Data 8 Spring 2024

The Bootstrap

Lab 07

March 2024

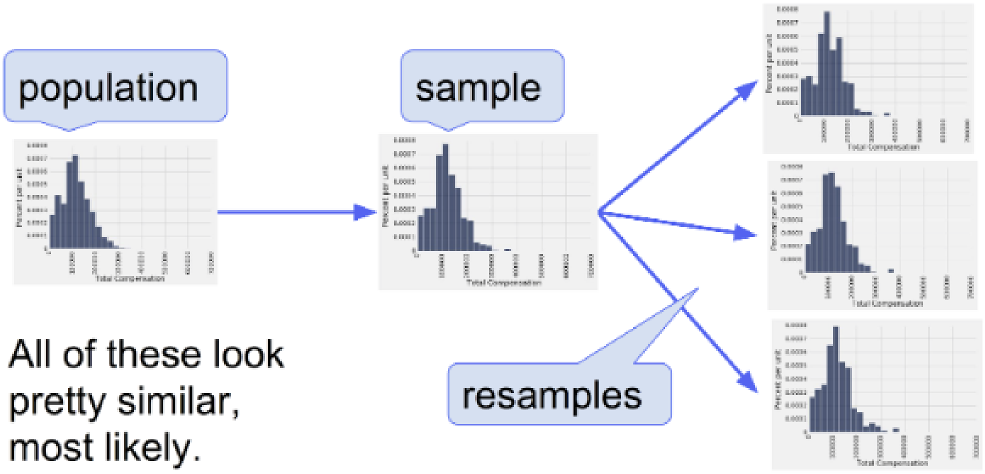
1. Mid-semester Check In

a. What has been your favorite topic/assignment/lecture/anything so far with the first half of the class done?

If you have any concerns about your performance in the class so far, feel free to bring it up to your lab TA.

2. Facts About the Bootstrap

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*.



a. When we conduct a bootstrap resample, what size resample should we draw from our sample? Why?

b. Why do we need to resample from our sample with replacement?

c. When we conduct a bootstrap resample, what is the underlying assumption/reasoning for resampling from our sample? Why does it work?

3. Thirsty

Warmup: What is the difference between a parameter and a statistic? Which of the two is random?

You are interested in investigating the liters of water consumed every day by UC Berkeley students. In particular, you want to study the proportion of students drinking less than 3 liters of water per day. You contact 150 random students from the directory and obtain the amounts of water each one of them drinks, storing them in the table `water`. The table has 1 column, `amount`, which stores the number of liters of water drunk by each student.

a. What is the parameter and what is the statistic in this scenario?

b. Write a line of code to calculate the proportion of students in your sample who drank less than 3 liters of water per day.

c. Write a line of code to perform a single bootstrap resample of the data stored in the `water` table.

d. Fill in the following blanks to conduct 10000 bootstrap resamples of your data, calculating the proportion of students in each resample that drink less than 3 liters of water per day, then plotting the distribution of those proportions using an appropriate visualization.

```
proportions = ________________________

for i in np.arange(10000):

      resampled_table = ________________________________________________

      resampled_statistic = ________________________________________________

      proportions = ________________________________________________

proportions_table = Table().with_column('Resampled proportion', proportions)

proportions_table.________________________________________________
```

4. Tennis Time

Ciara is interested in the heights of female tennis players. She's collected a sample of 100 heights of professional women's tennis players. She wants to use this sample to estimate the true interquartile range (IQR) of all heights of professional women's tennis players.

Hint: We defined the interquartile range (IQR) to be: 75th percentile - 25th percentile

a. In order to construct a 99% confidence interval for the IQR, what should our upper and lower percentile endpoints be?

b. Define a function `ci_iqr` that constructs a 99% confidence interval for the IQR as follows. The function takes the following arguments:

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

Hint: To find the 25th and 75th percentile of an array, you can use the percentile function

```
def ci_iqr(tbl, reps):

     stats = ________________________

     for ____________ :

          resample_col = ____________________________________

          new_iqr = ____________________________________

          stats = ____________________________________

     left_end = ________________________

     right_end = ________________________

     return make_array(left_end, right_end)
```

c. Say Ciara recruited 500 of her friends to perform the same bootstrapping process she did. In other words, each of her friends drew a large, random sample of 100 heights from the population of professional women's tennis players and constructed their own 99% confidence intervals. Approximately how many of these CI's do we expect to contain the actual IQR for the heights of professional women's tennis athletes?

Note how in this example, we obtain different random samples from the population for each confidence interval, rather than each person re-using the same original sample. Why is this distinction important?