Assignment #4

MACS 30000, Dr. Evans
Due Wednesday, Oct. 31 at 11:30am

1. **Non-probability sampling phone survey (7 points).** In GitHub Issue #18, you are assigned a U.S. area code. Within this area code you are to conduct a simple phone survey. The file `PhoneSurvey.xlsx` has a list of 200 randomly generated 7-digit phone numbers for you to call as well as three extra columns for you to record your responses. You first dial your assigned 3-digit area code, followed by the 7-digit phone number in the Excel spreadsheet. You are to stop calling people as soon as you either have 10 *Response* = 1 respondents OR you have called all 200 numbers OR you have spent three hours calling phone numbers. For each call, you are to carry out the following survey design.

    - Dial the first number on your list.
        - If nobody answers the phone or if the number is not an active phone number or if you are not able to talk to anyone 18 years old or older, mark a 0 in the "Response" column in the spreadsheet, and move to the next number in your list.
        - If somebody answers the phone:
            * State the following introductory information: "Hello, my name is [YourName], and I am a graduate student at the University of Chicago conducting a brief, three-question research survey."
            * "Are you 18 years old or older?"
                · If not, ask: "May I speak with the head of household who is over 18 years old?"
                · If yes, wait for new respondent, repeat your identification information and ask the following two questions.
            * "In the 2016 U.S. Presidential Election, did you vote Democrat (Clinton), Republican (Trump), Other, or Did Not Vote?"
            * "What is your age?"
        - Close with: "Thank you for taking the time to take this survey."
        - If they ask you what it will be used for, tell them: "These survey results will be used for survey design training at the University of Chicago."
    - Go to the next number on your list and repeat the process.

    The variable *Response*=1 if a person answers the phone, is over 18 years old, and allows you to begin asking the voting question.

(a) Submit your filled out version of the `PhoneSurvey.xlsx` spreadsheet.

(b) How many numbers did you call? How many people responded according to your *Response* variable? How many people did not respond according to your *Response* variable? What is your response rate?

(c) What fraction of those for whom $Response = 1$ answered the voting question? What fraction of those for whom $Response = 1$ answered the age question?

(d) What time of day was it in the area codes you called when you called them? What role did the time of day play in your response rate?

(e) What is the median age of your respondents? How does that compare to the average age in the state of the phone numbers you called?[1] What are some reason's why your sample median does or does not match the State data?

(f) What percent of your respondents voted Republican (Trump) in the 2016 U.S. Presidential election? What percent of your respondents voted Democrat (Clinton)? How do those percentages compare to the actual voting percentages from the 2016 election?[2] How might you test if the order in which you say the candidates or categories in the survey question influences the results?

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.

(a) Of the eight variables reported from the respondents, which three from the Xbox sample are the least representative of the data and which three are the most representative? For the three least representative variables, why do you think the Xbox sample would be so different from the broader voting population?

(b) What two data sources do the authors use to perform a post-stratification re-weighting of the respondents? Figure 5 shows the results of degree of representativeness of the re-weighted data.

(c) Consider Xbox raw (unweighted) data, Pollster.com forecast data, and Xbox post-stratified data. And consider a prediction of a 2012 U.S. Presidential election outcome as Obama wins, Romney wins, or uncertain. What would Xbox raw have predicted in the last three weeks of the election? What would Pollster.com have predicted during the last three weeks of the election? What would Xbox post-stratified have predicted in the last three weeks of the election?

[1] You can look up the median age for each U.S. state in 2016 using this U.S. Census Bureau Factfinder table.

[2] Politico has a nice election results site where you can check the percentages by state.

References

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