**Lab 3: Confidence intervals and Sampling**

Geog 4300/6300-Shannon

Fall 2016

Value: 20 points

**Overview**: This lab covers two main topics: basic spatial statistics and probability distributions. We will be using individual level “microdata” from the Current Population Survey (CPS). It is designed as an ongoing (collected monthly) set of data on financial and demographic characteristics. One main use of the CPS is to calculate national levels of food insecurity. Each December, a food security supplement is added to the regular survey, and data from the supplement is included here.

The data for this lab are in a zipped file on the ELC site named “IPUMS\_CPS\_FoodSec.” This contains a csv file with microdata from the CPS that is de-identified and publically available through the Minnesota Population Center (<https://cps.ipums.org/cps/index.shtml>). There is also a Word codebook describing each of those variables.

Your answers to the lab questions should be typed in a separate Word document and turned in using the Assignment Dropbox on the ELC site.

1. Look at the information on the YRIMMIG variable in the codebook, which describes when foreign born respondents arrived in the U.S. Much statistical data comes with a “missing variable” identifier, which in this case you want to leave out of your analysis. It’s labeled here as “NIU”. Create a script that subsets this data so that only records with meaningful responses are included. Your script should create a histogram of the resulting subset to show the distribution. Also calculate the mean, median, standard deviation, and IQR (see lab 1) of this variable.

***Question 1 (3 points):*** *Copy and paste your script, histogram, and the four statistics into your lab response. Describe the distribution using the results and note possible explanations for any trends you notice.*

1. The FSSTATUS variable describes the food security of respondents. While food security status is often grouped into “low” and “very low” food security, these two are often just combined to a single measure: food insecure. Using the information on this variable in the codebook (p. 12), create a subset of records without missing data. Then use the table function to create counts of respondents in each category. Use those counts to calculate what percentage of respondents are food insecure.

***Question 2 (4 points):*** *List the estimated food insecurity rate based on this survey data, and create a confidence interval for the true food insecurity rate with 95% confidence. Copy and paste the R script you used to do so in your response.*

1. Now select just those records with food security data from Georgia. Compute the same confidence interval as you did above for just these records. Then do the same for Colorado.  
     
   ***Question 3 (4 points):*** *List this food insecurity rate along with the confidence interval (95% confidence) for each state, including the scripts you use to determine them. Can we say that the real means in these three cases (national, Colorado, and Georgia) are probably different from one another? How do you know?****Question 4 (2 points):*** *Is the size of confidence interval different between the state and national data? If so, why?*
2. A related study is being developed to determine whether new food shelves in the Atlanta metropolitan area are connected to reduced rates of food insecurity. The research question is whether living within a mile of a food pantry lowers food insecurity for households.

***Question 5 (4 points):*** *Pick a sampling strategy (or combination of strategies) discussed in our text or in lecture that would be appropriate for this research question: convenience, snowball, quota, random, systemic, stratified, and cluster. Describe how this strategy would be used to create a sample from this sampling frame. What would be the strengths and weaknesses of this approach? Your response should be around 100-150 words in length.*

***Question 6 (3 points):*** *City health officials would like to do a related survey of household food insecurity with enough responses to allow for margins of error under 2% (with 95% confidence). How big a sample would they need? Explain how you arrived at your answer.*