Revised July 2018

Histograms and Boxplots

This set of notes describes how to use the computer program Stata to produce histograms and boxplots. It assumes that you have set Stata up on your computer (see the "Getting Started with Stata" handout), and that you have read in the set of data that you want to analyze (see the "Reading in Stata Format (.dta) Data Files" handout).

In Stata, most tasks can be performed either by issuing commands within the "Stata command" window, **or** by using the menus. Because Stata's graphics commands offer so many options, we often use the menus, but these notes illustrate both approaches, using the data file "GSS2016.DTA" (this data file is posted here: https://canvas.harvard.edu/courses/53958).

Obtaining histograms

The "histogram" command produces simple histograms. The basic syntax that you issue in the "Stata Command" window is:

```
histogram <varname>, <options>
```

where you fill in the variable name for which you want a graph (for "varname"), and add some "options" to control the presentation of the graph. There are a great many options in Stata. This handout does not review all of them.

You proceed somewhat differently for categorical (nominal, ordinal) and for continuous variables.

The following command produces a histogram for the categorical variable "marital" (marital status) in GSS2016.DTA:

```
histogram marital, discrete percent xlabel(1/5,valuelabel)
```

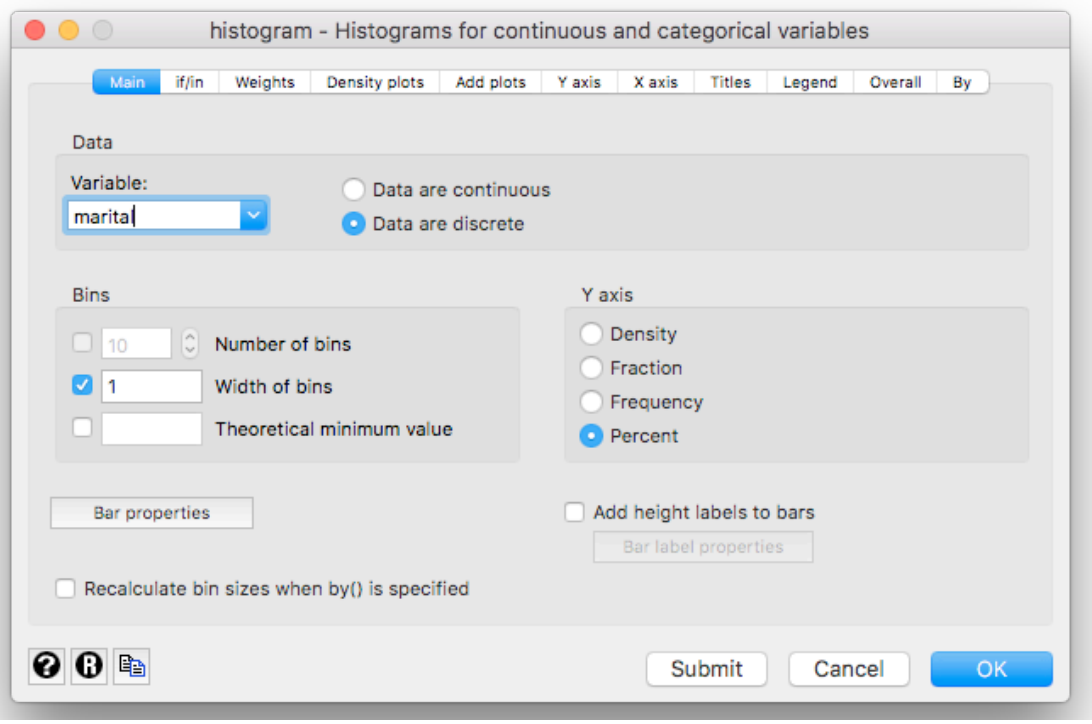
The "discrete" option tells Stata that "marital" is a categorical variable; "percent" scales the vertical axis in terms of the percentage or relative frequency in each category; and the "xlabel" option says to label the horizontal axis using the stored verbal labels for categories 1 through 5.

When you execute the command, a new "Stata Graph" window opens and the graph shown later in this handout appears in it.

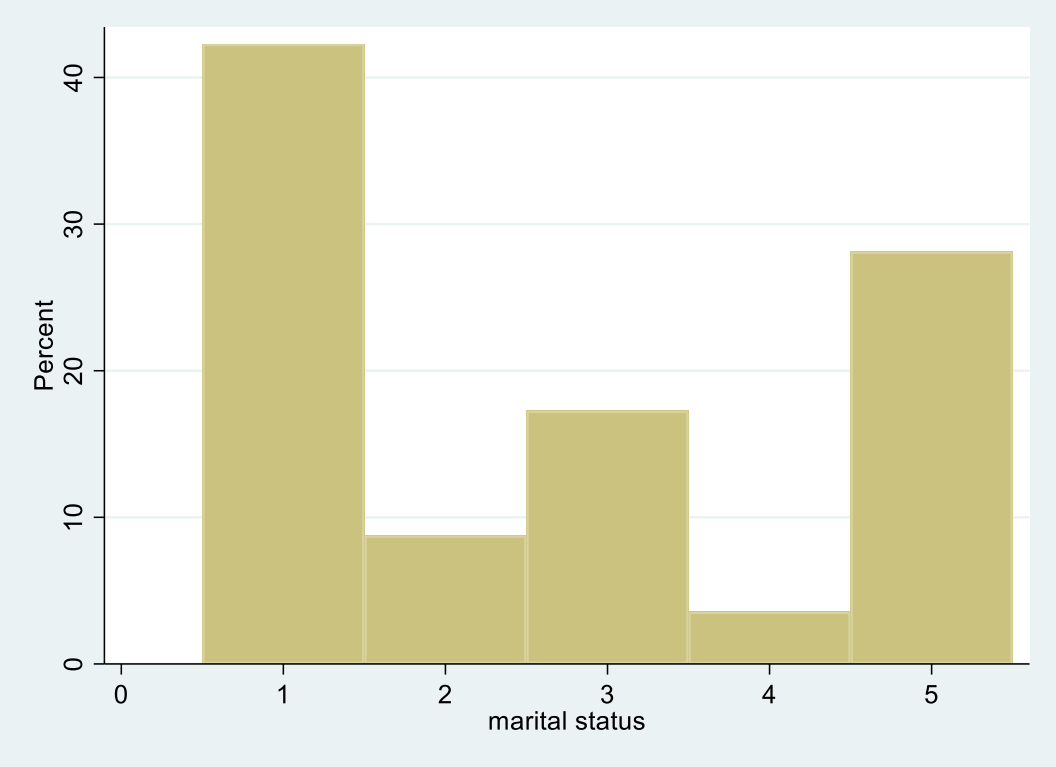
To obtain a histogram for a categorical variable using the Stata menus, proceed as follows:

click on "Graphics"
click on "Histogram"

A window like this opens up:



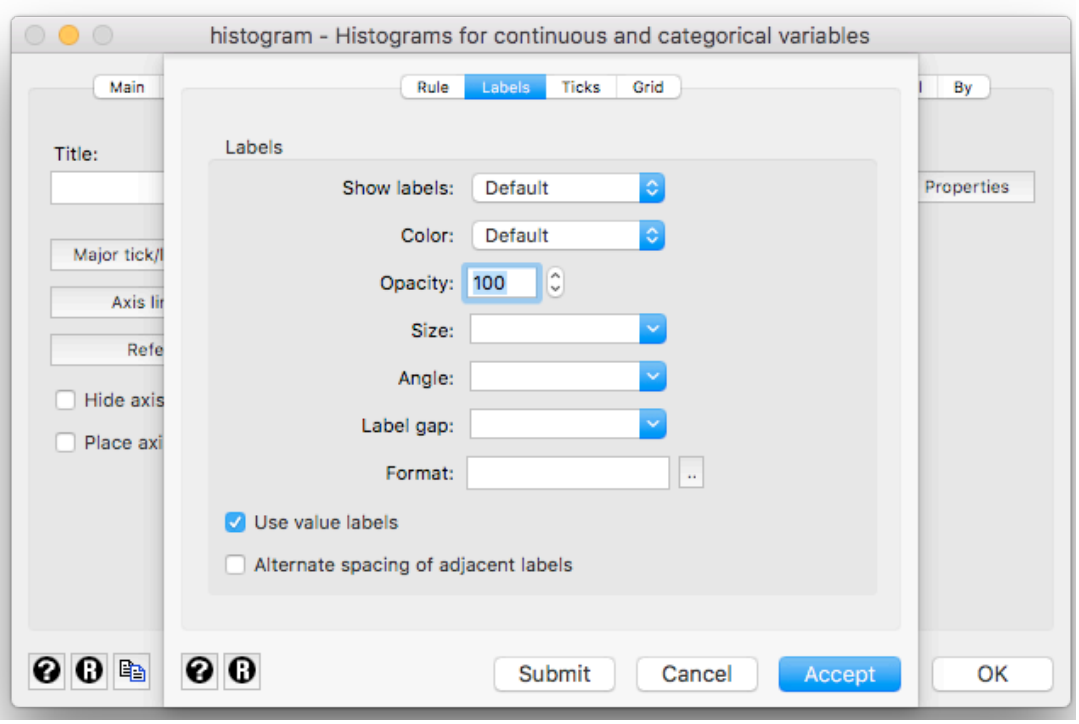
Fill in the variable you wish to represent in your histogram in the "Variable:" box and click the "Data are Discrete" button (if the variable is categorical, or if it is a quantitative variable that can take integer values only [like the number of siblings or children that someone has]). Other options on this screen control the display. Select "Width of Bins" and type "1" in the corresponding field to specify that each bin (i.e. each bar) has a width of one. Select the "Percent" button to set percentages as the scaling for the vertical axis. Then click "OK" or "Submit" to obtain the following display (if you use "Submit", the graphics command window will remain open, which can be helpful if you may be revising a graph; using "OK" closes the graphics command window).



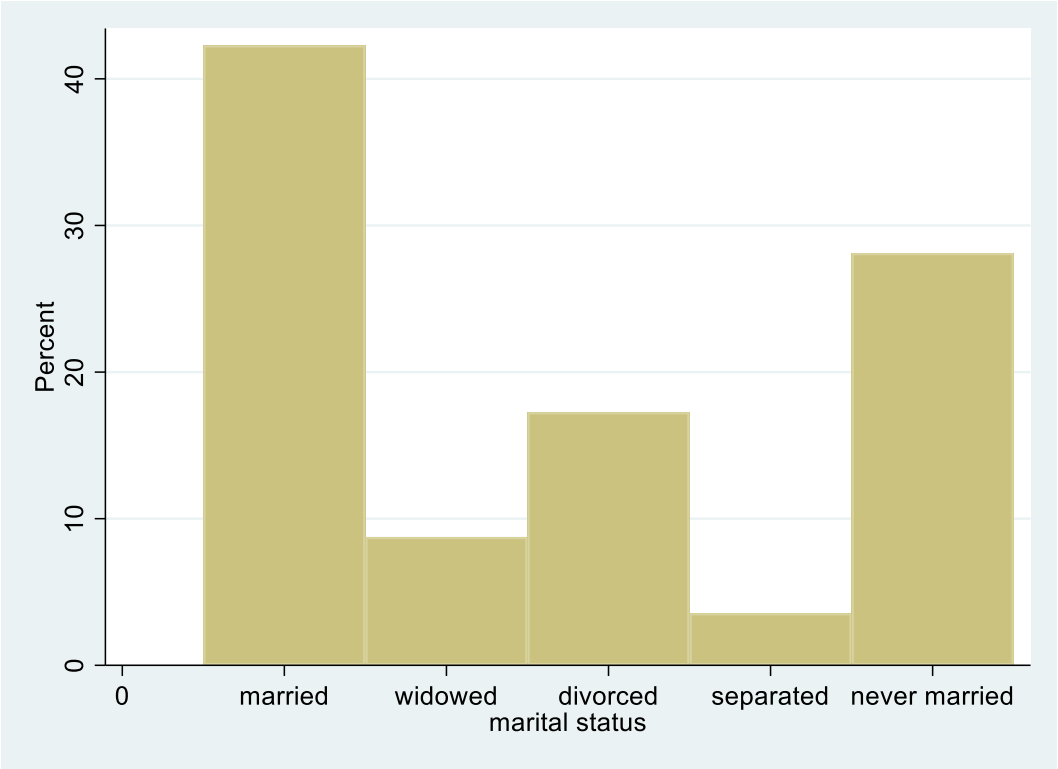
Other tabs on the screen allow you to further control the display.

Note that there are no "value labels" along the horizontal axis of the preceding graph. Such labels are helpful to show what the five codes for marital status mean.

One could add them via the "X Axis" tab of the above screen. Click "Major tick labels/properties," then click on the "Labels" tab in the pop-up window. Under "Labels" check the "Use value labels" box and click "Accept." Then click "OK" in the main histogram window.



This yields the following, somewhat more informative, graph:



The “histogram” command shown earlier produces this same result.

If instead you have a quantitative variable with many (scored) values, you proceed slightly differently. The data then should be grouped into a number of equal-width "bins", each of which will show up as a separate bar in the histogram. A suitable command for the "Stata Command" window is

```
histogram <varname>,bin(#) fraction
```

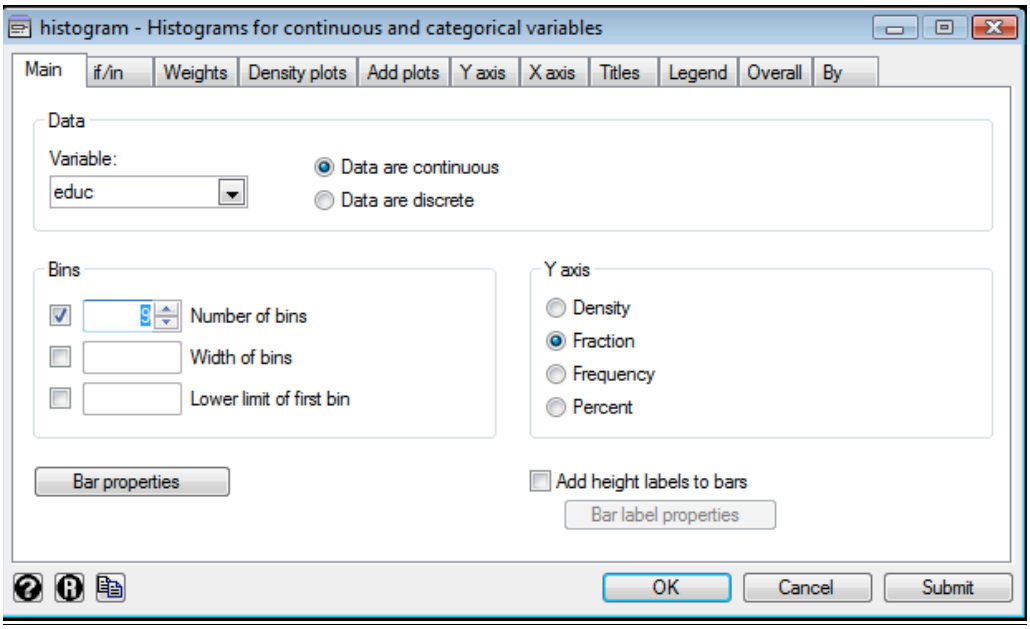
where you fill in the variable name of interest to you (for "varname"), and the number of bins you want in your histogram (for "#"). The "fraction" option sets the scale of the vertical axis in terms of proportions.

For example, a 9-bin histogram for the variable education in GSS2016.DTA is produced as follows:

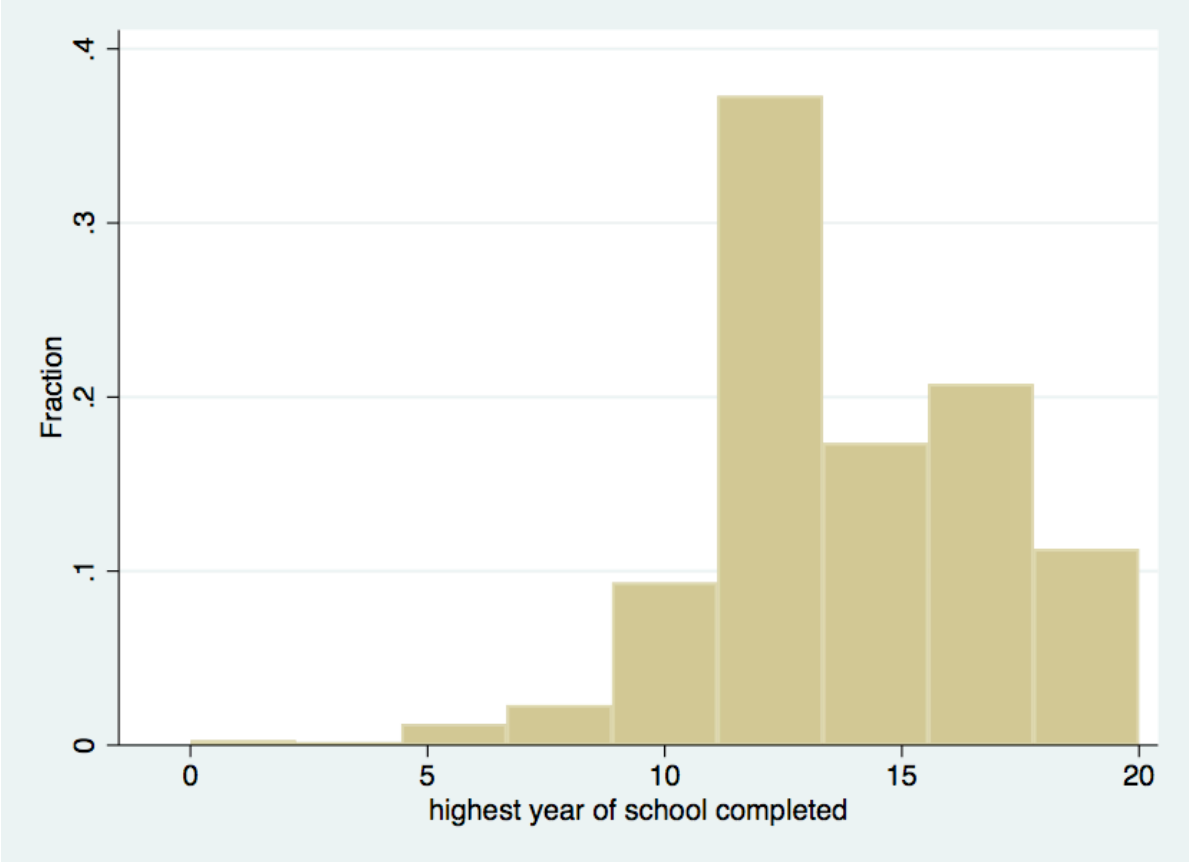
```
histogram educ,bin(9) fraction
```

The result is displayed on the next page.

Using the menus, you begin by clicking on "Graphics", then on "Histogram." You will then see the following window:



Fill in the variable name of interest (e.g. "educ", for education) in the "Variable" box and leave the "Data are continuous" button checked. Check the box beside "Number of bins" and select the number of bins you would like within that box. (Or, choose "Width of bins" to control the interval width.) Select the "Fraction" button to scale the "Y-axis" so bars will show the proportion of people in each bin. Hit "OK" and the graph on the next page appears:



Clearly people are concentrated in the bin that includes 12 years of education; other amounts of education are less common.

Boxplots

To obtain a boxplot via the command line, type the command

```
graph box <varname>
```

in the "Stata command" window, insert the variable name for your boxplot in place of "<varname>", and hit "enter". For example, to get a boxplot for the number of hours spent watching television in GSS2016.DTA, use the command

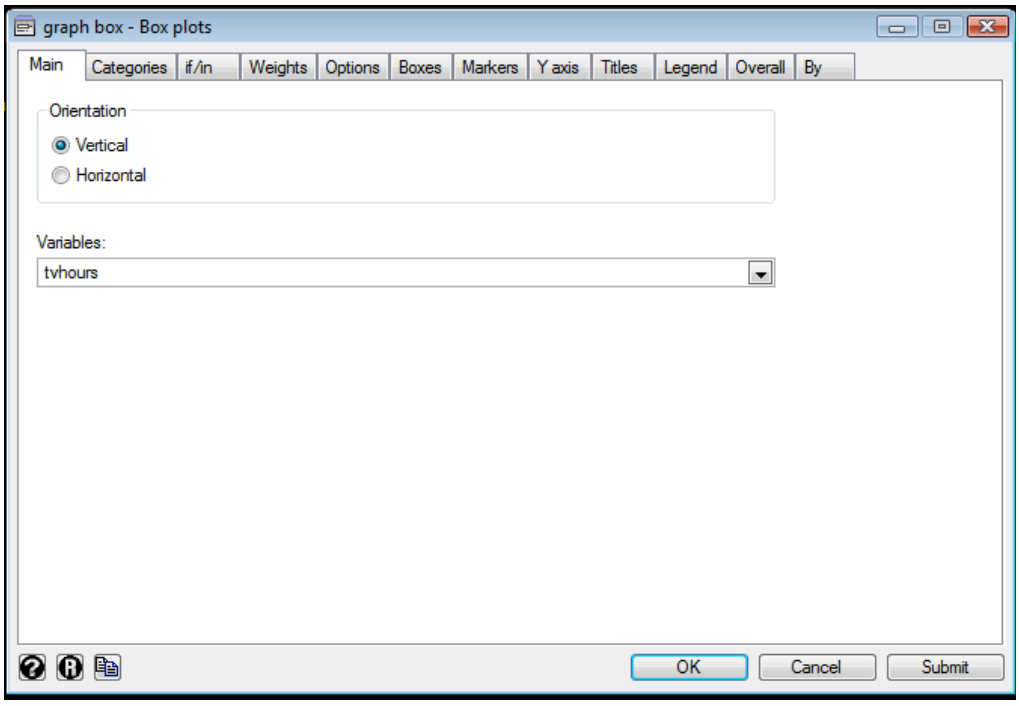
```
graph box tvhours
```

Options are available for this command, but none are key for a basic boxplot.

To get a boxplot via the menus,

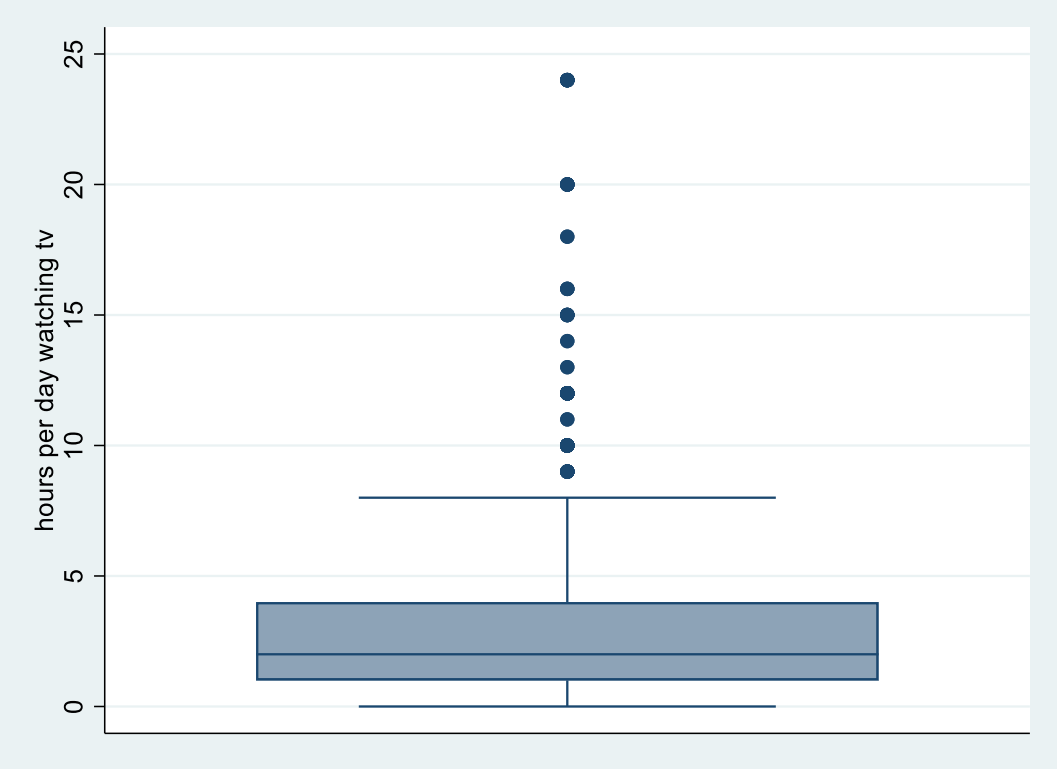
click on "Graphics"
click on "Box plot"

A window like this opens up:



Fill in the variable name of interest to you in the "Variables:" box; choose whether you want the plot to have a horizontal or vertical orientation, and then and click "OK"; boxplot options are available, but none are necessary for a straightforward plot.

Either way–via the menu or via the command line approach–you will then see the following graph in the "Stata Graph" window:



The shaded area contains the interquartile range (IQR); about 50% of people watch between 1 and 4 hours of TV per day, with a median of 2. 25% watch less than 1 hour, and the other 25% watch more than 4 hours. Persons watching more than 8 hours per day are identified as "outliers" (note that at least one person says that they watch TV all the time!).

Saving Graphs

To save your graph(s), see the "Saving Graphs" handout.