Many interesting ways to decide on Quartile values

In a recent blog posting on “The median and the interquartile range (IQR)”, we have discussed the concepts of quartile and the calculation of its interquartile range IQR which are easy to be apprehended.

For finding quartile values, we first arrange the data set in ascending order and then divide them into four parts, with each part containing *approximately* one-fourth, or 25% of the observations. The division points are referred to as the quartiles and are defined as:

Q1 = first quartile, or 25th percentile (sometimes also called lower quartile)
Q2 = second quartile, or 50th percentile (it is also the median)
Q3 = third quartile, or 75% percentile (sometimes also called upper quartile)

Confusion starts to rein in when we try to decide the cut-off values on Q1 and Q3. There is no argument on picking up the Q2 value as it is known to be the median of the data set. When you have 5 data in ascending order, the third value is the median and when you have 6 data set in ascending order, the average figure of 3rd and 4th value is the median.

Upon scouting around text books and software on statistics, it is amazing to find many different conventions and methods to decide on Q1 and Q3 values. Different methodologies produce may slightly varied quartile results and they may affect the final outcome on the data analysis.

So, how then do we choose the “best” value for these quartiles? It is interesting to note from a research paper which says: *It depends in part on the statistician’s fancy in finding quartile values!*

Let’s see how some of these methods work and I shall leave it to you to decide on the best method that suits your objective in data analysis.

1. An easiest and straightforward arithmetic method is to pick up the Q1 and Q3 values directly from the rearranged ascending order of the data using a general equation for the value at position i as below:

$$i = \left(\frac{p}{100}\right)n$$

where p = the percentile and n = the number of observations

There are 2 situations to be considered:

(a) If i is not an integer, we have to round it up and move to the next integer greater than the position obtained. For example, when p = 25% and n=9, then i = 2.25 and the position of the 25% percentile (Q1) is the next integer higher than 2.25, the 3rd position. Hence, we have:

when n = 9

Quartile	Actual	i^{th} position
Q1	2.25	3
Q2	4.5	5
Q3	6.74	7

(b) If i is an integer, the p^{th} percentile is the average of the values in position i and i +1.

For example: if we have 12 data in a set as below:

{ 272, 276, 279, 283, 287, 289, 290, 297, 301, 303, 304, 310}

The Q1 is at 3rd, Q2, at 6th and Q3, at 9th positions, respectively.

In this case, Q1 (25th percentile) is the average of the 3rd and 4th data values, thus, Q1 = (279+283)/2 = 281. Similarly, Q3 (75th percentile) is the average of 9th and 10th data values, giving Q3 = ((303-301)/2 = 302.

Of course, the median at 6th position is to have Q2 = (290-289)/2 = 289.5.

2. The Tukey's method for finding the quartile values is to find the median value first. For example in a data set of 9 (odd) values:

{23, 28, 29, 32, 35, 37, 41, 44, 47}

it is obvious that the value 35 is the median at Q2 (50th percentile). To find the Q1 and Q3, we sub-set the data set into:

{23, 28, 29, 32, 35} and {35, 37, 41, 44, 47}

Again, it is easy for us to tell that Q1 = 29 and Q3 = 41.

3. Moore and McCarby (sometimes referred to "-M and M-") method is similar to Turkey's but *do not* include the median M value in either half of the subsets when finding the quartile values.

So, looking back at the above example, we only have to consider 4 values in each subset without involving the median:

{23, 28, 29, 32}   and   {37, 41, 44, 47}

Since each of these data sets has an even number of elements (4), we average the middle two values, giving the Q1 value as (28+29)/2 = 28.5, and the Q2 value as (41+44)/2 = 42.5.

4. Mendenhall and Sincich suggest another way of finding the quartile values. When we have a data set with n observations, we find the lower quartile position L^{th}, being Q1, rounding to the nearest integer, as follows:

$$L = \frac{1}{4}(n+1)$$

If L falls halfway between two integers, round *up* to its larger figure.

For the upper quartile (Q3) value, we use the following equation, rounding U^{th} to the nearest integer:

$$U = \frac{3}{4}(n+1)$$

If U falls halfway between two integers, round *down* to its lower figure.

Hence, for our example data set with 9 elements :

{23, 28, 29, 32, 35, 37, 41, 44, 47}

We have: L = (1/4)(9+1) = 2.5 ~ 3 after rounding up
and U = (3/4)(9+1) = 7.5 ~ 7 after rounding down

Hence, Q1 = 29 and Q3 = 41.

The statistical software Minitab® adopts the same method but does not round up the L and U values. Instead, it uses linear interpolation between the two closest data points. When L = 2.5 in the above example, the software calculates the Q1 value as the middle value of 2nd and 3rd of the data set in ascending order.

5. The popular MS Excel® spreadsheet software uses a method described by Freund and Perles. It has been noted that almost no one else uses such approach but Excel. Many statistical software such as XLSTAT® using Excel® as the base for data analysis adopt the same approach.

The equations used for L and U defined as above are:

$$L = \frac{1}{4}(n+3)$$

$$U = \frac{1}{4}(3n+1)$$

If L or U is not an integer, Excel® uses linear interpolation to calculate the quartile values. For example, if a L value is equal to 3.25, that means the quartile value must be between the 3rd and 4th value of the dataset and calculated by adding the 3rd data to 0.25 of the difference between the 3rd and 4th data in the array.

Hence, for our example data set with 9 observations :

{23, 28, 29, 32, 35, 37, 41, 44, 47}

We have: L = (1/4)(9+3) = 3
and U = (1/4)(3x9+1) = 7

Hence, Q1 = 29 and Q3 = 41.

In summary,

Quartile	Tukey	Moore, et al	Mendenhall, et al	Excel
Q1	29	28.5	29	29
Q3	41	42.5	41	41

It can be shown that when there are even number of elements in a dataset, the differences in the quartile results tend to be more prominent amongst the methods because of the rounding up or rounding down requirements and linear interpolation by some of these methods. However, although the actual values reported for quartiles may vary slightly depending on the convention used, the basic objective for computing quartiles by dividing the data into four equal parts remains the same.