Displaying Distributions

- Histograms- A histogram uses adjacent bars to show the distribution of values in a quantitative variable. Each bar represents the frequency (relative frequency) of values falling in an interval of values.
    - The standard rule for a value that falls exactly on a bin boundary is to put it into the next higher bin.
- Stem-and-Leaf Displays- a stem-and-leaf display shows quantitative data values in a way that sketches the distribution of the data. It's best described in detail by example.

Shape

- Shape- The visual appearance of the distribution. To describe the shape, look for: simple vs. multiple modes and symmetry vs. skewness
- Modes- A peak or local high point in the shape of the distribution of a variable. The apparent location of modes can change as the scale of a histogram is changed
    - Unimodal- Having one mode. Generally mound-shaped.
    - Bimodal- Distributions with two modes
    - Multimodal- Distributions with more than two modes.
    - A distribution whose histogram doesn't appear o have any mode and in which all the bars are approximately the same height is called uniform.
- Symmetry
    - Symmetric- a distribution is symmetric if the two halves on either side of the center look approximately like mirror images of each other.
        - The thinner ends of a distribution are called he tails.
        - If one tail stretches out farther than the other, the distribution is said to be skewed to the side of the longer tail.
- Outliers- Extreme values that don't appear to belong with the rest of the data. They may be unusual values that deserve further investigation or just mistakes; there's no obvious way to tell.

Center

- Sigma (Σ) means "sum" $\bar{y} = \text{total}/n = \Sigma y/n$ this gives you the mean.
    - The mean is a natural summary for unimodal, symmetric distributions, it can be misleading for skewed data or for distributions with gaps or outliers.
- Median- the middle value with half of the data above it and half below it.

Spread of the Distribution

- Range- defined as the difference between the extremes. Range= max-min
- Interquartile Range (IQR)- The difference between the first and third quartiles. IQR= Q3 – Q1
- The average of the squared deviations is called the variance.

$$s^2 = \frac{\sum(x_i - \bar{X})^2}{n-1}$$

- Standard Deviation- A measure of spread found as

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Shape, Center, and Spread

- If the shape is skewed, point that out and report the median and IQR. May want to include the mean and standard deviation as well, explaining why the mean and median differ.
- The fact that the mean and median do not agree is a sign that the distribution may be skewed. Histogram will help make your point
- If the shape I unimodal and symmetric, report the mean and standard deviation and possibly the median and IQR as well. For unimodal symmetric data, the IQR is usually a bit larger than the standard deviation.
- Always pair the median with the IQR and the mean with he standard deviation.

Five-Number Summary and Boxplots

- Five-Number Summary- A five-number summary for a variable consists of the minimum and maximum, the quartiles Q1 and Q3, the median.
- Boxplot- A boxplot displays the 5-number summary as a central box with whiskers that extend to the non-outlying values. Boxplots are particularly effective for comparing groups. SEE NOTES ON HOW TO MAKE A BOXPLOT