Exercise 2: Collecting and Organizing Data

PS 270: Understanding Political Numbers

Due Wednesday, March 13

In this exercise, we will practice collecting and organizing real data from an online source, organize it in a spreadsheet, and then read it into R for analysis. These are important skills to practice for working with real data, and they will be useful for your final project.

Backstory: All elections are implemented by humans and human-designed systems. The vote tallies you hear about on election night are rarely the numbers that become certified as official election results. While this is entirely ordinary, sometimes these processes attract attention and controversy. Alabama's 2002 gubernatorial election is one example. The incumbent governor, Donald Siegelman (D), was initially declared the winner on election night by the AP, using uncertified vote totals. Around 10:30 p.m., computer results from Baldwin County showed Siegelman winning roughly 19,000 votes in the county, but later print-outs from ballot counting machines in Baldwin County listed Siegelman with only 13,000 votes. This turn of events, which was attributed to a computer "glitch," flipped the election in favor of challenger Bob Riley (R). This led to a series of legal challenges and allegations of partisan interference in the election. Riley eventually became the governor and was later reelected.

Figure 1 shows the magnitude of the change in Baldwin County. The left panel compares 2002 to 1998; overall the Democratic candidate did worse in 2002 compared to 1998, but the change in Baldwin County was among the larger shifts within a single county. The right panel compares just the 2002 results before and after the Baldwin County revision. The change in Baldwin County led to a 9 percentage point drop in Siegelman's vote share.

Your Task: A similar event occurred in Wisconsin's Supreme Court election in 2011.[1] Election night results indicated that the race was too close to call, with liberal challenger JoAnne Kloppenburg holding a slight lead. Two days later, an updated vote count from Waukesha County gave conservative incumbent Justice David Prosser a slight lead. Was something amiss? Did the change in Waukesha County have a big impact on the final tally in the election? Your job is to collect data on the 2011 Supreme Court election, comparing election night to the officially certified results by recreating the *right-side* panel of Figure 1.

1. The Milwaukee Journal-Sentinel published a table[2] comparing vote tallies from election night ("AP, April 6") to the certified election results (marked "Updated"). Copy the table and paste it into a spreadsheet program like Excel or something similar.

[1] Read more: http://archive.jsonline.com/news/statepolitics/119410124.html/

[2] http://archive.jsonline.com/news/statepolitics/119497684.html/

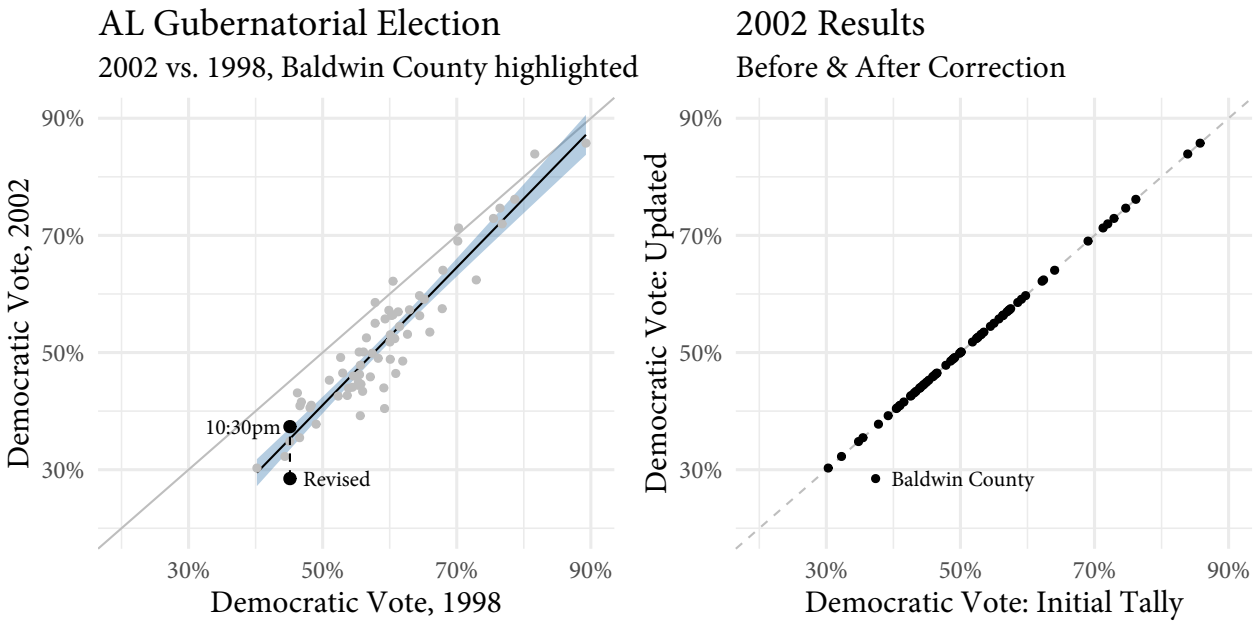

Figure 1: Election Data from Alabama. Left: comparison of 2002 and 1998. Right: comparison of 2002 results before and after changes in Baldwin County.

2. The table will appear disorganized in Excel, but salvageable. Clean the data so it looks like the data tables we've been using so far: one row for each observation (county), and one column per variable. Make sure that variable names are in the top row of the table, and ONLY the top row. There are also obnoxious rows midway through the table that remind you of the variable names; delete these rows as well as the bottom row ("Totals"). Rows can be deleted by right-clicking and selecting "Delete."
3. Your objective should be a table with 5 columns: county (1), election night vote counts for Prosser and Kloppenburg (2, 3), and certified results for Prosser and Kloppenburg (4, 5). You can delete the gain/loss columns. When your data look right, save the table into your `data` folder in `.csv` format.[3]
4. Open R, and load the `tidyverse` and `here` packages.
5. Read the WI data into R using the `read_csv()` function like we did in Exercise 1.[4] Examine the data type for each column. You want the vote tallies to be either `<int>` or `<dbl>` format. If they are `<chr>`, use `mutate()` to convert each variable to numeric by passing a variable to the `parse_number()` function.
6. Calculate Republican vote percentages from the raw vote variables, with separate vote percentages for election night and for the final vote. Use the same formula as in Exercise 1.
7. Create a scatter plot that compares election night to the final vote. Do something to identify Waukesha County, using aesthetics or labeling. Make it look good, and save it.
8. At the bottom of your R script, write a comment with your conclusion about how big the change in Waukesha County appears.
9. Upload your graphic and your R script to Canvas.

[3] You can use the Exercise 1 data file as a model for what the data should look like.

[4] If the data contain weird columns like `X7` in R, you can delete them using `select(-starts_with("X"))`.