**EPI 510: Stata assignment 3**

For this assignment, please submit two files: (1) **a file containing text, tables and figures** (.doc, .docx, .pdf or .txt) responding to questions posed in the assignment, and (2) **a Stata script** that performs the requested operations (.do).

For this assignment I'm going to have you build a dataset from component pieces and conduct a simple posterior simulation. I'll be asking you to write code similar to what I work with every day, so this is a very real-world exercise. I don’t expect you have conducted posterior simulations before (or even know what one is), so don’t worry if the method is new to you. If you are unfamiliar with the method, posterior simulation is a flexible method for propagating uncertainty. Whereas we most typically will define a distribution by its type and parameters (e.g. normally distributed with a mean of 40 and standard deviation of 5; or Poisson distribution with a mean of 6), with a posterior simulation, we pull random draws from the distribution and work with those draws. When we're done with all of the estimation work we can then get the point estimate as the mean of draws, and the bounds of a 95% uncertainty interval as the 2.5th and 97.5th percentile of the draws. This is more computationally expensive, but more flexible than fully parameterized approaches. Again, don't worry too much about whether or not you understand the methods here – hopefully the approach will be clear by the time you've finished the assignment, but the coding steps and the practice you'll get should work out either way. I have provided you with five sets of files:

1. incidence\_\*.dta are each location-specific datasets that contain 100 draws from the distributions of incidence estimates for a typhoid by year, age and sex;
2. case\_fatality.csv contains 100 draws from the distribution of case fatality estimates for typhoid fever by age;
3. age\_groups.csv contains details about the variable “age\_group\_id”;
4. locations.csv contains details about the variable “location\_id”; and
5. populations.csv contains age/sex-specific population estimates for every location and year.

There are a bunch of files. You can either download them individually, or download the file stataHW3.zip which contains all of the needed files. Before getting started, take a moment to look through the files and acquaint yourself with their structures.

1. For the case fatality, age groups, locations and populations files, import each file and store each in its own tempfile. You can do this with four separate import and save commands, or use a loop – your choice (the loop is quite a bit more challenging here, but I recommend that you consider trying it if you have the time and inclination, both for the practice and because it would be the "better" way to do it). **(8 points)**
2. There are two versions of this question: the standard version, and a bonus version. The bonus version is far more difficult, and you'll have to work to connect things that we've not explicitly connected in the class. I recommend that you finish the assignment and then come back to it if you’re so inclined. You can earn double credit if you choose the bonus version (note: bonus points are not available on resubmission, but you can still earn all of the standard points).  
     
   Standard Version: Append the six year-specific incidence datasets (.dta format) to create a complete incidence dataset. As above, you may do this with six separate commands or use a loop. As above, the loop is preferable. **(8 points)**Bonus Version: Append the six year-specific incidence datasets to create a complete incidence dataset, but use the .csv versions (NOT the .dta files). To get the bonus points you MUST use a loop. **(16 points)**
3. Merge the case fatality, age group, locations, and populations datasets (the tempfiles that you created in question 1) into the incidence dataset to create a master dataset that has all of the location-specific incidence draws, merged with age-specific case fatality draws, population estimates, and age group and location details. You should end up with a dataset that has 211 variables (year\_id, location\_id, location\_name, super\_region\_id, super\_region\_name, region\_id, region\_name, sex, age\_group\_id, age\_group\_name, population, cf\_0 – cf\_99, and inc\_0 – inc\_99) and 53,820 observations, where each observation represents a single location/year/age/sex. **(16 points)**
4. Now we're going to do draw-level calculations. Up to this point loops have been optional (both in terms of the assignment instructions, and in terms of practicality). That's about to change, and the importance of loops should become very obvious. Note, you can do parts a and b of this question in separate loops, or use a single loop for both parts – it's your choice, but the latter is easier, more efficient, and uses less code. So, read both parts before attempting the first.
   1. Generate 100 new variables (cases\_0 – cases\_99), each containing a draw of the number of cases. To do this, you'll loop over the values 0 – 99, and calculate each case draw as the product of a draw of incidence and population (i.e. cases\_0 = inc\_0 \* population; cases\_1 = inc\_1 \* population; …; cases\_99 = inc\_99 \* population). **(10 points)**
   2. Now, we're going to estimate deaths as the product of cases and case fatality, and we're going to do this as above with 100 draws (i.e. deaths\_0 = cases\_0 \* cf\_0; deaths\_1 = cases\_1 \* cf\_1; …; deaths\_99 = cases\_99 \* cf\_99). By doing these calculations at the draw level, the draws of deaths will include uncertainty from both the estimates of incidence and the estimates of case fatality (which is what we want!) **(6 points)**
   3. Save the current version of the dataset to a new file (don’t overwrite any original data!). I'll call my file hw3Master.dta. **(2 points)**