Notebook

February 3, 2020

0.0.1 Question 2a)

Let n be a positive integer and let s be an integer such that $0 \leq s \leq n$. Consider a sample of size n drawn at random with replacement from a population in which a proportion p of the individuals are called successes.

Provide a math expression for the probability that the number of successes in the sample is at most s.

In probability classes this probability will typically be denoted $P(S \leq s)$ where S denotes the random number of successes in the sample. Formal definitions of the pieces of this notation aren't particularly helpful for our purposes. Just read it as "the probability that the number of successes is at most s."

Solution

$\sum_{k=0}^{s} \binom{n}{k} p^k (1-p)^{n-k}$

Part 1 If we're trying to predict the results of the Clinton vs. Trump presidential race, what is the population of interest?

Voters who is eligible and willing to vote.

Part 2 What is the sampling frame?

Valid responses.

0.0.2 Question 5

Why can't we assess the impact of the other two biases (voters changing preference and voters hiding their preference)?

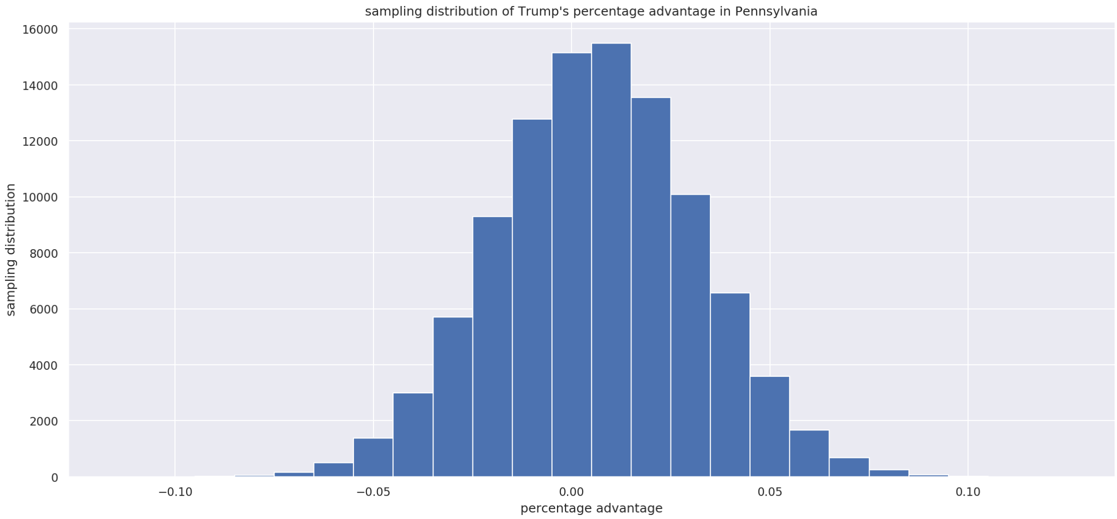
Note: You might find it easier to complete this question after you've completed the rest of the homework including the simulation study.

Because to assess these biases of sampling frame would require data out of the sampling frame, i.e. after the actual vote.

Part 4 Make a histogram of the sampling distribution of Trump's percentage advantage in Pennsylvania. Make sure to give your plot a title and add labels where appropriate. Hint: You should use the `plt.hist` function in your code.

Make sure to include a title as well as axis labels. You can do this using `plt.title`, `plt.xlabel`, and `plt.ylabel`.

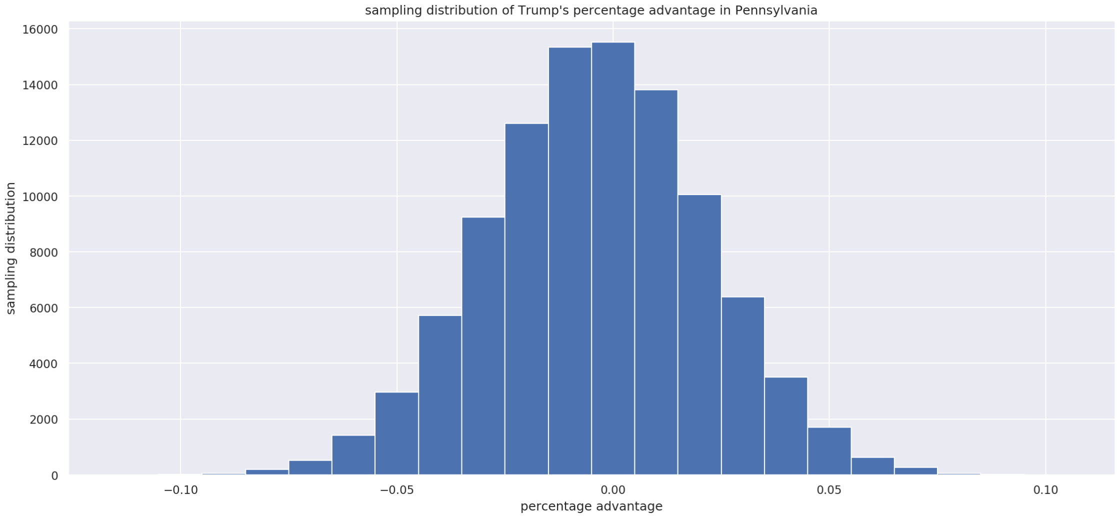
```
In [38]: plt.hist(simulations, bins=np.arange(round(min(simulations),2)-0.005,round(max(simulations),2)
         plt.xlabel('percentage advantage')
         plt.ylabel('sampling distribution', rotation=90)
         plt.title('sampling distribution of Trump\'s percentage advantage in Pennsylvania');
```



Part 2 Make a histogram of the new sampling distribution of Trump's advantage now using these biased samples. That is, your histogram should be the same as in Q6.4, but now using the biased samples.

Make sure to give your plot a title and add labels where appropriate.

```
In [45]: plt.hist(biased_simulations, bins=np.arange(round(min(biased_simulations),2)-0.005,round(max(bi
         plt.xlabel('percentage advantage')
         plt.ylabel('sampling distribution', rotation=90)
         plt.title('sampling distribution of Trump\'s percentage advantage in Pennsylvania');
```



Part 3 Compare the histogram you created in Q7.2 to that in Q6.4.

0.5% bias in favor of Clinton in each of these states will give opposite results.

Write your answer in the cell below.
Increased sample size lower sampling error but magnifies the bias.
Can get unbiased cases up to 99% success with sufficient large samples, but not biased.
About 35000.
Because larger sample size lower sampling error but magnifies the bias.

0.0.3 Question 9

According to FiveThirtyEight: ”... Polls of the November 2016 presidential election were about as accurate as polls of presidential elections have been on average since 1972.”

When the margin of victory may be relatively small as it was in 2016, why don’t polling agencies simply gather significantly larger samples to bring this error close to zero?

1. It’s hard, as a little bias, even unintentionally, would change the prediction significantly.
2. They may have their own interest which acts as an intentional bias.
3. It may be meaningless since larger sample magifies bias.
4. It costs.