Robert Morris

Data Analytics – Project 2

Comparing Stability Of Estimators

In this project, I was to evaluate the stability of the Variance Maximum Likelihood Estimator and the Sample Variance Estimator. In oder to do this I first performed a theoretical proof to test the biasedness of the Variance MLE. After making my way through this proof on the 3 page of the pdf that was submitted, I have concluded that the Variance MLE is biased. I arrived at this conclusion because the expected value of the Variance MLE does not exactly return that of the Variance. Instead it ends with the Variance times the sample size minus 1 dived by n. Although it is biased it could still be a good estimator as sample sizes grow larger or approach infinity because the deviation from the original variance will get closer and closer to 0.

Apon coding this project in matlab, I approached it from a few angles. First, I started with 10 fixed sample sizes that were iterated over with each time providing a new calculation for the Sample Variance and the MLE Variance. On this plot you can see that there is little difference between the Samp variance and the MLE variance yet they vary around the original population variance until the sample size grows larger.

For the second graph I choose to compare the convergence of the MLE variance by treating it like a random variable itself. For each 10 different sample sizes I would ran 500 tests and from each test I calculated the Maximum, Minimum, and Mean of those MLE variances variables. As it is ploted it is clear to see that in very small sample sizes, when the maximum MLE outliers deviate quite a lot form the original variance. It is also very easy to see the convergence of this with less and less extreme outliers once the sample sizes grow much larger which shows that MLE Variance is consistent with larger sample sizes. For this reason I would prefer to use the sample variance on smaller sample size due to the correction using n-1.

For the third and forth graphs I chose histograms to display the normal distribution of the MLE Variance with a 10k sample size (left), compared to the normal distribution of the Sample Variance with a 10k sample size (right), the overlay tall bin in the middle of each histogram represents the actual overall population variance.

If the data included a small number of extreme outliers I would use Sample Variance because it would be more stable with smaller sample sizes. Although, if their were extreme outliers I believe it would be sufficient to use the MLE variance because the sample sizes would more accurately reflect the population.