Assignment 4

Jiaxu Han

1. Non-probability sampling phone survey

(b) I called 200 numbers. 0 people responded. 200 people did not respond. The response rate is 0.

(c) Since 0 people responded, 0 fraction of the respondents answered the voting question and 0 fraction of the respondents answered the age question.

(d) One time I called at 9.30 am - 10.00 am (Los Angeles local time), another time I called at around 4 pm - 4.45 pm (Los Angeles local time). Since I got 0 response rate, at what time that I called did not make a difference in response rate in my case. However, I assume that call time would play an important role in response rate. If I have more valid phone numbers, calling at day time such as from 9 am to 4 pm would get higher response rate because most people are awake during this time than calling during the night such as from 12 am to 4 am when people are sleeping.

(e) Since I have 0 respondents, the median age of my respondents is 0 as well. The median age for California is 36 (https://factfinder.census.gov/faces/nav/jsf/pages/community_facts.xhtml) and that is very different from my sample median. The reason for that difference is that I got 0 response rate. Even if I have several more respondents, it is likely that the sample median would still deviate from state median age, because we have unrepresentative sample of the state of California.

(f) 0 percent of my respondents voted Republican (Trump) comparing to 62.9% of voters voted for Trump in the 2016 U.S. Presidential election. 0 percent of my respondents voted Democrat (Clinton) comparing to 34.6% of voters voted for Clinton in the 2016 U.S. Presidential election (https://www.politico.com/mapdata-2016/2016-election/results/map/president/).

To test if the order of the candidate presented would influence the results, I would randomly assign a number to one of the two conditions (either Trump first or Clinton first). After getting enough responses (say > 100) for each condition, we can do a t test to see if the ratio of voting Clinton in Clinton first Condition is significantly higher than the ratio of voting Clinton in Trump first condition, or vise versa. However, in order to avoid the effect of order of presentation on response, I would simply ask “who did you vote in 2016 election”.

Similarly, to test if categories in the survey question would influence the results, I would randomly assign a number to one of the two conditions: ask age first and ask age last. After getting enough responses (i.e. people answered the phone and agreed to do the survey), then I would calculate the response rate of the question “who did you vote in 2016 election?” . We can do a t test to see if the response rate of the question of who you

vote is significantly higher in "asking age last" condition than the response rate of that question in "asking age first" condition.

2. Predicting elections survey, Wang, Rothschild, Goel, and Gelman (2015).

Using unrepresentative sample for election forecasts may lead to very biased if not wrong predictions. However, in this paper, the authors used sample of Xbox players to demonstrate that, with proper statistical adjustment, "unrepresentative sample can be used to generate accurate election forecasts" (Wang, Rothschild, Goel & Gelman, 2015, p982).

According to Fig 1 (Wang, Rothschild, Goel & Gelman, 2015, p 982), age, gender and education level are the three least non-representative variables. On the other hand, Fig 1 also shows that race, state, and who they voted for in 2008 are the most representative variables.

It is not a surprise that age, gender and education level among Xbox sample would be very different from the broader voting population. Considering the consumer makeup of Xbox, we would expect that young males are dominating this user population. Indeed, as pointed out in the paper, "18- to 29-year-olds 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" (Wang, Rothschild, Goel and Gelman, 2015, p981). The reason for discrepancy in education level among Xbox sample and general voting population was not explicitly stated in the paper. We would assume that due to the age makeup of the Xbox user population, Xbox sample has a higher proportion of people who are still in high school or college comparing to the general voting population. As a result, the proportion of college graduates in Xbox are significantly lower than the general voting population.

Since the Xbox sample is biased, in the paper, the authors decided to "*poststrafify* the raw Xbox data to mimic a representative sample of likely voters" (Wang, Rothschild, Goel & Gelman, p 982). To achieve that, the authors used two data sources: the Xbox data including the demographic information and exit poll data from the 2008 presidential election.

After re-weighting the data and applying MRP, the prediction from Xbox sample has changed. From Fig 2 (Wang, Rothschild, Goel & Gelman, 2015, p 982) we can see that the last three weeks of Xbox raw data (the red line) indicating Two-party Obama Support is below 50%. Therefore the Xbox raw data is predicting Romney would win in 2012 election. Pollster.com forecast data (blue dashed line) though, in both Fig 2 and Fig 3, shows that Two-party Obama Support from the last three weeks of election is floating around 50%. Therefore, the prediction would be uncertain based on pollster.com forecast data. However, after post-stratifify Xbox raw data, the prediction is very close to the real result of 2012 election. In Fig 3 (Wang, Rothschild, Goel and Gelman, 2015, P984), the red line indicating MRP Ajusted Xbox Estimates and black dashed line indicating the actual two-party Obama vote share (Wang, Rothschild,

Goel and Gelman, 2015, P984) are almost on top of each other. Therefore, Xbox post-stratified not only predicted that Obama would win in 2012, but also had a very close prediction of vote share to the real results of 2012 election.

Reference

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