MACS30000 Assignment 4

Dr. Richard Evans
Submitted by: Junho Choi
Due October 31, 2018 (11:30 AM)

Problem 1

(a) For this question, I refer to the `.xlsx` file uploaded together with this `.pdf` file. Note that I have renamed the file to be `PhoneSurvey_junhoc.xlsx` for identification.

(b) For this assignment, I called all 200 numbers given by the initial `.xlsx` file, together with the area code I was assigned to (206, which corresponds to the Seattle, Washington area). According to how the *Response* variable is defined, no one responded; that is, I have had 0 people responding, 200 people not responding, and therefore my response rate is 0%.

(c) Due to the fact that I did not have any responses, I am unable to answer this question.

(d) Before going further, I note that times of day indicated in this question are all in Pacific Daylight Time (PDT), which is the time zone that Seattle, Washington belongs to. I have called 71 numbers in the morning (approximately at around 10AM to 12PM), 98 numbers in the afternoon (around 12PM to 5PM), and 31 numbers in the evening (around 7PM to 9PM). Because my response rate for this assignment is 0%, I cannot say for certain whether a specific time of day has affected my response rate.

However, using a chi-square test of independence, I was able to find out that the rate of people not answering the phone is much likely to be independent of time of day a phone call was made.[1] I was able to identify that there are four major types of nonresponse: 1) a number is invalid, 2) a number requires additional extension, 3) not answering the phone, and 4) declining to participate in the survey. Because the first two cases are those in which human beings cannot answer at all, let us exclude these from the test. Using the remaining two types, I was able to make a two-way table of times of day and nonresponse types (Table 1). The chi-square statistic is approximately 2.3783, with the degrees of freedom being 2 ($= (2-1) \times (3-1)$). This yields a p-value of approximately 0.304475, which is not statistically significant even at $\alpha = 0.1$.

[1] I note that in order to make this test more rigorous, I would have needed to taken measures such as equal spacing of times of day phone calls were made, and so forth.

Table 1: Two-way Table for Times of Day and Nonresponse Type

	10AM−12PM	12PM−5PM	7PM−9PM	Total
Not answering	16	23	9	48
	(0.333)	(0.479)	(0.188)	
Declining survey	9	9	1	19
	(0.474)	(0.474)	(0.053)	
Total	25	32	10	67

NOTE: Column variables indicate the times of day survey calls were made at; row variables indicate the types of nonresponse (excluding cases for invalid numbers and numbers requiring additional extension). Row fraction in parentheses.

(e) Because my response rate was 0%, I am not able to calculate the median age of my respondents. Furthermore, I cannot give reasons to why my sample median does (or does not) match the state-level data, which would correspond to the data of Washington state. However, I was able to find out that the median age in the state of Washington is 37.6 (U.S. Census Bureau n.d.).

(f) Once again, because my response rate was 0%, I am not able to calculate the percentages of respondents who voted Republican or Democrat. However, the actual voting results in the state of Washington are well-known. Out of the 2,954,903 total votes, 1,610,524 (54.5%) voted Democrat, 1,129,120 (38.2%) voted Republican, and 215,259 (7.28%) voted for other parties and candidates (Politico 2016).

Regarding the case in which the order of categories affecting survey results, I would tackle this issue by attempting to mix up the order of categories presented to the survey participants. For instance, in the case of this problem, the second question has four options: Democrat, Republican, Other, and Did Not Vote. Therefore, there can be 24 (= 4!) combinations in which the order of options could be presented. By cycling through the combinations one-by-one for each respondent, I will attempt to minimize the impact from the order of categories presented.

Problem 2

In their 2015 paper from the International Journal of Forecasting, the authors Wang, Rothschild, Goel and Gelman (abbreviated as Wang et al.) showcase how the method of multilevel regression poststratification can be utilized to make valid forecasts or predictions even with a non-representative sample. Wang et al. begin their discussion by elaborating two main reasons why it may be advantageous for one to use a non-representative sample over a representative one. The first is due to the increasing tendencies of non-response rates for random digit dialing, a popular representative-sampling method; the other is due to the ease and cost-effectiveness of gathering non-representative samples through modern technology (Wang et al. 2015, p. 982).

Of course, the authors fully realize the potential dangers of using a non-representative sample. Therefore, they suggest implementing poststratification, which is described as a "method for correcting for known differences between sample and target populations" (Wang et al. 2015, p. 983). Basically, the method can be described as applying appropriate weights from the target population (or a representative sample) onto the (non-representative) sample so that the features shown in the target can roughly be exhibited. However, the authors also point out a potential dilemma from only relying on poststratification. On the one hand, while estimates after poststratification has been applied to are assumed to be randomly sampled, this "is only reasonable when the partition is sufficiently fine" (Wang et al. 2015, p. 984). On the other hand, if the said partitions are "too fine," there can be large gaps between estimates. A way to cope with this problem is to combine poststratification with a regularized regression technique called "multilevel regression," in which the entire process is named multilevel regression poststratification (MRP) (Wang et al. 2015, p. 984).

To illustrate the usefulness of the MRP strategy, the authors use a large set of non-representative data collected from Xbox users collected 45 days before to the day of U.S. presidential election (Wang et al. 2015, p. 983).[2] The authors quickly point out that the two least representative variables from the Xbox data are age and sex, where the ratio of males (compared to females) and ratio of the young (compared to the older) are much higher than in what the population would look like (Wang et al. 2015, p. 983). Studying Figure 1 of the paper, one may further argue that education is another least representative variable, as the college graduates are extremely underrepresented. On the other hand, it can be observed from the same figure that race, state (categorized by "battleground," "quasi-battleground," "solid Obama," and "solid Romney" states), and 2008 vote variables are relatively well-represented (Wang et al. 2015, p. 984). The reason why the least representative variables behave differently from the voting population is most likely due to the fact that most users of Xbox are male and are young adults (Corden 2017). In addition, for education levels, it may be that those who graduated from college have less time to play Xbox due to reasons such as work and post-graduate education.

In applying the poststratification part of the MRP strategy, the authors use 2008 presidential election exit poll data to re-weight the non-representative Xbox data (Wang et al. 2015, p. 984). In other words, the two data sources used for poststratification are the Xbox data as well as the 2008 presidential election exit poll data. They further mention that they have considered using other datasets, such as the Current Population Survey (CPS). The CPS was not used due to "miss[ing] some key postratification variables, such as party identification" (Wang et al. 2015, p. 984). In addition, the authors point out a caveat in using 2008 exit poll data for re-weighting – that there may have been changes in the demographics between 2008 and 2012 –, while arguing for the limit of scope to the said data "for the sake of simplicity and transparency" (Wang et al. 2015, p. 984).

The prediction made with the re-weighted (using MRP) Xbox data is not perfect;

[2]For convenience's sake, this dataset will be referred to as "Xbox data."

for instance, as seen from Figure 3 of the paper, it predicts that the two-party Obama support will be approximately 52% during the last three weeks of the election (since around October 16th), whereas Pollster.com had predicted that it would approximately be around 50% (Wang et al. 2015, p. 985). In other words, in terms of Obama winning the election, the prediction using poststratified Xbox data would have been a bit stronger than that by Pollster.com However, this is a surprising result considering the fact that, prior to adjustments using MRP, the raw data was quite not representative of the population. Indeed, as observable from Figure 2 of the paper, predictions using only the unweighted, raw Xbox data would state that Obama's support is below 50% – at around 46-47% on average – during the last three weeks of the election, which leans towards Romney winning the election (Wang et al. 2015, p. 983).

In summary, Wang et al. (2015) present convincing arguments and evidence for using a non-representative sample combined with the technique of multilevel regression poststratification. Having learned of the MRP technique, it is my personal hope to be able to utilize it in future research endeavors for the sake of validity and academic rigor.

References

Corden, Jez. 2017. "Internal Microsoft Research Provides Insight on Xbox One Owners (Exclusive)." *Windows Central*. https://www.windowscentral.com/heres-some-interesting-stats-about-xbox-one-owners-microsoft-shared-partners.

Politico. 2016. "2016 Presidential Election Results." https://www.politico.com/mapdata-2016/2016-election/results/map/president/

U.S. Census Bureau. n.d. "2012-2016 American Community Survey 5-Year Estimates." https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. "Forecasting Elections with Non-representative Polls." *International Journal of Forecasting* 31 (3): 980-981.