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.

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}

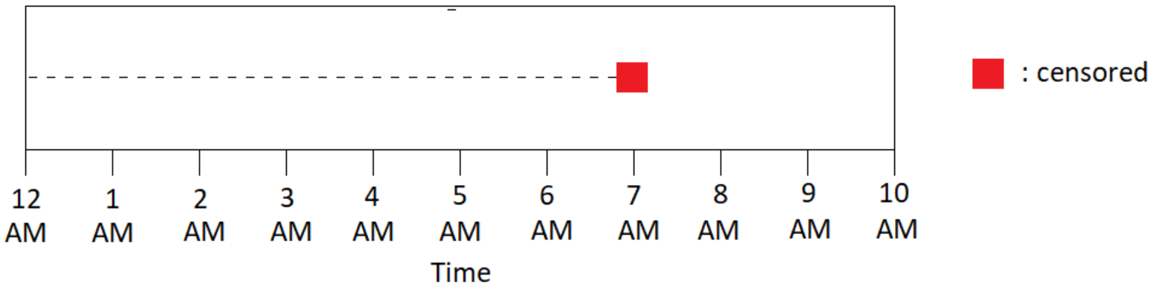

Figure 1: Left Censoring Example

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 [@Hall2020].

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.

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 [@Chen2011]. 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 [@May2012]. 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 [@Lafleur2011].

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 [@Canales2018].

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 [@Lee2005]. 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 [@Crovelli1993]. 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 [@Chen2011]. 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 [@Hornung1989]. 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 [@Lee2005]. 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.

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 [@Ganser2010]. 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 [@Bolks2014].

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 [@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 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) [@Klein2003].

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:

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

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.

[@EPA2009] 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 [@EPA2009], 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 [@Lee2005].

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 (?)]