The Halving Method for Sample Quartiles[*]

ANWAR H. JOARDER

Dept of Mathematical Sciences, King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia 31261, Email: anwarj@kfupm.edu.sa

An attempt is made to put the notion of sample quartiles on a mathematical footing in the light of ranks of observations, and equisegmentation property that the number of ranks below that of the first quartile, that between the consecutive quartiles, and that above the third quartile are the same. Ranks of sample quartiles provided by the proposed Halving Method, based on hinges, does satisfy the property.

1. Introduction

There are many methods available for calculating sample quartiles in different elementary text books on statistics without any explanation. The most popular one, called Popular Method hereinafter, is described here. The rank of the $i(i=1,2,3)$ th quartile is given by

$$i\ (n+1)/4 = l + d, \quad \mathrm{i} = 1, 2, 3 \tag{1.1}$$

where l is the largest integer not exceeding $i(n+1)/4$. Then the Popular Method uses the following linear interpolation formula for the calculation of sample quartiles

$$Q_i = x_{(l)} + d\,(x_{(l+1)} - x_{(l)}) = (1-d)\,x_{(l)} + d\,x_{(l+1)}, \quad (\mathrm{i} = 1, 2, 3), \tag{1.2}$$

where $x_{(l)}$ is the l-th ordered observation (Ostle, Turner, Hicks and McElrath, 1996, 38).

However, students and instructors alike are curious to know why the formulae for quartiles in (1.1) contain the quantity $n+1$. Why not n or $n-1$? Though the formulae for the median in the literature appear to be different, they all are equivalent. It is given by $Q_2 = (n+1)/2$ th observation. In case n is odd, $(n+1)/2$ will be an integer so that the median will be an observation with integer rank. If however, n is even, $(n+1)/2$ will lie between $n/2$ and $n/2+1$. Then the median can be calculated by the use of linear interpolation. Because of the success of the quantity $(n+1)$ in equation (1.1) to find the median, the idea of proportional weight given by (1.1) or (1.2) has possibly been popular to find other quartiles by the above method.

[*] To be Published in *International Journal of Mathematical Education in Science and Education*, 2003, London, UK]

It would be clear down the road that $n+1$ is the total of the ranks for the largest and smallest observations in the sample, and that the rank of the median is the average of the ranks of the observations.

To write out the ranks exhaustively let us denote the sample size by the following remainder-modulus representation

$$n = r \bmod 4 = 4m + r,\ (r = 0,1,2,3), \qquad (1.3)$$

so that the number of observations in each of the $4 \le n$ segments is given by $m = (n-r)/4$. With this representation of the sample size the ranks and quartiles of a sample will be denoted respectively by R_{ir} and Q_{ir} ; $i = 1,2,3;\ r = 0,1,2,3$. Though quartiles Q_{ir} ; $i = 1,2,3;\ r = 0,1,2,3$ are usually denoted by Q_i ; $i = 1,2,3$, we will not suppress r as it plays an important role in the proposed Halving Method for quartiles. The ranks in (1.1) given by the Popular Method can be rewritten as

$$R_{ir} = i(4m + r + 1)/4 = im + i(r+1)/4,\ i = 1,2,3;\ r = 0,1,2,3 \qquad (1.4)$$

which is the the rank of the i th quartile corresponding to the sample size with remainder r. Then the ranks for sample quartiles provided by the Popular Method can be written out exhaustively as:

$$R_{10} = m + 1/4,\ R_{20} = 2m + 2/4,\ R_{30} = 3m + 3/4$$
$$R_{11} = m + 2/4,\ R_{21} = 2m + 1,\ R_{31} = 3m + 1 + 2/4$$
$$R_{12} = m + 3/4,\ R_{22} = 2m + 1 + 2/4,\ R_{32} = 3m + 2 + 1/4$$
$$R_{13} = m + 1,\ R_{23} = 2m + 2,\ R_{33} = 3m + 3$$

We propose the new criterion of equisegmentation property that the number of ranks below that of the first quartile, that between the consecutive quartiles, and that above the third quartile are the same. However this will divide the ordered sample observations into four segments leaving the same number of observations in each if all the observations are distinct. Let the number of integers in each segment be m_i $(i = 1,2,3,4)$. Then the equisegmentation property guarantees that $m_1 = m_2 = m_3 = m_4$. In case $1 \le n \le 3$, the above formulae can also be used to calculate quartiles with $m = 0$.

It is interesting to note that though the Popular Method is not based on good mathematical reasoning, the equisegmentation property is satisfied by the quartiles provided by this method for all sample sizes except for $n = 4m + r$, $m \ge 1,\ r = 2$. For $r = 2$, the number of observations in four segemnts are $m, (m+1), (m+1)$ and m respectively.

Thus it is essential to modify the formulae of ranks so that the equisegmentation property is satisfied by quartiles provided by the Popular Method for any sample size. It is observed that, whenever $n = 4m + 2$, simple arithmetic rounding of ranks provided by this method would satisfy the equisegmentation property.

The Halving Method discussed in this paper demonstrates in an accessible way that the set of formulae for quartiles offered by this method is based on good mathematical reasoning. The ranks provided by the Halving Method written out exhaustively by the remainder-modulus representation of the sample size help prove that the corresponding quartiles satisfy the equisegmentation property. Moreover, ranks provided by the Halving Method guarantee that the remainder r of the sample size is the number of quartiles having integer ranks. Linear interpolation should be used to find quartiles with noninteger ranks.

2. The Halving Method for Sample Quartiles

The method, developed in the spirit of Tukey (1977, p32-35), is based on hinges which finds the median first, and then finding the medians of upper and lower halves of the data. Usually median is included in both halves while calculating the hinges. But we observe that if median of the whole data set is ignored in the calculation of hinges, then the two extreme hinges and median enjoy equisegmentation property. We develop algebraic expressions for ranks of quartiles based on this argument and call this method the Halving Method. It thus resolves the difference between quartiles and hinges. The ranks for the quartiles given by the Halving Method are developed below in terms of r and m where the sample size $n = 4m + r,\ (r = 0,1,2,3)$:

(a) Ranks of quartiles for $n = 4m$

The observations have ranks $1, 2, ..., 2m, 2m+1, ..., 4m$. The rank of the median is

$$R_{20} = \frac{1}{4m}(1 + 2 + \cdots + 4m) = \frac{1}{4m}\frac{4m(1+4m)}{2} = \frac{1+4m}{2} = 2m + 0.5$$

which is between $2m$ and $2m+1$ so that the ranks of extreme quartiles are given by

$$R_{10} = \frac{1+2m}{2} = m + 0.5,\ R_{30} = \frac{(2m+1)+4m}{2} = 3m + 0.5\,.$$

It is worth mentioning that in this case none of the quartiles has integer ranks.

(b) Ranks of quartiles for $n = 4m + 1$

The observations have ranks $1, 2, ..., 2m, 2m+1, 2m+2 ..., 4m+1$. The rank of the median is

$$R_{21} = \frac{1+(4m+1)}{2} = 2m + 1$$

which is between $2m$ and $2m+2$ so that the ranks of extreme quartiles are given by

$$R_{11} = \frac{1+2m}{2} = m+0.5,\ R_{31} = \frac{(2m+2)+(4m+1)}{2} = 3m+1.5\,.$$

It is worth mentioning that in this case the median has an integer rank.

(c) Ranks of quartiles for $n = 4m+2$

The observations have ranks $1, 2, ..., 2m, 2m+1, 2m+2 ..., 4m+2$. The rank of the median is

$$R_{22} = \frac{1+(4m+2)}{2} = 2m+1.5$$

which is between $2m+1$ and $2m+2$ so that the ranks of extreme quartiles

$$R_{12} = \frac{1+(2m+1)}{2} = m+1,\ R_{32} = \frac{(2m+2)+(4m+2)}{2} = 3m+2\,.$$

It is worth mentioning that in this case the extreme quartiles have integer ranks.

(d) Ranks of quartiles for $n = 4m+3$

The observations have ranks $1, 2, ..., 2m, 2m+1, 2m+2, 2m+3, ..., 4m+3$. The rank of the median is

$$R_{23} = \frac{1+(4m+3)}{2} = 2m+2$$

which is between $2m+1$ and $2m+2$ so that the ranks of extreme quartiles are

$$R_{13} = \frac{1+(2m+1)}{2} = m+1,\ R_{33} = \frac{(2m+3)+(4m+3)}{2} = 3m+3\,.$$

In practice one may simply use the above argument to calculate ranks of quartiles. The other alternative is to find r and $m = (n-r)/4$ and then use the ranks of quartiles given below to calculate quartiles.

$$R_{10} = m+2/4,\ R_{20} = 2m+2/4,\ R_{30} = 3m+2/4$$
$$R_{11} = m+2/4,\ R_{21} = 2m+1,\ R_{31} = 3m+1+2/4$$
$$R_{12} = m+1,\ R_{22} = 2m+1+2/4,\ R_{32} = 3m+2$$
$$R_{13} = m+1,\ R_{23} = 2m+2,\ R_{33} = 3m+3$$

The rank of the median, in Popular Method as well as in Halving Method, is the average of the first and third quartiles. The remainder r here is also the number of quartiles having integer ranks in the Halving Method but not in the Popular Method. It is easy to check that equisegmentation property is satisfied by the quartiles offered by the Halving Method for $n = 4m + r,\ r = 0,1,2,3$ i.e. for all sample sizes. The explicit form of the ranks of quartiles by the two methods help us compare them. In fact each of the rank R_{10} , R_{30}, R_{12}, R_{32} given by the Popular Method differs from that given by the Halving Method by $1/4$.

We recommend to use the Halving Method as it is based on logic. The generalization of the method to deciles, percentiles or to any quantiles, in general, remains open.

3. An Illustration

The following ten value are sample weights (in grams) of coating materials used in a masking process:

5.3 5.4 5.7 6.0 6.1 6.2 6.3 6.4 6.5 6.6

(i) Calculation of Quartiles by Popular Method

Here the sample size $n = 10 = 4(2) + 2$ so that $m = 2$ and $r = 2$. Since $r = 2$ we will denote the ranks of quartiles by $R_{i2}\ (i = 1, 2, 3)$. The rank of the quartiles provided by the Popular Method are (see equation 1.1)

$$R_{12} = (n+1)/4 = 2.75,\ R_{22} = (n+1)/2 = 5.5,\ R_{32} = 3(n+1)/4 = 8.25$$

which can also be written equivalently as

$$R_{12} = m + 3/4 = 2.75,\ R_{22} = 2m + 1 + 2/4 = 5.5,\ R_{32} = 3m + 2 + 1/4 = 8.25$$

(see equation 1.4). Note that the consecutive ranks are apart by 3, and there are 2 ranks below R_{12} or above R_{32} . Then by linear interpolation (see equation 1.2) the quartiles are given by

$$Q_{12} = x_{(2.75)} = (1-0.75)\ x_{(2)} + 0.75\ x_{(3)} = 0.25(5.4) + 0.75(5.7) \approx 5.625$$
$$Q_{22} = x_{(5.5)} = (1-0.5)\ x_{(5)} + 0.5\ x_{(6)} = 0.5(6.1) + 0.5(6.2) = 6.15$$
$$Q_{32} = x_{(8.25)} = (1-0.25)\ x_{(8)} + 0.25\ x_{(9)} = 0.75(6.4) + 0.25(6.5.) \approx 6.425$$

To check the equisegmentation property, we show the ranks of the quartiles by downward arrows in the sample:

5.3 5.4 ↓ 5.7 6.0 6.1 ↓ 6.2 6.3 6.4 ↓ 6.5 6.6

We observe that there are $2\,(=m)$, $3(=m+1)$, $3(=m+1)$ and $2(=m)$ observations in the four segments i.e. the ranks of the quartiles do not satisfy the equisegmentation property.

(ii) *Calculation of Quartiles by Halving Method*

Instead of using the formulae provided by the Halving Method at the end of section 2, we prefer to use the idea of halving to find quartiles with the hope that it would provide more insight into the problem.

The rank of the median is $R_{22}=\dfrac{1+n}{2}=5.5$ so that $Q_{22}=x_{(5.5)}=(1-0.5)\,x_{(5)}+0.5\,x_{(6)}=0.5(6.1)+0.5(6.2)=6.15$. The first quartile is the median of the observations below the median of the whole data set i.e. is $R_{12}=\dfrac{1+5}{2}=3$ so that $Q_{12}=x_{(3)}=5.7$. The third quartile is the median of the observations above the median of the whole data set i.e. is $R_{32}=\dfrac{6+10}{2}=8$ so that $Q_{32}=x_{(8)}=6.4$.

To check the equisegmentation property, we show the ranks of the quartiles by downward arrows in the sample:

$$5.3\quad 5.4\quad \overset{\downarrow}{5.7}\quad 6.0\quad 6.1\quad \overset{\downarrow}{}\quad 6.2\quad 6.3\quad \overset{\downarrow}{6.4}\quad 6.5\quad 6.6$$

We observe that there are $2(=m)$ observations in each of the four segments i.e. the ranks of the quartiles do satisfy the equisegmentation property. The author also thanks an anonymous referee for constructive suggestions that have improved the readability and presentation of an earlier draft of the paper.

Acknowledgements

The author is grateful to the excellent research facilities available at King Fahd University of Petroleum and Minerals, Saudi Arabia. The author is also thankful to an anonymous referee for constructive suggestions on an earlier draft that have improved the readability and presentation of the paper.

References

Ostle, B., Turner, K.V. Hicks, C.R. and McElrath, G.W. (1996). *Engineering Statistics*: The Industrial Experience. New York: Duxbury Press.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison Wesley.

File: p96a.doc