# Statistics

## 1 Five Number Summary

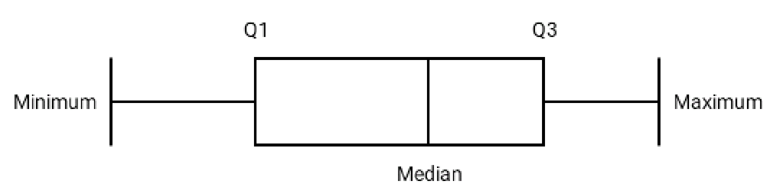
1. *minimum* - The Smallest Value
2. $Q_1$ - The First Quartile – the value such that 25% of the cases are less than or equal to $Q_1$
3. *median* - The value such that 50% of the cases are less than or equal to *median* and 50% of the cases are greater than or equal to *median*
4. $Q_3$ - The Third Quartile – the value such that 75% of the cases are less than or equal to $Q_3$
5. *maximum* - The Largest Value

### 1.1 Computing the Five Number Summary

1. Sort your data in ascending order.
2. Find the *minimum* and *maximum*, these will be the numbers at the beginning and end of your sorted data.
3. Find the *median* by observing the value which divides the list in two. If you have an even number of cases, average the two middle values.
4. Find the $Q_1$ value by drawing brackets around the numbers that are less than the median. Find the median of this set.
5. Find the $Q_3$ value by drawing brackets around the numbers that are greater than the median. Find the median of this set.

### 1.2 Box Plots

- Boxplots are graphical representations of the five number summary.
- They give an idea of the distribution of cases.
- Plotting boxes side by allow us to compare sets of data.



## 2 Mean

- The mean is another measure of center.
- The mean is the value each data item would hold if we evenly distributed the value across all the cases. (Like communism, but without the oppressive regimes!)
- The mean is computed by the formula:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The variables in this (and other statistics calculations) are:
  - $\bar{x}$ – The Mean
  - $X$ – The Set of All Cases
  - $x_i$ – The $i^{\text{th}}$ Case in $X$
  - $n$ – The Number of Cases in $X$

## 3 Measuring Spread About the Mean

- If we think of the mean as a measure of center, how can we measure the spread about the mean?
- We may attempt to measure the average deviation from the mean:

$$\frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x}$$

- Let's try it out on the exam data!
- What happened?
- Why do we get zero, or a number very close to it?
- We could fix this by squaring each difference (so now they are all positive)

$$\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Now, the problem becomes that we have the wrong units! Why?
- So we correct this, and arrive at the **standard deviation**:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- The above formula is the theoretical standard deviation, but is often biased toward extant data. So we typically make the deviation a bit wider by decreasing the denominator by one:

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

- $\sigma$ is the population standard deviation and $s$ is the sample standard deviation.
- From now on, we will always use $s$.
- Let's compute $s$ for the exam data! What does $s$ tell us?