INFO • Median, Quartiles, Percentiles

Median, quartiles, and percentiles are statistical ways to summarize the location and spread of experimental data. They are a robust form of **data reduction**, where hundreds or thousands of data are represented by several summary statistics.

First, **sort** your data from the smallest to largest values. This is easy to do on a computer. Each data point now has a **rank** associated with it, such as 1st (smallest value), 2nd, 3rd, ... n^{th} (largest value). Let $x_{(r)}$ = the value of the r^{th} ranked data point.

The middle-ranked data point [i.e., at $r = (1/2)\cdot(n+1)$] is called the **median**, and the data value x of this middle data point is the median value ($q_{0.5}$). Namely,

$$q_{0.5} = x_{(1/2)\cdot(n+1)} \qquad \text{for } n = \text{odd}$$

If n is an even number, there is no data point exactly in the middle, so use the average of the 2 closest points:

$$q_{0.5} = 0.5\cdot [x_{(n/2)} + x_{(n/2)+1}] \qquad \text{for } n = \text{even}$$

The median is a measure of the **location** or center of the data

The data point with a rank closest to $r = (1/4)\cdot(n+1)$ is the **lower quartile** point:

$$q_{0.25} = x_{(1/4)\cdot(n+1)}$$

The data point with a rank closest to $r = (3/4)\cdot(n+1)$ is the **upper quartile** point:

$$q_{0.75} = x_{(3/4)\cdot(n+1)}$$

These last 2 equations work well if n is large (≥100, see below). The **interquartile range** (**IQR**) is defined as $IQR = q_{0.75} - q_{0.25}$, and is a measure of the **spread** of the data. (See the Sample Application nearby.)

Generically, the variable q_p represents any **quantile**, namely the value of the ranked data point having a value that exceeds portion p of all data points. We already looked at p = 1/4, 1/2, and 3/4. We could also divide large data sets into hundredths, giving **percentiles**. The lower quartile is the same as the 25th percentile, the median is the 50th percentile, and the upper quartile is the 75th percentile.

These **non-parametric statistics** are **robust** (usually give a reasonable answer regardless of the actual distribution of data) and **resistant** (are not overly influenced by **outlier** data points). For comparison, the mean and standard deviation are NOT robust nor resistant. Thus, for experimental data, you should use the median and IQR.

To find quartiles for a small data set, split the ranked data in half, and look at the lower and upper halves separately.

Lower half of data: If n = odd, consider the data points ranked less than or equal to the median point. For n = even, consider points with values less than the median value. For this subset of data, find its median, using the same tricks as in the previous paragraph. The resulting data point is the **lower quartile**.

Upper half of data: For n = odd, consider the data points ranked greater than or equal to the original median point. For n = even, use the points with values greater than the median value. The median point in this data subset gives the **upper quartile**.

Sample Application (§)

Suppose the z_{LCL} (km) values for 9 supercells (with EF0-EF1 tornadoes) are:

1.5, 0.8, 1.4, 1.8, 8.2, 1.0, 0.7, 0.5, 1.2

Find the median and interquartile range. Compare with the mean and standard deviation.

Find the Answer:

Given: data set listed above.

Find: $q_{0.5}$ = ? km, IQR = ? km,
$Mean_{zLCL}$ = ? km, σ_{zLCL} = ? km

First sort the data in ascending order:

Values (z_{LCL}):	0.5,	0.7,	0.8,	1.0,	1.2,	1.4,	1.5,	1.8,	8.2
Rank (r):	1	2	3	4	5	6	7	8	9
Middle:					^				

Thus, the median point is the 5th ranked point in the data set, and corresponding value of that data point is **median = $q_{0.5}$ = $z_{LCL(r=5)}$ = 1.2 km**.

Because this is a small data set, use the special method at the bottom of the INFO box to find the quartiles.

Lower half:

Values:	0.5,	0.7,	0.8,	1.0,	1.2
Subrank:	1	2	3	4	5
Middle:			^		

Thus, the lower quartile value is $q_{0.25}$ = 0.8 km

Upper half:

Values:	1.2,	1.4,	1.5,	1.8,	8.2
Subrank:	1	2	3	4	5
Middle:			^		

Thus, the upper quartile value is $q_{0.75}$ = 1.5 km

The IQR = $q_{0.75} - q_{0.25}$ = (1.5km – 0.8km) = **0.7 km**

Using a spreadsheet to find the mean and standard deviation:

$Mean_{zLCL}$ = **1.9 km** , σ_{zLCL} = **2.4 km**

Check: Values reasonable. Units OK.

Exposition: The original data set has one "wild" z_{LCL} value: 8.2 km. This is the **outlier**, because it lies so far from most of the other data points.

As a result, the mean value (1.9 km) is not representative of any of the data points; namely, the center of the majority of data points is not at 1.9 km. Thus, the mean is not robust. Also, if you were to remove that one outlier point, and recalculate the mean, you would get a significantly different value (1.11 km). Hence, the mean is not resistant. Similar problems occur with the standard deviation.

However, the median value (1.2 km) is nicely centered on the majority of points. Also, if you were to remove the one outlier point, the median value would change only slightly to a value of 1.1 km. Hence, it is robust and resistant. Similarly, the IQR is robust and resistant.