Question 1

After calling all 200 numbers, I had one Response = 1. The other 199 numbers fell into three cases: either the number was invalid (only 61 numbers were valid), the individual did not pick up and it went to voicemail, or the individual chose to hang up before I began asking the voting question. My response rate was 0.5%.

As I had one Response = 1 who also responded to both the voting and age question, the fraction of those for whom Response = 1 answered the voting and age question is 1.

I conducted this part of the assignment on Sunday, 7pm CST. The area code for the numbers I called was near the western Washington state, outside of Seattle, which is 2 hours behind. This means I called the area at around Sunday, 5pm (with the assumption that the individuals are still located around the area in which they obtained their phone number from). I decided to call the numbers at this time for two reasons: Firstly, I believed that most people are at home on Sunday evenings, and were more likely to pick up the phone at this time. Additionally, I also considered that most individuals on Sunday are in a positive mood, which could mean that they are more likely to answer to my questions, especially if they receive an anonymous phone call to respond to a survey.

The age of my one respondent is 24 years old. The average age in Washington state is 37 years old, which is substantially higher than my respondent's age. I postulate that my sample size of 1 does not match the state data not only because it's a very small sample size, but also because my respondent self-selected into my sample. To elaborate, I suspect that my respondent chose to respond to my questions because he related to my description as a graduate student at University of Chicago as he is at the age where he or people in his network are either considering graduate school or are in one now. In comparison, others who chose to hang up sounded substantially older on the phone. As such, perhaps they did not relate to my description of being a graduate student as much, and had no qualms in refusing to participate in my survey as a result.

My respondent did not vote in the 2016 presidential elections, but he did indicate that he voted for Bernie Sanders in the Democratic presidential primaries. This is reflective of the demographic of Bernie Sanders' base during the presidential primaries– younger individuals.

To test if the order in which the interviewer mentions the candidates has an effect on the results of the survey, one can split the sample randomly into two groups – under each group, the interviewer asks a different order, while every other condition remains the same. For example, in the first group, the interviewer would ask "In the 2016 U.S. Presidential Election, did you vote Democrat (Clinton), Republican (Trump), Other, or Did Not Vote?" compared to the second group, in which the interviewer would ask "In the 2016 U.S. Presidential Election, did you vote Republican (Trump), Democrat (Clinton), Other, or Did Not Vote?". Assuming that these two groups are randomly assigned and everything else remains the same (for example, the time in which the interviewer calls both groups is the same), if there is a statistically significant difference in how respondents respond to the question between the two groups, we can then infer that this difference in results is due to the order in which the interviewer asked the question.

Question 2

In Wang et al.'s "Forecasting Elections with Non-Representative Polls" (2015), of the eight variables reported from the respondents, the three most representative variables compared to the 2012 electorate, are state, race, and who they voted for in the 2008 elections. In comparison, age, sex and education seem to be the least representative of the 2012 electorate- in figure 1, which shows the percentage differences between the Xbox sample and the 2012 electorate, age, sex and education seem to have the three largest differences between the sample and the electorate in general. Specifically, for age, the Xbox sample is generally much younger than the average 2012 electorate – with approximately 65% of the Xbox sample aged between 18-29, while about 20% of the 2012 electorate are aged between 18-29. For sex, the Xbox sample is again disproportionately male with about 5% of Xbox respondents who are female in the sample, while around 55% of the 2012 electorate are female. For education, although both the Xbox sample and the 2012 electorate's general trends of education from "Didn't graduate high school" to "Some College" are somewhat similar, there is around a 20% gap between the samples for "College Graduates", with most of the Xbox sample's level of education ranging from "High School Graduates" to "Some College". One can hypothesize then that the Xbox sample could differ from the 2012 electorate in age, sex and education because most Xbox gamers are typically young male college students or recent high school graduates who have the time and resources to play video games that are stereotypically male-oriented.

The two data sources that the authors used to perform post-stratification re-weighting of the respondents is the Xbox data itself and the exit poll data from the 2008 presidential election. Specifically, they used the exit poll data to reweight the Xbox data, by using the proportion of the electorate in each category and aggregating this number to an appropriate level, to obtain a representative sample from the Xbox data.

Based on the Xbox raw data alone, one would predict that Mitt Romney would win, with varying degrees, throughout the last three weeks before the 2012 presidential elections. One would also predict that Mitt Romney would win with a fairly large win at the end, as support for Obama would be predicted to be around 45%. In comparison, if one used the Pollster.com forecast data, one would predict a much closer race throughout the last three weeks, with support for Mitt Romney hovering around 50-51% until a few days before the elections, in which Obama wins at the end with a very slight victory. Based on the Xbox post-stratified data, one would predict that Obama would win three weeks before the elections and on the election day itself, with the predicted support for Obama at the end closer to the actual vote share in the 2012 presidential elections than the Pollster.com forecast data. This shows that after post-stratifying polling results from a non-representative sample, we can use it to generate very accurate election forecasts that may even surpass polling results taken from representative samples.

References

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting* 31 (3): 980-991.