Estimating the Bias in Self-Reported Turnout
Part III: Subsetting Variables and Creating Histograms

Let's continue working with the official and the self-reported ANES turnout data from 1980 to 2004. The dataset we will use is in a file called "ANES.csv". Table 1 shows the names and descriptions of the variables in this dataset, where the unit of observation is federal elections in the U.S.

variable	description
year	year of the election
presidential	whether it was a presidential election: 1=yes, 0=no
midterm	whether it was a midterm election: 1=yes, 0=no
ANES_turnout	proportion of ANES respondents who reported to have voted in the election (in percentages)
votes	number of ballots officially cast in the election (in thousands)
VEP	voting eligible population at the time (in thousands)
VAP	voting age population at the time (in thousands)
felons	number of felons not eligible to vote (in thousands)
noncitizens	number of non-citizens living in the U.S. (in thousands)

Table 1: Variables in "ANES.csv"

In this problem set, we practice creating new variables, visualizing the distribution of a variable, subsetting variables, and computing and interpreting means.

As always, we start by loading and looking at the data:

```
## load and look at the data
anes <- read.csv("ANES.csv") # reads and stores data
head(anes) # shows first observations
##   year presidential midterm ANES_turnout votes   VEP   VAP felons noncitizens
## 1 1980            1       0           71 86515 159635 164445    802        5756
## 2 1982            0       1           60 67616 160467 166028    960        6641
## 3 1984            1       0           74 92653 167702 173995   1165        7482
## 4 1986            0       1           53 64991 170396 177922   1367        8362
## 5 1988            1       0           70 91595 173579 181955   1594        9280
## 6 1990            0       1           47 67859 176629 186159   1901       10239
```

From the previous problem set, let's create the variable *VEP_turnout*, defined as the number of ballots officially cast in the election divided by the voting eligible population and multiplied by 100. This is the variable that we will assume measures the official voter turnout for each election (in percentages):

```
anes$VEP_turnout <- anes$votes / anes$VEP * 100 #creates new variable
```

1. Create a new variable called *turnout_bias* defined as the difference between *ANES_turnout* and *VEP_turnout*. Make sure to store this new variable in the existing dataframe named *anes* by using the $ character. (10 points)
2. Use the function head() to look at the first few observations again to ensure that you have created the new variable, *turnout_bias*, correctly. Is the first value of *turnout_bias* what one would expect, given the first values of *ANES_turnout* and *VEP_turnout*? What is the unit of measurement of *turnout_bias*? (5 points)
3. Create a visualization of the distribution of the variable *turnout_bias*. Are all the values positive? And, does this variable look normally distributed? (10 points)
4. Let's investigate whether the bias is larger in presidential elections than in midterm elections.
    a. For the presidential elections in the dataset, calculate the means of (i) *ANES_turnout*, (ii) *VEP_turnout*, and (ii) *turnout_bias*. Then, provide a substantive interpretation of what each of the averages mean, including the unit of measurement. (10 points)
    b. Now, for the midterm elections in the dataset, calculate the means of (i) *ANES_turnout*, (ii) *VEP_turnout*, and (ii) *turnout_bias*. Then, provide a substantive interpretation of what each of the averages mean, including the unit of measurement. (10 points)
    c. What can you conclude by comparing the results from question 4a to those from question 4b. (5 points)