hw08

April 3, 2020

1 Homework 8: Confidence Intervals

Reading: Estimation

Please complete this notebook by filling in the cells provided. Before you begin, execute the following cell to load the provided tests. Each time you start your server, you will need to execute this cell again to load the tests.

Homework 8 is due **Thursday, 4/2 at 11:59pm**.

Start early so that you can come to office hours if you're stuck. Late work will not be accepted as per the course policies.

Directly sharing answers is not okay, but discussing problems with the course staff or with other students is encouraged. Refer to the policies page to learn more about how to learn cooperatively.

For all problems that you must write our explanations and sentences for, you must provide your answer in the designated space.

[2]:
```
# 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)
```

1.1 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	Proportion
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*, or the proportion of the votes that each candidate received from the entire population. We will use confidence intervals to compute a range of values that reflects the uncertainty of our estimate.

The table `votes` contains the results of the survey. Candidates are represented by their initials. Undecided voters are denoted by `U`.

[3]:
```
votes = Table().with_column('vote', np.array(['C']*470 + ['T']*380 + ['J']*80 +␣
 ↪['S']*30 + ['U']*40))
num_votes = votes.num_rows
votes.sample()
```

[3]:
```
vote
T
C
T
J
C
T
T
C
U
U
... (990 rows omitted)
```

Question 1.1. Below, complete the given code that will use bootstrapped samples from `votes` to compute estimates of the true proportion of voters who are planning on voting for **Candidate C**. Make sure that you understand what's going on here. It may be helpful to explain `proportions_in_resamples` to a friend or TA.

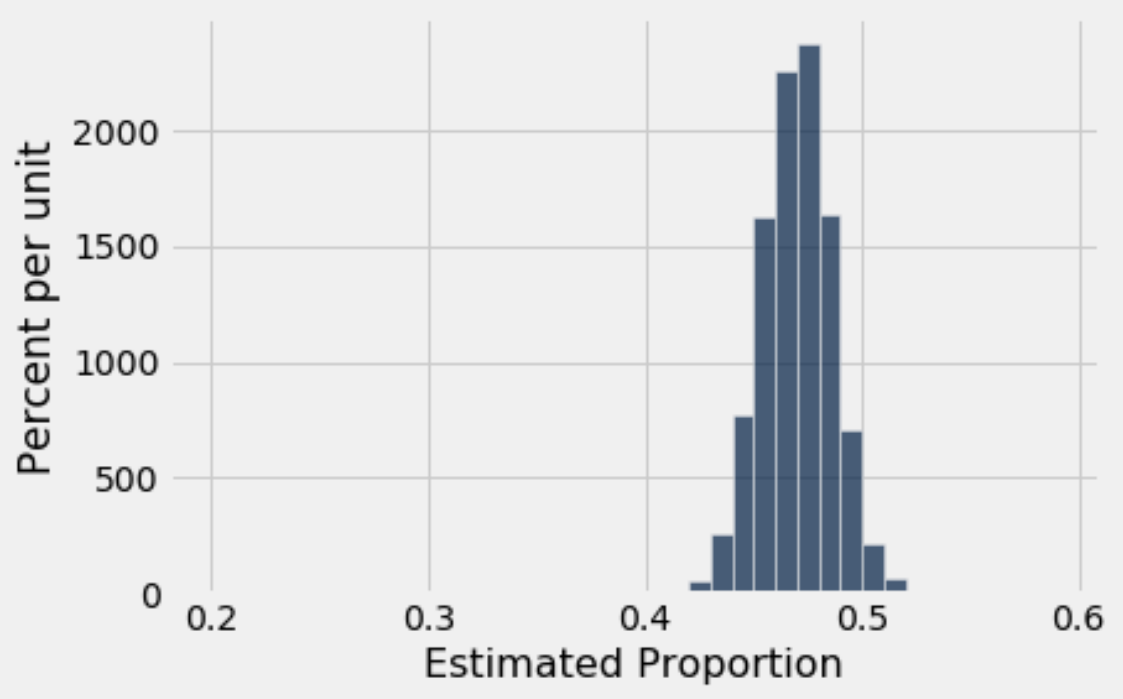
[ ]:
```
```

[4]:
```
def proportions_in_resamples():
    prop_c = make_array()
    for i in np.arange(5000):
        bootstrap = votes.sample()
        single_proportion = np.count_nonzero(bootstrap.where('vote', 'C').
 ↪column('vote')) / num_votes
        prop_c = np.append(prop_c, single_proportion)
    return prop_c
```

In the following cell, we run the function you just defined, `proportions_in_resamples`, and create a histogram of the calculated statistic for the 5,000 bootstraps. Based on what the original polling proportions were, does the graph seem reasonable? Talk to a friend or ask a TA if you are

unsure!

[5]:
```
sampled_proportions = proportions_in_resamples()
Table().with_column('Estimated Proportion', sampled_proportions).hist(bins=np.
 ↪arange(0.2,0.6,0.01))
```



Question 1.2. Using the array `sampled_proportions`, find the values that bound the middle 95% of the values in the data. (Compute the lower and upper ends of the interval, named `c_lower_bound` and `c_upper_bound`, respectively.)

[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 in␣
 ↪the population: [{:f}, {:f}]".format(c_lower_bound, c_upper_bound))
```

```
Bootstrapped 95% confidence interval for the proportion of C voters in the
population: [0.439000, 0.502000]
```

Question 1.3. The survey results seem to indicate that Candidate C is beating Candidate T among voters. We would like to use CI's to determine a range of likely values for her true *lead*. Candidate C's lead over Candidate T is:

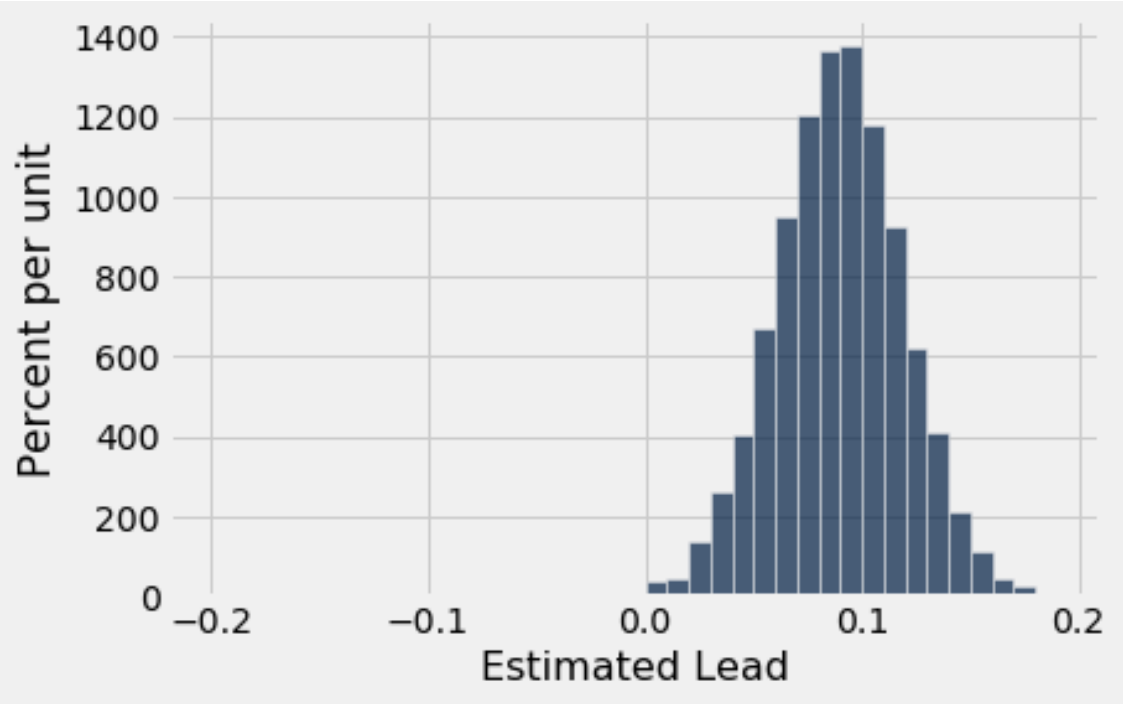
Candidate C's proportion of the vote − Candidate T's proportion of the vote.

Using the function `proportions_in_resamples` above as a *guideline*, use the bootstrap to compute an approximate distribution for Candidate C's lead over Candidate T. Plot a histogram of the the resulting leads.

[7]:
```
bins = np.arange(-0.2,0.2,0.01)

def leads_in_resamples():
    leads = make_array()
    for i in np.arange(5000):
        bootstrap = votes.sample()
        prop_c = np.count_nonzero(bootstrap.where('vote', 'C').column('vote')) /
 ↪ num_votes
        prop_t = np.count_nonzero(bootstrap.where('vote', 'T').column('vote')) /
 ↪ num_votes
        difference = prop_c - prop_t
        leads = np.append(leads, difference)
    return leads

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



[17]:
```
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␣
 ↪Candidate T: [{:f}, {:f}]".format(diff_lower_bound, diff_upper_bound))
```

```
Bootstrapped 95% confidence interval for Candidate C's true lead over Candidate
T: [0.031000, 0.147000]
```

1.2 2. Interpreting Confidence Intervals

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

$$[.439, .5]$$

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

Question 2.1 Can we say that 95% of the population lies in the range $[.439, .5]$? Explain your answer.

No we cannot say that. A 95% confidence interval in this case means that,we are estimating that 95% of the times, the population paramater in question, will be captured in 95% confidence intervals and not that 95% of the population lies in that range.

Question 2.2 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.

No, the statistic being tested is the proportion of voters for candidate C which is not random and the 95% is representative of the times the population parameter will be captured in 95% confident intervals.

A note about this question (this is outside of the scope of this class. If you don't already know what Bayesian and Frequentist reasoning are, don't worry about it!): You may recall that there are different philosophical interpretation of probability. The Bayesian interpretation says that it is meaningful to talk about the probability that the interval covers the true proportion, but a Bayesian would perform a different calculation to calculate that number; we have no guarantee that it is 95%. All we are guaranteed is the statement in the answer to the next question.

Question 2.3 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?

Assign your answer to `true_proportion_intervals`.

[ ]:
```
true_proportion_intervals = 95 * 10000
```

Question 2.4

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.

Respond next to each interval:

[.444, .495]:90 Medium Interval compared to the rest hence the confidence level is neither 99% nor 80%

[.450, .490]:88 Smaller interval; so the lower the confidence that the range contains the population paramter

[.430, .511]: 90 Wider interval; so the higher the confidence that the range contains the population paramter

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 2.5

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 staff confidence interval?

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

1. Reject the null
2. Fail to reject the null
3. Unable to tell using our staff confidence interval

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

[ ]:
```
candidates_tied = 2
```

Question 2.6 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 staff confidence interval?

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

1. Reject the null
2. Fail to reject the null
3. Unable to tell using our staff confidence interval

[ ]:
```
cutoff_one_percent = 1
```

Question 2.7 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?

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

1. Reject the null
2. Fail to reject the null
3. Unable to tell using our staff confidence interval

[ ]:
```
cutoff_ten_percent = 2
```

1.3 3. Submission

Once you're finished, submit your assignment as a .ipynb (Jupyter Notebook) and .pdf (download as .html, then print to save as a .pdf) on the class Canvas site.

[ ]: