Graphical Descriptions of Data

Part I: Before Class

1) Read this assignment all the way through;
2) Know the terms and understand the concepts of:
    - scatterplots
    - stemplots
    - distributions
    - histograms

Background: Descriptions of datasets are very important for at least two reasons. First, they are necessary tools for communicating your results to others. Second, these tools provide insight for the researcher into the structure of the data set. Descriptive tools are roughly divided into two categories: numerical and graphical. It is somewhat artificial to separate the two, because in any real investigation of data a statistician or researcher will use both simultaneously. However, there are enough nuances to both groups, particularly as far as Stata is concerned, to make it worth our while to focus on graphical and numerical techniques separately. In this lab, you will use graphical techniques to make some conclusions about a data set collected from the Old Faithful geyser.

Objective: This lab has two main tasks. One is to show you different ways of downloading data files into Stata. It will also teach you how to use graphs to represent data. By the end of this assignment, you will know different methods for displaying distributions. The take-home activity asks you to be an analyst and give advice to the National Park Service and to travel industry executives on making sure that tourists get to see an eruption of "Old Faithful" in Yellowstone National Park.

Part II: In-class Activities

There are four activties. Activity 1 is optional (and just provided for your information) and it shows how to download datasets from the web when you're at home. Activity 2 teaches about loading "flat" files, and Activity 3 gives an alternative method that is sometimes useful. After a glossary of some useful Stata terms, Activity 4 gives some tips for dressing up your graphics.

Loading Stata Objects

Sometimes, the datasets you will use will be "pre-digested" for Stata. These files all have the form "filename.dta." These are the easiest to download.

You've already seen one method for downloading such data files (using the "use" command). Here we'll show you another. To get started, you should open Stata.

Activity 1 (Optional): Downloading Stata objects from the WWW

NOTE: This method does not work here in the lab. To use it at home, you must configure your browser properly. See the instructions for your browser.

a) Open the Internet browser (Netscape or Internet Explorer). Go to http://www.stat.ucla.edu/~rgould/oldfaith.dta. Stata will start to automatically download the Stata object. You don't even need to have Stata running to do this. This also works when the Stata file is a hyperlink text.
b) Or if the Stata object is already on your hard drive, just double click on the file and Stata will automatically run it for you.
c) The other way of downloading this is to type

use http://www.stat.ucla.edu/~rgould/datasets/oldfaith.dta

in the command window.

Entering Data from a "Flat File"

A "flat file" is a text file that contains rows and columns of data. Each row represents a single observation, and each column represents a variable. The data might be separated by a comma, space, or tab.

Fairly often in this class, you will retrieve such flat files over the Internet. You will go to the file in your browser (Netscape or Internet Explorer), and then save the file to your hard drive. Other times, you may be given the flat file through other means (e.g., email or "ftp").

Now we will practice retrieving a file from the Internet and loading it into Stata.

Activity 2: Loading "flat files"

1) Before entering new data into Stata, you need to erase the previous dataset. Type **clear** in the "Command" window and then hit "Return."

Once you have emptied Stata, you can load the flat file.
You can view flatfiles on the WWW before downloading them. Point your browser to http://www.stat.ucla.edu/~rgould/datasets/oldfaith.raw

Note that this is almost exactly the same address above, except the extension "*.dta" is now "*.raw".

2) Now type:

insheet using http://www.stat.ucla.edu/~rgould/datasets/oldfaith.raw

3) Next, type:

describe

4) Hit "Return."

What do you see? Can you identify the name of the variables in this dataset? Unlikely. In such cases you need to name the variables *a priori*. You should do the following:

5) Type:

clear

6) Hit "Return."

7) Once you cleared Stata, download the same file but using the following command:

insheet duration length day using
http://www.stat.ucla.edu/~rgould/datasets/oldfaith.raw

8) Hit "Return."

9) Again, type:

describe

10) Hit "Return."

Notice then that whenever a dataset doesn't come with the variables identified it is up to you to name them. In some other cases, however, this might be unnecessary for the dataset already will come with the variables identified. For example, in the file **oldfaith.dta**, which we used earlier, the variables were already named.

Activity 3: Alternative way of loading flat files

1st Step: View the data

1) Open the Internet browser (Netscape or Internet Explorer) and type in the "address" window, the location file. Thus, type:

http://www.stat.ucla.edu/~rgould/datasets/oldfaith2.raw
(Notice that in the file **oldfaith2.raw** the variables are already named.)

2) Hit "Return." You will see the data, i.e., the three columns of numbers.

3) Next, you must save the data to your hard drive. Go to the "File" menu and select "Save As." A window opens that gives you the choice to rename the file. For

now, accept the default name: oldfaith2.raw

4) Click on the "Save" button or hit "Return." Result: the file will be saved on your hard drive.

2nd Step: Load "oldfaith2.raw" into Stata.

1) In the "Command" window, type:

insheet using

DO NOT HIT "RETURN" YET!

2) Select "Filename" from the "File" menu on the upper left side of the screen. A dialogue box will open.

3) In the "Files of type:" window, select "All Files (*.*)." Then double-click on "oldfaith2.raw" or click on it once and press the "Save" button.

(Note: in case you don't find the file, you'll have to look for it in other directories.)

4) Hit "Return."

Using Stata

Commands you type are in boldface ("x", "y", and "z" represent the name of the variables). For many of these commands, you can type as many variables as you want. Feel free to experiment.

use *filename* **->** loads a Stata-format file dataset from the Web. If filename is specified without the Stata extension, ".dta" is assumed.
clear -> erases dataset from the Stata program.
insheet using *filename* **->** reads unformatted ASCII (text) data. It is intended for reading files created by spreadsheet. This command is especially useful when you don't know the variables (their amount and names) of a dataset.
infile x y z **using** *filename* **->** reads into memory a dataset that is not in Stata format. However, this command is good only for more advanced data analyses. We will NOT be using it in our assignments. Instead, always prefer the **insheet** command.
dotplot x y **->** for comparative dotplots.
stem -> for stem-and-leaf displays.
graph x y**, box ->** specifies box-and-whisker plots for two variables. Up to six variables may be specified.
graph x y**, symbol ( ) ->** changes the symbols of the variables in scatterplots.
graph y z**, pen ( ) ->** changes the color of graphs and data points.
graph x **->** for histograms. Only one variable can be specified at a time.
graph x**, bin (** # **) ->** specifies the number of bins in a histogram.
graph y**, title ( ) ->** gives title to graphs.
chist x **->** computes histograms and frequency distribution for a continuous variable broken into class intervals.

Activity 4: Changing graphic settings

Stata allows you to modify and improve the settings of displayed graphs, including adding titles and changing labels and colors of data points in scatterplots. Using the same graphic output from question "5" do the following exercises:

(Look at the next page for lists of symbols and colors used in Stata)

1) Make a scatterplot of duration against length (graph duration length). Add a title to your graph.
2) Change the signs of the data points in your graph. Experiment with different symbols, but , as a general rule, you should avoid visual clutter.

3) Change the colors of the data points as well. Choose your favorite color.

List of Codes for Symbols

O	large circle
S	large square
T	large triangle
o	small circle
d	small diamond
p	small plus
.	dot
i	invisible
[*varname***]**	contents of variable to be used as a text symbol
[_n]	use observation numbers as symbol

List of Codes for Colors

1	blue
2	yellow
3	red
4	green
5	purple
6	dark blue
7	violet
8	lilac
9	pink

Note about the "Help" system

Stata has a comprehensive **Help** system. However, there are two methods for getting help that are particularly useful: **Help/Search** and **Search**. You should know when to use each of them.

1st Option: Help/Search

You use "Help/Search" when you know the name of the Stata command on which you want more information.

- ➔ From the main menu bar, go to "Window" and click on "Help/Search." A box will appear. You just need to type in the name of the command and click on "OK."
- ➔ Alternatively, you may go to "Help" on the main menu bar and click on "Stata Command…". Next, type in the name of the command and click on "OK."

2nd Option: Search

You should use "Search" when you need to find out the name of a command.

- ➔ Go to "Help" on the main menu bar and click on "Search..." Next, type in the word or expression on which you need information and click on "OK." Note that Stata commands will always appear in green. For example, if you type "graph" in this box, Stata will show you a long list of all graph related options. Just scroll down on the list and choose the topic that is closest to the information you need.

Part III: Take-Home Activity

Old Faithful

Old Faithful is a geyser in Yellowstone National Forest. This geyser earned its name from the regularity of its eruptions. This regularity, of course, also makes it ideal for a tourist attraction. Prior studies conducted by the travel industry have shown that tourists are (a) more likely to return to an attraction if their expectations have been met and (b) are more likely to recommend visiting the attraction if their expectations are met. If a geyser erupts infrequently, and with no discernable pattern, then how are tourists to know when to visit? (Ironically, Old Faithful is not the most faithful geyser in Yellowstone National Forest).

Ideally, as a travel industry analyst, you should be able to make recommendations to the National Park Service and to various travel industry leaders as to when tourists ought to make a visit to "Old Faithful" during their trip to Yellowstone National Forest.

The dataset

The dataset records the time between eruptions, the length of each eruption, and the day on which the observations were recorded. One possible use for these data is to see if we can predict how long before the next eruption occurs.

Data analysis

1) What are the variables in this dataset? How many observations are in these data? Tip: use the command **describe** to answer these questions.

2) How would you describe the variables? Are they continuous, discrete, or categorical?

3) Describe the distribution of the duration of the eruptions. Do you have an explanation for the shape you see?

4) Describe the distribution of the length of an eruption. How does it compare to the distribution of the duration?

5) What is the relationship between the length of an eruption and the time between eruptions? Describe it with as much detail as possible.

6) Make a histogram of either the length or duration variable. Explore the effects of changing the number of bins on the shape of the histogram. How many bins do you think should be displayed for these data?

7) In the preparation of your report, you decide to visit Yellowstone yourself (any good analyst will make it a point to schedule a site visit -- and get paid for it!). An eruption has just ended, and a tourist (who just missed it) comes up and asks the Park Ranger how long until the next eruption. The Park Ranger says she doesn't know. Using your graphs, can you provide a better answer? Be as detailed as possible.