CS2065 Spring 2022
**Discussion: Bootstrap and Confidence Intervals (Project 02)**

Suppose we are trying to estimate a population parameter. Whenever we take a random sample and calculate a statistic to estimate the parameter, we know that the statistic could have come out differently if the sample had come out differently by random chance. We want to understand the *variability* of the statistic in order to better estimate the parameter. However, we don't have the resources to collect multiple random samples. In order to solve this problem, we use a technique called *bootstrapping*.
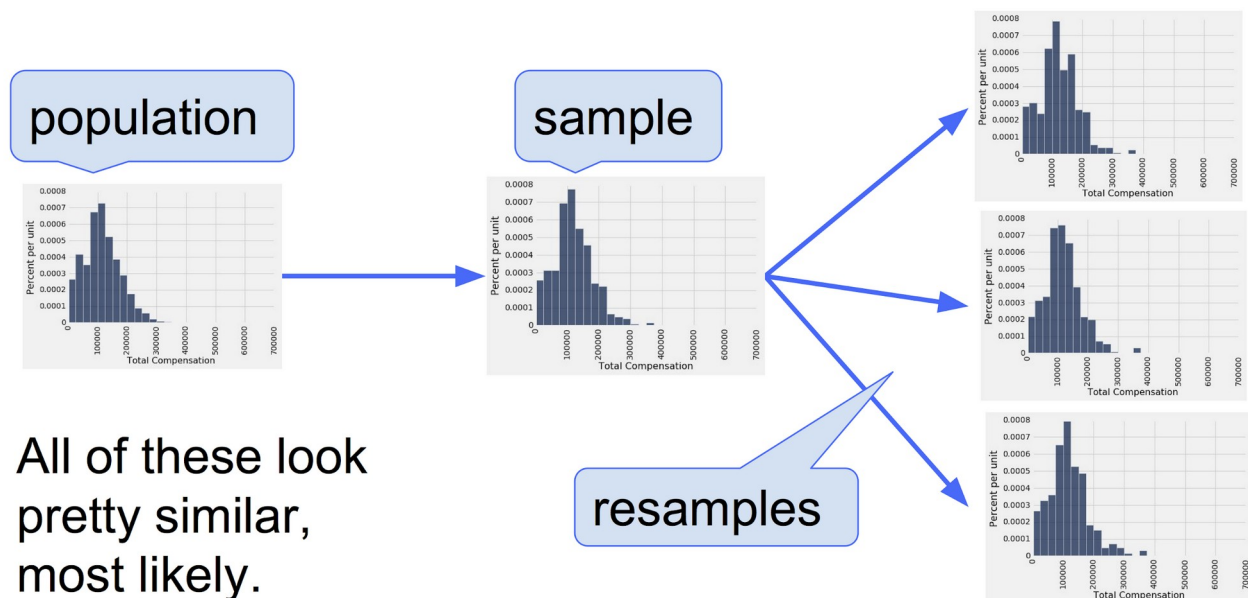
Question 1: What is the difference between a parameter and a statistic? Which of the two is random?

Question 2: Assume we have one large, random sample. How could we generate another sample that resembles the population if we don't have the resources to sample again from the population?

Tennis

Adith is interested in the height of tennis players. He's collected a sample of 100 heights of professional women's tennis players in the table `tennis`. The average of the sample is 69 inches and the standard deviation is 1.5 inches! He wants to use this sample to make some estimates about the population of the heights of professional women's tennis players.

Question 3: Adith wants to estimate the 50th percentile of heights in the women's professional tennis population. The 50th percentile has another name- the median! Define a function `ci_median` that constructs a 85% confidence interval for the median as follows. The function takes the following arguments:

Hint: to find the median of an array you can use the percentile function or the np.median function.

- tbl: A one-column table consisting of a random sample from the population; you can assume this sample is large
- reps: A number of bootstrap repetitions

```
def ci_median(tbl, reps):
    stats = _____
    for _____ :
        new_samp = _____
        new_median = _____
        stats = _____
    left_end = _____
    right_end = _____

    return make_array(left_end, right_end)
```

Question 4: If Adith calls `ci_median(tennis, 2000)` 500 times, approximately how many CI's do we expect to contain the actual median height of women's professional tennis athletes?

Adith and Jessica are arguing about the average height of women's tennis players. Jessica wishes she played tennis professionally, so she argues that the average height of all players is 68 inches. Adith doesn't know if Jessica's guess is too high or low but thinks that Jessica's guess is wrong. Help Adith conduct a hypothesis test about this!

Question 5: Write down the null and alternative hypotheses and test statistic for this test!

Null:

Alternative:

Test Statistic:

Question 6: Adith decides to do the simulation and the 95% confidence interval he computes is [68.6, 69.2]. What should his conclusion for the hypothesis test be?