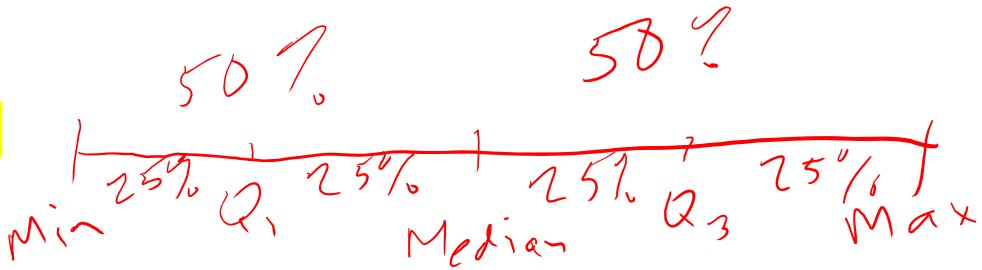
MATH 3307
Lesson 4

More measures of spread (or dispersion):

- Range – Full coverage of the data set from smallest to largest.
  Range = Max - Min

Drawbacks of range: sensitivity to **outliers**



- Percentiles:
  - 25^{th} percentile, Q1 – First, or Lower, Quartile. This is the median of the lower half of the data set.
  - 50^{th} percentile, Median or Q2 – Second Quartile, or Median, divides your data in half.
  - 75^{th} percentile, Q3 – Third, or Upper Quartile. This is the median of the upper half of your data.

The values of the minimum, Q1, Q2, Q3 and the maximum make up what is called our **five number summary**.

- IQR – Interquartile Range: The distance between the first and third quartiles in your data set.

  IQR = Q3 – Q1

  Relevance: It is the distance (or the range) of your middle 50% of data.

Five Number Summary: the list of data, in order: Minimum, First Quartile, Median, Third Quartile, Maximum

Example:

1. Twelve babies spoke for the first time at the following ages (in months):

8 9 10 11 12 13 15 15 18 20 20 26

Find Q1, Q2, Q3, the range and the IQR.

```
> fivenum(baby)
[1]  8.0 10.5 14.0 19.0 26.0
```

Based on the above five-number summary:
Min: 8
Q1: 10.5
Med: 14
Q3: 19
Max: 26

Range: Max – Min = 26 – 8 = 18
IQR: Q3 – Q1 = 19 – 10.5 = 8.5

To copy: 8 9 10 11 12 13 15 15 18 20 20 26

In R Studio, fivenum(list) gives the five point summary for a list.

The IQR is used to determine data classified as **outliers**. An outlier is an observation that is "distant" from the rest of the data. Outliers can occur by chance or be measurement errors so it is important to identify them. Any point that falls outside the interval calculated by Q1- 1.5(IQR) and Q3 + 1.5(IQR) is considered an outlier.

Outliers on the high side: Anything larger than: Q3 +1.5*(IQR)

Example: Outliers on the low side: Anything smaller than Q1 – 1.5*(IQR)

2. Are there any outliers in the data set given for example 1? If so, what are they?

```
> IQR=8.5
> 19+1.5*IQR
[1] 31.75
> 10.5-1.5*IQR
[1] -2.25
```

Q3 + 1.5*IQR: This means that any data points larger than 31.75 are outliers on the high side. (None for this data set)

Q1 – 1.5*IQR: This means that any data points lower than -2.25 are outliers on the low side. (None for this data set)

Hypothetical Example: If we added one additional point (value of 32) this would be an outlier on the high side of our data, because its value is larger than 31.75.

There are other percentiles as well. The ***k*th percentile** means that $k\%$ of the ordered data values are at or below that data value. For example, if the median is 100, then 50% of the ordered data values fall at or below 100. Also, $(100\text{-}k)\%$ represents the amount of ordered data that falls above the percentile data value.

If you are looking for the measurement that has a desired percentile rank, the $100P$th percentile, is the measurement with rank (or position in the list) of $nP+0.5$, where n represents the number of data values in the sample.

n: the total number of data points in the set of data.
P: is the percentile, given as a decimal.

The result to this equation will be the data point (its order in line) that has the desired percentile.

Example:

$n = 30$ $P = 0.30$

3. In a collection of 30 data measurements, which measurement represents the 30th percentile?

$nP + 0.5$

$(30)(.30) + 0.5$

$9.0 + 0.5$

$9.5 \longrightarrow 10$

Always Round up to the next whole number

This means that the 10th data point in an ordered of 30 data points is above 30% of the data (30th percentile).

Suppose you know the position (the order) of a value and want to know what percentile it is ranked at. In general, if you have n data measurements, x_1 represents the $100(1-0.5)/n^{\text{th}}$ percentile, x_2 represents the $100(2-0.5)/n^{\text{th}}$ percentile, and x_i represents the $100(i-0.5)/n^{\text{th}}$ percentile.

The result of this formula would be the percentile that your desired data point falls in.

i: is the position in line (known as your percentile rank)
n: sample size

Example:

4. Using the data in example 1, determine the percentile of the 4^{th} order statistic (x_4).

This was the example concerning when babies spoke for the first time.

8 9 10 11 12 13 15 15 18 20 20 26

$i = 4$

$n = 12$

$$\frac{100(i-0.5)}{n} = \frac{100(4-0.5)}{12} =$$

This means that the baby that spoke after 11 months (4^{th} item in our list) spoke after 29% of others in the data set.

29^{th} percentile