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?

The population of interest would be eligible voters who are likely to vote.

Part 2 What is the sampling frame?

The sampling frame would be people with a phone because this is how the surveys are sent out.

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.

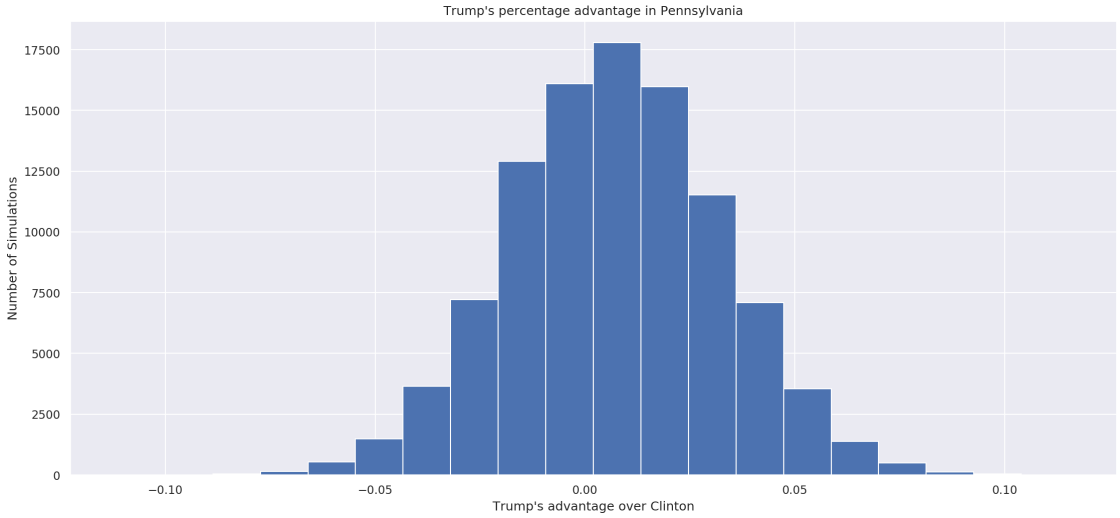
We can't assess the impact of those two biases because people don't always respond truthfully. You can't count on accounting for those two biases because you act under the assumption that people will answer truthfully.

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.title("Trump's percentage advantage in Pennsylvania")
         plt.xlabel("Trump's advantage over Clinton")
         plt.ylabel("Number of Simulations")
         plt.hist(simulations, bins = 20)
```

```
Out[38]: (array([3.0000e+00, 5.0000e+00, 4.0000e+01, 1.4500e+02, 5.2400e+02,
                 1.4830e+03, 3.6560e+03, 7.2180e+03, 1.2907e+04, 1.6089e+04,
                 1.7789e+04, 1.5966e+04, 1.1531e+04, 7.0810e+03, 3.5520e+03,
         … Omitting 3 lines …
                  0.05866667,  0.07      ,  0.08133333,  0.09266667,  0.104     ,
                  0.11533333]),
          <a list of 20 Patch objects>)
```

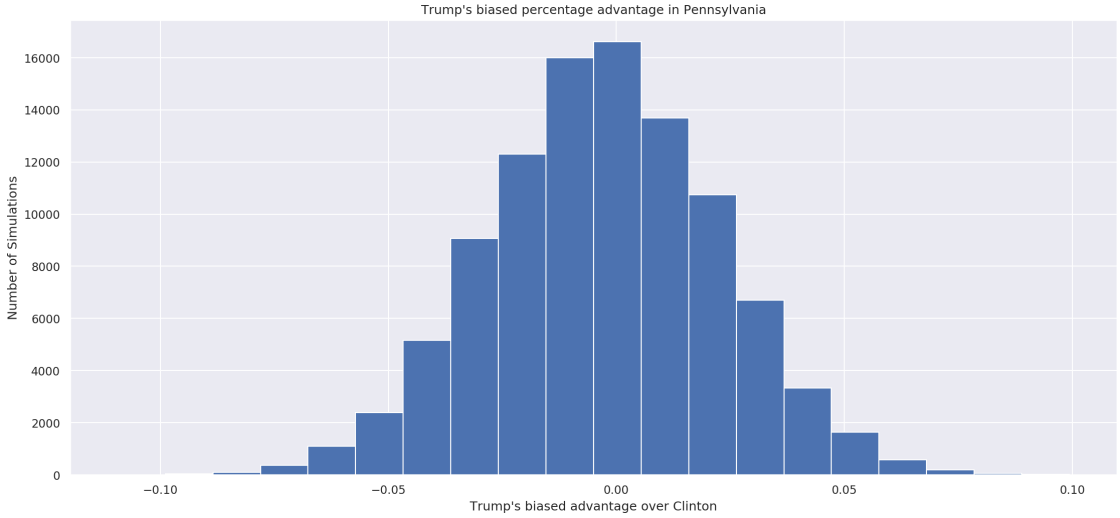


Part 2 Make a histogram of the new sampling distribution of Trump's advantage now using these biased samples. That is, your histogram should be the same as in Q6.4, but now using the biased samples.

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

```
In [45]: plt.title("Trump's biased percentage advantage in Pennsylvania")
         plt.xlabel("Trump's biased advantage over Clinton")
         plt.ylabel("Number of Simulations")
         plt.hist(biased_simulations, bins = 20)
```

```
Out[45]: (array([6.0000e+00, 2.9000e+01, 9.7000e+01, 3.6900e+02, 1.0940e+03,
                 2.3820e+03, 5.1520e+03, 9.0620e+03, 1.2295e+04, 1.5996e+04,
                 1.6614e+04, 1.3692e+04, 1.0738e+04, 6.7010e+03, 3.3210e+03,
         … Omitting 3 lines …
                  0.04716667,  0.0576    ,  0.06803333,  0.07846667,  0.0889    ,
                  0.09933333]),
          <a list of 20 Patch objects>)
```



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

The histogram based off of biased percentages shifted to the left in comparison to the histogram in Q6.4 due to the biased samples given. In this new histogram, there are less simulations in which Trump has the advantage over Clinton.

Write your answer in the cell below.

The observations from part 1 show that the biased percentages that caused the winning probabilities to be more even had a much smaller effect than the unbiased percentages. The higher sample size gave a much higher prediction for Trump to win for the unbiased percentages, showing that larger sampling sizes may decrease sampling error and bias. The prediction from the biased probabilities did not change much since the probabilities for Trump's or Clinton's victory was more even, showing that a bigger sample size does give more accurate results.

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?

It is very difficult for polling agencies to gather large amounts of data due to a lack of resources such as money and manpower. There is also the issue of non-respondents creating bias in the data and causing sampling error. Because of this, it is hard for polling agencies to bring the error to zero.