

## Measuring Spread

We use quartiles to measure the spread of a distribution. They are similar to the median which is used to measure center.

The median is the number such 50% of the observations are below it and 50% of the observations are above it.

The first quartile  $Q_1$  is the number such that 25% of the observations are below it and 75% of the observations are above it.

The third quartile  $Q_3$  is the number such that 75% of the observations are below it and 25% of the observations are above it.

The quartiles break the distribution up into quarters (hence the name). The median could also be called the second quartile.

We can combine these numbers with the minimum and maximum of the distribution to create the five-number summary of the distribution, which describes the distribution in some detail (center, spread).

Minimum    $Q_1$    Median    $Q_3$    Maximum

**Example** Here is a data set we can explore: 198, 167, 202, 190, 247, 238, 354, 234.

The first thing we want to do is order the list from highest to lowest (this is the opposite of what you would do for a stemplot, we shall see why in a minute):

354	Maximum
247	
	Third Quartile $= Q_3 = (247 + 238)/2 = 242.5$
238	
234	
	Median $= (234 + 202)/2 = 218$
202	
198	
	First Quartile $Q_1 = (198 + 190)/2 = 194$
190	
167	Minimum

There are  $n = 8$  observations.

### Median

The median is the number such that 50% of the observations are below, and 50% are above.

Since there is an even number of observations, there are two “middle” observations. We average these two middle observations to get the median.

### First Quartile

The first quartile  $Q_1$  is the number such that 25% of the observations are below, and 75% are above.

The first quartile will have  $25\% \times 8 = \frac{25}{100} \times 8 = 2$  observation below it and  $75\% \times 8 = \frac{75}{100} \times 8 = 6$  observation above it.

Since there is an even number of observations, there is no number from the data set for which this is true. We average two numbers from the data set to get the first quartile.

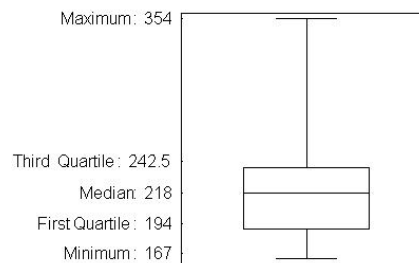
### Third Quartile

The third quartile  $Q_3$  is the number such that 75% of the observations are below, and 25% are above.

The first quartile will have  $75\% \times 8 = \frac{75}{100} \times 8 = 6$  observation below it and  $25\% \times 8 = \frac{25}{100} \times 8 = 2$  observation above it.

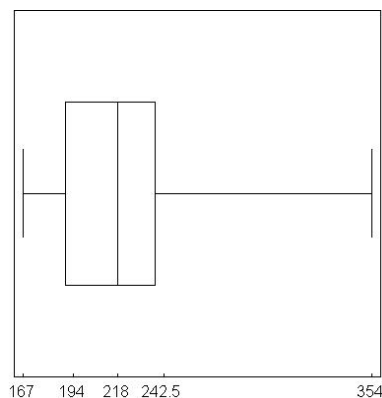
Since there is an even number of observations, there is no number from the data set for which this is true. We average two numbers from the data set to get the third quartile.

If we take the information we have above we can construct a box plot which visually represents the five number summary:



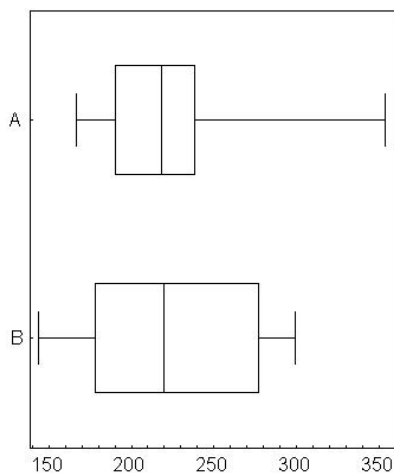
The box represents the spread of the middle half of the distribution; notice the median is NOT in the middle of the box.

In this representation, the vertical distances are important, and the horizontal distances meaningless. A boxplot could also be produced so the horizontal distances are the important ones:



Boxplots are useful for comparing more than one distribution at a time.

**Example** The boxplots for two distributions are given below.



We can instantly see that distribution B has a much larger spread than distribution A, although the medians of the two distributions are very close together, and the difference between the maximum and minimum values in distribution A is larger than for distribution B.

### Variance and Standard Deviation

There are other ways of describing the center and spread of a data set, which are more often reported. They involve the mean and standard deviation.

The standard deviation is a measure of how far the observations are from their mean. The standard deviation is the square of the variance.

$$\text{Mean} = \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\text{Variance} = s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

$$\text{Standard Deviation} = s = \sqrt{s^2}$$

You can use a calculator to determine the standard deviation if you need it, so don't worry about memorizing this formula.

Note that the standard deviation has the same units as the mean.

**Summary**

- two methods of describing center and spread of a distribution:
  - five number summary (min, first quartile, median, third quartile, max), leads to boxplots
  - mean and standard deviation
- the mean and standard deviation are greatly affected by outliers.
- the mean and standard deviation do not display the skewness of the distribution.
- the mean and standard deviation are best used for symmetric distributions without outliers.
- skewed distributions are best described by the five number summary, since the boxplot easily displays information about the skewness of the distribution.

The ideas used to construct quartiles can be used to construct:

- quintiles, which divides the distribution into 5 pieces,
- the deciles, which divides the distribution into 10 pieces, and the
- percentiles, which divides the distribution into 100 pieces.

For example,

The 90th percentile is equal to the 9th decile for a distribution.

The 75th percentile is equal to the 3rd quartile for a distribution.

The 40th percentile is equal to the 2nd quintile for a distribution.