Bootstrapping

Sujan Bandyopadhyay

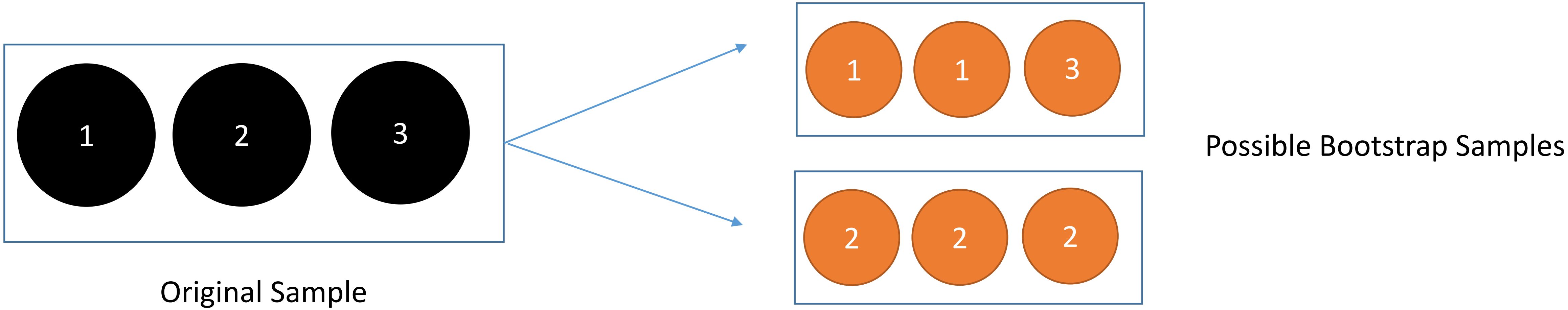
Motivation

- After estimating the estimator from a sample, we want to test the accuracy of it
- Usually, we employ assumptions about the nature of the distribution of the estimator; e.g. the estimator is normally distributed; in order to make inferences
- These are known as parametric methods
- Bootstrapping is a non-parametric method which requires no assumptions about the underlying distribution of the estimator; instead it uses the empirical sample to make inferences about the estimator

- B. Efron and R. Tibshirani (1986) – “*….all of the usual formulas for estimating standard errors…. are essentially bootstrap estimates carried out in a para- metric framework*.”
- Advantages: The researcher doesn’t need to make underlying parametric assumptions regarding the distribution of the estimator, and can be useful in cases when these assumptions are hard to make
- It would not work in cases when the sample is not representative of the actual population, and also in cases when there are dependence structures within the sample which makes independent sampling of the data fail

How does it work?

- You assume the sample to be a good approximation of the population, and then **sample with replacement,** generating the estimator for each statistic
- This will give you the distribution of the estimator, along with standard errors and confidence intervals
- Each bootstrap sample must have the sample size as the original data



- You have three balls with numbers 1,2,3
- You resample with replacement. So for each bootstrap sample, you pick up a ball randomly, with a probability of 1/3. Then you note the number of the ball down, and keep it back. Again, you pick up a ball random with a probability of 1/3. You keep going until you have 3 observations and then you move on to the next bootstrap sample.
- It is possible that observations are missing or repeated from the original sample.
- Over here, possible bootstrap samples include – 111, 112, 113, 122, 123, 133, 222, 223, 233, 333 (not accounting for order)

Example

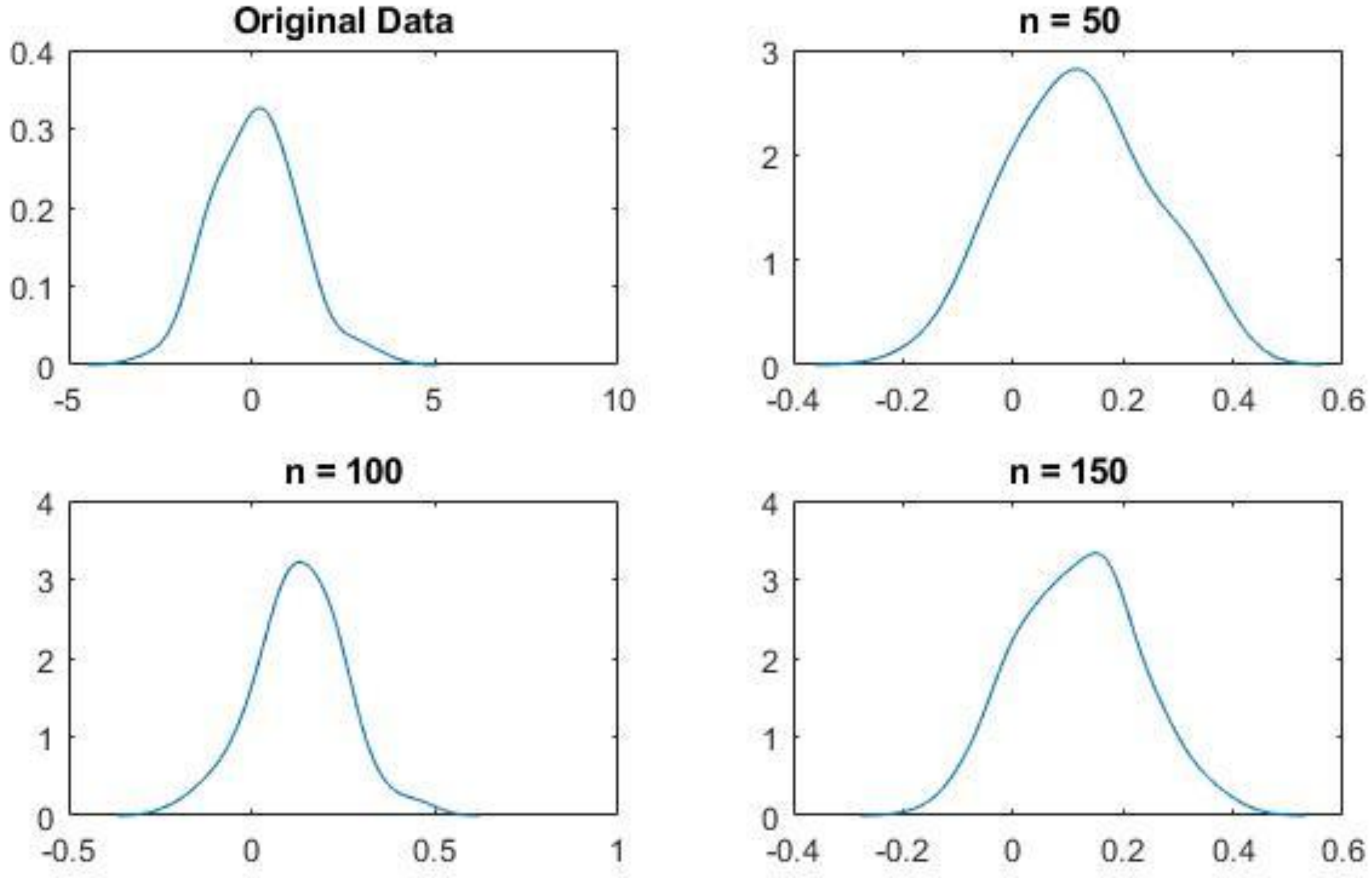
- Let’s test how bootstrap does in a case where we know the actual population statistic. We generate a random sample of a standard normal distribution (100 observations); Standard Normal has mean = 0, and standard deviation = 1
- Sample Mean = 0.1231, sample Std = 1.1624
- We know try out bootstrapping the sample to see what results we get, with no. of bootstraps (n) = 50,100,150

Results

	Actual	Sample	n = 50	n = 100	n = 150
Mean	0	0.1231	0.1212	0.129	0.1187
Std Dev	1	1.1624	1.1545	1.1621	1.1565
Std Error of the Mean			0.1247	0.1225	0.1099

Gives a good idea about how bootstrap works; keep in mind that we actually didn't need to bootstrap the data because we knew what the underlying distribution is

Distribution of the mean[1]



[1] Except for Figure 1 (Original Data) which has the distribution of the original sample

References

- B. Efron and R. Tibshirani (1986) , Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures ofStatistical Accuracy, *Statistical Science*, Vol. 1, No. 1 (Feb., 1986), pp. 54-75