Assignment 4

Fulin Guo

1. (a). The *PhoneSurvey.xlsx* spreadsheet is in the file A4.

   (b). I called 200 numbers. 0 people responded as there are 0 people whose *Response* variable is 1. 200 people did not respond as there are 200 people whose *Response* variable is 0. (19 people answered the phone but all of them refused to take the survey (Most of them hung up the phone before the vote question and the rest said they did not have time, or they were not interested in the survey)). The response rate is 0%.

   (c). Because the number of people whose *Response=1* variable is 0, no people answered the vote question or the age question. (I cannot answer the fraction question because no one has the Response=1)

   (d). I call the first ten phone numbers at about 5pm (Eastern Time, https://time.is/ET) on Sunday. I call the last 190 phone numbers from 11am to 2pm (Eastern Time, https://time.is/ET) on Tuesday. I think the time of day plays an important role in the response rate. Theoretically, calling in the daytime would result in higher response than calling after 10pm or before 6am because many people are sleeping during that time. On the other hand, many people are working or studying from 8am to 6pm, so they might not have the time to answer the survey, which will have a negative effect on response rate. I think this might be one reason that my response rate is zero. (My area code is 212, which is the area code of New York, https://en.wikipedia.org/wiki/Area_codes_212,_646,_and_332)

   (e). There is no median age in my survey because no one responded the survey. The median age of New York is 36.2. (https://datausa.io/profile/geo/new-york-ny/).
   Although I cannot compare the sample median and the real median. I can propose(guess) some factors that might cause a mismatch between the two **if** this happens. (1) the sample size is small (we can only get at most 10 responses according to the question's requirements), which means the variance of sample median is too big for us to accurately estimate the real

median age of New York based on the survey at every time. (2) The second reason is that there might be a correlation between people's age and whether answering the question. For example, young people might be busier than old people because they need to study and work more. Therefore, older people will have a higher possibility to have time to answer the question. (3) The third reason is that some old people might not have a phone number, which tends to lower the sample median age compared to the real median age. (4) The forth reason is that as the question required, we can only record the people's answer whose age is at least 18, which means we are calculating the sample median age of people who are 18 years old or older. This is another factor that can make the sample median higher than the real median.

(f). Because no people responded the voting question, there is no sample voting percent. The real situation is that 46.1% people voted Republican (Trump), while 48.2% people voted Democrat (Clinton). (https://en.wikipedia.org/wiki/United_States_presidential_election,_2016). In New York, 37.5% voted Republican and 58.8% voted Democratic. (https://www.politico.com/2016-election/results/map/president/new-york/)

To test whether the order that I say the two candidates influences the survey result, I can say Republican first Democrat second in the first 100 calls and reverse the order in the second 100 calls. Then I can test whether the difference between the results in first 100 and second 100 calls is significant. If there is a significant difference, the order may play a role in the results, if it is not significant, we cannot reject the assumption that the order does not influence the results.

2. (a) From the figure 1 of (Wang et al., 2015), the three least representative variables are sex, age and education. The three most representative variables are state, race and 2008 vote. For the sex variable, I think male likes playing video games more than female, so it is reasonable that most people that reply the survey in Xbox are male as more men use Xbox than women. For age, young people tend to play video games more than old people, so it would make sense that the percent of young people conducting the survey is higher than the percent in real voting situation. For education, there might

be some correlation between education level and the behavior of playing video games. For example, if the people who has college graduate education tend to play less video games (perhaps because they spend most of their time in life studying and working), it will make sense that the percent of college graduate students in the Xbox sample is much lower than that in the border election data.

(b) The author uses the (state and national) "exit poll data from the 2008 presidential election" (Wang et al., 2015, p984) and Xbox for poststratification.

(c) From the figure 1 of (Wang et al., 2015), the unweighted Xbox data predict Romney wins during the last three weeks as the percent of supporting Obama is less than 50%. We could also see from the figure 1 (Wang et al., 2015) that based on the Pollster.com, who win the election is uncertain in the last three weeks because the percent of Obama support is rough 50% during those time. From the figure 3 of (Wang et al., 2015), the post-stratified Xbox data predict Obama wins because the percent of Supporting Obama is higher than 50% during the last three weeks.

Reference:

https://time.is/ET
https://en.wikipedia.org/wiki/Area_codes_212,_646,_and_332
https://datausa.io/profile/geo/new-york-ny/
https://en.wikipedia.org/wiki/United_States_presidential_election,_2016
https://www.politico.com/2016-election/results/map/president/new-york/

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