MODULE 3: DESCRIBING DATA! – PRACTICE.

7316 - INTRODUCTION TO DATA ANALYSIS WITH R

Mickaël Buffart (mickael.buffart@hhs.se)

1. Exercise 1. Replicate Figure 1

In this exercise, you will replicate Figure 1 from the paper “Female Socialization: How Daughters Affect Their Legislator Fathers’ Voting on Women’s”. Along the way, you will also replicate a figure. Important: The **README** file describes the variables in the dataset.

Figure 1 shows the mean NOW score (a score that measures how closely a politician’s votes align with recommendations of the National Organization of Women) by party and the number of female children.

- Make sure the relevant variables have the correct data type.
- Recode all the independent politicians to “Democrat” (there’s only one).
- Remove observations where the number of daughters is missing.
- NOW scores are only available for the 105^{th} congress. Select this subset.

Figure 1 contains interesting information but it could be more pretty. Let’s make it look a bit more colorful.

- Use *ggplot2* to replicate Figure 1 (with slightly different coloring)
    1. Filter the data for the subset of observation from the 105^{th} congress with all politicians that have either 2 or 3 children
    2. Create a ggplot object with that data set.
    3. Add an *aes* layer with the x dimension being the number of daughters and the y dimension being the total NOW score. Set the *col* and *fill* option to *party* because we want Republicans and Democrats colored differently.
    4. Add a geometry layer: use stat_summary, inside which you choose the function “mean” and the geometry “bar”. Also set *inherit.aes* to TRUE, such that the filling and shading parameters are taken over from the *aes* layer.
    5. Add a facet grid that splits the chart along *party* (column) and *totchi* (rows). Use the *margins* option to generate the bar chart that combines Democrats and Republicans.

6. Add a color layer using *scale_color_manual* where the breaks are the parties (incl. "(all)"), and the values are the colors you want to assign each party. Use the *aesthetics* option to specify that you want to change the color and the filling.

- Display the figure.

2. Exercise 2: Create a descriptive statistics table

Entirely with R, for the $105th$ congress subsample, create a descriptive statistics table including means and standard deviations of the independent variables displayed in Table 2. Ensure that the table is designed following reasonable academic standards and that the labels are correctly displayed.

3. Exercise 3: Replicate another figure

- In this exercise, you will replicate Figure 4 from the paper "A Passage to America: University Funding and International Students" (Bound et al., 2020) in the *American Economic Journal: Economic Policy*. You may find more detailed instructions on what to do below.
- Load the dataset named `pub_pvt_scatters.dta` and `univ_names.xls`.
- Merge `pub_pvt_scatters` with `univ_names` using the appropriate merge command (think about which observations you want to keep).
- Create a common deflator *cpi_all* which is just the mean of *cpi* across all universities in the same year.
- Create the real value of the appropriation by dividing *nominal_approp* by *cpi_all*
- Use the `Private` variable to create a categorical variable differentiating public and private universities.
- Reorder your data by unit and by year.
- Create the log of the variable `real_approp` and `ENROLL_FRESH_NON_RES_ALIEN_DEG`
- Create the difference in log values by university between 2005 and 2012 for these two variables and save it in a new `data.frame`.
- Replicate Figure 4 using *ggplot*, including all the features such as fitted lines, labels, and colors. If you want to make it prettier, feel free to be creative.

4. References

- Bound, J., Braga, B., Khanna, G., & Turner, S. (2020). A passage to America: University funding and international students. *American Economic Journal: Economic Policy*, 12(1), 97-126.
- Washington, E. L. (2008). Female socialization: how daughters affect their legislator fathers. *American Economic Review*, *98*(1), 311-32.