

Measures of Center and Spread

There are three ways to measure the **center** of a dataset, to summarize a whole column of quantitative data using just one number:

- The **mean** of a dataset is the average of all the numbers.
- The **median** of a dataset is a value that is smaller than half the dataset, and larger than the other half. In an ordered list the median will either be the middle number or the average of the two middle numbers.
- The **mode(s)** of a data set is the value (or values) occurring most often. When all of the values occur equally often, a dataset has no mode.

In a **symmetric** dataset, values are just as likely to occur a certain distance above the mean as below the mean, and the median and mean are usually close together.

When a dataset is asymmetric, the median is a more descriptive measure of center than the mean.

- A dataset with **left skew**, and/or low outliers, has a few values that are unusually low, pulling the mean *below* the median.
- A dataset with **right skew**, and/or high outliers, means there are a few values that are unusually high, pulling the mean *above* the median.

When a dataset contains a small number of values, the mode may be the most descriptive measure of center.

Data Scientists can also measure the **spread** of a dataset using a **five-number summary** :

- The **minimum** – the lowest value in the dataset
- The **first, or “lower” quartile (Q1)** – the middle of the lower half of values, which separates the lowest quarter from the next smallest quarter
- The **second quartile (Q2)** – the middle value, which separates the entire dataset into “top” and “bottom” halves
- The **third, or “upper” quartile (Q3)** – the middle of the higher half of values which separates the second highest quarter from the highest quarter
- The **maximum** – the largest value in the dataset