Problem Set #4

MACS 30000, Dr. Evans
Nan Ge

Question 1
(a) Submit your filled out version of the PhoneSurvey.xlsx spreadsheet.
Please see to attachment.

(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?
I called 200 numbers. Get 0 valid response. Out of 200 assignmed numbers, 146 were invalid, 12 were business number, 33 got transferred to voice mail. 9 people answered the phone, but none of them offered valid response. Response rate is 0.

(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?
Number of Response = 1 was zero. Therefore, 0% answered the voting question. 0% 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?
My area is 209. I called twice. First time at 10 A.M., second time at 3 P.M pacific time. During the first round, one person hang up the phone because "it's 3 in the morning". Perhaps the respondent was travelling. During the second round, two people refused to answer because they are working. Therefore, timing might have negatively affected the 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? What are some reason's why your sample median does or does not match the State data?
There is no valid age data collected (one person answered "None of your business", and one person answered below 18). Area code 209 primarily serves Stockton City in California.[1] According to the United States Census Bureau, the median age in Stockton is predicted to be 31.8 in 2016.[2] Total population is 301,443. I can't compare between the sample and the population.

(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

[1] https://en.wikipedia.org/wiki/Area_code_209

[2] 2012-2016 American Community Survey 5-Year Estimates: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF

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?

I have no valid answer for the voting question. In the 2016 election, 61.6% of California citizens voted for Clinton, 32.8% boted for Trump. In order to test if the order in which I say the candidates or categories has an influence on the result, I'll generate a larger phone number list of valid numbers. Call the same list twice, at the same time on the same weekday. Keep all the other questions the same, but shift the order in which I read the candidates. Compare the results from the phone numbers that answered in both surveys and see if any body changed their answer the second time. There is still possibility that both the answers are fraudulent, but if the order does have an influence, we should see a pattern of changing answer in the second phone survey.

Question 2

In the Xbox poll, respondents were asked to provide information about their sex, race, age, education, state, party ID, political ideology, and who they voted for in the 2008 presidential election.(Wang et al., 2015, p.981) As we can see from Figure 1, sex, age and education are the least representative variables compared to conclusive national survey data. Race, state and 2008 vote are the most representative three variables. In the Xbox poll, a dominantly 93% of the respondents are male, compared to less than 50% in the 2012 Exit poll. (Wang et al., 2015, p.981) Demographic distribution is also skewed. Over 60% of the Xbox respondents are aged between 18 and 29, whereas 45-65 age group takes up less than 10% of the sample. In terms of education, college graduates take only about 30% of the Xbox respondents, 20% less than in the Exit poll. The sample differences are not very surprising. They've revealed the well-known self-selection problem in a survey. Young males are more addicted to digital games and thus constitute the majority of Xbox customers. Also, the less educated are more likely to be struggling to find a full-time job, and as a result, have more time playing Xbox games.

As we have seen from sample distribution, Xbox poll is a very unconventional and unrepresentative survey, which would have been discarded by old-school pollsters. However, with the help of new statistical method, it could also yield satisfactory predictions. In Wang et al. (2015), the authors used multilevel regression and poststratification (MRP) to gauge the results. First, they divide the sample into multi-level cells by considering all combinations of the previously mentioned eight variables. Then they fit a multilevel regression model within each cell. Finally, they aggregate cell-level estimates to the state- and national- level for an overall prediction. The last step requires measuring the proportional weight of each cell to the whole electorate. (Wang et al., 2015, p.982-983) The authors used the Current Population Survey (CPS) and the exit poll data from the 2008 presidential election to compute cell weights. (Wang et al., 2015, p.984) Exit poll is conducted outside the vote stations on the day of election. Although less conclusive, it contains critical information such as party identification, which makes it a perfect complement to the CPS data.

Figure 5 shows that MRP adjusted Xbox data has a very close prediction to the 2012 exit poll data the day before the election.

In figure 2 we can see that in the last three weeks before the election, Xbox raw data would have predicted a fluctuating but ”landslide” victory for Romney. (Wang et al., 2015, p.982) At the same time, Pollster.com poll result was very uncertain, with Obama’s supporting rate fluctuating around 50%. On the day before the election, Xbox participants were more prone to vote for Romney, while Pollster.com believed Obama would win with a slight advantage (52%).(Wang et al., 2015, p.982) With post-stratification Xbox data, Obama’s votes is steadily above 280 in three weeks before the election. On the day before the election, the possibility of Obama winning the election is predicted to be 88%. The median estimation is 312 electoral votes for Obama, very close to the actual number of 332. (Wang et al., 2015, p.989) Prediction from the raw data is the most biased, because the census population is inconclusive and non-representative. However, the authors used poststratification method to adjust for bias in the sample population. The results prove that even non-representative census could yield very accurate predictions. In the future, non-representative polls from various new platforms could serve as aids or even substitutes to exhaustive national polls.

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.