Lab 2D - Queue it up!

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

## Where we left off

* In the last lab, we looked at how we can use computer simulations to compute estimates of simple probabilities.
* Like the probability of drawing a song genre from a playlist.
* We also saw that performing *more* simulations:
  + Took *longer* to finish.
  + Had estimates that *varied less*.
* In this lab, we'll extend our simulation methods to cover situations that are more complex.
  + We'll learn how to estimate their probabilities.
  + We also look at the roll of sampling *with* or *without* *replacement*.

## Back to songs

* In R, simulate a *playlist of songs* containing 30 "rap" songs, 23 "country" songs and 47 "rock" songs.
  + *Assign* the combined playlist the name songs.
* Simulate choosing a single song 50 times. Then use your simulated draws to estimate the probability of choosing a *rap* song.
  + The actual (theoretical) probability of choosing a *rap song* in this case is 0.30.
  + **Write a sentence comparing your estimated probability to the actual probability.**

## With or Without?

* So far, you've selected songs *with replacement*.
  + We called it that, because each time you made a selection, you started with the same playlist. That is, you chose a song, wrote down its data, and then placed it back on the list.
* It's also possible to select *without replacement* by setting the replace option in the sample function to FALSE.
* Take a sample of size 100 from our playlist of songs *without replacement*. Assign this sample the name without.
  + **What do you notice if you run tally(~without)? Does something similar happen if you sample *with replacement*?**
  + **What happens if size = 101 and replace = FALSE?**

## Sample with? Or without?

* Imagine the following two scenarios.
  1. You have a coin with two sides: *Heads* and *Tails*. You're not sure if the coin is fair and so you want to estimate the probability of getting a *Head*.
  2. A child reaches into a candy jar with 10 *strawberry*, 50 *chocolate* and 25 *watermelon* candies. The child is able to grab three candies with their hand and you're interested in probability that all three candies will be chocolate.
* **Which of these scenarios would you sample *with replacement* and which would you sample *without replacement*? Why?**
  + **Write down the line of code you would run to sample from the candy jar. Assume the simulated jar is named candies.**

## Simulations at work

* In reality, songs from a playlist are chosen without replacement.
  + This way, you won't hear the same song several times in a row.
* Let's write a more realistic simulation and estimate the probability that if we select two songs at random, without replacement, that both are rap songs.
  + Use the do function to perform 10 simulated samples of size 2, with replacement and *assign* the simulations the name draws.

## Simulations and probability

* To estimate the probability from our simulations, we need to find the proportion of times that the event we're interested in occurs in the simulations.
* In other words, we need to count the number of times the desired events occurred, divided by the number of attempts we made (the number of simulations).
* The next slides will show you two ways to do this.

## Counting similar outcomes

* One way we can estimate the probability of drawing two songs of the *same* genre is to use the following trick to count the number of *rap* songs in each of the 10 simulations:

mutate(draws, nrap = rowSums(draws=="rap"))

* **For each of the lines of code below, describe how the output of the code changes as we move from line to line.**

draws == "rap"

rowSums(draws == "rap")

mutate(draws, nrap = rowSums(draws=="rap"))

## Counting other outcomes

* Another method we can use to estimate the probability of complex events is to use the following 2-step procedure:
  1. Subset the rows of the simulations that match our desired outcomes.
  2. Count the number of rows in the subset and divide by the number of simulations.
* The result that you obtain is an estimate of the probability that a specific combination of events occured.
* We'll see an example of this method on the next slide.

## Step 1: Creating a subset

* Fill in the blanks below to:
  1. Create a subset of our simulations when both draws were "rap" songs.
  2. Count the number of rows in this subset
  3. And divide by the total number of repeated simulations.

draws\_sub <- filter(draws, \_\_\_ == "rap", \_\_\_ == "rap")

nrow(\_\_\_) / \_\_\_

## Estimating probabilities

* **Calculate estimated probabilities for the following situations:**
  1. You draw two "rap" songs.
  2. You draw a "rap" song in the first draw and a "country" song in the 2nd.
* **Create a histogram that displays the number of times a "rap" song occurred in each simulation. That is, how often were zero rap songs drawn? A single rap song? Two rap songs?**

## On your own

* Using what you've learned in the previous two labs, answer the following question by performing two computer simulations with 500 repetitions a piece:

***If we draw 5 songs from a playlist of 30 rap, 23 country and 47 rock songs, how does the estimated probability of all 5 songs being rap songs change if we draw the songs with or without replacement?***

* For each simulation:
  + **Create a histogram for the number of *rap* songs that occurred for each of the 500 repetitions.**
* **Describe how the distribution of the number of *rap* songs changes depending on if we use replacement or not.**