Assignment #4

Sanittawan Nikki Tan

10/26/2018

Part 1

(a) please see the attached spreadsheet

(b) I called 200 numbers. 23 numbers (13 percent) are valid or active numbers. However, some of the numbers belong to companies or went straight to mailboxes. None of the people I called responded according to the *Response* variable. Six people who picked up my call did not respond according to the *Response* variable. My response rate is zero unfortunately.

(c) Because none of the people whom I spoke with answered any of the questions. My *Response* = 1 is equal to zero. I am not able to answer this question.

(d) I divided my list of phone numbers into two sets. I called number #1 to #133 on Wednesday, October 24, 2018, between 5:40 pm and 6:30 pm MDT. I called number #134 to #200 on Thursday, October 25, 2018, between 5:43 pm and 6:09 pm MDT. I intentionally controlled for time and decided to make phone calls approximately the same time on two days. I decided to make phone calls in the evening because I assume that people will leave work by that time. I also assume that if the owner of the phone number is a teenager, they should be back home from schools. None of the people I spoke to responded to the survey. Additionally, only 23 numbers are valid numbers and I controlled for the time I made phone calls. I am not sure if I would have gotten more responses had I made phone calls during the weekend. Nevertheless, I think that the time surveyors choose to make phone calls do have some effects on response rate. For example, if the phone number belongs to a person who is not older than 18 years old, the owner might be in school studying if the surveyor chooses to make phone calls in the afternoon. He/she might miss their chances of getting responses. In contrast, if the surveyors make phone calls on the weekends or in the evening, there might be a higher probability that the teenager is with his or her parents whom the surveyor can obtain responses from.

(e) Unfortunately, I do not have any respondents. I am not able to answer the question about the median age of my respondents. However, I would like to discuss the scenario where the median age of my respondents does not match with the state data. My area code (385) belongs to Salt Lake City, Utah which has a population of 186,440 according to the U.S. Census Bureau (unknown date). As the number of valid phone numbers on my list is 23, the sample from my survey represents less than one percent of the total population. I would argue that due to the small number of the sample size, it is very likely that the median age of my population will be different from the true population. Even if all my 200 numbers are valid, the sample is still less than one percent of the population. Thus, the median age of the respondents in my survey may still be different from the state data.

(f) Unfortunately, I am not able to answer questions about percentages of Trump and Clinton supporters. However, I would like to propose a way to test if the order in which surveyors say the candidate names or categories in the survey questions influences the results. I think we can test this hypothesis by randomly select individuals into how many groups we have for the voting choices. In this case, there will be four groups: Trump, Clinton, Other and No vote. Individuals in each group will hear the candidate names in a different order. If the results do not vary among groups, we may be able to conclude that the order does not affect the results. I think that the order of the candidate names should not substantially influence the results in the case of a major election like a presidential election and the fact that we are asking which candidate the individuals voted for in retrospect. I think orders of questions or candidate names should matter much more in the case of smaller or local election where candidates are not well-known or in the case of conducting surveys before an election.

Part 2

Forecasting elections have long been an interest of political scientists, the media and forecasters alike. As Wang, Rothschild, Goel, and Gelman mentioned in their 2015 paper "Forecasting Elections with Non-Representative Polls," modern polling methods rely heavily on representative sampling. However, it has become more difficult to obtain representative samples as the response rate is decreasing (Wang et al. 2015, 980). The authors proposed a statistical adjustment method to non-representative polls which are easier and less costly to obtain and showed that non-representative polls can be properly adjusted and achieved similar, or even better, results as traditional polls.

Wang et al. used Xbox data which is a survey of voting intention on the Xbox gaming platform. The survey collected data on eight variables, namely sex, race, age, education, state, party ID, political ideology, and which candidate the respondent voted for in the 2008 presidential election. It seems to me that the population of interest of the paper is American adults who are eligible to vote. Based on figure 1, the top three most representative variables are state, race, and whom the respondents voted for in the 2008 election. This is because the Xbox data showed the least difference from the data of 2012 exit poll which is more representative of the population of interest. By the same measure, the top three least representative variables are sex, age, and education. Sex and age are clearly vastly different from the 2012 exit poll. However, both education and party ID obtained from Xbox data are both different from the exit poll. Although I did not calculate the average differences of both variables to the population, figure 1 suggests that the magnitude of the difference of the education variable is higher than that of party ID. In addition, I think that the effect of undersampling college graduates on the election prediction result is higher than "other" group in party ID because college graduates tend to vote more and lean towards liberal candidates. If this is the case, the Xbox data will substantially undersample Obama supporters resulting in inaccurate prediction. Thus, although one can make a case that party ID is among the top three least representative sample, I am convinced by my analysis that education should be among the top three.

The reason why Xbox sample would be vastly different from the broader voting population is that the population of Xbox owners and players are fairly narrow; in other words, owners of Xbox tend to be young males who are not (yet) college graduates, i.e. currently in college or in high schools. In terms of gender, the Xbox data suggests that there are more male voters than female counterparts which are not true. According to the 2012 election data from Roper Center for Public Opinion Research (Roper Center unknown date), 47 percent of the voters in the 2012 presidential election is male while 53 percent is female. In terms of age, Xbox data have many younger respondents than older people. This also goes against the conventional understanding that old people tend to vote more than young people. The Xbox data oversampled the young who tend not to turn out to vote and undersampled the old. Roper Center (Roper Center unknown date) data also shows that 56 percent of all voters is 65 and over while only 37 percent is 18 to 29 years old. In terms of educational background, the Xbox data undersampled college graduates who are considered more highly educated than the rest of the sample. Conventional understanding and data from Roper Center show that college graduates, or people with higher education, voted more than other groups categorized by education level in the 2012 election.

After having gathered data from Xbox gaming platform, the authors leveraged demographic information of respondents from Xbox survey and the 2008 presidential election exit poll data to perform multilevel regression and poststratification (MRP) in order to adjust the data (Wang et al. 2015, 981-984). The authors used a variety of statistical techniques to arrive at the result shown in figure 3 and the degree of representativeness is shown in figure 5 (Wang et al. 2015, 984, 986). The adjustment really improved the two-party Obama support to be more in line with Pollster.com.

Without any adjustment to the Xbox data, the polls would have predicted that Mitt Romney would win in the last three weeks of the election. Pollster.com's forecast showed more uncertainty because the two-party Obama support trailed below 50 percent from October 22 to after October 29. However, Pollster.com result showed that Obama gained more support during the very last week of the election with more than 50 percent support. I tend to think that Pollster.com would have predicted that Obama would have won by a narrow margin and it was a close race. What is interesting is the post-stratified Xbox data shown in figure 3 which would have predicted that Obama would win the election with clearly over 50 percent support in the last three weeks of the election (Wang et al. 2015, 984). Based on the actual 2012 presidential election result, post-stratified Xbox data performed better than traditional polls by Pollster.com.

Ultimately, the result from this forecasting paper is illuminating because it presents a method that pollsters and academics can adopt if representative sampling is not possible or too costly.

Roper Center for Public Opinion Research. "How Groups Voted in 2012." Accessed October 26, 2018. https://ropercenter.cornell.edu/polls/us-elections/how-groups-voted/how-groups-voted-2012/.

United States Census Bureau. "Fact Finder." Accessed October 26, 2018. https://factfinder.census.gov/faces/nav/jsf/pages/community_facts.xhtml.