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

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

## A new direction

* For the past two labs, we've looked at ways that we can summarize data with numbers.
  + Specifically, you learned how to describe the *center*, *shape* and *spread* of variables in our data.
* In this lab, we're going to *estimate the probability* that a rap song will be chosen from a playlist with both rap and rock songs, if the choice is made at random.
  + The playlist we'll work with has 100 songs: 39 are rap and 61 are rock.

## Estimate what ... ?

* To *estimate the probability*, we're going to imagine that we select a song at random, write down its genre (*rock* or *rap*), put the song back in the playlist, and repeat 499 more times for a total of 500 times.
* The statistical question we want to address is: *On average, what proportion of our selections will be rap?*
* **Why do we *put a song back* each time we make a selection?**
* **What would happen in our little experiment if we did not do this?**

## Calculating probabilities

* Remember that a *probability* is the long-run proportion of time an event occurs.
  + Many probabilities can be answered exactly with just a little math.
  + The probability we draw a single rap song from our playlist of 39 rap and 61 rock songs is 39/100, or 39%.
* Probabilities can also be answered exactly if we were willing to randomly select a song from the playlist, write down its *genre*, place the song back in the list, and repeatedly do this *forever*.
  + Literally, *forever* ...
  + But we don't have that much time. So we're only going to do it 500 times which will give us an *estimate of the probability*.

## Estimating probabilities

* You might ask, *Why are we estimating the probability if we know the answer is 39%?*
  + Sometimes, probabilities are too hard to calculate with simple division as we did above. In which case, we can often program a computer to run an experiment to estimate the probability.
  + We refer to these programs as *simulations*.
* The techniques you learn in this lab could be applied to very simple probability calculations or very hard and complex calculations.
  + In both cases, your *estimated* probability would be very close to the *actual* probability.

## Getting ready

* Simulations are meant to mimic what happens in real-life using randomness and computers.
  + Before we can start simulating picking songs from a playlist, we need to simulate that playlist in R.
* To simulate our 39 rap songs, we'll use the repeat (rep) function.

rap <- rep("rap", times = 39)

* Use a similar line of code to simulate the rock songs in our playlist of 100.

## Put the songs in the playlist

* Now that we've got some different songs, we need to combine them together.
  + To do this, we can use the combine function in R, c().
* Fill in the blanks to combine your different songs:

songs <- \_\_(rap, \_\_\_\_)

* And with that, our playlist of songs should be ready to go.
  + Type songs into the console and hit enter to see your individual *songs*.

## Pick a song, any song

* Data scientists call the act of choosing things randomly from a set, *sampling*.
  + We can randomly choose a song from our playlist by using:

sample(songs, size = 1, replace = TRUE)

* Run this code 10 times and compute the proportion of "rap" songs you drew from the 10.
  + **Once everyone in your class has computed their proportions, calculate the *range* of proportions (The largest proportion minus the smallest proportion) for your class and write it down.**

## Now do() it some more

* Instead of running the same line of code multiple times ourselves we can use R to do() multiple repetitions for us.
  + Fill in the blanks below to do the sample code from the previous slide *50* times run:

do(\_\_\_) \* sample(\_\_\_, \_\_\_ = \_\_\_, \_\_\_ = \_\_\_)

* *Assign* the 50 selected songs the name draws. Then fill in the blank below to tally how often each genre was selected:

tally(~\_\_\_, data = draws)

* **Compute the proportion of "rap" songs for your 50 draws and find out if the *range* for your class' proportions is bigger or smaller than when we drew 10 songs.**

## Proportions vs. Probability

* To review, so far in this lab we've:
  + Simulated a "playlist" of songs.
  + Repeatedly simulated drawing a song from the playlist, noting its genre and placing it back in the playlist.
  + Computed the proportion of the draws that were "rap".
* These proportions are all *estimates* of the theoretical probability of choosing a rap song from a playlist.
  + As we increase the number of draws, the *range* of proportions should shrink.

*When using simulations to estimate probabilities, using a large number of repeats is better because the estimates have less variability and so we can be confident we're closer to the actual value.*

## Non-random Randomness

* We've seen that random simulations can produce many different outcomes.
  + Some estimated probabilities in your class were smaller/larger relative to others.
* There are instances where you might like the same random events to occur for everyone.
  + We can do this by using set.seed().
* For example, the output of this code will always be the same:

set.seed(123)  
sample(songs, size = 1, replace = TRUE)

## [1] "rap"

## Playing with seeds

* With a partner, choose a number to include in set.seed then redo the simulation of 50 songs.
  + Both partners should run set.seed(\_\_\_) just before simulating the 50 draws.
  + The blank in set.seed(\_\_\_) should be the same number for both partners.
  + Verify that both partners compute the same proportion of "rap" songs.
* Redo the 50 simulations one last time but have each partner choose a different number for set.seed(\_\_\_).
  + **Are the proportions still the same? If so, can you find two different values for set.seed that give different answers?**

## On your own

* Suppose there are 1,200 students at your school. 400 of them went to the movies last Friday, 600 went to the park and the rest read at home.

*If we select a student at random, what is the probability that this student is one of the one's who went to the movies last Friday?*

* **Answer this by estimating the probability that a randomly chosen student went to the movies using 500 simulations.**
  + **Write down both the estimated probability and the code you used to compute your estimate. You might find it helpful to write your answer in an R Script *(File -> New File -> R Script)***
  + **Include set.seed(123) in your code before you do 500 repeated samples.**