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

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)

   R code:

```
anes$turnout_bias <- anes$ANES_turnout - anes$VEP_turnout #creates new variable
```

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)

   R code:

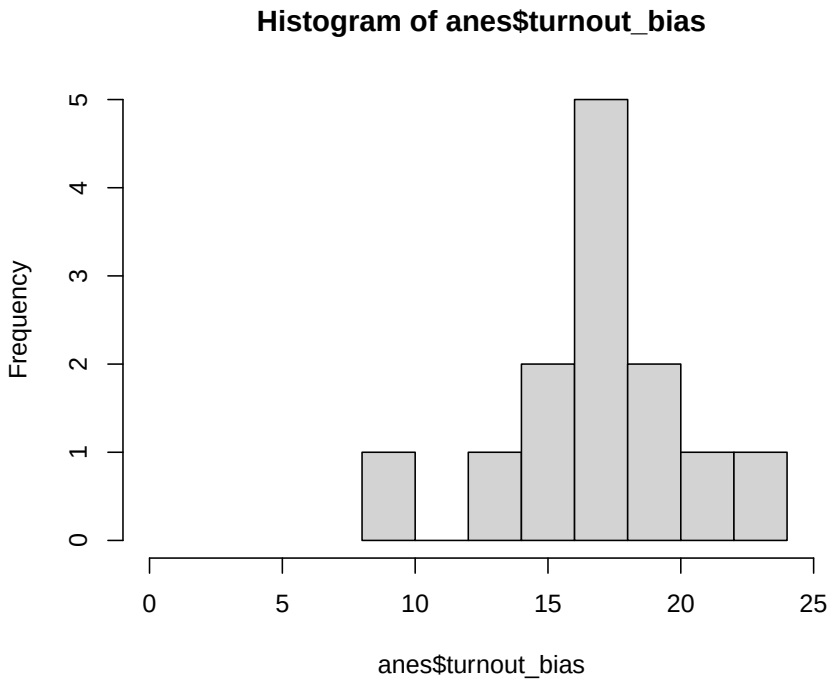
```
head(anes) # shows first observations
##   year presidential midterm ANES_turnout votes   VEP   VAP felons noncitizens
## 1 1980            1       0           71 86515 159635 164445   802        5756
## 2 1982            0       1           60 67616 160467 166028   960        6641
## 3 1984            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
##   VEP_turnout turnout_bias
## 1    54.19551    16.804491
## 2    42.13701    17.862987
## 3    55.24860    18.751404
## 4    38.14115    14.858846
## 5    52.76848    17.231520
## 6    38.41895     8.581054
```

   Answer: The first value of *turnout_bias* is 17 percentage points, which is what one would expect given that the first values of *ANES_turnout* and *VEP_turnout* are 71% and 54%, respectively (71% - 54% = 17 p.p.). The new variable *turnout_bias* is measured in percentage points because it is the difference between two percentages. Recall: percentage points is the unit of measurement for the arithmetic difference between two percentages (% - % = p.p.).

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)

   R code:

```
hist(anes$turnout_bias) # creates histogram
```



(Recall: the histogram of a variable is the visual representation of its distribution. The function in R to create a histogram is hist(). The only required argument is the code identifying the variable.)

Answers: Yes, all the values are positive. In other words, the self-reported turnout rate is always higher than the official turnout rate. The variable looks somewhat normally distributed because it is more or less symmetric and bell-shaped.

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)

   R code:

```
# compute mean of ANES_turnout for presidential elections
mean(anes$ANES_turnout[anes$presidential==1])
## [1] 73.28571
```

```
# compute mean of VEP_turnout for presidential elections
mean(anes$VEP_turnout[anes$presidential==1])
## [1] 55.1871
```

```
# compute mean of turnout bias for presidential elections
mean(anes$turnout_bias[anes$ presidential ==1])
## [1] 18.09862
```

   (Recall: We use [] to subset a variable; inside the square brackets, we specify the criterion

of selection. For example, we can use the relational operator == to set a logical test; only the observations for which the logical test are true will be extracted. In the first case above, we extract from *ANES_turnout* the observations that belong to the federal elections for which *presidential* equals 1 and then calculate the mean of those observations. The same logic applies to the other two means.)

Answer: Among the presidential elections in the database, about 73% of the ANES respondents reported to have voted, on average; yet, only about 55% of the voting eligible population was recorded as officially voting, on average. The bias in the self-reported presidential election turnout data is about 18 percentage points, on average. This means that the self-reported data overestimates turnout in presidential elections by 18 percentage points, on average. Put differently, in the presidential elections, close to 25% of the people who report to have voted are lying (18/73=0.25; 0.25*100=25%).

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)

R code:

```
# compute mean of ANES_turnout for midterm elections
mean(anes$ANES_turnout[anes$presidential==0])
## [1] 55
```

```
# compute mean of VEP_turnout for midterm elections
mean(anes$VEP_turnout[anes$presidential==0])
## [1] 39.5712
```

```
# compute mean of turnout bias for midterm elections
mean(anes$turnout_bias[anes$ presidential ==0])
## [1] 15.4288
```

(Note: These pieces of code are the same as above except that here we are interested in the observations for which *presidential* equals 0, or for which *midterm* equals 1. The subsetting criterion could, therefore, be either anes$presidential==0 or anes$midterm==1. Both would provide the same outputs.)

Answer: Among the midterm elections in the database, about 55% of the ANES respondents reported to have voted, on average; yet, only about 40% of the voting eligible population was recorded as officially voting, on average. The bias in the self-reported midterm election turnout data is 15 percentage points, on average. This means that the self-reported data overestimates turnout in midterm elections by 15 percentage points, on average. Put differently, in the midterm elections, close to 27% of the people who report to have voted are lying (15/55=0.27; 0.27*100=27%).

c. What can you conclude by comparing the results from question 4a to those from question 4b. (5 points)

Answer: Both the official and the ANES self-reported average turnout rates are higher in presidential elections than in midterm elections in the U.S. (Average *VEP_turnout*: 73% in presidential elections vs. 55% in midterm elections; Average *ANES_turnout*: 55% in presidential elections vs. 40% in midterm elections). However, both the average size of the bias and the average proportion of liars among those who claim to have voted are more or less the same in both kinds of elections (Average size of the bias: 18 p.p. in presidential elections vs. 15 p.p. in midterm elections; Average proportion of inaccurate information among those who claimed to have voted: 25% in presidential elections vs. 27% in midterm elections).