

## Lab 2

# Visual Display of Data in *Stata*

In this lab we will try to understand data not only through numerical summaries, but also through graphical summaries. The data set consists of a number of variables about students from a previous Statistics course. We will perform an exploratory analysis of student heights using not only means and standard deviations, but graphing tools such as boxplots and histograms.

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

Examine what this data set contains by using the `describe` command.

```
. describe
```

*Question 1: How many variables are in the data set? How many observations are in the data set?*

## Detailed Summary Statistics

You have been given a dataset and would like to gain an understanding of the distribution of students heights based on the sample. The first thing you

might think to do is calculate a mean and maybe a standard deviation. We have already seen how easy this is to achieve in *Stata*.

```
. summarize height
```

We now see that the average height is 67.3. Given this number, height must have been recorded in inches. We see that the standard deviation is about 4 inches, the maximum height in this sample is 76 inches and the minimum height is 60 inches. We also see that in this sample, all 82 students have a value recorded for height.

Whenever, we describe a distribution using the mean, we have to be careful because the mean (and the standard deviation) are affected by outliers. The median and the IQR are alternative measures of center and spread that are resistant to outliers and extreme values.

*Stata* makes it easy to find the median and the IQR. We still use the `summarize` command, but include an option. Options in *Stata* follow a comma. Options are command dependent, but frequently the same options work for a family of commands.

```
. summarize height, detail
```

In addition to the information that is given in the previous `summarize` command, we now see some of the percentiles for the distribution in the first column, the four largest and four smallest values in the dataset in the second column, and other measurements of variance in the third column, including kurtosis and skewness. Remember that the 50th percentile is the median, the 25th percentile is the first quartile, and the 75th percentile is the third quartile.

*Question 2: What is the five-number summary for the height variable? Are there any outliers?*

*Question 3: Based on the mean and the median, do you think that the distribution is skewed? If so, in what direction?*

## Visual Displays of Data

A boxplot is visual display of the five-number summary.

```
. graph box height
```

A glance at the boxplot helps us get an idea of the center, the spread and if the distribution is symmetric or not.

*Question 4: Use the boxplot to give a description of the distribution of the heights of the students in the sample. Make sure to discuss center, spread, and shape.*

*Stata* gives us many options to alter our boxplot, but one of the most obvious options is to give the plot a title.

```
. graph box height, title(Boxplot of student heights)
```

This `title` option can be used with *any* graph command.

Boxplots are great visualization tools. They allow us to easily identify possible outliers, and get a quick overall impression of the distribution of the dataset. But, it is important to remember that boxplots are created from only five numbers. You can create a boxplot without having the entire dataset. Important information can be hidden by summarizing the data with only five numbers.

To obtain a real understanding of the distribution of the data, it is better to use all of the data. Graphs such as stemplots and histograms do just that.

```
. stem height
```

*Question 5: Use the stemplot to give a description of the distribution of the heights of the students in the sample. Make sure to discuss center, spread, and shape. What have you learned (if anything) about the distribution that you didn't already know?*

Stemplots are nice graphs that allow us to see all of the individual values in the dataset. This can be very useful, but with a bigger dataset could become very cumbersome. Another drawback to stemplots is that they don't allow much control over the bin size. A histogram provides the same visual display of the distribution without including all the individual values.

```
. histogram height
```

*Stata* has many useful options for histograms. For example, you can change the number of bins.

```
. histogram height, bin(12)
```

Or suppose you want each bin to be of a certain width.

```
. histogram height, width(1) title(Histogram of student heights)
```

You can plot a normal curve over the histogram. Can you guess what mean and standard deviation the normal curve has?

```
. histogram height, bin(7) norm
```

You may have noticed that the  $y$ -axis says "Density". Remember that a density curve has an area of exactly one beneath it. A density histogram is one such that if you added up the area of all the rectangles, you would get a sum of one. Sometimes it is more useful to look at a frequency histogram...

```
. histogram height, freq bin(35)
```

The options `fraction` and `percent` will give you relative frequency histograms.

*Question 6: Play around with the number of bins or width of the bins for the histogram of height. Select the graph that you find most informational and add a title. Use this histogram to give a description of the distribution of the heights of the students in the sample. What have you learned (if anything) about the distribution that you didn't already know?*

## Comparing Subgroups

One thing you should have noticed is that the distribution of heights of students is bimodal with one peak around 63 inches and another around 67 inches. What could be the reason for that?

*Question 7: What might the reason be for two distinct peaks in the distribution of student heights?*

A logical guess is that one peak represents the women in the sample and the other peak represents the men in the sample. *Stata* allows us to easily compare the two subgroups of males and females using the variable `gender`.

```
. graph box height, over(gender)
```

*Question 8: Describe the distributions of the heights of female and male students. How do the two distributions compare?*

We can also make side-by-side histograms.

```
. histogram height, by(gender)
```

*Question 9: Describe the distributions of the heights of female and male students using the histograms. How do the two distributions compare? What have you learned (if anything) about the distributions that you didn't already know?*

## Additional Graphing Options and *Stata* help

You have already seen a number of options for graphs. Titles, selecting the number of bins or the width of the bins for histograms. You know

how to tell *Stata* if you want a frequency histogram or a relative frequency histogram instead of a density histogram. You can plot a normal curve over the histogram. You can separate plots using the **over** or the **by** options, but there are a number of other options that *Stata* allows for graphs.

Instead of listing out all the possible options for graphs, explore the *Stata* help menus.

```
. help histogram
```

At first glance, the help menu is a bit overwhelming, but with closer examination, things become clearer.

One of the first things you should see is the following text:

```
histogram varname [weight] [if exp] [in range] [,  
                [discrete_options | continuous_options] common_options ]
```

The **histogram** part means that you can type **hist** instead of the full word **histogram** when using the command. Things that are in brackets ‘[ ]’, are optional. So, you don’t need *weights* or *discrete\_options*, but you must have a variable input where it says *varname*.

Below that section, you should see a section titled *discrete\_options* and a section titled *continuous\_options*. Looking at the *discrete\_options*, you can see there are three options. We have already seen the **width** option.

*Question 10:* What does the **start** option do? (Hint: Scroll down for a lengthier description of the options.) Create a histogram of the heights that starts at 55.

Further down the help menu, you see the *common\_options*.

*Question 11:* What does the **addlabels** option do? Make a histogram using that option.

Remember that you could have typed **addl** instead of **addlabels** and gotten

the same result.

Titles and labeled axes prove to be useful when looking at a graph, so you might be interested in looking at the *twoway\_options*. Anything that is colored blue has its own help menu. But, before we look at the *twoway\_options*, continue through the rest of the help for histograms.

Examples are at the end of virtually every help item. You can click on them to run them or just look at them to get an idea of how the commands are used.

Now let's look at the *twoway\_options*.

```
. help twoway_options
```

As you can see, this gives you a list of more help menus.

*Question 12: Figure out how to create a histogram with a line through the x-axis at 67.3 (the average of all student heights).*

We also looked at box plots, suppose you are interested in the options for box plots.

```
. help box plot
```

What happens? You get a statement that says “help for **box** not found”. This is because the command **help** only works for specific *Stata* determined phrases. Not a problem though, *Stata* built in a more encompassing search program. Try the following command.

```
. search box plot
```

Now a series of menus appears. The one that we want is **graph\_box**. Look at the help menu for box plots.

*Question 13: Figure out how to make side-by-side boxplots of males and females horizontally instead of vertically.*

## Assignment

The U.S. Environmental Protection Agency rates vehicles according to their environmental performance in the Green Vehicle Guide. You can check out the Green Vehicle Guide at their website. <http://www.epa.gov/greenvehicles>

We imported a portion of the vehicle guide for 2003 into *Stata*.

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

Discuss the data set. How many variables are in the data set? How many observations? What do the variables represent?

Examine the highway mileage of the vehicles in this dataset. Use summary statistics and graphical tools to describe the distribution of highway mileage.

What is causing the bi-modality of the distribution of highway mileage? Hint: Consider the variable `vehclass`.

Write a report on your findings. Be sure to include answers to the specific questions mentioned above and include any graphs that are appropriate. **DO NOT** include tons of *Stata* output. Do not just copy the results of *Stata* into a document and turn it in. Don't say things like "the mean is 5", use sentences in context "the average mileage is 5 mpg".