Simulations

Having discussed the various methods to handle left-censored data in the previous chapter, we now turn to a simulation study. As encounters with missingness in data can be quite varied, our goal is to evaluate and confirm the strengths and weaknesses of each of the methods when working with datasets of varying censoring rates and sample size. We will also discuss the implementation of the methods, data generating mechanisms, and evaluation metrics to be used to assess the performance of each method.

Aims

[Sentence about how lots of papers have sought to try to compare and decide which method was the best to use for left censored data, but there is a general lack of consensus on the ideal method to use (apart from a common discourse discouraging the use of substitution. . .)]

Through our simulation study, we wish to identify settings where a method can be effective but also those in which the methods may not be able to perform quite as well.

Several investigators in this field have found issues with certain methods underperforming under certain conditions, and brings up the possibility of particular methods being more equipped than others to deal with different rates of censoring. Specifically, a study on methodologies to handling left-censored microbial risk assessment conducted by [@Canales2018] found that the substitution method seemed to work much better than expected while other methods, such as the MLE method, seemed to have trouble when applied to highly skewed data. Results from other works [@Antweiler2015] suggest that regardless of the method being utilized, obtaining reliable estimates from datasets where censoring was greater than 40% was unfeasible. This particular study also suggests that the size of the data had no influence on the result of the estimate in question.

Claims regarding the effectiveness of methods with regards to censoring rates, distribution of data, and sample size are all highly contentious. In order to get a better idea of sense of how these claims hold up, our goal is to evaluate the validity of those claims by conducting a simulation study of our own which will put those claims into practice. We must note that we don't expect one method to be significantly above the others in terms of performance for all settings, but we do hope to see which method might be best for certain settings. This approach and aim of our simulation study will then largely depend on how we will generate the data to be used in order to achieve our goal (of checking which methods work well with that settings)

Data-Generating Mechanisms

The data generated for use in our study will be obtained by using parametric draws from author-specified distributions (log-normal, exponential, and Weibull), the methods utilized can only be used with non-negative distributions [@Yavuz2017]. Our data-generating mechanism alters criterions such as the our sample size, $n_{obs} = \{10, 100, 1000\}$ and censoring rate $R = \{0.10, 0.20, 0.30, 0.40, 0.50\}$.

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

Estimands

Each of the four methods discussed in the previous chapter are designed for usage in obtaining summary statistics for left censored data [@Shoari2018]. in our simulation study, we want to evaluate just how well these methods are able to estimate a population quantity. We will be using the sample mean as an estimator for our estimand, the population mean, μ.

Performance Measures

Morris et al. (2019) define performance measures as numeric metrics used to assess the performance of the method in question. The criterions we will use to assess the performance of each of our four methods will

consist of: bias, variance, and mean squared error (MSE).

Variance

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

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

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

Estimators with low variances generally remain close in value throughout all samples, while those with high variance may wildly differ between samples. As such, it is generally preferable to have an estimator with low variance. However, it is important to note that precision measurements, such as the variance, are not a sole indicator of an estimator's performance [@Walther2005]. While precise estimators are ideal, it is also important to assess the estimator's bias, how close it is to the true value.

Bias

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

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

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

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

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

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

Mean Squared Error (MSE)

We generally would like estimators which have low bias and low variance, but it can be difficult to achieve both at once. As such, it is common to instead turn to a quantity known as the mean squared error (MSE), which is a quantitative measurement used to assess the accuracy of an estimator. The MSE measures how far away, on average, an estimator is from its true value.

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

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

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

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

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

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

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

Results

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

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

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

- km method for all censoring rates and all sample sizes gave large overestimates, had very large, positive bias values compared to all other methods

- variance is largely dependent only on sample size (to be expected)

- in the case of high censoring (0.50) and small sample size (10), km did not work as well as other methods

- in the case of medium censoring (0.30) with medium sample size (100) and high sample size (1000), ros performed the best, with mle following closely as compared to the other two methods. ros and mle had bias and mse values which were lower than the other two.

- in the case of high censoring (0.50) and large sample size (1000), ros performed SIGNIFICANTLY better than all other methods (low bias, low mse)

Discussion

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

Limitations

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

Study on Real Data

[connect back to chapter 1]