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.

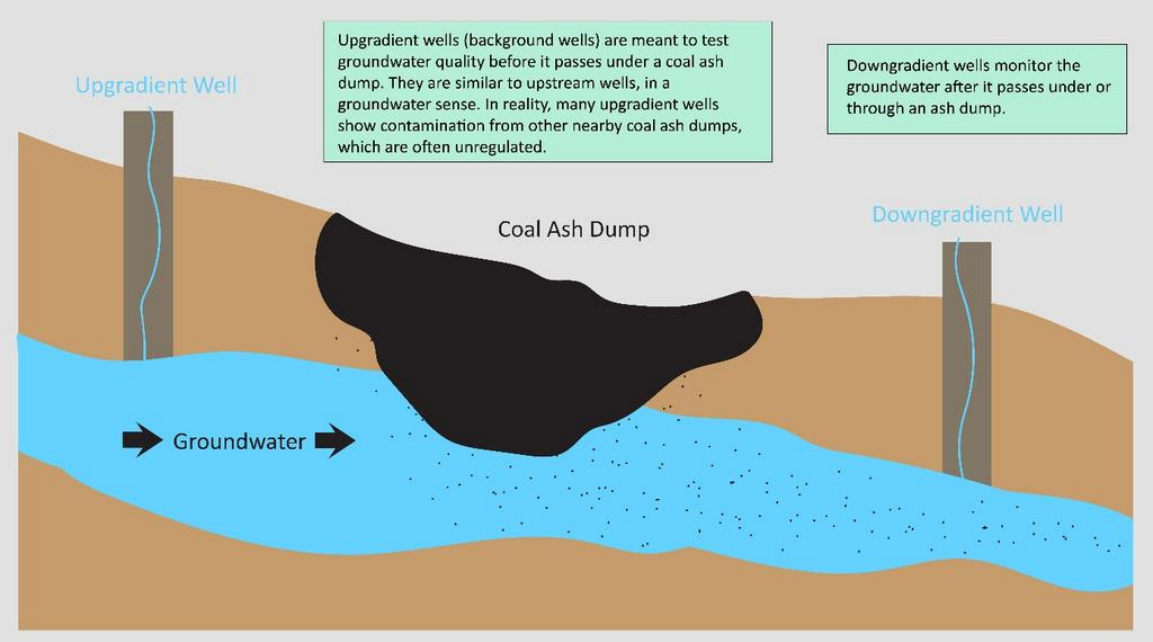

Figure 1.1: Difference Between Upgradient and Downgradient Wells

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 idea of missingness and incompleteness is commonplace throughout our world, and is even more prevalent within the statistical world. Known more formally as censoring, this condition exists when we have incomplete information regarding the values of a measurement within a dataset. Before we delve into an assessment of the methods commonly utilized to account for this, we will develop a general idea of what censoring and the problems it can cause.

2.1 Overview

Censored data is any data in which the value of a measurement is only known to a certain extent. 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 at which an device breaks, time at which birds migrate away from their homes, or even things like time at which 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."

There are a myriad of types of censoring which can be discussed, however the focus of my thesis deals specifically with left-censoring. This specific instance of censoring occurs when we do not know the the true value of a data point, but only know that it falls below a certain threshold which we call the "limit of detection." 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 wake up every morning far before the sun rises, at 3 A.M. and make sure to stay outside to witness the specific time at which the sunrise is able to be seen, recording that time. However, on the first day of the study, you oversleep and wake up at 7:00 AM, 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 value. All we know is that by 7:00 AM, the sun had already risen. The time at which the sunrise occurred could be any time before 7:00 AM, we have no knowledge on when that time could be. This form of censoring can often prove to be a thorn in one's side, as the lack of certainty that one have in the measurement adds a layer of complexity which must be accounted for within an analysis.

2.1.1 Common malpractices from non-statisticians

Expanding on the discussion on the limit of detection, which will be used interchangeably with the abbreviation *LOD* in the future, the LOD is a concept closely linked to the field missing data. LOD values are often overlooked as a result of misguided practices (in terms of statistical analysis) by some non-statisticians, which can bring about faulty analysis and/or conclusions which are heavily flawed.

One of the most common malpractices used in order to account for left-censored data is by omitting censored values from the analysis, which is of course the easiest way to deal with them – but this approach discards a myriad of useful information. In

a study conducted by (Berthouex, 2020), the researchers specifically gave instructions to participating laboratories to report a numeric value for each sample regardless of whether or not the value is below the detection limit which was followed by all but one laboratory. Not reporting the numeric values of those below the detection limit seems to be a common practice in the fields outside of statistics due to misinformation which is a practice which needs to be discontinued. 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).

Another issue lies within the art of reporting values which fall below the LOD. Amongst chemists, indicating whether or not if a value is below the LOD varies widely across labs and as reporting practices are not standardized. Some laboratories may explicitly record `ND`, others may put down `<` followed by the smallest recorded value to indicate that an observation's value is left-censored, others may simply omit the value completely (Berthouex, 2020). The lack of a universal reporting practice for values below the LOD is something which fosters a breeding ground for bad reporting habits among researchers.

On the same vein, reporting limits can often be misunderstood by non-statisticians. In an article published by the American Bar Association, a scenario is played out in which which a company seeking to purchase a gasoline service station obtains a laboratory report of the chemical concentrations of copper in that area, which was found to be `ND`, also known as non-detect. The company and its lawyers mistakenly interpret this `ND` as nullity, when in fact it only means that the measurement is not detected by the devices used by the laboratories to measure the chemical concentrations (Elias & Goodman, 1999).

2.2 Approaches

As discussed in section 2.1.1, there are a variety of sound and unsound statistical treatments for censored data which have been popularized in the statistical community to treat censored data. Discussed briefly previously, 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 *complete-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.

Apart from complete-case analysis, which is of course the most natural idea which pops up in our minds when discussing topics involving missingness and censoring. 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. Some of the most common methods to estimate descriptive statistics involving censored data include but are not limited to: substitution, maximum likelihood estimation, Kaplan-Meier, regression on order statistics, and of course distributional based multiple imputation methods (Lafleur et al., 2011).

2.2.1 Substitution Approach

Often condemned in papers, and rightly so, as a statistically unsound method to handle censored data, substitution methods are unfortunately ubiquitous in many fields outside of statistics as a way to handle censored data sets, often being cited

in environmental science papers as a method to work with left-censored chemical concentration data (Canales, 2018).

In analytical chemistry, a limit of detection is defined as:

$$LOD = \mu_{blank} + K\sigma_{blank}$$

where the distribution of the blank is assumed to be Gaussian, with mean μ_{blank}, standard deviation σ_{blank}, and K representing a "definition-specific constant," which is usually between the range 2.0 to 3.0. Ideally the blank will contain as little of the analyte of interest as possible, as it serves as the control and the basis as to which samples are being compared to. With a K = 3, it is to be expected that around 99.7% of the observations from a blank sample will be below the limit of detection as per the empirical rule for a Gaussian distribution (May, 2012).

Once the LOD is determined for the study, the substitution method simply involves imputing in a replacement value in lieu of the censored data point.

This replacement value used may differ between studies but common values include: $\frac{LOD}{2}$, $\frac{LOD}{\sqrt{2}}$, or LOD. Of course there may be more out there, but 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).

In a study performed by Glass to investigate the effectiveness of LOD approaches, they used a variety of naive substitution methods from the values listed previously. The investigated substitution enthusiasts' claims of certain replacement values being more apt for certain types of data sets. These proponents of the method claim that the replacement value $\frac{LOD}{2}$ is useful for data sets in which lots of data are below the LOD or when the data is highly skewed with a geometric standard deviation (a

measure of spread commonly used in tandem with log-normal distributions) of 3 or more. On the same note, they claim that the $\frac{LOD}{\sqrt{2}}$ is helpful for cases when there are only a few data points below the LOD or when the data is not highly skewed. From Glass' results, it was found that both of these methods are equally unsound in their reasoning and logic as they both introduce large errors and biases regardless of the data set being used (Glass & Gray, 2001).

2.2.2 Maximum Likelihood Estimation

- Maximum likelihood is a parametric technique which allows us to estimate the parameter values of a distribution/model.

- [needs more theoretical exposition, walk through general theory, if we have observed $x_i, ..., x_n$ from $X \sim f(x|\theta)$ and we want to estimate θ]

- As an example, given a few data points which we know have been generated from a Gaussian(0, 1) distribution, this technique seeks to calculate the maximum likelihood estimates of the paramter values for the Gaussian distribution, μ and σ.

- We need to calculate the total probability of observing all the data (the joint probability distribution of all data points). This is the product of the marginal densities assuming i.i.d.

This gives us a likelihood of

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

which we can maximize in order to obtain our MLE estimate for the parameter of interest. [CITE RICE TEXTBOOK]

- 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 (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 accomplishes this and allows us to calculate descriptive statistics for the entire dataset.

- Useful slides to refer to here [https://www.eurachem.org/images/stories/workshops/2017_10_PT/pdf/contrib/O05-Mancin.pdf]

2.2.3 Kaplan-Meier Estimate Approach

- Kaplan-Meier approach is a common nonparametric technique used to deal with right-censored data, but it can also be used for left-censored values as well – in the form of the reverse Kaplan-Meier estimator.

- What is the Kaplan-Meier estimator? In survival analysis studies when the focus is a type of "time to a certain event occurring," most often things like time to death, or time to failure.

- If we think about it from a mortality perspective, the survival function is a function which gives the probability of survival over time

- [INSERT PICTURE OF EXAMPLE SURVIVAL CURVE HERE]

- The Kaplan-Meier estimator is a nonparametric statistic which is used to estimate the survival function/survival curve from our 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.)

- KM method 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 WHEN TIME ALLOWS https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_survival/BS704_Survival_print.html]
- $\hat{S}(t) = \prod_{x_j \leq t}(1 - \frac{d_j}{y_j})$
- where x_j is the distinct event/death time, d_j is the number of event/death occurences 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 PROPERLY WHEN TIME ALLOWS https://www.youtube.com/watch?v=NDgn72ynHcM&t=398s&ab_channel=mathetal]
- This will give an estimator of the survival curve
- Suggests using the reverse Kaplan-Meier (KM) estimator to estimate the distribution function and population percentiles for data where there is "left-censored data" (data point is below a certain value but known by how much) (Gillespie et al., 2010)
- Reverse Kaplan-Meier 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 reverse Kaplan-Meier, our censor is now the event and the event is now censored [CITE PROPERLY WHEN TIME ALLOWS https://www.pharmasug.org/proceedings/2019/ST/PharmaSUG-2019-ST-081.pdf]

- The advantages of the KM method lie in its robustness(as a nonparametric method), it performs well with a wide range of distributions. It is also good to use when there are cases of extreme/severe censoring in the dataset due to its nonparametric nature [Canales2018]

- 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. 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)

2.2.4 Regression on Order Statistics

- 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 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.

- ROS method is based on simple linear regression, detected values are ordered from smallest to largest, and then quantiles are used to estimate the concentration of censored values

- In summary, ROS imputes the censored data using the estimated parameters from the linear regression model of the uncensored observed values versus their quantiles [CITE https://www.eurachem.org/images/stories/workshops/2017_10_PT/pdf/contrib/O05-Mancin.pdf]

- Values are plotted on a on a probability plot and a linear regression line is calculated, in order to estimate the parameters of the distribution from which the values came from. Then, values are pulled from the assumed distribution in order to impute in value for the censored measurements. These imputed values and combined with the known values in order to obtain descriptive statistics of interest for the data [CITE PROPERLY WHEN TIME ALLOWS https://www.itrcweb.org/gsmc-1/Content/GW%20Stats/5%20Methods%20in%20indiv%20Topics/5%207%20Nondetects.htm#:~:text=Robust%20ROS%20is%20semi%2Dparametric,are%20made%20for%20the%20nondetects

2.2.5 Distribution-Based Multiple Imputation Approach

- Study exploring different options to handle LOD laboratory data – specifically with regards to multiple imputation methods for left-censored data. They concluded that "the distribution-based MI method" worked well for bivariate data where the values were < LOD. (Chen et al., 2011)

- They also investigated distribution-based multiple imputation methods – they used MLEs to estimate distribution parameters based on all datas (< LOD and those not). They repeatedly imput the values to create multiple complete sets of data, and then analyzed each one individually (Chen et al., 2011)

- Mathematically, they created a log-likelihood function with all the data, then derived MLEs of each parameters on multiple bootstrapped datasets. Each bootstrap data gives different estimates for the mean, sd, etc. (refer to article for math) (Chen et al., 2011)

2.2.6 Write about other approaches here

- Another method is "Cohen's Method" where one extrapolates the left hand side of distribution based on the distribution of the uncensored data and then calculate the MLE estimate of the arithmetic mean – found to be unreliable with data with outliers, this method can ONLY be used with data where there is a single LOD. From their study they found that this method gave high, unlikely results of the mean (Glass & Gray, 2001)

- Imputing from a uniform is useful when you don't know the distribution of the data (Canales, 2018)

2.2.7 Comparison between different methods

- Compared 5 methods: found that in terms of performance: imputation method using MLE to estimate distribution parameters and then imputing censored data points with values from this distribution below the LOD > imputation from a uniform distribution > other 3 methods (substitution, log-normal MLE to estimate mean and SD, and kaplan-meier estimate) (Canales, 2018)

- KM method is better than MLE for data where there are TONS of missing data or if data is highly skewed (distribution not assumed in KM method) (Canales, 2018)

- Uses RMSE (root mean squared error) to see how close the estimated values are to the true values (lower RMSE means closer estimation to known values) (Canales, 2018)

2.3 Closing Remarks

- Substitution is commonly cited as the worst way, nonparametric ways do better, maximum likelihood methods are the best (Lafleur et al., 2011)

- Of course, these methods aren't the only way to deal with it, but it is what we have come up with for now. Detection limits are constantly changing as time passes and because of that – techniques and ways to combat it will gradually evolve and change (for the better) to deal with it. Detection limits are constantly changing: as technology improves, so too does our ability to accurately measure substances (Elias & Goodman, 1999)

2.3.1 Packages

- Probably will not be included, just here for my own reference!

- "MLE and KM methods were implemented using the NADA package (https://cran.r-project.org/web/packages/NADA/NADA.pdf) in R (`cenmle` and `cenfit` functions) where data is labeled as censored or uncensored, for censored values, LOD is used as a placeholder, since these methods aren't imputation methods – these censored values weren't replaced. Instead, summary statistics were generated with the entire data set (including the censored data)" (Canales, 2018)

- "The imputations methods used mostly followed the general ideas: we have to assume the entire data set follows a particular distribution. Then we use this distribution to impute in values for the censored data. The MLE imputation method uses MLE methods to estimate the parameters of a distribution to fit the dataset, then values lower than the LOD are imputed FROM this dataset

for all censored values (they used the function `fitdistcens` from the R package `fitdistRplus`). The second uniform imputation method assumes a uniform distribution with minimum 0, maximum LOD – for all values less than the LOD, then the left-censored values are replaced with a number randomly selected from this uniform distribution." (Canales, 2018)

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. (2020). 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

Elias, D., & Goodman, R. C. (1999). When Nothing Is Something : Understanding Detection Limits, *13*(4), 519–521.

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

Glass, D. C., & Gray, C. N. (2001). Estimating mean exposures from censored data: Exposure to benzene in the Australian petroleum industry. *Annals of Occupational Hygiene*, *45*(4), 275–282. http://doi.org/10.1016/S0003-4878(01)00022-9

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

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