All About Distributions

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

In the beginning . . .

- Most of the labs thus far have been about learning how to visualize, summarize, and manipulate data.
- Now we'll start to learn some more formal and advanced techniques.
- We'll start this lab by learning some commands for *numerical summaries*.
- We'll also cover how we talk about the *distribution* of data.

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 is, how **spread** out it is, and what sort of **shape** it has.

Let's begin!

- Start by answering the following questions about the *Personality Color* survey you took:
  - **What was your *predominant Personality Color*? What was your point-total for that color?**
  - **Do you think the point-total for your predominant color was *typical* for your class?**
- Now: Download, upload, and load your class' *Personality Color* data.
  - Name your data `colors` when you load it.

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 variables that contain the point-totals for each color.**
  - **How many variables are in the data set?**
  - **How many observations are in the data set?**

Centers

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

```
dotPlot(~x, data = colors, cex = 1.0)
```

- Pro-tip: You can make the dots in your plot smaller/larger by changing the value of `cex`.
- **Based on the plot, what would you say was the *typical* score for your class for this color?**
- **Explain why you chose 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 (remember to replace the `x` with the variable name of your predominant color):

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

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

- **Do the *mean* and *median* give roughly the same answer?**
- **Explain why you think they do or do not.**

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 for this color? 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 or 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 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*.

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 using the following code and **record the value**:

```
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.**

Standard Deviation

- A similar method for quantifying the *variability* of our data is the **standard deviation**.
- We abbreviate the *standard deviation* as *SD*.
- Want to know why *SD* was invented? . . . Take calculus!
- Calculate the SD by using the following code:

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

- **Is the gender with the larger SD the same as the gender with the larger MAD?**
- **Is this surprising? Why or why not?**

Shapes

- Now we know how to estimate a distribution's *typical* value and its *variability*.
- The last way we usually describe what data look like is by describing the distribution's *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 do the bulk of our data fall?
  - Between what values do the data appear relatively scarce (or thin)?
  - Is the data equally distributed on both sides of the center?
  - Is more data on one side of the center than the other?

Shape is different than center and spread

- Create a `dotPlot` for your *predominant color*, but split it into two graphs based on whether your classmates are active in sports or not.
  - **Most of the data are contained between which values?**
  - **Describe which values don't have very much data.**
  - **Are the shapes of the distributions drastically different for classmates who participate in sports versus those who don't?**
  - **Would you say that either distribution is *skewed*? Or *symmetric*? Explain and provide your graph.**

On your own

- Using the scores of your *secondary* personality color, answer the following questions in your journals:
  - **What scores are the most/least common for this color? 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 variability than those of your predominant color? How do you know? Use a graph and a numerical summary to support your answer.**