Exercise No. 3 : Dropping Variables & Observations, Central Limit Theorem

New Stata Commands	Old Commands Reviewed		
browse	describe		
drop *varname*	graph box		
drop if(*condition)*	label variable		
save	ylabel(....)		
tabulate	xlabel(....)		
qnorm	use		
histogram	summarize		
over(...)			
set seed			
bs			

1 Introduction

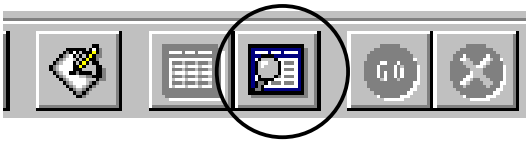
This exercise introduces some data management tools and illustrates a fundamental principle of statistical sampling theory. You will learn how to eliminate some variables from a data set, and you will also learn how to eliminate the missing values from the data set.

First, you need to download a data set from the course web site. You will be using the data set **'film_modified.dta'.** So, if you haven't already done so, download the data now. Then, start Stata and bring in this data set. You may remember these data from an earlier handout that instructed on how to create a boxplot. In that handout the above data set was created; it contained data on the gross revenues of 9 motion picture studios during a recent three-year period. In order to create a boxplot for the combined data of all nine studios, I created a new variable called "**studios**"; the creation process also created another variable "**_stack**" that contains a number from one to nine, which shows from which of the nine studios each revenue number comes. Remember, each observation in **studios** is the gross revenue of one motion picture from one of the nine studios.

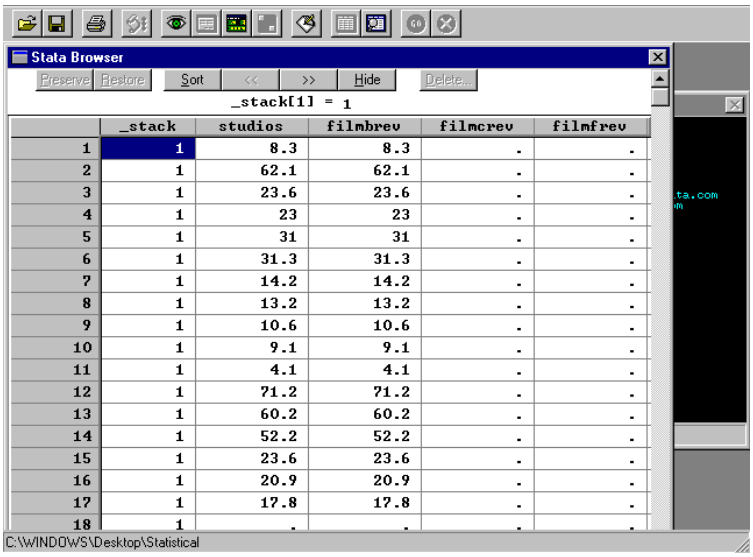
[NB: Be sure to open a log and record and save all your activities during this exercise]

2 Data management

1. First, **describe** the data. Paste the results into your exercise. How many observations are in the data set? How many variables?
2. Next, look at the data by clicking the "**browse**" button in Stata's toolbar:



The data spreadsheet will look like this:



	_stack	studios	filmbrev	filmcrev	filmfrev
1	1	8.3	8.3	.	.
2	1	62.1	62.1	.	.
3	1	23.6	23.6	.	.
4	1	23	23	.	.
5	1	31	31	.	.
6	1	31.3	31.3	.	.
7	1	14.2	14.2	.	.
8	1	13.2	13.2	.	.
9	1	10.6	10.6	.	.
10	1	9.1	9.1	.	.
11	1	4.1	4.1	.	.
12	1	71.2	71.2	.	.
13	1	60.2	60.2	.	.
14	1	52.2	52.2	.	.
15	1	23.6	23.6	.	.
16	1	20.9	20.9	.	.
17	1	17.8	17.8	.	.
18	1	.	.	.	.

3. Notice the dots under the column variable named **filmcrev.** They represent observations that have *missing values*. Scroll through the data. You'll see that there are many missing values in these data. They arose when I created the variable **studios** and they take up valuable space.

4. Since the exercise below will only make use of the **_stack** and **studios** variables, let's get rid of the other variables in the data set. This is very easy in Stata.

5. Close the data browser. Execute the command `drop film*`

6. Then **describe** the data and paste the results into your exercise. What's happened to the data set? What is the role of the "*" in the command?

7. Now, let's get rid of all the observations in the variable **studios** that contain missing values:

```
drop if(studios==.)
```

8. **describe** the data again, and explain what's happened to the dataset. Why do we use "==" instead of "=" in the command? Notice that the **drop** command can be used in more than one way. Look up **drop** in Hamilton, *Statistics with Stata* for a discussion of this command.

9. Open the data browser again and look over your data. Have the changes been correctly made? Are there any missing values left in the data? **describe** the data again and paste the

result into your exercise. How many observations? How many variables?

10. Finally, let's save our modified data set under a new name: **film_reduced.** The easiest way to do this is to go to the "file/Save As/" menu and pick a new name for the data. Be sure to put the data in the right directory on your hard drive. Name the revised data set "**film_reduced**" and click *Save.*

    The other way to save the data set under a new name is to type the Stata command (please substitute your own path; I used my own here):

    ```
    save "D:\amifiles\Econ400\data\film_reduced.dta"
    ```

11. You'll note that when you used the first option Stata actually placed this command in the "Review" window. In what comes next, we'll be using **film_reduced.dta.**

3 Describing the "studios" variable

Now, let's look more closely at the **studios** variable. First, let's determine how many observations come from each of the nine studios. We can do this two ways:

1. First, let's create a table and count the number of times each value of the _stack variable occurs using command **tabulate**:

   ```
   tabulate _stack
   ```

   Paste the resulting table into your exercise and answer the following questions:

   a) What proportion of the films were produced by studios 1-4?

   b) Which studio put out the most films during this period? How many?

   c) What proportion of all films did studios 7-9 produce?

2. Secondly, let's do a histogram on the _stack variable:

   ```
   histogram _stack,bin(9)
   ```

   This graph is pretty basic; let's pretty it up a bit by:

   a) putting an informative label on the "_stack" variable:

   ```
   label variable _stack "Studio Number"
   ```

   b) labelling the axes and adding a title and other improvements to the graph:

   ```
   histogram _stack, ylabel(0 2 to 18) xlabel(1 2 to 9) xtitle(" "
   "Studio Number" ) ti("Proportion of films produced by each
   studio") discrete percent addl
   ```

   paste this graph into your exercise.

3. Now, let's compare film revenues for all 9 studios by doing a comparative boxplot. We can use the "over(_stack)" option to produce a series of boxplots, one for each studio.

4. First, **summarize** the data to find the largest value for the studios variable:

   ```
   summarize studios
   ```

5. Then, submit a graph command (notice that I've prettied it up a bit):

   ```
   graph box studios, over(_stack) ylabel(0 25 to 225) ti("Film Revenue by Studio")
   ```

   Paste this graph into your exercise.

6. Now, do a boxplot for the entire set of films without differentiating by studio. Label the y-axis and provide an informative title for the graph. Paste this graph into your exercise.

7. Now, do a histogram of film revenues for the entire industry. Note, to do a histogram, you need to make some decisions in order to get the number of bars you want. From the boxplot we have a good idea of the distribution of the data. The data range from 0.46 million to 220.8 million dollars (what would you need to do to discover the minimum and maximum values? the median? the mean? the four smallest and largest values? *Do it and paste the results into your exercise).* Let's say we want each bar to represent a range of $25 million; how many bars would we need? (225/25 = 9).

   Now, applying informative labels to the x and y axes and adding a title, let's produce a histogram of the **studios** data:

   ```
   histogram  studios, bin(9) ylabel(0 10 to 60) xlabel(25 50 to 225) ti("Distribution of film Revenue") percent
   ```

   Finally, let's see how this distribution compares to a normal distribution:

   ```
   histogram  studios, bin(9) ylabel(0 10 to 60) xlabel(25 50 to 225) ti("Distribution of Film Revenue") percent norm
   ```

   Paste this last graph into your exercise.

8. Finally, let's do a more informative analysis of the shape of the distribution of the **studios** variable by doing a **quantile-normal plot.** As you remember, the quantile normal plot shows how well the distribution of a particular variable corresponds to a theoretical normal distribution with the same mean and standard deviation. If film revenues follow a normal distribution the data points would lie on a diagonal line. We already know from the last exercise that these data are not normally distributed. Let's confirm that by doing a quantile-normal plot (these plots are also known as **normal probability plots**):

   ```
   qnorm  studios, grid
   ```

   Paste this graph into your exercise and, right below it, explain the ways in which the **studios** data differ from a normal distribution.

4 The Central Limit Theorem

We know a lot about our data from the nine different film studios because of all the descriptive work we've just done. But, what if these nine studios constituted an *unknown population about which we were trying to learn more?* In particular, suppose we were only able to randomly sample from this unknown population of 304 film revenues. How confident could we be that the mean (or average) calculated from a small sample of these data would accurately reflect the true mean of the 304 observations? We can use Stata to do an experiment to answer this question.

We know that the **studios** data are highly skewed, but what about the sampling distribution of a sample mean computed from these data? We'll do the following experiment:

1. Randomly sample (with replacement) 5 values out of **studios** and compute the sample mean.
2. Store this sample mean and randomly sample again, computing and storing the sample mean. Do this 100 times so that we have a data set consisting of 100 sample means from **studios** computed from random samples of size 5.
3. **summarize** (with the *detail* option) the means data and paste into your exercise.
4. Do a **quantile-normal plot** and paste the resulting graph into your exercise.

You're going to do this exercise 4 times, changing the size of the sample used to compute the sample mean each time. At the end, you'll have learned something quite important about the sample mean computed from a population of an arbitrary distribution. Here are the commands necessary to complete the 4 steps above.

`set seed` **`yourPIDnumber`** Do this only the first time through this experiment. After that proceed directly to the next command.

START HERE for all subsequent passes through this part of the exercise.

```
bs "su  studios" "r(mean)", reps(100) size(5)
saving(d:\amifiles\econ400\exercises\mean5) replace
```

NB: *No* spaces between the items in parentheses and the text just in front of them. The two lines above should be typed as *one command* in Stata. The path in the above command, `d:amifiles\econ400\exercises`, is my path. You need to substitute your own path to the directory (which you've already created) where you want to store data set "mean5.dta". ***{Windows Users: Do Not create your folder under "My Documents." This can lead to problems when Stata tries to find your saved data. Instead, create a folder directly on your "C:" (or "D:") drive, say something like this:*** `C:\econ400\exercises\` ***}***

The command itself is a variant of the "bootstrap" command, a procedure in Stata that lets you take a large number [100 in this case] of random samples of a given size [`size(5)`]. After drawing a sample **bs** executes the command "su studios" which is shorthand for "summarize studios", and computes the mean [`r(mean)`], which is stored in data set "mean5.dta". **bs** does this 100 times.

```
clear

use "d:\amifiles\econ400\exercises\mean5"
```

Again, use your own path name in the above command to get to your data set "mean5".

```
summarize, detail
```

The above command (obviously) summarizes the means data from the 100 samples (paste it into your exercise).

```
qnorm _bs_1, grid
```

This command above computes the quantile normal plot (paste the plot into your exercise)

```
use "D:\amifiles\econ400\data\film_reduced.dta", clear
```

This last command reads your "film_reduced.data" data set back into memory so that you can repeat the experiment using different sample sizes.

Now, repeat this experiment for sample sizes of 15, 25, and 50. [change the "size( ) option in the **bs** command each time to reflect the desired sample size]. Rename your sample means data sets *mean15, mean25, mean50*. At the end you'll have 4 complete experiments in which you have computed distributions of sample means computed from samples of different sizes.

Now, in your exercise create a table that looks like the following and enter the appropriate numbers from your four experiments: To create a table in Word, go to the "Insert" tab and choose "Table." Then, specify a table with 6 rows and 5 columns. Once the table is in the text, select the entire first row and right-click "Merge Cells" in that row to get a single column for the title.

Results for Central Limit Theorem Experiments				
Exp. No.	Sample Size	Mean of Distribution	Std. Deviation of Distribution	How does distribution of means compare with normal distribution? (be specific)
Exp. 1				
Exp. 2				
Exp. 3				
Exp. 4				

Now, Answer the following questions about the experiments.

1. As sample size rises, what happens to the shape of the distribution of means calculated at each sample size?
2. How close does the distribution of sample means come to the normal distribution as the size of each sample rises?
3. Why is the result of your series of experiments so important to the analysis of sample means drawn from unknown populations?