Data Exploration: Making Decisions

Owen Bernstein

September 9, 2021

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

You may have noticed that the questions on the survey you took during class last week were based on the Kahneman (2003) reading you did for this week. The goal for this set of questions is to examine those data to see if you and your classmates exhibit the same cognitive biases that Kahneman wrote about. The data you generated is described below.

Data Details:

- File Name: `bias_data.csv`
- Source: These data are from the in-class survey you took last week.

Variable Name	Variable Description
`id`	Unique ID for each respondent
`rare_disease_prog`	From the rare disease problem, the program chosen by the respondent (either 'Program A' or 'Program B')
`rare_disease_cond`	From the rare disease problem, the framing condition to which the respondent was assigned (either 'save' or 'die')
`linda`	From the Linda problem, the option the respondent thought most probable, either "teller" or "teller and feminist"
`cab`	From the cab problem, the respondent's estimate of the probability the car was blue
`gender`	One of "man", "woman", "non-binary", or "other"
`year`	Year at Harvard
`college_stats`	Indicator for whether or not the respondent has taken a college-level statistics course

Before you get started, make sure you replace "file_name_here_1.csv" with the name of the file. (Also, remember to make sure you have saved the .Rmd version of this file and the file with the data in the same folder.)

```
# load the class-generated bias data
bias_data <- read_csv("C:/Users/owenb/OneDrive/Documents/Harvard Sem 5/Gov 1372/data/bias_data.csv")
```

Question 1

First, let's look at the rare disease problem. You'll recall from the Kahneman (2003) piece that responses to this problem often differ based on the framing (people being saved versus people dying), despite the fact that the two frames are logically equivalent. This is what is called a 'framing bias'.

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?

```
# Calculating the percentage of people that chose program a under die and save
# conditions

q1 <- bias_data %>%
  group_by(rare_disease_cond) %>%
  count(rare_disease_prog) %>%
  summarise(prop_program_a = n[rare_disease_prog == "Program A"] / sum(n))

# Calculating the difference in the proportion of people that chose program A
# under die and save conditions

prop_save <- q1[2] %>%
  slice(2)

prop_die <- q1[2] %>%
  slice(1)

diff <- prop_save - prop_die
```

66% of people chose program A under the save condition while only approximately 34% chose program A under the die condition. This is a difference of 0.317829457364341.

We do see the same pattern that Kahneman described. Specifically, participants were more likely to choose program A when it was described as saving lives and less likely to when it was described as people dying. This is because individuals are more risk adverse in the gain framing and more risk accepting in the loss framing.

EXTENSION: Report the 95% confidence interval for the difference in proportions you just calculated. Hint: the `infer` package has a function that is useful here. What does the 95% confidence interval mean?

Note that extensions to questions are not the same as data science questions. Complete this question if you like, but it is not required for data science students like actual data science questions.

```
# Proportion test of difference in proportion of participants who chose program
# A under each condition

prop_test(bias_data, rare_disease_prog ~ rare_disease_cond, order = c("save", "die"))
```

```
## # A tibble: 1 x 6
```

```
##   statistic chisq_df p_value alternative lower_ci upper_ci
##       <dbl>    <dbl>   <dbl> <chr>          <dbl>    <dbl>
## 1      7.36        1 0.00666 two.sided     0.0928    0.543
```

The results show a confidence interval of {0.09, 0.54}. In the frequentest interpretation, this means that if we conducted the same experiment many times, 95% of the sample differences in proportion would be between these two values. This means that the results are significantly significant and we can reject the null hypothesis that the difference in proportions is 0.

Question 4: Data Science Question

Now we will take a look at the taxi cab problem. This problem, originally posed by Tversky and Kahneman in 1977, is intended to demonstrate what they call a "base rate fallacy". To refresh your memory, here is the text of the problem, as you saw it on the survey last week:

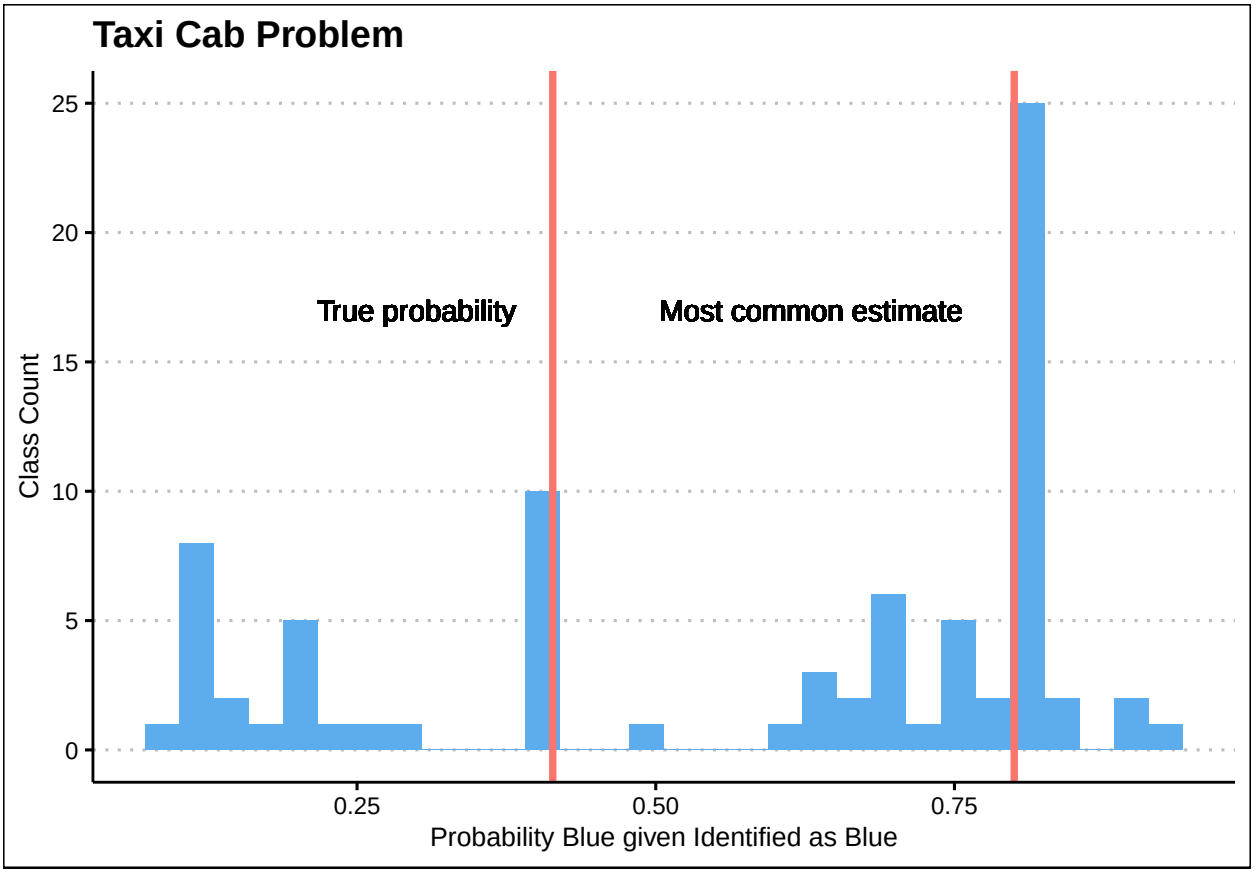
> A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. 85
>
> A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colours 80
>
> What is the probability that the cab involved in the accident was Blue rather than Green knowing that this witness identified it as Blue?

The most common answer to this problem is .8. This corresponds to the reliability of the witness, without regard for the base rate at which Blue cabs can be found relative to Green cabs. In other words, respondents tend to disregard the base rate when estimating the probability the cab was Blue.

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?

```
# Calculating true probability using Bayes Theorem

pr_w <- .8 * .15 + .2 * .85

pr_b_w <- .8 * .15 / .29

# Creating histogram of class guesses

hist_q4 <- bias_data %>%
  ggplot(aes(cab)) +
  geom_histogram(fill = "steelblue2") +
  labs(title = "Taxi Cab Problem", x = "Probability Blue given Identified as Blue", y = "Class Count") +
  geom_vline(aes(xintercept = pr_b_w, color = "indianred"), size = 1.3) +
  geom_vline(aes(xintercept = 0.8, color = "indianred"), size = 1.3) +
  geom_text(x = 0.3, y = 17, aes(label = "True probability")) +
  geom_text(x = 0.63, y = 17, aes(label = "Most common estimate")) +
  guides(size = "none") +
  guides(colour = "none") +
  theme_clean()

hist_q4
```

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

## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



The true probability that the car was actually blue given that the witness identified it as blue is 0.4137931. This is significantly less than the most common estimate of the probability by the class which was 0.8.