Statistics 104 – Autumn, 2004 – Assignment 1

Due Friday, October 1. Homework is handed in at the Friday's lecture. Please include your name, your TF's name, and your section time at the top of your assignment.

Readings (Moore and McCabe)

- Sections 1.1 and 1.2
- Sections 2.1 and 2.2

Against All Odds videotape

The relevant tapes for this week and next week are numbers 1 (what is statistics?), 2 (picturing distributions), 3 (numerical summaries), 8 (relationships), and 9 (correlation). This is completely optional supplementary viewing!

Written Assignment (Moore and McCabe)

Please show your work, especially on odd-numbered problems with answers in the back.

MM 1.22, 1.28, 1.32, 1.38, 1.42, 1.44, 1.54, 1.76

Computer Assignment (STATA)

This assignment is intended to get you familiar with using Stata, the statistical software for this course. Stata is a very powerful software package and is especially popular in economics and medical sciences. The time you put into learning this software will benefit you throughout your undergraduate career, and beyond. You may find this software useful for other classes, and for practically any senior thesis involving data analysis.

This assignment also should demonstrate that there often is more than one way to approach the problems we will be addressing in this class. Applied statisticians often need to be creative investigators!

To complete this assignment you will only need to refer to the *Basics of Stata* handout. You will be asked to *read in* data, *summarize* it, and make *graphs*. When you are finished, you will *print* the graphs along with a transcript of the your session.

Remember, any time you hand in computer printouts they should be cut or edited to include only the relevant parts. Please do not just hand in a pile of output as it comes out of the computer! Follow the instructions given below. I suggest running a word processor at the same time as Stata and copy and paste the generated tables and graphics into the word processor document.

The data set you will examine is called `homeruns.dta` and contains information on homeruns for seven baseball players. You can access the data from the class homepage
`http://www.courses.fas.harvard.edu/~stat104`.

The players are Babe Ruth (played 1914-1935, primarily in New York), Hank Aaron (1954-1976, Milwaukee), Roger Maris (1957-1968, New York), Mark McGuire (1986-2001, St. Louis), Ken Griffey Jr. (1989-2002, Cincinnati), Sammy Sosa (1989-2002, Chicago), and Alex Rodriguez (1994-2002,

Seattle). The baseball season runs from March to September. These seven players are known for, among other things, hitting homeruns. The data set records four variables:

Variable name	Variable meaning
player	The name of the baseball player.
year	The season that the following homerun were hit in
homerun	Number of homeruns in one baseball season.
atbats	Number of batting attempts in one baseball season.

We want to compare the performance of these seven players focusing on homerun hitting. You should turn in graphs and answers to the questions.

1. Download the data from the class homepage: Go to the Stat 104 homepage and click on Assignments. Click on the link for `homeruns.dta` and save it to disk. (You might have to select Save As under File if clicking on the link does nothing on your machine.) After you have downloaded the file, you can close your web browser if desired.

2. If you will be installing Stata on your personal machine, do it following the instructions online. If you do not already have it installed, you will also need to download and install the key server program. Wherever you are using Stata (personal machine, one of the Science Center machines or via the network from your dormitory), practice getting into Stata following the instructions on the *Basics of Stata* handout.

3. Double click on the icon for `homeruns.dta`. It should open in Stata. At other times, you will get into Stata by following the instructions on the *Basics of Stata* handout. If the data set does not open when you click on it, get into Stata and try to open it there. If you are opening it within Stata, you will need to find the list of files in the directory where you have stored the data file. To bring various windows into the foreground, look under Window.

4. Log your session to obtain a transcript. You should do this *first* to make sure that all of your output goes to a file. You can either do it via the menus, by going to `File` $>$ `Log` $>$ `Begin` and giving a log file name, or by typing the command

   ```
   . log using hw1out
   ```

   in the Stata Command window. Note that most Stata commands can be done by the menu system or by typing a command. All commands run are displayed in the Review window and stored in a log file if requested.

   When you are logging your session type the command `log close` or go to the menu `File` $>$ `Log` $>$ `Close`

5. Summarize the data with the `summarize` command. This can be performed via the menu system with `System` $>$ `Summaries, tables,& tests` $>$ `Summary statistics` $>$ `Summary statistics`, or just by typing the command `summarize`. Why doesn't Stata compute numerical summaries for the variable `player`? Type `tabulate player` to get a

summary of the variable `player`. Why does the frequency of *Aaron* have the value 23? Type `list in 1/10` to see the first ten values of the data set.

6. Make graphs of the data. Briefly describe the distribution of homeruns and batting attempts and the relation between the variables.

   You should produce *histograms* of the two variables (homeruns and atbats) and then plot one variable against the other using the following commands. To create the histograms, go to the Graphics menu and select Histogram under Easy graphs or choose Histogram directly (this version has more options). For the two histograms, set both of them to have 8 bins. Look through the different options in the dialogue box to find where to set the number of bins, or click on the question mark to go to the help page. Next you should create a scatterplot with atbats on the x axis and homeruns on y axis. This can be be done with the Scatter plot entry under Easy graphs (`Graphics` $>$ `Easy graphs` $>$ `Scatter plot`). Or the following lines can be typed in the Command window. The saving option saves the graphs as Stata graph files. This can also be done by right clicking on the plot (in Windows) or clicking and holding (with a Mac) and selecting save graph.

   ```
   . graph homeruns, hist bin(8) saving(histhr)
   . graph atbats, hist bin(8) saving(histbat)
   . scatter homeruns atbats, saving(plothrbat)
   ```

   Note: If you omit the `saving(`*`filename`*`)` modifier above the plot will still be displayed on the screen, but not saved in a file for later printing. You should omit this modifier when you are exploring data sets, and use it only when you find a plot you wish to print later.

7. Find the correlation between two variables, `homerun` and `atbats`. Find out how to do this by typing `help`, and exploring. This can also be done through the menu system (`Statistics` $>$ `Summaries, tables,& tests` $>$ `Summary statistics` $>$ `Correlations & covariances`). Why is the correlation positive? Why is the correlation less than 1?

8. If a player is on a better team, the player might get more chances to bat in a season, and thus might get more homeruns (since the correlation is positive). Let's calculate the *rate* at which players hit homeruns. What does the distribution of rates look like?

   ```
   . generate hrrate=homerun/atbats
   . graph hrrate, hist bin(8) saving(histrate)
   ```

   The generate command can also be done with the menu system (`Data` $>$ `Create or change variables` $>$ `Create new variable`).

   Here is a trick for reducing the amount of paper you use in printing. Graphs can be combined on the same page. If you print the graph named plothw1, you will get the four previous plots together.

```
. graph using histhr histbat plothrbat histrate, saving(plothw1)
```

Note that this will only work if you save the graphs. Also look around to see if you can figure out how to do it through the menu system.

9. Now let's compare the players. Calculate and examine the average number of homeruns and average homerun rates for the seven players. How do the players compare? Does looking at rates versus homeruns make any difference? What difference does it make? This can be examined by tabulating the data. Go to (`Statistics` $>$ `Summaries, tables,& tests` $>$ `Tables` $>$ `Tables of summary statistics`). Set the row variable to be `player` and request the statistics `frequency, mean`, and `standard deviation`. Then generate a boxplot of hrrate for each player. This can be done by going to (`Graphics` $>$ `Easy graphs` $>$ `Box plot`). Set the variable to `hrrate` and the Over1 variable to be `player`. Or you can do everything by the following commands. Note that these commands will be slightly different than the commands generated in the Review window if you use the menuing system to answer this question.

```
. sort player
. tabulate player, summ(homerun)
. tabulate player, summ(hrrate)
. graph hrrate, box by(player)
```

10. One consequence of averaging the homerun rates is that it treats all seasons the same and does not give more weight to seasons in which a player has more times at bat. A different summary is dividing the total homeruns by total times batting over a player's entire career. We will enter this data and then compare the players one more time. Type the command `clear` removes the previous data. Then select `Data editior` under the `Data` menu and add the following data to the data window. To change a variable name, double click on the column heading and type in the change. For the variable `player` make sure the format is `%9s` (you can also use a number bigger than 9). Also create a variable `hrrate = homerun/atbat` similarly to before. Once the data set is created you should save it.

    How do the players compare when homerun hitting ability is measured this way? Find reasonable summaries to do this comparison.

```
player homerun atbat
Aaron 755 12364
Griffey 468 6913
Maris 275 5101
McGuire 583 6187
Ruth 714 8399
Sosa 499 7026
Rodriguez 298 4382
end
```

11. Exit Stata and print output.

12. Organize your output and make comments.

    A transcript of your session is in the file '`hw1out`', or the file name you chose. Remove any mistakes and other extraneous text. You can do this electronically by editing the file, or by cutting, with scissors, and taping the worthwhile output to standard sized paper. On the graphs, comment on any interesting features you see.