Bias developed due to Censorship & Truncation of Big Datasets

Cecilia Ferrando
University of Massachusetts, Amherst
cferrando@cs.umass.edu

Gaurang Chaudhary
University of Massachusetts, Amherst
gaurang@umass.edu

Abstract

The goal of this study was to determine how much bias do we get after we either censor a dataset or truncate it. It is quite important to understand what the consequences are of these two different processes and how our new dataset behaves in comparison to the original. We approach this problem by developing a good understanding how our dataset changes over different trials.

Background

The focus of this research is to know how the process of truncation and censorship shifts us away from the true values. But before we do that, it is crucial to understand what is done by these processes.

Censorship: Values above (or below) a certain threshold in a dataset are transformed (or replaced) by the threshold value.

- Censorship = threshold T, any value > T we make it equal to T

Truncation: Values above (or below) a certain threshold in a dataset are dropped (or removed) from the dataset.

- Truncation = threshold T, any value > T we drop it

Bias: Difference between the true value and the estimated value. It is important to note that the way we calculate bias is by taking mean of all the values in the dataset and then calculating the difference between our estimate (lamba_hat, mean_hat) and the true value.

Analysis

The research centered on two main factors of the dataset, the threshold limit and size of the sample. We wanted to analyze their behavior and how it affects our results with different values. Throughout our trials, we used a random poisson dataset and observed the results we got. The dataset was then run through two separate experiments, one focusing on different thresholds with a fixed sample size and the other one taking a fixed threshold with different sample sizes.

Experiments

Experiment #1:

- Take different levels of poisson parameter: T = 5, 6, 8, 10, 20, 50
- Generate a poisson dataset (N > 1000)
- Censor/Truncate the dataset
- Take the scalar average and find the estimates
- Compare the estimates with the true parameter to find the bias

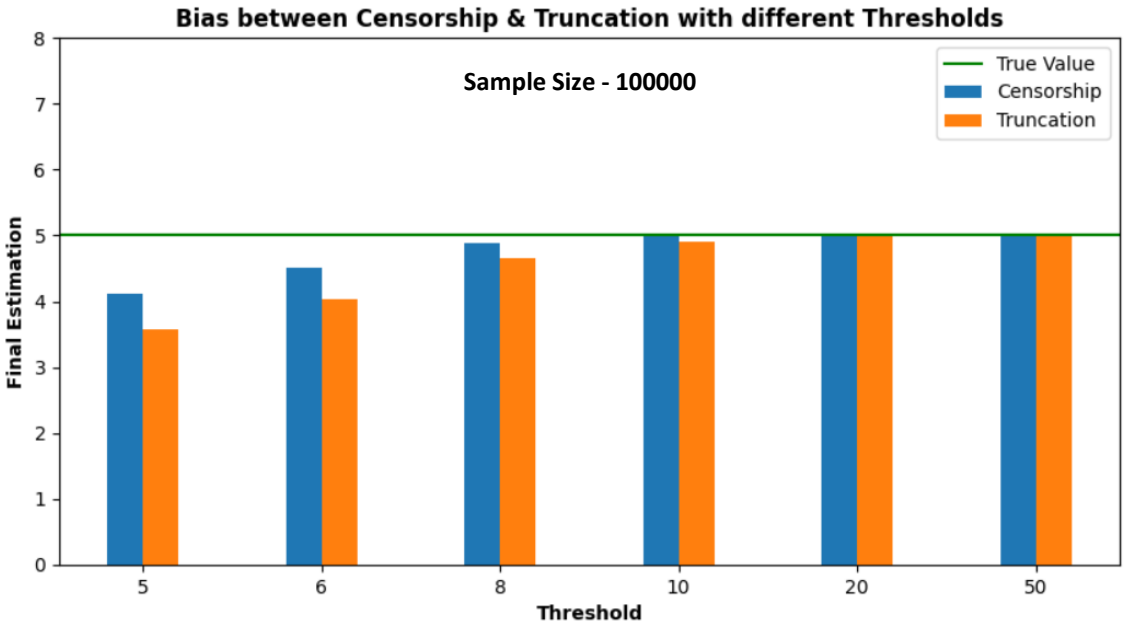

Fig 1: Analyzing the bias with different thresholds for a given dataset.

Experiment #2:

- Take different sizes for the sample: N = 10, 100, 1000, 10000, 100000
- Generate the poisson datasets
- Censor/Truncate the dataset for a fixed threshold for all the samples
- Take the scalar average and find the estimates
- Compare the estimates with the true parameter to find the bias

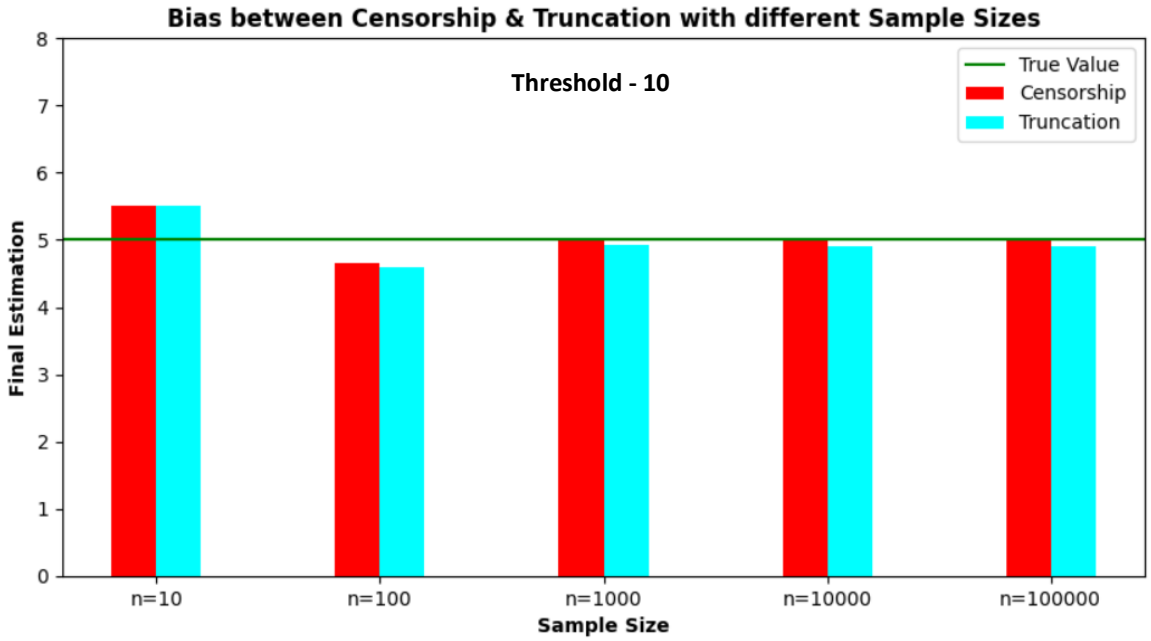

Fig 2: Analyzing the bias with different sample sizes with a fixed threshold.

Conclusion

We developed a better understanding of two main factors of the dataset, the threshold value and the size of the sample. First, for a particular dataset, the closer the threshold value is to our true parameter, the more bias we develop in both censorship and truncation. Although censorship works better and develops less bias, especially if the threshold value is closer to the true value. For threshold values further away from the true, both processes work the same as there is not a lot of bias developed. Second, for datasets of different sizes, the larger the dataset, the less bias it develops. The bias developed for samples larger than 1000 was really low for both censorship and truncation but censorship was just a bit better.