Descriptive Statistics II

---

5 Number Summary

1. `Minimum`: The smallest number in the dataset.
2. `Q1`: The value such that 25% of the data fall below.
3. `Q2`: The value such that 50% of the data fall below.
4. `Q3`: The value such that 75% of the data fall below.
5. `Maximum`: The largest value in the dataset.

Range

The **range** is then calculated as the difference between the **maximum** and the **minimum**.

IQR

The **interquartile range** is calculated as the difference between `Q3` and `Q1`.

Standard Deviation and Variance

The **standard deviation** is defined as **the average distance of each observation from the mean**.

The variance is **the average squared difference of each observation from the mean**.

Important Final Points

1. The variance is used to compare the spread of two different groups. A set of data with higher variance is more spread out than a dataset with lower variance. Be careful though, there might just be an outlier (or outliers) that is increasing the variance, when most of the data are actually very close.
2. When comparing the spread between two datasets, the units of each must be the same.
3. When data are related to money or the economy, higher variance (or standard deviation) is associated with higher risk.
4. The standard deviation is used more often in practice than the variance, because it shares the units of the original dataset.

The distribution of The data

1. Right-skewed : Mean greater than Median

2. Left-skewed : Mean less than Median

3. Symmetric (frequently normally distributed) : Mean equals Median

The `mode` of a distribution is essentially the tallest bar in a histogram.

Outliers

`outliers` are points that fall very far from the rest of our data points. This influences measures like the mean and standard deviation much more than measures associated with the five number summary.

Identifying Outliers

1. Sorting your values from low to high and checking minimum and maximum values.
2. Visualizing your data with a box plot and looking for outliers.
3. Using the interquartile range to create fences for your data.

---

1. **Population** - our entire group of interest.
2. **Parameter** - numeric summary about a population
3. **Sample** - subset of the population
4. **Statistic** numeric summary about a sample

`Inferential Statistics` **is about using our collected data to draw conclusions to a larger population**.