Text: Descriptive Statistics Summary

Recap

Variable Types

We have covered a lot up to this point! We started with identifying data types as either `categorical` or `quantitative`. We then learned, we could identify quantitative variables as either `continuous` or `discrete`. We also found we could identify categorical variables as either `ordinal` or `nominal`.

Categorical Variables

When analyzing categorical variables, we commonly just look at the count or percent of a group that falls into each **level** of a category. For example, if we had two **levels** of a dog category: `lab` and `not lab`. We might say, 32% of the dogs were `lab` (percent), or we might say 32 of the 100 dogs I saw were labs (count).

However, the 4 aspects associated with describing quantitative variables are not used to describe categorical variables.

Quantitative Variables

Then we learned there are four main aspects used to describe `quantitative` variables:

1. Measures of **Center**
2. Measures of **Spread**
3. **Shape** of the Distribution
4. **Outliers**

Measures of Center

We looked at calculating measures of `Center`

1. **Means**
2. **Medians**
3. **Modes**

Measures of Spread

We also looked at calculating measures of `Spread`

1. **Range**
2. **Interquartile Range**
3. **Standard Deviation**
4. **Variance**

Shape

We learned that the distribution of our data is frequently associated with one of the three **shapes**:

1. Right-skewed

2. Left-skewed

3. Symmetric (frequently normally distributed)

Depending on the shape associated with our dataset, certain measures of center or spread may be better for summarizing our dataset.

When we have data that follows a **normal** distribution, we can completely understand our dataset using the `mean` and `standard deviation`.

However, if our dataset is **skewed**, the `5 number summary` (and measures of center associated with it) might be better to summarize our dataset.

Outliers

We learned that outliers have a larger influence on measures like the mean than on measures like the median. We learned that we should work with outliers on a situation by situation basis. Common techniques include:

1. At least note they exist and the impact on summary statistics.

2. If typo - remove or fix

3. Understand why they exist, and the impact on questions we are trying to answer about our data.

4. Reporting the 5 number summary values is often a better indication than measures like the mean and standard deviation when we have outliers.

5. Be careful in reporting. Know how to ask the right questions.

Histograms and Box Plots

We also looked at histograms and box plots to visualize our quantitative data. Identifying outliers and the shape associated with the distribution of our data are easier when using a visual as opposed to using summary statistics.

What Next?

Up to this point, we have only looked at **Descriptive Statistics**, because we are describing our collected data. In the final sections of this lesson, we will be looking at the difference between **Descriptive Statistics** and **Inferential Statistics**.