Introduction to Stata

August 27, 2014

The Basics

Most of the commands that you will need to know in Stata are available through the use of drop-down menus and also through typing commands. We will emphasize typing commands because it is a quicker way to get the answer you want and often times more straightforward; but the drop down menus are also an option if you feel more comfortable working with them.

Note: To denote use of drop-down menus and their sub-menus and items in this lab session, we will write: *Menu/Sub-menu/Menu Item*.

Getting Started

After you have logged onto the computer, point to and click on the Start button, then on "All Programs" at the bottom, then on the "Stata 13" folder, and then on the file "StataSE 13 (64-bit)"

Stata Windows

The Stata interface is divided into five windows, the first four of which appear when you open Stata. (Note: if they are not all there click on *Prefs/Default Windowing*):

1. *Command* Window

   This is where you type in commands into Stata in order to perform statistical analyses, plot graphs, etc. However, many of the commands that you will use in this course are available in the *Data*, *Statistics* and *Graphics* menus found on the toolbar at the top of the Stata screen. Using these menus will eliminate having to type the commands into the command window.

2. *Results* Window

   This is where the results of any command performed by Stata will appear.

3. *Variables* Window

   When you have a dataset open, or if you have created a new dataset, the variable names will be listed here.

4. *Review* Window

   The *review* window is useful because as you send commands to Stata they appear there even if you use the *Statistics* and *Graphics* menus to perform your analyses. If you are interested in learning the command-driven version of Stata,

then noticing the commands in the *review* window is a good place to start. If you are feeling ambitious, you can also call back the commands in the *review* window by clicking on them, then editing them in the *command* window, then running them by hitting the Enter key.

5. *Graph* Window

   This window appears when you create a graph. The resulting graph can then be included in a word-processed document, which we will talk about later.

Stata Menus

As mentioned earlier, most of the commands that you will use in this course are available through the use of drop-down menus. The menus can be found on the Menu bar at the top of the Stata window. The following is a description of the Stata menus:

- *File* – Opens and saves Stata data files. Opens and closes Log files. Saves or prints graphs. Imports and exports ascii files, exits Stata.
- *Edit* – Allows you to copy output from the *Results* or *Graphs* windows to a word processor or other application.
- *Data* – Open the Data Editor and Data Browser. Summarize data. Label datasets and variables. Replace and generate data.
- *Graphics* – Contains all of Stata's graphing tools.
- *Statistics* – Data summaries and all statistical tests.
- *User* – Place to store any user-generated commands.
- *Window* – Controls the windows opened in Stata.
- *Help* – A good resource if you have questions about how to use Stata.

Opening Existing Stata Data Files

If you are working in the MicroLab, you can get Stata data files for BIO 200 from the g: drive at *g:\SHARED\BIO200\Stata data for Prin of Bio Ed 2*. Note that any changes you make to the file will *not* be saved to the g: drive.

Stata data files have the extension .dta. All of the data files you will need are contained in the *g:\SHARED\BIO200\Stata data for Prin of Bio Ed 2\exercise* folder. There is also a folder for each chapter in the textbook. These folders contain the data sets used in the examples in the book, appearing by chapter. Raw text files can be opened in Stata by clicking on *File/Import ASCII.*

The data set named "bed.dta" contains data with the number of nursing home beds per 1000 population for each state in the US in 1980 and 1986. The number of beds in 1980

is stored in the variable name bed80 and for 1986 in bed86. The state identifiers are stored in state.

◊ Open the data set “bed.dta” using *File/Open* and selecting the correct file. (located in *g:\SHARED\BIO200\Stata data for Prin of Bio Ed 2\exercise*)
- You will see the list of variable names in the data set appear in the variable window.
- Also note that the command
  use "*G:\SHARED\BIO200\Stata data for Prin of Bio Ed 2\exercise* bed.dta", clear

appears in the *Results* and *Review* windows. Typing this command into the *Command* window instead of using *File/Open* would have also opened the data set. We will now look at the data in the *Data Editor*.

Data Editor

The Data Editor is where you can enter new data, make changes to the current dataset, or create new variables.

◊ Open the Data Editor by clicking the Data Editor button on the toolbar at the top of the screen. The button looks like a little spreadsheet with a pencil. Alternatively you could choose *Data/Data editor* from the menu bar.

Note: The Data Editor is different from the Data Browser. The Browser only allows you to look at the data and doesn’t allow you to make changes to the dataset. The button for the Data Browser looks like a little spreadsheet with a magnifying glass, which is also located on the toolbar.

◊ Sort the data on the variable bed80.
- Select the *Data/Sort* in the Data Browser
- Select bed80 in the Variables box, then click OK. Watch how the order of the observations changes.

Inputting Data into Stata

◊ Now close the Data Editor and clear out the current dataset by typing clear in the *command* window. What happened in the *variable* window?

◊ Refer to the data in the table on the right. It contains observations of gestational age and birth weight for 10 infants. We will manually input this data into Stata.

◊ Open the Data Editor and type in the values of gestational age into the first column. Assign the

Gestational Age (weeks)	Birth Weight (grams)
39	3544
39	3430
41	3770
35	2522
39	3997
36	3530
41	3402
44	4355
39	2977
39	3474

variable a name by double-clicking where it says “var1”. Type gestage for the variable name.
Note that Stata variable names need to be less than 32 characters and cannot start with a number.

◊ Enter the values for birth weight and name the variable weight.

◊ Now rather than use the Preserve button, click on the close (‘X’) button on the window’s upper right corner.

◊ This data file can be saved for future use by clicking *File/Save As*. Practice saving this dataset to disk calling it “baby.dta”. Note that if you save it to your folder in the p: drive, you will be able to access this dataset from any computer in the MicroLab.

Labels

For the sake of clarity you may want to give labels to your data. Labels can be given to the entire dataset or to the individual variables. This is helpful when you have given abbreviated names to your dataset and/or variables and want to keep a more detailed explanation of the dataset’s/variables’ contents. For example, you might want to label the variable gestage with “Gestational Age (weeks)”. You can label the data at the same time you give the variable a name (when you double click on “var1”), but you can also label data using the command window.

◊ Assign a label to the baby dataset by typing

label data “<data title>”

in the *Command* window and give an appropriate title in place of <data title> (e.g. Introduction to Stata Data).

◊ To label a variable type

label variable varname “<variable label>”

An example for the variable gestage:

label variable gestage “Gestational Age (weeks)”

These labels will appear in any tables or graphs that you make with this data. Now label the variable “weight.”

◊ Type describe in the *Command* window.

◊ You will see a table of information appear in your *Results* window. Generally this command provides the number of observations and variables in the dataset, the storage and display type for each variable and any special labels. There is (potentially) a label for the entire dataset, a label for each variable, and special label values used in storing a variable.

◊ Once you know your labeled baby dataset is saved, open the “bed.dta” dataset and look at its thorough labeling.

Log Files

Now that you have data to work with, you will want to do some statistical analyses. The log file is where you save all the work you do during your Stata session. Basically, as long as a log file is open everything that appears in the *Results* window is saved in the log file. Graphs, however, are not saved to the log file.

◊ Create a log file by clicking on the button that looks like a spiral notebook. Alternatively, you could pull down *File/Log/Begin*.

◊ In the window that pops up, enter a file name, for example you could call it "session1." Choose the *.log* file type, and click on Open. The *.log* extension is used so that the log file can be opened in most word processing applications.

◊ Look at what appears in the *Results* window.

Once you have opened a log file, it will record everything you do until you suspend it or close it. You can do this by clicking on the same button you clicked on to create the file, then selecting the appropriate option. When you choose the close option, you close the log file and have to open a new one before you can start recording again. However, with the suspend option you pause recording until you click the log button again, at which point Stata starts recording in the *same* log file. You can view the current contents of the log file by choosing the *snapshot of log file* option.

◊ Leave the log file open and go on to the next step. We will come back to saving and printing it after it has recorded something. Note also that you may want to just include portions of your work in a Word document (see Printing, below).

Summaries

In this section, you will calculate some descriptive statistics for this data using various Stata commands.

◊ Open the dataset "lowbwt.dta," which is also found in the *Exercise* folder on the g: drive

(*G:\SHARED\BIO200\Stata data for Prin of Bio Ed 2\exercise*).

◊ Describe the variables in the dataset (use the command describe).

◊ What is the mean systolic blood pressure (sbp) for this group?

- Type summarize sbp in the *Command* window.

◊ What is the standard deviation of sbp?

◊ What is the mean gestational age for boys?

- Here you have to type by sex, sort: summarize gestage, detail.
- You may have to hit the spacebar to scroll through all the entries.

◊ How many girls are there in this data set?

- Use the command tabulate sex.

◊ How many women who gave birth to boys experienced toxemia (tox)

- Now use tabulate sex tox.

◊ Now suspend your log file by clicking the Log button.

◊ List all values for sex and apgar5 using *list sex apgar5*. Hit the space bar in order to scroll through the observations.

Printing

To avoid wasting paper and added noise in lab, do not print any of the session now. Thanks!

Here are some steps to follow: View the current log snapshot (by clicking on the log button), and select a portion of text from your results. Select *Edit /Copy Text*. Switch to an open Microsoft Word document, place text cursor at your target location, and choose *Edit / Paste*. Notice that text does not necessarily have the same appearance. Highlight the portion you pasted, and change the font, making sure you use a *fixed-width* font such as Courier New.

◊ Resume suspended log file

Creating New Variables with Formulas

Sometimes you may need to create a new variable that is a function of one or more existing variables. You will see later your courses why you may want to do this.

◊ Create a new variable called *gestsq* that is the square of gestational age (gestage^2).

- Type generate gestsq = gestage^2.

Graphs

All of the graph options in Stata are located in the *Graphics* menu.

◊ Still using the lowbwt.dta data, create a box plot for sbp.

- Select *Graphics/Boxplot.*
- Click in the box under “Variables”. Type sbp or select it from the Variables window and hit “OK”.

◊ Create a histogram for gestage (consider it a continuous variable).

- Select *Graphics /Histogram.*
- Click in the box under “Variables”. Type gestage or select it from the Variables window and hit “submit”.

◊ Now create separate histograms of gestage for boys and girls. This can be done under *Graphics/Histogram* menu. Separate histograms can be plotted using the “By” option tab and choosing sex as the variable. How do the histograms differ between the two genders?

◊ Create a scatterplot of gestage vs. sbp. Is there a relationship between the two variables?
- Here you have to type scatter gestage sbp.
- Alternatively, select *Graphics/Twoway graph (scatter, line, etc.)*, then click create and in the menu select scatter and choose the X and Y variables.

Note:
- What happens when you hit submit?
- What happens when you hit ok?

Saving Graphs

While the graph window is active, under the File menu you have the option to either print it or save it (you can recall the saved graphs by *Graphs / View saved graphs*). Again, note that graphs do *not* appear in your log file. You can therefore copy and paste graphs to word processing programs by choosing *Copy Graph* from the *Edit* menu while the graph window is active and then pasting it into a Word document.

- Practice cutting and pasting your scatterplot into your open Word document.

Before exiting from Stata, close your open Log file by clicking on the Log button and selecting "Close log file".

Congratulations! You have successfully completed your Introduction to Stata.

In case you are uncomfortable typing commands in Stata here is a summary of menu paths to the same commands:

For labeling data sets and variable (using the same data as earlier in the lab for an example):

◊ Assign a label to the baby dataset by clicking on *Data/Data Utilities/Label Utilities/Label dataset*, and in the resulting dialog box type in an appropriate label.

◊ Use the *Label variable* option under the *Data/Data Utilities/Label Utilities* sub-menu to label your variables (it's also done by double-clicking column headings when you're in the Data Editor). Move the cursor and click in the box under "Variable". Type gestage or select gestage from the *Variables* window. Enter an appropriate variable label. Do the same thing for the variable "weight".

◊ Now pull down the *Data/Describe data/Describe variables in memory or in a file* menu. The "describe" dialog box will appear. Click "OK" in order to describe all the variables in the dataset.

To get summary statistics about variables:

For basic summary statistics:

- Click on the *Statistics/Summaries, tables, and tests/Summary and descriptive statistics/Summary statistics* sub-menu.
- Click in the box under “Variables”.
- Select the variable from the *Variables* window

For more summary statistics (including median):

- Repeat as above but choose the “Display additional statistics” option.

To break down the summaries by some other variable (such as gender):

- Here you will have to go to the “by/if/in” tab of the *Summary statistics* pop-up menu.
- Choose the “Repeat command” option and type sex (or other variable of interest) in the window.

To get counts of variables (for example gender):

- Click on *Statistics/Summaries, tables, and tests/Frequency tables/One-way tables* to get the One-way table dialog window.
- Type sex in the box under “Categorical Variable” and hit “OK”.

To create a table of counts broken down by more than 1 variable (for example sex and toxemia status):

- Now use *Statistics/Summaries, tables, and tests/Frequency tables/Two-way tables with measures of association*.
- In the resulting dialog window, select sex as your row variable and tox as the column variable.

To create new variables with formulas (using the same example as from the lab):

- Select *Data/Create or change data/ Create new variable* to bring up the “Generate a New Variable” dialog box.
- Type gestsq into the box under “Generate Variable”.
- Type gestage^2 into the “Contents” box.

Session Notes

MENUS

- File
  - Open
  - Save Graph
  - Print Graph
  - Log
    - Begin…
  - Print Results

- Edit
  - Copy Text
  - Copy Graph

- Data
  - Describe data
    - Describe variables in memory
  - Data editor
  - Labels & notes
    - Label dataset
    - Label variable
    - Label values
  - Create or change variables
    - Create new variable

- Graphics
  - Easy graphs
    - Scatter plot
    - Box plot
    - Histogram

- Statistics
  - Summaries, tables, & tests
    - Summary Statistics
      - Summary Statistics
    - Tables
      - One-way tables
      - Two-way tables with measures of association

WINDOWS

- Command
- Results
- Variable
- Review
- Graph
- Log

BUTTONS

- Data Editor
  - Preserve
  - Restore
  - Sort
  - Delete
- Data Browser