Approximate Sampling Distribution of Sample Correlation Coefficient Under Type I Censoring

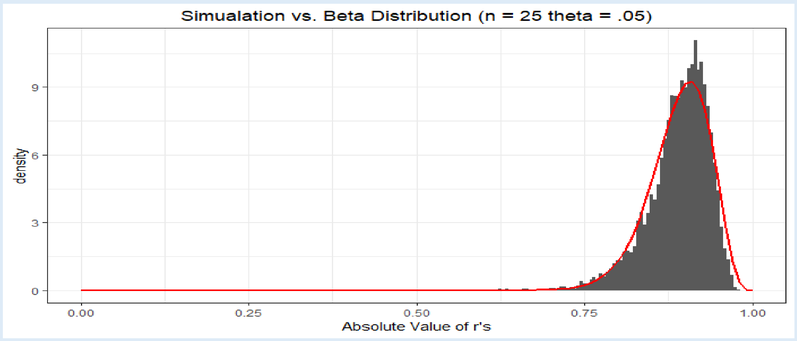
Jacob Tarnowski
St. John Fisher College
Faculty Mentor: Dr. Scott Linder

Abstract

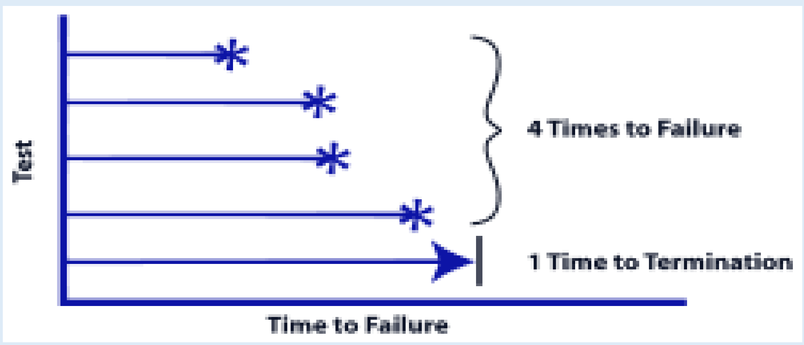
Suppose a random sample of size *n* is selected from a bivariate normal population and exposed to Type I (time constrained) censoring on one of the variates, so that cases associated with the values of one of the variates beyond time *T* are censored. The presence of censoring in this context renders intractable the sampling distribution of the sample correlation coefficient. Using simulation, we systematically examine the impact of censoring on the sampling distribution of the absolute correlation coefficient, $|r|$. We propose approximation of the sampling distribution of this statistic by the Beta distribution, whose parameters are determined as functions of the experimental conditions (*n*, proportion censored (θ), and *r*). These functions are regression models fit to the average maximum likelihood parameter estimates obtained through simulation. We examine the goodness of fit of this approximate sampling distribution, and also consider the relative error of estimation of percentiles of this distribution commonly necessary for inference.

Background

The beta distribution is a very flexible distribution whose shape depends upon two parameters, alpha and beta. The flexibility of of this distribution allows for one to easily fit a sample distribution with a beta distribution. The distribution also has a domain of (0, 1) which makes an ideal distribution to fit $|r|$ with.



Type I censoring has its benefits, such as time being fixed in the experiment, however it forces dependency upon data. This restricts one from generating statistics, like confidence intervals, that assume independence in the data.



https://en.wikipedia.org/wiki/Censoring_(statistics)

Motivation

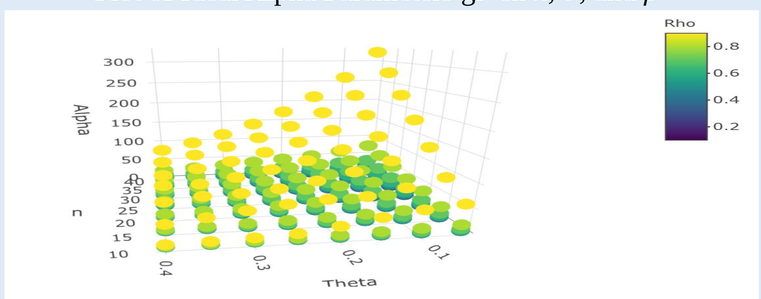
Type I censoring is very commonly used in industry, but in reality, the sampling distributions of commonly used statistics (correlation, regression slope, etc.) are simply unknown when data have been censored. Researchers have often (wrongly) applied traditional statistical inference methods to censor-impacted samples, effectively pretending that their sample was complete (and smaller). This is not something that has been widely studied. The sampling distributions critical to correct inference are typically mathematically intractable and require approximation via extensive simulation. We aim to provide the practical researcher with a tool that will enable inference for correlation coefficient that accounts for the impact of censoring.

Simulation

Example of Resulting Data Frame

n	Theta	MLE_Alpha	MLE_Beta	Rho
40	0.30	1.2358833	6.307155	0.1
40	0.35	1.2292776	6.041297	0.1
40	0.40	1.2210713	5.739114	0.1
10	0.05	1.1595629	2.570481	0.2
10	0.10	1.1437384	2.468638	0.2

Plot of Ideal Alpha Parameters given *n*, θ, and ρ

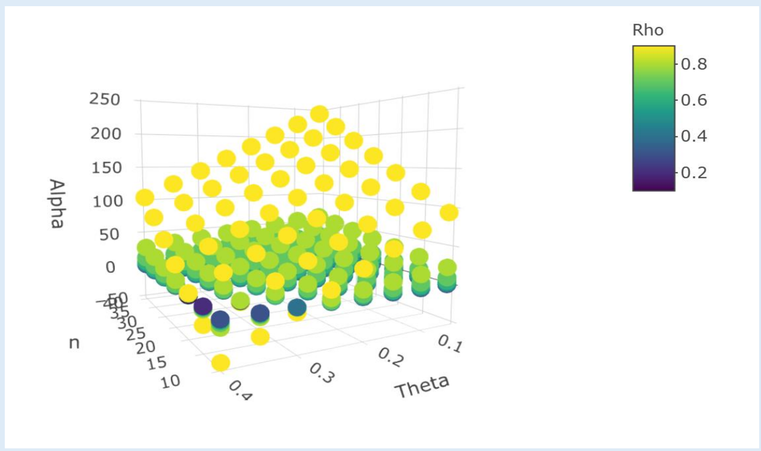


Fixing *n*, θ, and ρ, we simulated *n* (*x*,*y*) pairs from a bivariate normal population. We censored (removed) all pairs for which *x* exceeded the $1-\theta$ percentile of the marginal distribution of *x*. In this way we simulated 10,000 values of $|r|$. From this we computed maximum likelihood estimates of Beta distribution parameters. This process was repeated 400 times, and the average parameter estimates computed in order to minimize the impact of variation. After this process, we have a record of 'typical' Beta distribution parameter across the range of experimental conditions.

Modeling

The model we generated was created solely based of functions of *n*, θ and ρ. That being said, our model is totally created of terms such as $\rho^2\theta$ and $\rho^3 n$. It is important to note that we were not worried about overfitting the parameter estimates. Therefore our goal was to fit the model as tightly as possible to the parameter estimates.

Alpha Model Predicting Alpha at Select Values of *n*, θ, and ρ



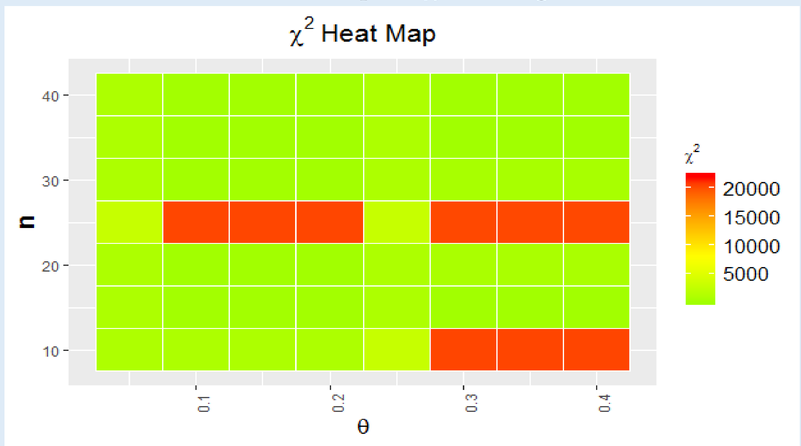
Acknowledgements

This project was funded by the National Science Foundation under Grant No. 1658998 (OWU REU).

Ohio Wesleyan University | NSF | St. John Fisher College

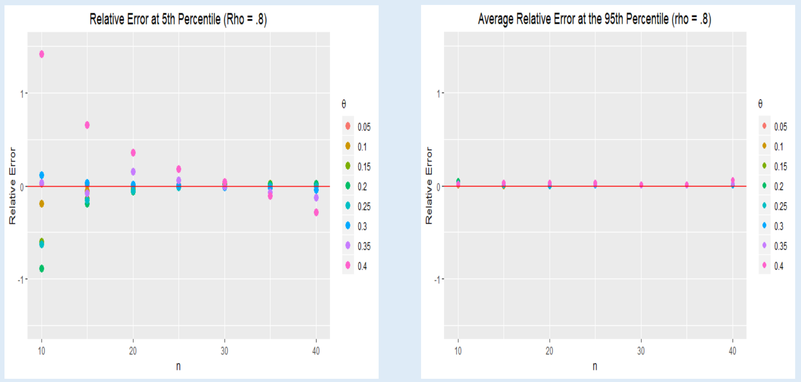
Goodness of Fit

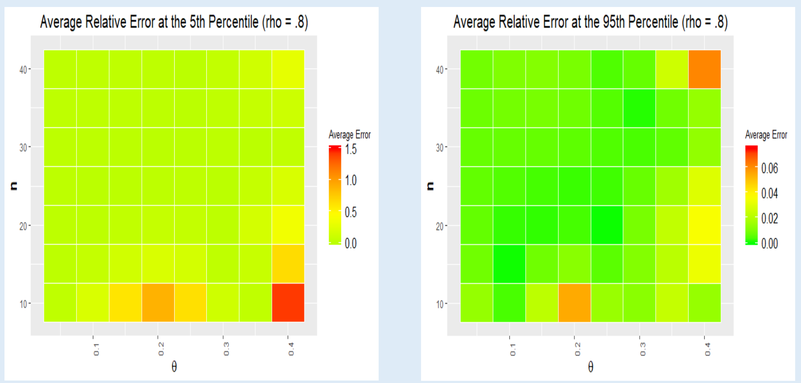
Heat Maps of $\chi 2$ Testing



For our first goodness of fit test we conducted a χ^2 test to measure the overall goodness of fit of our model. The ideal χ^2 would be 10 because when building the test we broke our model down into ten different sections. Taking that into account, the overall fit of our model is very poor.

Plots of Relative Error





For our second goodness of fit test we compared the beta distribution to simulated data with the same *n*, θ, and ρ parameters, and found the relative error of our model at various percentiles. By looking at the graphs above, the relative error tests suggest that our models fit performs well at given combinations of *n*, θ and ρ, even though the χ^2 suggest against it.

Future Work

In the future we will continue our work by continuing to build a better model. In an ideal world we would be able to build a model that would not produce any negative alpha values. If we were able to produce a model with out any negative alpha values would also allow us to get a better look at the goodness of fit our model. Aside from building a better model we also are looking to generate confidence interval for the absolute value of the correlation coefficient.