Oh the summaries. . .

Unit 2 - Lab 2

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

Just the beginning

- Means, medians, MAD are just a few examples of **numerical summaries**
- **Numerical summaries** are numbers that describe characteristics of the data.
  - Means and medians described the *center* of the data.
  - MAD describes the *spread* of the data.
- **What other numbers might describe your data?**

To start

- Load your personality color data again and name it: `colors`.
- In the lines of code that appear on the following slides:
  - Replace any of the `~x` that appear in the code with the name of the variable for your *predominant* color's score.

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 calculate the `min` and `max`.

```
min(~x, data=colors)
```

```
max(~x, data=colors)
```

Range

- Using the `min` and `max` values, we can calculate the `range`.
- The range is another, but often less informative, measure of *spread*.
  - Calculate it by taking: `max - min`.

Calculating the range

- Try the following

```
max(~x, data = colors) - min(~x, data = colors)
```

```
range(~x, data = colors)
```

- Notice how we can treat `max(~x, data = colors)` and `min(~x, data = colors)` just like we would a number.
- **Why did we say that the `range` is less informative than the `MAD`?**
  - **Can you think of examples where the `MAD` will give you a better idea of the *variability* than the `range`?**

Quartiles (Q1 & Q3)

- We often use the `median` to describe the *center* of our data because half of the data is smaller than the median and the other half is larger.
- If instead we found a value that was larger than just 25% of our data, we would have computed the *1st quartile*.
- If we found the value that was larger than 75% of our data, we call that the *3rd quartile*.
- **Why do you think we use the names *'1st* and *3rd quartiles*?**

The Inter-Quartile-Range (IQR)

- Just like we used the `min` and `max` to comupute the `range`, we can also use the *1st* and *3rd* quartiles to compute the `IQR`.
- The `IQR` is another way to describe *spread*.
  - It describes how *wide* or *narrow* the middle 50% of our data are.
  - If the `IQR` is a small number, then the middle 50% of our data is close to the `median`.
  - Otherwise, the middle 50% of our data is further away from our median.

Finding the IQR

- Make a histogram of your predominant color's scores.
- Visually:
  - 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

- Calculate the IQR by using either of the following

```
IQR(~x, data = colors)
```

```
iqr(~x, data = colors)
```

- **How close was your visual estimate to the actual IQR?**

Other quantiles

- The median, 1st and 3rd *quartiles* can also be called the the 50th, 25th and 75th *quantiles*.
  - They're called *quantiles* becuase they're the *quantity* of data that is smaller than that value.
  - The 25th quantile is the value that is larger than 25% of the data.
- We can compute quantiles too!

```
qdata(~x, data = colors, p = 0.35)
```

- Where `p` stands for the *percentage* of data you'd like our value to be larger than.

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 a **boxplot**.
- Try making one of your predominant color.

```
bwplot(~x, data=colors)
```

- **Sketch your boxplot in your journal. Label the `min` `max`, `Q1`, `Q3` and `median`.**
- **How would you interpret your boxplot? Where is the bulk of your data? Where is it centered? Can you say anything about its shape?**

Our favorite summaries

- Numerical summaries are brief ways to describe our data, using numbers.
- Computing lots of different summaries though can be tedious.
- Use the following command to compute some of our *favorite* summaries

```
favstats(~x, data=colors)
```

- **Which summaries are displayed?**
- **What do you think `n` stands for?**