Percentiles

One way to look at quartile points is to say that, for a sorted list of values, Q_1 is the value that has 25% of the rest of the values that are less than it, Q_2 is the value that has 50% of the values that are less than it, and Q_3 is the value that has 75% of the values that are less than it. Along with Q_0, the lowest value, and Q_4, the highest value, these quartile points give us an idea of, a feeling for, the spread of the values. It is reasonable to ask if there is something special about using these points, 25%, 50%, and 75%, rather than some other points. Why not use 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90%? The answer is that with just the few quartile points we get a good feel for the spread with just a few values, whereas, with the nine points just listed, although we would get a much better idea of the distribution of values, we would have to find and consider many more points.

Just how hard would it be to find such points? We have dealt with the case of a list of ages for the 12,608 students at the college in a particular term. If we have those ages in a sorted list, then to find the 10% point we need to find the 1260.8th item, starting from the lowest value. There is no 1260.8th item but there is the 1260th item, and the 1261st item. The 1260th item actually has less than 10% of the values before it in the sorted list, even if we include the 1260th item in that sub-list. There are 1260 such values and 1260/12608=9. 99365% . The 1261st item certainly has at least 10% of the values before it in the list (as long as we include the 1261st item in that sub-list): there are 1261 such values and 1261/12608=10.0016%. Since we want to identify the first point in the list of sorted values that has 10% of the values less than or equal to it, we will choose the 1261st item. We see that we can get the value 1261 by taking 10% of the number of items, in this case that is 10% of 12608, giving 1260.8, and rounding up to the next higher whole number. It is nice to have such a rule, but as we have seen before, in a large data list either of our two proposed items would be the same value. In fact, for the case at hand, items 1260 and 1261 are both instances of age 18, an age that fills positions 640 to 1831 in the sorted list.

The rule is to take the desired percentage of the total number of items and then round up to the next higher whole number. Doing that for the rest of the percentages that we have gives:

Percent	Position expression	Position value	Rounded position	Value at the position
10%	10% of 12608	1260.8	1261	18
20%	20% of 12608	2521.6	2522	19
30%	30% of 12608	3782.4	3783	20
40%	40% of 12608	5043.2	5044	21
50%	50% of 12608	6304	6304	23
60%	60% of 12608	7564.8	7565	26
70%	70% of 12608	8525.6	8526	28
80%	80% of 12608	10086.4	10087	34
90%	90% of 12608	11347.2	11348	44

Go back to the 50% value. We know this is the median. With an even number of items we would compute the median as the average of item 6304 and 6305. The computation for the 50%

point gave us 6304 which we decided to leave as that instead of rounding up to the next higher whole number, 6305. If we strictly apply the rule we should have rounded up, but, as usual, with a large data list it would not have made any difference given that both the 6304th and 6305th items are age 23.

There are two more issue; we will define the 0th percent point as the lowest value and we will define the 100th percent point as the highest value. The "rule" works for the 0% point since 0% of 12608 is 0 and we could round that up to 1 and the 1st item in the sorted list is the lowest value. The "rule" does not work for the 100% point since 100% of 12608 is 12608 and if we were to round that up to 12609 we would be in trouble since there is no 12609th value.

Getting the eleven values that we just found, from 0% to 100% in steps of 10%, is nice, and it gives us a better feel for the distribution of ages in the data list, but it does so at the cost of having many more values to examine than we had with the simple quartile experience.

All of this is nice but it merely sets the stage for what we really want, namely percentiles. To do percentiles we simply expand our process to go from 0% to 100% but this time to do so in steps of 1%. We follow the same rule. To find the 73%tile we find 73% of 12608, namely 9203.84, round that up to the next higher whole number, 9204, and look at the value in position 9204 in the sorted list, for us that will be age 30. Our statement is that the 73rd percentile of the list is age 30, or, more commonly, 30 is the 73rd percentile of the list. You might note that this does not really mean that there are 73% of the values that are less than age 30. In fact, there are only 9042 age items less than 30 and 9024/12608=71.72%. However, there are 9303 values that are age 30 or less, representing 73.79% of all the values, so our 73rd percentile point is one of the 271 ages that are all 30. In common usage, we would still say that 30 is the 73rd percentile, and we would interpret that to mean that 73% of the values are less than 30. Clearly this is wrong, but it is what is generally done.

What if we want to go in the other direction? Let us say that we have the set of values, that we are looking at one of those values and we want to assign a percentile to that value. In general, percentiles are given to whole number values. If we are looking at age 37 in our large list example, then we find that age 37 occupies positions 10448 through 10586, representing percentages 10448/12608=82.868% to 10586/12608=83.963%. It seems that it is safe to say that 37 is in the 83rd percentile. On the other hand, what if we want to give a percentile ranking to age 25? Age 25 holds the positions 7286 through 7742 in our sorted list. This corresponds to percentages of 7286/12608=57.789% to 7742/12608=61.405%. That is to say, we have age 25 items that are at the 58th percentile, 59th percentile, 60th percentile, and even 61st percentile. What should we say about age 25 in general? The safest thing to say is that 25 occupies all four percentiles. There is no standard on what to say beyond that.

We saw that quartiles give us a feel for the data. Expanding the number of points (originally we went from 3 to 9) gave us a better feel, but at the expense of having to consider many more points. Expanding the number of points to percentages (now we are up to 101 points if we include the endpoints) gives us a really good feel for the data, but that is just too many points to “hold” in our mind at once. Percentiles, however, when given for a specific value, such as 37 is in the 83rd percentile, gives a feel for where that specific value resides in the ordered list of all values.