Introduction to Stata

Lecture V and VI

Tomas R. Martinez

UC3M

September, 2019

Quote of the Day

"I'm not an outlier. I just haven't found my distribution yet." - Unknown

Describing the Data

- "I'm not an outlier. I just haven't found my distribution yet." - Unknown
- Now let's learn how to inspect the data
- Also, some statistical routines
- **Example:** CPS 2005 (US labor survey data)

Describing the Data

- Description of the entire data
- **describe**, **codebook** and **summarize**
- But the real power is when we use it in specific variables

Describing the Data

- Probably the two of the most used commands: **tabulate** and **summarize**
- **tabulate** varname [if] $\rightarrow$ frequency table
- **summarize** varlist [if] $\rightarrow$ basic statistics (use the details option!)
- Very powerful when combined with [if] and **by**!
- Allow sample weights
- Other describing commands: **count, tabstat, tab1, table**
- Do some examples

Describing the Data

- You might want to save some statistics
- Stata save internally some of these results
- **Example:** summarize incwage, details
- Write *return list* to see how to have access
- Then you can use it: *dmean_wage* = *incwage* - *r(mean)*
- Check whether this new variable has mean 0!

Statistics and tests

- All the basic statistics have easy to use routines in Stata
- Correlations: **correlate var1 var2**
- Spearman test of correlations: **spearman var1 var2**
- T-test: **ttest varname=null**
- Chi-Square: **tabulate var1 var2, chi2**

Statistics and tests

- T-test is particularly powerful and has many options
- Choose you
- You can test whether the mean is equal to some value: **ttest varname=null**
- Also test whether different groups have different means: use the option **by(varname)**
- Usually with **unequal** to take into account different variances

Manipulating Multiple Data sets

- One limitation by Stata is to have only one data in your memory
- If your data is not related you can just have two different do-files, use preserve-restore, and etc...
- When your data is related you can just put everything together
- $\rightarrow$ **append**, **merge** and **joinby**
- We will focus in the first two

Append

- Let's say you have two data where one has the same subset of variables than the other
- Combining both data $\rightarrow$ **append**
- It is a relatively simple command:
  - Open your data1.dta
  - To append data2.dta just write **append using filename**

Merge

- Combine two data sets with different information shared by the same "key"
- **Example:** One data set has GDP and the other has population for the all countries (key: country)
- **Example:** Or you have the GDP growth in one and unemployment rate for the years 1990-2010 (key: year)
- merge 1:1 *varlist* using *filename*
- This merges the data *filename* with the current data based on the keys *varlist*

Merge

- Combine two data sets with different information shared by the same "key"
- **Example:** One data set has GDP and the other has population for the all countries (key: country)
- **Example:** Or you have the GDP growth in one and unemployment rate for the years 1990-2010 (key: year)
- merge 1:1 *varlist* using *filename*
- This merges the data *filename* with the current data based on the keys *varlist*

Merge

- What if the keys are the same for many observations?
- **Example:** One data set is household survey and the other you have State data
- merge **m:1** *state* using *statedata.dta*
- If the merging data has many observations for each key value: **1:m**

Merge

- After the merge, the command creates one variable: _merge
- It is useful to inspect this variable to check whether your data has the expected structure
- Check the options! **help merge**
- You may want to use: **nogen, update, replace, assert, keep** are all very useful

Exercise

1. Open the data *cps05.dta* and append the survey years: 06, 07, 08, 09, save your data
2. Take a look at the data index_ONET.dta. It has three measures of task content for different occupations: routine, abstract and manual. The highest is the measure more "routine / abstract / manual" tasks that occupation requires
3. The variable *occ1990dd* is a recode occupation by David Dorn
4. Merge the ONET.dta with your newly combined CPS data
5. Use the variable _merge to check whether the operation was successful

Exercise

1. Select only individuals between 18 and 65 years old
2. Which gender perform more "routine tasks"? What about "manual" and "abstract"?
3. Test whether the average routine content of occupations by men is the same as women.
4. Does "routine" occupations correlate with lower or higher wages? What about 'abstract"?
5. Calculate the average wage by occupation, what is the highest paid occupation?
6. What is the gender distribution in that occupation?