Assignment #4

MACS 30000, Dr. Evans
Ying Sun

1. Non-probability sampling phone survey

(a) See the attached PhoneSurvey.xlsx

(b) I called all 200 numbers. But among these 200 numbers, only 5 are valid numbers. Among these 5 numbers, three of them were directly transferred to voicemail. Unfortunately, only one call was connected in the left two numbers. But he refused to answer. So according to my sampling phone survey, no people responded. My response rate is 0.

(c) For the only people respond to my phone, he hung up when I asked about whether he is over 18 years old. I could not continue my survey on voting question. So there is no variable Response =1 in my survey result. As a result, I may cannot answer the questions like what fraction of those for whom Response = 1 answered the voting question or what fraction of those for whom Response = 1 answered the age question.

(d) My area code is 339 for Massachusetts (GMT-4). And time I describe below are all converted to GMT-4 time zone. I called these numbers in to two days (Saturday and Sunday) and tried to make these calls at different times in two days. On Saturday I called 100 calls during 7:00pm – 8:00 pm while on Sunday I called the other 100 calls during 3:00pm – 4:00 pm. Because I think people may have more time on weekends and these two time periods may could lead to more responses. The only call was connected was on Saturday night even though he refused to answer my questions, it seems that 7:00 pm – 8:00 pm is a relatively better period to have more chance to get response. But because of the really limited sample and the 0 response in both two days, so I cannot conclude that 7:00 pm – 8:00 pm would plays role in response rate.

(e) I have no response so I do not have any age data to answer this question so I cannot compare it with the average age in Massachusetts and further answer the question that what are some reason's why my sample median does or does not match the state data. According to the data[1], the median age of Massachusetts is 39.4. But basically, I think some certain factors may cause this difference, such like the time you called these numbers because old people may have more available time in the daytime of working days while young people may not.

(f) I have no response so I do not have voting data to answer these questions. According to the actual voting percentages from 2016 election,[2] Hillary Clinton wined 60.8% electoral votes while Donald Trump wined 33.5% electoral votes in Massachusetts. To test if the order in which I say the candidates or categories in the survey question influences the results, we can select another random sample which may include more valid numbers and has a relatively

[1] https://factfinder.census.gov/faces/nav/jsf/pages/community_facts.xhtml
[2] https://www.politico.com/mapdata-2016/2016-election/results/map/president/

higher response rate, and redo this survey. Just randomly mention the names of Donald Trump and Hillary Clinton or Republican and Democrat in different orders and analyze the result to see if the order will cause different results and test if this relationship (if have) is significant.

Reference

1. Data Access and Dissemination Systems (DADS), 2016. Retrieved from https://factfinder.census.gov/faces/nav/jsf/pages/community_facts.xhtml

2. 2016 Election Results: President Live Map by State, Real-Time Voting Updates. Retrieved from https://www.politico.com/mapdata-2016/2016-election/results/map/president/

2. Predicting elections survey, Wang, Rothschild, Goel and Gelman (2015)

Different from traditional election forecasts which mostly based on representative polls, this paper tries to show that with proper statistical adjustment, non-representative polls are able to generate accurate election forecasts (Wang, Rothschild, Goel, and Gelman, 2015, p.980). Basically, the authors conduct election survey on the Xbox gaming platform during 45 days leading up to the 2012 US presidential race (Wang, Rothschild, Goel, and Gelman, 2015, p.981). Then they construct daily estimates of voter intent via multilevel regression and poststratification (MRP) and transform voter intent into projections of vote share and electoral votes (Wang, Rothschild, Goel, and Gelman, 2015, p.981).

The authors choose eight variables: sex, race, age, education, state, party ID, political ideology and who they voted for in the 2008 presidential election to reflect basic demographic information. Comparing the demographic composition of the Xbox participants and general electorate of 2012 national exit poll, we can find, among the eight variables, the three least representative of the data are age, sex and education. 18-29-year-olds comprises 65% of the Xbox dataset, compared to 19% in the exit poll; and men make up 93% of the Xbox dataset but only 47% of the electorate (Wang, Rothschild, Goel, and Gelman, 2015, p.981). In terms of education, Xbox players have relatively low education level compared with the voting population. In the contrast, the three most representative variables are race, state and 2008 vote. As for the three least representative variables, why the Xbox sample would be so different from the broader voting population. I think this can be explained by the totally different population composition between Xbox players and general electorates. Same as our expectation, most of Xbox participants are young men, which are very different from the population composition of general electorates. According to research results, both age and sex are strongly correlated with voting preferences (Wang, Rothschild, Goel, and Gelman, 2015, p.981).

Based on the Xbox data and the data of 2008 national exit poll, the authors perform a post-stratification re-weighting of the response. The weights of each cell are calculated by the proportion in the real electorate data (Wang, Rothschild, Goel, and Gelman, 2015, p.982). More specifically, the authors use exit poll data in the 2008 presidential election to calculate these weights (Wang, Rothschild, Goel, and Gelman, 2015, p.984). After the calculation, the authors make a comparison between the Xbox estimates of Obama support for each level of categorical variables on the day before the election with the actual voting behaviors of those same groups estimated by the 2012 national exit poll, we find that the Xbox estimates are remarkably accurate (Wang, Rothschild, Goel, and Gelman, 2015, p.985).

According to figure 2 and figure 3, Xbox estimates of the two-party Obama support during the 45 days leading up to 2012 presidential election, which suggest a landslide victory for Mitt Romney (Wang, Rothschild, Goel, and Gelman, 2015, p.981). While Pollster.com gives the uncertain prediction. Because we could find the two- party Obama support fluctuates around 50%. And Xbox post-stratified have predicted that Obama wins since the two-part Obama support was obviously above 50%.

Reference

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