Problem Set #4

MACS 30000, Dr. Evans
Keertana V. Chidambaram
(UCID: 12211266)

Problem 1: Non-probability sampling phone survey

(a). The filled out version of the `PhoneSurvey.xlsx` file has been pushed. Unanswered questions (because of various reasons including but not limited to direction of calls to voicemail, no pick-up, hanging-up mid conversation, and refusal to provide answers) have been filled with 'N.A.' or Not Applicable.

(b). All the 200 random numbers provided were called. 5 respondents have *Response* = 1. 195 respondents have *Response* = 0. *Response Rate* = 5/200 * 100 = 2%.

(c). Only 1 respondent with *Response* = 1 answered the voting question and the age question. Thus, for respondents with *Response* = 1, fraction answering voting question = fraction answering age question = 1/5 = 20%.

(d). The calls were made during the time period between 6 pm to 9 pm, Chicago time. The area code provided was for Connecticut and the corresponding Connecticut timings were between 7 pm to 10 pm on a Friday night. Although $\sim$ 40 numbers from the spreadsheet were valid, only 5 picked the call. Probably more number of responses could have been received on a weekday evening. It can be said with certainty that the time of the call impacts the response rate, but this impact would vary from group to group. For example, if the target is a working population, then they would probably be more recheable in the evenings as compared to the afternoons, but if the target population is a set of housewives, they might be more available during afternoons instead.

(e). The median age of respondents is 38 and that of Connecticut is 41[1]. Although the sample median is close to the state median, only 1 person in a state with population $\sim$ 3.6 mil[2] was sampled. Hence the small sample can not be treated as a representation of the area's population even though median ages are close.

(f). 0% respondents voted for Trump and 0% voted for Clinton. The actual percentage of votes in Connecticut is 41.2% for Trump and 54.5% for Clinton[3]. Two methods to test if the survey design influences respondent results are:

(i) The first is to split the respondents into random groups and to subject each group to a different administration of the same survey. The variation between groups can be used to identify the effect of survey design. Possible source of

[1] https://datausa.io/profile/geo/connecticut/
[2] https://www.census.gov/quickfacts/ct
[3] https://www.politico.com/mapdata-2016/2016-election/results/map/president/

error: when the group assignment is not random or representative, it inherently contributes to differences in response.

(ii) The second method is to survey the same set of people at different times using different survey methods. The variation between individual answers over time can be attributed to the tweaks in the survey design. Possible source of error: when individuals change opinions over time, it creates a variation in result which can not be attributed to the change in survey design.

Problem 2: Predicting elections survey

The paper "Forecasting Elections with Non-Representative Polls"makes a case for using non-representative polls in forecasting election outcomes. Data is collected from daily polls administered via the Xbox gaming platform over 45 days leading to the 2012 US Presidential Elections. Eight variables: *Race*, *State*, *Sex*, *Age*, *Education*, *Party ID*, *Ideology*, and *2008 Vote* are collected. While Xbox data is representative of the electorate with respect to a few variable, it is evidently off for the others.

From visual analysis of Fig. 1. of the paper, it is clear that *Race* and *State* are the top two and *Sex* and *Age* the bottom two representative variables. But between *Education*, *Party ID*, *Ideology*, and *2008 Vote*, clear representativeness ranking based on visual inspection of Fig. 1. alone is not possible. Therefore, we proceed to analytically estimate the representativeness by calculating the RMS error of the Xbox data w.r.t. the Exit Poll data (approximate values of each variable is estimated from the graph). `Table 1` gives the RMS error for each of the variables:

Table 1: RMS Error of Xbox Vote Share for Obama w.r.t Exit Poll Data

Variable	Categories	Xbox%	Exit Poll%	RMS Error
Education	Did Not Graduate HS	9.8	3.5	10.5
	School Graduate	24.0	20.0	
	Attended Some College	36.5	28.8	
	College Graduate	29.0	47.0	
Party ID	Democrat	32.3	37.3	4.8
	Republican	30.8	32.5	
	Other	35.8	29.3	
Ideology	Liberal	28.3	24.3	3.5
	Moderate	35.8	40.3	
	Conservative	34.8	34.5	
2008 Vote	Obama	56.3	52.0	5.1
	McCain	38.0	45.0	
	Other	5.0	1.5	

* Note: only variables with indecisive ranking order are included in the analysis

Thus the top 3 most representative variables are: *Race*, *State*, and *Ideology*, the

bottom 3 are: *Sex*, *Age*, and *Education*. The reason for this offset is because the sampling is biased. The demographic of people who engage in Xbox games are disproportionately dominated by young males. This explains the variation observed in the *Sex* and *Age* variables. Quoting from a study conducted by the PEW Research Center[4]:

> "A person′s education level is another predictor of video game play. Some 57% of respondents with at least some college education play games, significantly more than high school graduates (51%) and those who have less than a high school education (40%)."

Thus there is an observable positive correlation between the likelihood of engaging in video games and education levels. This implies that Xbox players also may not accurately represent the education levels of the population, causing the drift in prediction. In order to account for this non-representativeness of the Xbox data, suitable weights were administered. The authors consider two possible datasets to estmiate these weights: Current Population Survey (CPS) and historical Exit Poll data (years: 2008, 2004, and 2000). However, only the latter dataset has been used in this study.

To illustrate how non-representative data without proper adjustments can significantly bias results, the election outcome was predicted by using both the raw and adjusted Xbox data and it was compared with the aggregate Exit Poll data obtained from *www.pollster.com*. Considering a window containing the final three weeks before the elections i.e. from 16$^{\text{th}}$ Oct (day of the 2$^{\text{nd}}$ Presidential Debate) to 5$^{\text{th}}$ Nov (day before the elections), the three datasets exhibit three different prediction trends in the days leading up to the Presidential Elections. Pollster predictions hover around 50%, initially predicting a close vistory for Obama, then a close victory for Romney before finally predicting a marginal victory for Obama. The raw Xbox data throughout predicts a very comfortable victory for Romney (with the exception of 25$^{\text{th}}$ and 26$^{\text{th}}$ Oct where Romney wins only by a very small margin). But the adjusted Xbox forecasting has its 95% confidence interval always above the 50% line, consistently predicting Obama to win. If we look at the closing vote share data, the adjusted Xbox ($\sim$52.5%) outperforms both Pollster ($\sim$50.5%) and raw Xbox($\sim$46%) estimates in predicting the actual result (52%).

Thus the paper makes a very persuasive argument for the substitute of suitable weighed non-representative polls in the place of representative polls in forecasting election outcomes. Backed by compelling empirical evidence, this work of research shows the potential to shape the polling mechanisms of the future.

[4] http://www.pewinternet.org/2008/12/07/adults-and-video-games/