Dhruval Bhatt
Assignment #4
October 31st, 2018

Part 1: Non-Probability Sampling Phone Survey

a) A filled spreadsheet with responses and additional notes is included in the submission folder.

b) I called 200 numbers, out of which, only 77 were operational, not a business or some automated voice message. Of the 77 valid numbers, 19 picked up the phone. About 10 people hung up as soon as they heard it was survey and 7 people said they were busy or not interested in taking the survey. 2 people responded to all survey questions and were classified as response =1.

The response rate based on all numbers dialed is:

$$\frac{2}{200} = 0.01 \; or \; 1\%$$

Excluding all the invalid numbers, the response rate can be calculated as:

$$\frac{2}{77} = 0.026 \; or \; 2.6\%$$

Of the 19 people that picked up the phone, a response rate is calculated as:

$$\frac{2}{19} = 0.105 \; or \; 10.5\%$$

c) Response = 1 is defined as a when a person picks up the phone, is over 18 and allows to ask the question about voting. Of the 19 people that picked up the phone, either they let me ask survey questions or not so none of them heard the survey question and then decided not to answer. With that in mind both the responses that counted as 1, answered all three questions.

d) I made the calls between 4 and 6.30pm central time on a weekday. The area code I was calling was New Jersey so it was between 5 and 7.30pm in eastern time zone. I think it was a decent time to call. People were not interested in taking the survey but were not annoyed or angry that someone gave them a call. The two people that responded to the survey were willing to do it as they seemed to be leaving work or just got home and didn't mind spending a minute.

As seen in the table below, it is interesting to note that more people picked up the phone between 5 to 6pm ET and both the responses were received were during that time. This might be because later in the evening people are busier with dinner or after work day activities and less likely to pick up a random number.

Table 1 Calls Picked Up Versus Time of the Evening

		Valid	**Picked Up**	
5 to 6 ET	**First 100**	46	14	30%
6 to 7 ET	**Second 100**	31	5	16%

The table below shows the data for the 2 responses to the survey received.

Table 2 Survey Responses Received

Phone number	Response	Republican, Democrat, Other, No Vote	Age	Notes
___-744-6604	1	No Vote	31	*Ineligible to vote - citizenship
___-744-1722	1	No Vote	24	*Ineligible to vote - citizenship

e) The median age is

$$\frac{31 + 24}{2} = 27.5$$

According to data available by US Census Bureau's Factfinder table, the median age in the state of New Jersey is 39.5. The median age of the responses I received is lower than the state's median age. One of the main reasons for this discrepancy would be that the sample is too small. Two responses cannot give an accurate representation of the actual population. It could be that younger people were more available to talk at that time versus the older demographic or the area being contacted was predominantly a younger demographic and the contact list didn't capture the broader population spectrum.

f) According to Politico's election result reporting, 55% of New Jersey voters, voted for Clinton, 41.8% voted for Trump and 2.9% voted for other. However, 0 % of my respondents voted. Both made a note that they were ineligible to vote as they are not US citizens. The responses I received were not representative of the broader New Jersey population. Based on the responses and the interaction on the calls made, it appears that the area code and numbers might correspond to a community of newly immigrated population and are ineligible to vote based on citizenship criterion. Therefore, the survey does not accurately project the actual election results. The phone survey would be enhanced by calling diverse areas of New Jersey and include more participants.

Since none of the respondents voted, it is difficult to quantitatively gauge what the effect of order of candidates or criteria would be. Considering this a survey of what has already occurred, I suppose, it should not have a major effect on the responses. However, for more open ended or longer responses a survey designer should be careful to mix up the order. People might have the tendency to pick the first one or last one consistently without putting actual thought into the options. One way to test if that is the case, would be to have use different order of responses for similar respondent subgroup types and measure if there is a difference.

Part 2: Predicting Election Surveys

The paper, "Forecasting Elections with Non-Representative Polls" by Wang, Rothchild, Goel and Gelman, presents some interesting work that is proposing the using data from non-representative population to make predictions for broader population, for instance, general election results. The paper highlights how data from Xbox users is used to forecast 2012 presidential elections. Based on the initial collection of demographic data of the Xbox users survey, it is evident that most of the users were younger men. As the authors note, "young men dominate the Xbox population: 18- to 29-year old comprise 65% of the Xbox dataset, compared to 19% in the exit poll; and men make up 93% of the Xbox sample but only 47% of the electorate" [1]. This shows Xbox users is not what typical surveys using representative techniques would look at, but the paper goes on to demonstrate how using mathematical techniques, using this set of data could be viable.

The researchers ask for eight variables from each Xbox respondent: Sex, Age, Race, Education, State, Party ID, Ideology and 2008 vote. Based on Figure 1 in the paper, the least representative is age, sex and education and the most representative is race, state and 2008 vote. Xbox games are appealing and heavily marketed to a subsection of the population demographics. Many of the games are catered to what is typically associated as masculine interests so it is not a surprise that more men tend to play them while the voting population is a balanced gender divide. The age is also a major point of difference as typically it is known that people tend to vote more regularly as they grow older as they may have more interest or awareness in the political issues. However, Xbox is a form of entertainment introduced relatively recently and enjoyed more by younger demographic – possibly due to time and resource availability. Finally, education is a correlated factor. With Xbox users being younger, they may not have completed graduation while more educated people feel the need to participate in democratic process, causing the point of difference.

Clearly there are some notable differences between Xbox users and voting population. The authors, "post stratify the raw Xbox responses to mimic a representative sample of likely voters" [1]. To do so, they need to use additional data is used. While typically Current Population Survey (CPS) is used, the authors use only 2008 presidential election exit poll data for post stratification. In addition, the authors of the paper recognize that there are differences between intention of voters and actual voting outcome. To adjust for this, the authors "collect historical data from three previous US presidential elections, in 2000, 2004, and 2008" to calibrate the data correctly [1].

Having applied the "Multilevel Regression and Post Stratification" (MRP), the authors were able to frame some predictions for the 2012 elections. The results are graphed to demonstrate the differences between Xbox raw data, data from pollster.com and Xbox post-stratified data. Figure 2 and 3 of the paper illustrate the predictions over time with the vote share gained by Barack Obama in a two-party vote as reference. The Xbox raw data, seen in figure 2, clearly does not match the reality. It is varying highly and only matches pollster.com's data in few instances. However, the scenario is quite different once the Xbox data is processed. In figure 3, the Xbox predictions follow the trends predicted by pollster.com for a few weeks before the elections. However, in the last three weeks before the election, Pollster.com predicts close to 50% vote share for Obama but Xbox post – stratified predicts higher vote share, around 52%. Post-stratified Xbox data matches the actual vote share more than even an aggregate of traditional polls.

These results help solidify the authors' viewpoint that processed, non- representative data can be as good, if not better for certain situations.

References

[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.