*note that all writing in this section is throwing ideas from head to paper w/ no concern for grammar, spelling, cohesiveness, etc. very, VERY, rough draft – 100% just to get writing down

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

The aim of our study follows a proof-of-concept idea. We are hoping 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. Literature [INCLUDE LITERATURE HERE] have frequently discussed things such as {method] being unviable in cases when censoring is greater than ___%, or [another example]. 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 producing parameteric draws from a author-specified distribution. Our DGM will alter things(?) such as the our sample size, n_{obs}, censoring rate R, and distributional parameters, such as the mean and standard deviation for the normal distribution.

Censored values are generated and determined by first arranging the uncensored observatiosn in ascending order, and then from the censoring rate (ex: 0.15), we will make the lowest 0.15 of the observations censored, while leaving the remaining, uncensored.

The methods developed to handle left censoerd data can only be used with non-negative distributions (cite here), and as such the distributions we are used will only be the distributions such as the lognormal distribution, weibull distribution, exponential distribution, etc.

Estimands

each of the four methods discussed in the previous chapter are designed for usage in obtaining summary statistics for left censored data [FIND SOURCE]. in our simulation study, we want to evaluate just how well these methods are able to estimate a population quantity. Our estimand in this case will be 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.

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 an estimate which have low bias (accurate) and have low variance (precise), 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 makes use of both variance and bias in its calculation.

The MSE measures how far away, on average, an estimator is from its true value.

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

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

It is desirable to attain low MSE values.

DELETE NOTES BELOW

as the estimand in our simulation study is the population mean μ, let $\hat{\mu}$ represent the estimator for our estimand.

Results

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

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]