Geog 4/6300: Lab 3

Confidence Intervals and Sampling

Due: Monday, Oct. 16

Value: 30 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.

To load these data, load the csv file:

```
library(tidyverse)
cps_data<-read_csv("https://github.com/jshannon75/geog4300/raw/master/Labs/Lab%203-Confidence%20interva
```

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 codebook available on Github describing each of those variables.

Part 1: Calculating national food insecurity

The n() function can be used to count records in a group. See this example for immigration:

```
cps_data %>%
  group_by(YRIMMIG) %>%
  summarise(count=n())
```

```
## # A tibble: 23 x 2
##    YRIMMIG  count
##      <int>  <int>
##  1       0 466161
##  2    1949    743
##  3    1959   2071
##  4    1964   1815
##  5    1969   2147
##  6    1974   2981
##  7    1979   3682
##  8    1981   2634
##  9    1983   1660
## 10    1985   2264
## # ... with 13 more rows
```

For this response, look at the FSSTATUS variable, which 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.

Question 1 (4 points) *Using the information on this variable in the codebook (p. 12), create a subset of records without the NIU or missing response records. Then use group_by and summarise to calculate the number of individuals grouped in each status as shown above. Use the resulting data frame to calculate an estimate of the national food insecurity rate.*

Question 2 (3 points) *Using the formula for confidence intervals for proportions shown in class, calculate a confidence interval for the rate you identified in question 3. Interpret what that confidence interval tells you.*

Part 2: Analyzing state food insecurity data

We can also use the "STATE" variable to calculate rates for each state. To do this, we'll need to use the *spread* function from tidyverse. Spread converts "long" data to "wide." Here's an example using the immigration data. First, let's filtering all records with a 0 and use ifelse to create a dummy variable for all individuals immigrating in the year 2000 or after.

```
cps_data_2000<-cps_data %>%
  filter(YRIMMIG!=0) %>%
  mutate(yr_2000=ifelse(YRIMMIG>=2000,1,0))

cps_data_2000
```

```
## # A tibble: 64,872 x 28
##     YEAR    CPSID STATEFIP STATECENSUS       STATE METAREA FSSTATUS
##    <int>    <dbl>    <int>       <int>       <chr>   <int>    <int>
##  1  2010 2.01e+13        1          63     Alabama    3440        1
##  2  2010 2.01e+13        1          63     Alabama    3440        1
##  3  2010 2.01e+13        1          63     Alabama    3440        3
##  4  2010 2.01e+13        1          63     Alabama    3440        3
##  5  2010 2.01e+13        1          63     Alabama    3440        3
##  6  2010 2.01e+13        9          16 Connecticut    1161        2
##  7  2010 2.01e+13        9          16 Connecticut    1161        2
##  8  2010 2.01e+13        9          16 Connecticut    1161        2
##  9  2010 2.01e+13        9          16 Connecticut    1161        2
## 10  2010 2.01e+13        9          16 Connecticut    1161        2
## # ... with 64,862 more rows, and 21 more variables: FSSTATUSA <int>,
## #   FSSTATUSC <int>, FSFDSTMP <int>, FSWIC <int>, FSFDBNK <int>,
## #   FSSOUPK <int>, FSPOOR <int>, AGE <int>, SEX <int>, RACE <int>,
## #   MARST <int>, BPL <int>, YRIMMIG <int>, CITIZEN <int>, HISPAN <int>,
## #   EDUC <int>, EMPSTAT <int>, IND <int>, EARNWEEK <dbl>, DIFFANY <int>,
## #   yr_2000 <dbl>
```

Then we can use group_by, summarise, and spread to count the number of immigrants before and after 2000.

```
cps_data_2000_wide<-cps_data_2000 %>%
  group_by(STATE,yr_2000) %>%
  summarise(count=n()) %>%
  spread(yr_2000,count)

cps_data_2000_wide
```

```
## # A tibble: 51 x 3
## # Groups:   STATE [51]
##                   STATE   `0`   `1`
##  *                <chr> <int> <int>
##  1              Alabama    95   108
##  2               Alaska   375   302
##  3              Arizona   672   298
##  4             Arkansas   180   128
##  5           California  8917  3755
##  6             Colorado   745   487
## 7          Connecticut  1210   625
## 8             Delaware   441   358
## 9 District of Columbia   637   405
## 10              Florida  2856  1750
## # ... with 41 more rows
```

To convert this to a rate, we can use mutate with the above function. Here's the percentage of the immigrant population arriving since 2000. Note that in this case, since your variables will be numbers, R requires tags around them: 1, not just 1.

```
cps_data_2000_wide<-cps_data_2000 %>%
  group_by(STATE,yr_2000) %>%
  summarise(count=n()) %>%
  spread(yr_2000,count) %>%
  mutate(total=`0` + `1`,
         rt2000=`1`/total*100)

cps_data_2000_wide
```

```
## # A tibble: 51 x 5
## # Groups:   STATE [51]
##                   STATE   `0`   `1` total   rt2000
##                   <chr> <int> <int> <int>    <dbl>
##  1              Alabama    95   108   203 53.20197
##  2               Alaska   375   302   677 44.60857
##  3              Arizona   672   298   970 30.72165
##  4             Arkansas   180   128   308 41.55844
##  5           California  8917  3755 12672 29.63226
##  6             Colorado   745   487  1232 39.52922
##  7          Connecticut  1210   625  1835 34.05995
##  8             Delaware   441   358   799 44.80601
##  9 District of Columbia   637   405  1042 38.86756
## 10              Florida  2856  1750  4606 37.99392
## # ... with 41 more rows
```

Question 3 (5 points) *Adapting the above function, create an estimated food insecurity rate for each state from these data. To do so, you'll need to create counts for each response (food secure, low food security, very low food insecurity), sum the latter two, and divide by the total responses within each state.*

Question 4 (3 points) *Create a command that estimates the margin of error for each state based on the national rate you calculated in question 1 and the total responses for each state.*

Question 5 (3 points) *Compare the margin of error you calculated for Georgia to the national margin of error. How do they differ? Mathematically, why are they different?*

Question 6 (5 points) *Create a column in your state food insecurity estimates that converts each state's food insecurity rate to a z score based on the whole population. What are the z scores for Wisconsin, Washington, and Mississippi? What do those z scores tell you?*

Part 3: Sampling

A new study is being developed to determine whether new food shelves in the Atlanta metropolitan area are reducing rates of food insecurity. The research question is whether living within a mile of a food pantry lowers food insecurity for households.

Question 7 (4 points) *Pick a probabilistic sampling strategy (or combination of strategies) discussed in our text or in lecture that would be appropriate for this research question: random, systemic, stratified, and* *cluster. Describe how this strategy could be used to create a sample for use in this proposed study. Describe one strength and weakness of this approach.*

Question 8 (3 points) *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). Assume that the rate is similar to the one you identified for Georgia in question 5. Use R to compute how big a sample they would need.*