Lab 2A - All About Distributions

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

## In the beginning ...

* Most of the labs thus far have covered how to visualize, summarize, and manipulate data.
  + We used visualizations to explore how your class spends their time.
  + We also learned how to clean data to prepare it for analyzing.
* Starting with this lab, we'll learn to use R to answer statistical questions that can be answered by calculating the mean, median and MAD.

## How to talk about data

* When we make plots of our data, we usually want to know:
* Where is the *bulk* of the data?
* Where is the data more *sparse*, or *thin*?
* What values are *typical*?
* How much does the data *vary*?
* To answer these questions, we want to look at the *distribution* of our data.
  + We describe *distributions* by talking about where the *center* of the data are, how *spread* out the data are, and what sort of *shape* the data has.

## Let's begin!

* *Export*, *upload* and *import* your class' *Personality Color* data.
  + Name your data colors when you load it.
* Before analyzing a new data set, it's often helpful to get familiar with it. So:
  + **Write down the names of the 4 variables that contain the point-totals, or *scores*, for each personality color.**
  + **Write down the names of the variables that tell us an observation's *birth gender* and whether they participated in playing *sports*.**
  + **How many variables are in the data set?**
  + **How many observations are in the data set?**

## Estimating centers

* Create a dotPlot of the scores for your *predominant color*.
  + Pro-tip: If the dotPlot comes out looking wonky, try changing the value of the *character expansion* argument, cex.
  + The default value is 1. Try a few values between 0 and 1 and a few more values larger than 1.
* Based on your dotPlot:
  + **Which values came up the most frequently? About how many people in your class had a score similar to yours?**
  + **What, would you say, was a *typical* score for a person in your class for your predominant color? How does your own score for this color compare?**

## Means and medians

* *Means* and *medians* are usually good ways to describe the *typical* value of our data.
* Fill in the blank to calculate the mean value of your predominant color score:

mean(~\_\_\_\_, data = colors)

* **Use a similar line of code to calculate the median value of *your* predominant color.**
  + **Are the mean and median roughly the same? If not, use the dotPlot you made in the last slide to describe why.**

## Comparing birth\_genders

* Make a dotPlot of your *predominant color* again; but this time, facet the plot based on gender.
* Use a line of code, using similar syntax to how you facet plots, to *calculate* a value that describes the *center* of *each* birth gender.
  + **Do males and females differ in their typical scores for your predominant color? Answer this statistical question using your dotPlot.**
* **Assign the mean values a name. Then place the name into the diff() function to calculate the difference. How much more/less did one birth gender score over the other for your predominant color?**

## Estimating Spread

* Now that we know how to describe our data's *typical* value we might also like to describe how closely the rest of the data are to this *typical* value.
  + We often refer to this as the **variability** of the data.
  + Variability is seen in a histogram or dotPlot as the horizontal *spread*.
* **Look at the spread of the dotPlot you made for your predominant color then fill in the blank:**

*Data points in my plot will usually fall within* \_\_\_\_ *units of the center.*

* **Which birth gender, if either, seem to have values that are more spread out from the center?**

## Mean Absolute Deviation

* The **mean absolute deviation** finds how far away, on average, the data are from the mean.
  + We often write *mean absolute deviation* as *MAD*.
* Calculate the MAD of your *predominant color* by filling in the blanks:

MAD(~\_\_\_\_\_, data = colors)

* **Based on the MAD, which birth gender has more variability for your predominant color's scores?**
  + **Does this match the answer you gave for the last question in the previous slide?**

## On your own

* Do boys and girls in your class differ in their color scores?
  + **Perform an analysis that produces *numerical summaries* and *graphs*.**
  + **Then, write a few sentence interpretations that addresses this statistical question and considers the *shape*, *center* and *spread* of the distributions of the graphs you create.**