Oeconomica Research Workshop 1: Summary Statistics in Research (R)

In this meeting, you and your group will be working with R to understand and utilize summary statistics in an economic context. Good luck and have fun!

Nominal, ordinal and interval variables/data

There are different kinds of data/variables that economists work with. One way we can think about them is nominal, ordinal and interval/ratio.

*What is a nominal variable?*

*What is an ordinal variable?*

*What is an interval/ratio variable?*

Now you will work with R. After having downloaded R Studio

We will start out with making sure we understand the nature of the variables and data we are working with. In R, you should see 5 different variables. Feel free to rename your variables.

*Fill out the table for each variable, both in type and measurement:*

|  |  |  |
| --- | --- | --- |
| Variable | What does the variable measure? | Variable Type (i.e. Ordinal, Nominal, or Interval/Ratio) |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Many economists and researchers use something called *operationalization* to transform abstract concepts that are interesting (ex: happiness) into something measurable (ex: survey asking “how happy are you”?).

*What do you think the operationalization was with these variables? I.e. what do you think the data is trying to measure? What do you think it actually measures?*

Data analysis and research with nominal and ordinal variables poses a problem because they are not continuous (mathematical definition). Another way to think about it is when cannot describe a mathematical relationship or distance between the gender male and female, for example.

*How can we address nominal variables in general data analysis? What should we do in R? Do that! (Hint: You only really need to address the president party variable; ask a board member for help if you want.)*

Now that we have the variables as we want them, we will summarize them. Let’s go over what summary of data is.

Summarizing Data

Before we actually compare variables (the nitty gritty and fun part of research and data analysis), we need to understand what we are working with.

*Fill out the following table, putting appropriate summary statistics for each variable indicating the measure:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Measure of Center | Measure of Spread | Minimum | Maximum |
| EX: Variable 1 | 15 (mean) | 2 (SD) | 0 | 28 |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

*Provide the relevant graph for each of your variables below (it will either be a histogram or box plot). This will get you practice with exporting images/graphs from R!:*

What’s the point of summarizing data?

This is a good question; big time economists rarely summarize data in their papers – it is basically assumed that the economists know what they are working with and that variables/data are sufficient for the analysis being presented. That being said, it is still extremely important and something everyone needs to know – it can lead to big problems if you don’t understand your variables well enough.

*What did you learn about your variables based on summary statistics? Anything interesting?*

*What did you learn about your variables based on your graphs that you couldn’t get in the summary statistics?*

*In general, what conclusions regarding might you have after this process?*

Other/extra:

Cleaning data, removing outliers, changing data specifications, scatterplots, introduction to relationships between variables (correlations, tablature)