Data Exploration: Making Decisions

Yanxi Fang

September 9, 2021

```
library(tidyverse)
library(estimatr) # difference_in_means()
```

In this Data Exploration assignment, you have two separate data sets with which you will work. The first involves the data generated by you and your classmates last week when you took the in-class survey. The second involves some of the data used in the Atkinson et al. (2009) piece that you read for class this week. Both data sets are described in more detail below.

If you have a question about any part of this assignment, please ask! Note that the actionable part of each question is **bolded**.

Part 1: Cognitive Biases

```
# read data
bias <- read.csv("bias_data.csv")
```

Question 1

Did you all exhibit this bias? Since the outcomes for this problem are binary, we need to test to see if the proportions who chose Program A under each of the conditions are the same. Report the difference in proportions who chose Program A under the 'save' and 'die' conditions. Do we see the same pattern that Kahneman described?

```
prop.table(table(bias$rare_disease_prog, bias$rare_disease_cond), margin = 1)
```

```
##
##                    die      save
##   Program A 0.3488372 0.6511628
##   Program B 0.6666667 0.3333333
```

In the table above, the proportions of the class that chose Program A are displayed. 0.3488 of the students chose A with the "die" language, while 0.6512 of the students chose A with the "save" language.

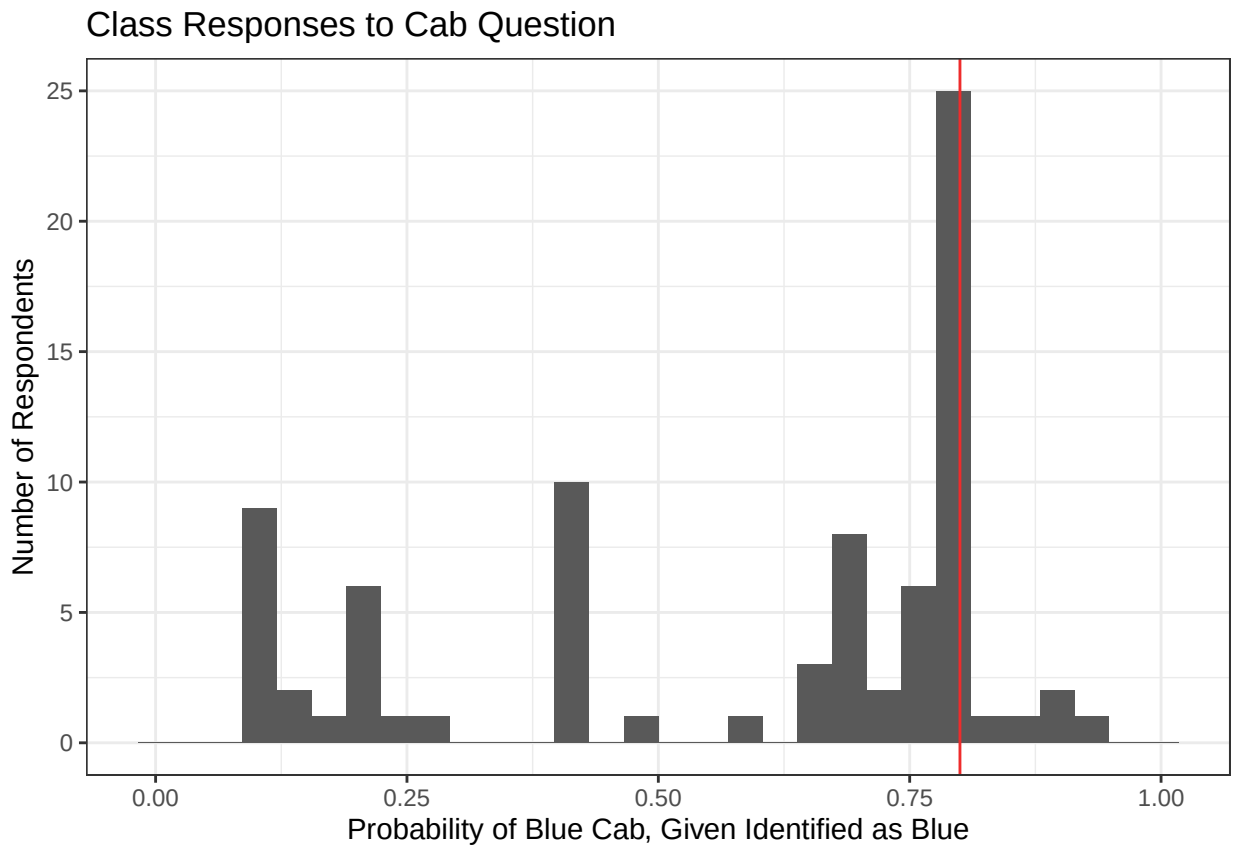
Question 4: Data Science Question

What is the true probability the cab was Blue? Visualize the distribution of the guesses in the class using a histogram. What was the most common guess in the class?

```
# clean data for cabs
bias_clean <- bias %>%
  filter(cab != "NA")

# visualize distribution w/ histogram
ggplot(data = bias_clean, aes(x = cab, xmin = 0, xmax = 1)) +
  geom_histogram() +
  theme_bw() +
  xlab("Probability of Blue Cab, Given Identified as Blue") +
  ylab("Number of Respondents") +
  ggtitle("Class Responses to Cab Question") +
  geom_vline(aes(xintercept = 0.8), col = "firebrick2")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# most common guess in the class
table(bias_clean$cab)
```

```
##
```

```
##  0.1 0.12 0.15 0.17  0.2 0.22 0.25 0.29  0.4 0.41  0.5  0.6 0.64 0.65 0.68 0.69
##    1    8    2    1    5    1    1    1    4    6    1    1    1    2    2    1
##  0.7 0.72 0.74 0.75 0.76 0.77  0.8 0.81 0.83 0.85  0.9 0.94
##    5    1    1    2    2    2   23    2    1    1    2    1
```

The histogram of the class guesses is shown above. The most common guess was 0.8, with a total of 23 responses.

For the true probability, we are given both the information about the witness (80 accuracy) and the information about the characteristics of the cabs (85 Green, 15 Blue) in the city.

Here, the probability of interest is the probability of the cab being blue, given that the witness says the cab is blue. This can be written as $P(A|B)$, where A is the cab being blue, and B is the witness saying that the cab is blue.

Bayes' Rule is needed to solve this problem. The unknown is $P(A|B)$, and the rule states that $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. Here, we know that $P(B|A)$ is 0.8, since it is the probability that the cab is identified as blue given that it is actually blue. We also know that $P(A)$ is 0.15, the proportion of blue cabs in the city.

Finally, $P(B)$, the probability that the witness identified a blue cab, is found by adding the probabilities that the witness identified a blue cab when the cab was blue, and that the witness identified a blue cab when the cab was green. Thus, $P(B)$ equals 0.8×0.15 plus 0.2×0.85.

Thus, plugging everything in to Bayes' Rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.8*0.15}{0.8*5*0.15+0.2*0.85} = 0.4138$. The true probability is 0.4138, which is slightly more than half of the probability that the class guessed.