EPIDEMIOLOGY 340.600: STATA PROGRAMMING AND DATA MANAGEMENT

Assignment 3

Due date: 11:59 PM, Monday, May 17, 2021 via CoursePlus dropbox

Due to constraints of final grade submission to the registrar, late assignments will not be accepted

Overview: write a .do file which performs the tasks described below. Your .do file should be called assignment3_yourname.do (e.g.: assignment3_allanmassie.do). Remember to write comments for full credit! Your .do file should follow conventions for .do file structure described in class (e.g.: header, logging, commenting, etc). Make sure your script will run on our machines, even if we are using a different version of Stata. Do not submit your log files as part of the assignment.

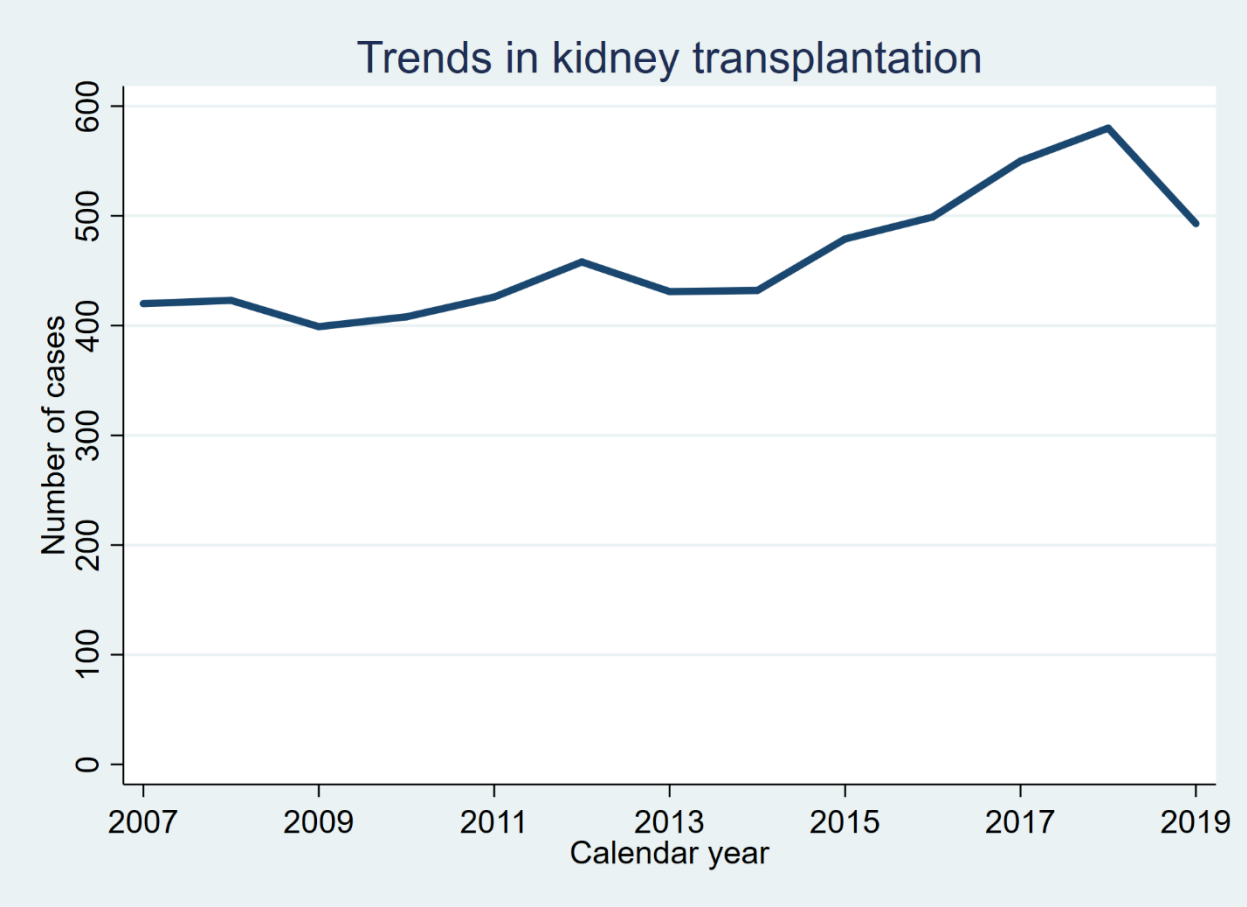
Codebook:

Variable	Description	Values
transplants.dta		
ctr_id	Center ID	Integer
transplant_date	Date of transplant	Date
prev_ki	Recipient has history of previous kidney transplant?	Binary (1=Yes/0=No)
age	Recipient's age	Integer
bmi	Recipient's BMI	Numeric

Question 1. Clear the memory and load `transplants.dta`. Drop all observations with missing values on `transplant_date`. Reproduce the following graph as precisely as possible.

This graph shows a line plot of the number of cases performed in each calendar year. Be mindful about the thickness of the line, the main title, and the axis labels and titles.

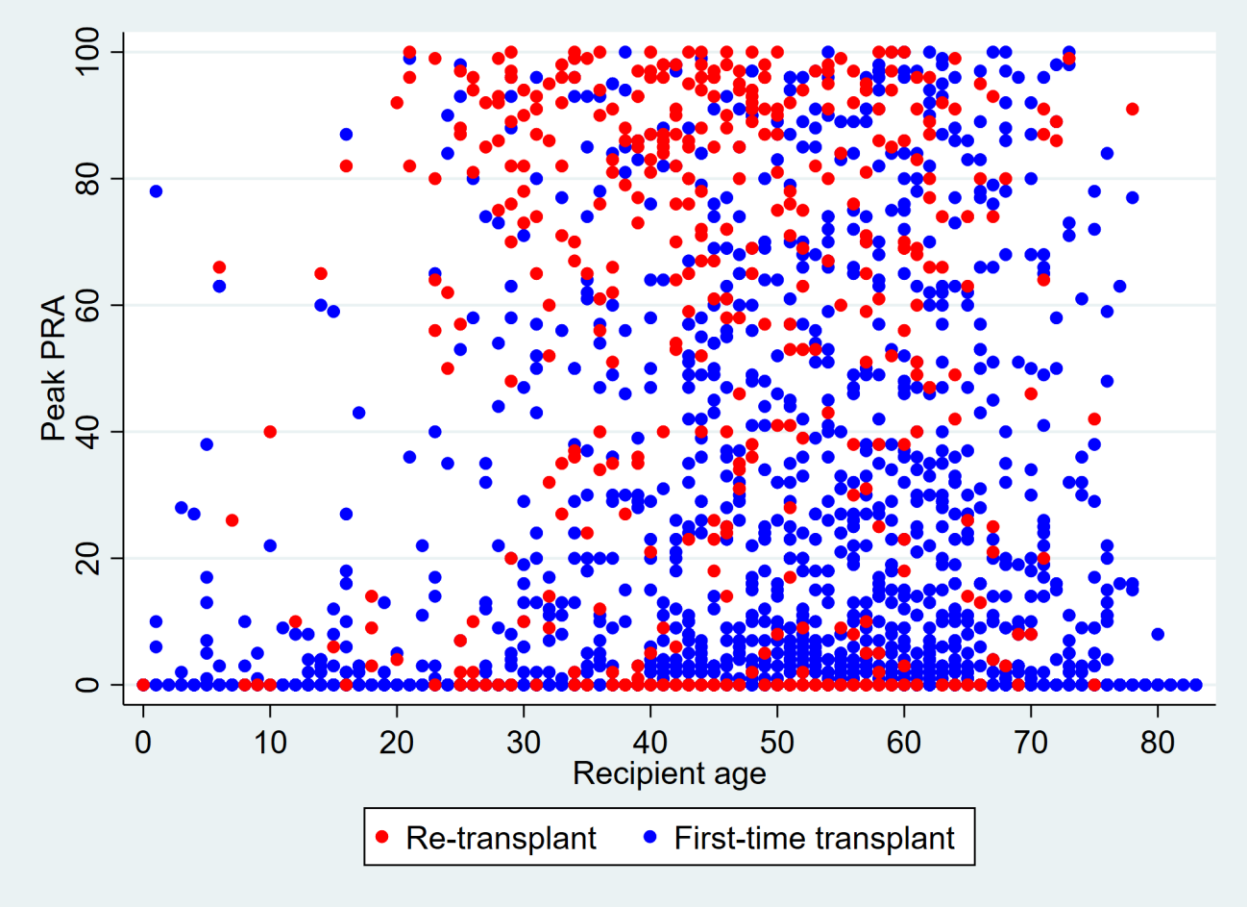
Save the graph as `q1_[yourname].png` (For example, `q1_allanmassie.png`)



Question 2. Clear the memory and load `transplants.dta`. Drop all observations with missing values on `peak_pra`. Reproduce the following graph as precisely as possible.

This graph shows a scatter plot of `peak_pra` and `age`, stratified by `prev_ki`. Observations with `prev_ki==1` are shown in red. Be mindful about the color ("blue" and "red" in Stata color codes) and size of the markers, the axis titles, and the legend.

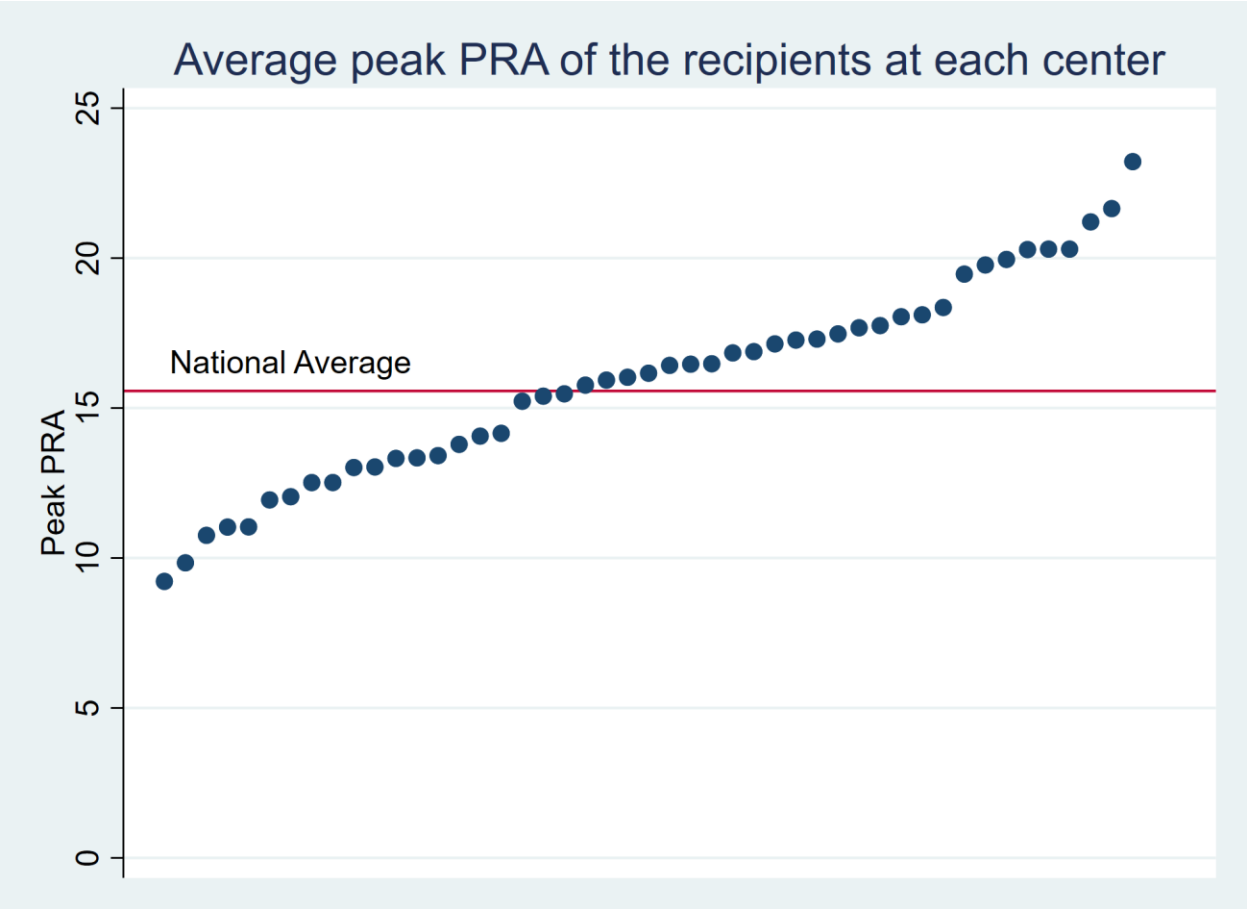
Save the graph as q2_[yourname].png (For example, q2_allanmassie.png)



Question 3. Clear the memory and load `transplants.dta`. Drop all observations with missing values for `peak_pra`. Reproduce the following graph as precisely as possible.

Calculate the average of `peak_pra` at each transplant center (`ctr_id`). Display the average value of `peak_pra` after sorting. The location of the text label "National Average" can be hardcoded (i.e. manually specify the location; about x=7 and y=17). Other markers and lines should NOT be hardcoded and should be derived by your script using the dataset.

Save the graph as `q3_[yourname].png` (For example, `q3_allanmassie.png`)



Question 4. Print the following text: `"Question 4: I estimate that it took me xxxx hours to complete this assignment."` For example, if it took you six hours, your .do file will contain the line

```
disp "Question 4: I estimate that it took me 6 hours to complete this assignment."
```

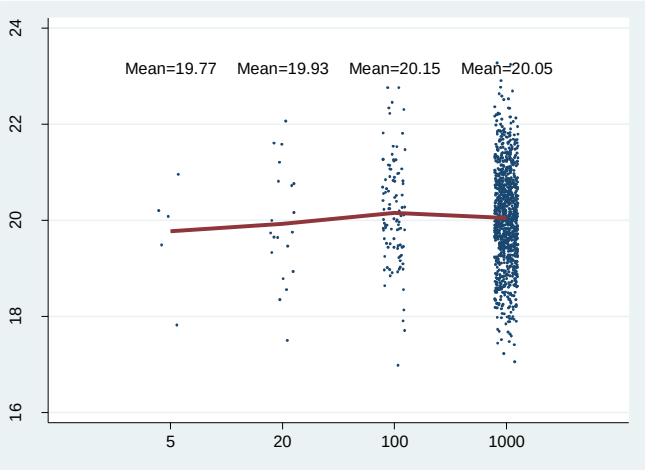
Give an honest answer; this is just for our data collection purposes. Do not include time spent on the Extra Credit Challenge (see below). Everyone who answers question 4 will receive full credit for this question. However, this question is worth some points, so don't skip it.

Extra Credit Challenge

Problem 5: write a program called `sampmean` to plot random data drawn from a normal distribution. `sampmean` takes a list of numbers representing different sample sizes. It also takes (optionally) a mean and standard deviation for the normal distribution. If we run

```
sampmean, at(5 20 100 1000) mean(20)
```

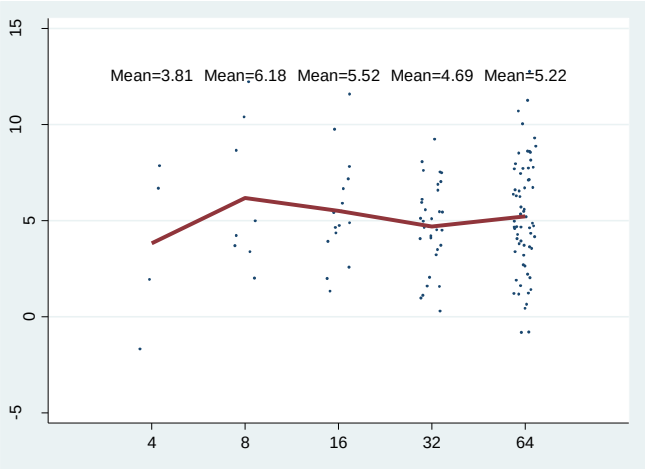
We get a graph like this:



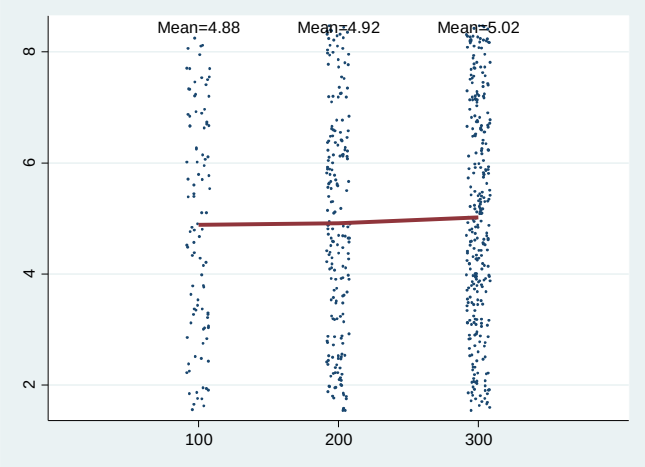
In this example, the program generates four sets of normally distributed random numbers (one set of 5 numbers, one set of 20, one set of 100, and one set of 1000) and calculates the mean for each set. It also plots each randomly generated number (as points) and the group mean (as a red line). The group mean also appears above each group as text.

Here are some more examples. The exact output will depend on the random number seed you use.

```
sampmean, at (4 8 16 32 64) mean(5) sd(3)
```




```
sampmean, at(100 200 300) mean(5) sd(2) uniform
```



In the last example, the distribution is a uniform distribution instead of a normal distribution.

Hints:

- You can use the keyword `numlist` (for a list of numbers) just as you do with a `varlist` (list of variables)
- The uniform distribution from 0 to 1 has mean 0.5 and standard deviation sqrt(1/12).

Note: this problem is pretty hard! We expect that few people will solve the whole thing, but we will give partial credit for a partial (working) solution. If your program only works partly, then explain in the comments, like this:

```
//NOTE: my program does not display the means in the graph.

//instead it prints them to the screen.
```

```
//Also "uniform" doesn't work,
//and my program runs for only one number (not a list)
```