Introduction to Stata

Ista Zahn

Harvard MIT Data Center

The Institute *for* Quantitative Social Science at Harvard University

Outline

Topic

Documents for today

USERNAME: dataclass PASSWORD: dataclass

- Find class materials at `http://j.mp/intro-stata`
- Download and extract to your desktop!

Organization

- Please feel free to ask questions at any point if they are relevant to the current topic (or if you are lost!)
  - There will be a Q&A after class for more specific, personalized questions
- Collaboration with your neighbors is encouraged
- If you are using a laptop, you will need to adjust paths accordingly
- Make comments in your Do-file rather than on hand-outs
  - save on flash drive or email to yourself

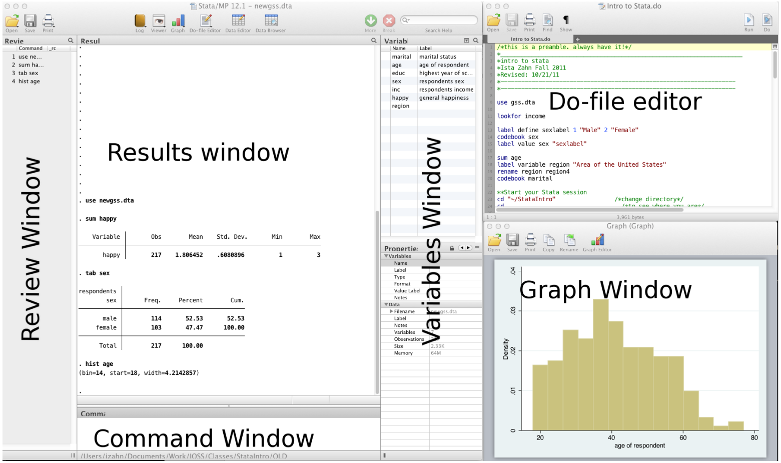
Workshop descripton

- This is an introduction to Stata
- Assumes no/very little knowledge of Stata
- Not appropriate for people already well familiar with Stata
- Learning Objectives:
    - Familiarize yourself with the Stata interface
    - Get data in and out of Stata
    - Compute statistics and construct graphical displays
    - Compute new variables and transformations

Why stata?

- Used in a variety of disciplines
- User-friendly
- Great guides available on web (as well as in HMDC computer lab library)
- Student and other discount packages available at reasonable cost

Stata interface



```
1  use ne...
2  sum ha...
3  tab sex
4  hist age
```

```
. use newgss.dta

. sum happy

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       happy |       217    1.806452    .6080896          1          3

. tab sex

respondents |
        sex |      Freq.     Percent        Cum.
------------+-----------------------------------
       male |        114       52.53       52.53
     female |        103       47.47      100.00
------------+-----------------------------------
      Total |        217      100.00

. hist age
(bin=14, start=18, width=4.2142857)

.
```

Name	Label
marital	marital status
age	age of respondent
educ	highest year of sc...
sex	respondents sex
inc	respondents income
happy	general happiness
region	

```
/*this is a preamble. always have it!*/
*_______________________________________________________________________
*intro to stata
*Ista Zahn Fall 2011
*Revised: 10/21/11
*-----------------------------------------------------------------------
*-----------------------------------------------------------------------

use gss.dta

lookfor income

label define sexlabel 1 "Male" 2 "Female"
codebook sex
label value sex "sexlabel"

sum age
label variable region "Area of the United States"
rename region region4
codebook marital

**Start your Stata session
cd "~/StataIntro"                 /*change directory*/
```

- Review and Variable windows can be closed (user preference)
- Command window can be shortened (recommended)

Do-files

- You can type all the same commands into the Do-file that you would type into the command window
- BUT. . . the Do-file allows you to save your commands
- Your Do-file should contain ALL commands you executed – at least all the "correct" commands!
- I recommend never using the command window or menus to make CHANGES to data
- Saving commands in Do-file allows you to keep a written record of everything you have done to your data
    - Allows easy replication
    - Allows you to go back and re-run commands, analyses and make modifications

Stata help

- Easiest way to get help in Stata - just type help followed by topic or command, e.g., help regress
- Falls back to "search" if command not found
- Generally, if you google "Stata [topic]," you'll get some helpful hits
- UCLA website: `http://www.ats.ucla.edu/stat/Stata/`

General Stata command syntax

- Most Stata commands follow the same underlying principles Command varlist, options, e.g., sum var1 var2, detail
  - CAUTION - in some cases, if you type a command and don't specify a variable, Stata will perform the command on all variables in your dataset
- You can find command-specific syntax in the help files

Commenting and formatting syntax

- Start with comment describing your Do-file and use comments throughout
- Single line and block comments

```
// comment
 describe var
 /*
 comment block comment block comment block comment
 block comment block comment block
*/
```

- Use / to break varlists over multiple lines:

```
// break commands over multible lines
describe var1 var2 var2 ///
var4 var5 var6
```

Let's get started

- Launch the Stata program (MP or SE, does not matter unless doing computationally intensive work)
  - Open up a new Do-file
  - Run our first Stata code!

```
// change directory
cd "C://Users/dataclass/Desktop/StataIntro"
// start a log file to record your stata session
log using myStataLog.txt, text replace
// Pause / resume logging with "log on" / "log off"
// Close lot with "log close"
```

How to start every do-file

1. Describe what the file does
2. Change directory
3. Begin log file
4. Call up data
5. Do stuff: Data manipulation, statistics etc.
6. Save data under new name (if making changes to dataset)

Topic

Data file commands

- Next, we want to open our data file
- Open/save data sets with "use" and "save":

```
cd dataSets
// open the gss.dta data set
use gss.dta
// saving your data file:
save newgss.dta, replace
// the "replace" option tells stata it's OK to
//   write over an existing file
```

A note about path names

- If your path has no spaces in the name (that means all directories, folders, file names, etc. can have no spaces), you can write the path as is
- If there are spaces, you need to put your pathname in quotes
- Best to get in the habit of quoting paths

Where's my data?

- Data editor (browse)
- Data editor (edit)
    - Using the data editor is discouraged (why?)
- Always keep any changes to your data in your Do-file
- Avoid temptation of making manual changes by viewing data via the browser rather than editor

What if my data is not a Stata file?

- Import delimited text files

```
/* import data from a .csv file */
insheet using gss.csv, clear
/* save data to a .csv file */
outsheet using gss_new.csv, replace comma
```

- Import data from SAS and Excel

```
/* import/export SAS xport files */
clear
import sasxport gss.xpt
export sasxport newFileName, replace

/* import/export data from Excel */
import excel using gss.xlsx, firstrow clear
export excel newFileName.xls, replace
```

What if my data is from another statistical software program?

- SPSS/PASW will allow you to save your data as a Stata file
    - Go to: file > save as > Stata (use most recent version available)
    - Then you can just go into Stata and open it
- Another option is StatTransfer, a program that converts data from/to many common formats, including SAS, SPSS, Stata, and many more

Exercise 1: Importing data

1. Close down Stata and open a new session
2. Go through the three steps for starting each Stata session that we reviewed
   - Begin a log file
   - Open your Stata dataset (gss.dta)
   - Save your Stata dataset using a different name
3. Try opening the following files:
   - A comma separated value file: gss.csv
   - A SPSS file: gss.sav
   - A SAS transport file: gss.xpt

Topic

Basic graphing commands

- Univariate distribution(s) using hist

```
/* Histograms */
hist educ
// histogram with normal curve; see 'help hist' for other options
hist age, normal
```

- View bivariate distributions with scatterplots

```
/* scatterplots */
twoway (scatter educ age)
graph matrix educ age inc
```

The "by" command

- Sometimes, you'd like to generate output based on different categories of a grouping variable
- The "by" command does just this

```
/* By Processing */
// tabulate happy separately for men and women
bysort sex: tab happy
// summarize eudcation by marital status
bysort marital: sum educ
```

Exercise 2: Descriptive statistics

1. Use the dataset, gss.dta
2. Examine a few selected variables using the describe, sum and codebook commands
3. Tabulate the variable, "marital," with and without labels
4. Summarize the variable, "income" separately participants based on marital status
5. Cross-tabulate marital with region and show gender percent by region
6. Summarize the variable, "happy" for married individuals only
7. Generate a histogram of income
8. Generate a second histogram of income, but this time, split income based on participants sex and ask Stata to print the normal curve on your histograms

Topic

Labels

- You never know why and when your data may be reviewed
- ALWAYS label every variable no matter how insignificant it may seem
- Stata uses two sets of labels: variable labels and value labels
- Variable labels are very easy to use – value labels are a little more complicated

Variable and value labels

- Variable labels

```
/* Labelling and renaming */
// Label variable inc "household income"
label var inc "household income"

// change the name 'educ' to 'education'
rename educ education

// you can search names and labels with 'lookfor'
lookfor household
```

- Value labels are a two step process: define a value label, then assign defined label to variable(s)

```
/*define a value label for sex */
label define mySexLabel 1 "Male" 2 "Female"
/* assign our label set to the sex variable*/
label val sex  mySexLabel
```

Exercise 3: Variable labels and value labels

1. Open the data set gss.csv
2. Familiarize yourself with the data using describe, sum, etc.
3. Rename and label variables using the following codebook:

var	rename to	label with
v1	marital	marital status
v2	age	age of respondent
v3	educ	education
v4	sex	respondent's sex
v5	inc	household income
v6	happy	general happiness
v7	region	region of interest

1. Add value labels to your "marital" variable using this codebook:

value	label
1	"married"
2	"widowed"
3	"divorced"
4	"separated"
5	"never married"

Working on subsets

- It is often useful to select just those rows of your data where some condition holds–for example select only rows where sex is 1 (male)
- The following operators allow you to do this:

==	equal to
!=	not equal to
>	greater than
<	less than
>=	greater than or equal to
<=	less than or equal to
&	and
|	or

- Note the double equals signs for testing equality

Generating and replacing variables

- Create new variables using "gen"

```
// create a new variable named mc_inc
//   equal to inc minus the mean of inc
gen mc_inc = inc - 15.37
```

- Sometimes useful to start with blank values and fill them in based on values of existing variables

```
/* the 'generate and replace' strategy */
// generate a column of missings
gen age_wealth = .
// Next, start adding your qualifications
replace age_wealth=1 if age<30 & inc < 10
replace age_wealth=2 if age<30 & inc > 10
replace age_wealth=3 if age>30 & inc < 10
replace age_wealth=4 if age>30 & inc > 10

// conditions can also be combined with "or"
gen young=0
replace young=1 if age_wealth==1 | age_wealth==2
```

Exercise 4: Manipulating variables

1. Use the dataset, gss.dta
2. Generate a new variable, age2 equal to age squared
3. Generate a new "high income" variable that will take on a value of "1" if a person has an income value greater than "15" and "0" otherwise
4. Generate a new divorced/separated dummy variable that will take on a value of "1" if a person is either divorced or separated and "0" otherwise

Topic

Help us make this workshop better!

- Please take a moment to fill out a very short feedback form
- These workshops exist for you – tell us what you need!
- `http://tinyurl.com/6h3cxnz`

Additional resources

- IQSS workshops:
  http://projects.iq.harvard.edu/rtc/filter_by/workshops
- IQSS statistical consulting: http://rtc.iq.harvard.edu
- The RCE
  - Research Computing Enviroment (RCE) service available to Harvard & MIT users
  - www.iq.harvard.edu/research_computing
  - Wonderful resource for organizing data, running analyses efficiently
  - Creates a centralized place to store data and run analysis
  - Supplies persistent desktop environment accessible from any computer with an internet connection