Assignment 4

Sanja Miklin

10/31/2018

1. Non-probability sampling phone survey

(a) Submit your filled out version of the PhoneSurvey.xlsx spreadsheet.

Find the filled out spreadsheet in the main folder.

For more detailed information about my phone-endeavour, you can see 'ValidatedPhones.xlsx' spreadsheet in the 'phone numbers' folder. This spreadsheet contains the following:

- information about the phone numbers I got through Numverify.com website, using an R script (also in the 'phone numbers' folder). I was inspired to do this by Li Liu, writing in this thread on GitHub.
- notes about which numbers were out of service, which reached a fax or an automated service, or which individuals answered the phone but declined participation. I thought all of this actually was interesting information when thinking about phone surveys and response rates.

(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 all 200 numbers, and according to my Response variable, 4 individuals responded, and 196 did not respond. That makes my response rate $\frac{4}{200} = 2\%$.

However, of the 200 numbers, 112 were not in service, 7 could not connect, and 9 were either fax numbers or called to an automated service. That means that 128 numbers were basically invalid. Of the remaining 72 numbers, 64 calls went unanswered, 8 individuals answered the phone and four of those were willing to participate in the survey. Therefore, if taking into account only valid numbers, my response rate is a bit higher, at $\frac{4}{72} = 5.56\%$, and 50% of individuals who answered the phone were willing to participate.

(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?

All the individuals who were willing to participate in the survey answered both questions.

(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?

I called most numbers in the morning (10AM-12PM) local time on Monday, and called the rest in the afternoon (4-5PM) local time. Calling during work hours meant that any residential land-line phones would likely go unanswered, while the business line would be answered. Many individuals might also be less likely to answer their cellphones while working, especially when receiving a call from an unknown number.

(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?

The median age of my respondents is 59. According to Factfinder, the average age in Maryland, which is the state my phone numbers were in, is 38.

It is likely that my sample is much older because older individuals are more likely to answer their phone (especially during the workday) and participate in a survey. Additionally, this survey in its very design would produce an older sample, as it excludes individuals under 18, though this did not emerge as an issue in my data.

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 voting percentages from the 2016 election? How might you test if the order in which you say the candidates or categories in the survey question influences the results?

25% (N=1) of my respondents voted Republican, while 75%(N=3) voted Democrat. According to Politico, 60.5% of voters voted Democrat, and 35.3% voted Republican, so my sample is not significantly different—35.5% of an N=4 sample is 1.42.

To test if the order in which I say the candidates/categories influences the result, I would randomly assign a number to a condition (e.g. Republican first vs. Democrat first) and compare the results. Of course, I would want my number of responses to be significantly greater than $ in order to make any conclusions about the effects of the option ordering on individual's answers.

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

In their paper, Wang et al. (2015) use a large data-set of opt-in poll responses from an Xbox platform conducted in the 45 days preceding to the US Presidential elections in 2012, to argue even non-representative data can be used to predict election results.

The demographics of the Xbox gaming platform users is, in fact, fairly similar to the US population as a whole with respect to variables such as race, 2008 vote, and state (distinguishing between 'battleground states,' 'quasi-battleground states', 'solid Romney' and 'solid Obama'). However, with regard to other variables, such as age, sex and education, the sample is quite unrepresentative. The Xbox population is much more likely to be male (93% compared to 47%), age 18-29 (65% compared to 19%) and to be a college graduate (around 50% compared to around 30%). This is nor surprising, as Xbox gamers tend to be young males, that have the dispensable time and income that allows them to play Xbox and participate on the gaming platform, which might be more likely among college graduates.

In order to use this non-representative sample to forecast election results, the authors performed a post-stratification re-weighing of the respondents. That is, they partitioned their data-set into 176,256 distinct cells that contained unique combinations of sex, race, age, education, state, party ID, ideology and 2008 vote, estimate votes for each cell and aggregate the data. To compute cell weights, the authors use a different data-set—the cross-tabulated population data from the 2008 presidential election exit poll. Additionally, the authors also use historical data from 2000, 2004 and 2008 to map their estimated voted intent onto actual election outcomes.

Ultimately, during the three weeks before the election, the raw Xbox data would have predicted that Romney would win the presidential election. In contrast, during that time frame, the Pollster.com data seem to

indicate the election really could go either way, as the estimates are very close to 50% throughout, marginally in Romney s favour some 10 days before the election, and then marginally in Obama's favour the few days before the election. (Figure 1.) In contrast, the weighed Xbox data would have predicted that Obama would win (Figure 3.)—with the estimate of intent on the day of election being 0.6% off from the actual outcome.

The authors, therefore, seem to succeed in showing that a non-representative sample, if appropriately adjusted, can provide fairly accurate predictions, that can even outperform polls constructed for the exact purpose of predicting election outcomes.