

2007 Income Data by County

In the United States, there are over 3,000 counties. The data set that will be used in this lab provides a listing of all counties and several variables that were measured in each county in 2007. Our variable of interest for this lab is the Adjusted Gross Income.

In this lab we will be using Rstudio to find a random sample, summary statistics, and confidence intervals. To begin you will need to download the data set: `Tax_Year_2007_Income.csv`. Data sets can be found in Blackboard in the Data folder. You can rename the data set if you wish from the Import Dataset screen. In the following, the data set is called “x”. The Rstudio Script that contains the commands is the file `CIRstudioLab.R`.

1. Upload the data set.
2. What is the variable? _____
3. What is the population? _____
4. Set the seed in RStudio. This allows you to replicate your experiment if necessary. Do this by `set.seed(#)` where # is your favorite number. What seed did you set?

5. Obtain a random sample of size $n=100$; `samp<-sample(x[[7]], 100)`. The notation `[[7]]` indicates to RStudio that you want to take a sample from the 7th column of the data set. OR `sample(x$AdjustedGrossIncome, 100)`
 - a. Obtain the summary statistics for your sample.
 - a. Mean; `mean(samp)`

 - b. Median; `median(samp)`

 - c. Standard Deviation; `sd(samp)`

4. Create a 95% Confidence Interval for your sample.

```
> xbar <- mean(samp)  #(This is your point estimate for the mean.)
> s <- sd(samp)        #(This is your point estimate for the standard deviation.)
> n <- length(samp)    #(This is the length of your sample.)
> error <- qt(0.975, df = n - 1)*s/sqrt(n)  # (This calculates the error term. That is,  $t_{\alpha/2} \frac{s}{\sqrt{n}}$ .)
> lower <- xbar - error      # (This is the lower bound of your CI.)
> upper <- xbar + error     #(This is the upper bound of your CI.)

> lower

> upper
```

5. Record and interpret your answer.