Lab 2E - The Horror Movie Shuffle

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

## Playing with permutations

* *Slasher* films are notoriously gory and are said to contain recurring biases.
  + One such bias, is that women in slasher films are more likely to survive than men.
* This lab will focus on the statistical question: *Are women in slasher films more likely to survive until the end of the film than men?*
* To answer this question, we'll learn how to use permuted data to gauge how likely an event occurs by chance.
* To begin, use the data function to load the slasher data file.
  + The data contains information about 485 characters from a random sample of 50 *slasher* horror films.

## Initial thoughts...

* To familiarize yourself with the data, answer the following:
  + **How many variables and observations are contained in the data and what are the possible values of the variables?**
  + **Which gender had more survivors? Write down a few sentences as to how you came to your conclusion. Be sure to look at both the *counts* and *proportion* of survivors before deciding.**
  + **Calculate the difference between the proportion of females who survived and the proportion of males who survived. Is the difference large enough to conclude that women tend to survive more often than men?**

## Tally whoa ... !

* Something you might have noticed is that these two lines of code aren't equivalent:

tally(gender ~ survival, data = slasher)

tally(survival ~ gender, data = slasher)

* One of these lines takes the group of *survivors* and tells us how many of them were Male or Female.
* The other takes the group of *females* and tells us how many of them Dies or Survives.
* **The last question on the previous slide can be answered using the 2nd line of code. Why?**

## Examining differences

* When we're comparing the difference between two quantities, such as survival rates of slasher films, it can be difficult to decide how *different* two values need to be before we can conclude that the difference didn't just happen by chance.
  + To help us decide when a difference is not due to chance, we'll use repeated random shuffling.
* By using repeated random shuffling, we'll estimate how often our *actual* difference occurs by *chance*.

## Do the shuffle!

* When we shuffle data, we use our original data set as a starting point.
  + Run the following and write down the resulting table on a piece of paper.

tally(survival ~ gender, data = slasher)

* Now run the following to randomly reassign each survival status to each observation. Compare the resulting table to the one you wrote down.

tally(shuffle(survival) ~ gender,   
 data = slasher)

## Let's compare ...

* **How many people survived, in total, the slasher film before shuffling? How many people survived after shuffling?**
* **How has shuffling our data changed the proportion of women who survived compared to men who survived?**
  + **Is the difference in proportions from your shuffled data larger or smaller than the difference from the original data? Interpret what this means.**
* **Explain why shuffling our data one time is not enough to decide if the difference seen in our *actual* data occurs by chance or not.**

## Detecting differences

* To help us decide if the difference in proportions in our *actual* data occurs by chance or not, we can use the do() function to shuffle our data many times and see how often our *actual* difference occurred by chance.
* Use do, tally and shuffle functions to shuffle the survival variable and tally the proportion of women who survived 500 times. Assign your 500 shuffles the name shuffles
* **View your shuffled data and explain what the rows and each column represents.**
* **For the first row of shuffled data in the shuffles, what is the difference between proportion of females who survived and the proportion of males who survived?**

## Now what?

* The next step to find out how often our *actual* difference occur by chance is to compare it to the differences in our shuffled data.
* To compute the differences for each shuffle we can use the mutate function.
  + Fill in the blanks to add the difference between Survives.Female and Survives.Male to our shuffles data.

shuffles <- mutate(shuffles,   
 diff = \_\_\_\_ - \_\_\_\_)

## Time to decide

* Create a histogram of the differences in our shuffles data. Based on your plot, answer the following
  + **What was the typical difference in proportions between men and women survivors?**
  + **Locate the value of the *actual* difference in the plot. Does the actual difference occur very often by chance alone?**
* **Does gender play a role in whether or not a character will survive in a horror film? Explain your reasoning.**
* **If you wanted to survive in a horror film, would you want to play a female character or a male character?**

## Summary

* By shuffling the survival label, we made it so that the proportion of males and females who survived the slasher film was random.
  + The males and females survived by chance alone.
* If surviving the film occurred purely by chance, then most of the time the difference in survival proportions was close to zero.
  + Notice how most values in the histogram occur close to zero.
* When we look to see how often our actual difference occurs in our shuffled data, if the actual difference doesn't occur very often then perhaps there is something more going on than just chance alone ...

## On your own

* Carry out another 500 simulations but this time shuffle the gender variable instead of the survival variable.
  + Include the code set.seed(1) before your 500 simulations to make your answer reproducible.
* **Does shuffling the gender variable instead of the survival variable change your answer to the question *Does gender play a role in whether or not a character will survive in a horror film?***
  + **Why or why not?**