




Tutoring Section 10

Bootstrapping and Confidence Intervals

Slides by Kevin Miao

Logistics

- How did the midterm go?
- Last half of the semester, how are we feeling?
- **Next Week:** U.S. Presidential Elections 
 - *Tutoring section will be rescheduled. Following options:*
 - Monday 11am-1pm (PST)
 - Wednesday 11am-1pm (PST)
 - Online Walkthrough

Per usual:

- **Feedback Form:** <https://tinyurl.com/feedbackD8Kevin>

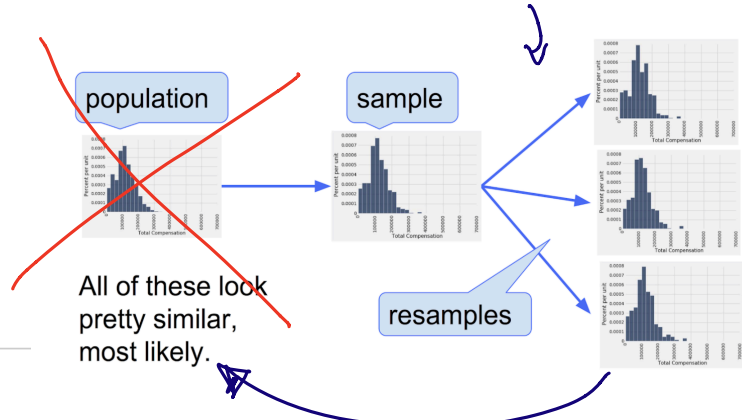
All resources can be found on kevin-miao.com

Today

- Bootstrapping
 - Conceptual Overview
 - Methodology
 - *Worksheet*
 - Confidence Intervals
 - Defining Confidence Intervals
 - Generating CIs
 - *Worksheet*
-

Bootstrapping

- **Setup:** A sample can always turn out different. If we were to calculate the average number of Trump vs Biden voters of two different surveys, the averages might turn out differently.
- **Goal:** Instead of generating one value, we want to generate a range of **plausible** values that could be the parameter of interest.
- **Related concepts:**
 - CI/Hypothesis Testing
 - Law of Averages



Bootstrapping

representative
procedure

- Obtain one large simple random sample
- For i in range `np.arange(number of resamples)`:
 - Resample **original sample** with replacement
 - What happens if we resample without replacement?
 - Calculate statistic and append it to your array of interest.
- **Result:** Array of items
- What can we do with this **result**?
 - Confidence Intervals
 - Collecting averages of resamples has some cool properties
 - Central Limit Theorem (Lecture later this week)

shuffles everything
↑

Worksheet

Link: <https://tinyurl.com/d8tutweek10>

Q1

1.1 Suppose we have a sample of the heights of 100 UC Berkeley students contained in a table `heights`, drawn as a simple random sample from the population of UC Berkeley students. We would like to estimate the average height of the UC Berkeley population. To do so, we will use the average height in our sample as our estimate.

a. Identify the population parameter and sample statistic for this particular experiment.

Which of these two quantities is random?

parameter: avg height of all UC Berkeley student
statistic: " of all 100 students in sample

b. We would like to use the bootstrap to generate an empirical distribution for our statistic. Fill in the code such that `parameter_estimates` contains 10,000 bootstrapped sample statistics.

Hint: You may find the table method `sample` helpful in this problem.

```
parameter_estimates = make_array()
```

```
for i in np.arange(10000):
```

```
parameter_estimate = np.average(height.sample(100, with_replacement = True).column(0))
```

```
parameter_estimates = np.append(parameter_estimate, parameter_estimate)
```

height.sample(c)
by default
• w/ replacement = True
• sample size = len(table)

Here are the first 5 rows of the table `heights`:

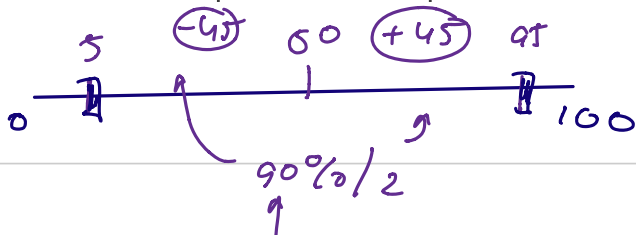
Height
67
72
75
65
62
(95 more)

c. Why do we have to sample with replacement? What would the histogram of `parameter_estimates` look like if we sampled without replacement?

Without replacement gives the same table!

Confidence Intervals

- Cool, so what else can we do with bootstrapping?
- **Bootstrapping** tells us how different a certain statistic can turn out through giving us an interval.
- **Confidence Intervals** allow us to give a measure of confidence.
- For instance, “With 95% confidence, we can say that Berkeley students spend between \$10 and \$15 on lunch”



Confidence Intervals

more

p-value: 10% 5% 3%
↑ ↑ ↑

- **Generating Confidence Interval**

- Determine what confidence you want: 90%, 95%, 97%?

- **Always remember:**

- For a $x\%$ confidence interval, you take the **middle x percent**

- For instance, 95% confidence interval starts at 2.5th percentile to the 97.5th percentile.

→ ⊕ More confident about your interval

- **What are the pros/cons of a wider confidence interval?**

- **Sins:**

- Never say “There is a 95% chance that this is the value”

not related to probability → ⊖ It might not be informative

- Sample Size \neq Number of Resamples

Q2.1

2.1 Tonight is the Monster Mash. We're trying to determine the median spookiness level of ghosts. We are given a sample of ghosts in the form of a one column table called `spooky_sample` that contains 200 numbers. Each number describes how spooky a ghost is on a scale from 0 to 10. You can assume that the sample is a simple random sample from the population of all ghosts. Fill in the code below to create a function that returns a 95% confidence interval as a 2-item array for the median spookiness level in the population of ghosts.

```
def candy_confidence_interval(spooky_sample, replications):  
    result_medians = make_array()  
    for i in np.arange(replications):  
        resample_median = spooky_sample.sample()  
        median = np.median(resample_median.column(0))  $\leftrightarrow$  np.percentile(50,  
        result_medians = np.append(result_medians, median) resample_median.  
        left_end = percentile(2.5, result_medians) column(0))  
        right_end = percentile(97.5, result_medians)  
    return make_array(left_end, right_end)
```

a. If we run the function you wrote above multiple times, will it always return the same interval? Why or why not?

Small chance, but most it turns out differently
due to randomness.

Q2.2

population parameter
g

2.2 There exists a median spookiness level for the population of all ghosts. We call this the "true median spookiness level" of the population.

a. What percent of the population will be contained in the confidence interval?

NO, the CI estimates a population parameter.
you can't say anything about the population.

b. What is the probability that the interval we calculate will contain the "true median spookiness level" of the population?

There's no probability! a value can be or not be on interval. There's no chance involved.

c. If we were to compute 100 confidence intervals with the function from 3.1, how many of those confidence intervals would we expect to capture the true median spookiness level?

• 95%
CI
~95 times, we should see the true value being captured by our confidence interval.

Old exam question (FA19 Final)

6. (5 points) Final Exam Studying

You ask a random sample of 250 Data 8 students from last semester how long they spent studying for the final exam in minutes. The median of the data in your sample is 9.2 hours. To quantify the uncertainty in your estimate, you create a 90% confidence interval by bootstrapping the 250 sampled students. The interval you obtain is [8.6 hours, 9.9 hours].

(a) (3 pt) In the blank below, describe one way to decrease the width of your interval. Your answer must fit in the blank provided.

(b) (2 pt) Suppose every single one of the 1300 students in the course this semester repeats the bootstrapping process above, and each one obtains a confidence interval. How many of the confidence intervals would you expect to **not** contain the population's median time spent studying for the final? **Choose the closest answer from the choices below.**

- 650
 - 130
 - 65
 - 0
-

Old exam question (FA19 Final)

6. (5 points) Final Exam Studying

You ask a random sample of 250 Data 8 students from last semester how long they spent studying for the final exam in minutes. The median of the data in your sample is 9.2 hours. To quantify the uncertainty in your estimate, you create a 90% confidence interval by bootstrapping the 250 sampled students. The interval you obtain is [8.6 hours, 9.9 hours].

- (a) (3 pt) In the blank below, describe one way to decrease the width of your interval. Your answer must fit in the blank provided.

Decrease confidence level, OR increase sample size.

- (b) (2 pt) Suppose every single one of the 1300 students in the course this semester repeats the bootstrapping process above, and each one obtains a confidence interval. How many of the confidence intervals would you expect to **not** contain the population's median time spent studying for the final? **Choose the closest answer from the choices below.**

- 650
 - 130
 - 65
 - 0
-

End of Section

- Please complete the anonymous Feedback form so I can improve my teaching:
 - <https://tinyurl.com/feedbackD8Kevin>
 - Solutions and notes will be posted as soon as possible.
 - Email me if you have any questions: kevinmiao@berkeley.edu
-