Introduction to Stata

Biostat 511 - Fall 1999

Table of Contents

	pg.
Accessing Stata ..……..	1
Entering commands; StataQuest ….……..	1
Getting Help …...……..	2
Log files ……	2
Inputting and Saving Data ……	2
Documenting a Dataset ……	4
Creating and transforming variables ……	4
Summarizing data ……	4
Graphics ……...	5
Getting Output ……..……..	5
A sample Stata session ……..	6

Accessing Stata

Stata can be acquired for home use (see "Aquiring Stata" on the class web page) or it may be accessed at the Health Sciences Computer Microlab (3rd floor T-wing, Health Sciences building). The following instructions will assume you are working at the Microlab.

To start Stata, click on "Intercooled Stata" from the Windows Start … Programs menu (or click on Start … Run and type in H:\stata\wstata.exe). When you start up Stata you get one large "Stata" window. Contained within that are 4 smaller windows. These smaller windows include a command window (for typing in commands), a command history window (listing the commands you have entered), a data window (listing the variables in the current dataset) and a results window (giving results of commands). Additional windows, such as a help window, a log window (see log files, below) and a graphics window are created as needed by the system.

To exit Stata, type

exit

In the command window.

Entering commands; StataQuest

Stata is primarily a command driven package, meaning that you must type in each command. Such packages tend to be more flexible than packages that rely soley on menus but the problem is that you need to know the syntax of key commands before you can get anything done. This can slow you down when you first start using the program. However, all is not lost! There is an add-on package called "StataQuest" that adds a partial menuing system onto Stata. By "partial menuing system" I mean that not all Stata commands are available through the StataQuest menus. However, there is enough there to get you started and to do simple tasks. If you have purchased Stata for your personal computer, you can download the StataQuest additions from the Stata website (www.stata.com). The StataQuest additions have already been added to the Health Sciences Microlab computers.

To start up the StataQuest menu system type

quest on

in the command window. If you want to turn the StataQuest menus off, type

quest off

Getting Help

Stata comes with the following documentation (which I think are available at the Microlab). The User's Guide is on reserve at the Health Sciences library reserve desk.

Getting Started [GSW]
User's Guide [U]
Reference Manual (4 vols) [R]
Graphics Manual [G]

Stata also has an on-line help system that can be accessed by choosing Help from the menu bar. When you click on Help you get the following submenu

Contents	See the help table of contents
Search…	Search for help on a particular topic
Stata Command…	Get help for a particular Stata command
:	
:	

If you are a new user it is useful to browse through the table of Contents just to get an idea of what the package can do. Search is useful if you have a general topic you are interested in but don't know the actual command.

Each command's help page uses three colors: blue for the command's syntax; black for documentation/explanation of the command and green for references to other related commands/topics. When you see a word in green you can click on that and you will see the help screen for that word.

Log files

You should usually start a log file when using Stata. This keeps a record of everything you type and everything Stata displays so you can print out the results. You can edit the log file and use it to create "do" files that can be used to repeat commands at a later date. There is a Log button at the top left of the Stata window (beneath the menu bar; it looks like a scroll). Click on this button. You will now be prompted to choose a name and location for your log file. The default folder is probably labeled Data, which is a folder belonging to the Stata program. You should change this to your own folder. Click on the small triangle to the right of the folder name to get a list of choices and select Hard Drive (C:), then My Documents, then your own folder. Now click on the box labeled File Name and type the name for your log file (eg tutorial.log). Click Ok to continue.

A log window will now appear. To make it go away, click the cross button at its top right. This does not turn off the log file. To close the log file or to temporarily suspend logging click on the Log button again.

Inputting and saving data

The basic unit of analysis in Stata is the dataset, which consists of a number of variables. You have one dataset active at a time. The names of the variables in the current dataset are given in the Variables window. To get data into Stata you can type it in using the data editor, read the data in from a text file, or recall a previously saved Stata data file (typically with extension .dta). These options are described further below. You can also create data with the generate command. Once data has been entered, you typically will want to save the data. To save the current dataset to disk, type

save "<path\filename>"

on the command line or

save "<path\filename>", **replace**

if filename already exists or choose File …. Save As from the menu. If <path> is not given then the data is saved to the Stata Data directory (usually C:\DATA). If you are working on the Health Science Microlab system you need to change <path> to your temporary folder e.g.

save "c:\temp\fev"

To clear a dataset from memory type

clear

on the command line. If the dataset has not been previously saved to disk, this will erase the data permanently.

I. Typing in data

To enter data manually into Stata, type

edit

in the command window or choose Data … Edit data from the StataQuest menu. This starts the Stata data editor. The data editor uses a spreadsheet format with columns representing the variables and rows representing the observations in your dataset. Simply type the data into the spreadsheet. Strings do not need to be quoted and missing values are entered as a period (.) for numeric data or blank for character data. The default names for the variables (the columns) are var1, var2, etc. To change any variable's name, double-click on any entry in that column. You will also be given the opportunity to enter a label for the variable (see Documenting a Dataset, below).

II. Inputting from a text file

Often data will be available in the form of a text file. Stata has 3 command to read text files. These are

insheet <var1> <var2> …. <varn> **using** <path\filename> * for comma or tab delimited
infile <var1> <var2> …. <varn> **using** <path\filename> * for space-delimited free format
infix … * for fixed format

where <var1>, etc are variable names. See the Reference manual, User's Guide, or online help for details. In general, strings or character data should be quoted (there are exceptions) and missing values are input as period (.) for numeric data and blank ("") for character data. Once you have input the data, you can view it with the

browse

command. When you are in browse, you cannot alter the data. Use edit if you want to change the data using the data editor. To close the editor, click on the X in the upper right hand corner.

III. Using an existing Stata dataset

To open an existing Stata dataset type

use <path\filename>

or File … Open from the menus.

Documenting a Dataset

Usually, you will want to make variable names short since you will be typing them a lot (e.g. wt for weight). However, it is convenient to have more complete descriptions of the data attached to the dataset so that when you come back a year latter you can remember what everything means. The label and note commands are used for this purpose. You can use label to give extended descriptions of the variables, e.g.

label variable wt "Weight (lbs)"
label variable age "Age Groups"
label variable case "Case/Control Status"

The label command may also be used to label the *values* of the variables, e.g.

label define agegroups 1 "25-24" 2 "35-44" 3 "45-54" 4 "55-64" 5 "65-74" 6 "75+"
label define status 1 "Case" 0 "Control"
label values age agegroups
label values case status

You may also use

note: <note phrase>

to add a note to the entire dataset or

note <variable name>: <note phrase>

to add a note to any variable in the dataset.

Creating and Transforming Variables

You can create new variables with the generate command or replace existing variables with the replace command (also available through the StataQuest Data … Generate/Replace menu). The recode command is used to code variables into categories. E.g.

generate lwt = ln(wt)	* makes a new variable, lwt, equal to the natural log of wt
replace wt = ln(wt)	* replaces wt with the natural log of wt
generate rnum = uniform() < (.5)	* generate 5 random 0/1 random numbers
generate htgrp = ht	* copy ht to htgrp
recode htgrp 0/71=1 72/83=2 84/100=3	* recode htgrp into categories

Summarizing Data

The following commands are useful for describing and summarizing data.

describe	***** describes the dataset, giving variables names and labels
list	***** lists the data, using value labels if they have been supplied
list, nolabel	***** lists the data; value labels are not used
notes	***** prints out any notes you have added to the dataset
summarize	* prints number of obs., mean, standard dev., min. and max. for each variable
tabulate <var1><var2>…	* crosstabulates <var1> by <var2> by …; if only one <var> is listed a frequency table is given

stem <var> * generates a stem and leaf plot for <var> (not in a graphics window)

Graphics

The basic graphing command in Stata is graph, which has many options. The basic form of the command is

graph <varlist>, <graph_type> <options>

where graph_type can be histogram, twoway, matrix, oneway, box, star, bar, or pie. If you fail to specify graph_type, the default is histogram. See graph in the Reference Manual and online help. There is also a separate Graphics Manual. A few examples are

```
graph fev,  xlab bin(8) ylab      * histogram (the default) of fev, include labels on x and y axis, make
                                  * 8 bins in the histogram
graph fev, box ylab by(sex)       * boxplot of fev by sex, include a label on the y-axis
graph wt ht, xlab ylab bor gap(4)     * scatterplot of wt by ht, include labels on x and y axis, put a
                                      * border around the plot, set the gap between y-axis and its
                                      * label
```

Using the Edit menu you can copy the graph to the clipboard (and hence to other Windows applications). Using the File menu you can print the graph. An additional graphics related command you should be aware of is

```
set textsize 150                  * Increases text size in the graph by 50%
```

Getting Output

When working at the Health Sciences microcomputer lab you can get printed and graphic output by sending output to the printer using the Stata print functions (found under the File menu). The cost of printed output is 5 cents/page and you can use a library copicard or cash. The Microlab personnel can provide further instruction on obtaining your output.

A sample Stata session

********** Use * at the beginning of the line to enter a comment **********
*
* Start a log file by clicking on the log button

. quest on

. edit

. * enter the following data for var1
. * 72 65 84 73 68
. * name this variable ht, add label "height (in.)"
. * enter the following data for var2
. * 165 120 210 180 125
. * name this variable wt, add label "weight (lbs)"

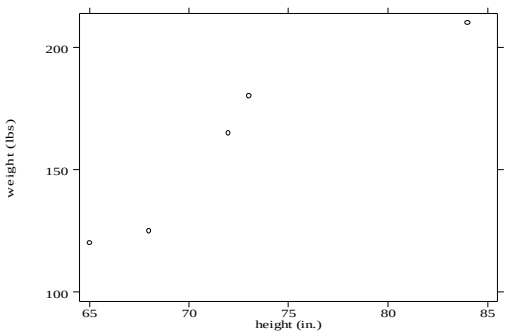
. describe

```
Contains data
  obs:             5
 vars:             2
 size:            35 (99.9% of memory free)
------------------------------------------------------------------------------
   1. ht        byte   %8.0g                 height (in.)
   2. wt        int    %8.0g                 weight (lbs)
------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

. summarize

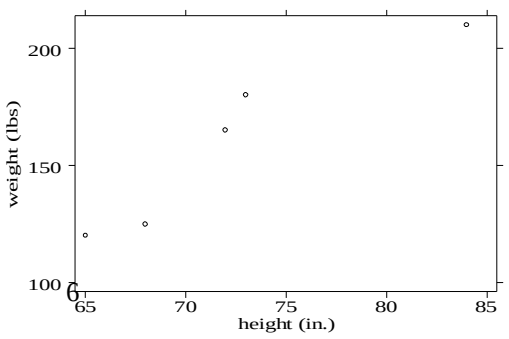
```
Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+--------------------------------------------------------
    ht   |       5        72.4   7.231874        65         84
    wt   |       5         160    37.91438       120        210
```

. graph wt ht, xlab ylab bor



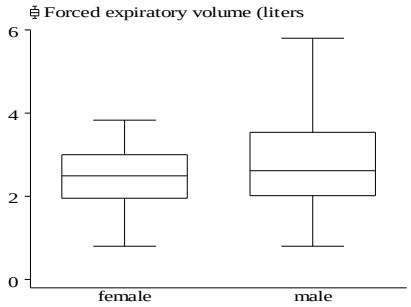
. set textsize 150

. graph wt ht, xlab ylab bor gap(4)



```
. clear

. infile id age fev ht sex smoke using fev.dat
(654 observations read)

. label var age "Age (years)"
. label var fev "Forced expiratory volume (liters)"
. label var ht "height (inches)"
. label var smoke "Smoking status"
. label define status 0 "noncurrent smoker" 1 "current smoker"
. label define gender 0 "female" 1 "male"
. label values smoke status
. label values sex gender

*. save "C:\DATA\fev.dta"
*file C:\DATA\fev.dta saved

. describe

Contains data
  obs:           654
 vars:             6
 size:        18,312 (96.5% of memory free)
------------------------------------------------------------------------------
   1. id        float  %9.0g
   2. age       float  %9.0g                 Age (years)
   3. fev       float  %9.0g                 Forced expiratory volume
                                               (liters)
   4. ht        float  %9.0g                 height (inches)
   5. sex       float  %9.0g       gender
   6. smoke     float  %17.0g      status     Smoking status
------------------------------------------------------------------------------

. summarize age - smoke

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+--------------------------------------------------------
     age |     654    9.931193   2.953935          3         19
     fev |     654     2.63678   .8670591       .791      5.793
      ht |     654    61.14358   5.703513         46         74
     sex |     654    .5137615   .5001931          0          1
   smoke |     654    .0993884   .2994118          0          1

. tabulate sex smoke, chi2

           |    Smoking status
       sex | noncurren  current s |     Total
-----------+----------------------+----------
    female |       279         39 |       318
      male |       310         26 |       336
-----------+----------------------+----------
     Total |       589         65 |       654

          Pearson chi2(1) =   3.7390   Pr = 0.053

. sort sex
```

. graph fev, box ylab by(sex)



* Close the log file
. exit