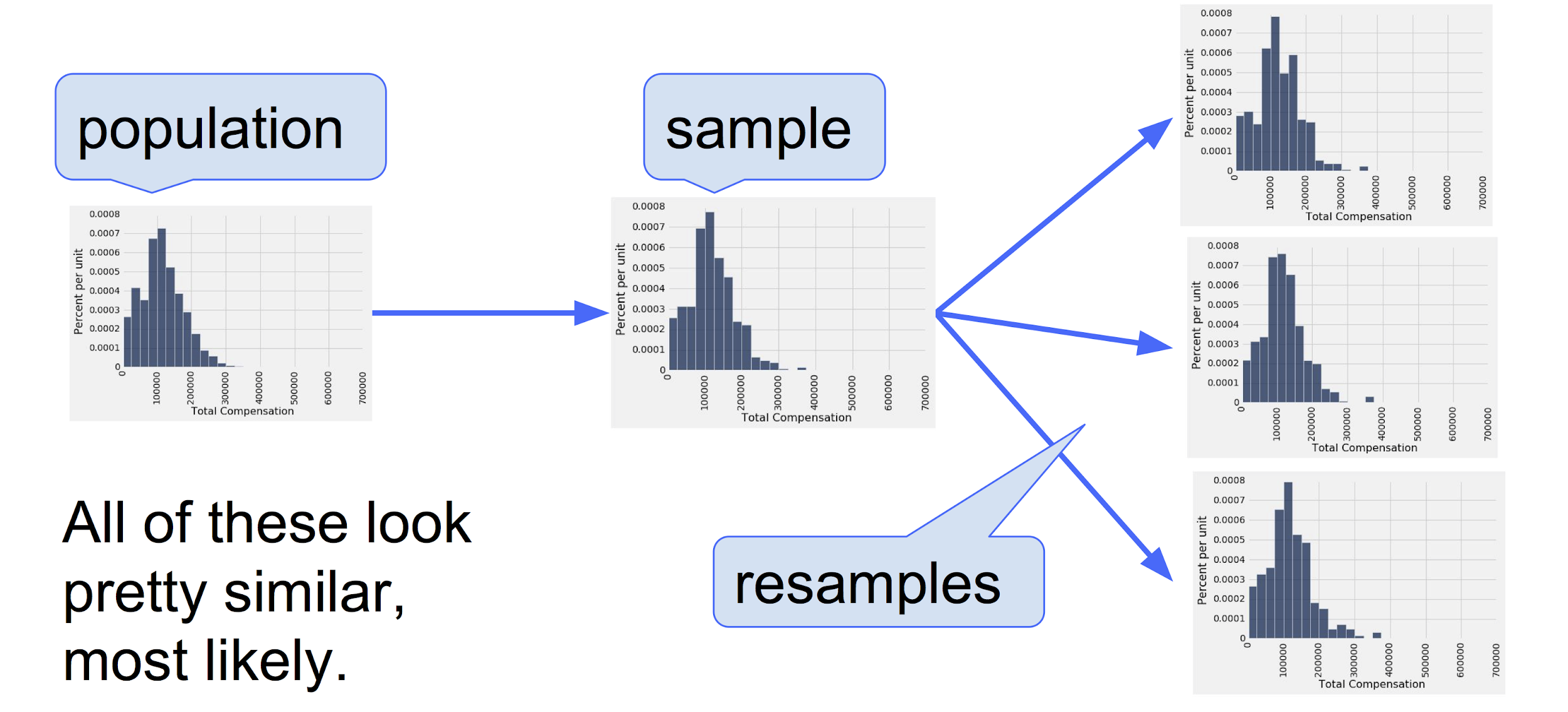
**Data 8 Spring 2020**

**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?

A parameter is a property of the population, so it is fixed and doesn’t change. However, we calculate statistics from samples, which are often random. Typically, we want to use statistics in order to estimate population parameters. Therefore, a statistic is random and a parameter is not 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?

If we didn’t have the resources to sample again from the population, we can do the “next best thing” in order to generate what would seem like a sample from the population, which is to take the *bootstrap*. In order to create a bootstrapped sample, we would sample with replacement using the same sample size as the original sample from the sample. This creates a new sample which is representative of 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 = make\_array()

for i in np.arange(reps) :

new\_samp = tbl.sample()

new\_median = percentile(50, new\_samp.column(0))

stats = np.append(stats, new\_median)

left\_end = np.percentile(7.5, stats)

right\_end = np.percentile(92.5, stats)

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?

There’s a 85% chance that using this method we produce a confidence interval that contains the actual median tennis athlete height. So we have 500\*(0.85) = 425 confidence intervals.

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: The average height of professional women’s tennis players is 68 inches.

Alternative: The average height of professional women’s tennis players is not 68 inches.

Test Statistic: The average heights of player in the sample

**Question 6:** Adith doesn’t have access to Python and unfortunately can’t use a simulation for this hypothesis test. Can he still figure out what the distribution of average heights is? Why?

*Hint: What do we know about the population?*

Yes! Since we are computing the distribution of means, we can apply the CLT here and say that the distribution of means is normal. Since we are given the mean and SD, we can say that the distribution is normal with a mean of 69 inches. We also use the formula SD of sample means = populationSD/sqrt(sample size), so we have 1.5/sqrt(100) = 0.15in

Normal distribution, mean = 69in, SD = 0.15in

**Question 7:** 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?

Since 68 lies outside of the bounds he should reject the null!