Does Social Pressure Affect Turnout? Part I: Loading and Making Sense of Data, and Computing and Interpreting Means (with Solutions)

(Based on Alan Gerber, Donald Green, and Christopher Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review*, 102(1): 33-48.)

In part II, we will estimate the average causal effect of receiving a message designed to induce social pressure to vote on the probability of voting. For this purpose, we will analyze data from an experiment conducted in Michigan in 2006, were registered voters were randomly assigned to receive a postcard with the following message:

> Dear Registered Voter:
>
> WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED? ... We're sending this mailing to you and your neighbors to publicize who does and does not vote. The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not. DO YOUR CIVIC DUTY–VOTE!

MAPLE DRIVE	Aug 2004	Nov 2004	Aug 2006
9993 [YOU]	Didn't vote	Voted	???
9995 JOSEPH JAMES SMITH	Voted	Voted	???
9997 RICHARD B JACKSON	Didn't vote	Voted	???
9999 KATHY MARIE JACKSON	Didn't vote	Voted	???

The dataset we will use is in a file called "voting.csv". Table 1 shows the names and descriptions of the variables in this dataset, where the unit of observation is registered voters.

variable	description
birth	year of birth of registered voter
message	whether registered voter was assigned to receive the social pressure message: "yes", "no"
voted	whether registered voter voted in the August 2006 election: 1=voted, 0=didn't vote

Table 1: Variables in "voting.csv"

In this problem set, we practice (i) how to load and make sense of data and (ii) how to compute and interpret means.

1. Use the function read.csv() to read the CSV file "voting.csv" and use the assignment operator <- to store the data in an object called *voting*. (Do not forget to set the working directory first.) Provide the R code you used (without the output). (5 points)

R code:

```
voting <- read.csv("voting.csv") # reads and stores data
```

(Recall: to the left of the assignment operator <-, we specify the name of the object, voting in this case; to the right of the assignment operator <-, we specify the contents, which, in this case, are produced by reading the CSV file "voting.csv". Also, we do not use quotes around the name of an object such as voting or around the name of a function such as read.csv(), but we do use quotes around the name of a file: "voting.csv".)

2. Use the function head() to view the first few observations of the dataset. Provide the R code you used (without the output). (5 points)

   R code:

```
head(voting) # shows first observations
##   birth message voted
## 1  1981      no     0
## 2  1959      no     1
## 3  1956      no     1
## 4  1939     yes     1
## 5  1968      no     0
## 6  1967      no     0
```

3. What does each observation in this dataset represent? (5 points)

   Answer: Each observation represents a registered voter. (Note: We know this because, as stated above, the unit of observation in this dataset is registered voters.)

4. Please substantively interpret the first observation in the dataset. (5 points)

   Answer: The first observation in the dataset represents a registered voter who was born in 1981, who was not assigned to receive the social pressure message, and who did not vote in the August 2006 election. (Note: the first observation consists of the following values: *birth*=1981, *message*="no", *voted*=0; we can interpret each value by using the description of the variables in Table 1.)

5. For each variable in the dataset, please identify the type of variable (character vs. numeric binary vs. numeric non-binary) (5 points)

   Answer: *birth* is a numeric non-binary variable, *message* is a character variable, and *voted* is a numeric binary variable. (Recall: binary variables can only take two values, here 0s and 1s, and non-binary variables can take more than two values.)

6. How many observations are in the dataset? In other words, how many registered voters were part of this experiment? (Hint: the function dim() might be helpful here.) Provide the R code you used (without the output) and provide the substantive answer. (5 points)

R code:

```
dim(voting) # provides dimensions of dataframe: rows, columns
## [1] 229444      3
```

Answer: There were 229444 registered voters that were part of this experiment. (Recall: the first number provided by dim() corresponds to the number of observations in the dataframe, the second number corresponds to the number of variables. Based on the output of dim(), there are 229444 observations in the dataframe *voting*. Since the unit of observation is registered voters, the dataframe *voting* has data on 229444 registered voters.)

7. Now, let's use the function mean() to calculate the average of the variable *birth*. Please provide a full substantive interpretation of what this average means. Make sure to provide the unit of measurement. (10 points)

R code:

```
mean(voting$birth) # calculates the average of birth
## [1] 1956.18
```

Answer: The average birth year among the registered voters in the experiment is 1956. (Note: the word "average" appears in the interpretation of this mean. Recall: The mean of a non-binary variable should be interpreted as an average and in the same unit of measurement as the variable itself, which is calendar year in this case.)

8. Finally, use the function mean() to calculate the average of the variable *voted*. Please provide a full substantive interpretation of what this average means. Make sure to provide the unit of measurement. (10 points)

R code:

```
mean(voting$voted) # calculates the average of voted
## [1] 0.3101759
```

(Recall: We use $ to access a variable within a dataframe. To its left, we specify the name of the object where the dataframe is stored, voting in this case; to its right, we specify the name of the variable, voted in this case. Also, we do not use quotes around the names of functions, the names of objects, or the names of elements within an object such as variables.)

Answer: 31% of the registered voters in the experiment voted in the August 2006 election. (Note: The word "average" does not appear in the interpretation of this mean because it should be interpreted as a proportion. Recall: The mean of a binary variable should be interpreted as the proportion of the observations that have the characteristic identified by the variable. Here, *voted* is binary and identifies the registered voters that voted. The unit of measurement is %, after multiplying the output by 100. Here: $0.31 \times 100 = 31\%$.)