

NOTE N: RANKING, QUARTILES, MEDIANS AND PERCENTILES

INTRODUCTION:

These are all computations on a univariate data set, and they are all conventions. There are two constructions:

1. The first is where the order of the data as it comes is not important, and reordering the data can be done without losing any information.
2. The second is where the initial order is important, or that the order is that of a second measure such as scores. This may be a more complicated arrangement of data, including categories, counts within intervals, etc.

The median calculation has an accepted convention. Quartiles and percentiles have no generally accepted convention, and there are many accepted methods for calculating them. Since there is no standard (Hyndaman and Fan 1996), any one of them represents acceptable usage. Excel's method is one of the accepted methods. Moore (2003) on p 43, "For example, the first quartile is \$19.27 by hand and \$19.06 by software. Software often uses more elaborate rules that aren't suited for hand calculation. Moreover, not all software uses the same rule. For example, the Microsoft Excel spreadsheet gives the first quartile as 19.3275. These differences are too small to affect conclusions based on the data. Just use the values that your software gives you."

Criticism about Excel's method is common, but it is an accepted method. All the methods however all depend on the data coming from a continuous distribution. Methods to calculate ranks and percentiles also have different interpretations, and again there is no accepted standard.

The issue about percentiles, quartiles and the median with respect to an integer number of values or values from a discrete distribution needs to be discussed here. The issue is how one interprets a set of discrete data. The question is, do values that are calculated as interpolations between discrete values exist or not.

If they do not exist, such as integers from nominal, ordinal and absolute measurements, then the data set can be represented by a series of columns as a histogram. Here a median value only exists for an odd number of data values in a set. For an even number of data points, a median represented as halfway between the two center values (which is the accepted convention here) would only exist if the two central values were the same. If they were different, one would have to interpolate, and obtain a value that may theoretically not be possible. In this case a median value could not be defined. The data can however be ranked. Percentiles of each point can be calculated from the ranked data.

If they exist, then the data set can be represented by a scatter plot graph, where each data value is a point on a uniform interval (horizontal) scale, and straight lines can be drawn through each point. Any value represented by points on the straight lines represents a valid interpolation. On this basis, exact quartiles and

median values can be calculated. This interpretation is more generally taken by statisticians, and is the basis for evaluating Excel outcomes. The ability to interpolate means that the data is from a continuous distribution.

Ranks, percentiles, quartiles and median values are derived from a sorted column of univariate data, sorted internally from smallest to largest. Identical values are taken as ties and are treated as being equal without any order within the tied group. To fully evaluate ranked data with ties, one first needs to be sure that the ties are true ties. If there is a difference, then a small value (epsilon) needs to be added to each of the identical values to form the correct sorted order.

MEDIAN AND QUARTILES:

For data from a continuous distribution of real numbers, the 25th percentile is the first quartile (Q1) and the 75th percentile is the third quartile (Q3). There is one accepted method for calculating the median (Q2), but it is not generally the 50th percentile. The accepted method for the median is:

If the number of data is odd, the median is the middle value of the sorted data list.

If the number of data is even, the median is the average of the two central data values in the sorted table.

The Excel median function follows this convention.

There are at least 10 known methods of calculating quartiles as the 25th and 75th percentiles identified by Hyndaman and Fan (1996). A summary of the Q1 method (as the 25th percentile) is in tables H-1 and H-2.¹

The methods all locate a rank value from the ranked values table, and either use that value as the quantile or calculate a value from the locations of two ranks

¹ This part of note H has been extensively rewritten on August 30, 2005, based entirely on a message on the EDSTAT list by
Daniel J. Nordlund
Research and Data Analysis
Washington State Department of Social and Health Services
Olympia, WA 98504-5204
His message was appreciated.

Table H-1: Basic Quartile Methods, No Interpolation

Method	Name/Source	First Quartile Location	Third Quartile	Interpolation between two values	Comments
1	Step	$n/4$	$3n/4$	no	If $g > 0$ select $x[(j)+1]$
2	Average Step	$n/4$	$3n/4$	no when $g > 0$, yes when $g = 0$	When $g = 0$, $Q = [x(j)+x(j+1)]/2$
3a	Nearest Integer to np	$n/4$	$3n/4$	no	Even choice when $g = 0$
3b	Nearest Integer to np	$n/4$	$3n/4$	no	Odd choice when $g = 0$

j = integer part of the quartile location calculation. (Hyndman and Fan's symbols)

g = fractional part of the quartile location calculation

Table H-2: Basic Quartile Methods, With Interpolation

Method	Name/Source	First Quartile k value	Third Quartile k value	Comments
4	Parzen	$n/4$	$3n/4$	Method 1 with interpolation
5	Hazen	$(n+2)/4$	$(3n+2)/4$	Value midway between method 1 steps
6	Weibull	$(n+1)/4$	$3(n+1)/4$	Sample space is $n+1$ regions
7	Gumbell	$(n+3)/4$	$(3n+1)/4$	
8	Median Position	$(3n+5)/12$	$(9n+7)/12$	Median unbiased
9	Bernard & Bos-Levenbach	$(n/4)+0.31$	$(3n/4)+0.6$	
10	Blom's plotting position	$(4n+7)/16$	$(12n+9)/16$	A better approximation when the distribution is normal
11	Moore-1	$(n+1/2)/4$		
12	Moore-2	$(n+1)/4$		Sample space remains as n

j = integer part of k . (Hyndman and Fan's symbols)

g = fractional part of k

Interpolation is from j to $j+1$ with g as the proportion between j and $j+1$.

Method 13 is Tukey's hinge and method 14 is Moore's variation on Tukey's hinge.

Tukey's hinge and Moore's hinge are not 25 percentiles, but depend on direct division of the data set into halves and the use of the accepted median method to the divided parts.

Excel does not calculate these hinge values. They are however, extensively used in box plots.

Hyndman and Fan (1996) identified six desirable properties of quantile measures in their table 1. Method 5 was the only method that met all six requirements. However method 5 is one of the least used methods in commercial software. See Table H-3.

Table H-3: Methods Used By Some Common Statistical Programs

Software References	Method Number
SAS #1	4
SAS #2	3a
SAS #3	1
SAS #4	6
SAS #5	2
Excel	7
Minitab (DESCRIBE)	6
Minitab (%DESCRIBE)	2
GLIM (percentile)	2
GLIM (interpolate)	5
BMDP	6
SPSS (frequencies)	6
S-PLUS-3.1 (quantile)	7
R	7

For small data sets, there are appreciable differences between the methods. For large data set the differences are small. This is illustrated in table H-4.

Table H-4, First Quartile on Different Data Sets

Data Set	10-40	10-50	10-60	10-70	1-100	1-99	1-98	1-97	1-96
Number in Set	4	5	6	7	100	99	98	97	96
Interval Size	10	10	10	10	1	1	1	1	1
Step	10	20	20	20	25	25	25	24	24
Average step	15	20	20	20	25	25	25	24.5	24.5
Nearest Integer	10	10	20	20	25	24	24	24	24
Method 4	10.00	12.50	15.00	17.50	24.75	24.50	24.25	24.00	24.00
Method 5	15.00	17.50	20.00	22.50	25.25	25.00	24.75	24.50	24.50
Method 6	12.50	15.00	17.50	20.00	25.00	24.75	24.50	24.25	24.25
Method 7	17.50	20.00	22.50	25.00	25.50	25.25	25.00	24.75	24.75
Method 8	14.17	16.67	19.17	21.67	25.17	24.92	24.67	24.42	24.42
Method 9	14.00	16.50	19.00	21.50	25.15	24.90	24.65	24.40	24.40
Method 10	14.38	16.88	19.38	21.88	25.19	24.94	24.69	24.44	24.44
Method 11	11.25	13.75	16.25	18.75	24.88	24.63	24.38	24.13	24.13
Method 12	20.00	20.00	20.00	25.00	25.50	25.00	25.00	25.00	25.00
Method 13	15.00	20.00	20.00	25.00	25.50	25.00	25.00	24.50	24.50
Method 14	15.00	15.00	20.00	20.00	25.00	25.00	24.50	24.50	24.50

Note: Each data set consists of integers with equal intervals. Data sets 10-40 through 10-70 have intervals of 10. Data sets 1-100 to 1-96 have intervals of 1. There are no ties. For example 10-40 consists of the four numbers, 10, 20, 30, 40.

The differences between maximum and minimum first quartile values range from 0.75 to 1.25. Some of the methods are consistently on the minimum side, some always in the middle and some on the large side. This is shown in table H-5

Table H-5: First Quartile Calculated Values, Ranked by Value, Reported as Method Number.

100	99	98	97	96	95	4	5	6	7	8	9
1	4	3	3	3	4	1	3	4	4	4	3
3	11	4	4	4	11	3	4	11	11	3	4
4	1	11	11	1	3	4	11	6	6	1	11
13	3	6	6	11	1	11	6	9	3	11	6
6	6	9	14	6	6	6	14	8	14	6	14
9	2	8	9	9	14	9	9	10	1	9	9
8	14	10	8	8	2	8	8	3	2	8	8
10	9	1	10	10	9	10	10	14	9	10	10
2	8	2	5	14	8	2	5	5	8	14	5
5	10	14	1	5	10	5	1	1	10	2	1
13	5	5	2	2	5	13	2	2	5	5	2
14	13	13	13	13	13	14	13	13	13	13	13
7	7	12	12	7	7	7	7	12	12	7	7
12	12	7	7	12	12	12	12	7	7	12	12

The first row is the number in the data set, from row 2 of table H-4. Each data set has unity interval width. The top rows represent the smaller quartiles and the bottom rows represent the larger quartiles. The number represents the computation method as given in tables H-1 and H-2. Some of the columns have considerable ties at the top or bottom, so that only very general comparisons can be made

The Excel method is method 7. (Gumbell)

Excel calculates correct quartiles and median values, even when tied values occur. You do not have to presort the data to get correct values. These are the functions QUARTILE and MEDIAN.

The first quartile calculates the interpolation corresponding to the position of $(n+3)/4$ in the sorted list (may be one of the values in the list or may be in between two values in the list). The third quartile calculates the interpolation corresponding to the position of $(3n+1)/4$ in the sorted list.

The Excel method generally gives the larger quantile values and is close to Tukey's hinge value, which is good.

RANKS:

When there are no tied values, then ranking order is done in accordance with the order based on a sorted list and the location in the list. The list can be sorted either ascending or descending, and the rank depends of which way the list was sorted. There are no accepted conventions here. The problem of ranking with ties is not logically solvable if ranks always have to be integers. Several approaches have been taken in the literature.

When there are tied values, the ranking can be done several ways.

The current Excel method: Microsoft says, "RANK gives duplicate numbers the same rank. However the presence of duplicate numbers affects the ranks of subsequent numbers. For example, in a list of integers, if the number 10 appears twice and has a rank

of 5, then 11 would have a rank of 7 (no number would have a rank of 6).” Table H-6 shows how this is done:

Table H-6: Excel Ranking Method Results on a Data List of 12 Values

Data	Table Order	Downward Rank	Upward Rank
2.3	1	1	12
2.8	2	2	11
3.1	3	3	10
3.6	4	4	9
4.2	5	5	6
4.2	6	5	6
4.2	7	5	6
5.6	8	8	5
6.0	9	9	4
6.5	10	10	2
6.5	11	10	2
7.3	12	12	1

Several sources criticize Excel on this ranking method, and state that Excel is wrong. One wanted the ranking to give the smallest value a rank of 1 and the largest value a rank of 12, and to distribute ranking values of the values in-between, such that the sum of ranks equals the sum of (1,2,3,4,5,6...) or 78 (in this case). The sum of downward ranks is 74 and the sum of upward ranks is 74, both not equal to 78.

Table H-7: Equal Sum of Ranks Method on a Data List of 12 Values

Data	Downward Rank	Upward Rank
2.3	1	12
2.8	2	11
3.1	3	10
3.6	4	9
4.2	6	7
4.2	6	7
4.2	6	7
5.6	8	5
6.0	9	4
6.5	10.5	2.5
6.5	10.5	2.5
7.3	12	1

You end up however with half ranks in order to preserve the proper sequence, the presence of duplicates, and the total sum of ranks.

Other critics just stated Excel is wrong, without giving any information or basis for their claim of error. Hesse (2006) is in this group.

The Excel method while not preserving ranks, “pins” the tied group to an adjacent point. If the PERCENTRANK function is plotted (see figure1 below), the interpolation between points follows a more correct point-to-point relationship, than if the tied value ranks “floated”. For the example above, the 4.2 ties would be at a rank of 6, which would not be tied to the previous or next point.

PERCENTILES:

There are different views on how to calculate percentiles. There is the view from the “statistical” communities, and the view from the “educational research and social research” communities. The later heavily use “scores” as measures, and the former use a variety of measures. Both views will be described here.

A VIEW FROM THE STATISTICAL COMMUNITY

Table H-8 gives the percentile methods from Hydaman and Fry (1996), which basically are from the “statistical” communities. There is no standard method, and any of these is considered acceptable.

Table H-8: Basic Percentile Methods

Met	Name/Source	P value	First percentile	Last percentile
1	Step	$= k / n$	$100/n$	100
2	Average Step	$= k / n$	$100/n$	100
3	Nearest Integer	$= k / n$	$100/n$	100
4	Parzen	$= k / n$	$100/n$	100
5	Hazen	$= (k - 1/2) / n$	$50/n$	$100-50/n$
6	Weibull	$= k / (n + 1)$	$100/(n+1)$	$100n/(n+1)$
7	Gumbell	$= (k - 1) / (n - 1)$	0	100
8	Reiss	$= (k - 1/3) / (n + 1/3)$	$200/(3n+1)$	$100(3n-1)/(3n+1)$
9	Bernard & Bos-Levenbach	$= (k - 0.3) / (n + 0.4)$	$350/(5n+2)$	$100(n-0.3)/(n+0.4)$
10	Blom	$= (k - 3/8) / (n + 1/4)$	$500/(8n+2)$	$100(8n-3)/(8n+2)$

Note: Methods 1, 2 and 3 do not interpolate, method 3 is a step functions at the ½ point, and all the rest use linear interpolation between X(k) and X(k+1).

Excel uses method number 7 (Gumbell) to relate the sorted table order to percentiles. For example:

EXCEL REPRESENTATION OF RANK AND PERCENT RANKS

Table H-9: Percent Rank values for a Data List of 12 Values

Values	Table order	Downward Rank	Upward Rank	RANK	PERCENTRANK
2.3	1	1	12	1	0
2.8	2	2	11	2	0.090909
3.1	3	3	10	3	0.181818
3.6	4	4	9	4	0.272727
4.2	5	5	6	5	0.363636
4.2	6	5	6	5	0.363636
4.2	7	5	6	5	0.363636
5.6	8	8	5	8	0.636364
6.0	9	9	4	9	0.727273
6.5	10	10	2	10	0.818182
6.5	11	10	2	10	0.818182
7.3	12	12	1	12	1

k in the equation for method 7, is the value in the RANK column, and the value in the PERCENTRANK column is the result of the equation. Microsoft refers to the p value

from 0 to 1 as a percentile in these functions, which is not correct. Percentiles run from 0 to 100. For correct percentile values, you will have to multiply the cell values by 100.

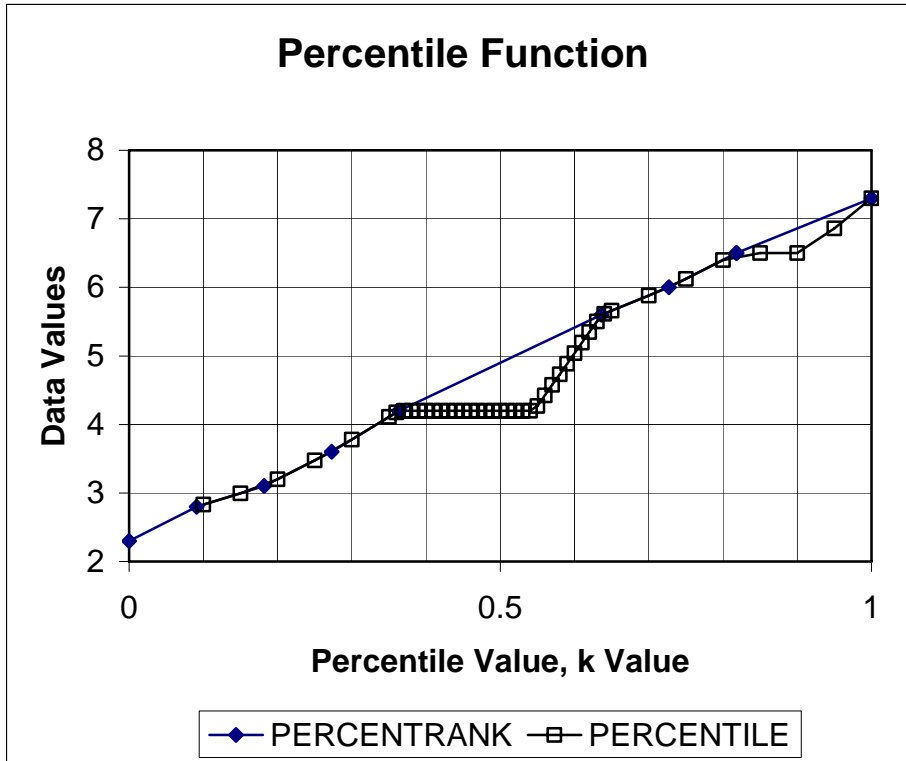
For the table H-9 values from above, the function PERCENTILE returns the following values

Table H-10: Outputs of the PERCENTILE Function on a Data List of 12 Values.

K Value	PERCENTILE output
0.10	2.8300
0.15	2.9950
0.20	3.2000
0.25	3.4750
0.30	3.7800
0.35	4.1100
0.40	4.2000
0.45	4.2000
0.50	4.2000
0.55	4.2700
0.60	5.0400
0.65	5.6600
0.70	5.8800
0.75	6.1250
0.80	6.4000
0.85	6.5000
0.90	6.5000
0.95	6.8600

They are interpolated rank values (not table sorted order) corresponding to an interpolation on RANK values. They have a sort of odd characteristic. See Figure H-1.

Figure H-1: Percentile Function Values



Instead of dealing with the ties as a step function, PERCENTILE puts in a “ramp” at the end of the tie region (as shown in the figure). The logic is that the tie continues from the table order 5 ($p=0.3636$) to table order 7 ($p=0.5253$) and a linear transition is taken from table order 7 to table order 8 ($p=0.6363$), which is at rank 8.

There is a Data Analysis Routine called “Rank and Percentage”. This gives a table of four columns, calculated from an input range. “Rank and Percentages” essentially does what the functions do, but the percentages are rounded down.

Table H-11 shows how the RANK function works and how it compares to the Data Analysis output. The last two columns are the last two columns of the Data Analysis output. The first two columns of the Data Analysis Output are the location of the sorted value in the original table.

A VIEW FROM THE EDUCATIONAL RESEARCH COMMUNITY

The educational research community heavily uses “scores” as a primary measurement. Scores here are typically integer values, and a concern is about meanings when values in-between score values are to be interpreted.. Paul Barrett (2004) reviews several textbooks and other sources, and concludes, “A percentile is the point in a distribution **at or below** which a given percentage of scores is found -or- The value **below** which P% of the values fall is called the P^{th} percentile. In fact, both definitions are correct. What is at fault is the lack of clarity in some cases over what constitutes a “score”.

There are then two terms here, “percentile” and “percentile rank”

Barrett (2004) gives equations for values as follows:

PERCENTILE

$$P_i = ll + \left(\frac{np - cf}{f_i} \right) \cdot w$$

where

P_i = the i^{th} percentile

ll = the exact lower limit of the interval containing the percentile point

n = the total number of scores

p = the proportion corresponding to the desired percentile

cf = the cumulative frequency of scores below the interval containing the percentile point

f_i = the frequency of scores in the interval containing the i^{th} percentile point

w = the width of the class interval

PERCENTILE RANK

$$PR_x = \left[\frac{\left(cf + \left(\frac{x - ll}{w} \right) \cdot f_i \right)}{n} \right] \cdot 100.0$$

where

PR_x = the percentile rank of score x

ll = the exact lower limit of the interval containing the score x

n = the total number of scores

cf = the cumulative frequency of scores below the interval containing the score x

f_i = the frequency of scores in the interval containing x

w = the width of the class interval

To see how these differ, lets look at Barrett's data

Table H-11: Frequency Table, LONG_E(EPQR100M.STA)

Score	Interval Lower Limit	Interval Upper Limit	Midpoint	Basic Count	Cumulative Count
0	-0.5	0.5	0	9	9
1	0.5	1.5	1	12	21
2	1.5	2.5	2	13	34

3	2.5	3.5	3	17	51
4	3.5	4.5	4	16	67
5	4.5	5.5	5	12	79
6	5.5	6.5	6	15	94
7	6.5	7.5	7	16	110
8	7.5	8.5	8	22	132
9	8.5	9.5	9	26	158
10	9.5	10.5	10	32	190
11	10.5	11.5	11	31	221
12	11.5	12.5	12	36	257
13	12.5	13.5	13	31	288
14	13.5	14.5	14	29	317
15	14.5	15.5	15	33	350
16	15.5	16.5	16	39	389
17	16.5	17.5	17	35	424
18	17.5	18.5	18	29	453
19	18.5	19.5	19	31	484
20	19.5	20.5	20	34	518
21	20.5	21.5	21	39	557
22	21.5	22.5	22	33	590
23	22.5	23.5	23	20	610
Total				610	

The Excel function 75th percentile on column 1 is 17.25 and the 75th percentile on column 6 is 431.25. By interpolation the column 6 percentile corresponds to a score of 17.25. The Excel functions do not give the score value directly.

Barrette's 75th percentile here is a score of 18.645. This is the score at which 75% of the observations will be observed to be below this score. However this score is not attainable because this was from an integer scored test. For a score of 18 the percentile is 71.89, and for a score of 19, the percentile is 76.80%. For these two values, the linear interpolated 75% corresponds to a score of 18.37. The value 18.37 is a linear interpolation, and given Barrett's equations, is not a valid check of the validity of the 18.645 value.

CONCLUSIONS

The Excel percentile measure does not correspond to Barrett's percentile measure. The Excel value is 17.25 and Barrett's equations give 18.645. However Excel gives Gumbell's measure. **Both are based on logic, but arrive at different values.**

We have then at least eleven different computations for rank and percentile. They do not give the same values, but give different results. Since Excel correctly gives values for one of the eleven, it cannot be faulted for giving incorrect rank and percentile values.

ISSUES ON RANKING, QUARTILES, PERCENTILES, HINGES AND TIES:

This is a review of what was said in terms of Hyndman and Fan's comments (Hyndman and Fan 1996) and on the Queensland Web page <http://exploringdata.cqu.edu.au/ticktack.htm>

The issue of Tukey's hinges came up. The Queensland web site talks about hinges, and concludes that Tukey was very flexible about it. Viewed as a quick, no-computer method to estimate a data set characteristic. Although cited as being the median of the lower half of the sorted data set, there are different interpretations of this statement, and consequently there is an uncertainty on how a computer should get a value of Q1, Tukey's lower hinge. Four short data sets are given with Q1 values

Table H-12: Tukey's Hinge Values

Data Set	Tukey's Hinge	Minitab Q1	Moore & McCabe (1998)
12, 20, 28, 36	16	14	16
10, 20, 30, 40, 50	20	15	15
10, 20, 30, 40, 50, 60	20	17.5	20
10, 20, 30, 40, 50, 60, 70	25	20	20

The Minitab Q1 is $(n+1)/4$

Tukey's Hinge is as follows: m , n integers, and k real

$$m = n \setminus 2$$

$$\text{If } (n - 2 * m) > 0 \text{ then } m=m+1$$

$$m = m + 1$$

$$k = m / 2 \text{ 'value is an interpolated value at data set index } k$$

Tukey's hinge is logical.

The issue of ties. Ties represent equalities as far as the data set is concerned. The practice of assigning one rank value to all equal ties is another issue and another problem in interpretation. Under Hyndman and Fan (1996), ties have different rank values. This presents problems in setting tied values to different percentile values. Under Hyndman and Fan (1996) percentiles are "pinned" to rank values, and percentiles in-between take on values according to the defined "model".

Hyndman and Fan (1996) talk about plotting positions. The basic approach is to use the equation $p = (k-a)/(n-a-b+1)$ to define a "percentile" for each point k (k from 1 to n). Both a and b are arbitrary constants. With $a=0$ and $b=1$, there is linear interpolation between points for definitions 1 and 4. For each definition, appropriate a and b values for linear interpolation between points can be found.

Donald Burrell proposed 3 models which lead to some difficult areas in interpretations. He also replied that, "The $(n+1)/4$ quartile which you state is derived from definition 6 and can be identified as Gumbell's method. The $(n/4 + 1/2)$ quartile in your message is derived from definition 5, which came from Weibull. These were obtained by setting the

p equations $= 1/4$. In H&F, they start from the first value in the sorted list where $j = 1$. He also raises the issue of integer data and limited ability to resolve measurements except into bins, which complicates things. I don't have any resolution here at all.

David Moore's comment that "any old" quartile value is OK has some validity. However one always has to have the ability to say, yes that approximation (value) is acceptable and that one is "wrong". However in the game of analyzing black box statistical software outputs, you don't have the tools (except in very limited cases where published "ok" values are available in the literature) to determine what is OK and what is not OK. Accuracy becomes the primary evidence of acceptable software. Consequently if the evaluator puts in a known data set that has a known quartile, and the output is not close enough to the evaluator's expectation, the software is at fault because it does not give the right answer. This is the game.