

## Biostatistics 208 Lab #1, 01/07/10

The purpose of this lab is to familiarize you with basic univariate descriptive statistics, both numerical and graphical. Our main tools will be the `summarize` and `graph` commands in Stata, each of which is flexible, with a number of options. You can produce graphs in Stata either by typing commands in the command window or by using pull-down menus. Below you'll find the commands in Courier font, and the sequence of menu selections in italics. Start by reading the online help for the `summarize` and `graph` commands. The Stata graphics manual is a valuable resource for learning Stata's considerable graphical capabilities.

Before you start this lab

- Download the dataset `lab1.dta` from the "Syllabus" link on the following page:  
<http://www.biostat.ucsf.edu/biostat208/>
- Start a Stata log on your external drive or disk. Go to *File, Log, Begin* or click on the log icon, which resembles a scroll.

The dataset is from an HIV clinical trial, ACTG 333. The study measured HIV viral load at baseline and after eight weeks of treatment. In addition, the HIV-1 protease gene was sequenced for all study participants. The objective of the study was to link the changes in viral load to mutations in this gene at codons 10, 71, and 90. The 5 variables in the dataset are:

1. `logv1ch`: change in log-10 HIV viral load from baseline to week 8
2. `tx`: treatment assignment
3. `codon10`: mutation at codon 10 (1=present, 0=absent)
4. `codon71`: mutation at codon 71 (1=present, 0=absent)
5. `codon90`: mutation at codon 90 (1=present, 0=absent)
6. `logv10`: baseline log-10 HIV viral load.

### 1. NUMERICAL DESCRIPTION

A convenient place to start is the `summarize` command, which gives summary statistics for all numerical variables, including the number of observations, mean, standard deviation, minimum and maximum:

```
summarize
```

*Statistics; Summaries, tables & tests; Summary statistics; Summary statistics*

To get more detailed information about baseline HIV viral load, including the median, quartiles, and other percentiles, use the `detail` option. Check for any unusual values (remember that they are on the log-10 scale):

```
summarize logv10, detail
```

*Statistics; Summaries, tables & tests; Summary statistics; Summary statistics;*

*-Display additional statistics*

Even though the `tx` and `codon` variables appear when you issued the `summarize` command (because they are encoded as numbers with “value labels,” not characters) they are of course categorical. Use `tabulate` to obtain more useful descriptive statistics for `tx` and `codon10`:

```
tabulate tx
tab codon10
```

*Statistics; Summaries, tables & tests; Tables; One way tables*

## 2. SINGLE VARIABLE GRAPHS

### 2.1 Histograms

Various graph commands can be used to depict the distribution of a variable. To get histograms of baseline log-10 HIV viral load as well as the changes in viral load from baseline to 8 weeks, type

```
histogram logv10
hist logvlch
```

*Graphics; Easy graphs; Histogram*

A graphics window appears with a histogram of the distribution of log-10 HIV viral load values. Use the `bin()` option to change the number of bars on the histogram. What number of bins appears to give the most informative histogram? **HINT:** If you use the menus to do graphs, pressing “submit” instead of “OK” will allow you to add options and rerun the graph without having to start from scratch. Either way, the actual command gets printed in the results window, if you want to learn the code.

### 2.2 Boxplots

Recall that a boxplot gives a graphical representation of key features of the distribution of a continuous variable. The upper and lower limits of the box represent the 75<sup>th</sup> and 25<sup>th</sup> percentiles of the data, while the median (50<sup>th</sup> percentile) is included as a line through the box. The height of the box marks the interquartile range (IQR) of the observations, and the “whiskers” mark the largest (smallest) observations 1.5 times the IQR outside of the box in either direction. Points falling outside this range are typically represented as dots, and can be viewed as outlying observations.

To obtain a boxplot of the changes in log-10 viral load from baseline to week 8, type:

```
graph box logvlch
```

*Graphics; Easy graphs; Box plot*

Can you read off the approximate 25, 50 and 75th percentile of the data? Do there appear to be outliers?

### 2.3 Q-Q Normal Plots

Histograms and boxplots are useful for assessing distributional shapes, but the `qnorm` command is best for assessing whether a variable approximately follows a Normal distribution. This is useful since later in the course we will use statistical techniques that work best for variables that are Normally distributed (at least approximately). Try the `qnorm` command on baseline log-10 viral load:

```
qnorm logv10
```

*Graphics; Distributional graphs; Normal quantile plot.* For variable, type `logv10`

Recall that the diagnostic for Normality is that the plotted data should fall approximately on a straight line; Stata superimposes a diagonal line to help you make this judgment. Real data never fall exactly on a straight line; the pattern in this case reflects the short tails noticeable in the histogram.

How close is close enough to a straight line? This is a hard question and depends on what technique you want to use, how sensitive it is to the Normality assumption, the nature of the violation (short tails are relatively benign), and on the number of observations (you can get away with more serious violations in a big dataset). One useful exercise is to see what Q-Q plots look like when the data are really from a Normal distribution. You can do this by issuing the commands:

```
gen rannor=invnorm(uniform())
qnorm rannor
```

*Data; Create or change variables; Create new variable*  
*Graphics; Distributional graphs; Normal quantile plot*

This generates a Normally-distributed variable with the same number of observations as the original data set. Because of random variation, the Q-Q plot will not be completely linear, especially in small datasets. Do this a couple of more times to get idea if your Q-Q plot is close to that expected for a plot for exactly Normal data. You will need to first drop the `rannor` variable (or give the new one a different name). Using the PageUp key or the Review window is handy for rerunning commands.

```
drop rannor
gen rannor=invnorm(uniform())
qnorm rannor
```

*Data; Variable utilities; Eliminate variables or observations. In drop: type rannor*  
*Data; Create or change variables; Create new variable*  
*Graphics; Distributional graphs; Normal quantile plot*

To control the number of observations, since this affects the appearance of the Q-Q plot, you can use the following commands. Try 25 observations, since the plots tend to look messier in small datasets:

```
clear
set obs 25
gen rannor=invnorm(uniform())
qnorm rannor
```

You will need to reload lab1.dta after doing this.

### 3. MODIFYING, SAVING, AND PRINTING GRAPHS

#### 3.1 Adding titles and other options to existing graphs

If you don't like the way a graph looks you can improve its appearance using options. The graphics menus make this particularly easy. Remember that if you press "Submit" rather than "OK", the window will not disappear. You can then add options for the title, axis, and so forth, as many times as you like. In particular, explore the options in the histogram pop-up window (titles, axes, and options).

You can also develop a better-looking graph from the command line. Use the PageUp key or click in the Review window to recall the current version of the graph command. Then type options at the end of the command (adding a comma before the first option if necessary). Suppose you want to add a title to the graph you obtained using the command `histogram logvlch`. Recall this command, then add the option for the title:

```
histogram logvlch, title(Histogram of Changes in Log-10 HIV Viral Load)
```

Editing and re-running a do-file is another handy method for refining a plot.

#### 3.2 Saving, pasting, and printing graphs

Graphs can be saved, printed, and copied and pasted into Word and other documents. This is quite useful for producing your homework reports. When the plot is in the graphics window, it can be saved to a file by selecting on *File, Save Graph*. After a graph has been saved, double clicking on it from Windows Explorer or the Mac Finder will bring up a STATA window with that graph. To cut and paste a graph into a Word document using a PC, choose "Copy" from the edit menu, and paste the graph directly into an open Word document. Alternately, you can save the graph as a .wmf file, then use *Insert, picture* in Word. On a Mac, you can copy the graph using Edit; Copy; then use the *Paste* or *Paste Special* commands under the Edit menu in Word. Try this now by opening a Word document and pasting a graph into it. Note that the "Copy Graph" edit menu item in Stata will place the graph in the Mac clipboard in pdf format. This may or may not work for pasting directly into Word depending on the version you are using. However, the clipboard contents can be opened in Preview or other Apple applications such as Keynote and Pages.

Graphs can be printed directly using the *File; Print Graph* command from within Stata, or by printing the Word document into which the graph has been pasted.