POLI502 Methods of Political Analysis, 2024 Fall
Instructor: Howard Liu

2024.10.11

Midterm Exam

Due 2024.10.11 5:00pm

Rules (5 points each × 3 = 15 points)

- Submit your midterm on your Github repository under a folder called "midterm502." Late submissions will not be accepted.
- Submit following files: (1) your qmd code script (e.q., POLI502-midterm-howardLiu.qmd). (2) a knitted pdf so I can write comments on it (POLI502-midterm-howardLiu.pdf). You'll earn 5 points if you do all of these correctly.
- Make sure your script has a proper header like the my qmd samples. Also Make sure your script runs without an error. I will execute your file to check if need be. You'll earn 5 points if you do all of these correctly.
- Add comments and code annotations to everything you do. Try to make your script look like mine. If your output doesn't have proper annotations, you'll lose 5 points. Don't copy and paste all the questions into your script, but do show me the question number for each question. You'll earn 5 points if you do this correctly.

Notes

- The last question is probably more difficult and time consuming than others. I advise you not to spend too much time on q17 if you find it too difficult.

Tasks (5 points each × 17 = 85 points)

1. Load the world dataset (`world.csv`), and store it as an object named `world.data`. It can be downloaded [here]
2. The data set contains a dummy variable (i.e., a nominal variable with two categories) named oecd that classifies countries into two groups, OECD member countries and non-member countries. One way to describe and summarize the information contained in a nominal variable is to describe the distribution numerically. As we learned during the past weeks, we describe the distribution of a nominal variable numerically by creating a frequency table. Create a frequency table of this variable and store it into a data frame object `ft.oecd`. The table has to have three columns: values (initially called `Var1`), frequency (called `Freq`), and percentage (should be called `Percentage`). Change the column name of the first column to "OECD Member?".
3. According to the frequency table you created above, (A) how many countries in the data set are OECD members? (B) How many countries in the data set are not? (C) What percentage of countries are OECD members? (D) What percentage of countries are non-members? Give me four answers (four numbers) as a comment. Note: for this task, you don't need an R command. Just read the table and tell me the numbers.

- Hint 1: Don't forget to load the package using the library function. It's usually a good idea to do so at the beginning of your R script.
- Hint 2: Don't forget to change the axis labels using the xlab and ylab options. The appropriate label for the X axis would be "OECD membership", whereas the label for the Y axis could be "Number of countries".

4. Another way to describe and summarize a nominal variable is to draw a frequency distribution graph. For nominal variables, we draw a bar chart. Using the functions available in the ggplot2 package (e.g., `geom_bar`), draw a bar chart of the dummy variable that measures OECD membership.

5. List three countries that are coded as OECD member states. List three countries that are non-democratic according to the democracy dummy variable.

6. The data set contains a numerical variable (interval-level variable) named `gdp_10_thou` that records a country's per capita GDP in 10,000 US dollars. Note that this variable measures per capita GDP in 10,000 dollars, not in dollars. This means that, when this variable takes a value of 4, for example, then that country's per cpaita GDP is 40,000 dollars, not 4 dollars. Describe this variable numerically by calculating the following statistics:

   - Range (minimum and maximum), median, mean, 1st and 3rd quartile values (Hint: this can be done at once with one command)
   - Standard deviation (Hint: you need to take care of missing values using the `na.rm` option)
   - *Note*: You need to provide R commands, not just numerical answers for this one.

7. It appears that the mean and the median of this per capita GDP variable are far apart: the mean is 6,018 dollars whereas the median is 1,897 dollars. Given that the mean is much higher than the median, the distribution of this variable is very skewed (i.e., not symmetric). In which way does the skew go? Answer this question by choosing between two options: (A) negatively skewed (skewed to the left) or (B) positively skewed (skewed to the right). *Note*: Give me your answer in words, not in R commands.

8. Describe this per capita GDP variable graphically by drawing a histogram.

   - Hint: Don't forget to change the axis labels using the xlab and ylab options. The appropriate label for the X axis would be "Per capita GDP (in 10,000 US dollars)", whereas the label for the Y axis could be "Number of countries".

9. There are two countries in the data set whose per capita GDP is greater than 40,000 US dollars. Identify these two countries. For this task, I need an R command that gives us the name of the two countries. Your command will probably generate the `NA` symbols (14 of them), along with the name of the two countries, but that's fine.

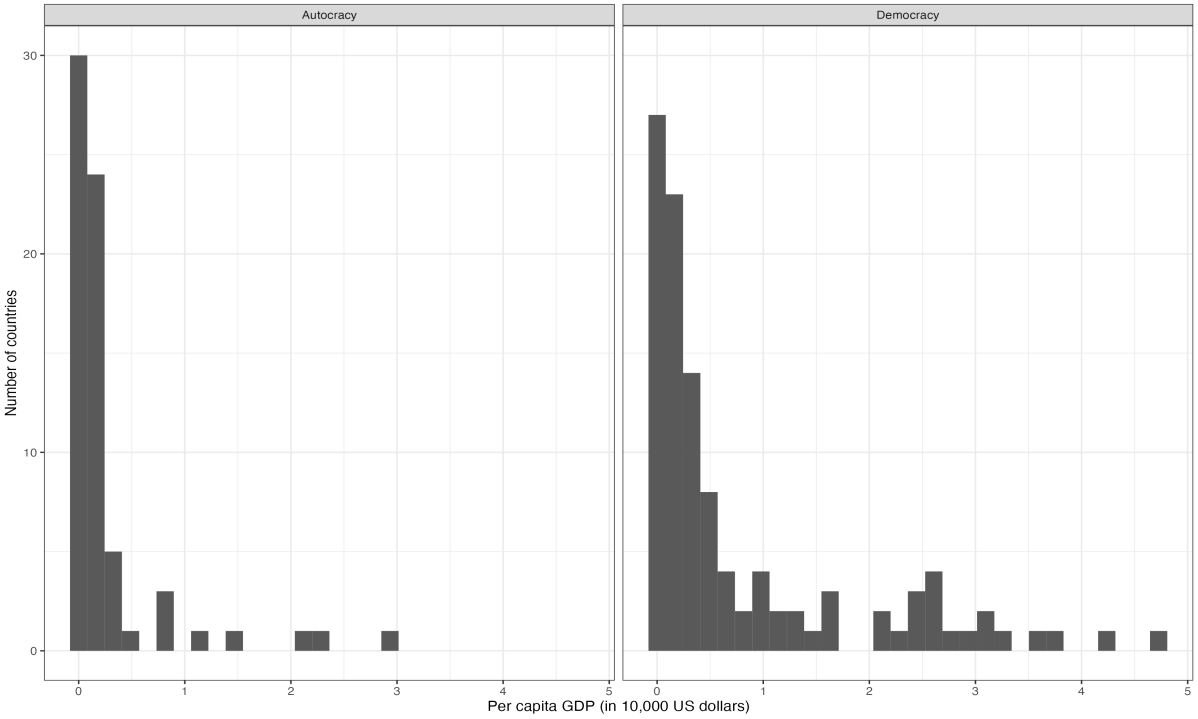
10. We have calculated the sample mean of this per capita GDP variable in task 6. We have also calculated its standard deviation. We also know from task 6 that there are 14 observations (countries) where this variable is missing, so we have 191 (total number of countries in the data set) $-$ 14 = 177 observations (i.e., n = 177). Therefore, we have all the building blocks to calculate the standard error of the mean. Calculate the standard error (the answer should be 0.07091015). *Note:* I need R commands, not just the numerical answer.

11. Using the calculated standard error and the mean value, construct the 95% confidence interval of the sample mean of `gdp_10_thou`. For this, I need both R commands and the numerical answer.

12. Draw histograms of per capita GDP variable, one for democracies and the other for non-democracies.

    - Hint 1: Use the democ regime variable to classify the countries into democracies and non-democracies.
    - Hint 2: Use the facet wrap option. For this task, you may actually have three histograms (No, Yes, and NA), and that's OK. We will correct it below.

13. We find (I mean, I find) a few things about this graph unsatisfactory. First, it is a little bit aesthetically unpleasing that we have a blank graph on the far right. This happens because there are missing values. Second, the labels "No" and "Yes" are not intuitive at all (readers can't know what "Yes" and "No" mean simply by looking at the graph). So let's now fix these two things. Create a new data frame named dem.gdp that excludes those rows where the democ regime variable is missing. Use the is.na function for this. Then, create a new variable dem.dum within this new data set, which has two nominal values, "Democracy" and "Autocracy", instead of "Yes" and "No". Then, recreate the histograms you drew in 12. It should look like the following:



14. The graph above appears to suggest that democracies tend to have higher per capita GDP. Let's document this relationship by calculating the mean value of per capita GDP for each group. In doing so, report the 95% confidence intervals as well. For task 14, calculate the mean of per capita GDP for democracies, along with the 95% confidence interval. Please provide both the commands as well as the results (numbers).

15. Similarly, calculate the mean of per capita GDP for autocracies (non-democracies), along with the 95% confidence interval.

16. On a mild October morning in Columbia SC, you spot foreboding clouds gathering in the sky. Given that there's a 30% chance of rain on any day in October, and on the days it does rain, 95% of them begin with such clouds, but on dry days, 25% still experience dark clouds, what's the likelihood of rain after seeing dark clouds coming?

    - Hint 1: Now you have information on $Pr(R), Pr(C|R), Pr(C| \sim R)$.
    - Hint 2: Use Bayes rule to calculate your likelihood

17. Suppose a professor wants to know if students like his class, but he doesn't have a good sense (a weak prior). He decides to do a small survey of students who attended his class before to have a better understanding. Suppose 50 individuals are randomly chosen from a vast group of students used to join his class. They're asked a question with potential responses being "Yes, I liked it" or "No, I hated it." Let the proportion in the population who would answer "Yes" be θ. Also let the professor's prior distribution for θ to be a beta(1.5,1.5) distribution with know parameters. From the selected sample, 37 responded with "Yes" while 13 picked "No." What can we infer about θ after observing data?

    (a) Find the prior mean and prior standard deviation of θ
        - Hint 1: If X follows a distribtion of beta, that is X $\sim$ beta(a, b), then the mean of X is $E(X) = \frac{a}{a+b}$
        - Hint 2: The variance of X is var(X) = $\frac{ab}{(a+b+1)(a+b)^2}$

    (b) Find the prior probability that $\theta < 0.6$:
        - Hint: If $X \sim beta(a, b)$ then you can use a command such as the following in R to find $Pr(X < c)$: `pbeta(`θ`,a,b)`

    (c) Find the likelihood $f(x|\theta)$:
        - Hint: For a beta distribution, it is $\theta^{n\ yeses}(1-\theta)^{n\ nos}$

    (d) Find the posterior distribution of $\xi(\theta|x)$:
        - Hint: $\underbrace{\xi(\theta|x)}_{\text{posterior dist.}} \propto \underbrace{f(x|\theta)}_{\text{data/Likelihood}} \underbrace{\xi(\theta)}_{\text{prior dist.}}$
        - Hint: The probability density function of a $beta(a, b)$ distribution is $\propto \theta^{a-1}(1-\theta)^{b-1}$
        - Hint: The prior density of $\xi(\theta)$ is $\propto \theta^{1.5-1}(1-\theta)^{1.5-1}$

    (e) Plot a graph showing the prior and posterior probability density functions of $\xi(\theta|x)$ on the same axes.
        - Hint: To plot the prior and posterior probability densities you may use R commands such as the following.
        - ```
          theta<-seq(0.01,0.99,0.01)
          prior<-dbeta(theta,a,b)
          posterior<-dbeta(theta,c,d)
          plot(theta,posterior,xlab=expression(theta),ylab="Density",type="l")+
          lines(theta, prior,lty=2)
          ```

    (f) Explain what you have achievd by generating the plot.
          ```