Section 1.4
Range, IQR and Finding Outliers

In earlier sections we discussed measures of center (mean and median) and measures of spread (variance and standard deviation). In this section, we will introduce more measures of spread (range and interquartile range) as well as other measures of location (percentiles). Using this information we will be able to learn more about our data and will be able to identify outliers.

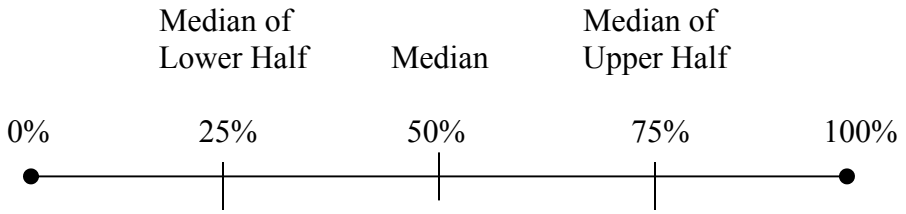
The **minimum** is the smallest data value. The **maximum** is the largest data value.

The **range** of a set of data is: Maximum – Minimum (Sensitive to outliers.)

Another measure of dispersion is called the **interquartile range**, or **IQR**. Before we can discuss how to determine the IQR we will need to understand some more measures of location, specifically **percentiles**.

The most common percentiles are the 25^{th} percentile, the 50^{th} percentile and the 75^{th} percentile. **Quartile 1 (Q1)** is the 25^{th} percentile of ordered data or median of lower half of ordered data. **Quartile 2 (Q2)**, is the median or 50^{th} percentile of ordered data. **Quartile 3 (Q3)** is the 75^{th} percentile of ordered data or median of upper half of ordered data. Note that half of the data will fall between Q1 and Q3.

Here is a quick picture:



The IQR = Q3 – Q1. The value Q3 – Q1 is also called the middle 50%.

The IQR is used to determine data classified as outliers. An **outlier** (extreme value) 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. We will learn how to calculate outliers later in this section. The mean, range, variance and standard deviation are sensitive to outliers, but IQR is not (it is resistant to outliers). The median and the mode are also not affected by extreme values in the data set.

Given a data set with positive values, if the largest data value is doubled the interquartile range WILL NOT increase, but range, mean, variance, and standard deviation WILL increase. If the smallest data value is divided by 2, IQR and mean decrease while the standard deviation and range increase.

The minimum, Q1, Q2, Q3, and the maximum makes up the **five number summary,** or **five statistical summary**.

The R command is: fivenum("name")

The results represent: MIN Q1 Q2 (median) Q3 MAX

Any point that falls outside the interval calculated by Q1 – 1.5(IQR) and Q3 + 1.5(IQR) is considered an outlier. Write outlier boundaries in interval notation, (lower bound, upper bound).

Example 1: Height measurements from a sample of adults were taken. The results (in inches) are: 66, 68, 63, 71, 67, 73, 70, 65, 69, and 68. Find the five number summary, the range and the IQR.

Commands:
heights=c(66,68,63,71,67,73,70,65,69,68)

Result of fivenum:

Min: Q1: Q2 (median): Q3: Max:

Range:

IQR:

Outlier Boundaries (write them in interval notation):
Q1 – 1.5(IQR)

Q3 + 1.5(IQR)

Boundaries?

Any outliers?

Other Percentiles

There are other percentiles as well. The **kth 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.

> *For example, suppose your Math SAT score is at the 80th percentile of all Math SAT scores. This means your score was higher than 80% of all other test takers.*

If you are looking for the measurement that has a desired percentile rank, the 100P^{th} percentile, this 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 and P is the percentile in decimal form.

Example 2: A cereal manufacturer claims that his cereal consists of 80 mg of sodium. To check his claim, we take a small sample from each box of cereal in a shipment and determine its sodium content. The results of 25 such measurements are as follows:

77, 81, 76, 76, 79, 79, 80, 77, 89, 77, 78, 85, 80, 75, 79, 88, 81, 78, 82, 80,
76, 83, 81, 85, 79

Before we begin we must first sort the list.
Commands:
cereal=c(77,81,76,76,79,79,80,77,89,77,78,85,80,75,79,88,81,78,82,80,76,83,81,85,79)

Result of the sort: 75, 76, 76, 76, 77, 77, 77, 78, 78, 79, 79, 79, 79, 80, 80, 80, 81, 81, 81, 82, 83, 85, 85, 88, 89

Determine the 96th percentile.
Recall: $nP + 0.5$
The 96th percentile will be located at position:

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 $[100(1-0.5)/n]^{th}$ percentile,
x_2 represents the $[100(2-0.5)/n]^{th}$ percentile,
…etc,
and x_i represents the $[100(i-0.5)/n]^{th}$ percentile.

Example 3: Use the data in Example 2 to find the percentile of the 4th order statistic (that is, the data value in the 4^{th} position of the list when the list is in order).

Example 4: Given a data set consisting of 33 unique whole number observations, its five-number summary is: [12, 24, 38, 51, 64].

a. What is the IQR?

b. How many observations are less than 38?