# Simulation of Type I Error in Bootstrapping

*Daniel Hodges*

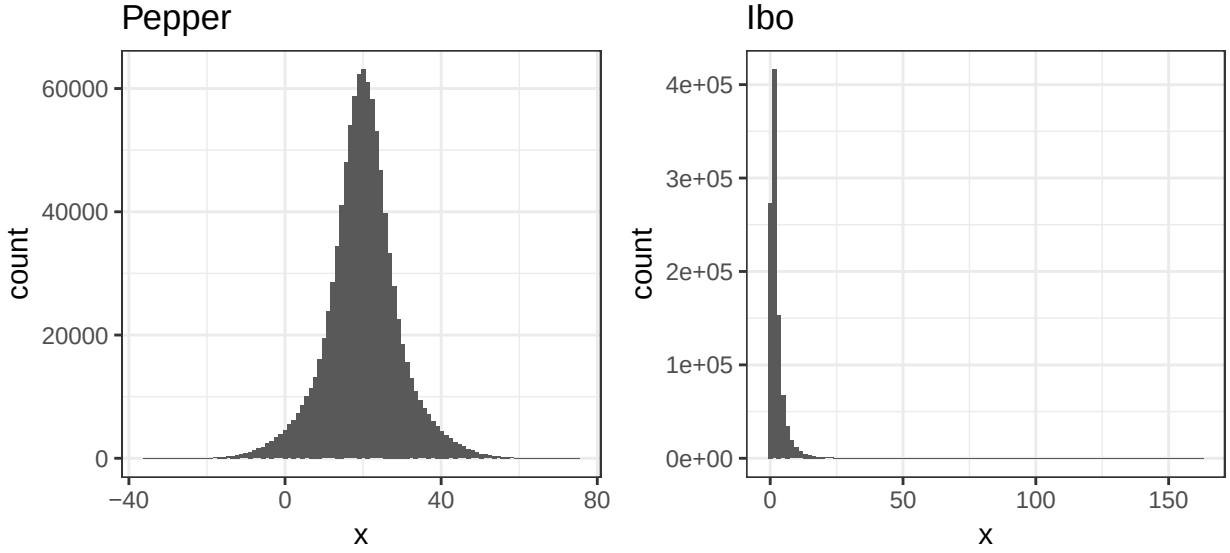
*12/2/2018*

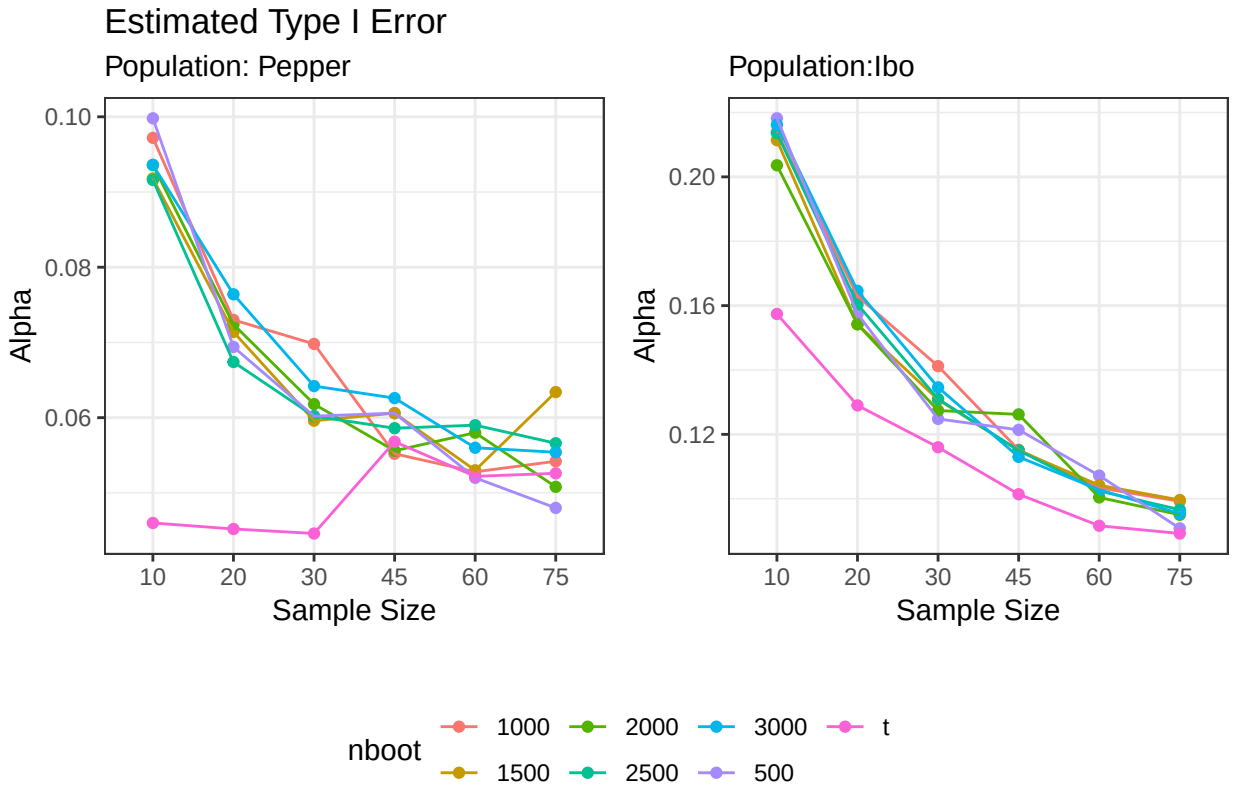
Bootstrapping is a non-parametric, simulation-based, alternative to traditional hypothesis testing. It is done by resampling from your original sample with replacement. By taking the desired sample statistic (e.g. mean, median, etc) of each of these bootstrap samples the distribution of that statistic is estimated and used to conduct a hypothesis test. The purpose of this project will be to estimate the Type I error rate of the bootstrapping testing procedure and comparing that to a more traditional test. What I've done involves bootstrapping the sample mean and comparing it to the t-test.

Bootstrap methods are becoming more commonly taught in entry level statistics courses. From my experiences courses based on the textbook "Unlocking the Power of Data" by the Lock familiy will teach bootstrapping, such as ST 351 and 352 here at Oregon State University. Critics are concerned that the details of the method's performance are left out of the education and strong assumptions are made when using a bootstrap distribution to estimate a population distribution.

In order to estimate the Type I error of the bootstrap procedure I've generated two data sets for use the in the simulation, named after my two cats. The first one (Pepper) is roughly symmetric based on a mixture of two normal distributions with the same mean, different variance. I expect the t-methods to have Type I error rate no larger than $\alpha = .05$ in this setting. So I would like to see how bootstrapping does in comparison and wether or not $\alpha$ is controlled at the desired rate for all sample sizes or simulation sizes. The second data set (Ibo) is based on a lognormal distribution and is very skewed. I expect the t-methods to perform poorly here at controlling $\alpha$ and am curious to see how bootstrapping performs.

The simulation was done by drawing a sample from a distribution, performing a bootstrap test of a given size, and recording the p-value. Type I error was estimated by the count of tests that rejected the true null hypothesis divided by the simulation size. This was done 5,000 times for samples of sizes 10, 20, 30, 45, 60, and 75; with bootstrap simulation sizes 500, 1000, 1500, 2000, 2500, 3000. This was a somewhat intensive simulation and a significant speed up was made. My original code was iterative rather than vectorized, and vectorization made the code over 80 times faster. Note that this speed up was for each individual bootstrap test, so the significance of this cannot be emphasized enough.






The results of the simulation are actually somewhat surprising. I was expecting the different bootstrap sizes to perform differently. I expected bootstraps with a low number of resamples like 500 or 1000 to have type I error rates larger than those with larger numbers of resamples. It seems like a bigger bootstrap isn't always better. If the resulting type I error is roughly the same then maybe the cost in computation time isn't worth it. Although with my code the cost in computation time is negligable.

One important result of the simulation to notice is that when the t-test is appropriate (i.e. in the symmetric population) it controls for type I error quite well. I believe this is due to the degrees of freedom in the t-test making the tails of the t distribution wider. Bootstrapping does not have that same protection. Type I error rates are quite inflated for sample sizes 10, 20 and 30.

I think the results of this simulation demonstrate that the bootstrap testing method will have type I error rate equal to $\alpha$, but only for large enough samples. It seems we will need a sample size roughly as large as the Law of Large Numbers requires in order for the type I error rate to be what we desire. Bootstraps based on smaller sample sizes resulted in higher type I error rates. I currently stand with the critics of teaching bootstrapping to introductory statistics students. This technique does not have all the same properties of other tests and understanding when it is appropriate to use is crucial. We should not gloss over these details if we are going to teach these methods. Doing so only teaches students enough information to be dangerous when doing research or studies of their own.