markdown

benjamin chu

3/8/2017

Introduction

In biomath 204, we studied numerous data analysis techniques to draw useful information from given datasets. Surveys constitute a main source of data we may have to work with, and many industries and corporations rely on survey results to make business decisions. If we are unaware of bad surveying techniques and potential biases, we are volunerable to misleading datasets and may eventually make the wrong decision based on trashy information.

Method

Data Description

Results

Identifying Non-response Bias through Callback

An excellent and most common way of identifying non-response bias is to compare initial and late respondents. In the selected sample, people who did not participate in the survey (e.g. refusal, not at home...etc) may be asked again some time later. Those who agreed to take the survey only after several attempts are called late respondents. It is assumed that these late respondents are similar to the non-respondents, and the difference between respondents and nonrespondents 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="data2.txt", sep=",", header=T)
sum(survey$rcnt==0)
```

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

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

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

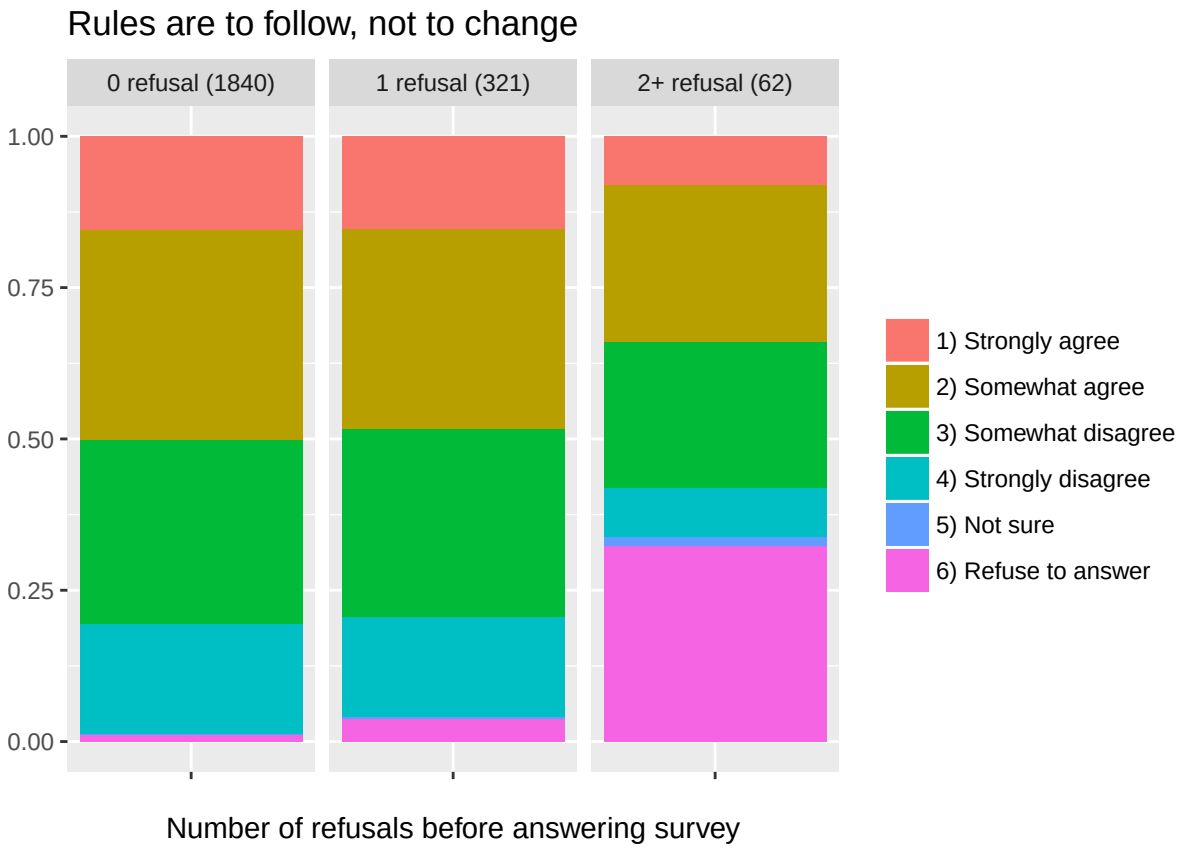
Thus there were 1840 people who agreed to take the survey when they were first reached, and 62 people who eventually took the survey despite refusing to do so at least 2 times. Let us compare whether they responded differently to the question "Rules are to follow, not change".

```
library(ggplot2)
library(sqldf)
test <- read.table(file="data2.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 distribution of opinions among initial and late respondents is not exactly the same. In particular, the proportion of people who refused to answer this question is significantly higher in the late respondents than initial respondents. Is this statistically significant? Let us test out another example.

```
test <- read.table(file="data2.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
```