

# THE UNIVERSITY OF TEXAS AT AUSTIN

## SCHOOL OF INFORMATION

### MATHEMATICAL NOTES FOR LIS 397.1

### INTRODUCTION TO RESEARCH IN LIBRARY AND

### INFORMATION SCIENCE

Ronald E. Wyllys  
Last revised: 2004 June 20

## QUARTILES AND *MICROSOFT EXCEL*

*Microsoft Excel* handles quartiles with its quartile function. The syntax of this function is given by the *Excel* help file, as follows:

#### Syntax

QUARTILE(array,quart)

Array is the array or cell range of numeric values for which you want the quartile value.  
Quart indicates which value to return.

If quart equals QUARTILE returns

- 0 Minimum value
- 1 First quartile (25th percentile)
- 2 Median value (50th percentile)
- 3 Third quartile (75th percentile)
- 4 Maximum value

#### Remarks

- If array is empty or contains more than 8,191 data points, QUARTILE returns the #NUM! error value.
- If quart is not an integer, it is truncated.
- If quart < 0 or if quart > 4, QUARTILE returns the #NUM! error value.
- MIN, MEDIAN, and MAX return the same value as QUARTILE when quart is equal to 0 (zero), 2, and 4, respectively.

#### Example

QUARTILE({1,2,4,7,8,9,10,12},1) equals 3.5

With the above help-file definition in mind, we can put the numbers 1,2,4,7,8,9,10,12 into an Excel spreadsheet. We will find that Excel reports  $Q_2$  as 7.5, in accord with the definitions we discussed in class. However, Excel reports  $Q_1$  as 3.5 and  $Q_3$  as 9.25, and these values do not accord with the definitions used in our class, which set  $Q_1$  as 3.0 and  $Q_3$  as 9.5.

(As a reminder, the definitions used in our class are based on what I like to call the "basic intuitive definition" of a quartile. By "basic intuitive definition" I mean definitions along the following lines: For a set of observations, the first quartile is a number having the property that one-fourth of the set of observations lie below it and three-fourths of the set of observations lie above it. Definitions for the second

and third quartiles are analogous. These basic definitions can fail to yield unique values for the quartiles, e.g., when any number within an interval has the desired property. In such cases, we employ a widely used convention: viz., to agree to label, as the quartile, the midpoint of the interval.)

As another example, for the set 1,2,3,4, *Excel* reports  $Q_1$  as 1.75,  $Q_2$  as 2.5, and  $Q_3$  as 3.25. Again only the median accords with our definitions.

As still another example, for the set 20,30,40,50 *Excel* reports  $Q_1$  as 27.5,  $Q_2$  as 35, and  $Q_3$  as 42.5 (our definitions yield 25 and 45 for  $Q_1$  and  $Q_3$ ).

We have to conclude that *Microsoft Excel* does some seemingly odd things with quartiles. The formula for the quartile calculation is not explained in the *Excel* help file, unlike the case with many of *Excel*'s other statistical functions, such as the standard deviation.

In the absence of an explanation by *Excel* of what it does in actually calculating quartiles, I am left having to guess what it does. My guess is that *Excel* constructs a quartile according some formula that tries to take the overall set of numbers into account. Here is such a formula<sup>1</sup>:

$$Q_k = L + \frac{w}{f_k}(kn - cf_b)$$

where  $Q_k$  = the  $k$ th percentile  
 $L$  = lower limit of the class interval that includes the  $k$ th percentile  
 $n$  = total frequency  
 $cf_b$  = cumulative frequency for all class intervals below the  $k$ th percentile class  
 $f_k$  = frequency of the class interval that includes the  $k$ th percentile  
 $w$  = width of the interval containing the  $k$ th percentile

Whether *Excel* actually follows this formula, I leave it to you, the reader, to test, in case you happen to be interested in pursuing the matter to that extent. For me, it suffices that *Excel* is almost surely employing some formula of this general nature, i.e., that *Excel* is following one of several possible conventions that can be used to handle situations in which the basic intuitive definition of quartile fails to yield a unique value.

Some years after I originally wrote the four paragraphs above (i.e., beginning "In the absence of an explanation"), a kind reader of this essay, Mr. Anders Langmyr, emailed me on 2003 April 22 as follows:

I am a student from Norway. I was going to use the quartile-function in Microsoft Excel in some of my work. I tried it out, but the results didn't make sense according to the definition I used for quartiles. So I started wondering about what kind of algorithm was behind this function, and searched for the answer on the web. I [found] your notes "Quartiles and Microsoft Excel", where you also wondered about this.

After I read through your notes I [discovered] some support pages from Microsoft where they explained the algorithm. So I just thought to send you this link, so that you can update your notes until the next time you use them.

<http://support.microsoft.com/?kbid=214072>

Regards  
Anders Langmyr

In short: Microsoft's quartile-algorithm

- 1 Find the  $k$ th smallest member in the array of values, where:  
 $k = (\text{quart}/4) * (n-1) + 1$   
If  $k$  is not an integer, truncate it but store the fractional portion ( $f$ ) for use in step 3.  
 $n$  = number of values in the array  
 $\text{quart}$  = value between 0 and 4 depending on which quartile you want to find
- 2 Find the smallest data point in the array of values that is greater than the  $k$ th smallest -- the  $(k+1)$ th smallest member.
- 3 Interpolate between the  $k$ th smallest and the  $(k+1)$ th smallest values:  
Output =  $a[k] + (f * (a[k+1] - a[k]))$   
 $a[k]$  = the  $k$ th smallest  
 $a[k+1]$  = the  $k+1$ th smallest

Example

Array of values 0,2,3,5,6,8,9

Finding the third quartile

1  $k = \text{TRUNC}((3/4 * (7-1)) + 1) = 5$ ,  $f = (3/4 * (7-1)) - \text{TRUNC}(3/4 * (7-1)) = 0.5$

2  $a[5] = 6$ ,  $a[5+1] = 8$

3 Output =  $6 + (0.5 * (8-6)) = 7$

My sincere thanks go to Mr. Langmyr for finding this explanation and taking the trouble to share it with me. Now we know what Microsoft is doing in the *Excel* quartile function. But we still have to wonder why Microsoft settled for using a rather poor algorithm—"poor" in the sense that the algorithm clearly yields results, even in certain simple cases, that fail to accord with what most people who use quartiles expect to see.

A very plausible explanation of why Microsoft chose the algorithm it did has been offered by Mr. David Shammai, who emailed me on 2004 June 16 to say:

I read your note on the algorithm used by *Excel* to calculate quartiles with great interest. I [have come] across this issue before. It seems to me that the methodology used by *Excel* is outside what one may call 'mathematical conventions' (do you agree?). I think (and this is my own personal theory) the reason this algorithm was used is so that when quartiles '0' and quartile '4' are calculated by *Excel*, the formula will produce *min* and *max* values of the sample respectively. This [makes] the algorithm . . . neater from an IT syntax perspective, but less accurate mathematically.

I suspect that Mr. Shammai has hit the nail on the head with this analysis, and I thank him for sharing it with me.

---

<sup>1</sup>This formula is adapted from: Ott, L. An Introduction to Statistical Methods and Data Analysis. North Scituate, MA: Duxbury Press; 1977. ISBN:0-87872-134-7.