Exercise No. 2: Entering Data, Prettying Up Graphs, etc.

New Stata Commands		Old Commands Reviewed	
Data Editor	label variable	describe	
rename	graph set print logo on/off	summarize	
graph twoway scatter	graph hbox	use	
file/save	by		
clear	label define		
gsort / sort	label values		
list	graph twoway lfit		

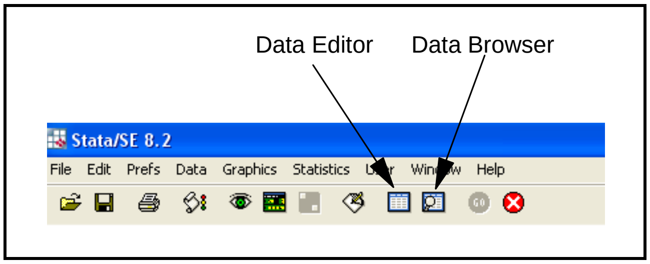
{Be sure to open a log and keep it for your records}

As you complete your exercise, be sure to include in the exercise the material requested in paragraphs that are check-marked (✓).

1 Graphing two time series variables versus year

Try the following exercise:

- Start Stata and click the Data Editor icon (Note: the Data Browser allows you to look at a data set, but it allows no changes to the data)



- Type in the 4 columns of data (but not the headings) as shown:

Medical School Enrollment 1986 to 1992
(in Thousands)

year	men	women	total
1986	10.5	5.6	16.1
1987	10.2	5.8	16
1988	10.1	5.9	16
1989	9.8	6	15.8
1990	9.8	6.2	16
1991	9.8	6.4	16.2
1992	9.5	6.8	16.3

Source: Mendenhall, Beaver & Beaver, p. 63.

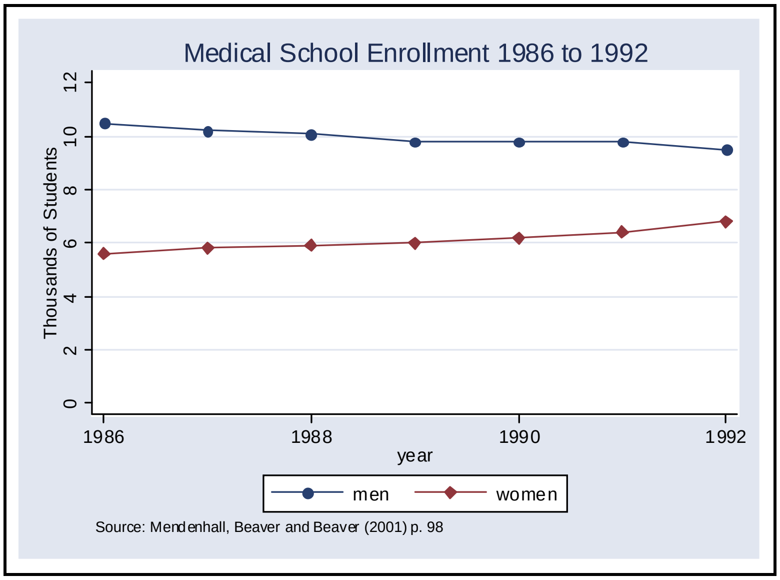
- When you close the data editor, the 4 variables will be named “var1”, “var2”, “var3”, and “var4”. Rename each variable, giving it the name in the first row of the above table. (Type "help rename" in the command line to see how to do this)

- Create the graph using this command”

```
graph twoway scatter men women year, ylabel(0 (2) 12) xlabel(1986 (2) 1992)
ytitle("Thousands of Students") c(l l) msymbol(O diamond) ti("Medical School
Enrollment 1986 to 1992") note("Source: Mendenhall, Beaver and Beaver (2001) p.
98")
```

Note: In a Stata command, you will often see key words with a parentheses "( )" following. In every case, make sure that the key word and the following parentheses are <u>not</u> separated by one or more spaces.

- Do you get this graph?



Medical School Enrollment 1986 to 1992

Source: Mendenhall, Beaver and Beaver (2001) p. 98

- ✓ You will use the EC400tp.dotx template to create an exercise to turn in. Paste this graph and the command used to create it into the document.

- ✓ Write a short paragraph explaining how each of the options in the command were used to create this graph. Use Hamilton, Chapter 3 and Stata’s own “Help” facility to explain how each of the options are used to create the graph. (This would be a quite acceptable graph to include in a term paper, article, etc.)

You might need to use the medical school data again, so save the data set to a place where you can find it. Call the dataset “medschool”. Click “file/save” on Stata’s toolbar and save it to a folder of your choosing.

Then, type “clear” in Stata’s command window to clear Stata’s memory.

2 Doing Some Analysis on ACC Basketball Statistics

In 2001 and 2002, at the end of the men’s basketball season, the *ACC Area Sports Journal* published a numerical index rating of each player who played a minimum number of minutes in league games. The index was comprised of a set of offensive and defensive statistics on each player, for league games only, and purports to show the player’s overall “effectiveness” with high values implying “very effective” and low values the opposite.

- Download “acceff.dta” from the course web site and load it into Stata.

- Type “describe” to observe the variables.

- Use the “gsort” command to sort the data first by the variable “year” and then (in descending order) by the variable “-eff” (a minus sign in front of “eff” produces a top-down ordering), which is the player efficiency rating. (Stata has two commands which sort the data by the size of one or more variables, “gsort” and a more limited command “sort” that only sorts from smallest to largest values of one or more variables.) Use Stata’s “help” system to figure out how to get the sorting done correctly.

- “describe” the data set again and notice that Stata now tells you that the data are sorted by year (2001 first) and then by player efficiency score. (Strangely, it will *not* indicate that the data are also sorted in descending order by “eff”. If the “eff” data had been sorted in *ascending* order then the “describe” command would indicate that the data had been sorted by both variables ... go figure)

- Now, list the observations in the data set. First list the year 2001 data by using the command:

```
list if year==2001
```

✓ Then, list the observations from the year 2002, and paste both years’ listing into your exercise. Note: Be sure to present these data using the Courier New typeface. That way all the names, columns, lines, etc. will line up properly.

- Let’s label some of the variables so that they are more informative:

```
label variable eff “ACC Player Efficiency Score”
```

Then, using the above command as a model label the other variables “Year”, “College/University”, and “Player Name” (See Hamilton pp. 17-19, “Getting Started with Stata”, chapter 8, and the Stata “help” file for more information)

✓ “describe” the data so that you can see that each variable now has an informative label, and paste the result into your exercise.

- Now, for the year 2001, let’s produce a set of box plots that show, for each team, the spread of individual efficiency ratings. But first, let’s suppress a Stata “feature”: When printing graphs directly from Stata, the program automatically includes the Stata logo on the graph. I find this annoying; if I want to advertise for Stata, I’ll do it. Otherwise, don’t bug me. So, enter the following command into Stata:

```
graph set print logo off {enter}
```

This turns off the automatic logo printing. If you want it back on at a later date, just enter the command again with “on” instead of “off.”

- Let’s build a complete box plot in steps:

```
graph hbox  eff if year==2001, over( schl)
```

This command does a horizontal box plot for year 2001 only and produces a separate box for each school.

- Now, let’s add a line showing the overall median player efficiency score for the 2001 season. First, we need to determine what that value is:

```
summarize eff if year==2001, detail
```

You’ll find that the median value for 2001 was 0.679. Let’s add that line to the box plot we just did:

```
graph hbox  eff if year==2001, over( schl) yline(.679) note("Median
efficiency score for all the players is 0.679")
```

The graph would be easier to interpret if we ordered the schools by the efficiency scores of their median players, and added a title and subtitle:

```
graph hbox  eff if year==2001, over( schl, sort(1) descending)
title("ACC Player Efficiency Scores by School") subtitle("2000-2001
Season") yline(.679) note("Median efficiency score for all the
players is 0.679")
```

The “over” option now orders the boxes over the schools, sorting them by the median value of their efficiency scores, in descending order. I also added a title and subtitle for the graph. Notice that we’ve built up a pretty complex graph, one step at a time. This is a good way to go about making Stata graphs, especially the complicated ones.

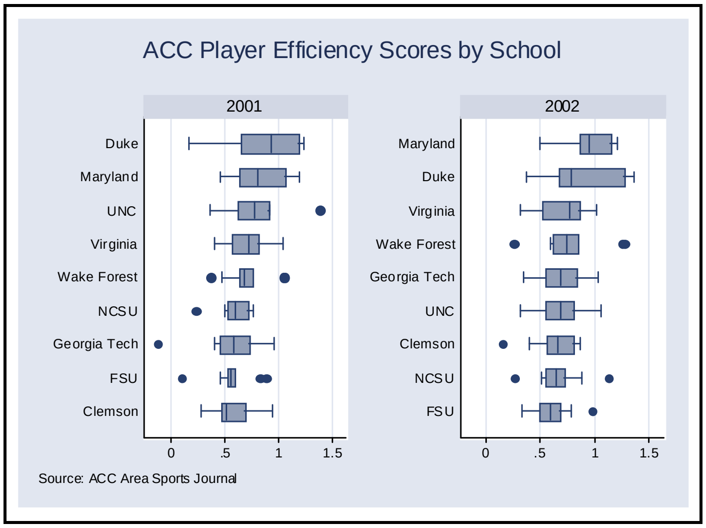
✓ Copy this last graph into your exercise, and, right below the graph, write a sentence or two about each school’s player efficiencies, noting similarities and differences based on your reading of the box plots

✓ Now, go through the same procedure to produce a graph for the 2001-2002 season. (Make sure the labels accurately reflect which season you’re graphing.) Copy the resulting graph into your exercise and, again, write a short description of each school’s player efficiencies. What seems to have happened to UNC between 2001 and 2002?

- Finally, we’ve done box plots for each year individually. Suppose we want to do a comprehensive box plot that shows both years in one graph. That’s a large and potentially messy graph. Stata provides a neat way to combine a large number of box plots efficiently and attractively:

```
graph hbox  eff, over( schl, sort(1) descending) ylabel(,grid)
ytitle("") by(year, title("ACC Player Efficiency Scores by
School",span) subtitle("  ") note(Source: ACC Area Sports Journal))
```

✓ You should get the following graph. I don't really expect you to grasp how to do this yet. I just want to show you some pretty cool stuff can be done. Paste the graph into your exercise.



Source: ACC Area Sports Journal

3 Creating A New Aggregate ACC Data Set

Until now, we've been using an individual level data set to describe player efficiencies for each team. Now let's construct a new data set for each team for each of the two years. There are nine teams in the league, so we'll have 18 observations. Each observation (or case) will contain the following variables:

1. School
2. Median Player Efficiency
3. ACC Winning Percentage
4. Season (i.e., either 2001 or 2002)

- To generate the data, we need first to sort the data by year and by school:

```
gsort year schl
```

- Then we need to compute the median efficiency score for each school for each year

```
by year schl: summarize eff, detail
```

- Then we need to compute each school's winning proportion (in ACC play only) based on the following league results:

2001		
Duke	13	3
N Carolina	13	3
Maryland	10	6
Virginia	9	7
Wake Forest	8	8
Georgia Tech	8	8
NC State	5	11
Florida State	4	12
Clemson	2	14
2002		
Maryland	15	1
Duke	13	3
Wake Forest	9	7
NC State	9	7
Virginia	7	9
Georgia Tech	7	9
N Carolina	4	12
Florida State	4	12
Clemson	4	12

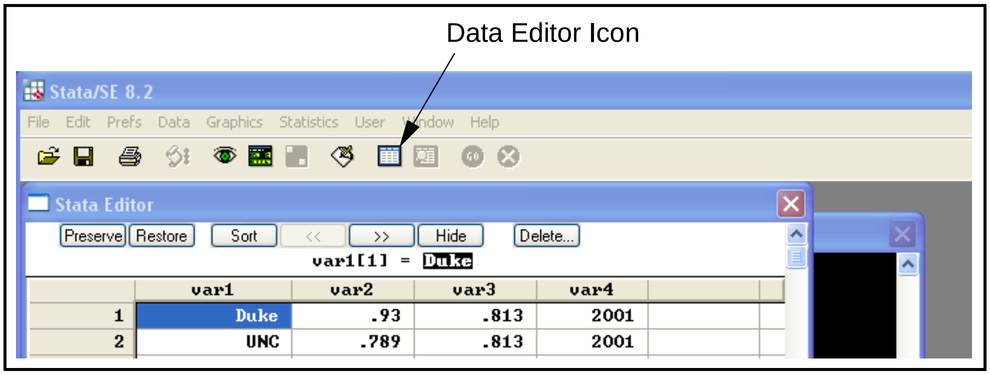
You need to convert each team's won-loss records into a proportion of wins. Each team plays 16 league games so Duke's proportion of wins in 2001 is $\frac{13}{16} = 0.813$. Calculate this for each team for each year (to 3 decimal places).

- Clear the current data set (acceff.dta) out of Stata's memory by typing the command:

```
clear {enter}
```

- Now, you need to enter your data into a new Stata data set. Click the Stata Editor Icon (see picture below) and a blank spreadsheet-like table will appear. You will now enter 18 rows of data (one for each team for each year), each row containing 4 variables. ***Do not enter the variable <u>names</u>; enter only the <u>numerical data.</u> If you do enter the variable names, Stata will think that the variable represented in that column is a "string" variable, not a numerical variable and you won't be able to continue the exercise. If that happens, exit the data editor and clear all data from Stata before starting over again.*** Starting in the top left cell enter — for the first school in the first year — into the first row the 4 variables in this order: School Name (var1), Median Efficiency Score (var2), Proportion of Wins (var3), and Season (var4). To move from column to column (on the same row) strike the Tab key. To move from row to row, press the Enter key. Enter all 18 observations, following the format shown below. Stata assigns a variable name to each column of data (var1, var2, etc.). Remember, *do not* enter the variable names into the editor, you'll change them when you're through entering data.

  When you're done click "Preserve" and exit the Stata Editor.


Data Editor Icon

var1[1] = Duke

	var1	var2	var3	var4
1	Duke	.93	.813	2001
2	UNC	.789	.813	2001

- Save the new data set, calling it “ACC_team_data”. Then rename the variables “var1”, “var2”, etc., and save the data set again. The appropriate commands are:

```
rename var1 School
rename var2 mef
rename var3 wins
rename var4 Season
```

- Now we’re in a position to see how a team’s player efficiency relates to its success in league games. Have Stata do a scatter plot using the command:

```
graph twoway scatter wins mef if(Season==2001)
```

✓ Paste this graph into your exercise.

This is a rather plain graph. Let’s pretty it up by adding a title (“Relating League Wins to Player Efficiency, 2000-2001 Season”) and labelling each point with the school name. The problem is that some of these school names are rather long and might not look too good on the graph.

- Let’s create a new, numerical variable identifying each school to which we can attach a short label. Open the Data Editor as before and click in the header of the “School” column. Now, click the “Sort” button in the editor. This sorts the data alphabetically by school.

- In the blank column to the right of the last column of data enter numbers starting with one for the first school (Clemson), 2 for the second school (Duke), and so on. Click “preserve” and exit the Data Editor. Rename “var5” as “shortname”.

```
rename var5 shortname
```

- Now, we’re going to define labels for each value of “shortname”. We define a label “shrtnames” as follows:

```
label define shrtnames 1 "Clem" 2 "Duke"  3 "FSU" 4 "GATech"  5 "MD"
6 "NCSU" 7 "UNC" 8 "VA" 9 "WF"
```

- Then we attach the labels in “shrtnames” to each value of “shortname” as follows:

```
label values  shortname shrtnames
```

✓ Now enter the command “list” and paste the resulting table of data into your exercise.

- Now try the improved graph again with the following command.

```
graph twoway scatter wins mef if(Season==2001), mlabel( shortname)
ti("Relating League Wins to Player Efficiency, 2000-2001")
```

Now, we’ve got a title and each point is labelled with the school name.

- Finally, we can get a rough idea about how well player efficiency affects team performance by fitting a least squares line through the data. Also let's make the axes more informative by labelling them:

```
graph twoway lfit wins mef if(Season==2001)||scatter wins mef
if(Season==2001), mlabel(  shortname) title("Relating League Wins to
Player Efficiency, 2000-2001 Season")  ytitle("Season ACC Winning
Proportion" " ")  xtitle(" " "Median Team Player Efficiency Score")
```

✓ Paste the preceding command and the resulting graph into your exercise.

This is a quite complicated command that produces a good looking (but not quite perfect) graph. It superimposes two graphs (the least squares line and the scatter plot) upon each other. (The || symbol does the superimposing). It labels each school with a short version of its name (mlabel(shortname)), and it labels the axes with a better description (The " " quotes add some space between the axis labels and the axis titles).

✓ Create a similar graph for the 2002 season, being sure to get all the titles and labels right. Paste the resulting graph into your exercise.

4 Questions to Answer

1. Based on their median player efficiency scores, which teams would you have expected to have the best chance at winning the NCAA championship in 2001 and 2002? Did they, in fact, win?
2. During league play in 2001, which teams apparently performed better than their median player efficiencies would have suggested? Performed worse?
3. During league play in 2002, which teams apparently performed better than their median player efficiencies would have suggested? Performed worse?
4. In 2002, which team appears to have gotten the least out of the players that they *did* have?
5. Using the two linear fit graphs, how would you assess the effect of factors other than player efficiency on the teams' success in the ACC? I.e., could you discern a "coaching effect" from this graph? Why?/Why not?