All About Distributions

Unit 2 - Lab 1

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

How to talk about data

- When we make plots of our data, often times 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.
  - Where's the **center** of our data?
  - How is it **spread**?
  - What sort of **shape** does it have?

Let's begin!

- Start by answering the following questions in your journal:
  - **What was your *predominant color*? And what was your score for that color?**
  - **Do you think your score for your predominant color was *typical* in your class?**
  - **Which color's score do you predict had the most variation?**
- Now: Download, upload and load your class' personality color data.
  - Name your data: `colors`

Exploring the data

- Before analyzing a new data set, it's often helpful to get familiar with it.
- Answer the following in your journal:
  - **Write down the `names` of the 4 different *color score* variables.**
  - **How many variables are there in the data?**
  - **How many observations are there?**

Centers

- Use the following code to make a `dotPlot` of the scores for your *predominant color*
  - Replace `x` with your predominant color's variable name.

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

- **Based on the plot, what would you say was the *typical* score for your class for this color?**
- **Why did you choose this value as the *typical* value for your class?**

Means and medians

- **Means** and **medians** are usually good ways to *estimate* the *typical* value of our data.
- Type the following commands into the console (But replace `x` by your predominant color variable):

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

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

- **Do the *mean* and *median* give roughly the same answer?**

Comparing birth_genders

- Make a dotPlot of your *predominant color* again but this time split (facet) the plot based on gender.
  - **Do males and females differ in their typical scores? Answer by giving the values of the centers for each distribution.**
  - Use the following code to check your estimates:

```
mean(~x | birth_gender.label,
     data = colors)
```

- **Is the bulk of one gender's data closer/further-away from the center? What about the plot makes you think that?**

Estimating Spread

- We already saw how to calculate an estimate for the data's *typical* value by finding its *mean*.
- We might also like to know how closely the rest of the date is to this *typical* value.
  - We often refer to this as the **variability** of the data.
  - Variability is seen in a histogram as the horizontal *spread*

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 by using the following code:

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

- **Which gender has a larger MAD? Are they both the same?**
- **What does `MAD` tell us about the values of our data?**
- **Write down the code you used to calculate each gender's MAD.**

Shapes

- Now we know how to estimate a distribution's *typical* value and it's *variability*.
- The last way we usually describe what data looks like is by describing its *shape*.

Shape is different than center and spread

- To describe a distribution's shape, we don't calculate a number.
  - We just look at a plot!
  - This means that shape is often up to interpretation.
- When describing the shape of a distribution we want to know:
  - Between what values is the bulk of our data.
  - Between what values is the data relatively scarce.

Shape is different than center and spread

- Create a `dotPlot` for your predominant color but split it into two graphs based on whether your classmate played a sport or not.
  - **Most of our data is contained between which values?**
  - **Describe which values don't have very much data.**
  - **Are the shapes of the distributions drastically different?**
  - **Would you say that either distribution is *skewed*? Or *symmetric*?**

On your own

- Using your the scores of your *secondary* color, answer the following questions in your journals:
  - **What scores are the most/least common? Write down any code you used to find your answers.**
  - **What was the typical score for your secondary color? Did males/females have different typical scores?**
  - **Do the values of your secondary color have more or less variablility than those of your predominant color? How do you know? Use a graph and a numerical summary to support your answer.**