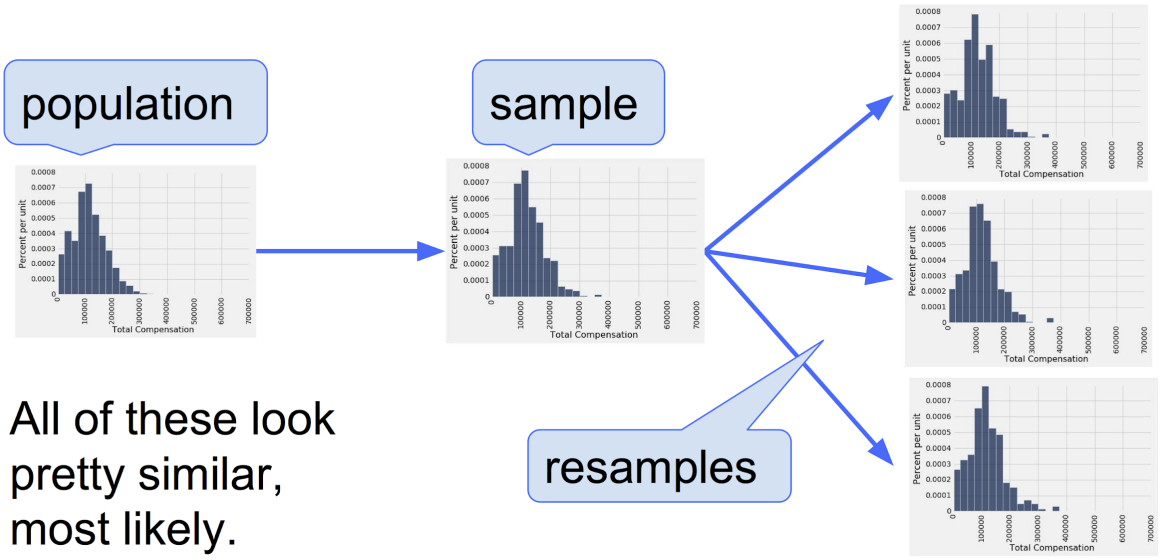
Lab 08: Bootstrap, Confidence Intervals

Data 8 Discussion Worksheet

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*.



1. **Warm-Up:** 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.

2. Sampling Techniques: 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.

3. 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?

Our lower endpoint should be 0.5 and upper endpoint should be 99.5

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)
```

```
def ci_iqr(tbl, reps):
     stats = make_array()
     for i in np.arange(reps) :
          resample_col = tbl.sample().column(0)
          new_iqr = percentile(75, resample_col) -
percentile(25, resample_col)
          stats = np.append(stats, new_iqr)
     left_end = percentile(0.5, stats)
     right_end = percentile(99.5, stats)

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?

We interpret a 99% confidence interval to mean that we are 99% confident **in the process** used to construct that given interval. In other words, 99% of the time we use this process we expect to construct an interval that contains the true population parameter. Since we have 500 CIs, each at a 99% confidence level, we find that since 500*(0.99) = 495, we expect to have 495 of these CIs containing the actual IQR of heights.