

Introduction to Computing for Sociologists

Neustadtl

Graphing Data

Stata has a large number of graphic types and options upon options for each graphic. Two good resources are Mitchell's book *A Visual Guide to Stata Graphics* (Stata Press 2012) and the following UCLA Web site: www.ats.ucla.edu/stat/Stata/library/GraphExamples/default.htm.

We will examine four types of graphic and some of the complications. Our goal is not to produce publication ready graphics (yet), but to learn about the basic graphic types and related issues. Most but not all Stata graphics are subsumed under the general command `graph twoway` followed by a specific graphic type (e.g. `scatter`, `histogram`, etc.) However, not all graphics are twoway graphics (i.e. `graph box`, `hist`, `kdensity`, `spikeplot`, `avplot`, etc.).

`histogram` (`help graph twoway histogram` or `help histogram`)

- `histogram` produces a frequency histogram and by default makes guesses about how many intervals and how wide the intervals should be.
- You can control the number and width of the intervals or use the `discrete` option to force one interval for each unique value (useful for variables with a small number of response outcomes).
- You can scale the histogram by density (the default), fractions, percentages, or frequencies)
- *kdensity* is a popular alternative.

`box` (`help graph box`)

- AKA box-and-whisker plot.
- Summarizes five numbers: the smallest and largest observations, Q1, Q2, and Q3.
- Also shows outliers.
- Usually used with the `over` option to make across group comparisons.

`bar` (`help graph bar`)

- `graph bar` draws vertical bar charts.
- The y -axis is numerical, and the x -axis is categorical.
- by default the height of the bars is equal to the mean but many statistics are allowed
 - mean, median, percentiles, standard deviation, standard error of the mean, sum, count, max, min
- Like `box` the `over` option can be used

`scatter` (`help graph twoway scatter`)

- Cartesian coordinates are used to display values for two variables
- Typically used for two continuous variables.
- Can be difficult to interpret if the variables are not truly continuous (but, plot means or jitter).

hi stogram

Histograms provide an overview of the distribution of a single variable. They can be used to compare a single measure across subgroups (e.g. men compared to women) and can provide visual confirmation of the normality assumption of the dependent variable in regression analyses. See the help file for `kdensity` for a popular alternative.

Often the default histogram in Stata is sufficient but Stata makes assumptions about how to collapse your data. The simplest command is `hi st varname` and I use it often. But, `graph twoway` provides more options. Consider the following examples:

```
#delimit ;  
graph twoway (hi st age), title("No Options") name(hi st1, replace);  
  
graph twoway (hi st age, bin(10)), title("Bin=10")  
    name(hi st2, replace);  
  
graph twoway (hi st age, width(5)), title("Width=5")  
    name(hi st3, replace);  
  
graph twoway (hi st age, discrete), title("Discrete Option")  
    name(hi st4, replace);  
  
graph twoway (hi st age, discrete percent),  
    title("Discrete/Percent Options") name(hi st5, replace);  
  
graph combine hi st1 hi st2 hi st3 hi st4 hi st5, nocopies;  
  
graph twoway (hi st age, by(race) width(10) discrete percent),  
    title("By Option") name(hi st6, replace);  
  
#delimit cr
```

Note that names have been assigned to every graph. Without this option the default name is *Graph*. Each time a graph command is executed it will write over *Graph*. This is avoided by assigning names to each graph. These graphs also use the *title* option to describe the difference between these examples. You can control the axis labels and use other annotation on every graphic. The command `graph combine` puts individual graphs into a single graph.

Read the online help (`help twoway histogram`) to understand the other options.

1. Create and interpret a frequency histogram of the sociability scale you created.
2. Compare histograms of sociability for white and black respondents (use the `by` option).

graph box

Box plots are interesting plots that provide information about the distribution of a variable. Typically box plots are used to compare distributions across groups.

Use the following `graph box` commands to produce four plots:

```
graph box age, over(mari tal) name(box1, repl ace)
graph box age, over(mari tal) over(sex) name(box2, repl ace)
graph box age, over(sex) over(mari tal) name(box3, repl ace)
graph hbox age, over(mari tal) over(sex) name(box4, repl ace)
graph hbox age, over(sex) over(mari tal) name(box5, repl ace)
```

The box represents the middle 50% of the distribution of the variable. The line in the middle of the box is the median—half of the distribution lies below this line and half above the line. The whiskers extend no more than 1.5 times the interquartile range (the distance between the 25th and 75th percentiles—the bottom and top of the box). Any points that exceed the whiskers are plotted as points and considered extreme outliers.

3. Create a box plot analogous to *box1* (above) with sociability. Interpret the distribution of the sociability variable across marital groups.
4. The box plots *box2* and *box3* are difficult to read so were created using `hbox`. Replicate the plots *box4* and *box5* using the sociability scale measure. Describe the difference between these plots.
5. Interpret the information contained in your last box plot (i.e. sociability over sex over marital).

bar

In Stata, bar graphs are not the same as histograms. Bar graphs typically require a continuous variable—a measure where the arithmetic average is meaningful—and a discrete variable (used as the `over` variable). For example, a bar graph could be used to graphically show the difference in average age (continuous) over marital status (discrete). Consider the following examples:

```
graph bar age, over(sex) over(mari tal) name(bar1, repl ace)
graph bar age, over(sex) over(mari tal) asyvars name(bar2, repl ace)
graph bar (p75) age, over(sex) over(mari tal) asy name(bar3, repl ace)
graph hbar (p75) age, over(sex) over(mari tal) asy name(bar4, repl ace)
```

The first two plots show the average of age as the height of the bar over marital status and then over sex. The second plot treats sex as a y-variable. The second two plots use the 75th percentile as the height of the bar. The last plot used `hbar` to create a horizontal bar graph.

6. Create and interpret a horizontal bar graph for the median of the sociability scale over race and over marital status.

scatter

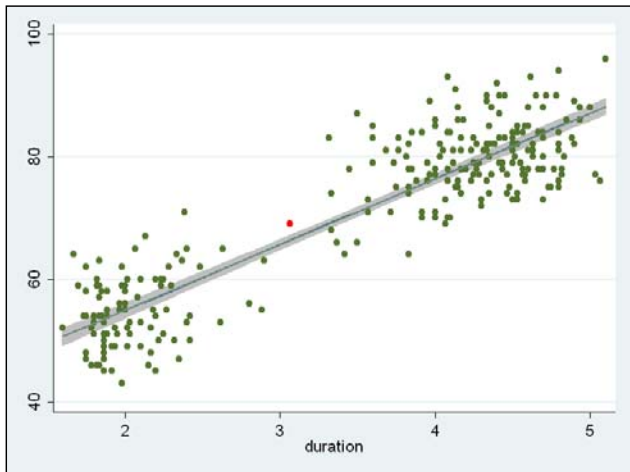
Producing scatterplots can be easy if you have truly continuous data. But, they are more complicated with typical survey data like the GSS. I will run through several examples to show some of the issues with accompanying Stata code.

Scatterplots in Stata

Old Faithful

The following Stata code and scatterplot show many of the features available in Stata. The data were taken from a standard dataset showing the relationship between the duration of an Old Faithful eruption and the interval to the next eruption.

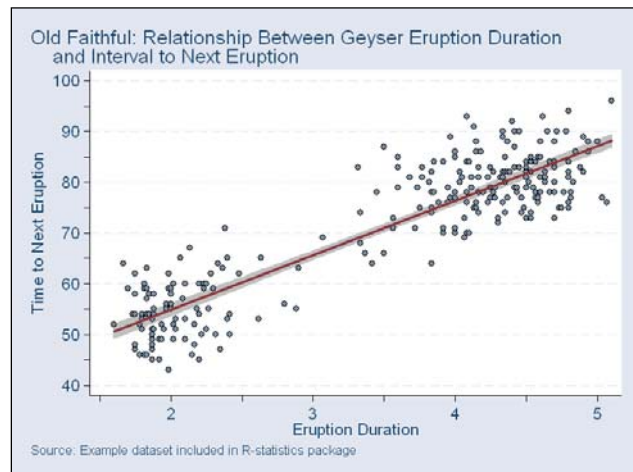
```
#delimit ;  
graph twoway (lfitci interval duration)  
             (scatter interval duration),  
             legend(off);  
graph export faithful.tif,  
             replace width(800) height(600);  
#delimit cr
```



Each point is placed using a coordinate. For example, the red point in this plot was placed at $X=3.067$ and $Y=69$. The code above will not produce this “red” point to keep the example simple. The line through the scatterplot corresponds to the linear best fitting line based on a regression analysis of interval regressed on duration. The grey shaded area represents the 95% confidence interval for the predicted values of this regression (i.e. the regression line).

This example shows the power of `graph twoway` by combining two separate graphs in one. The first graph, `lfitci`, shows the confidence interval. The second graph, `scatter`, plots the points in two-dimensional space. The order matters. If the confidence interval was plotted last it would cover the points under it.

As you can imagine, almost every aspect of this graph can be changed—the colors (points, line, confidence interval, background, etc.); the shape and size of the marker symbols; titles; axis marking and labels, and more and more and more! Here is a version using some of those features:

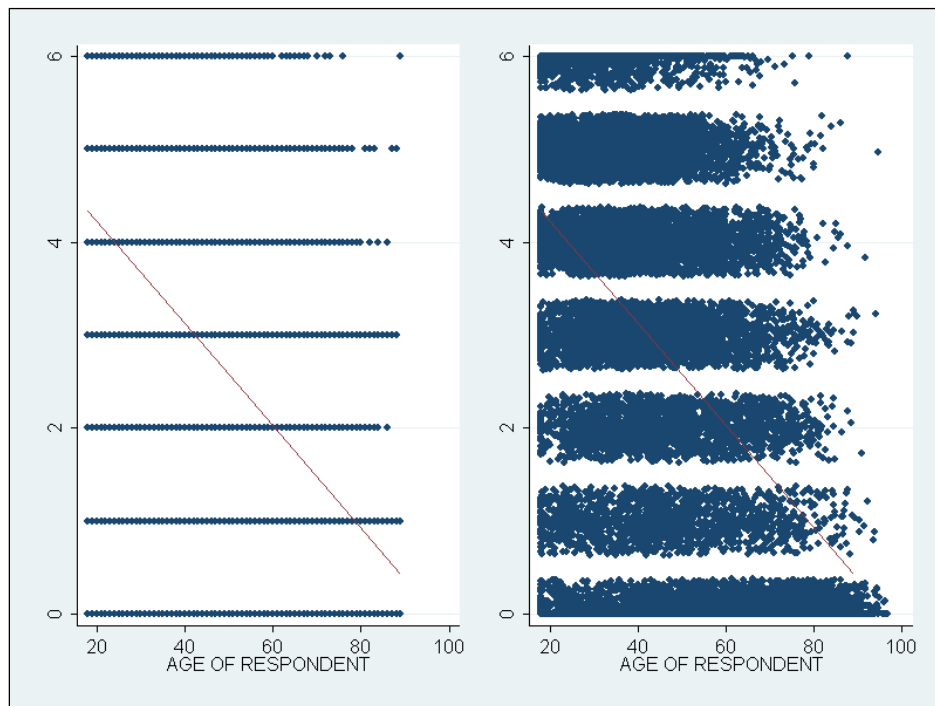


GSS: Sexual Frequency and Age

Typical Scatterplot

Consider the following Stata code and resulting graphic:

```
#delimit ;  
graph twoway (scatter sexfreq age) (lfit sexfreq age),  
    legend(off)  
    nodraw  
    name(sexfrq1, replace);  
  
graph twoway (scatter sexfreq age, jitter(10))  
    (lfit sexfreq age),  
    legend(off)  
    nodraw  
    name(sexfrq2, replace);  
  
graph combine sexfrq1 sexfrq2, nocopies name(sexfrqcmb, replace);  
#delimit cr
```



From a Stata perspective two scatterplots were created and combined into one graphic for easier comparison. Several new options are used. The *nodraw* option prevents the graph from being displayed. Since the plots will be combined they did not have to be displayed. The *nocopies* option (in `graph combine`) is undocumented but speeds up drawing the combined graph. Finally, the plot on the right uses the *jitter* option.

This example shows one of the difficulties of plotting discrete measures (sexual frequency) that aspire to be continuous variables. The fitted line shows that as age increases there is an associated decrease in average reported sexual frequency. However, without the fitted line could you determine that from the plot on the left? Sometimes symbols are plotted on top of one another so much that the relationship is obscured—the left plot. “Jittering” introduces random perturbations to provide some “spread” to

the plotted points. Sometimes this will help clarify the relationship. In this example—on the right—we do learn a little more by jittering.

Plotting Means

There are other ways to clarify this relationship. First let's create a new variable that converts the GSS measure to a yearly reported sexual frequency measure.

FREQUENCY OF SEX DURING LAST YEAR	Freq.	Percent	Cum.	#delimit ;
NOT AT ALL	4,403	22.34	22.34	recode sexfreq (0= 0)
ONCE OR TWICE	1,519	7.71	30.04	(1= 2)
ONCE A MONTH	2,058	10.44	40.49	(2= 12)
2-3 TIMES A MONTH	3,141	15.94	56.42	(3= 30)
WEEKLY	3,473	17.62	74.04	(4= 52)
2-3 PER WEEK	3,895	19.76	93.80	(5=156)
3+ PER WEEK	1,222	6.20	100.00	(6=208), gen(sexfreq1);
Total	19,711	100.00		#delimit cr

This Stata code converts the GSS categories “NOT AT ALL”, “ONCE OR TWICE”, etc. into numbers that approximately reflect the category. These are estimates and definitely inaccurate but useful with qualification. For example, all respondents who reported have sex “ONCE A MONTH” were assigned the value of 12 in the variable *sexfreq1* which is a yearly estimate. Obviously if we plotted this measure by age we would have the same overlap problem shown above. However, plotting the mean reported sexual frequency for each age group (every age) will reduce the clutter of the original scatterplot.

The Stata command `tabstat sexfreq1, by(age) format(%6.1f)` produces these values some shown below:

Summary for variables: sexfreq1	
by categories of: age (AGE OF RESPONDENT)	
18	60.8
19	61.3
20	76.8
.	.
.	.
.	.
87	4.4
88	5.2
89 or older	2.1

Total	58.3

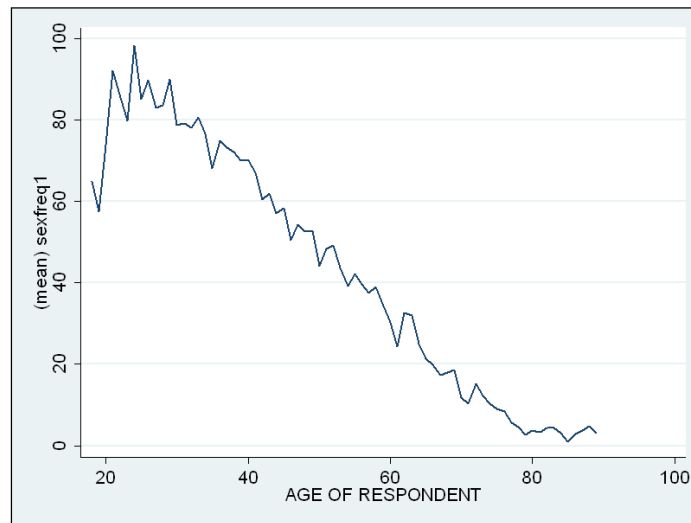
We want to plot age on the X-axis and average reported sexual frequency on the Y-axis without data input or too much trouble. Stata provides a way to do this but you need two new commands—`preserve/restore` and `collapse`.

We are going to create a new dataset that contains the values shown by `tabstat`. To do this we need to write over (i.e. destroy) the data in memory. But I don't want to lose these data. The commands `preserve` and `restore` deal with the programming problem where the data must be changed to achieve the desired result but, when the program concludes, the damage done to the data is undone. The following example 1) preserves the data, 2) creates a new variable, and 3) drops an existing variable (and anything else you might want to do). When the program ends the dataset you began with is restored. Anything between `preserve` and `restore` is undone at the end of the program. The indentation is just to improve the readability of the program.


```
preserve
  generate age=age^2
  drop age
  ...MORE STATA COMMANDS...
restore
```

The `collapse` command converts the dataset in memory into a dataset of means, sums, medians, etc. Putting these two commands together produces the following Stata code and plot:

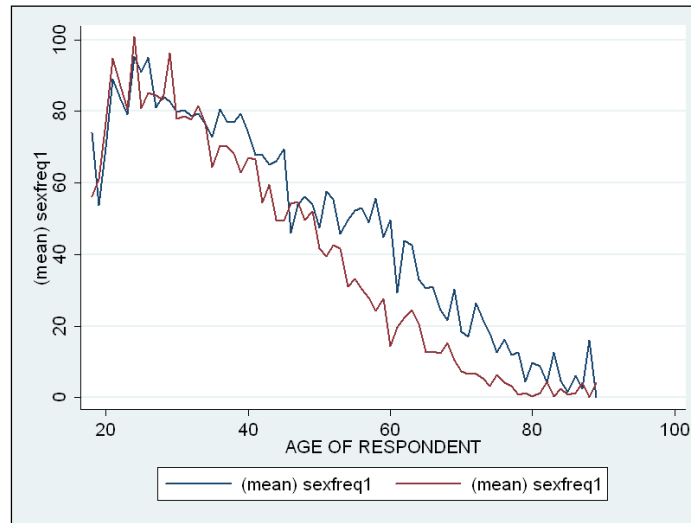
```
preserve
  collapse sexfreq1, by(age)
  graph twoway (scatter sexfreq1 age, msymbol(i) connect(l))
restore
```



This plot can be cleaned up (titles, labels, ticks, colors, etc.) but clearly shows the decline in average yearly reported sexual frequency as age increases. How is this line plot a scatter plot? Note the options `msymbol(i)` and `connect(l)`. The first suppresses the display of the marker symbols (“i” means invisible) and the second connects (the invisible) markers with a line.

This command is very flexible. Look at the following Stata code and accompanying graph that shows how to plot parallel trend lines:

```
preserve
collapse sexfreq1, by(sex age)
#delimit ;
graph twoway (scatter sexfreq1 age if sex==0, msymbol(i) connect(l))
              (scatter sexfreq1 age if sex==1, msymbol(i) connect(l));
#delimit cr
restore
```



7. Create and interpret a scatterplot of the sociability scale over time.
8. Replicate this plot but plot two series—one for married respondents and one for respondents who have never been married.