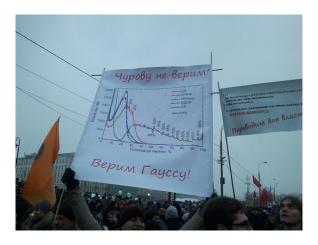# CIV 337 Homework 5

due: **10/2/24**

Instructions: Please submit solutions on canvas. Only a knitted pdf of an `Rmarkdown` file will be accepted.

Problem 1: The paper Rich State, Poor State, Red State, Blue State: What's the Matter with Connecticut? discusses the peculiar trend of Republican voting share's relationship to income within and among states. See, for example, Figures 3 and 4 in the paper. Without worrying about the modeling details, describe qualitatively what Figure 4 is showing (the different axes, open and solid circles, etc.). Why are the lines different slopes on the left panel? What other variables are potentially driving the difference in these slopes? Please be as descriptive and specific as possible, either through additional quantitative evidence outside of the paper or developing a testable hypothesis.

Problem 2: The poster below says, "We don't believe Churov! We believe Gauss!" Churov is the head of the State Electoral Commissions and Gauss refers to a 18th century German mathematician, Carl Friedrich Gauss, whom the Gaussian (Normal) distribution was named after.

In this exercise, we use the rules of probability to detect election fraud by examining voting patterns in the 2011 Russian State Duma election. (The State Duma is the federal legislature of Russia.)[1] The ruling political party, United Russia, won this election, but faced many accusations of election fraud, which the Russian government denied. Some protesters highlighted irregular patterns of voting as evidence of election fraud, as shown in the poster. In particular, protesters pointed out the relatively high frequency of common fractions such as 1/4, 1/3, and 1/2 in the official vote shares.

We use official election results, contained in the `russia2011` data frame in `fraud.RData` to investigate whether there is any evidence for election fraud. This file can be loaded using the `load` function. In addition to `russia2011`, the file contains election results from the 2003 Russian Duma election `russia2003`, the 2012 Russian presidential election `russia2012`, and the 2011 Canadian election `canada2011` as separate data frames. Each of these data sets has the same variables, described in the table below.

---

[1]This exercise is based on: Detecting Election Fraud from Irregularities in Vote-Share Distributions. *Political Analysis.* Arturas Rozenas (2017).

Variable	Description
N	Total number of voters in a precinct
turnout	Total number of turnout in a precinct
votes	Total number of votes for winner in a precinct

a. To analyze the 2011 Russian election results, first compute United Russia's vote share as a proportion of the voters who turned out. Identify the 10 most frequently occurring fractions for the vote share. Create a histogram that sets the number of bins to the number of unique fractions, with one bar created for each uniquely observed fraction, to differentiate between similar fractions like 1/2 and 51/100. This can be done by using the `breaks` argument in the `hist` function. What does this histogram look like at fractions with low numerators and denominators such as 1/2 and 2/3?

b. The mere existence of high frequencies at low fractions may not imply election fraud. Indeed, more numbers are divisible by smaller integers like 2, 3, and 4 than by larger integers like 22, 23, and 24. To investigate the possibility that the low fractions arose by chance, assume the following probability model:

   - Turnout for a precinct is binomially distributed, with size equal to the number of voters in the precinct and success probability equal to its observed turnout rate.
   - Conduct a Monte Carlo simulation under these assumptions. 1000 simulated elections should be sufficient. (Note that this may be computationally intensive code. Write your code for a small number of simulations to test before running all 1000 simulations.)

   What are the 10 most frequent vote share values? Create a histogram similar to the one in the previous question. Briefly comment on the results.

c. To judge the Monte Carlo simulation results against the actual results of the 2011 Russian election, we compare the observed fraction of observations within a bin of certain size with its simulated counterpart. To do this, create histograms showing the distribution of part (b)'s four most frequently occurring fractions, i.e., 1/2, 1/3, 3/5, and 2/3, and compare them with the corresponding fractions' proportion in the actual election. Briefly interpret the results.

d. We now compare the relative frequency of observed fractions with the simulated ones beyond the four fractions examined in the previous question. To do this, we choose a bin size of 0.01 and compute the proportion of observations that fall into each bin. We then examine whether or not the observed proportion falls within the 2.5 and 97.5 percentiles of the corresponding simulated proportions. Plot the result with vote share bin on the horizontal axis and estimated vote share on the vertical axis. This plot attempts to reproduce the one held by protesters in the figure. Now count the number of times an observed precinct vote share falls outside its simulated interval. Interpret the results.

e. To put the results of the previous question in perspective, apply the procedure developed in the previous question to the 2011 Canadian elections and the 2003 Russian election, where no major voting irregularities were reported. In addition, apply this procedure to the 2012 Russian presidential election, where election fraud allegations were reported. No plot needs to be produced. Briefly comment on the results you obtain.

Note: This question requires computationally intensive code. Write a code with a small number of simulations first and then run the final code with 1000 simulations.

Problem 3: Using built-in `R` functions, simulate three random variables that display a confounded relationship. You are free to choose the distributions these random variables follow, and be sure to provide visual and numerical evidence of confounding.