

3 Getting Started with *Stata*

Dealing with data by hand or even with a calculator can be tedious. Working with appropriate statistical software enables us to explore the data and to deepen our understanding of statistics. In this lab we will consider demographic data from Los Angeles County. *Stata* encourages us to focus on the story that the data are telling us.

Opening *Stata*

After you have logged on, using your lab ID as your name and your nine digit UCLA ID as your password, click on the *Stata* icon. You should see four windows: Review, Variables, *Stata* results, and Command. You may want to resize the windows to fit the screen. To change the font size, you need to go to Preferences or Prefs.

To begin, type

```
. describe
```

in the Command window and press enter. All commands (written in bold type following the dot) will be typed in the Command window. *Stata* is line command oriented which makes it fast with lots of memory free for data. You must be careful to type each command exactly as written, but without the dot.

*Question 1: Since there is no data, what is written in the *Stata* results window?*

Now we will open a file that contains data that we will use several times in our course.

```
. use http://www.stat.ucla.edu/labs/datasets/smallcen.dta
```

Green words mean that the typing is fine while red words mean that the typing needs to be modified.

Exploring the Data

Type

```
. describe
```

or take a short cut by using the page up option on the keyboard.

Question 2: What is this data set? How many variables are there? How many observations?

Type return to scroll line by line and space bar to scroll down page by page.

We want to get a feeling for the data. We can look at individual observations by typing

```
. list
```

Question 3: What do we know about these individuals?

Scrolling through 2500 observations is tedious. To break the scroll, we use

```
. q
```

for quit or press Apple and period at the same time.

It is hard to concentrate with so many variables and cases, so let us focus on two variables. Suppose we are interested in whether men make more money than women. We can focus on the two variables gender and monthly income. You can look at just this data by typing

```
. list gender income
```

Question 4: What does the 0 represent?

Instead of listing all of the data in the dataset, *Stata* also allows us to subset

the data and concentrate on particular observations. For example, we can look at the income for ten of the youngest people. Type

```
. sort age
```

And then type

```
. list age income in 1/10
```

The *Stata* command **list** is a great way to look at individual observations and to get a feeling for what the data looks like, but with a dataset as large as this one, it is difficult to get an overall feeling about the data. The **summarize** command gives a quick numeric summary of all the observations. By listing the data and summarizing the numerical variables, we begin to understand what the data look like.

Describe what you see when you type

```
. summarize
```

Similarly to the **list** command, *Stata* will limit its summaries to variables you select.

```
. summarize income
```

Question 5: What is the typical income?

The average of the variable income may not actually tell us what the typical income is, there are many people that do not work and hence their income is recorded as 0. We want to look at the mean income for those who have an income. To remove the 0s, we will code them as missing values.

```
. mvdecode income, mv(0)
```

Question 6: How did this change the summarize income output?

Remember, we are interested in comparing the incomes of men and women.

We can first look at the summary of income for men and then the summary of income for women by typing the following:

```
. summarize income if gender==1  
. summarize income if gender==2
```

In this data set, men are coded with a 1 and women are coded with a 2.

Question 7: What do you see? What differences or similarities do you notice about the income variable for men and women?

The previous command requires that you know exactly what values the variable gender consists of and what they represent. In this case, you only know because we told you. Another way to obtain the same output, but without prior information on how men and women are coded is to do the following:

```
. sort gender  
. by gender: summarize income
```

The **by** prefix is something we will use quite frequently in this course and whenever it appears in a *Stata* command, the data must first be sorted using the same variable.

Question 8: How does this display differ from the previous output?

Numerical summaries tell us some general things about the data, but graphical summaries can help us understand the overall distribution of the data. One way we can compare the incomes of men and women visually is by using side-by-side boxplots. *Stata* provides assistance in understanding commands with its **search** and **help** commands. The following commands could be used to determine how to make a boxplot.

```
. search boxplot  
. help grbox
```

The command we ultimately use is

```
. graph income, box by(gender)
```

Question 9: Write a paragraph to describe the differences in the income earned by men and women. Use both summary statistics and the graph to assist you in your description.

Here is a list of the commands we used in the Getting Started with *Stata* Lab. Use the space next to each command to make notes on what that command does.

describe

list

sort

summarize

graph, box

search

help

Here is a list of conditions:

in

if

by

To leave Stata, type

```
. clear
```

```
. exit
```

Assignment

We looked at income differences with respect to gender. Select a categorical variable other than gender, such as marital status, race, educational level, or type of household. Write a brief summary which compares income with respect to the variable of your choice. Make sure to include appropriate summary statistics and/or graphs.