The Horror Movie Shuffle

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

Some background...

- For this lab, we will be working with the `slasher` data file, which contains information about 485 characters from a random sample of 50 slasher horror films.
- `gender` and `survival` are the two variables in the dataset, each with only two possible values:
- `gender`: Female, Male
- `survival`: Survives, Dies

Initial thoughts...

- We would like to know if a character's gender can give us information about whether or not he/she will survive through the end of the horror film.
- **Write down a hypothesis about the statement above. Do you think males and females are equally likely to die in a horror film? Is one gender more likely to die than the other? Why do you think this?**

Take a look at the data

- After you've loaded the `slasher` data file, create a two-way frequency table of the two variables.

```
tally(survival~gender, data = slasher,
      format = 'count')
```

- **Is it fair to compare the counts of survivors from males and females? Or does it matter that there are more males than females? Explain.**
- **Record the total number of males and females. How many people survived overall? Use RStudio to calculate the *proportion* of each gender who survived.**

First, look at proportions

- Instead of tallying just the counts, let's use proportions.
- What is a proportion? $\frac{\#\text{ observed}}{\text{total}}$

```
tally(survival~gender, data = slasher,
      format = 'proportion')
```

- **Record the proportion of females who survived and the proportion of males who survived. Is there a difference? Is this difference larger or smaller than what you expected?**
- **How can you tell you're calculating the proportion of females/males from the syntax above?**

But we still don't know how to answer...

- Again, how can we know if there's actually a difference between the groups? How far apart do the values need to be?

NA - Survival is like drawing a blue marble from a bag with 18 blue marbles and 82 red marbles. - In our case, men and women draw from the same bag; but, because of chance, the proportions of characters who survive are not exactly the same for each gender.

Take a chance

NA - To see what Horror Movie survival is like in Chance World, we could answer by randomly shuffling the `survives` and `dies` values in the data set. - Imagine that the 485 characters are in a line, just in the order they appear in our data set. NA NA

Do the shuffle!

- When we shuffle data, we still use our original data set as a starting point.
- But, when we want to create the new samples, we have to rearrange (or shuffle) one column of our data.
- We can do that by leaving the `gender` variable as is, but shuffling the `survival` variable.

A little more explaining please...

- If four rows of our `slasher` data looked like:

gender	survival
Female	Dies
Male	Dies
Female	Survives
Male	Dies

- **Which rows differ from the original?**

- We can shuffle `survival` and possibly get this:

gender	survival
Female	Dies
Male	Survives

gender	survival
Female	Dies
Male	Dies

How do I create a whole shuffled data set?!

- Use the `resample()` function!

```
shuffle1 <-
  resample(slasher, shuffled = "survival")
```

- **How does this shuffled sample compare to the original `slasher` data set?** (Hint: Use the `tally()` function)
- **Record the proportion of females who survived and the proportion of males who survived. Is there a difference? Is your value the same as your neighbor's? Why might they be different?**

Let's compare. . .

- Recall that the original `slasher` data returns the following results when using the `tally` function:

Female

Male

Dies

0.77

0.87

Survives

0.23

0.13

- **Is the difference in proportions from your shuffled data larger or smaller than the difference from the original data? What might this mean?**

Detecting differences

- In order to know if there is a difference in proportions in our actual data, we have to create many shuffled data sets (not just one).
- We can use the `do()` function to help us with this!

```
shuffles <-
  do(300)*resample(slasher,
             shuffled="survival")
```

- **How many samples did we create? What part of the code tells you this?**
- **Look at the data file. Can you tell which observations belong to which shuffle?**

We need to correct this!

- In order to sort our shuffled data sets, we can simply tell RStudio to tally the proportions for each shuffle.

```
tally300 <- do(300) *
             tally(survival~gender,
             data = resample(slasher,
             shuffled = "survival"),
             format = 'proportion')
```

- Note that our `data` in this function is a new shuffle since we're using the `resample()` function.
- Look at the top of the data file (using `head(tally300)`) to see what this new tallied data looks like.

Now we have all the proportions!

- Each of the 300 shuffled data sets are contained in this table by giving us the proportions of:
- Females who Died
- Females who Survived
- Total percent of females
- Males who Died
- Males who Survived
- Total percent of males
- **What are the variable names associated with this new data frame?**

Now what?

- Our original `slasher` data showed the difference in proportions of gender survival to be about 0.10 (females - males = 0.23 - 0.13 = 0.10).
- **Looking at your first row of simulated data in the `tally300` data frame, what is the difference between females who survived and males who survived?**
- **How does this value compare to the true difference in proportions?**

But don't we have 300 rows?!

- Instead of having to calculate each of the 300 differences by hand, we can use RStudio!
- We can simply calculate the differences between the females who survive and the males who survive by using the `transform()` function introduced in Lab 1.7b.

```
tally300 <- transform(tally300,
            difference = Survives.Female -
                         Survives.Male)
```

- **What new variable was added to our `tally300` data set?**

I need a picture...

- Just seeing all these values for the differences isn't helpful. Let's look at a histogram!

```
histogram(~difference, data = tally300)
```

- **What is the range of your differences?**
- **What is the typical difference?**
- **Recall that the difference in the actual data was 0.10. How does that value fit into your histogram? (Is it near the center of the distribution? Is it in the tails? Is it included at all?)**

What does all this mean?

- We created 300 random shuffles of our data to see if gender can actually tell us something about whether a character survives through the end of a horror film.
- In these shuffles, we assumed that there was NO difference between genders and each gender had an equal likelihood of surviving.
- If the actual value falls in the center of the distribution, then the difference between genders just happened by chance.
- If the actual value falls in the tails, or is not in the distribution at all, then there is an actual difference between the genders, meaning that one gender has a higher likelihood of surviving a horror film.

So what do you think?

- **Does gender play a role in whether or not a character will survive in a horror film? Explain your reasoning and include a plot.**
- **If you wanted to survive in a horror film, would you want to play a female character or a male character?**