between an estimator's expected value and the true value of the parameter. In our case, we are using the estimator $\hat{\mu}$ to "estimate" the true population mean, μ, in each of our samples. As such, bias can be defined in our case as:

$$Bias = E(\hat{\mu}) - \mu$$

It is important to note that bias is a metric which only informs us on the difference of the estimator from the true parameter, and tells us nothing regarding accuracy nor precision.

If the bias of an estimator were to be equal to zero, we would define the estimator to be *unbiased*, meaning that the estimator produces parameter estimates which are on average, equal to the true value.

However, it is important to note that just because an estimator is unbiased, does not necessarily tell us anything about the quality of our estimator (of being good or bad). An unbiased estimator could have high variance, which would mean that the estimator in each sample would be significantly different from one another, but on average – they equal the true population estimand.

On that same note, it would not be very useful if an estimator had low variance but high bias, either – as this would mean that each sample would consistently produce similar estimates which are very far away from the true population estimand in question.

3.4.3 Mean Squared Error (MSE)

We generally would like estimators which have low bias and low variance, but it can be difficult to achieve both at once. As such, it is common to instead turn to a quantity

known as the mean squared error (MSE), which is a quantitative measurement used to assess the accuracy of an estimator. The MSE measures how far away, on average, an estimator is from its true value.

(NOTE: section below is VERY messy, need to clean up, play around with math mode, blah)

$MSE = E[(\hat{\mu} - \mu)^2] = Var(\hat{\mu}) + [Bias(\hat{\mu})]^2$

We can show that the MSE of estimator can be rewritten in terms of its variance and bias:

$$E[(\hat{\mu} - \mu)^2] = E(\hat{\mu}^2) + \mu^2 - 2E(\hat{\mu})\mu$$

Since we know bias to be $Bias = E(\hat{\mu}) - \mu$, it follows that $Bias^2 = E^2(\hat{\mu}) + \mu^2 - 2E(\hat{\mu})\theta$. We already know variance to be $Variance = E[(\hat{\mu} - E(\hat{\mu}))^2] = E(\hat{\mu}^2) - E^2(\hat{\mu})$. Thus, combining the square of the bias with variance yields:

$Bias^2 + Var = [E^2(\hat{\mu}) + \mu^2 - 2E(\hat{\mu})\theta] + [E(\hat{\mu}^2) - E^2(\hat{\mu})]$ the $E^2(\hat{\mu})$ terms cancel out, and we are left with: $E^2(\hat{\mu}) + \mu^2 - 2E(\hat{\mu})\theta = E[(\hat{\mu} - \mu)^2] = Bias$.

As the MSE is always positive, MSE values closer to zero are more desirable – as it is an indicator that the estimator is accurate.

3.5 Results

[place figures/tables from results of simulation study here, along with explanation]

From the results of our simulation study, we can see that with the data from the log-normal distribution...

- the methods which obtained the largest underestimates of the average sample means was with mle and substitution when censoring rates were $> 50\%$ (regardless of sample size)

- km method for all censoring rates and all sample sizes gave large overestimates, had very large, positive bias values compared to all other methods
- variance is largely dependent only on sample size (to be expected)
- in the case of high censoring (0.50) and small sample size (10), km did not work as well as other methods
- in the case of medium censoring (0.30) with medium sample size (100) and high sample size (1000), ros performed the best, with mle following closely as compared to the other two methods. ros and mle had bias and mse values which were lower than the other two.
- in the case of high censoring (0.50) and large sample size (1000), ros performed SIGNIFICANTLY better than all other methods (low bias, low mse)

We can see with the exponential...

We can see with the Weibull...

To sum it all up...

3.6 Discussion

[discuss findings from the simulation study. are the results expected from knowledge gained from literature search? are they different? compare/contrast with literature. what things seem to be shared between our findings and that in literature?

3.6.1 Limitations

[discuss some limitations of the simulation study – ideas include things such as how simulated data =/= real life data, discuss some limitations, future plans?]

Shortcomings in the results presented in this study very well may come from the fact that we generated data with known distributional parameters. It could be the case that the effectiveness of our methods were only due to having such artificial data. Alterations in our study to instead generate data from methods such as randomized pulls from an a real-world dataset of interest via. methods such as bootstrapping could provide different insights.

3.7 Study on Real Data

[connect back to chapter 1]

[can write this chapter after some preliminary exploration with coal groundwater data]

Chapter 4 Conclusion

[write a few paragraphss to wrap up entire thesis]

Corrections

A list of corrections after submission to department.

Corrections may be made to the body of the thesis, but every such correction will be acknowledged in a list under the heading "Corrections," along with the statement "When originally submitted, this honors thesis contained some errors which have been corrected in the current version. Here is a list of the errors that were corrected." This list will be given on a sheet or sheets to be appended to the thesis. Corrections to spelling, grammar, or typography may be acknowledged by a general statement such as "30 spellings were corrected in various places in the thesis, and the notation for definite integral was changed in approximately 10 places." However, any correction that affects the meaning of a sentence or paragraph should be described in careful detail. The files samplethesis.tex and samplethesis.pdf show what the "Corrections" section should look like. Questions about what should appear in the "Corrections" should be directed to the Chair.

References

10 Antweiler, R. C. (2015). Evaluation of Statistical Treatments of Left-Censored Environmental Data Using Coincident Uncensored Data Sets. II. Group Comparisons. http://doi.org/10.1021/acs.est.5b02385

Bolks, A., DeWire, A., & Harcum, J. B. (2014). Baseline Assessment of Left-Censored Environmental Data Using R. *Technotes*, *9*(2), 153–172. Retrieved from https://www.epa.gov/sites/production/files/2016-05/documents/tech_notes_10_jun2014_r.pdf

Byer, D., & Carlson, K. H. (2019). Real-time detection of intentional chemical contamination in the distributional system, *97*(7), 130–133.

Canales, R. (2018). Methods for Handling Left-Censored Data in Quantitative Microbial Risk Assessment, *84*(20), 1–10.

Chen, H., Quandt, S. A., Grzywacz, J. G., Arcury, T. A., Environmental, S., Perspectives, H., ... Arcury, T. A. (2011). A Distribution-Based Multiple Imputation Method for Handling Bivariate Pesticide Data with Values below the Limit of Detection, *119*(3), 351–356. http://doi.org/10.1289/ehp.l002124

Crovelli, R. A. (1993). An Objective Replacement Method for Censored Geochemical Data, *25*(1), 59–80.

Environmental Integrity Project. (2020). Coal Ash Groundwater Contamination: Documenting Coal Ash Pollution. Retrieved from https://environmentalintegrity.org/coal-ash-groundwater-contamination/

Environmental Protection Agency. (2009). *STATISTICAL ANALYSIS OF GROUNDWATER MONITORING DATA AT RCRA FACILITIES UNIFIED GUIDANCE*.

Environmental Protection Agency. (2020). Disposal of Coal Combustion Residuals from Electric Utilities Rulemakings. Retrieved from https://www.epa.gov/coalash/coal-ash-rule

Ganser, G. H., & Hewett, P. (2010). An Accurate Substitution Method for Analyzing Censored Data, (April), 233–244. http://doi.org/10.1080/15459621003609713

Hall, L. W., Perry, E., & Anderson, R. D. (2020). A Comparison of Different Statistical Methods for Addressing Censored Left Data in Temporal Trends Analysis of Pyrethroids in a California Stream. *Archives of Environmental Contamination and Toxicology*, *79*(4), 508–523. http://doi.org/10.1007/s00244-020-00769-0

Hornung, R., & Reed, L. (1989). Estimation of Average Concentration in the Presence of Nondetectable Values. *Applied Occupational and Environmental Hygiene*, *5*(1), 46–51.

Kelderman, K., Kunstman, B., Roy, H., Sivakumar, N., Mccormick, S., & Bernhardt, C. (2019). Coal's Poisonous Legacy: Groundwater Contaminated by Coal Ash Across the U.S.

Klein, J. P., & Moeschberge, M. L. (2003). *SURVIVAL ANALYSIS: Techniques for Censored and Truncated Data* (2nd ed., pp. 92–104). New York: Springer-Verlag.

Lafleur, B., Lee, W., Billhiemer, D., Lockhart, C., Liu, J., & Merchant, N. (2011).

Statistical methods for assays with limits of detection: Serum bile acid as a differentiator between patients with normal colons, adenomas, and colorectal cancer. *Journal of Carcinogenesis*, *10*, 1–8. http://doi.org/10.4103/1477-3163.79681

Lee, L., & Helsel, D. (2005). Statistical analysis of water-quality data containing multiple detection limits : S-language software for regression on order statistics $, *31*, 1241–1248. http://doi.org/10.1016/j.cageo.2005.03.012

May, R. C. (2012). Estimation Methods for Data Subject to Detection Limits, 82.

Shoari, N. (2018). Toward Improved Analysis of Concentration Data : Embracing Nondetects, *37*(3), 643–656. http://doi.org/10.1002/etc.4046

Walther, B. A., & Moore, J. L. (2005). The concepts of bias , precision and accuracy , and their use in testing the performance of species richness estimators , with a literature review of estimator performance, *6*(July).

Yavuz, Y., Tekindal, M. A., & Dog, B. (2017). Evaluating Left-Censored Data Through Substitution , Parametric , Semi-parametric , and Nonparametric Methods : A Simulation Study, 153–172. http://doi.org/10.1007/s12539-015-0132-9