# Notebook

February 9, 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?

people who have the right to vote in the us presidential election.

**Part 2** What is the sampling frame?

people who have the right and use telephone.

### 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)?

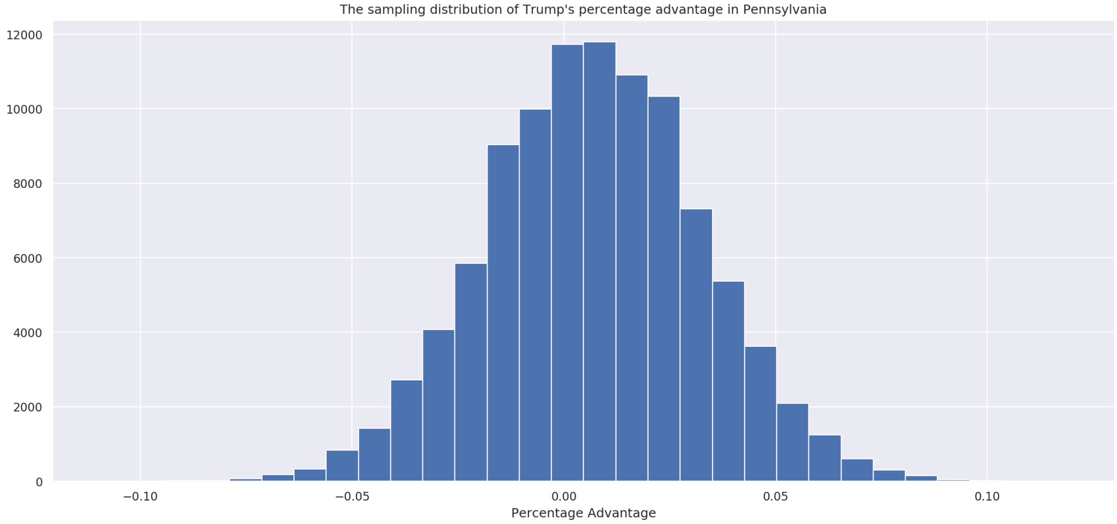
Note: You might find it easier to complete this question after you've completed the rest of the homework including the simulation study.

I think we cannot determine the number of biased peopel. One side can have more portion of biases than other one, so we cannot apply constant portion to those. Therefore, if we want to access the impact, we have to gather samples of biases (infinitely), or just assume the portion. Both cases would not give good effect for the sampling.

**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`.

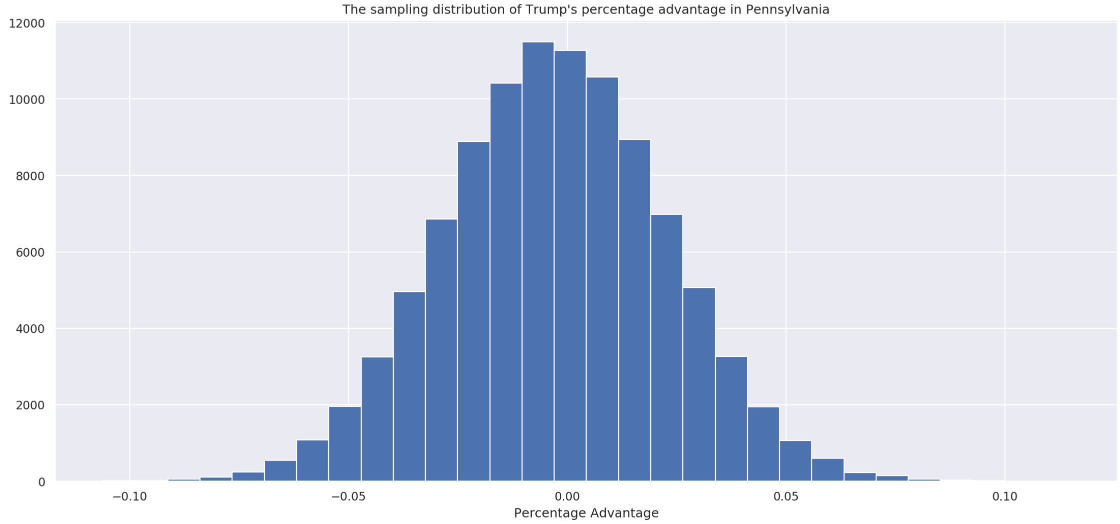
```
In [38]: plt.hist(simulations, bins = 30);
         plt.title("The sampling distribution of Trump's percentage advantage in Pennsylvania");
         plt.xlabel("Percentage Advantage");
```



**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 = 30);
         plt.title("The sampling distribution of Trump's percentage advantage in Pennsylvania");
         plt.xlabel("Percentage Advantage");
```



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

The historam in Q7.2 generally moves to the left. Based on 0.00, the first histogram shows there are more blue space at the right side, but the second one seems there are more blue spce at the left side. It means there is less possibility to win for Trump in the second case.

Write your answer in the cell below.

I think the large sample size makes the possibility of win (or lose) extreme. I tested the larger sample size, and it shows Trump wins 99.99% when there is no bias. Originally, there is a little difference in percentage in voters, but if we test many times with large sample sizes, we can get a result that shows high possibility.

### 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?

I thought about the case we gather the larger sample with replacement than the population of interest (like square of the population). In that case, the result shows one person wins with extremely high possibility, and it cannot represent real world. I am not sure about the case without replacement, but I guess it is better to find a good samples than gather large sample sizes. It would be efficient to find the way to gather balenced samples who represent each group according to race, sex, age, ect.