Lab 3D - Are you sure about that?

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

## Confidence and intervals

* Throughout the year, we've seen that:
  + Means are used for describing the typical value in a sample or population, but we usually don't know what they are, because we can't see the entire population.
  + Means of samples can be used to *estimate* means of populations.
  + By including a margin of error with our estimate, we create an interval that increases our confidence that we've located the correct value of the population mean.
* Today, we'll learn how we can calculate *margins of error* by using a method called the *bootstrap*.
  + Which comes from the phrase, *Picking yourself up by your own bootstraps*.

## In this lab

* Load the built-in atus (*American Time Use Survey*) data set, which is a survey of how a sample of Americans spent their day.
  + **The United States has an estimated population of 327,350,075. How many people were surveyed for this particular data set?**
* The statistical question we wish to investigate is:

*What is the mean age of people older than 15 living in the United States?*

* **Why is it important that the ATUS is a random sample?**
* **Use our atus data to calculate an estimate for the average age of people older than 15 living in the U.S.**

## One bootstrap

* A *bootstrapped* sample is when we take a random sample() of our original data (atus) *WITH* replacement.
  + The size of the sample should be the same size as the original data.
* We can create a single *bootstrapped* sample for the mean in 3 steps:
  1. Sample the number of the rows to use in our *bootstrap*.
  2. slice those rows from our original data into our *bootstrap* data.
  3. Calculate the mean our our *bootstrapped* data.

## Our first bootstrap

* Fill in the blanks to sample the row numbers we'll use in our *bootstrapped* sample.
  + Be sure to re-read what a *bootstrapped* sample is from the previous slide to help you fill in the blanks.
  + Use set.seed(123) before taking the sample.

bs\_rows <- \_\_\_\_(1:\_\_\_\_, size = \_\_\_\_, replace = \_\_\_\_)

* We can use the slice function to create a new data set that includes each row from our sample

bs\_atus <- slice(atus, bs\_rows)

## Take a look

* Look at the values of bs\_rows and bs\_atus.
  + **Write a paragraph that explains to someone that's not familiar with R how you created bs\_rows and bs\_atus. Be sure to include an explanation of what the *values* of bs\_rows mean and how those values are used to create bs\_atus. Also, be sure to explain what each argument of each function does.**

## One strap, two strap

* Calculate the mean of the age variable in your bootstrapped data, then use a different value of set.seed() to create your own, personal *bootstrapped* sample. Then calculate its mean.
  + Compare this second *bootstrapped* sample with three other classmates and write a sentence about how similar or different the *bootstrapped* sample means were.

## Many bootstraps

* To use *bootstrapped* samples to create *confidence intervals*, we need to create many *bootstrapped* samples.
  + Normally, the more *bootstrapped* samples we use, the better the *confidence interval*.
  + In this lab, we'll do() 500 *bootstrapped* samples.
* To make do()-ing 500 *bootstraps* easier, we'll code our 3-step bootstrap method into a function.
  + Open a new R script (File -> New File -> R Script) to write your function into.

## Bootstrap function

* Fill in the blank space below with the 3-steps needed to create a *bootstrapped* sample mean for our atus data.
  + Each step should be written on its own line between the curly braces.

bs\_func <- function() {  
   
   
   
}

* Highlight and *Run* the code you write.

## Visualizing our bootstraps

* Once your function is created, fill in the blanks to create 500 *bootstrapped* sample means:

bs\_means <- do(\_\_\_\_) \* bs\_func()

* **Create a histogram for your bootstrapped samples and describe the *center*, *shape* and *spread* of its distribution.**
  + These bootstrapped estimates no longer estimate the average age of people in the U.S.
  + Instead, they estimate how much the estimate of the average age of people in the U.S. varies.
* In the next slide, we'll look at how we can use these bootstrapped means to create *90% confidence intervals*.

## Bootstrapped confidence intervals

* To create a 90% confidence interval, we need to decide between which two *ages* the middle 90% of our bootstrapped estimates are contained.
* **Using your histogram, fill in the statement below:**

The lowest 5% of our estimates are below \_\_\_\_\_\_\_ years and the highest 5% of our estimates are above\_\_\_\_\_\_\_ years.

* Use the quantile() function to check your estimates.
* **Based on your bootstrapped estimates, between which two ages are we 90% confident the actual mean age of people living in the U.S. is contained?**

## On your own

* Using your *bootstrapped* sample means, create a 95% confidence interval for the mean age of people living in the U.S.
  + **Why is the 95% confidence interval wider than the 90% interval?**
  + **Write down how you would explain what a 95% confidence interval means to someone not taking *Introduction to Data Science*.**