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

It is people who are eligible to vote, registered to vote and willing to vote.

Part 2 What is the sampling frame?

The sampling frame is people who have telephones because only these people can be reached by random digit calling.

0.0.1 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.

Firstly, we cannot assess the impact of voters changing their preferences because they do so after the poll is complete and before the election so there is no additional follow-up poll for the voters' to fill out in case their preferences change. So this bias is not taking into account while conducting the poll.

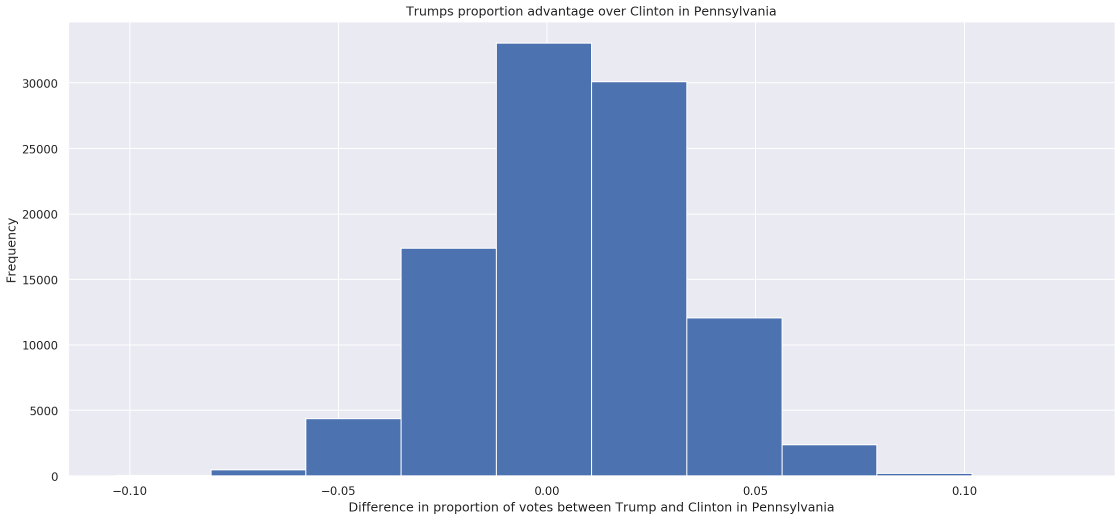
Secondly, we cannot assess the impact of voters hiding their preferences because there is no way to tell if someone is telling the truth or lying. Furthermore, since these polls are not conducted in person, you can't even employ psychologists to study individuals to guage whether their response is true or not. Thus, this bias can also not be taken into account while conducting random-digit-calling for polls.

Part 4 Make a histogram of the sampling distribution of Trump's proportion 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 [136]: plt.hist(simulations)
          plt.title('Trumps proportion advantage over Clinton in Pennsylvania')
          plt.xlabel('Difference in proportion of votes between Trump and Clinton in Pennsylvania')
          plt.ylabel('Frequency')
```

```
Out[136]: Text(0, 0.5, 'Frequency')
```

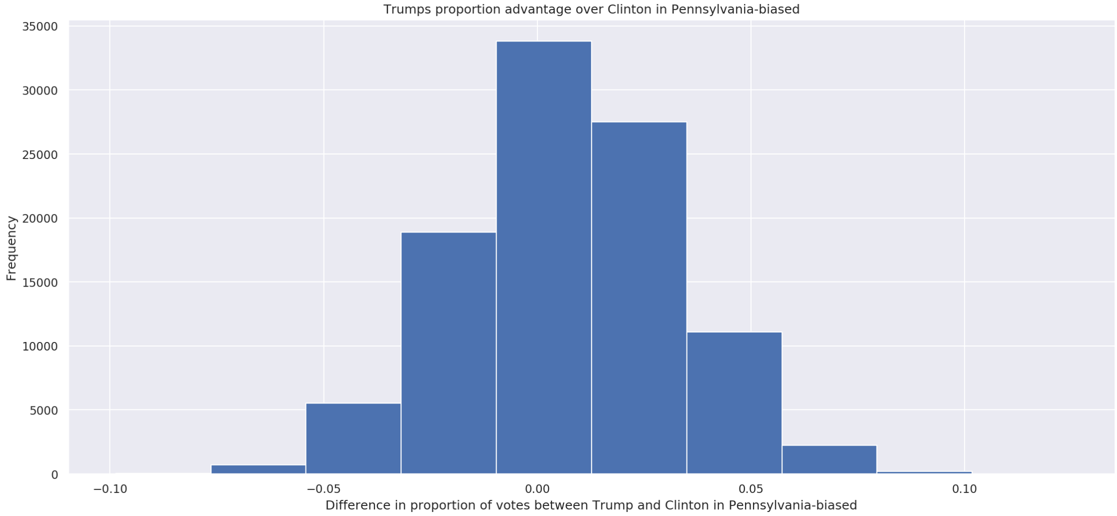


Part 2 Make a histogram of the new sampling distribution of Trump's proportion advantage now using these biased samples. That is, your histogram should be the same as in Q6.4, but now using the biased samples.

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

```
In [143]: plt.hist(biased_simulations)
          plt.title('Trumps proportion advantage over Clinton in Pennsylvania-biased')
          plt.xlabel('Difference in proportion of votes between Trump and Clinton in Pennsylvania-biased
          plt.ylabel('Frequency')
```

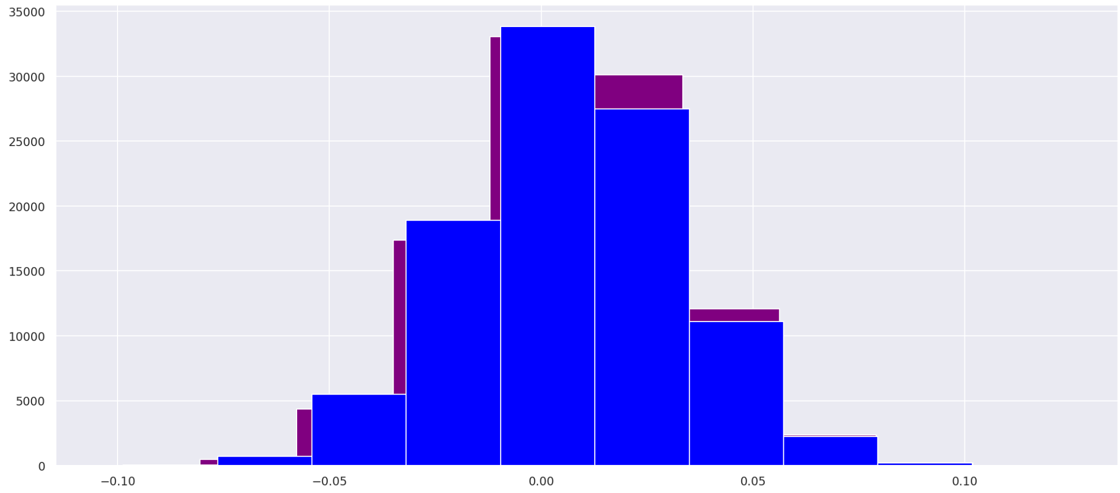
```
Out[143]: Text(0, 0.5, 'Frequency')
```



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

```
In [144]: plt.hist(simulations, color='purple')
          plt.hist(biased_simulations, color='blue')
```

```
Out[144]: (array([4.8000e+01, 7.1000e+02, 5.5090e+03, 1.8880e+04, 3.3816e+04,
                  2.7492e+04, 1.1084e+04, 2.2430e+03, 2.0500e+02, 1.3000e+01]),
           array([-0.09866667, -0.0764    , -0.05413333, -0.03186667, -0.0096    ,
                   0.01266667,  0.03493333,  0.0572    ,  0.07946667,  0.10173333,
                   0.124     ]),
           <a list of 10 Patch objects>)
```



The histogram using the biased samples shifted to the left as compared to the from Q6.4. In Q6.4, the highest frequency proportions were much towards the right of 0, but now they are only slightly right of 0, or at 0.

Write your answer in the cell below.

Using a larger sample size increases the proportion of Trump wins for the unbiased sample and decreases the proportion of Trump wins for the biased sample. The unbiased sample of 5000 had a Trump win rate of about 82%, while the unbiased sample of 10000 had a Trump win rate of about 91%. This illustrates a decrease in the sampling error due to larger sample size.

On the other hand, the biased sample of 5000 had a Trump win rate of about 44%, while for the biased sample of 10000, the Trump win rate went down to about 42%, which is only 2%. So the larger sample size did not have a significant effect on the bias.

0.0.2 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?

A larger sample size is time consuming and expensive to study. Even if they were to use a larger sample size, this does not guarantee that the error would come closer to zero because there are often factors that make the study inaccurate which cannot be eliminated by using a larger sample size. An example would include a Republican supporter in California who would not openly voice their political opinions, because most people around them would be in support of the Democratic party, so they would feel the need to lie in political polls in order to conform to societal norms. Furthermore, these polls are often conducted too early before the elections and events occurring after the poll and before the election have the ability to change the voters’ political opinions. An example would include COVID-19 and Trump’s reaction to COVID-19 may affect a voters’ opinion of him. Furthermore, the technique of conducting the polls may not always be representative of the whole population. E.g. if the poll is being conducted on the internet, not everyone has access to the internet so it excludes low income individuals who do not have access to the internet.