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.

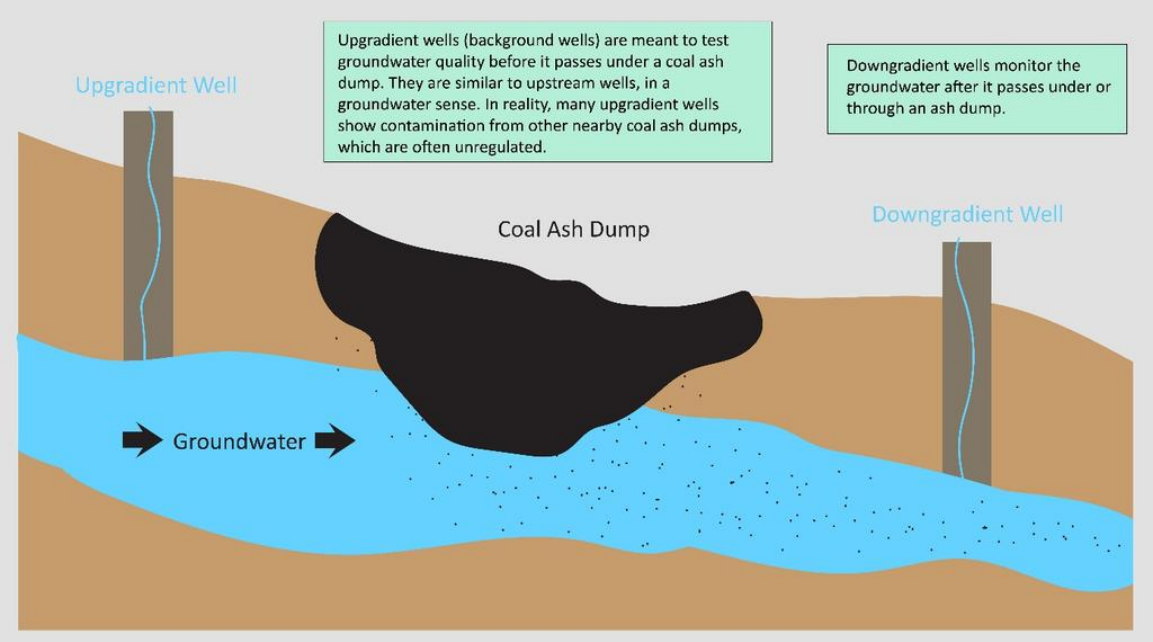

Figure 1.1: Difference Between Upgradient and Downgradient Wells

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 bootstrapping and imputation techniques to correct for these measurements by accounting for the innate contamination which may be caused by factors such as retired and unregulated wells that were mentioned before.

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. From the 19 different contaminants (antimony, arsenic, boron, etc.) a major problem is that some wells only have measurements for certain chemicals and don't have them for others.

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 which frequently complicates any 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. One of the most glaring issues with missingness is how it can introduce bias within an analysis, which can more often than not, invalidate a study if not accounted for and handled properly.

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, challenges with the reporting of censored data, alongside a discussion on 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.

2.1.1 Right Censoring

Right censoring is a specific instance in which we only know that the true value of a data point lies above a certain threshold, but it is unknown by how much. Suppose a study on income and mortality is conducted with the variable of interest, T, being the time measured from the start of the study to the death of the participant. The study has a duration of 5 years, in which participants are expected to submit a form regarding their annual income. The value for the participant would be considered to be right-censored if at any point during the study, they failed to follow-up, or if the participant was still alive at the conclusion of the 10 year study. In this design study, several possibilities can occur, illustrated in figure blah.

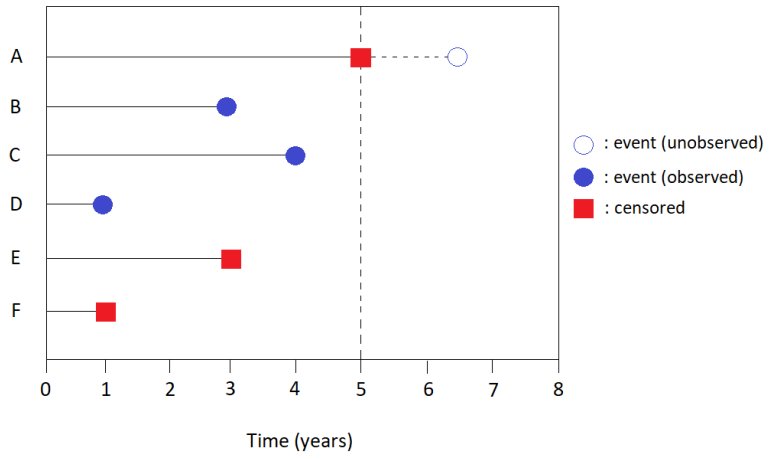

Figure 2.1: Right Censoring Example

If the individual passes to the termination of the study, the only information we have is $T > 5$.

If an individual passes away at some point, t_i during the study, then $T = t_i$.

Suppose an individual stopped submitting a form during the third year of the study. As we have no information about whether or not if they died or simply did not submit their form, all we know is that the individual died/will die at some point after year three of the study. In this instance, $T > 3$.

Right censoring is the most common type of censoring and can often be found in clinical trial studies, mortality studies, and other forms of surival analyses.

2.1.2 Left Censoring

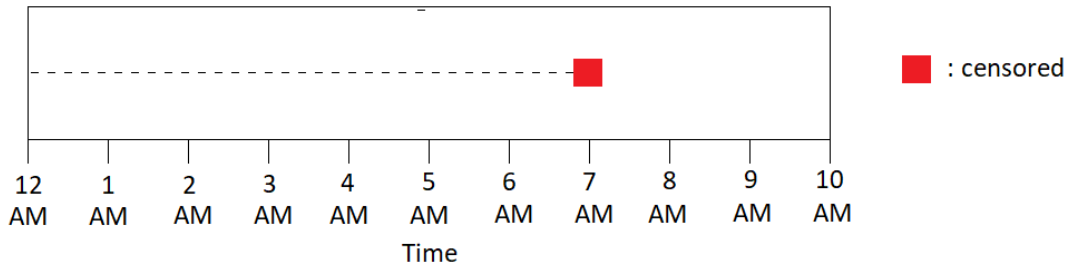

Figure 2.2: Left Censoring Example

In contrast with right censoring, 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.

2.1.3 Interval Censoring

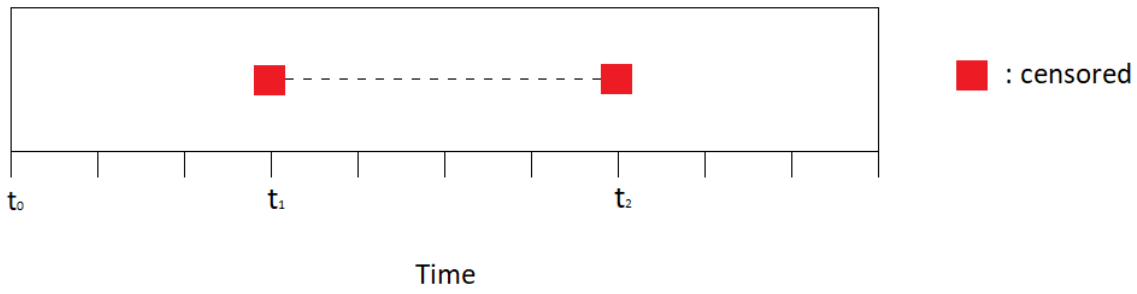

Figure 2.3: Interval Censoring Example

Interval censoring is another form of censoring in which the random variable of interest is known to be between an interval of two values. Considering a random variable T, which denotes the survival time of interest, if interval censoring is at hand, we can denote the interval containing T to be $I = [L, R]$, with L being the beginning of the interval and R being the end of the interval.

Left and right censoring are special cases of interval censoring. In the case of left censoring, $L = 0$; and conversely in the case of right censoring, $R = \infty$.

To conceptualize interval censoring, we can consider a example study on virus testing, in which participants get their blood drawn in order to detect whether or not if they test positive for the virus or not. The random variable in question is T,

which represents the exact timepoint at which the subject contracted the virus. If an individual was first tested at time t_1 and tested negative, but was tested again at a later time t_2 and tested positive, the specific time t at which the subject contracted the virus is unknown. All we know is that it lies somewhere between the interval is $I = [t_1, t_2]$, but not the exact time at which they contracted it.

There are a myriad of types of censoring which can be discussed, however the focus of my thesis deals specifically with the challenges of reporting these values alongside methods of handling left-censored data.

2.2 Challenges of Reporting Censored Data

[CURRENTLY WORKING ON REORGANIZING HERE]

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 this disparity in practice within chemists and their labs and describe in the following list as:

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 the "less than" value, which is the numerical LOD preceded by a "<"
5. Reporting some value between 0 and the LOD, 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 [FIND PAPER BY Gilbert, 1987; Hunt and Wilson, 1986; and Rhodes, 1981 – backtrace through berthouex when time allows? by what metrics?].

Berthouex 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 the numeric values of measurements, even when those values are below the limit of detection. (Berthouex, 1993)

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. [WHAT SORT OF USEFUL INFORMATION IS DISCARDED, INCLUDE HERE, omitting is a valid method which is used by default in a lot of statistical software, when should care be used?]

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 Approach

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. These 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 are used often in non-rigorous

statistical settings due to them being quite easy to implement (Chen et al., 2011). As such, there have seen several studies in order to investigate the effectiveness of the method.

One particular investigation was conducted by Glass and Gray (2001) to investigate the effectiveness of LOD approaches, they used a variety of naive substitution methods from the values listed previously, with $\frac{LOD}{2}$ and $\frac{LOD}{\sqrt{2}}$. 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 usage of these substitution techniques.

2.3.2 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) 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 MLE and how it functions, given a random, independently and identically distributed (*i.i.d.*) set of random variables $X_1, X_2, ..., X_n$ from distribution $f(x|\theta)$.

For every observed random sample $x_1, ..., x_n$, we can define the joint density function to be:

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

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

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

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

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.

Maximum likelihood estimation is widely thought to be optimal, but only if one

knows the proposed model and underlying distribution of the dataset in advance, hence its classification as a parametric technique. In a study comparing methods to handling missing data, Canales found that the MLE method underperformed when the data in question was highly skewed, in which overinflated mean squared errors were often obtained (Canales, 2018).

The MLE method we will be utilizing is actually performed by obtaining regression estimates of slope(s) and intercepts through maximum likelihood with censored data. The `cenmle` function in the `NADA` package 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.

- Useful slides to refer to here [https://www.eurachem.org/images/stories/workshops/2017_10_PT/pdf/contrib/005-Mancin.pdf]

2.3.3 Kaplan-Meier Estimate Approach

[As a phenomenon, censoring is most often discussed in the branch of statistics known as survival analysis, which concerns itself with techniques to analyze a time to an event variable. As their name suggests, these variables measure the time which passes until some sort of event occurs. The type of event being observed need not be related to issues related to mortality, but it is certainly is most commonly employed in the health-care field. These types of events can be as innocuous as the time until device breaks, time until birds migrate away from their homes, or even things like time until an ice cream scoop falls onto the pavement. Regardless of which, all of these scenarios share a common flaw in terms of the possibility of the data being "censored."] [need to fix bracketed into existing paragraph]

The Kaplan-Meier method is a common nonparametric technique used to deal

with censored data. Originally developed to handle right-censored survival analysis data, an offshoot method in the form of the Reverse Kaplan-Meier Estimator have sprung up as a way to handle left censored data as well (Gillespie et al., 2010). The advantages of the KM method lie in its robustness as a nonparametric method; it performs well with a wide range of distributions. Many recommend its usage for when there are cases of extreme/severe censoring as a result of this [Canales2018].

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 "time to a certain event occurring", often being cases such time to death, or time to failure.

- [INSERT PICTURE OF EXAMPLE SURVIVAL CURVE HERE]

The KM-estimator is a nonparametric statistic used to estimate the survival curve from the empirical data while accounting for the possibilities of certain values being censored (participants in a mortality study could drop out, die during the study, become unavailable to contact after a certain time, etc.). It does this by assuming that censoring is independent from the event of interest (death) and that survival probabilities remain the same in observations found early in the study and those recruited later in the study [CITE PROPERLY https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_survival/BS704_Survival_print.html]

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

$$\hat{S}(t) = \prod_{x_j \leq t} (1 - \frac{d_j}{y_j})$$

where x_j is the distinct event/death time, d_j is the number of event/death occurrences at time x_j, and y_j is the number of followup times (t_i) that are $\geq x_j$ (how many observations in sample survived at least/or past the time t_i). [CITE PROP-

ERLY WHEN TIME ALLOWS https://www.youtube.com/watch?v=NDgn72ynHcM&t=398s&ab_channel=mathetal]

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 yields the reverse Kaplan-Meier approach, which allows for the estimation of the survival curve with left-censored values. This approach follows exactly the same logic as the Kaplan-Meier estimate of the survival curve, except we reverse the censored indicator and event of interest indicator. In other words, our censor is now the event and the event is now censored. This allows us to estimate the distribution function and population percentiles for data containing left-censored values (Gillespie et al., 2010).

For our analysis, we will be using the `cenfit` function from the `NADA` package in R to estimate the empirical cumulative distribution function (survival curve) for our left-censored data using the reverse Kaplan-Meier method. Similarly to the MLE method, the KM method is not an imputation method, so we are not replacing censored values with an imputed value, but rather estimating descriptive statistics for the entire dataset – including the censored concentrations (Canales, 2018).

2.3.4 Regression on Order Statistics

In between both the parametric nature of the MLE approach and nonparametric of the Kaplan-Meier estimator is the regression on order statistics (also known as ROS) method. As its name suggests, ROS is a semi-parametric method. It assumes that the censored measurements (emphasis on ONLY the censored, this what makes it semi-parametric) in the data comes from a normal or lognormal distribution.

In the ROS method, in order to model the distribution of the censored values, a linear regression model is created by plotting the uncensored observed values (ordered

from smallest to largest) vs. the quantiles (also known as "order statistics"), which is then used estimate and impute the values of the censored data (Lee & Helsel, 2005). These imputed values for the censored portions of the data are then combined with the known values of the uncensored bits, which allows for the computation of the descriptive statistics of interest. In summary, ROS imputes the censored data using the estimated parameters from the linear regression model of the uncensored observed values versus their quantiles.

There are of course, some requirements which must hold in order for ROS to be utilized: at minimum, there needs to be at least 3 known values and more than half the values within the data set must be known. As regression is utilized in this method, additional assumptions in the ROS method are shared with those necessary for linear regression to be performed as well. The response variable must 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.

Chapter 3 Simulations

[short passage describing what we hope to gain from performing a simulation study]

3.1 ADEMPS

[discuss ademps approach to designing a simulation study] (Morris, White, & Crowther, 2019)

3.1.1 Aims

3.1.2 Data-Generating Mechanisms

3.1.3 Estimands

3.1.4 Methods

3.1.5 Performance Measures

3.2 Results

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

3.3 Discussion

[discuss findings from the simulation study. are the results expected from knowledge gained from literature search? are they different?

3.3.1 Limitations

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

3.4 Study on Real Data

[connect back to chapter 1]

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

Berthouex, P. (1993). A Study of the Precision of Lead Measurements at Concentrations Near the Method Limit of Detection, *65*(5), 620–629.

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.1002124

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. (2020). Disposal of Coal Combustion Residuals from Electric Utilities Rulemakings. Retrieved from https://www.epa.gov/

coalash/coal-ash-rule

Gillespie, B. W., Chen, Q., Reichert, H., Franzblau, A., Lepkowski, J., Adriaens, P., ... Garabrant, D. H. (2010). Estimating Population Distributions When Some Data Are Below a Limit of Detection by Using a Reverse Kaplan-Meier Estimator. *Epidemiology*, *21*. http://doi.org/10.1097/EDE.0b013e3181ce9fD8

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.

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.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. http://doi.org/10.1002/sim.8086