

Quartiles for Discrete Data

KEYWORDS:

Teaching;
Remainder;
Modulus.

A.H. Joarder and M. Firozzaman

King Fahd University of Petroleum and Minerals,
Dhahran, Saudi Arabia.

e-mail: anwarj@kfupm.edu.sa
fzaman@bdonline.com

Summary

Different methods for finding quartiles for discrete data that are available in the literature are examined in the light of a desirable property that quartiles divide the ordered sample observations into four segments having the same number of observations in each segment. A set of formulae that satisfy this property is proposed and illustrated with examples.

◆ INTRODUCTION ◆

The formulae for quartiles, or more generally percentiles, are uniquely defined for continuous random variables. Let X be a continuous random variable with probability density function $f(x)$ and cumulative distribution function $F(x)$. Then the quartiles Q_1 , Q_2 and Q_3 are defined by $F(Q_1) = \frac{1}{4}$, $F(Q_2) = \frac{1}{2}$ and $F(Q_3) = \frac{3}{4}$ respectively.

In the discrete case, however, it is difficult to define quartiles. The most popular idea is based on finding the median first and then finding the medians of the upper and lower halves of the data. In doing so, roughly 25% of the observations remain below the lower quartile and 25% above the upper quartile. There is, however, considerable confusion about the details, and many definitions are offered for the lower and the upper quartiles. Further, different definitions are used by some of the popular statistical computing packages.

An alternative method of dividing ordered sample observations into approximately equal quarters employs ‘hinges’. The lower hinge is the median of that portion of the data which is less than or equal to the median of the whole data set. The upper hinge is the median of that portion of the data which is greater than or equal to the median of the whole data set. See Tukey (1977, p. 34) or Mayer and Sykes (1996, p. 25) for an illustration.

In this article we impose the following restriction on the quartiles for discrete data:

Property I: Quartiles divide the ordered sample observations into four quarters having the same number of observations (m) in each quarter.

The remainder of this article depends on this property and on the notion that it is desirable and it should hold. Different methods available in the literature have been examined in the light of the property. We propose a set of refined rules for quartiles that enjoys the property. The set of rules can be considered as a precise formulation for ‘hinges’. If the median of the whole data set is ignored in the calculation of hinges, then the two hinges and median enjoy Property I. It thus resolves the difference between quartiles and hinges.

◆ A REVIEW OF WELL-KNOWN ◆ FORMULAE FOR QUARTILES

In this section we survey the formulae for extreme quartiles (i.e. the lower and upper quartiles, Q_1 and Q_3) available in the literature.

Method 1

The oldest and most popular method to find extreme quartiles is based on finding the median first and then finding the medians of the upper and lower halves of the data. Some authors count the median in both halves if the size of the sample is odd. As mentioned earlier, Tukey (1977) then called the quartiles ‘hinges’. We observe that Method 1 guarantees Property I if the median is

always ignored in the calculation of the lower and upper quartiles.

Method 2

Another popular method to find extreme quartiles is the following: $Q_1 = (n+1)/4$ th observation, $Q_3 = 3(n+1)/4$ th observation. (See, for example, Mendenhall and Sincich 1995, p. 54.)

It is interesting to note that Q_1 and Q_3 will have integer ranks for sample sizes of the form $n = 4i + 3$ ($i = 1, 2, \dots$), and the above formulae satisfy Property I.

Quartiles Q_1 and Q_3 will have non-integer ranks (positions) for other sample sizes. It is common practice to make arithmetic rounding in these cases. We observe that it is not arithmetic rounding but rather linear interpolation that ensures Property I for sample sizes $n = 4i, 4i + 1$ ($i = 1, 2, \dots$). Arithmetic rounding is only needed for $n = 4i + 2$ ($i = 1, 2, \dots$) to ensure Property I.

Method 3

The formulae given by Vinning (1998, p. 44) can be simplified as

$$Q_1 = \begin{cases} (n+3)/4\text{th observation if } n \text{ is odd} \\ (n+2)/4\text{th observation if } n \text{ is even} \end{cases}$$

$$Q_3 = \begin{cases} (3n+1)/4\text{th observation if } n \text{ is odd} \\ (3n+2)/4\text{th observation if } n \text{ is even} \end{cases}$$

The example Vinning provides with $n = 35$ divides the observations into four segments with 9, 8, 8 and 9 observations in them. The median has an integer rank, namely the 18th position. It is easy to check that the above formulae divide ordered sample observations into four quarters only if n is even.

Method 4

The formulae provided for percentiles by Smith (1997, p. 36) can be specialized for quartiles as

$$Q_1 = \begin{cases} (n+2)/4\text{th observation if } n/4 \text{ is not an integer} \\ [(n/4\text{th} + \text{the next}) \text{ observation}]/2 \text{ if } n/4 \text{ is an integer} \end{cases}$$

$$Q_3 = \begin{cases} (3n+2)/4\text{th observation if } 3n/4 \text{ is not an integer} \\ [(3n/4\text{th} + \text{the next}) \text{ observation}]/2 \text{ if } 3n/4 \text{ is an integer} \end{cases}$$

Smith suggests rounding the ranks to the nearest integer. The example provided for $n = 12$ does satisfy Property I with $m = 3$. In fact, the above formulae satisfy Property I for any sample size except for $n = 4i + 1$ ($i = 1, 2, \dots$).

Method 5

A popular method for finding quartiles is linear interpolation; see, for example, Lapin (1997, pp. 45–6) and Ostle *et al.* (1996, p. 38). The quartile Q_α ($\alpha = 1, 2, 3$) is usually interpolated by the following steps.

- Step 1. Determine $R_\alpha = \alpha(n+1)/4$, $\alpha = 1, 2, 3$.
- Step 2. Separate the integer (i) and decimal (d) parts of R_α and write $R_\alpha = i + d$.
- Step 3. The quartile is finally given by

$$Q_\alpha = x_{(i)} + d(x_{(i+1)} - x_{(i)}) = (1-d)x_{(i)} + dx_{(i+1)} \quad (\alpha = 1, 2, 3)$$

where $x_{(i)}$ is the i th ordered observation.

The above formulae are given in different forms in numerous books and satisfy Property I for any sample size except $n = 4i + 2$ ($i = 1, 2, \dots$).

◆ THE PROPOSED FORMULAE ◆ FOR QUARTILES

Although the formulae for the median in the literature appear to be different, they are all equivalent. The median is given by $Q_2 = (n+1)/2$ th observation, where n is the sample size. In the case where 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 using linear interpolation the median is given by $Q_2 = [(n/2\text{th} + \text{the next}) \text{ observation}]/2$.

Now define r by $n \equiv r \pmod{4}$ (so that r must be 0, 1, 2 or 3). It is not surprising that the formulae for quartiles for different sample sizes with the same r are all the same. We need to consider the four different sample sizes in terms of r , namely $n = 4m + r$ ($r = 0, 1, 2, 3$). If the sample size is $n = 4m$ (i.e. $r = 0$), none of the quartiles have integer ranks. If the sample size is $n = 4m + 1$ (i.e. $r = 1$), one of the quartiles has an integer rank. Since n here is odd, the integer rank must be for the median. If the sample size is $n = 4m + 2$, two of the quartiles have integer ranks. Since n here is even, the median does not have an integer rank and consequently the two extreme quartiles have integer ranks. If the sample size is $n = 4m + 3$ (i.e. $r = 3$), all three quartiles have integer ranks.

In practice we need to divide the sample size by 4 and denote the remainder by r . The number

r	Q_1	Q_3
0, 1	$[(m\text{th} + \text{the next}) \text{ observation}]/2$	$[(n - m)\text{th} + \text{the next}] \text{ observation}]/2$
2, 3	$(m + 1)\text{th} \text{ observation}$	$(n - m)\text{th} \text{ observation}$

Table 1.

of observations in each quarter is given by $m = (n - r)/4$. Then we may use the 'Remainder Rule' given in table 1 to calculate extreme quartiles in such a way as to satisfy Property I.

◆ AN ILLUSTRATION ◆

An independent consumer group tested radial tyres from a major brand to determine expected tread life. The ordered observations (in thousands of miles) are as follows (see Vinning, 1998, p.193):

42	43	47	48	50
51	51	52	53	54
56	56	56	58	61

To illustrate the proposed formulae, we make four different data sets with sample sizes $n = 12, 13, 14$ and 15 and label them as Data 1, Data 2, Data 3 and Data 4, as follows:

Data 1:	42	43	47	48	50	51
	51	52	53	54	56	56
Data 2:	42	43	47	48	50	51
	51	52	53	54	56	56
						56
Data 3:	42	43	47	48	50	51
	51	52	53	54	56	56
		56	58			
Data 4:	42	43	47	48	50	51
	51	52	53	54	56	56
		56	58	61		

Quartiles are calculated below for each of these data sets.

Data 1. Here $n = 12, r = 0, m = 3$. Since $r = 0$, all the quartiles have ranks between two integers. The extreme quartiles are given by

$$Q_1 = \frac{1}{2}(m\text{th} + \text{the next}) \text{ observation}$$

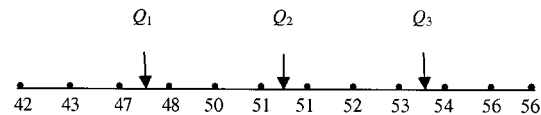
$$= \frac{1}{2}(3\text{rd} + 4\text{th}) \text{ observation} = \frac{1}{2}(47 + 48) = 47.5$$

and

$$Q_3 = \frac{1}{2}[(n - m)\text{th} + \text{the next}] \text{ observation}$$

$$= \frac{1}{2}(9\text{th} + 10\text{th}) \text{ observation} = \frac{1}{2}(53 + 54) = 53.5$$

There are three ($=m$) observations below the first quartile 47.5, three observations between the first quartile and the median $(51 + 51)/2 = 51$, three observations between the median and the upper quartile 53.5, and three observations above the upper quartile.



Data 2. Here $n = 13, r = 1, m = 3$. Since $r = 1$, the first and the third quartiles have ranks between two integers, and they are given by

$$Q_1 = \frac{1}{2}(m\text{th} + \text{the next}) \text{ observation}$$

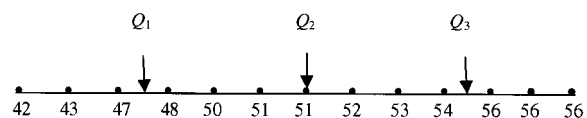
$$= \frac{1}{2}(3\text{rd} + 4\text{th}) \text{ observation} = \frac{1}{2}(47 + 48) = 47.5$$

and

$$Q_3 = \frac{1}{2}[(n - m)\text{th} + \text{the next}] \text{ observation}$$

$$= \frac{1}{2}(10\text{th} + 11\text{th}) \text{ observation} = \frac{1}{2}(54 + 56) = 55$$

There are three ($=m$) observations below the first quartile 47.5, three observations between the first quartile and the median 51, three observations between the median and the upper quartile 55, and three observations above the upper quartile.

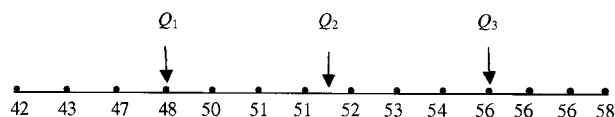


Data 3. Here $n = 14, r = 2, m = 3$. Since $r = 2$, the first and the third quartiles have integer ranks, and are given by

$$Q_1 = (m + 1)\text{th} \text{ observation} = 4\text{th} \text{ observation} = 48$$

$$Q_3 = (n - m)\text{th} \text{ observation} = 11\text{th} \text{ observation} = 56$$

There are three ($=m$) observations below the first quartile 48, three observations between the first quartile and the median $(51 + 52)/2 = 51.5$, three observations between the median and the upper quartile 56, and three observations above the upper quartile.



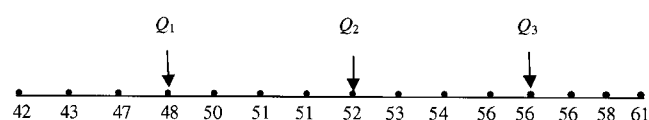
To illustrate the calculation of hinges, consider the following. Including the median in the calculation of the lower hinge means that it is between the 4th and the 5th observations. If, however, the median was not included, then the lower hinge would be the 4th observation, coinciding with the lower quartile as given by our definition.

Data 4. Here $n = 15$, $r = 3$, $m = 3$. Since $r = 3$, the first and the third quartiles have integer ranks, and are given by

$$Q_1 = (m + 1)\text{th observation} = 4\text{th observation} = 48$$

$$Q_3 = (n - m)\text{th observation} = 12\text{th observation} = 56$$

There are three ($=m$) observations below the first quartile 48, three observations between the first quartile and the median 52, three observations between the median and the upper quartile 56, and three observations above the upper quartile.



◆ CONCLUSION ◆

The set of proposed formulae for quartiles satisfying Property I will be of interest to some readers. We observe that the extreme quartiles are more easily figured out using the remainder (r) of the sample size with modulus 4. We note also that the position of the median as given by our proposed formulae is simply, as expected, the average of that of the extreme quartiles for any sample size.

The existing formulae in the literature based on consideration of the evenness or oddness of the sample size behave in the same way for some sample sizes. Method 1 is the simplest and enjoys Property I if the median of the whole data set is ignored in the calculation of extreme quartiles.

Acknowledgements

The authors acknowledge the excellent research facilities provided by King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia. This article is based on the authors' Technical Report (Firozzaman and Joarder 2000).

References

- Firozzaman, M. and Joarder, A.H. (2000). *A Refinement over the Usual Formulae for Quartiles*. Technical Report No. 254, Department of Mathematical Sciences, King Fahd University of Petroleum and Minerals, Dhahran.
- Lapin, L. (1997). *Modern Engineering Statistics*. Belmont, CA: International Thomson Publishing.
- Mayer, A.D. and Sykes, A.M. (1996). *Statistics*. London: Edward Arnold.
- Mendenhall, W. and Sincich, T. (1995). *Statistics for Engineering and the Sciences*. Englewood Cliffs, NJ: Prentice-Hall.
- Ostle, B., Turner, K.V. and Hicks, C.R. (1996). *Engineering Statistics – The Industrial Experience*. New York: Duxbury.
- Smith, P.J. (1997). *Into Statistics*. Heidelberg: Springer.
- Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison Wesley.
- Vinning, G.G. (1998). *Statistical Methods for Engineers*. Brooks/Cole Publishing Company.