**Assignment 4**

Angela Zorro Medina

October 31, 2018

**Question #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.

**Answer #1**

I did my first round of calls between 7pm and 8pm central time, which is 6pm-7pm in Utah, on Monday 29. Because only few numbers were active, I called on Tuesday 30 in the morning (9AM central time) and again in the afternoon (2 pm central time) and still did not get any response.

After calling the 200 numbers, I did not get any response. Most of my numbers were not active numbers, and in the few phone numbers that I did not get a machine telling me that the number was inactive or disconnected, the person did not pick up the phone. I learned some interesting things though. First, there are different ways in which each company tells the caller that the number is not active. For example, Verizon allows the phone to ring several times, usually three or four times, and after that, a machine tells you that “the number you are calling is not active or it has been disconnected.” Other companies play an advertising message before telling the caller that the number has not been assigned.

Other interesting things that I noticed while calling is that most of the machine voices are female voices, although there were some male machines, the vast majority of the machines had female voices. Besides the fact that I did not get any response, I got three interesting experiences. First, I got a machine asking me in Spanish if I wanted to proceed in English to press 1. Second, I got a number that is a “text number” which means that the number is only used to received or send text messages and that I should text the person instead of calling. And finally, I got a machine asking my name before connecting me to the person; I replied my name and the person did not pick up the phone anyway.

In summary, I did not get any response (all my observations are 0), I did not have the opportunity to even talk to someone (most of my numbers were disconnected, and for the few working numbers the person did not pick up the phone). In light of this, I have no information to answer the questions of this point of the assignment.

**Question #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 the 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?

**Answer #2**

One of the major problems of survey data is the representativity sampling. Researchers struggle to find a sample that is a “miniature” version of the entire target population. Sampling representativity is necessary to guarantee that our estimators are not biased (Kruskal and Mosteller 1979), and therefore that our results are providing useful information about the target population. While in the past, our only option to guarantee representativeness was to make sure that our sampling method was good enough to capture a “miniature” version of the large population, Wang et al. (2015) show that non-representative samples can be accurate after proper statistical adjustments. In their paper “Forecasting elections with non-representative poll,” Wang et al. present an interest proposal to adjust a non-representative presidential poll conducted in 2012 through an Xbox platform. Using the Xbox data and their adjusting method, authors attempt to demonstrate that it is possible to have high predictive surveys even without having a representative sample of the target population by showing how precise their “corrected” Xbox sample predicted the 2012 presidential election outcomes.

In the first section of the mentioned paper, after the introduction, Wand et al. describe the Xbox data and compare the demographic information of the individuals in their sample with the demographic characteristics of the electoral population that voted in 2012 presidential elections. According to the information presented by the authors in Figure 1. we observed some similarities by comparing both populations. Specifically, it is possible to observe that the Xbox population is very similar to the target population regarding race, state, and 2008 vote information. Additionally, although not as similar as these three variables, the ideology variable presents significant similarities too.[[1]](#footnote-1)

On the other hand, the variables in which both populations diverge the most are sex, age, and party identification[[2]](#footnote-2). It is not surprising to see large differences among these variables. The age difference can be explained by the availability of video games during the childhood of the different generations. For example, a person that was 45 years old during the 2012 presidential elections was 34 years old when Xbox was launched, while a person that was 29 by 2012 was 18 when Xbox was launched. The gender difference between both populations can be explained by the fact that men play more video games than women (Lucas and Sherry 2004).[[3]](#footnote-3) And finally, the party identification could be strongly correlated to the age variable because younger generations tend to have a weaker level of partisanship (Jennings and Markus 1984).

After describing the data, Wang et al. (2015) explain their statistical techniques to adjust the estimations done using their non-representative Xbox data. Their method is divided into two parts. In the first part, the authors divided their data into demographic cells by the target population’s demographic composition and estimate voter intent at the cell level using multilevel regression. In the second part, Wang et al. (2015) aggregate the estimations obtained from the multilevel regression analysis at the cell level to produce a final estimation of the voter intent for the 2012 presidential election. The aggregation is a weighted sum of every cell-estimate by the proportion of the electorate in each cell by state and national levels. To do the weighted aggregations, Wang et al. (2015) use data from the Current Population Survey and the exit poll data from the 2008 presidential election.[[4]](#footnote-4)

By looking at Figure 2 of the Wang et al. (2015) paper, we observe the differences between the predictions based on the non-representative Xbox data without statistical adjustment, and the 2012 electoral information published by Pollster.com. By comparing these two results, it is obvious that the non-adjusted Xbox data made an inaccurate prediction of the 2012 U.S. presidential election by foreseeing that Romney would win the elections, when in fact Obama won those presidential elections. Moreover, Pollster.com data shows a more accurate prediction by foreseeing that Obama would win the presidential elections in 2012.[[5]](#footnote-5)

Despite the inaccurate results of the Xbox data before the statistical adjustment, Wang et al. (2015) show that after using their proposed statistical correction their Xbox data is capable of predicting the results of the 2012 U.S. presidential election. These results are striking taking into consideration that both age and sex are some of the variables that the literature has identified as strongly correlated variables with voting preferences (Kaufmann and Petrocik 1999) and those variables have large differences between the Xbox sampled population and the target population.

The results presented by Wang et al. (2015) offer a new alternative to overcome the sampling representativeness challenge that researchers in different fields faced. By showing that even a highly unrepresentative sample can produce accurate estimators, Wang et al. open the door to exploit different survey data even when representativity of the sample cannot be achieved, or the cost of generating a representative sample is so high that deters researcher to do it. In summary, by using the new technology and advanced statistical techniques, social researchers have now more tools to produce insightful empirical studies around a variety of social phenomenon that could not be studied before or to exploit new sources of data that do not have the representativity requirements that we needed in the past.

**References**

Griffiths, M. D. (1991). Amusement machine playing in childhood and adolescence: A comparative analysis of video games and fruit machines. Journal of adolescence, 14(1), 53.

Jennings, M. K., & Markus, G. B. (1984). Partisan orientations over the long haul: Results from the three-wave political socialization panel study. American Political Science Review, 78(4), 1000-1018.

Kaplan, S. J. (1983). The image of amusement arcades and differences in male and female video game playing. The Journal of Popular Culture, 17(1), 93-98.

Kaufmann, K., & Petrocik, J. (1999). The Changing Politics of American Men: Understanding the Sources of the Gender Gap. American Journal of Political Science, 43(3), 864-887. doi:10.2307/2991838

Kruskal, W., & Mosteller, F. (1979). Representative Sampling, III: The Current Statistical Literature. International Statistical Review / Revue Internationale De Statistique, 47(3), 245-265. doi:10.2307/1402647

Lucas, K., & Sherry, J. L. (2004). Sex Differences in Video Game Play:: A Communication-Based Explanation. Communication Research, 31(5), 499-523.  <https://doi.org/10.1177/0093650204267930>

Phillips, C. A., Rolls, S., Rouse, A., & Griffiths, M. D. (1995). Home video game playing in schoolchildren: A study of incidence and patterns of play. Journal of Adolescence, 18(6), 687-691.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. International Journal of Forecasting, 31(3), 980-991.

Wright, J. C., Huston, A. C., Vandewater, E. A., Bickham, D. S., Scantlin, R. M., Kotler, J. A., ... & Finkelstein, J. (2001). American children's use of electronic media in 1997: A national survey. Journal of Applied Developmental Psychology, 22(1), 31-47.

1. Wang et al. (2015) conducted a survey through Xbox platform during 45 days before the 2012 U.S. presidential election. Besides asking respondents about their intention to vote that day, the researchers also asked different demographic characteristics (race, sex, age, education level, state, party identification, political ideology and who they voted in 2008 U.S. presidential election). At the end, they performed 750,148 interviews, with 345,858 unique respondents. [↑](#footnote-ref-1)
2. While the differences are not so dramatic as in the three variables mentioned, education also is a variable in which both populations have significant differences, although similar trends for those persons who have not graduated from college. [↑](#footnote-ref-2)
3. Lucas and Sherry (2004) offer an interesting explanation of why women and men have different tendencies to play videogame. Some of the reasons presented by the authors are related with the videogames as “domain of the boys”, and the design of the videogames for male audiences. In their paper they also cited a vast literature that supports the fact that videogames are more likely to be played by males than females (Griffiths 1991; Kaplan 1983; Rolls et al. 1995; Wright et al. 2001) [↑](#footnote-ref-3)
4. The researchers explain that the Current Population Survey does not have information for some key poststratification variables, like party identification, so they used the data of the exit poll from the 2008 presidential election to include those important variables. [↑](#footnote-ref-4)
5. These analyses are based on the data taken during the 45 days leading up to the 2012 U.S. presidential election. [↑](#footnote-ref-5)