King Fahd University of Petroleum & Minerals

DEPARTMENT OF MATHEMATICAL SCIENCES

Technical Report Series

TR 254

May 2000

A Refinement over the Usual Formulae for Quartiles

M. Firozzaman, A.H. Joarder

DHAHRAN 31261 ● SAUDI ARABIA ● www.kfupm.edu.sa/math/ ● E-mail: mathdept@kfupm.edu.sa

A Refinement over the Usual Formulae for Quartiles

M Firozzaman[1] and A H Joarder
Department of Mathematical Sciences
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia 31261

Emails: fzaman@bdonline.com and anwarj@kfupm.edu.sa

Abstract

The formula for the mid-quartile (median) is well defined in the literature. But the formulae for the upper and the lower quartiles available in the literature are conflicting in the sense that the quartiles do not divide the ordered sample observations into four segments having the same number of observations in each segment. The problems in the formulae for the upper and the lower quartiles available in the literature are pinpointed in this note. A set of formulae that does overcome above problems is proposed. The notion is also illustrated with an example.

Keywords and Phrases: Quartiles, remainder, modulus, prime number.

1. 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 the cumulative distribution function $F(x)$. Then the quartiles Q_1, Q_2 and Q_3 are defined by

$$F(Q_1) = \frac{1}{4},\ F(Q_2) = \frac{2}{4},\ \text{and}\ F(Q_3) = \frac{3}{4}\ \text{respectively.}$$

In the discrete case, it is, however, difficult to define quartiles by formulae that satisfy the following property:

Property I: Quartiles divide the ordered sample observations having the same number of observations (m) in each segment.

The literature is full of contradictory formulae for the extreme quartiles i.e. the lower and the upper quartiles. It seems the notion is hitherto neglected. In this note we propose a set of new formulae that takes care of the above problem.

[1] On leave from the university of Dhaka, Bangladesh.

2. A Review of the Well-known Formulae of Quartiles

The quartiles may be defined as the three numbers that divide the ordered sample observations into four segments having the same number of observations in each of them. In this section we survey the formulae for extreme quartiles available in the literature.

Method 1 The oldest and most popular method to find extreme quartiles is the following:

$$Q_1 = \frac{1}{4}(n+1) \text{ th observation}$$

$$Q_3 = \frac{3}{4}(n+1) \text{ th observation}$$

(see e.g. Mendenhall and Sincich, 1995, p54). It is interesting to note that Q_1 and Q_3 will have integer ranks for sample sizes of the form $n = 4i+3,\ (i = 1, 2, \ldots)$, and the above formulae satisfy **Property I.**

The quartiles Q_1 and Q_3 will have non-integer ranks (positions) for other sample sizes. It's a common practice to make arithmetic rounding in these cases. We observe that it is not the arithmetic rounding rather the linear interpolation that ensures **Property I.** for sample sizes $n = 4i, 4i+1,\ (i = 1, 2, \ldots)$. We note that arithmetic rounding is only needed for $n = 4i+2,\quad (i = 1, 2, \ldots)$ to ensure **Property I.**

Method 2 The formulae given by Vinning (1998, p 44) can be simplified as

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

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

The example he provides with $n = 35$ divide the observations into four segments with 9, 8, 8 and 9 observations among 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 4 equal parts only if n is even.

Method 3 To calculate any quartile Q_α $(\alpha = 1, 2, 3)$ find $n\alpha/4$. If it is an integer then the required quartile is the average of $n\alpha/4$th and $(n\alpha/4+1)$th observations. If $n\alpha/4$ is not an integer then round it to the next integer and the required quartile is the observation with rank equal to this rounded integer.

It is easy to check that the above formulae do not divide the sample observations into four equal parts if $n = 4i+1$.

Method 4 The formulae provided for percentiles by Smith (1997, p36) can be specialized for quartiles as

$$Q_1 = \begin{cases} \dfrac{n+2}{4} \text{ th observation if } n/4 \text{ is not an integer} \\ \dfrac{1}{2}\left(\dfrac{n}{4}th + \text{the next}\right) \text{observation if } n/4 \text{ is an integer} \end{cases}$$

$$Q_3 = \begin{cases} \dfrac{3n+2}{4} \text{ th observation if } 3n/4 \text{ is not an integer} \\ \dfrac{1}{2}\left(\dfrac{3n}{4} + \text{the next}\right) \text{observation if } 3n/4 \text{ is an integer} \end{cases}$$

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

Method 5 A popular method for finding quartiles is the linear interpolation. See e.g. Lapin (1997, pp. 45-46) and Ostle, Turner, Hicks and McElrath (1996, p 38). The quartile Q_α $(\alpha = 1, 2, 3)$ is usually interpolated by the following steps:

Step 1. Determine $R_\alpha = \alpha\,\dfrac{n+1}{4}$, $\alpha = 1, 2, 3$.

Step 2. Separate the integer (i) and decimal part (d) of R_α and write $R_\alpha = i + d$.

Step 3. The quartile is finally given by

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

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

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

3. The Proposed Formulae for Quartiles

Though the formulae for the median in the literature appear to be different, they all are equivalent. It is given by

$$Q_2 = \frac{n+1}{2}\text{th observation}.$$

In case n is odd, $\frac{n+1}{2}$ will be an integer so that the median will be an observation with integer rank. If however, n is even, $\frac{n+1}{2}$ will be an observation between $\frac{n}{2}$ th observation and the next one and by linear interpolation the median is given by

$$Q_2 = \frac{1}{2}\left[\frac{n}{2}\text{th observation} + \text{the next observation}\right].$$

The proposed formulae for the lower and upper quartiles are given below for even and odd sample sizes $n \geq 4$:

Case 1. (n is odd)

$$Q_1 = \begin{cases} \frac{1}{2}\left(\frac{n-1}{4}\text{ th observation} + \text{the next observation}\right) \text{ if } n = 4i+1 \\ \frac{n+1}{4}\text{ th observation} \text{ if } n = 4i+3 \end{cases}$$

$$Q_3 = \begin{cases} \frac{1}{2}\left(\frac{3n+1}{4}\text{ th observation} + \text{the next observation}\right) \text{ if } n = 4i+1 \\ 3\left(\frac{n+1}{4}\right)\text{th observation} \text{ if } n = 4i+3 \end{cases}$$

Case 2. (n is even)

$$Q_1 = \begin{cases} \frac{n+2}{4}\text{ th observation} \text{ if } n = 4i+2 \\ \frac{1}{2}\left(\frac{n}{4}\text{ th observation} + \text{the next observation}\right) \text{ if } n = 4i \end{cases}$$

$$Q_3 = \begin{cases} \frac{3n+2}{4}\text{ th observation} \text{ if } n = 4i+2 \\ \frac{1}{2}\left(\frac{3n}{4}\text{ th observation} + \text{the next observation}\right) \text{ if } n = 4i \end{cases}$$

where $i = 1, 2, 3, \ldots$

It is interesting to note that if sample sizes are of the form $n = 4i$, $n = 4i+1$, $n = 4i+2$, and $n = 4i+3$, ($i = 1, 2, 3, \ldots$) respectively, then the number of observations falling in each segment is given by

$$\frac{n}{4}, \frac{n-1}{4}, \frac{n-2}{4} \text{ and } \frac{n-3}{4}$$

respectively. Since the sample size can be represented by $n = r \bmod 4$, $(r = 0, 1, 2, 3)$, the number of observations in each of the four segments may be denoted by

$$m(r) = \frac{n-r}{4}, \quad (r = 0, 1, 2, 3)$$

or m for short.

Thus we have the following **Remainder Rule** to find quartiles:

r	Q_1	Q_3	$m(r) = \frac{n-r}{4}$
0	$\frac{1}{2}\left(\frac{n}{4}\text{th} + \text{the next}\right)$	$\frac{1}{2}\left(\frac{3n}{4}\text{th} + \text{the next}\right)$	$\frac{n}{4}$
1	$\frac{1}{2}\left(\frac{n-1}{4}\text{th} + \text{the next}\right)$	$\frac{1}{2}\left(\frac{3n+1}{4}\text{th} + \text{the next}\right)$	$\frac{n-1}{4}$
2	$\frac{n+2}{4}$th	$\frac{3n+2}{4}$th	$\frac{n-2}{4}$
3	$\frac{n+1}{4}$ th	$\left(3\,\frac{n+1}{4}\right)$th	$\frac{n-3}{4}$

The position of quartiles can also be expressed in terms of r and m. If $r = 2, 3$ (prime), the first quartile Q_1 has the $(m+1)$ th integer position and the third quartile has the $(n-m)$ th integer position. If $r = 0, 1$ (composite), the first quartile Q_1 has the position between m th and $(m+1)$ th observation, and the third quartile has the position between $(n-m)$ th and $(n-m+1)$ th observation. These are summarized below:

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

4. An Illustration

An independent consumer group tested radial tires from a major brand to determine expected tread life. The data (in thousands of miles) are given below:

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

(cf. Vinning, 1998, 193)

The ordered sample observations are given by

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 $n=12$, $n=13$, $n=14$, $n=15$ 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 the data sets above.

Data 1: Here $n = 12$, $r = 0$ and $m = 3$. Since $r = 0$, the first and the third quartiles have ranks between two integers, and are given by

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

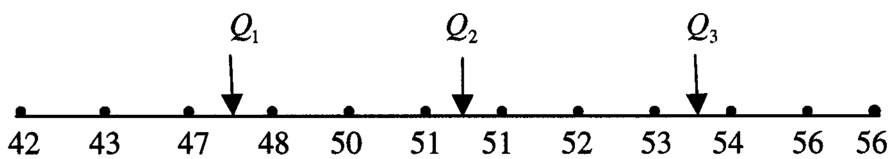
$$= \frac{1}{2} \text{ (3rd} + 4 \text{ th) observation} = \frac{1}{2} \ (47 + 48) = 47.5$$

and

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

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

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, three observations above the upper quartile.



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

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

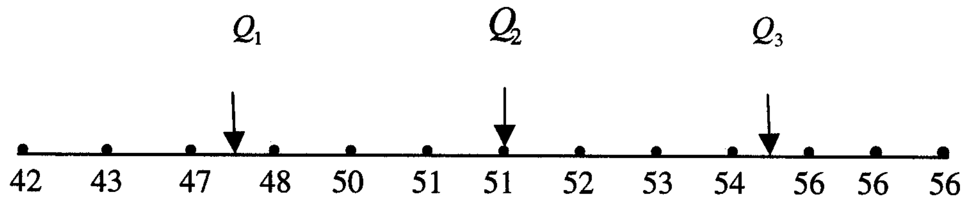
$$= \frac{1}{2} \text{ (3rd} + 4 \text{ th) observation} = \frac{1}{2} \ (47 + 48) = 47.5$$

and

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

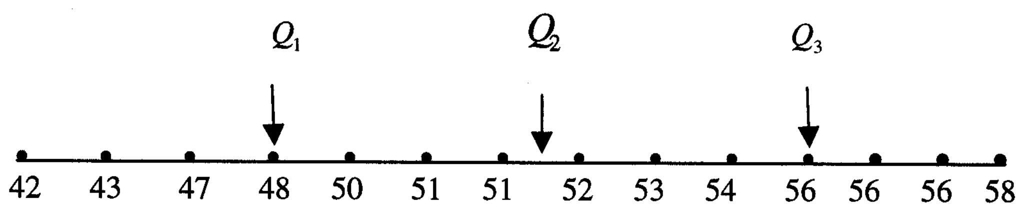
$$= \frac{1}{2} \ (10 \text{ th} + 11 \text{ th) observation} = \frac{1}{2} \ (54 + 56) = 55$$

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, three observations above the upper quartile.



Data 3: Here $n = 14$, $r = 2$ and $m = 3$. Since $r = 2$, the first and the third quartiles have integer ranks, and are given by
$Q_1 = (m+1)$ th observation $= 4$ th observation $= 48$ and
$Q_3 = (n-m)$ th observation $= 11$ th observation $= 56$

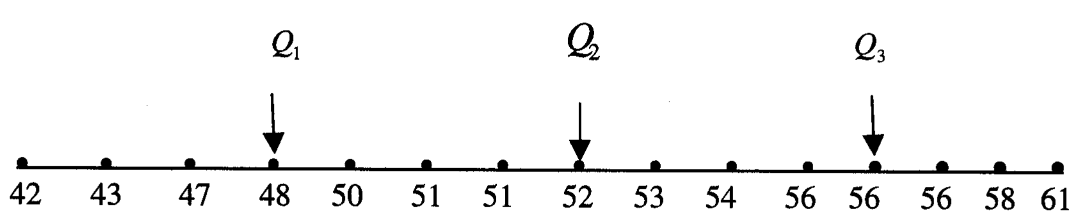
There are three $(= m)$ observations below the first quartile 48, three observations between the first quartile and the median (51+52)/2=51.5, three observations between the median and the upper quartile 56, three observations above the upper quartile.



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

$Q_1 = (m+1)$ th observation $= 4$ th observation $= 48$ and
$Q_3 = (n-m)$ th observation $= 12$ 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, three observations above the upper quartile.



5. Conclusion

The proposed formulae for quartiles satisfies **Property I** i.e. they divide the ordered sample observations into four segments having the same number of observations in each. In the proposed formulae, the position of the median is also, as expected, the average position of the extreme quartiles for any sample size. The existing formulae in the literature behave the same way but for some particular sample sizes. Most of the existing formulae for extreme quartiles are based on the consideration of the evenness or oddness of the sample size. We observe that the extreme quartiles are better figured out if the remainder (r) of the sample size with modulus 4 are categorized as prime and composite. The proposed formulae are thus an improvement over the usual ones.

Acknowledgement

The authors acknowledge the excellent research facilities provided by King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.

References

1. Lapin, L. (1977). *Modern Engineering Statistics.* International Thomson Pub. Belmont, CA.

2. Mendenhall, W. and Sincich, T. (1995). *Statistics for Engineering and the Sciences.* Prentice Hall. New Jersey.

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

4. Smith, P.J. (1997). *Into Statistics.* Springer-Verlag. Heidelberg.

5. Vinning, G.G. (1998). *Statistical Methods for Engineers.* Brooks/Cole Publishing Company. CA, USA.