Quartile Points

The mean, median, and mode of a data list give us some ideas of what we could define as the “center” of the data. Now, we want to look at how we can get an idea of the “spread” of the data. That is, we want some way to describe how the data is arranged **around** these measures of the “center”. One way to look at the “spread” of the data is to simply look at the lowest and highest values of the set, the “range.” Expanding on that a bit, we can also look at the quartile points of the data.

Range: The range of the data is just the low and the high values of the data. Naturally, to do this there must be some order to the data. For the list of values, which we will call A, A=[4, 8, 6, 13, 8, 2, 8], we can see that the range gives the low value as 2 and the high value as 13. We might recall that the mean of these values is 7, the median is 8, and the mode is 8. Here is another list of data B=[1,8,2,20,8,2,8]. Again, the mean is 7, the median is 8, and the mode is 8, but now the range is 1 to 20. Knowing the range gives an idea that this data could be more spread out than was the data in list A. Or, for the list C=[-44,8,-25,100,8,-6,8], which still has the mean=7, the median=8, and the mode=8, the range is now -44 to 100, giving us an idea of the spread of the data. The range is highly influenced by extreme values, it’s actually DEFINED by them. At best it gives us an idea of the possible values in the data list.

Let us look at a much larger data list. In the discussion of the median we found out that the range of the age of the 12,608 students in a particular semester at the college was 13 to 84.

Quartile Points: While the range gives us the highest and lowest values, it still leaves us guessing about how the values are spread within that range. Quartile points help give us a better idea of the way the values are spread. We will start with three quartile points Q_1, Q_2, and Q_3. Q_2 is the median value. Remember that the median, now known as Q_2, is the value that has half the other values below it and half above it in a sorted list of the values. We will choose Q_1 to be the value in the sorted list that has a quarter of the values below it and three quarters of the values above it. In essence, Q_1 is the "median" of the sub-list of values that are less than Q_2. In the same way, the third quartile, Q_3, is the "median" of the sub-list of values that are greater than Q_2. Thus, Q_3 is the value in the sorted list that has ¾ of the other values below it and ¼ of the values above it.

Let us look at an example. Consider the sorted data list D=[2, 2, 4, 5, 6, 7, 7, 9, 11, 14, 16, 16, 21, 23, 24, 26, 28, 28, 29]. There are 19 values in this list so Q_2 is the tenth item, or 14. That leaves us with 9 items less than Q_2 and the median of those 9 items will be the fifth item, namely $Q_1=6$. The median of the 9 items above Q_2 will be the fifth item in that sub-list, namely, $Q_3=24$. In general, Q_1 has a quarter of the values less than it and three quarters of the values greater than it. Q_2 has half the values less than it and half the values greater than it. Q_3 has three quarters of the values less than it and one quarter of the values greater than it.

Now that the concept of the quartile points makes some sense, let us recall that we had a special case computing the median. In particular, if we had an even number of items in the list then the median became the average of the two middle items. If we have a list E that is the same as list D but with the 29 removed, then list E would have 18 elements and the median, Q_2, would be the average of the 9th and 10th items, or (11+14)/2=25/2=12.5. Then we would still have 9 items below the median and 9 above so Q_1=6. However, even though there are 9 elements above the median, the median is now 12.5 and the first element greater than the median is item 10. Therefore, the fifth item in the sub-list of items greater than Q_2 is now Q_3=23. But, what if we had list F that is the same as list D but with the value 31 added to the list. Then list F has 20 elements and the median would be the average of the two middle elements, namely Q_2=(14+16)/2=30/2=15. But that leaves us with an even number of values below Q_2 and an even number above Q_2. Now, Q_1 will be computed as the average of the two middle elements in the lower half of the list, namely, Q_1=(6+7)/2=13/2=6.5. In a similar fashion, Q_3=(24+26)/2=50/2=25.

Having explained all of that we need to add that the method described here is not the only method of calculating the quartile points. There are at least three different methods, the one described here being the easiest to implement. The complexity of the calculation masks the fact that in the real world, for a large data set, the methods will all amount to the same thing. For example, in our data list of 12,608 students we know that the median age, 23, was buried in the pile of 683 students of age 23. If we compute Q_1, using our methodology, we need to find the median of the student ages less than that median value, that is, the median of the 6304 youngest students. That would make Q1 equal to the average of the ages in positions 3152 and 3153 of the sorted list of student ages. That means that Q_1 is buried in the 1333 students age 19 whose ages occupy positions 1832 to 3164 of the sorted list. The average of 19 and 19 is just 19, so Q_1=19. In the same way, Q_3 would be the average of items 9456 and 9457, but those two values are buried in the 254 students of age 31 whose ages occupy positions 9304 to 9557 of the sorted list. The average of 31 and 31 is just 31 so Q_3=31.

We can slightly expand the quartile system by introducing Q_0 and Q_4. We define Q_0 as the lowest value and Q_4 is defined as the highest value. Knowing these quartile values tells us a great deal about the "spread" of the values in the data list. Naturally, this is more useful in a real data set. For our student age example we have Q_0=13, Q_1=19, Q_2=23, Q_3=31, and Q_4=82. From this we get a good idea that we have a huge concentration of student ages between 19 and 23, a bit more spread out ages both from 13 to 19 and from 23 to 31, and a more disbursed list of ages between 31 and 82. We should note that what we have is a "feeling" about this, not a certainty. The concept of quartiles gives us the feeling whereas a compilation of the data, perhaps a table giving each age and the number of students at that age, would tell us the reality.