Assignment #4
MACS 30000, Dr. Evans
Due Wednesday, Oct. 31 at 11:30am
Haowen Shang

1. Non-probability sampling phone survey

(a) Please see 'PhoneSurvey-HaowenShang.xlsx'.

(b) I have called 200 numbers. 110 numbers were disconnected or no longer in service. Among the 90 valid numbers, most of them directly went to voice mail boxes. 18 numbers have been connected but 14 people refused to answer the questions and 2 people said they were under 18 years old and their parents were not around them. Fortunately, 2 guys were willing to answer the questions.
In conclusion, 2 people responded according to my Response variable and 198 people did not respond according to my Response variable. My response rate is 1.0%.

(c) Among the 2 respondents, the fraction of answering the voting question is 100% and the fraction of answering the age question is 100%.

(d) I called the numbers from 6 pm to 9 pm on Friday night. At that time, people are not busy at work or study and not sleeping, so the response rate are supposed to be higher than working hours or sleeping hours.

(e) The median age of my respondents is 54.5 years old while the median age in Indiana is 37.4 years old [1].

The median age of my respondents is older than the state data because people who are under 18 years old are not considered as respondent but they are included in the state data. Also, the sample size of respondents is too small and the sample is not representative.

(f) 50% of my respondents voted Republican (Trump) in the 2016 U.S. Presidential election and 50% of my respondents voted Democrat (Clinton).
The actual voting percentages of the 2016 election in Indiana are that 57.2% of voters voted Republican (Trump) and 37.9% of voters voted Democrat (Clinton) [2].

Compare to the actual voting percentages, Republican voting percentage is lower while Democrat voting percentage is higher in my sample.
In order to test if the order in which I say the candidates or categories in the survey question influences the results, I may do the survey several times again and each time randomly selected 200 numbers from the same area code (317). I'll ask the survey question in this order of candidates and categories (Republican in the first and then Democrat) in half of my experiments

and in different order of candidates and categories (Democrat in the first and then Republican) in other half of my experiments and find whether the voting percentages are different in these two groups.

Reference:

1. U.S. Census Bureau: American FactFinder
   https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_16_5YR_GCT0101.ST04&prodType=table
2. Politico: 2016 Presidential Election Results
   https://www.politico.com/mapdata-2016/2016-election/results/map/president/

2. Predicting elections survey, Wang, Rothschild, Goel, and Gelman (2015) (3 points). Read the paper Wang et al. (2015), and write a one-to-two-page responding to the following questions.

In the paper "*Forecasting Elections with Non-Representative Polls*", the authors indicated that, compared with traditional survey methods, when using "proper statistical adjustment", non-representative polls can generate accurate election forecasts with faster time and lesser expense (Wang, Rothschild, Goel, and Gelman, 2015, p.980).

In the last 45 days of the 2012 US presidential election, the authors conducted an opt-in poll on the Xbox gaming platform (Wang, Rothschild, Goel, and Gelman, 2015, p.981). Based on the results of this opt-in poll, we can see that although the sample size is very large, the Xbox sample is still not representative for the broader voting population (Wang, Rothschild, Goel, and Gelman, 2015, p.981). In Figure 1, we can compare the Xbox sample with electorate estimated by the 2012 national exit poll. In this figure, we can see that the Age variable, Sex variable and Education variable from the Xbox sample are the least representative of the data, and the Race, State and 2008 Vote are the most representative. For the Age variable, people who are between 18 to 29 years old comprised 65% of the Xbox sample, while just comprised 19% of the 2012 electorate. For the Sex variable, men comprised 93% of the Xbox sample, while only comprised 47% of the electorate. For the Education variable, there were more low-level educational population in Xbox sample than in electorate because college graduate student comprised just 25% of Xbox sample but comprised 50% of the electorate in exit poll (Wang, Rothschild, Goel, and Gelman, 2015, p.981-982). The Age, Sex and Education variables from the Xbox sample are the least representative, because Xbox poll was conducted on a gaming platform, whose users are mostly comprised by young men with lower education level.

The raw Xbox poll data is not representative for general electorate, so the authors used the methods of "multilevel regression and poststratification" to make the Xbox data much more representative: "The core idea is to partition the population into cells based on combinations of various demographic and political attributes, use the sample to estimate the response variable within each cell, and finally aggregate the cell-level estimates up to a population-level estimate by weighting each cell by its relative proportion in the population." (Wang, Rothschild, Goel, and Gelman, 2015, p.983). Cross-tabulated population data is required to compute the weights of each cell, while the Current Population Survey (CPS) data missed some key poststratification variables, the authors used exit poll data from the 2008 presidential election instead (Wang, Rothschild, Goel, and Gelman, 2015, p.984). Thus, the two data sources for performing the post-stratification re-weighting of the respondents are exit poll data from 2008 presidential election and Xbox poll data.

In order to show the prediction results of 2012 U.S. Presidential election and compare the accuracy of different methods, the authors provided some figures. The Figure 2 showed that Xbox raw(unweighted) data predicted that Mitt Romney would win because in the last three weeks of the election, the two-party support rate for Obama was below 50%. The Pollster.com

forecast data predicted that Obama's support rate maintained around 50% during the last three weeks of the election, so the election result was uncertain (Wang, Rothschild, Goel, and Gelman, 2015, p.982). According to Figure 3 and Figure 5, we can see that Xbox post-stratified data's prediction was much more reasonable and representative. The Xbox post-stratified data predicted that Obama would win the election, because in the last three weeks, the predicted two-party support rate of Obama was always above 50% (Wang, Rothschild, Goel, and Gelman, 2015, p.984, p.986).

Reference:

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