# Data 100/200 Homework 2 Written

Jackie Hu

TOTAL POINTS

**10 / 11**

QUESTION 1

## Question 4 2 pts

### 1.1 4.1 1 / 1

✓ **+ 1 pts** **The population is everyone who voted in 2016 election.**

**+ 0 pts** Incorrect/blank

### 1.2 4.2 0 / 1

**+ 1 pts** Anyone who has the phone (the exclusion would be a bonus)

✓ **+ 0 pts** **Incorrect / blank**

QUESTION 2

## 2 Question 5 1 / 1

✓ **+ 1 pts** ****Correct: ** We can't determine whether voters changed preferences or hid their preferences until after the election.**

**+ 0 pts** **Blank/Incorrect**

QUESTION 3

## Question 6 1 pts

### 3.1 6.4 1 / 1

✓ **+ 1 pts** **Correctly drawn histogram and labels (title, axis) used when appropriate.**

**+ 0.5 pts** Correctly drawn histogram.

**+ 0 pts** Incorrect/blank

QUESTION 4

## Question 7 3 pts

### 4.1 7.2 1 / 1

✓ **+ 1 pts** **Correct graph and labels**

**+ 0.5 pts** Correct graph but missing/incorrect labels

**+ 0 pts** Incorrect/Blank

### 4.2 7.3 2 / 2

✓ **+ 2 pts** **Reasonable comparison of shapes/symmetry and identification of left shift**

**+ 1 pts** Reasonable comparison of shapes/symmetry or left shift

**+ 0 pts** Incorrect/Blank

QUESTION 5

## Question 8 2 pts

### 5.1 8.2 2 / 2

✓ **+ 2 pts** **Fully correct**

**+ 1 pts** Random error _is_ affected and reduced by larger sample sizes, making the unbiased one more likely to be correct

**+ 1 pts** Bias/shift is not affected by a larger sample size and so the biased one remains inaccurate despite decreases in random error (the bias is more pronounced)

**+ 0 pts** Incorrect/blank, or described the effects on accuracy without alluding to the sampling error or bias

QUESTION 6

## 6 Question 9 2 / 2

✓ **+ 2 pts** **Explains how sample bias is not eliminated by taking a larger sample**

**+ 1 pts** Gives a reason such as lack of time/money but does not explicitly identify sample bias

**+ 0 pts** Incorrect/blank

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

Everyone who has voted in the presidential race.

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

Everyone who has voted in the presidential race and participated in the survey, aka the people who are eligible to be sampled.

### 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.

Because first of all, we have no way to measure these biases, second even if we have data about these potential biases, we have no way knowing how would these biases affect our data and our prediction; since a small percentage of bias could dramatically change the prediction results.

**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.

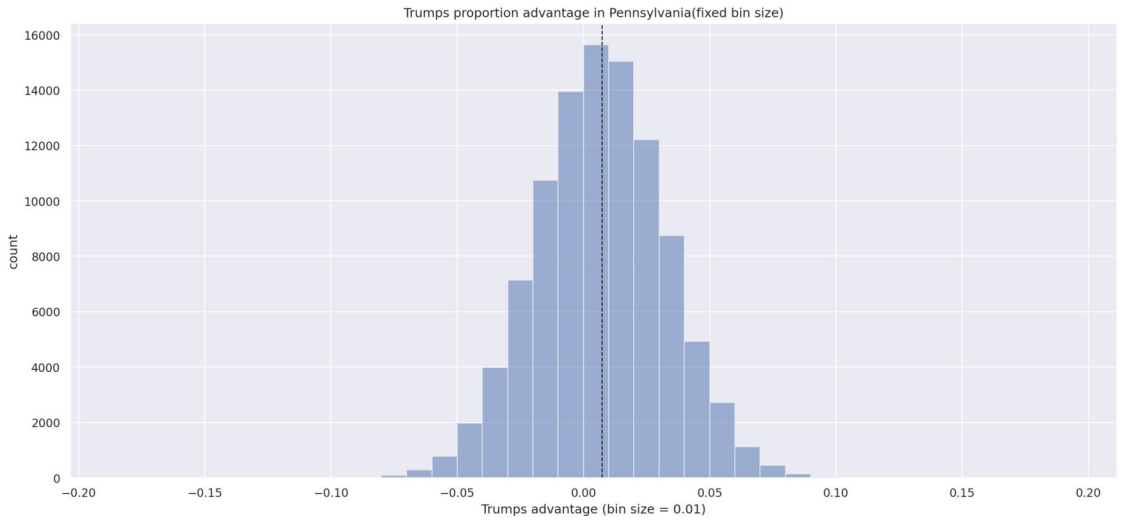
Make sure to include a title as well as axis labels. You can do this using `plt.title`, `plt.xlabel`, and `plt.ylabel`.

```
In [47]: #calculate a mean value
         data = simulations
         mean_val1 = sum(data)/100000

         # fixed bin size
         bins = np.arange(-0.2, 0.2, 0.01)

         plt.xlim([min(data)-0.1, max(data) + 0.1])
         plt.hist(data, bins=bins, alpha=0.5)
         plt.title('Trump''s proportion advantage in Pennsylvania(fixed bin size)')
         plt.xlabel('Trump''s advantage (bin size = 0.01)')
         plt.ylabel('count')
         plt.axvline(mean_val1, color='k', linestyle='dashed', linewidth=1)

         plt.show()
```



**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.

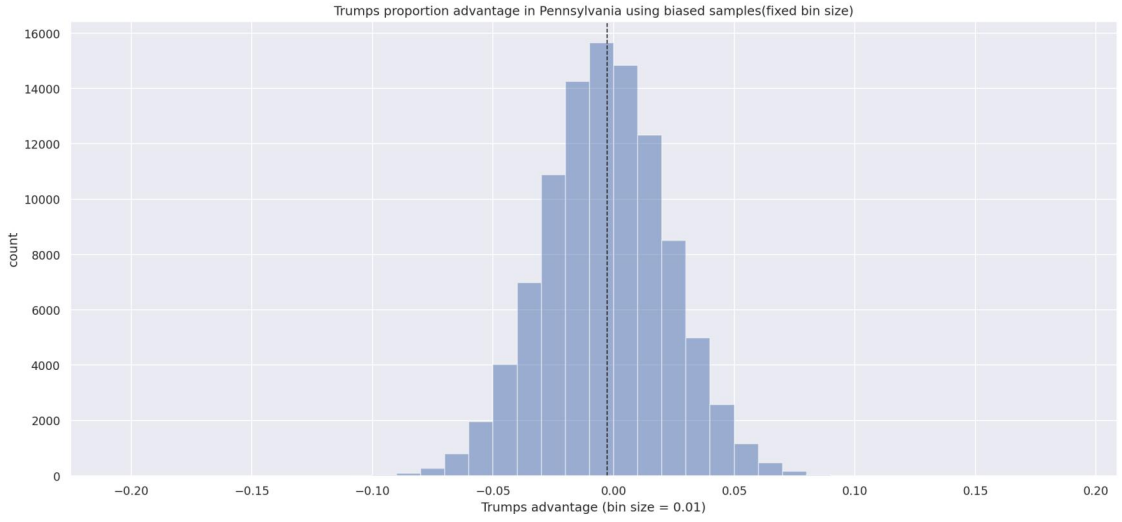
In [123]:
```
#calculate the mean value for the bias simulations
data_bias = biased_simulations
mean_val = sum(data_bias)/100000

# fixed bin size
bins = np.arange(-0.2, 0.2, 0.01)

plt.xlim([min(data_bias)-0.1, max(data_bias) + 0.1])

plt.hist(data_bias, bins=bins, alpha=0.5)
plt.title('Trump''s proportion advantage in Pennsylvania using biased samples(fixed bin size)
plt.xlabel('Trump''s advantage (bin size = 0.01)')
plt.ylabel('count')
plt.axvline(mean_val, color='k', linestyle='dashed', linewidth=1)

plt.show()
```



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

From the graph we can see the biased data has a negative mean line, that shows Trump does not have a proportion of advantage in Pennsylvania; where in the original data we have a positive mean line, that shows Trump has an advantage in Pennsylvania.

This discrepancy between the mean value shows that bias data can provides the wrong prediction, even the bias is minor.

Write your answer in the cell below.

With the unbiased proportion rate, the sample prediction can get really close to the reality, in that case, increasing sample size can increase the prediction accuracy;

But with the biased data, up the sample size made the prediction results more biased, it doese not help with the prediction accuracy.

### 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?

As we see in the high sample size prediction model using biased and unbiased data, the biased rate even just 0.5% results in a big difference in prediction and the reality, simply increase sample size did not help with increase the prediction accuracy nor lower the bias rate, it only made the predicction more biased.

And because we can never 100% neglect the potential bias during our survey and initial data gathering, we can't eliminate the bias rate by just up the sample size.