Methodology

ESTABLISH PROBLEM OF MISSING DATA (start with how it might work?) WHAT IS CENSORED DATA (BROADLY) STRUCTURE OF THIS CHAPTER

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.* Missing data

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 is and the problems it can cause.

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.

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 [@Berthouex2020], 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 [@Chen2011].

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 [@Berthouex2020]. 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 [@Elias1999].

Approaches

As discussed in section @ref(malpractices), 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 [@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.

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 [@Lafleur2011].

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 an appropriate (and even recommended) method to work with left-censored chemical concentration data [@Canales2018].

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 [@May2012].

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. Different disciplines have their own suggested "best," replacement value to use [@Lee2005]. 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 [@Chen2011].

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 [@Glass2001].

Maximum Likelihood Estimation

Maximum likelihood is a parametric technique which allows us to estimate the parameter values of a distribution/model and one which is useful when encountering censored data.

To give a brief introduction as to the the mechanisms of MLE and how it functions, given a random *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 (mle) 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 [@Canales2018].

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/O05-Mancin.pdf]

Kaplan-Meier Estimate Approach

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 [@Gillespie2010]. 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 PROPERLY 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 [@Gillespie2010].

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 [@Canales2018].

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 [@Lee2005]. 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 [@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.