Module 4: Bootstrapping

Nikita Lakhotia, Vishwa Bhuta, Stuti Madaan

August 1, 2016

Standard Error and Sampling Distribution

Standard Error

$SE(\bar{x}) = \frac{\hat{\sigma}}{\sqrt{n}}$

where,

SE: *Estimate of population standard deviation*

$\hat{\sigma}$: *Standard deviation of the sample*

n : *sample size*

We calculate standard error as a measure of our confidence in our answer. That confidence is tied to our certainty that we would derive a similar answer given a slightly different sample. This is how we quantify how stable a statistic(like mean, median) is under repeated sampling from the population.

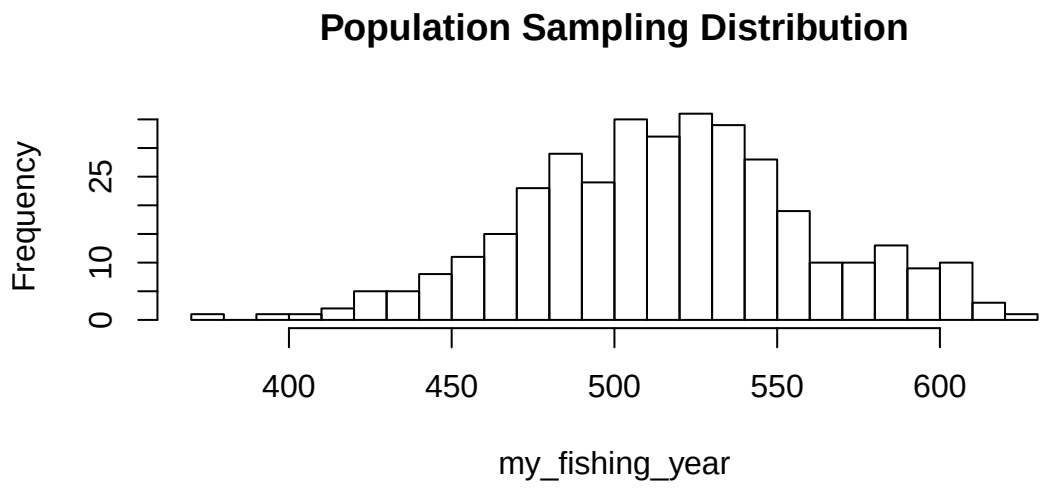
Sampling Distribution

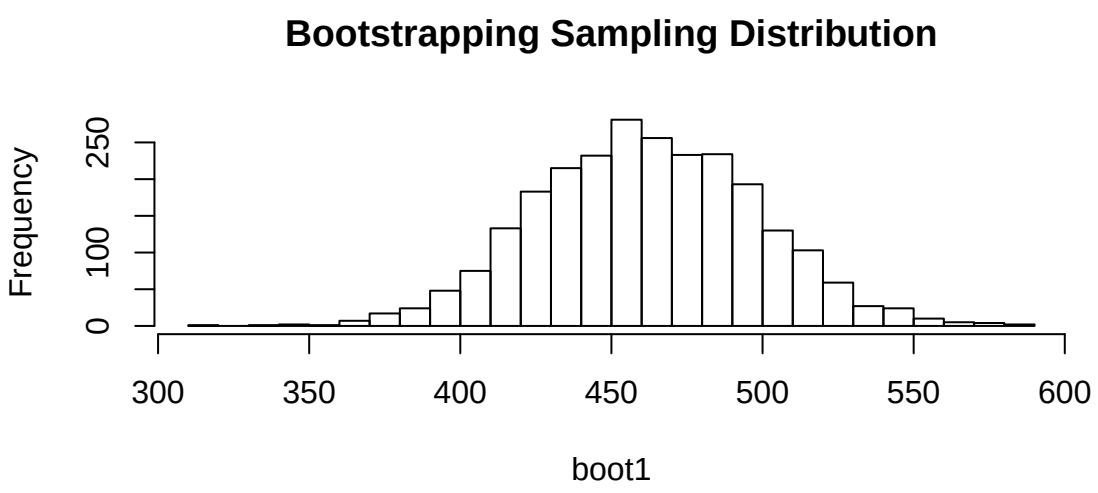
In an ideal world:

We could take an infinite number of samples from our population and derive the test statistic (eg. mean height of UT students) from each of those samples. The histogram of all the different $\hat{\theta}$'s that we get represents the "*sampling distribution*". Refer to Figure 1 in Appendix.

The Problem: it's unrealistic to generate that many samples. Our best estimate of the population, therefore, is our sample.

Bootstrapping






```
## Standard Deviation for Population Sampling Distribution:  44.06296
```

```
## Standard Deviation for Bootstrapping:  36.47343
```

How does Bootstrapping help?

- It allows us to create a sampling distribution using our one sample.
- It captures variability in the data.
- Has very few assumptions, and therefore can be used on situations of varying complexity
- As the number of *original* samples increases, the statistic is more likely to resemble the population parameter

Assumption: the original sample is representative of the population

How does Bootstrap work?

1. Original sample size = n
2. We take x number of resamples of size n from the original sample

a. Resample with replacement: think about this as a bag filled with n marbles that are marked. For each resample, you pull out one marble, jot down the marking, and put the marble back in the bag. You keep doing this until you have n markings on your list. That's your first resample. Because you put the marble back in the bag, there is a chance that you could pick the same marble multiple times in your resample.

   i. If you do this x number of times, you will have x number of resamples that are slightly different from each other. X should be very large.
   ii. Each resample will have some duplicates, but the combination will be different across the resamples.
   iii. Across all the resamples, we are replicating the variability we might see in the population.

3. For each resample, we calculate the test statistic.
4. The histogram of these test statistics gives us the "bootstrapped sampling distribution"
5. The standard deviation of the bootstrap becomes our standard error, i.e. our confidence in our answers resembling the population.

Refer to Figure 2 in Appendix.

What do I report?

The test statistic (eg. mean) should be the test statistic of just the original sample. The standard error is the standard error we get from the bootstrapping.

When is bootstrapping useful?

The short answer: always. Bootstrapping allows us to simulate the population sampling distribution.

The longer answer: bootstrapping is especially useful when the SE of the sample is not known or easy to calculate outright. For instance, in our fishing scenario, the SE of the sample would have simply been the formula we mention above. But if our example included complex tools like decision trees, or required multiple transformation, the SE becomes difficult to assess. Bootstrapping allows us to go around this.

Bootstrap Example

Standard Correlation vs Spearman's correlation

Standard/Pearson's Correlation:

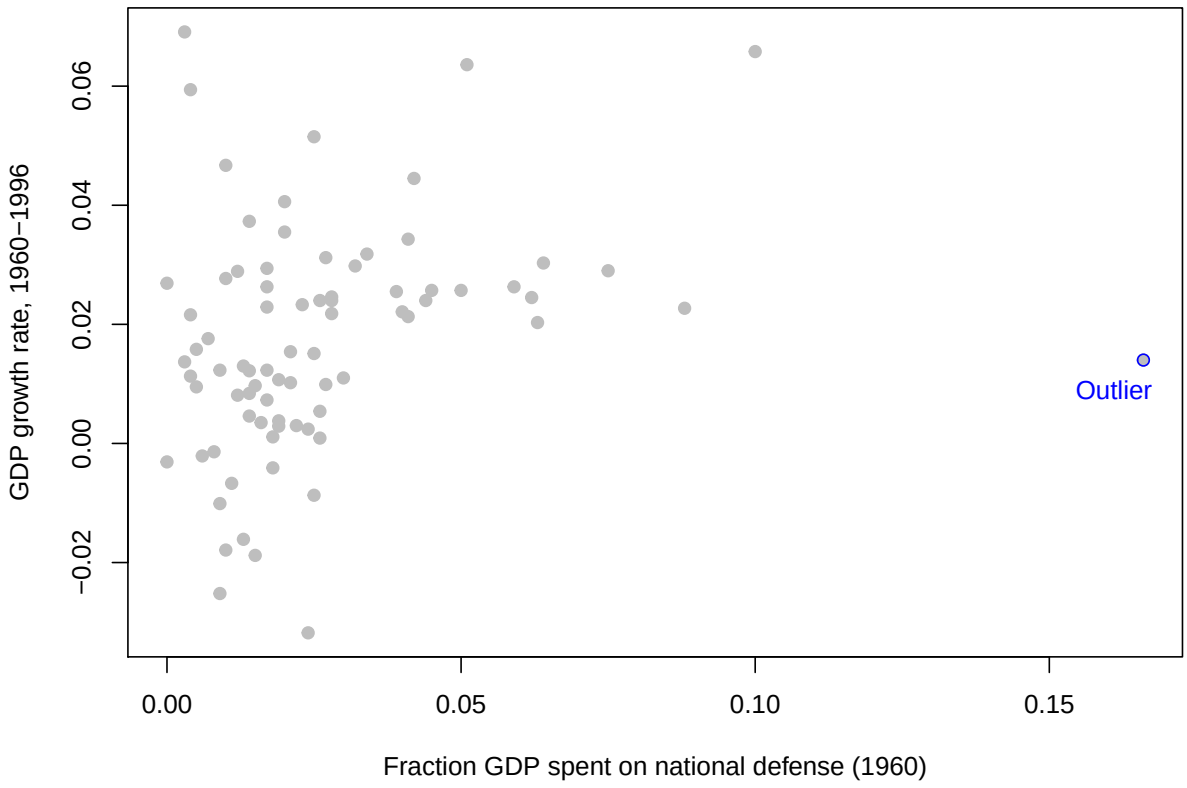
- linear relationship between the values of two variables
- gets impacted by the outliers in the dataset

Spearman's Correlation:

- calculates relationship between the 'ranks' of the two variables
- does not get impacted by the outliers in the dataset

GDP Growth Example

Scatter Plot for GDP Data



Pearson's Correlation including Outliers:

```
## [1] 0.2683152
```

Pearson's Correlation excluding Outliers:

```
## [1] 0.3608357
```

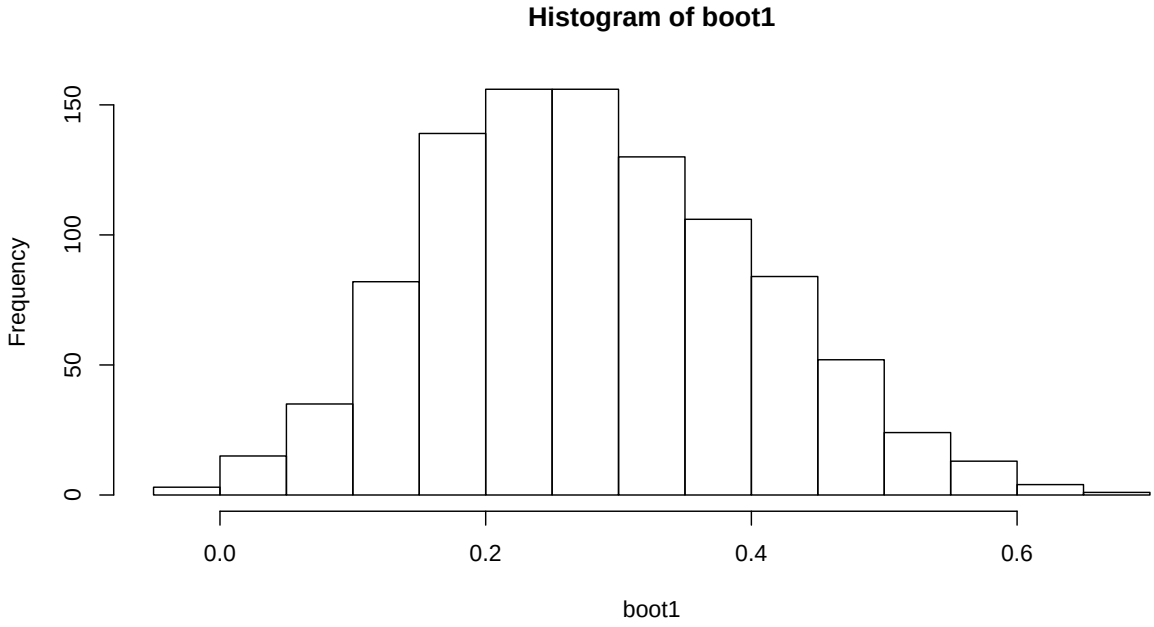
Spearman's Correlation including Outliers:

```
## [1] 0.3381575
```

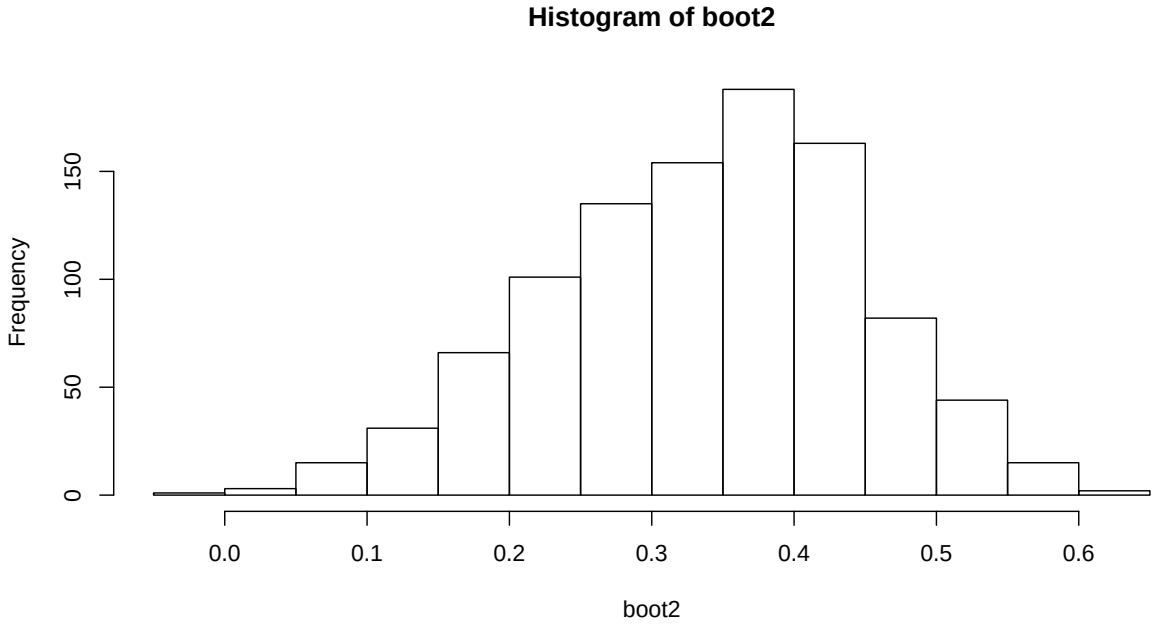
Spearman's Correlation excluding Outliers:

```
## [1] 0.3451648
```

Bootstrap ordinary correlation



Bootstrap Spearman's correlation



Explanation:

It can be observed in above histograms that the distribution of Spearman's Correlation is narrower as compared to the distribution of Pearson's correlation. This is because Spearman's Correlation is not impacted by the Outlier present in the data. However, Pearson's Correlation distribution gets inflated due to the outliers. The outlier is present in some samples and absent in others when bootstrapped. This creates a wide range of variations in the statistic values obtained from each sample,thus, causing this inflation.

Once we are comparing the ranks of these variables, SE is not as intuitive to calculate. However, with bootstrapping, do not need to calculate the SE of the sample. Instead, we just run the correlation many times with the resamples, we can easily calculate the standard deviation of the resulting correlations, thereby allowing us to quantify our uncertainty.

Conclusion:

Bootstrapping provides a unified way to quantify uncertainty across many different statistical scenarios.

Appendix

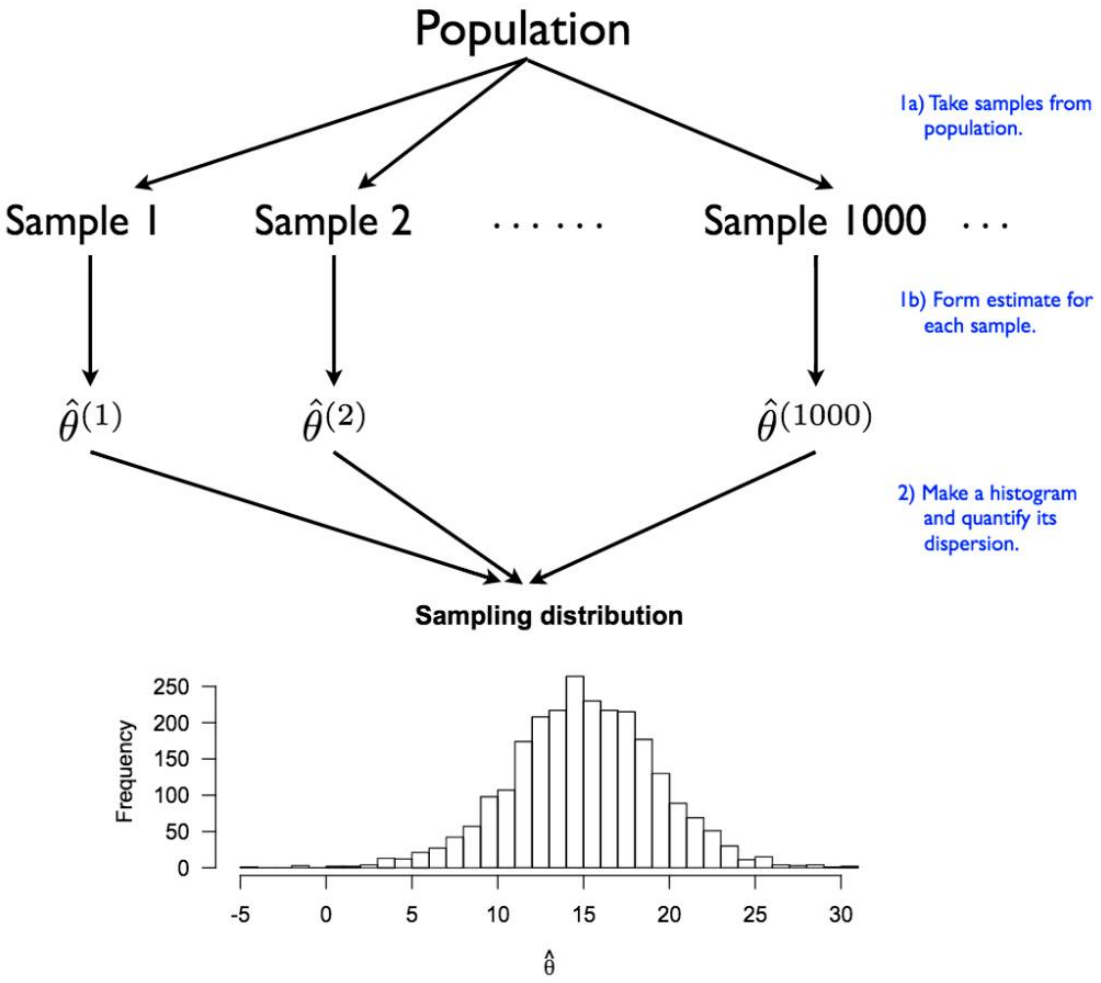


Figure 1: Population Sampling Distribution

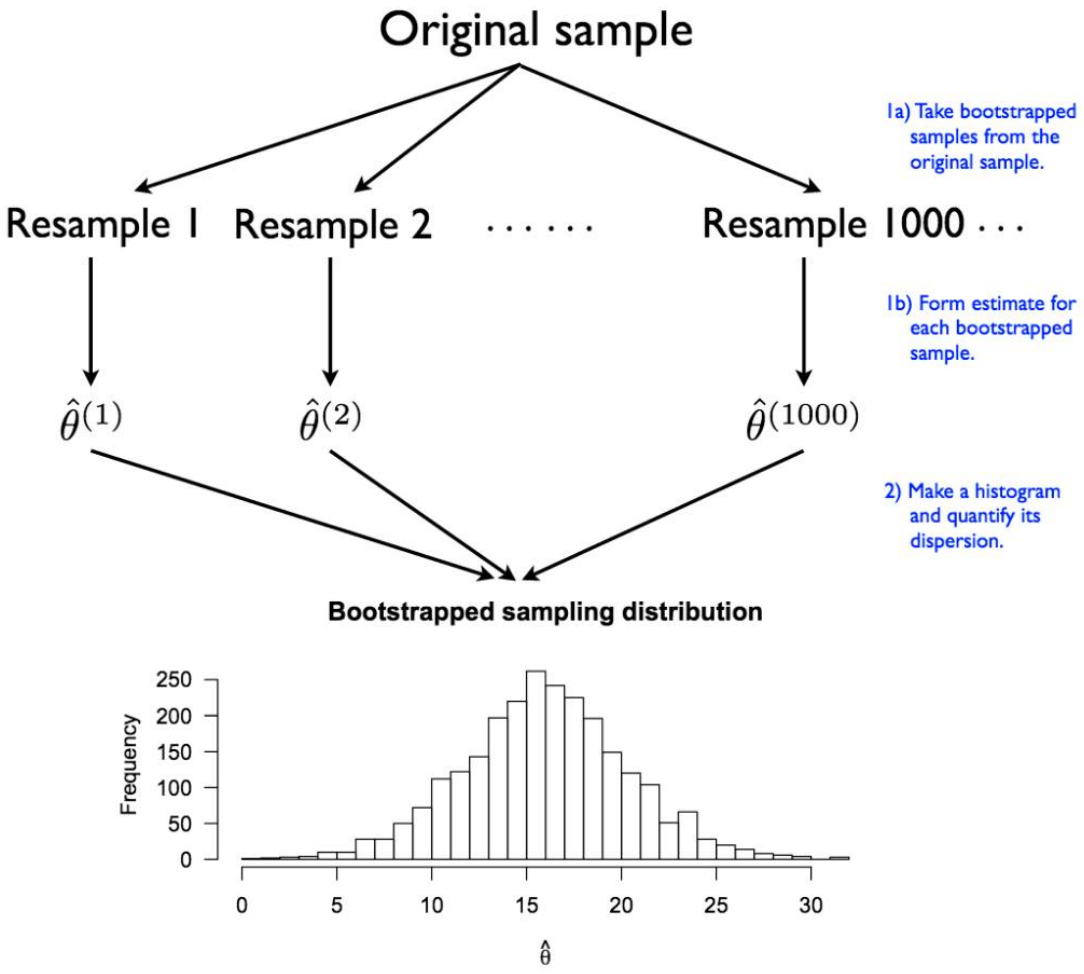


Figure 2: Bootstrapped Sampling Distribution