The Color Shuffle

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

The benefit of computers

- In class, you shuffled your group's color labels only a handful of times.
- Each time, you looked at the difference in the median color scores for the class' most occurring predominant color.
- Today, we will shuffle the group labels hundreds of times to see what the median difference looks like when something is caused, solely, by chance.
- **NOTE**: This lab is written assuming the class' most occurring predominant color is *Green* . If your class' most occurring primary color is NOT green, be sure to change questions and code accordingly.

Before we begin

- If we shuffle our data many times and compare the medians:
- **Write down what you think the smallest median difference will be.**
- **What do you think the largest difference will be?**

- **What do you think the typical difference will be?**
- **In class, you found the real difference in median *Green* scores between the true Green group and everyone else. Write down that difference.**

Step 1: Upload your data

- For this lab, we will be looking at your class' actual *Personality Color* data.
- So head to the Mobilize Web Front-end and then *Download*, *Upload*, and *Load* your class' data.
- Assign your data the name: `colors`.
- If you have forgotten how to *Download*, *Upload*, and *Load* your data, refer back to how you did this during the Unit 1 labs.

Greens and Others

- Just like in the classroom activity, we want to group people whose predominant personality color is *Green* together and those whose predominant personality color is not Green together.
- An easy way to do this is with the `ifelse` function.
- Run the following code to create a new variable called `pc_group`.

```
colors <-
  transform(colors, pc_group =
    ifelse(p_color.label == "Green",
           "Green", "Other"))
```

What did we just do?

```
colors <-
  transform(colors, pc_group =
    ifelse(p_color.label == "Green",
           "Green", "Other"))
```

- This code is pretty complex. So let's break it down:
- First, we tell R that we want to `transform` our dataset called `colors`...
- Specifically, we want to create a new variable called `pc_group` (*predominant color group*).
- To assign which group (*Green* or *Other*) each student belongs in, we use the `ifelse` function.

Using ifelse:

```
colors <-
  transform(colors, pc_group =
    ifelse(p_color.label == "Green",
           "Green", "Other"))
```

- **IF** a person's `p_color.label` variable has the value of `"Green"`
- That is, if `p_color.label == "Green"`
- **then** give the new `pc_group` variable the value `"Green"`
- ...

Using ifelse:

```
colors <-
  transform(colors, pc_group =
    ifelse(p_color.label == "Green",
           "Green", "Other"))
```

- **ELSE** (or otherwise) if the student's `p_color.label` is *NOT* `"Green"`
- **then** assign their `pc_group` variable the value `"Other"`

Now we learn to shuffle

- The term data scientists use for *shuffling* data is *resampling*.
- If we want to calculate the medians for our `"Green"` and `"Other"` groups we should write:

```
median(~green | pc_group,
       data=colors)
```

- To randomly assign the *predominant color* labels and compute each group's (randomized) median scores:

```
median(~green | pc_group,
       data=resample(colors,
              shuffled="pc_group"))
```

What just happened?

```
median(~green | pc_group,
       data=resample(colors,
              shuffled="pc_group"))
```

- By writing: `data=resample(colors, shuffled="pc_group")`
- R takes our `colors` data . . .
- And resamples (or 'shuffles') it up . . .
- By shuffling all of the `pc_group` values.
- All of the other values of the data stay the same.
- After resampling, the median *Green* scores for the different groups are completely random.

Shuffling many times

- So why should we bother resampling?
- If we resample lots and lots of times, we can see how often our actual observed difference occurs by chance.
- Knowing this will help us decide if people with a *Green* predominant color typically have large *Green* color scores.
- Use a `do`-loop to compute the shuffled-medians 300 times.

```
shfl_colors <-
  do(300)*median(~green | pc_group,
       data=resample(colors,
        shuffled="pc_group"))
```

What have we got now?

- Now that we have shuffled our data 300 times . . .
- And each time computed the medians for our randomized data.
- We can type the following to see the first few randomized medians:

```
head(shfl_colors)
```

- And we can calculate the difference in color score between our randomized `"Green"` and `"Other"` groups.

Finding the difference

- Similar to how we used `transform` to add our new variable `pc_group` to our `colors` data, we can use `transform` on our shuffled data to compute the difference in median values.
- To do so, run:

```
shfl_colors <-
  transform(shfl_colors,
    Diff = Green - Other)
```

- **Explain, in your own words, what this code does exactly.**

On your own

- **Make a visualization of the difference in medians.**

- **What was the typical difference? What was the largest difference? What was the smallest?**
- **How does the true median difference compare to this distribution of randomized differences?**