GV207 2022

Instructor: Ryan Bakker

Homework Assignment 2

Due Week 7

For this assignment, I want you to write R code and save the file as gv207-HW2.R. The easiest way to write an R code file is to start with a pre-existing one (like the R Lab example code/practice exercises) and to modify the contents accordingly.

RULE:

-submit 2 files and 2 files only. That is, submit the coversheet (located in the HW tile on Moodle) and the R code file (gv207-HW2.R). You will earn 5 points if you do this correctly (submit 2 files!)

-Make sure your name does NOT appear in the R code. (5 points)

-Execute the entire file before you submit it and make sure the file runs without error. I will execute the file to check if you did this correctly. (5 points if your R file runs without error)

-Be sure to add comments (using the # symbol) to everything you do. Try to make your code file look like the example files in Moodle. Don't copy and paste the questions from the HW into your R file, but do show me the question number for each question. (5 points if done correctly)

TASKS (5 points each)

1. Load the "world" data set (world.csv) and store it as an object named world.data
2. The data set contains a dummy variable (categorical variable with 2 categories) named 'oecd' that classifies countries into 2 groups, OECD members and non-members. One way to describe and summarise the information contained in a nominal variable is to describe the distribution numerically. As we've learned during the past weeks, we describe a nominal/categorical variable numerically by creating a frequency table. Create a frequency table of this variable and store in into a data frame object called 'ft.oecd'. The table has to have 3 columns: values (initially called 'Var1', frequency (called 'Freq'), and percentage (should be called 'Percentage'). Change the name of the first column (Var1) to 'OECD Member?'.
3. According to the frequency table you created above, (A) how many countries in the data are OECD members? (B) How many countries in the data are not?. (C) what percentage of countries are OECD members. (D) what percentage of countries are not OECD members? Give me four answers as a comment (using #) Note—you do not need an R command for this bit. Just read the table and report the numbers. Don't forget to comment them out!!
4. Another way to describe a nominal variable is to draw a graph. For nominal variables, we use a bar graph. Using the package 'ggplot2' and the command geom_bar (as demonstrated in the R Lab/practice exercises), draw a bar chart of the OECD variable.
    a. Don't forget to load the package ggplot2 using the library function.
    b. 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", and for the y-axis would be "Number of countries"

5. 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 thisvariable takes a value

of 4, for e xample, 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.

6. 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.

7. 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".

8. 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 that I need R commands, not just the numerical answer.

9. 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.

10. Draw histograms of per capita GDP variable, one for democracies and the other for non-Democracies. Be sure to remove NAs first That is, create a new data frame named dem.gdp that excludes rows where the democ_regime variable is missing. (using the is.na function). Then create a new variable called dem.dum, which has 2 nominal values “Democracy” and “Autocracy” instead of “Yes” and “No”.

    --Hint 1: Use the democ regime variable to classify the countries into democracies and non-democracies.
    -- Hint 2: Use the facet wrap option.

11. 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 11, 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).

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