Identify and Adjust for Non-response Bias

benjamin chu

3/24/2017

Contents

Abstract 1
Introduction 1
Method and Data Description 2
Identifying Non-response Bias through Callback 2
Adjusting for Unit Non-response with Weights 4
Discussion 7
References 7

Abstract

In the field of surveying, there are many causes for one to mistakenly collect a bad dataset and obtain fruitless or even misleading results. This project uses the 1991 Race and Politics Survey to illustrate a common type of unintentional bias called non-response bias. In particular, we identify non-response bias by analyzing the difference between initial and late respondents. Then using the "second phase surveying" data as described in methods section, we show how to adjust for non-response bias by designing appropriate weights. Finally, we compare the adjusted data set with the non-adjusted data and discuss their difference. This project is carried out in an effort to supplement our studies in biomath 204 at UCLA, since we covered numerous techniques in data analysis but had little discussion in surveying techniques.

Introduction

In biomath 204, we focused on data analysis techniques to draw useful information from given datasets. Another important aspect of data analysis is surveying. If we are unaware of bad surveying techniques and potential biases, we could work with a terrible dataset and get nothing (junk in junk out) at best. At worst, we could give wrong recommendations even though our analysis is completely correct. Among the pool of surveying techniques and biases, I focused on how to identify and adjust for **non-response bias** in the context of a **random sampling** survey, both of which are standard and compelling issues today.

Random Sampling

A subset of individuals (sample) chosen from a larger set (population) to be surveyed. This is generally the strategy chosen unless it is not feasible.

Non-response Bias (unit)

Error due to a subset of the chosen sample not responding to the survey. This becomes a problem when a significant population have reasons to avoid responding to a survey.

- When present, **no amount of data can negate its effect**.
- e.g. In the 1936 U.S. Presidential Election, 10 million surveys were sent out and 2.3 million were returned, predicting Alf Landon would win with 370/521 electoral votes. He got 8.

Method and Data Description

From Survey Documentation and Analysis (SDA) archive, I obtained the 1991 Race and Politics Survey results. This is a telephone survey containing 178 questions, which collected a total of 2223 respondents with an impressive 65.3% response rate. This dataset is particularly suited for non-response bias analysis for two reasons. First, they perform up to 4 callbacks, which enabled me to employ a standard technique to identify non-response bias known as callback. Secondly, participants were sent additional survey questionnaires through mail after they completed the phone interview, in which the researchers received another 1198 "second round" responses. Because the first round of telephone survey already collected certain background information about the respondents, we can view this second round of sampling as a separate survey, and develop weights to adjust for non-response bias using demographic data such as age and race collected from the first round.

Identifying Non-response Bias through Callback

A common technique to identify non-response bias is by using callbacks. In the selected sample, people who did not participate in the survey (e.g. refusal, not at home…etc) may be randomly asked again to participate some time later. Those who agreed to take the survey only after several attempts are called late respondents. In survey methodology, this process is called **callback** and it is an excellent way to quickly identify non-response bias. We must assume:

- These late respondents are similar to the non-respondents
- The difference between early and late respondents is captured in the metric we used to measure them.

To begin, among the 2223 respondents, let us first determine how many people refused to participate in the survey at least 2 times:

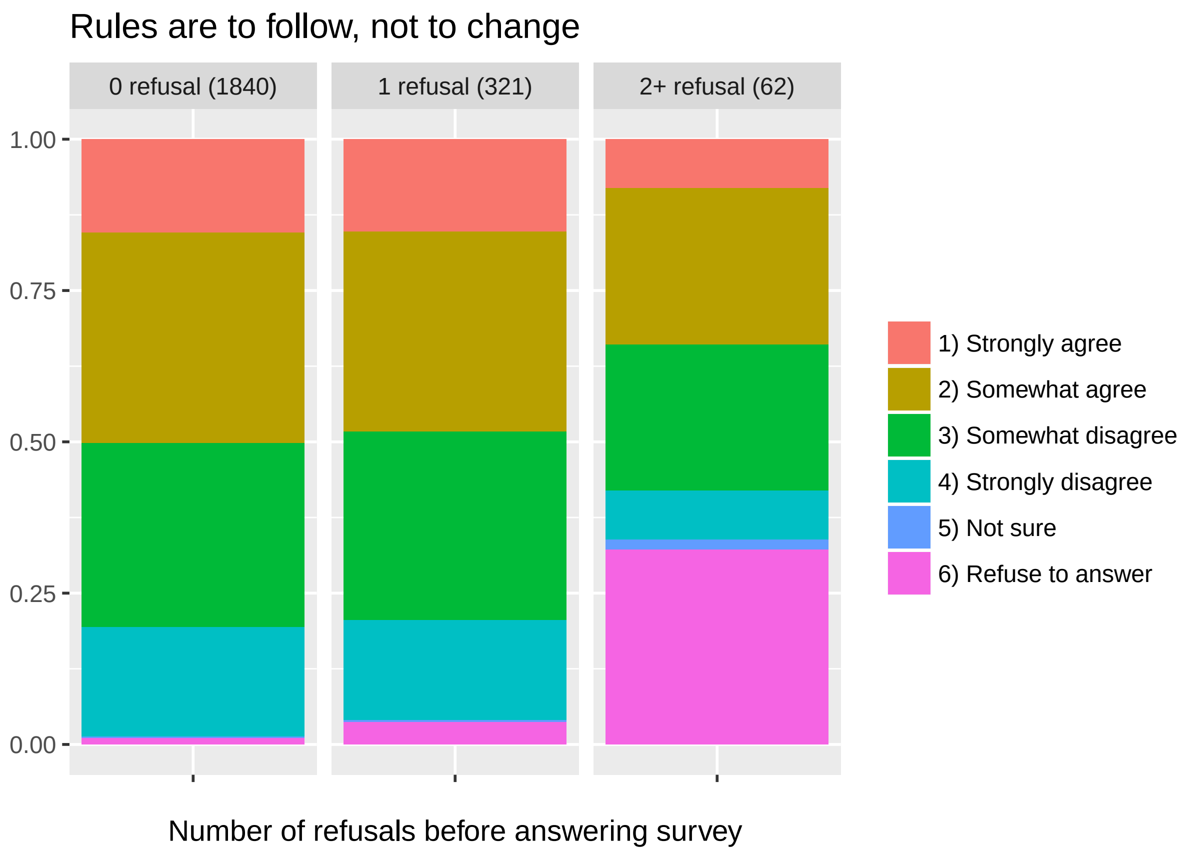
```r
survey <- read.table(file="data3.txt", sep=",", header=T)
sum(survey$rcnt==0)
```

```
## [1] 1840
```

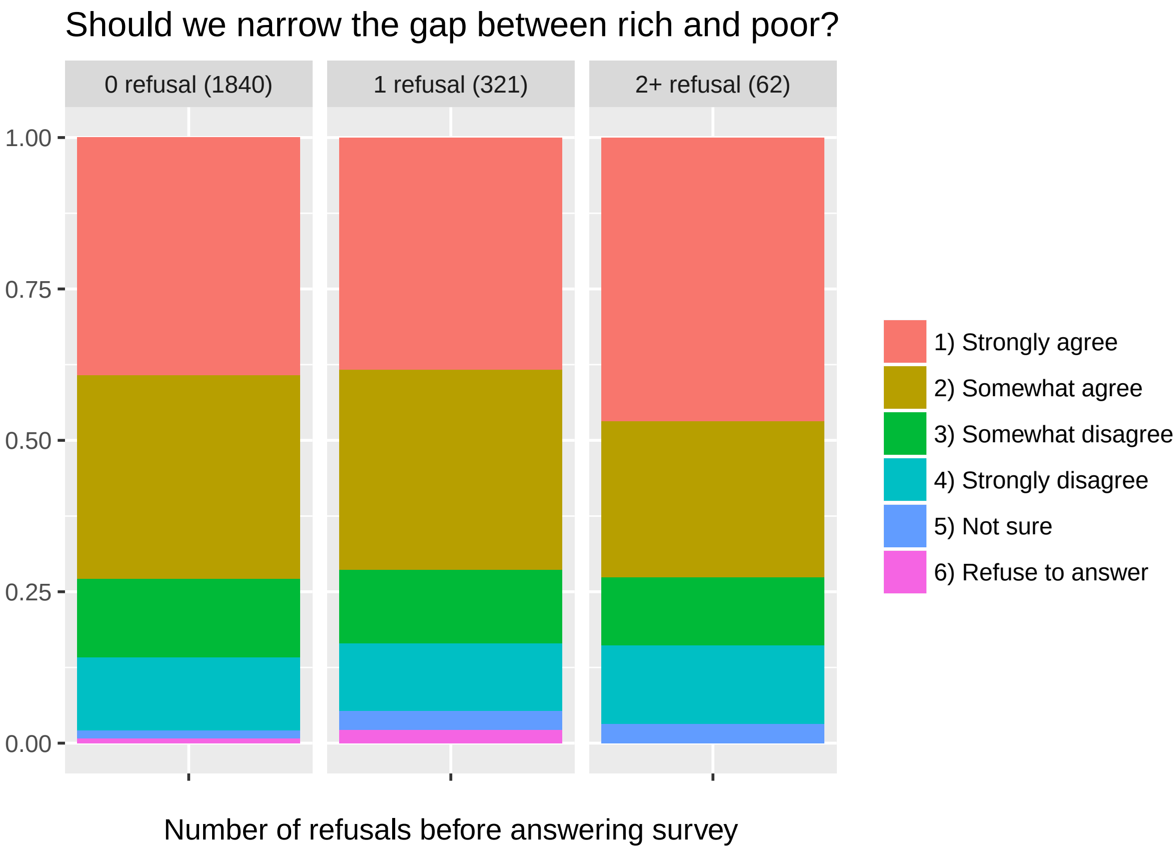
```r
sum(survey$rcnt>=2)
```

```
## [1] 62
```

Thus there were 1840 people who agreed to take the survey when they were first reached, and 62 people who eventually took the survey despite refusing to do so at least 2 times. Let us compare whether these early and late respondents held different opinion to the question "Rules are to follow, not change" (code hidden):



As we can see from the bar graph above, the proportion of people who refused to answer this question is significantly higher in the late respondents than initial respondents. Specifically, with 1840 initial respondents, 19 refused to answer this question, while 20 refused to answer this among the 62 late respondents. Clearly we have non-response bias, but do we immediately toss this dataset? Let us check out another example (code hidden again).



Here the graph for 0 refusal and 1 refusal look almost identical, while for the 2+ refusals a proportion of the somewhat agree went to the strongly agree. However, given that the proportion of responses agreeing to the statement as a whole is almost completely the same, I think it is safe to conclude that there is no non-response bias for this question.

So out of 2 questions, one definitely contained non-response bias, while the other probably did not. What does this tell us about the overall credibility of the survey? If we want to be more certain, we can easily look at more questions, as a total of 178 were asked. On the other hand, this method is a rather qualitative notion for exposing non-response biases. Suppose we have a dataset we really want to use, even if there are some non-response bias. Can we adjust the data cleverly so that our analysis is meaningful? In the next section, we show how to do so by weighting.

Adjusting for Unit Non-response with Weights

Suppose we wish to determine what percentage of population agrees with the statement “Most people who don’t succeed are lazy”. Given a set of responses, one thing we could do is to simply take the average: divide total number of responses that agrees with the statement by the total number of responses. Let us visualize the descriptive statistics (1~4 = Strongly agree ~ Strongly disagree, 9 = refuse to answer).

However, the above data visualization indicated that people around the age of 35 mostly voted for "3, somewhat disagree" to the above statement, as the density for that row appears to be the darkest. If our survey respondents somehow contained a larger than usual pool of people around 35 years old, we might underestimate the percentage of people who agrees with this statement. How do we adjust for this bias?

Overview

Intuitively, we decide who's response is "more important" in some way, and make their votes count more. In order to apply this technique, however, we must satisfy a very difficult criteria before applying it. That is, we must already have some basic information of the samples (e.g. age, income, weight…etc) *before* we send out our questionnaires. We use this information to categorize respondents and non-respondents into disjoint cells *after* we conclude our studies. Each cell relative to each other is equally important, so that individuals in a small cell is relatively more important than individuals in a large cell.

Note in our case, we can apply this technique because our first round of telephone survey already asked for certain information about our respondents. In general, this information is hard to obtain.

Definitions

- $\beta_i = 1$ if response is strongly agree or somewhat agree.

- S_i represents the number of respondents in each cell.
- $y_i = \beta_i S_i$ be the variable of interest.
- π_i be the probability to be drawn (design weight).
- p_i be the response probability .
- $w_i = (\pi_i p_i)^{-1}$ the non-response-adjusted weight for observation i.

Pre-weighting

First let us determine our sample's response to the statement "Most people who don't succeed are lazy".

```
##    Strg Agr   Smwat Agr Smwat Disag  Strg Disag      No Ans
##          51         341         563         229          14
```

We can naively take the average by dividing total number of positive response by the total number of response:

$$\overline{Y} = \sum y_i / \sum S_i = \frac{51 + 341}{51 + 341 + 563 + 229 + 14} = 32.7\%$$

Thus among all the respondents, only 32.7% of the population agrees with the statement. Quite a pessimistic society.

Post weighting by age

From the density scatter plot, we know that age affects people's opinion. So first we divide the respondents into disjoint cells by age.

Among the 2223 samples and 1198 respondents, their age distribution arranged in 10 years starting from 18 is summarized in the following table (cell size = 10 years):

```
##                           18~28 29~38 39~48 49~58 59~68 69~78 79~88 89+
## Total Sample                444   606   466   259   221   159    52  16
## Respondents                 213   320   256   148   136    95    28   2
## Respondents who agreed       66   103    76    54    36    45    10   2
```

Now to compute the adjusted percentage of people who agrees ($\hat{\overline{Y}}$):

$$\hat{\overline{Y}} = \frac{\sum w_i y_i}{S_i} = \frac{66\frac{444}{213} + 103\frac{606}{320} + 76\frac{466}{256} + 54\frac{259}{148} + 36\frac{221}{136} + 45\frac{159}{95} + 10\frac{52}{28} + 2\frac{16}{2}}{444 + 606 + 466 + 259 + 221 + 159 + 52 + 16} = 33.0\% \quad (1)$$

Thus the difference is quite small. We estimate that approximately 0.3% more people agrees with the statement than by averaging. This is not too unexpected, however, because the original survey employed a very sophisticated system of telephone calling to ensure a sample as random as possible. So non-response bias should already be minimized if not existant. Afterall, the best way to adjust for non-response bias is to prevent it with a good survey design. But that is another story.

Discussion

Surveying is an important source of data to undertsand general human behavior. As participants are not controlled as in a science experiment, being able to account for human biases is crucial to drawing meaningful inferences from this kind of data. In this project, I showed how to quickly identify non-response bias through callbacks, which is a tedious but standard technique used in almost every good survey. Then I showed how to adjust for non-response bias through weighting, which is very intuitive idea but requires information about participants before we conduct our survey. Hopefully these two techniques will be a useful starting point when one is faced with a real survey challenge some time in the future.

Unfortunately, I did not find any published analysis on this dataset, so I could not compare my results with any professional work. I was quite surprized that this is the case since the data collected is very complete.

References

- Sniderman, Tetlock, Piazza, "1991 Race and Politics Survey" http://sda.berkeley.edu/D3/Natlrace/Doc/nrac.htm (UC Berkeley, 1991)
- Survey Documental and Analysis (4.0) Archive http://sda.berkeley.edu/archive.htm (2014, UC Berkeley)
- Sharon L. Lohr "Sampling: Design and analysis" 2nd edition
- Jan Wang "Non-response in the Norwegian Business Tendency Survey" (2004)
- Roderick Little "Survey Nonresponse bias" (1984)
- Carlson and Williams, "A comparison of two methods to adjust weights for non-response: propensity modeling and weighting class adjustments" (2001)
- http://www.trchome.com/research-knowledge/white-paper-library/227-situational-use-of-data-weight
- http://www.trchome.com/65-market-research-knowledge/white-paper-library/215-non-response-bias-in-survey-sampling-complete
- http://fluidsurveys.com/university/how-to-avoid-nonresponse-error/
- Kristofer Jennings Lecture notes 7,8, and 10 for Stat 522 (Purdue University)
- Nathan Berg, "Non-Response Bias" (2005)
- Many many pages of Wikipedia