MATH 3307
Lesson 4

More measures of spread (or dispersion):

- Range –

Drawbacks of range: sensitivity to **outliers**

- Percentiles:
  - 25^{th} percentile, Q1 –
  - 50^{th} percentile, Median or Q2 –
  - 75^{th} percentile, Q3 –

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

- IQR –

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.

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.

Example:

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

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-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 100Pth 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.

Example:

3. In a collection of 30 data measurements, which measurement represents the 30^{th} 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.

Example:

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

Using RStudio Packages:

RStudio data packages are pre-stored sets of data contained in the program. You can use them just like any list of data that you assigned. (The assign step is done for you!)

The data set LakeHuron gives the level (in feet) of LakeHuron at yearly levels from 1875 to 1972.

Find the mean, median, variance, and standard deviation of this data set.