Lab 3C - Random Sampling

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

## Learning by sampling

* In many circumstances, there's simply no feasible way to gather data about everyone in a *population*.
  + For example, the Department of Water & Power (DWP) wants to determine how much water people in Los Angeles use to take a shower. They've created a survey to pass out to collect this information.
  + **Write down two reasons why getting *everyone* in Los Angeles to fill out the survey would be difficult. Also, write a sentence why the DWP might consider using a sample of households instead.**
* In this lab, we'll learn how *sampling methods* affect how *representative* a sample is of a *population*.

## Loading a population

* In previous labs, we used the cdc data as a sample for young people in the United States.
  + In this lab, we'll consider these survey respondents to be our population.
* Load the cdc data into R and fill in the blanks to take a *convenience* sample of the first 50 people in the data:

s1 <- slice(\_\_\_\_, 1:\_\_\_\_)

* **Why do you think we call this method a *convenience* sample?**

## Comparing your convenience sample

* A convenience sample is a sample from a population where we collect data on subjects because they're easy-to-find.
* Using your convenience sample, create a bargraph for the number of people in each grade.
  + **Do you think the distribution of grade for your sample would look similar when compared to the whole cdc data?**
  + **Which groups of people do you think are over or under represented in your convenience sample? Why?**
* Create a bargraph for grade using the cdc data.
  + **Compare the distributions of the cdc data and your convenience sample and write down how they differ.**

## Using randomness

* Fill in the blanks below to create a sample by randomly selecting 50 people in the cdc data, without replacement. Call this new sample s2:

\_\_\_ <- sample(\_\_\_, size = \_\_\_, replace = \_\_\_)

* **Write a sentence that explains why you think the distribution of grade for this *random sample* will look more or less similar to the distribution from the whole cdc data.**
  + Create a bargraph for grade based on this *random sample* to check your prediction.

## Increasing sample size

* Create bargraphs for grade based on each of the following sample sizes: 10, 100, 1,000, 10,000.
  + Compare each distribution to that of the population.
* **How do the distributions change as the size of the sample increases? Why do you think this occurs?**
* tally() the proportion of grades for your *convenience* sample and all your *random* samples.
  + **Which set of proportions looks most similar to the proportions of the population?**

## Lessons learned

* The mean, or proportion, from a *random* sample might not always be closer to that of the true population when compared to a *convenience* sample.
* However, as sample sizes get larger:
  + *Random* samples will tend to be better estimates for the population.
  + With *convenience* samples, this might not be the case.
* **Write down a reason why estimates based on *convenience* samples might not improve even as sample size increases.**