Homework 8: Confidence Intervals

Reading:

- Chapter 13: Estimation

Please complete this notebook by filling in the cells provided.

For all problems that you must write our explanations and sentences for, you **must** provide your answer in the designated space. Moreover, throughout this homework and all future ones, please be sure to not re-assign variables throughout the notebook! For example, if you use `max_temperature` in your answer to one question, do not reassign it later on.

```
# Don't change this cell; just run it.

import numpy as np
from datascience import *

# These lines do some fancy plotting magic.
import matplotlib
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
import warnings
warnings.simplefilter('ignore', FutureWarning)
```

1. Thai Restaurants

Ben and Frank are trying see what the best Thai restaurant in Massachusetts is. They survey 1500 Massachusetts college students selected uniformly at random, and ask each student what Thai restaurant is the best (*Note: this data is fabricated for the purposes of this homework*). The choices of Thai restaurant are Lucky House, Imm Thai, Thai Temple, and Thai Basil. After compiling the results, Ben and Frank release the following percentages from their sample:

Thai Restaurant	Percentage
Lucky House	8%
Imm Thai	52%
Thai Temple	25%
Thai Basil	15%

These percentages represent a uniform random sample of the population of Massachusetts college students. We will attempt to estimate the corresponding *parameters*, or the percentage of the votes that each restaurant will receive from the entire population (the entire population is all Massachusetts college students). We will use confidence intervals to compute a range of values that reflects the uncertainty of our estimates.

The table `votes` contains the results of the survey.

```
from google.colab import drive
drive.mount('/content/drive')
```

```
# Just run this cell
votes = Table.read_table('/content/drive/MyDrive/votes.csv')
votes
```

Vote
Lucky House
Lucky House
Lucky House
Lucky House
Lucky House
Lucky House
Lucky House
Lucky House
Lucky House
Lucky House

... (1490 rows omitted)

Question 1. Complete the function `one_resampled_percentage` below. It should return Imm Thai's **percentage** of votes after simulating one bootstrap sample of `tbl`.

Note: `tbl` will always be in the same format as `votes`.

```
def one_resampled_percentage(tbl):

    resampled_tbl = tbl.sample(with_replacement=True)

    percentage_imm_thai = np.count_nonzero(resampled_tbl == 'Imm Thai') / len(resampled_tbl) * 100

    return percentage_imm_thai
    one_resampled_percentage(tbl)
```

Question 2. Complete the `percentages_in_resamples` function such that it returns an array of 2500 bootstrapped estimates of the percentage of voters who will vote for Imm Thai. You should use the `one_resampled_percentage` function you wrote above.

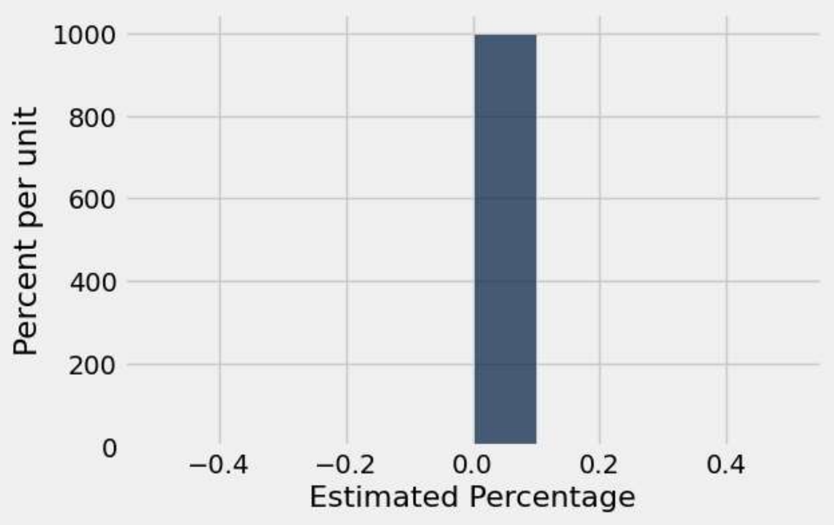
```
def percentages_in_resamples():
    percentage_imm = make_array()
    for i in range(2500):
        percentage = one_resampled_percentage(votes)

        percentage_imm = np.append(percentage_imm, percentage)

    return percentage_imm
```

In the following cell, we run the function you just defined, `percentages_in_resamples`, and create a histogram of the calculated statistic for the 2,500 bootstrap estimates of the percentage of voters who voted for Imm Thai. Based on what the original Thai restaurant percentages were, does the graph seem reasonable? Talk to a friend or ask your instructor if you are unsure!

```
resampled_percentages = percentages_in_resamples()
Table().with_column('Estimated Percentage', resampled_percentages).hist("Estimated Percentage")
```



Question 3. Using the array `resampled_percentages`, find the values at the two edges of the middle 95% of the bootstrapped percentage estimates. (Compute the lower and upper ends of the interval, named `imm_lower_bound` and `imm_upper_bound`, respectively.)

```
imm_lower_bound = np.percentile(resampled_percentages, 2.5)
imm_upper_bound = np.percentile(resampled_percentages, 97.5)
print("Bootstrapped 95% confidence interval for the percentage of Imm Thai voters in the population: [{:f}, {:f}]".format(imm_lower_bound, i
```

```
    Bootstrapped 95% confidence interval for the percentage of Imm Thai voters in the population: [0.000000, 0.000000]
```

Question 4. The survey results seem to indicate that Imm Thai is beating all the other Thai restaurants combined among voters. We would like to use confidence intervals to determine a range of likely values for Imm Thai's true lead over all the other restaurants combined. The calculation for Imm Thai's lead over Lucky House, Thai Temple, and Thai Basil combined is:

$$\text{Imm Thai's \% of the vote} - (\text{Lucky House's \% of the vote} + \text{Thai Temple's \% of the vote} + \text{Thai Basil's \% of the vote})$$

Define the function `one_resampled_difference` that returns **exactly one value** of Imm Thai's percentage lead over Lucky House, Thai Temple, and Thai Basil combined from one bootstrap sample of `tbl`.

```
def one_resampled_difference(tbl):
    resampled_tbl = tbl.sample(with_replacement=True)

    imm_percentage = np.count_nonzero(resampled_tbl == 'Imm Thai') / len(resampled_tbl) * 100
    lh_percentage = np.count_nonzero(resampled_tbl == 'Lucky House') / len(resampled_tbl) * 100
    tt_percentage = np.count_nonzero(resampled_tbl == 'Thai Temple') / len(resampled_tbl) * 100
    tb_percentage = np.count_nonzero(resampled_tbl == 'Thai Basil') / len(resampled_tbl) * 100

    imm_lead = imm_percentage - (lh_percentage + tt_percentage + tb_percentage)

    return imm_lead
```

Question 5. Write a function called `leads_in_resamples` that finds 2,500 bootstrapped estimates (the result of calling `one_resampled_difference`) of Imm Thai's lead over Lucky House, Thai Temple, and Thai Basil combined. Plot a histogram of the resulting samples.

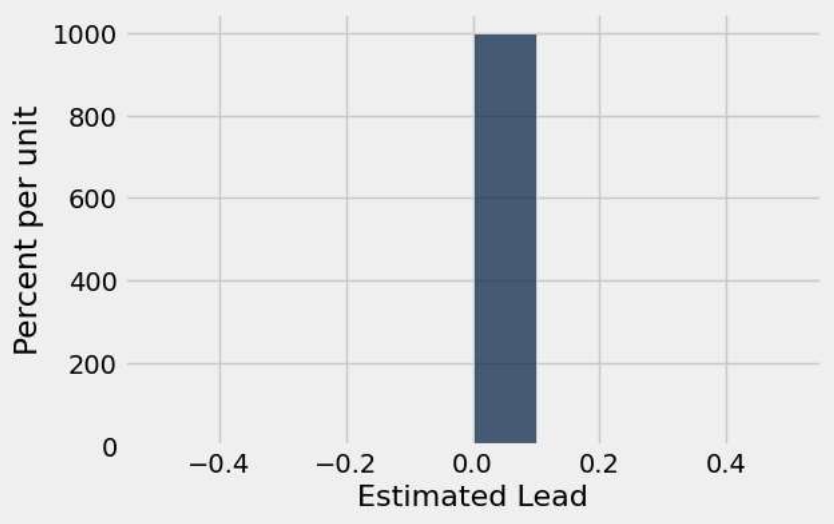
Note: Imm Thai's lead can be negative.

```
def leads_in_resamples():
    sampled_leads = make_array()
    for i in range(2500):
        lead = one_resampled_difference(votes)

        sampled_leads = np.append(sampled_leads, lead)

    return sampled_leads

sampled_leads = leads_in_resamples()
Table().with_column('Estimated Lead', sampled_leads).hist("Estimated Lead")
```



Question 6. Use the simulated data from Question 5 to compute an approximate 95% confidence interval for Imm Thai's true lead over Lucky House, Thai Temple, and Thai Basil combined.

```
diff_lower_bound = np.percentile(sampled_leads, 2.5)
diff_upper_bound = np.percentile(sampled_leads, 97.5)

print("Bootstrapped 95% confidence interval for Imm Thai's true lead over Lucky House, Thai Temple, and Thai Basil combined: [{:f}, {:f}]".f
```

```
Bootstrapped 95% confidence interval for Imm Thai's true lead over Lucky House, Thai Temple, and Thai Basil combined: [0.000000, 0.00000
```

2. Interpreting Confidence Intervals

Your instructor computed the following 95% confidence interval for the percentage of Imm Thai voters:

$$[49.40, 54.47]$$

(Your answer may have been a bit different; that doesn't mean it was wrong!)

Question 1

The instructor also created 70%, 90%, and 99% confidence intervals from the same sample, but forgot to label which confidence interval represented which percentages! Match each confidence level (70%, 90%, 99%) with its corresponding interval in the cell below (e.g. __ % CI: [49.87, 54.0] $\rightarrow$ replace the blank with one of the three confidence levels). **Then**, explain your thought process and how you came up with your answers.

The intervals are below:

- [49.87, 54.00] $\rightarrow$ ___
- [50.67, 53.27] $\rightarrow$ ___
- [48.80, 55.40] $\rightarrow$ ___

Explain your reasoning: *Write your answer here, replacing this text.*

Question 2

Suppose we produced 5,000 new samples (each one a uniform random sample of 1,500 voters/students) from the population and created a 95% confidence interval from each one. Roughly how many of those 5,000 intervals do you expect will actually contain the true percentage of the population?

Assign your answer to `true_percentage_intervals`.

```
true_percentage_intervals = 4750
```

Recall the second bootstrap confidence interval you created, which estimated Imm Thai's lead over Lucky House, Thai Temple, and Thai Basil combined. Among voters in the sample, Imm Thai's lead was 4%. The instructor's 95% confidence interval for the true lead (in the population of all voters) was

$$[-0.80, 8.80]$$

Suppose we are interested in testing a simple yes-or-no question:

> "Is the percentage of votes for Imm Thai tied with the percentage of votes for Lucky House, Thai Temple, and Thai Basil combined?"

Our null hypothesis is that the percentages are equal, or equivalently, that Imm Thai's lead is exactly 0. Our alternative hypothesis is that Imm Thai's lead is not equal to 0. In the questions below, don't compute any confidence interval yourself - use only the instructor's 95% confidence interval.

Question 3

Say we use a 5% P-value cutoff. Do we reject the null, fail to reject the null, or are we unable to tell using our instructor's confidence interval?

Assign `restaurants_tied` to the number corresponding to the correct answer.

1. Reject the null / Data is consistent with the alternative hypothesis
2. Fail to reject the null / Data is consistent with the null hypothesis
3. Unable to tell using our staff confidence interval

Hint: If you're confused, take a look at this chapter of the textbook.

```
restaurants_tied = 3
```

Question 4

What if, instead, we use a P-value cutoff of 1%? Do we reject the null, fail to reject the null, or are we unable to tell using our insturctor's confidence interval?

Assign `cutoff_one_percent` to the number corresponding to the correct answer.

1. Reject the null / Data is consistent with the alternative hypothesis
2. Fail to reject the null / Data is consistent with the null hypothesis
3. Unable to tell using our instructor's confidence interval

```
cutoff_one_percent = 2
```

Question 5

What if we use a P-value cutoff of 10%? Do we reject, fail to reject, or are we unable to tell using our confidence interval?

Assign `cutoff_ten_percent` to the number corresponding to the correct answer.

1. Reject the null / Data is consistent with the alternative hypothesis
2. Fail to reject the null / Data is consistent with the null hypothesis
3. Unable to tell using our instructor's confidence interval

```
cutoff_ten_percent = 2
```

3. Submission