Estimating the Bias in Self-Reported Turnout
Part II: Computing and Interpreting Means (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 (i) using arithmetic operators to create new variables as well as (ii) 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
```

1. Create a new variable called *VEP_turnout* defined as the number of ballots officially cast in the election divided by the voting eligible population and multiplied by 100. Make sure to store this new variable in the existing dataframe named *anes* by using the character $. (See page 41 of DSS, to learn how to use the character $ to identify a variable inside a dataframe not just to access it but also to create it.) (5 points)

R code:

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

(Recall: When creating a new object or a new element within an object (as is the case here), we use the assignment operator <-. To the left of the assignment operator <-, we specify the name of the new object or the name of the new element within an object, here: anes$VEP_turnout since we are creating a new variable inside the existing dataframe *anes*; to the right of the assignment operator <-, we specify the contents, which, in this case, are produced by dividing the values of the variable *votes* by the values of the variable *VEP* and multiplying the result by 100. When specifying each variable, we use the $ character to identify the name of the object where the variables are either already stored, as is the case with *votes* and *VEP*, or should be stored, as is the case with *VEP_turnout*. Had we ran VEP_turnout <- anes$votes / anes$VEP instead, R would have created a new, separate object named *VEP_turnout*. However, what we want is to create a new element/variable called *VEP_turnout* inside the existing object/dataframe named *anes*, so we should add anes$ in front of the name of the new variable VEP_turnout.)

2. Use the function head() to look at the first few observations again to ensure that you have created the new variable, *VEP_turnout*, correctly. Is the first value of *VEP_turnout* what one would expect, given the first values of *votes* and *VEP*? What is the unit of measurement of *VEP_turnout*? (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
## 1    54.19551
## 2    42.13701
## 3    55.24860
## 4    38.14115
## 5    52.76848
## 6    38.41895
```

Answer: The first value of *VEP_turnout* is 54%, which is what one would expect given that the first values of *votes* and *VEP* are 86,515 thousand and 159,645 thousand, respectively (86515000 / 159645000 $\times$ 100 = 54%). The new variable *VEP_turnout* is measured in percentages because it is a proportion, the proportion of voters among the voting eligible population.

3. Now, create a new variable called *VAP_turnout* defined as the number of ballots officially cast in the election divided by the voting age population and multiplied by 100. Make sure to store this new variable in the existing dataframe named *anes*. (5 points)

R code:

```
anes$VAP_turnout <- anes$votes / anes$VAP * 100 #creates new variable
```

4. Use the function head() to look at the first few observations again to ensure that you have created the new variable, *VAP_turnout*, correctly. Is the first value of *VAP_turnout* what one would expect, given the first values of *votes* and *VAP*? What is the unit of measurement of *VAP_turnout*? (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 VAP_turnout
## 1    54.19551    52.61030
## 2    42.13701    40.72566
## 3    55.24860    53.25038
## 4    38.14115    36.52780
## 5    52.76848    50.33937
## 6    38.41895    36.45217
```

Answer: The first value of *VAP_turnout* is 52.6%, which is what one would expect given that the first values of *votes* and *VAP* are 86,515 thousand and 164,445 thousand, respectively (86515000 / 164445000 $\times$ 100 = 52.6%). The new variable *VAP_turnout* is also measured in percentages because it is a proportion, the proportion of voters among the voting age population.

5. Looking at the first few observations of the two new variables, *VEP_turnout* and *VAP_turnout*, shown by the function head() above, can you tell whether one of them always contains higher values than the other? Why do you think that is? Which of the two variables do you think most accurately measures turnout? (5 points)

Answer: The values of *VEP_turnout* are always slightly higher than the values of *VAP_turnout*. This is likely because among the people counted in the voting age population (VAP) variable, there are many who are not eligible to vote, such as felons in some states and non U.S. citizens. As a result, the values of *VAP* will be higher than the values of *VEP*, and thus, *VAP_turnout* will be lower than *VEP_turnout* (since *VAP* and *VEP* are in the denominator of the formula that calculates the turnout). *VEP_turnout* reflects turnout in a more accurately way since it only takes into account the population who are eligible to vote.

6. Use the function mean(), to compute the average value of *VEP_turnout* among the 13 federal elections in the dataset. Please provide a full substantive interpretation of what this average means and make sure to provide the unit of measurement. (10 points)

R code:

```
mean(anes$VEP_turnout) #computes average
## [1] 47.97976
```

Answer: Among the 13 federal elections in the dataset, on average, about 48% of the voting eligible population voted. (Note: this is an average of percentages. In other words, the unit of measurement is percentages and it is an average value.)

7. Use the function mean(), to compute the average value of *ANES_turnout* among the 13 federal elections in the dataset. Please provide a full substantive interpretation of what this average means and make sure to provide the unit of measurement. (10 points)

   R code:

   ```
   mean(anes$ANES_turnout) #computes average
   ## [1] 64.84615
   ```

   Answer: Among the 13 federal elections in the dataset, on average, about 65% of the ANES respondents reported to have voted. (Note: this is an average of percentages. In other words, the unit of measurement is percentages and it is an average value.)

8. When comparing the average value of *ANES_turnout* to the average value of *VEP_turnout*, do you find any evidence of people lying about their voting behavior? (5 points)

   Answer: There is quite a substantial difference between the average ANES self-reported turnout rate and the average official turnout rate, among the 13 federal elections in the dataset. (Specifically, the average difference is of 65% - 48% = 17 percentage points.) We can, therefore, state that we have found some evidence of people lying about their voting behavior.