STAT371: Introductory Applied Statistics for the Life Sciences **Spring 2018**

Chapter 2 — Descriptive Statistics Summary

1. Histograms

   - Histograms are created by dividing the data into bins and plotting the number of observations in each bin as a bar.
   - Histograms help us see the shape of data.
   - Histograms can be either frequency (how many observations are in each bin) or relative frequency (what proportion of the observations in the whole dataset are in each bin).

2. Numerical Summary Measures - Location

   - The **median** is a measure of middle and is the observation in the ordered list such that half the observations are larger and half are smaller.
     - If the dataset consists of an odd number of observations, the median is the middle value in the sorted list.
     - If the dataset consists of an even number of observations, the median falls halfway between the two middle values in the sorted list (the average of the two middle values).
   - There are three **quartiles**, the first, second, and third, which divide the data into approximately four equal groups. The second quartile is the same as the median.
     - If the dataset consists of an even number of observations, the first quartile is the median of the first half of the sorted list, and the third quartile is the median of the second half of the sorted list.
     - If the dataset consists of an odd number of observations, use the following procedures. To find the first quartile, make a new dataset consisting of the first half of the full sorted dataset, and include the median of the full dataset as one of the data points in this new dataset. (The ”first half” of the sorted list will contain the numerically smaller numbers.) Compute the median of this new dataset. The result is the first quartile. Similarly, to find the the third quartile, make a new dataset consisting of the second half of the full sorted dataset, and include the median of the full dataset as one of the data points in this new dataset. Compute the median of this new dataset. The result is the third quartile.

- The **mean**, or average, is defined as the sum of the observations divided by the number of observations. Let $y_1, y_2, ..., y_n$ denote the n observations in a dataset, then define:

  $$\text{Sum: } \sum_{i=1}^{n} y_i = y_1 + y_2 + ... + y_n$$

  and

  $$\text{Mean: } \bar{y} = \frac{\sum\limits_{i=1}^{n} y_i}{n}$$

3. Numerical Summary Measures - Spread

- The **range** is the largest value in a dataset minus the smallest value.

- The **interquartile range** or **IQR** is the third quartile of a dataset minus the first quartile.

- The **standard deviation**, or **SD**, is the square root of the average of the squared deviations from each observation in a dataset to the mean of the dataset. If we let $y_1, y_2, ..., y_n$ denote the individual observations, the formula is:

  $$\text{Standard Deviation: } s_y = \sqrt{\frac{\sum\limits_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$