Homework 8: Bootstraps and Confidence Intervals

Please complete this notebook by filling in the cells provided. When you're done:

1. Select `Run All` from the `Cell` menu to ensure that you have executed all cells.
2. Select `Download as HTML (.html)` from the `File` menu
3. Inspect your file to make sure it looks like it should.
4. Upload your file to Moodle.

In [1]:
```
# Don't change this cell; just run it.

import numpy as np
from datascience import *

%matplotlib inline
import matplotlib.pyplot as plots
plots.style.use('fivethirtyeight')
```

1. Plot the Vote

Four candidates are running for President of Dataland. A polling company surveys 1000 people selected uniformly at random from among voters in Dataland, and it asks each one who they are planning on voting for. After compiling the results, the polling company releases the following proportions from their sample:

|Candidate | Proportion| |:----------:|:----------:| |Candidate C | 0.47 | |Candidate T | 0.38 | |Candidate J | 0.08 | |Candidate S | 0.03 | |Undecided | 0.04 |

These proportions represent a uniform random sample of the population of Dataland. We will attempt to estimate the corresponding *population parameters* - the proportions of each kind of voter in the entire population. We will use confidence intervals to compute a range of values that reflects the uncertainty of our estimate.

The table `voters` contains the results of the survey. Candidates are represented by their initials. Undecided voters are denoted by `U`. Below we show you a sample from the survey results.

In [2]:
```
voters = Table().with_column('vote', np.array(['C']*470 + ['T']*380 + ['J']*
num_votes = voters.num_rows
voters.sample(10)
```

Out[2]:

vote
T
T
C
C
T
J
J
T
C
C

Question 1. Write a single line of code that will generate a bootstrap sample from our survey and call it "random_sample".

There is a table method that makes this very easy.

In [3]:
```
random_sample = voters.sample()
random_sample
```

Out[3]:

vote
U
C
T
T
S
C
T
T
J
C

... (990 rows omitted)

Question 2. For your random sample, please calculate the sample statistic that reports the proportion of reports who said they would vote for Candidate C.

In [4]:
```
proportion_for_C = random_sample.where("vote", are.equal_to("C")).num_rows/1
proportion_for_C
```

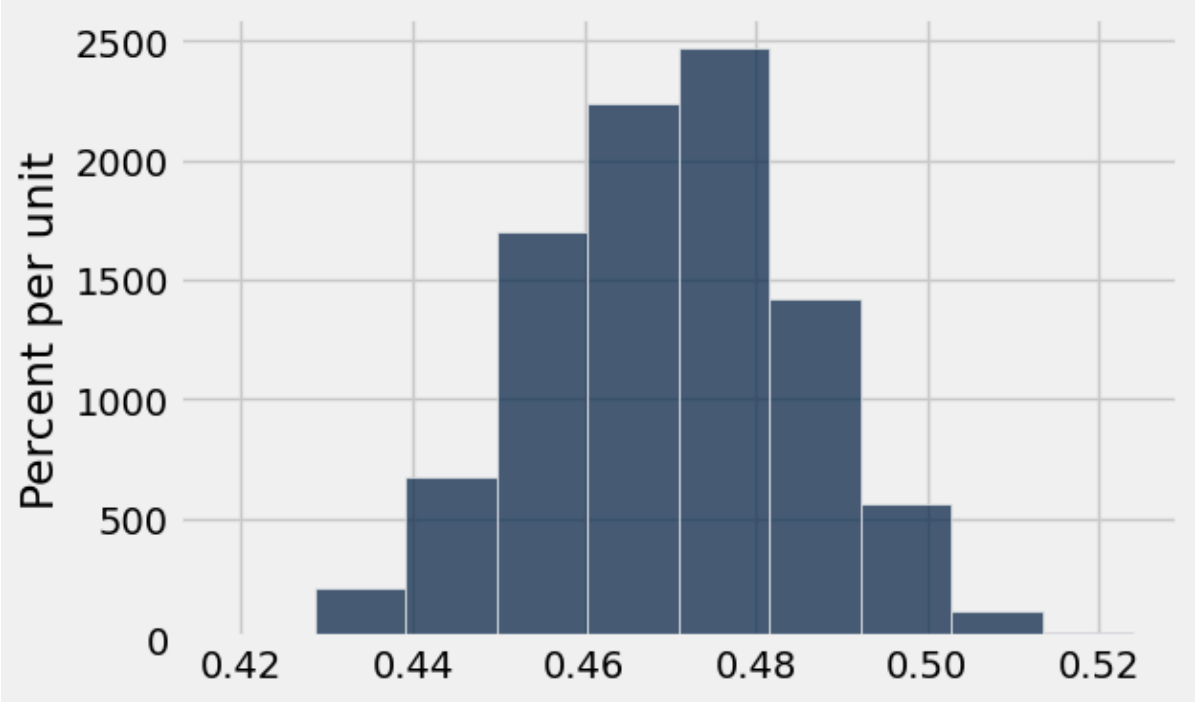
Out[4]:
```
0.463
```

Question 3. Please write code to generate 5000 bootstrapped sample statistics for the proportion of voters that select Candidate C and store this in an array called "sampled_proportions". Then create a histogram of your results.

In [5]:
```
sampled_proportions = make_array()

for i in np.arange(5000):
    bootstrap = voters.sample()
    proportion = bootstrap.where("vote", are.equal_to("C")).num_rows/1000
    sampled_proportions = np.append(sampled_proportions, proportion)

Table().with_columns("", sampled_proportions).hist()
```



Question 4. Using the array `sampled_proportions`, compute an approximate 95% confidence interval for the true proportions of voters planning on voting for candidate C. (Compute the lower and upper ends of the interval, named `lower_bound` and `upper_bound`, respectively.)

In [6]:
```
c_lower_bound = percentile(2.5, sampled_proportions)
c_upper_bound = percentile(97.5, sampled_proportions)
print("Bootstrapped 95% confidence interval for the proportion of C voters i
```

```
Bootstrapped 95% confidence interval for the proportion of C voters in the
population: [0.440000, 0.499000]
```

Question 5 Suppose the candidate asks you if it is possible that they will get over half the votes. How would you answer them?

It is possible, since the half of the votes are in the 95% confidence interval.

Question 6. The survey results seem to indicate that Candidate C is beating Candidate T among voters. We would like to use a confidence interval to determine a range of likely values for her true *lead*. Candidate C's lead over Candidate T is:

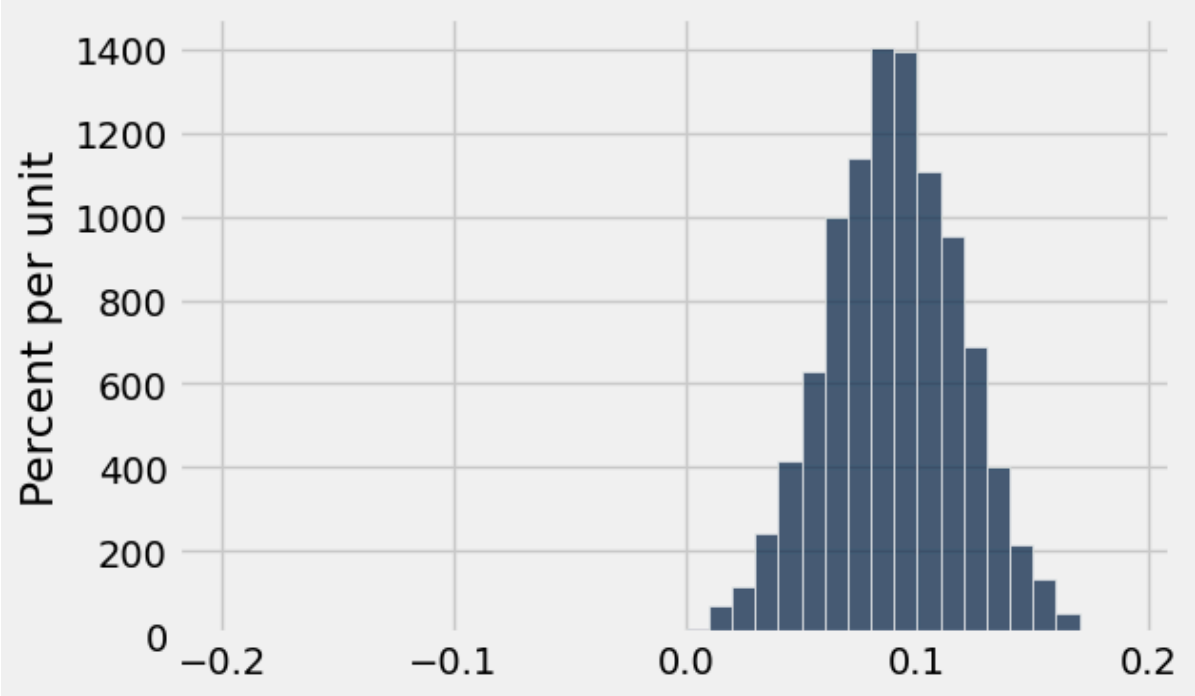
$$\text{Candidate C's proportion of the vote} - \text{Candidate T's proportion of the vote.}$$

Similar to your work above, use the bootstrap to compute an approximate distribution for Candidate C's lead over Candidate T. Plot a histogram of the resulting samples. Bins have been provided for you.

In [9]:
```
sampled_leads = make_array()

for i in np.arange(5000):
    bootstrap = voters.sample()
    c_vote = (bootstrap.where("vote", are.equal_to("C")).num_rows)/1000
    t_vote = (bootstrap.where("vote", are.equal_to("T")).num_rows)/1000
    lead = c_vote - t_vote
    sampled_leads = np.append(sampled_leads, lead)

Table().with_columns("", sampled_leads).hist(bins=np.arange(-0.2,0.2,0.01))
```



Question 7. Finally, create a 95% confidence interval for the true difference in proportions between Candidate C and Candidate T. Please interpret your results.

In [10]:
```
diff_lower_bound = percentile(2.5, sampled_leads)
diff_upper_bound = percentile(97.5, sampled_leads)
print("Bootstrapped 95% confidence interval for Candidate C's true lead over
```

```
Bootstrapped 95% confidence interval for Candidate C's true lead over Candi
date T: [0.033000, 0.149000]
```

We are 95% confident that Candidate C is leading over Candidate T.

2. Interpreting Confidence Intervals

The staff computed the following 95% confidence interval for the proportion of Candidate C voters:

$$[.439, .5]$$

(Your answer might have been different; that doesn't mean it was wrong.)

Question 1

Can we say that 95% of the population lies in the range $[.439, .5]$? Explain your answer.

No. The range is estimating that the average proportion of Candidate C votes is in the interval 0.439 and 0.5, but not 95% of the population lies in the range.

Question 2

Can we say that there is a 95% probability that the interval [.439, .5] contains the true proportion of the population who is voting for Candidate C? Explain your answer.

Yes. We can estimate by simulating samples through bootstrap mathod and we are 95% confident that the interval contains the true proportion of the population of Candidate C votes.

Question 3

Suppose we produced 10,000 new samples (each one a uniform random sample of 1,000 voters) and created a 95% confidence interval from each one. Roughly how many of those 10,000 intervals do you expect will actually contain the true proportion of the population?

9500 intervals are expected to contain the true proportion of the population.

Question 4

The staff also created 80%, 90%, and 99% confidence intervals from one sample, but we forgot to label which confidence interval represented which percentages! Match the interval to the percent of confidence the interval represents. (Write the percentage after each interval below.) **Then**, explain your thought process.

Answers:

$[.444, .495]$: 90%

$[.45, .49]$: 80%

$[.43, .511]$: 99%

For smaller confidence interval, the range have the bigger minumum and the smaller maximum.

Question 5

Recall the second bootstrap confidence interval you created, estimating Candidate C's lead over Candidate T. Among voters in the sample, her lead was .09. The staff's 95% confidence interval for her true lead (in the population of all voters) was

$$[.032, .15].$$

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

> "Are the candidates tied?"

Our null hypothesis is that the proportions are equal, or, equivalently, that Candidate C's lead is exactly 0. Our alternative hypothesis is that her lead is not equal to 0. In the questions below, don't compute any confidence interval yourself - use only the staff's 95% confidence interval.

Question: 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 confidence interval?

Reject the null hypothesis since the 95% confidence interval do not contain 0.

Question 6

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 confidence interval?

We are unable to tell even when it is 99% confident interval, since we do not know the value in the interval.

Question 7

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?

We reject the null since the interval is narrower than 5% interval that we already reject the null.