Bias in Self-reported Turnout

Autumn 2019

Surveys are frequently used to measure political behavior such as voter turnout, but some researchers are concerned about the accuracy of self-reports. In particular, they worry about possible *social desirability bias* where in post-election surveys, respondents who did not vote in an election lie about not having voted because they may feel that they should have voted. Why might they do that? Is such a bias present in the American National Election Studies (ANES)?

The ANES is a nation-wide survey that has been conducted for every election since 1948. The ANES conducts face-to-face interviews with a nationally representative sample of adults. The table below displays the names and descriptions of variables in the `turnout.csv` data file.

Name	Description
`year`	Election year
`VEP`	Voting Eligible Population (in thousands)
`VAP`	Voting Age Population (in thousands)
`total`	Total ballots cast for highest office (in thousands)
`ANES`	Self-reported turnout rate in the ANES (%)
`felons`	Total ineligible felons (in thousands)
`noncitizens`	Total non-citizens (in thousands)
`overseas`	Total eligible overseas voters (in thousands)
`osvoters`	Total ballots counted by overseas voters (in thousands)

Question 1

Before we begin, remember to set your working directory to the location where you saved the data. You can check your current working directory at any time with the function `getwd()`. And you can set the working directory with the `setwd()` function (e.g., `setwd("C:/R/")` if we wanted to set the working directory to a folder called "R" on a Windows-based computer.

Now let's load the data into R as an object called `turnout` using the `read.csv` function. First, let's check that you've done this properly using the `head()` function: `head(turnout)`. How many observations are contained in this dataframe?

Now let's check the summary statistics of the variables in this dataframe. We can do that with the `summary()` function: `summary(turnout)`. You can do this with small dataframes, but for larger ones with many variables, it's better to ask for summary statistics only for the variables you're interested in. To do this, we can use the `$` operator to let R know that we want to look within an object for a particular variable. What happens when you type the object `turnout$` with the `$` operator? Let's ask for summary statistics for the variable `total`: `summary(turnout$total)`.

Finally, let's see how many different years are covered in the dataframe. We can do this with the `table()` function: `table(turnout$year)`. Do you notice any missing years in this dataframe?

Question 2

You'll notice that the dataframe contains raw vote and population totals, which means that we can't really compare them from year to year because of changes in population over time. Instead, we need to calculate the turnout *rate* based on the voting age population (VAP). Note that for this dataframe, we must add the total number of eligible overseas voters since the *VAP* variable does not include these individuals in the count.

Turnout rate = total turnout / (VAP + overseas voters)

To do this in R, we need to create a new variable in the `turnout` dataframe by using the `$` operator: `turnout$VAP.tr <- turnout$total / (turnout$VAP + turnout$overseas)`

Now check that you've successfully added a new variable to the dataframe by using the `summary()` function for that variable.

If we want to convert this turnout rate from a proportion to a percentage, we can do that easily by using multiplying this rate by 100. We can either create a new variable using the procedure above or simply replace the `VAP.tr` variable using a slightly different line of code: `turnout$VAP.tr <- turnout$total / (turnout$VAP + turnout$overseas) * 100` Again, check that the variable is now a percentage rather than proportion.

What happens if we calculate the turnout rate using the voting eligible population (VEP) rather than VAP? What difference do you observe? Which do you think is a more accurate measure of voter turnout?

Question 3

Now let's create a new variable called `VAP.diff` to check the difference between the VAP and ANES estimates of the turnout rate. How big is the difference on average? What is the minumum and maximum of the difference? Conduct the same comparison for the VEP and ANES estimates of voter turnout. Briefly comment on the results.

Question 4

Now let's compare the VEP turnout rate with the ANES turnout rate separately for presidential elections (every 4 years beginning in 1980; e.g., 1980, 1984, 1988, etc.)
and midterm elections (every 4 years beginning in 1982; e.g. 1982, 1986, 1990, etc.). One fun thing about R is that there are many different ways to accomplish the same task. One way to break the data by election type is by subsetting the it using the `subset()` function, which can take three arguments. Arguments are used to specifiy exactly what the function should be doing. The first argument is required and consists of the dataframe that you want to subset – in this case, we want to subset the `turnout` dataframe. We can then specify how we want to do this with another argument using conditional logic.

For example, if we wanted to subset the `turnout` dataframe by a specific year – 1980 – we could do this with the following code: `turnout.sub <- subset(turnout, year == "1980")`. This would create a new dataframe stored as the object `turnout.sub` only with a single year – 1980. If we wanted to subset the data by additional years we could do so with the the `|` operator (i.e., 1980 or 1984). The new code would look like the following: `turnout.sub <- subset(turnout, year == "1980" | year == "1984")`.

Once we've subsetted the data by election type – presidential or midterm – we can check to see whether the self-reported bias from the ANES varies across election types (e.g., create a new variable in each dataset and check its summary statistics). Is there a differece in bias between presidential and midterms elections?

Question 5

Ok, now let's divide the data into half by election years so that you subset the data into two distinct periods (e.g., the first 7 election years in one dataframe vs. the second 7 election years). Create a new variable that calculates the difference between the VEP turnout rate and the ANES turnout rate separately for each year within each period. Now has the bias of the ANES increased over time?

Question 6

(For fun!) The ANES does not interview overseas voters and prisoners. Calculate an adjustment to the 2008 VAP turnout rate. Begin by subtracting the total number of ineligible felons and non-citizens from the VAP to calculate an adjusted VAP. Next, calculate an adjusted VAP turnout rate, taking care to subtract the number of overseas ballots counted from the total ballots in 2008. Compare the adjusted VAP turnout with the unadjusted VAP, VEP, and the ANES turnout rate. Briefly discuss the results.