Identify and Adjust for Non-response Bias

benjamin chu

3/8/2017

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 in two different ways. 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.

Definitions and examples

- Random sampling - a subset of individuals (sample) chosen from a larger set (population) to be surveyed.
- Non-response bias - 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, 2.3 million surveys predict 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, because 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 identify non-response bias is to compare initial and late respondents. 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.

Among the 2223 respondents, we can determine how many people refused to participate in the survey at least 2 times:

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

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

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

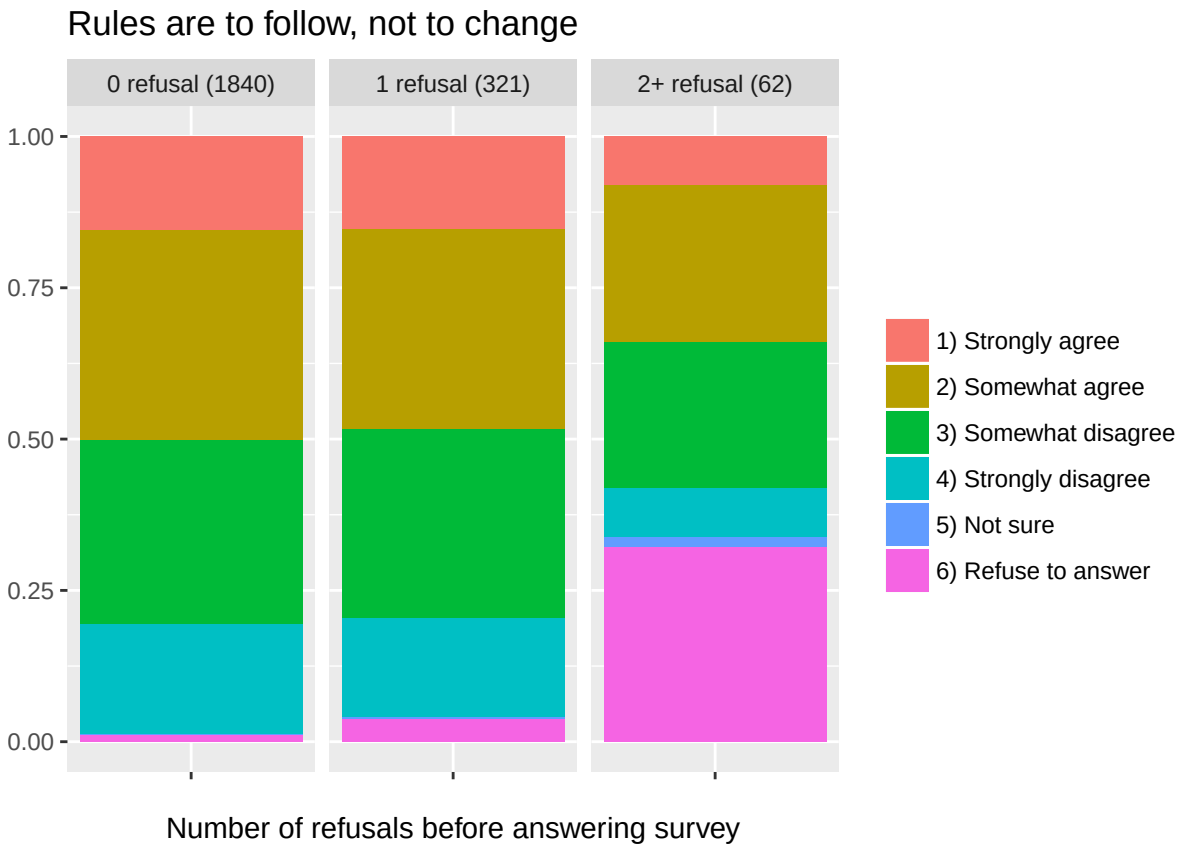
```
## [1] 62
```

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 they responded differently to the question "Rules are to follow, not change".

```
library(ggplot2)
library(sqldf)
test <- read.table(file="data3.txt", sep=",", header=T)
test$con2 <- factor(test$con2) # converts to a categorical variable
test$rcnt <- factor(test$rcnt)

test=sqldf("select CASEID,
  CASE WHEN rcnt==0 THEN '0 refusal (1840)'
       WHEN rcnt==1 THEN '1 refusal (321)'
       WHEN rcnt>=2 THEN '2+ refusal (62)'
  END rcnt,
  CASE WHEN con2==1 THEN '1) Strongly agree'
       WHEN con2==3 THEN '2) Somewhat agree'
       WHEN con2==5 THEN '3) Somewhat disagree'
       WHEN con2==7 THEN '4) Strongly disagree'
       WHEN con2==8 THEN '5) Not sure'
       WHEN con2==9 THEN '6) Refuse to answer'
  END con2 from test")

p = ggplot(data=test, aes(x="", stat="bin", fill=con2)) + geom_bar(position="fill")
p = p + ggtitle("Rules are to follow, not to change") + ylab("") + labs(fill='') + xlab("Number of refus
p = p + facet_grid(facets=. ~ rcnt) # Side by side bar chart
p
```



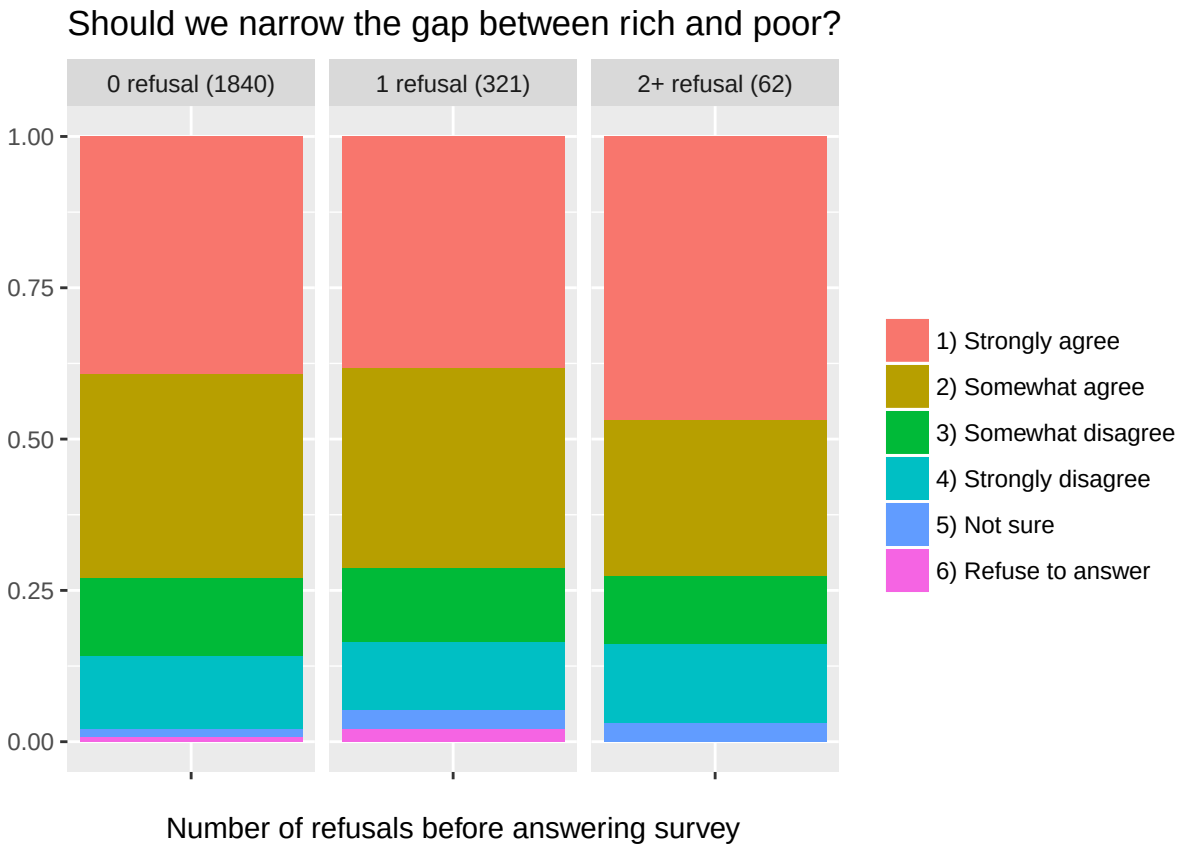
As we can see from the bar graph above, the proportion of people who refused to answer this question is significantly higher in the late respondents than initial respondents. With 1840 initial respondents, 19 refused to answer this question, while 20 refused to answer this among the 62 late respondents. Clearly there is non-response bias, but is this statistically significant? Let us check out another example.

chi squared test here?

```
test <- read.table(file="data3.txt", sep=",", header=T)
test$ef5 <- factor(test$ef5) # converts to a categorical variable
test$rcnt <- factor(test$rcnt)

test=sqldf("select CASEID,
  CASE WHEN rcnt==0 THEN '0 refusal (1840)'
       WHEN rcnt==1 THEN '1 refusal (321)'
       WHEN rcnt>=2 THEN '2+ refusal (62)'
  END rcnt,
  CASE WHEN ef5==1 THEN '1) Strongly agree'
       WHEN ef5==3 THEN '2) Somewhat agree'
       WHEN ef5==5 THEN '3) Somewhat disagree'
       WHEN ef5==7 THEN '4) Strongly disagree'
       WHEN ef5==8 THEN '5) Not sure'
       WHEN ef5==9 THEN '6) Refuse to answer'
  END ef5 from test")

p = ggplot(data=test, aes(x="", stat="bin", fill=ef5)) + geom_bar(position="fill")
p = p + ggtitle("Should we narrow the gap between rich and poor?") + ylab("") + labs(fill='') + xlab("Nu
p = p + facet_grid(facets=. ~ rcnt) # Side by side bar chart
p
```



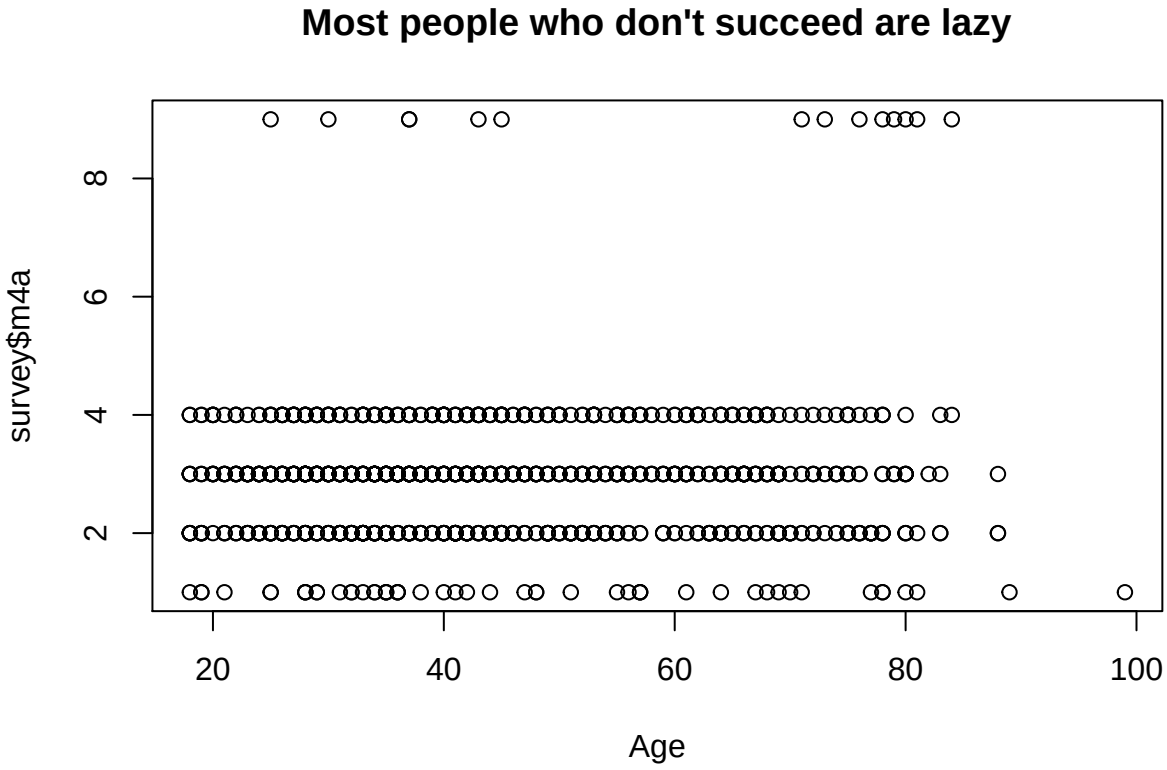
Here the graph for 0 refusal and 1 refusal look almost identical, while for the 2+ refusals a proportion of the Somewhat agree went to the Strongly agree. However with the responses both being affirmative, we conclude that there is no non-response bias.

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 analyze more questions as a total of 178 were asked. On the other hand, even if we looked at all of them, this method is a rather qualitative notion for exposing non-response biases. In the next section, we show how to actually adjust for them using weights.

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". 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:

```
plot(survey$age, survey$m4a, xlab="Age", main="Most people who don't succeed are lazy")
```



However, we may suspect that perhaps younger people are generally more optimistic about their future. These motivated youngsters, so confident in their ability, overwhelmingly voted yes to the above question. If our survey respondents somehow contained more young respondents due to random chance, we would overestimate the percentage of people who agrees with this statement. How do we adjust for this bias? Let us begin with definitions.

Overview

In order to apply the weighting technique, we must already have some basic information of the samples (e.g. age) *before* we send out our questionnaires. After we conclude our studies and obtain a list of respondents and non-respondents, we then compare the difference between them by categorizing respondents and non-respondents into disjoint cells. Based on the number of respondents in each cell, we scale each data point by their frequency. Intuitively, we decide who's response is "more important" in some way, and make their vote count more.

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's see people's response to the statement "Most people who don't succeed are lazy". (1 = Strongly agree, 2 = somewhat agree, 3 = somewhat disagree, 4 = Strongly disagree, 9 = missing data)

```r
table(survey$m4a)
```

```
##
##   1   2   3   4   9
##  51 341 563 229  14
```

So here

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

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

Post weighting by age

First we divide the respondents into disjoint cells by age (cell size = 10 years).

```r
#split(survey, cut(survey$age, c(18, 28, 38, 48, 58, 68, 78, 88, 150), include.lowest=TRUE))
msurvey <- survey[!is.na(survey$m4a),]
#split(msurvey, cut(msurvey$age, c(18, 28, 38, 48, 58, 68, 78, 88, 150), include.lowest=TRUE))
msurvey[msurvey$age>=18 & msurvey$age<=28 & msurvey$m4a<=2,]
msurvey[msurvey$age>=29 & msurvey$age<=38 & msurvey$m4a<=2,]
msurvey[msurvey$age>=39 & msurvey$age<=48 & msurvey$m4a<=2,]
msurvey[msurvey$age>=49 & msurvey$age<=58 & msurvey$m4a<=2,]
msurvey[msurvey$age>=59 & msurvey$age<=68 & msurvey$m4a<=2,]
msurvey[msurvey$age>=69 & msurvey$age<=78 & msurvey$m4a<=2,]
msurvey[msurvey$age>=79 & msurvey$age<=88 & msurvey$m4a<=2,]
msurvey[msurvey$age>=89 & msurvey$age<=150 & msurvey$m4a<=2,]
```

So we know that among the 2223 samples and 1198 respondents, their age distribution arranged in 10 years starting from 18 is summarized in the following table:

```r
sam_to_resp <- c(444, 606, 466, 259, 221, 159, 52, 16,
                 213, 320, 256, 148, 136, 95, 28, 2,
                 66, 103, 76, 54, 36, 45, 10, 2)
colname <- c("18~28", "29~38", "39~48", "49~58", "59~68", "69~78", "79~88", "89+")
rowname <- c("Total Sample", "Respondents", "Respondents who agreed")
age_dist <- matrix(sam_to_resp, nrow=3, ncol=8, byrow=TRUE, dimnames=list(rowname, colname))
age_dist
```

```
##                       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 ($\hat{\overline{Y}}$):

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

Discussion

Interestingly, I did not find any published analysis on this dataset, so I could not compare my analysis with any professional work.

References

- http://sda.berkeley.edu/D3/Natlrace/Doc/nrac.htm
- http://sda.berkeley.edu/archive.htm
- http://www.trchome.com/research-knowledge/white-paper-library/227-situational-use-of-data-weighting-complete
- http://www.trchome.com/65-market-research-knowledge/white-paper-library/215-non-response-bias-in-survey-sampling