

## 4 Visual Display of Data in *Stata*

Please remember to use your Lab ID# as your log in ID and your nine digit UCLA ID# as your password when you log-in.

In this lab we will plot histograms with various bin widths, add appropriate scales and titles, and look at other plots to help us better understand the demographic data collected in a recent census of Los Angeles County.

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

### Detailed Summary Statistics

Suppose we are curious about the ages of those in our sample. We would want to know the average age of the respondents, the standard deviation, and the range.

```
. summarize age
```

This tells us that the youngest respondent was 16 while the oldest was 87, with a mean of 43 years, with the standard deviation of 18 years. Now lets dig a little deeper into the ages of the respondents. The command

```
. summarize age, detail
```

gives us additional information about the distribution of ages. In addition to the information that is given in the pervious command we now see the percentiles for the distribution in the first column, largest and 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 1: What is the age range of the respondents in our sample? What is the median age of the respondents? What is the IQR of age?*

## Visual Displays of Data

*Stata* offers many options for graphing and plotting data. We will examine a few of them in this lab. There are many different options that can be used in graphs to make them more informative and easier to read; things like number of bins, axis scale, labels and titles. *Stata* offers many more options than what will be covered, so check out the help menus for more options. Let's start by looking at age again, except visually.

First look at a basic histogram of age with no special options selected.

```
. graph age
```

Notice that the range on the x-axis matches the range of the *age* variable (16 to 87), the default for *Stata* is to use the minimum and the maximum for the variable on the axis, for the range on the axes. *Stata* uses five as the default number of bins in a histogram.

If we plotted the data by hand, we might use a horizontal scale from 0 to 100 and have 10 bins. In *Stata*, we type:

```
. graph age, xscale(0, 100) bin(10)
```

Now try the command:

```
. graph age, xscale(0, 100) bin(10) norm
```

*Question 2: What does Stata do when we add norm to the command?*

We can get a better sense of the age plot, if we adjust the number of bins. We do not want too many bins or the plot will look like a city skyline, too few will give very little information.

*Question 3: Change the number of bins to 5, 20, and 35. Compare with the graph that has 10 bins. Which graph do you find most pleasing and informational?*

## Adding Labels and Titles

Labelling the horizontal axis may help us locate the center of the distribution. We can add as many labels along the axis as we like, however too many will begin to look cluttered and should be avoided. Adding labels on the axes only requires an extra option to the command that we have been using already.

```
. graph age, xscale(0,100) bin(10) xlabel(0, 10, 20, 30, 40,
50, 60, 70, 80, 90, 100)
```

It's a bit tedious to type all these numbers. If the intervals between the labels is equal, there is a shortcut to this command that can reduce typing.

```
. graph age, xscale(0, 100) bin(10) xlabel(0 10 to 100)
```

The first two numbers set an interval length that will be repeated until the third number in the series is reached. For example, suppose you want to label every 5th number from 20 to 65, you would use the command `xlabel(20 25 to 65)`.

These same rules apply to the  $y$ -axis. If we wanted to change the scale of the  $y$ -axis we would use the command `yscale` and if we wanted to change the labels on the  $y$ -axis we would use the command `ylabel` and fill in the appropriate values.

*Question 4: Repeat the same graph as above with different labels and scaling for the  $y$ -axis. What scale for the  $y$ -axis do you think is most appropriate and informative?*

Titles on our graph will help us identify the plot later after it has been printed and we have left the lab. To give your graph a title there is a simple, and obvious, command.

```
. graph age, xscale(0, 100) bin(10) xlabel(0 10 to 100)
title(Age of the Respondents in our sample)
```

This is okay, but lets suppose you don't want the title of your graph at the bottom, but at the top. *Stata* allows us to have two titles in four different

locations around the figure: top, bottom, right and left. Adding the command `t1title` will put a title in the first spot on the top of the figure and `r2title` will put a title in the second spot on the right of the figure. The fonts in the first spots on each of the four sides is larger than those in the second spots. Lets give our graph a title on the top. type:

```
. graph age, xscale(0, 100) bin(10) xlabel(0 10 to 100)
t1title(Age of the Respondents in our sample)
```

## Comparing Graphs

If we are interested in comparing the ages of men and women, we would want to look at two histograms, one for the men and one for the women. For us to do this we need to sort the data by gender to put the data in order for *Stata*.

```
. sort gender
```

If you type `browse` in the command window you can see the raw data that we are working with. Later you may learn the `edit` command which looks similar to `browse`. The difference is that using `browse` you can only look at the raw data, whereas using `edit` you can look at the raw data and edit it as well. `Browse` will keep you from accidentally editing your data. Scroll over to the *gender* variable and notice that it is sorted `Male` followed by `Female`. Exit the browse window. Now lets make our histograms.

```
. graph age, by(gender)
```

You now have a histogram showing the age distribution on men and another histogram showing the age distribution of women.

*Question 5: Reproduce these graphs comparing ages of men and women with appropriate labels, number of bins, and a title. Describe any similarities or differences that you see.*

## Decoding Data

Now we will consider the rent paid by those in our sample. Type:

```
. summarize rent, detail
```

Before we plot the data, we want to remove the zeros for those who paid no rent by storing the zeros as missing values. This helps us eliminate those who pay no rent, for whatever reason, from the calculations, and gives us a better estimate of the rent variable. If we were to leave them in, all of our calculations would be skewed. Type:

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

We just replaced the rent values that were zero with a missing value, which is displayed in *Stata* with a period. Let's take a look at what this looks like:

```
. list rent in 1/25
```

This lists the first 25 observations. Notice how some of them have periods, instead of values. These are missing values we just replaced. This dataset has a large number of zeros for rent. Now let's look at the rent variable with the zeros removed.

```
. summarize rent, detail
```

*Question 6: How do the median and mean rents compare? What does this tell us about our data?*

Next we plot our rent data. In this command we simplified the `xlabel` command.

```
. graph rent, xscale(0, 700) xlabel t1title(Monthly Rents)
```

*Question 7: Reproduce this graph again, changing the number of bins to ten. Does this change the distribution?*

Now we will consider several other questions about the rent data.

*Question 8: Do women and men in our sample pay comparable rents? What command is needed to find out? What are the findings?*

We may wonder if there is a relationship between age and rent? To display paired data in the form of a scatterplot, we type:

```
. graph rent age, xlabel ylabel
```

*Question 9: Do people pay more rent as they get older?*

A scatterplot is a good graph to use when comparing two continuous variables. Side-by-side boxplots or side-by-side histograms are good graphs to use when comparing one continuous variable to one categorical variable (like *rent* and *gender*).

*Question 10: Is there a relationship between the income someone earns and the rent that they pay? What graph command is needed to find out? What are the findings?*

## Assignment

In this lab, we looked at age, rent, and gender. We learned many new ways to describe data numerically and visually. Now let's put that to work, use these new skills to answer the following questions, remembering to use appropriate labels and bins.

*Question 11: Do men and women in our sample earn the same amount or do their earnings differ?*

*Question 12: Is there an income difference for race? (Note: the labels for race are as follows. 1:White, 2:Black, 3:American Indian, 4:Asian)*

*Question 13: Do single people pay more rent than married? (Note: the labels for marital status are as follows. 1: Married-spouse present, 2: Married-spouse absent, 3:Widowed, 4:Divorced, 5:Separated, 6:Never Married)*