Lab 2B - Oh the Summaries ...

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

## Just the beginning

* Means, medians,and MAD are just a few examples of *numerical summaries*.
* In this lab, we will learn earn how to calculate and interpret additional summaries of distributions such as: minimums, maximums, ranges, quartiles and IQRs.
  + We'll also learn how to write our first custom function!
* Start by loading your *Personality Color* data again and name it colors.

## Extreme values

* Besides looking at *typical* values, sometimes we want to see *extreme* values, like the smallest and largest values.
  + To find these values, we can use the min, max or range functions.
* **Which of the color scores had the smallest min value? Which had the largest max value?**
* **Use the range function to calculate the max and min values of your predominant color**
  + The difference between a variable's smallest and largest value is often refered to as the *range* of the variable.

## Calculating a range value

* We saw in the previous slide that the range function calculates the maximum and minimum values for a variable, but not the difference between them.
* We could calulate this difference in two steps:
  + Step 1: Use the range function to assign the max and min values of a variable the name values.
  + Step 2: Use the diff function to calculate the difference of values.
* **Use these two steps to calculate the *range* of your predominant color.**

## Introducing custom functions

* Calculating the *range* of many variables can be tedious if we have to keep performing the same two steps over and over.
  + We can combine these two steps into one by writing our own custom function.
* Custom functions can be used to combine a task that would normally take many steps to compute and simplify them into one.
* The next slide shows an example of how we can create a custom function called mm\_diff to calculate the absolute difference between the mean and median value of a variable in our data.

## Example function

mm\_diff <- function(variable, data) {  
 mean\_val <- mean(variable, data = data)  
 med\_val <- median(variable, data = data)  
 abs(mean\_val - med\_val)  
}

* The function takes two *generic* arguments: variable and data
* It then follows the steps between the curly braces {}
  + Each of the *generic* arguments is used inside the mean and median functions.
* Copy and paste the code above into a *R script* and *run* it.

## Using mm\_diff()

* After running the code used to create the function, we can use it just like we would any other numerical summary.
  + In the *console*, fill in the blanks below to calculate the absolute difference between the mean and median values of your predominant color:

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

* **Which of the four colors has the largest absolute difference between the mean and median values?**
  + **By examining a dotPlot for this personality color, make an argument why either the mean or median would be the better description of the *center* of the data.**

## Our first function

* Using the previous example as a guide, create a function called Range (*Note the capial 'R'*) that calculates the *range* of a variable by filling in the blanks below:

\_\_\_\_ <- function (\_\_\_\_, \_\_\_\_) {  
 values <- range(\_\_\_\_, data = \_\_\_\_)  
 diff(\_\_\_)  
}

* **Use a dotPlot or histogram to find the personality color with the largest difference between the max and min values. Then use the Range function you created to calculate its *range*.**

## Quartiles (Q1 & Q3)

* The *median* of our data is the value that splits our data in half.
  + Half of our data is smaller than the *median*, half is larger.
* *Q1* and *Q3* are similar.
  + 25% of our data is smaller than *Q1*, 75% are larger.
* Fill in the blanks to compute the value of *Q1* for your predominant color.

quantile(~\_\_\_\_, data = \_\_\_\_, p = 0.25)

* **Use a similar line of code to calculate *Q3*, which is the value that's larger than 75% of our data.**

## The Inter-Quartile-Range (IQR)

* Make a dotPlot of your *predominant* color's scores.
* Visually (Don't worry about being super-precise):
  + Cut the distribution into quarters so the *number* of *data points* is equal for each piece. (Each piece should contain 25% of the data.)
  + **Write down the numbers that split the data up into these 4 pieces.**
  + **How long is the interval of the middle two pieces?**
  + This length is the *IQR*.

## Calculating the IQR

* The IQR is another way to describe *spread*.
  + It describes how *wide* or *narrow* the middle 50% of our data are.
* Just like we used the min and max to compute the range, we can also use the *1st* and *3rd* quartiles to compute the *IQR*.
* **Use the values of *Q1* and *Q3* you calculated previously and find the *IQR* by hand**.
  + **Then use the iqr() function to calculate it for you.**
* **Which personality color score has the widest spread according to the *IQR*? Which is narrowest?**

## Boxplots

* By using the medians, quartiles, and min/max, we can construct a new single variable plot called the *box and whisker* plot, often shortened to just *boxplot*.
* **By showing someone a dotPlot, how would you teach them to make a *boxplot*? Write out your explanation in a series of steps for the person to use.**
  + **Use the steps you write to create a sketch of a *boxplot* for your predominant color's scores in your journal.**
  + **Then use the bwplot function to create a *boxplot* using R.**

## Our favorite summaries

* In the past two labs, we've learned how to calculate numerous *numerical summaries*.
  + Computing lots of different summaries can be tedious.
* Fill in the blanks below to compute some of our *favorite* summaries for your predominant color all at once.

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

## On your own

* **Create a function called myIQR that uses the *only* quantile function to compute the middle 30% of the data.**