# Quartiles: How to calculate them?

# A. Introduction

## 1. Purpose

The purpose of the document is to explain the way the percentiles (quartiles) are calculated by the different software available at NERC sites (such as Excel, Minitab, and SAS), as well as to explain the meaning of this statistical notion. Although this statistical tool is widely used by scientists, the values given by software can be a source of confusion for them as these values might differ according to the software used. This document therefore tries to clarify this concept.

## 2. Definition

To understand the quartiles, one needs to understand percentiles (or quantiles) as the quartile calculation depends on the percentiles definition. The First quartile is the 25th percentile (noted Q1), the Median value is the 50th percentile (noted Median), and the Third quartile is the 75th percentile (noted Q3).

Among all software, it's SAS which gives the clearest definition of a percentile:
'' A percentile is a value at or below which a given percentage or fraction of the variable values lie. For a set of measurements arranged in order of magnitude, the p-th percentile is the value that has p% of the measurements below it and (100-p)% above it. Thus, the 20th percentile is the value such that one fifth of the data lie below it. It is higher than 20% of the data values and lower than 80% of the data values.''
[Extract from SAS Help files]

The problem is then to give a computational method for the percentile. Indeed, it may be easy to work out the 50th percentile, but problems then come to roughly any other percentile. Ex: What is the 75th percentile of a data set of three observations?

Hence the useful remark given by WS Cleveland for the quantiles: ''The f quantile, q(f), of a set of data is a value along the measurement scale of the data with the property that approximately a fraction f of the data are less than or equal to q(f). The property has to be approximate because there might not be a value with exactly a fraction f of the data less than or equal to it. (…) An explicit rule is needed for computing q(f).''

This is why different rules (or methods) exist to calculate the percentiles. We will try to see a few of them in this document to illustrate the problem.

Note: SAS will be used here as a reference for this problem, as it is one of the most complete statistical software packages available.

## 3. *Example to be studied / Notations*

To better understand the different methods, we'll apply them on an simple example.
The data set studied is:

| Variable | X1 | X2 | X3 | X4 |
|---|---|---|---|---|
| Value | 2 | 1 | 4 | 3 |

Once ordered it becomes:

| Variable | X(1) | X(2) | X(3) | X(4) |
|---|---|---|---|---|
| Value | 1 | 2 | 3 | 4 |

The following notations described below will be used to illustrate the different methods available to calculate the quantiles.

Let n be the number of observations in a data set (here n=4), and x(1), … x(n) the ordered values of a data set. Let p be the p-th percentile we want to calculate (ex: p=0.25, 0.5, or .75).

For each of these methods, we'll need to calculate the product n*p (or a similar one).
The product n*p can be split up between j and g, where j is the integer part of n*p and g is the decimal part of n*p.

Ex: for p=0.5 ($50^{th}$ percentile), n=8,
n*p= 8*0.5= 4 = 4 + 0 , hence j=4 and g=0.

Ex: for p=0.75 ($75^{th}$ percentile), n=10,
n*p= 10*.75= 7.5 = 7 + 0.5, hence j=7 and g=.5 .

Let y be the percentile associated to p (so that, in fact, y= q(p)). We then have approximately p % of the data set values lying below y. In the following paragraphs, the p-th percentile will always be referred to as y.

# B. Methods

## *1. SAS Methods*

SAS provides us with five methods to calculate the percentiles. We'll see two of them to illustrate the percentile calculation. For the others, please refer to the manual (SAS Procedure Guide, Procedure Univariate, statement ''PCTLDEF= '' ).

- SAS Method 5 (default method)

This method is the default method of SAS and is based on the empirical distribution function. The p-th percentile is defined by:

$$\begin{cases} \left(x_{(j)} - x_{(j+1)}\right)/2 & \text{if } g = 0 \\ x_{(j+1)} & \text{if } g > 0 \end{cases}$$

where n*p= j + g .

Ex: For our example, we want to calculate Q3, p=0.75, n=4, hence:
n*p = 4*0.75 = 3 = 3 + 0 (j=3 and g=0).

So y= ( x(3) + x(4) ) /2 = (3+4)/2 = 3.5 .
The 75th percentile is 3.5 with SAS Method 5 .
The full results are :

| Method for percentile calculation | Q1 | Median | Q3 |
|---|---|---|---|
| SAS Method 5 | 1.5 | 2.5 | 3.5 |

In fact, each value x(j) is associated to a probability slightly above (j-1)/n . There is roughly (j-1)/n*100 % of the measurements that fall below x(j).

- SAS Method 4

For this method, we use (n+1) instead of n. In that case, the p-th percentile is defined by:

y= (1-g)*x(j) +g*x(j+1), where (n+1)*p= j + g (and x(n+1) is taken to be x(n)).

Ex: In our example, for Q1: p=0.25, n=4 :
(n+1)*p = 5*0.25 = 1.25 = 1 + .25 (j=1 and g=0.25).

So y= (1-0.25)*x(1) + 0.25*x(2) = 0.75*1+0.25*2= 1.25.
The 25th percentile is 1.25 with SAS Method 4.

The full results are:

| Method for percentile calculation | Q1 | Median | Q3 |
|---|---|---|---|
| SAS Method 4 | 1.25 | 2.5 | 3.75 |

In other words, each value x(j) is associated to a probability of (j/(n+1)). The percentiles for p values between these references (j/(n+1)) are obtained by simple interpolation (see formula above).
Here again, we can say that there is roughly (j/(n+1))*100 % of the data that fall below x(j).

## *2. Minitab Method*

Minitab's method is the same as the SAS Method 4. The results for the quartile calculation are therefore:

| Method for percentile calculation | Q1 | Median | Q3 |
|---|---|---|---|
| Minitab Method = SAS Method 4 | 1.25 | 2.5 | 3.75 |

## *3. Excel Method*

For this method, Excel uses (n-1) instead of n. the p-th percentile is defined by:

y= (1-g)*x(j+1) +g*x(j+2), where (n-1)*p= j + g (and x(0) is taken to be x(1)).

Ex: In our example, for Q1: p=0.25, n=4 :
(n-1)*p = 3*0.25 = 0.75 = 0 + .75 (j=0 and g=0.75).

So y= (1-0.75)*x(1) + 0.75*x(2) = 0.25*1+0.75*2= 1.75.
The 25th percentile is 1.75 with the Excel Method.

The full results are:

| Method for percentile calculation | Q1 | Median | Q3 |
|---|---|---|---|
| Excel Method | 1.75 | 2.5 | 3.25 |

In other words, each value x(j) is associated to a probability of ((j-1)/(n-1)). The percentiles for p values between these references (j/(n-1)) are obtained by simple interpolation (see formula).
There is roughly ((j-1)/(n-1))*100 % of the data that fall below x(j).

The method developed by Excel does not fall into any of SAS methods.

## *4. Other Methods*

Many other methods do probably exist. Depending on people's needs, some are more relevant than others. For example, Cleveland associates each x(j) with a probability of (j-0.5)/n , and interpolates to obtain the percentiles for p values between the references ((j-0.5)/n).

# C. Comparisons of the different methods

## *1. Test program*

A test was carried out by running the following program ‘quartile.sas’ against the Excel data file ‘myvard.xls’. This has to be run after having copied the two files under N:\sas612 (or by changing the paths at the beginning of the ‘quartile.sas’ file). The program displays Excel, Minitab and SAS 5 results, as well as compares Excel results with the corresponding theoretical algorithms.

Contents of the program quartile.sas follow:

```
/* These first two lines have to be modified to reflect the location (path) */
/* and name of the Excel file. The excel file, if modified, has to be saved */
/* as an Excel 5.0 type file.                                              */

FILENAME excelfil "N:\sas612\myvard.xls";
LIBNAME excellib "N:\sas612\";
/* End of user input */

/* The following is standard to any dataset   */
/* Proc ACCESS: reads an Excel Data file */
PROC ACCESS DBMS=XLS;
CREATE excellib.datac.access;
   GETNAMES=Y;
   PATH= excelfil;
   ASSIGN=Y;
   LIST ALL;
CREATE excellib.datvi.view;
   SELECT ALL;
   LIST VIEW;
RUN;

/* DATA step: creates a temporary dataset with the values of the xvar */
/* This dataset is not ordered   */
DATA indata (DROP= q1exc q2exc q3exc);
   SET excellib.datvi;
   IF XVAR NE . ;
RUN;

/* #############################################################################  */
/* All calculations below are created to check formulae for Quartile calculations */
/* carried out in Excel                                                           */
/* Proc Sort: sorts indata with increasing xvar values. The output dataset is temp1 */
PROC SORT DATA= indata
     OUT= temp1 ;
BY xvar;
RUN;

/* Proc Means: creates a temporary dataset with the number of elements of temp1 */
PROC MEANS DATA=temp1 NOPRINT;
   VAR xvar;
   OUTPUT OUT=temp2
          N= number;
RUN;

/* Proc Transpose: creates a temporary dataset with each element of temp1 as a
variable*/
PROC TRANSPOSE DATA= temp1
     PREFIX=xord
     OUT= temp3;
RUN;
```

```
/* Data step: merges temp3 and temp2 to create calcdata   */
/* For our example:                       calcdata has: xord1 xord2 xord3 xord4 number
*/
/*                                                        1     2     3     4     4
*/
DATA calcdata (DROP= _TYPE_ _FREQ_);
   MERGE temp3 temp2;
RUN;

/* Data step: calculates the quartiles according to Excel formulae */
DATA rescalc (KEEP= Q1 median Q3 calcname software);
   SET calcdata;
   ARRAY xord{100} ;

   FORMAT calcname $20.;
   FORMAT software $30.;

   /* Theoritical formula is (n-1)*p= j + g   */
   /* where n is the number of elements for the xvar (ie 'number' here) */
   /*       p is the desired quantile (ex: 0.25 for Q1, 0.5 for median, 0.75 for Q3)*/
   /*       j is integer((n-1)*p)  */
   /*       g is decimals((n-1)*p) */

   /* Creates an array which will contain the values Q1,median,Q3 */
   ARRAY Qval{3};

   /* Creates an array which contains the values of p with which we are calculating */
   /* Q1, median, and Q3  */
   ARRAY Pval{3};
   Pval{1}=0.25;
   Pval{2}=0.5;
   Pval{3}=0.75;
   calcname='Method MS Excel';
   software='MS Excel method check with SAS';

   /* Loop for calculations of Q1, median, and Q3 */
   DO i= 1 TO 3;
   real= (number-1)*Pval{i};
   intg= INT(real);
   dec= real-intg;
   Qval{i}= (1-dec)*xord{INT(intg+1)} +(dec)*xord{INT(intg+2)};
   END;

   Q1=Qval{1};
   median=Qval{2};
   Q3=Qval{3};
RUN;
/* End of calculations to check MS Excel method.                                   */
/* The output created rescalc contains all relevant information Q1, median, Q3.   */
/* #############################################################################  */

/* DATA step: creates resexcel with result of the Quartile function run under EXCEL */
DATA resexcel (DROP= xvar);
   SET excellib.datvi;
   RENAME Q1exc= Q1
          Q2exc= median
          Q3exc= Q3;
   IF _N_=1;
   calcname='Method MS Excel';
   software='MS Excel output';
RUN;

/* Proc Univariate: creates temp4 with results from Method 4 - SAS Method - Minitab */
PROC UNIVARIATE DATA=indata
                PCTLDEF= 4;
    VAR xvar;
    OUTPUT OUT= temp4
              MEDIAN= median
              Q1=Q1
              Q3=Q3;
RUN;

/* Data step: creates resmeth4 with results from Method 4 from SAS - Minitab method */
DATA resmeth4;
   SET temp4;
   calcname='Method 4 - SAS';
   software='Minitab output (=Method 4)';
RUN;

/* Proc Univariate: creates temp5 with results from Method 5 - SAS Method - Default */
PROC UNIVARIATE DATA=indata
                PCTLDEF= 5;
    VAR xvar;
    OUTPUT OUT= temp5
              MEDIAN= median
              Q1=Q1
              Q3=Q3;
RUN;

/* Data step: creates resmeth5 with results from Method 5 from SAS - Default method */
DATA resmeth5;
   SET temp5;
   calcname='Method 5 - SAS';
   software='SAS output (=Default Method 5)';
RUN;

/* Data step: resall contains all results from different methods */
DATA resall;
   SET rescalc resexcel resmeth4 resmeth5;
RUN;

PROC PRINT DATA=resall;
   VAR calcname software Q1 median Q3;
   FORMAT Q1 9.5 median 9.5 Q3 9.5;
RUN;
```

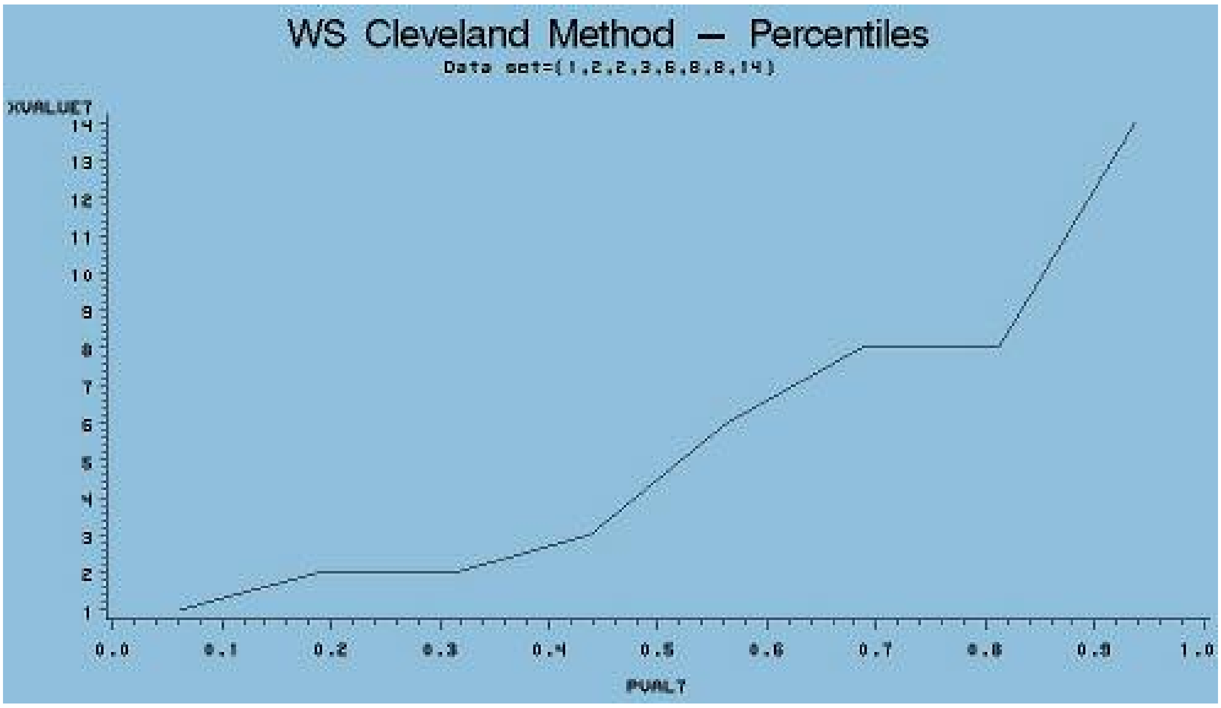
The following table displays the contents of the file myvard.xls as it was when the program was run:

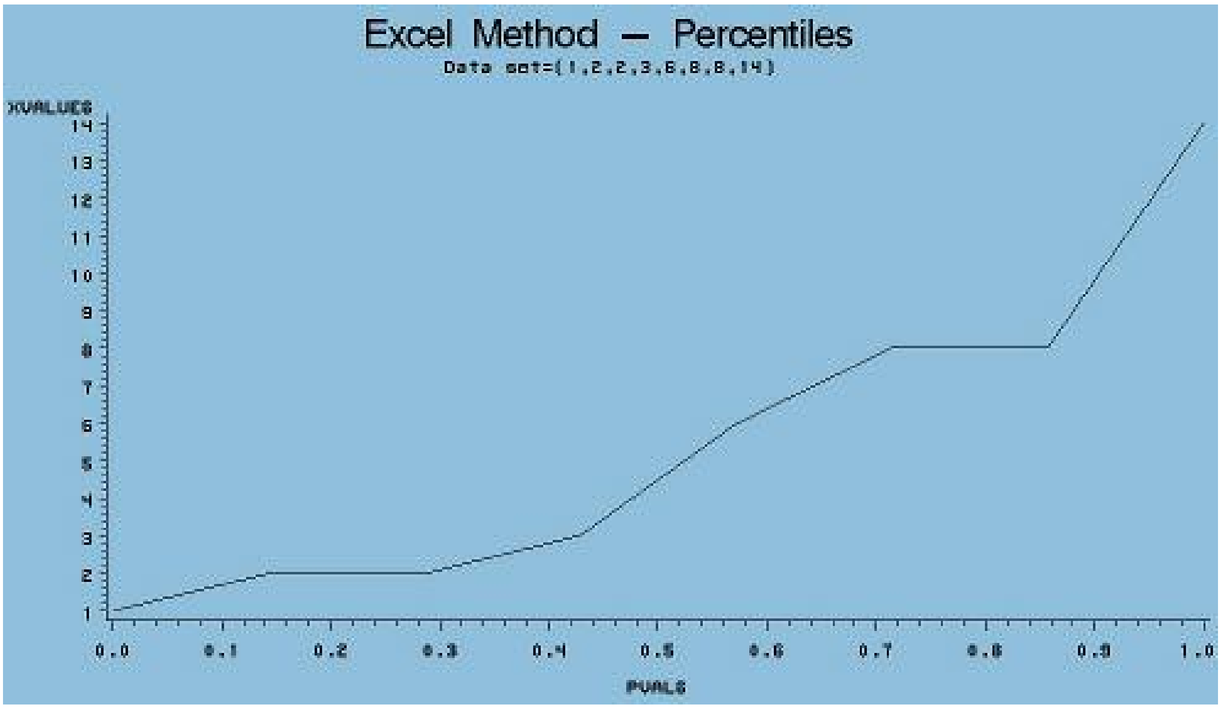
| xvar | Q1exc | Q2exc | Q3exc |
|---|---|---|---|
| 1.00 | 1.75000 | 2.500000 | 3.250000 |
| 2.00 | | | |
| 3.00 | | | |
| 4.00 | | | |

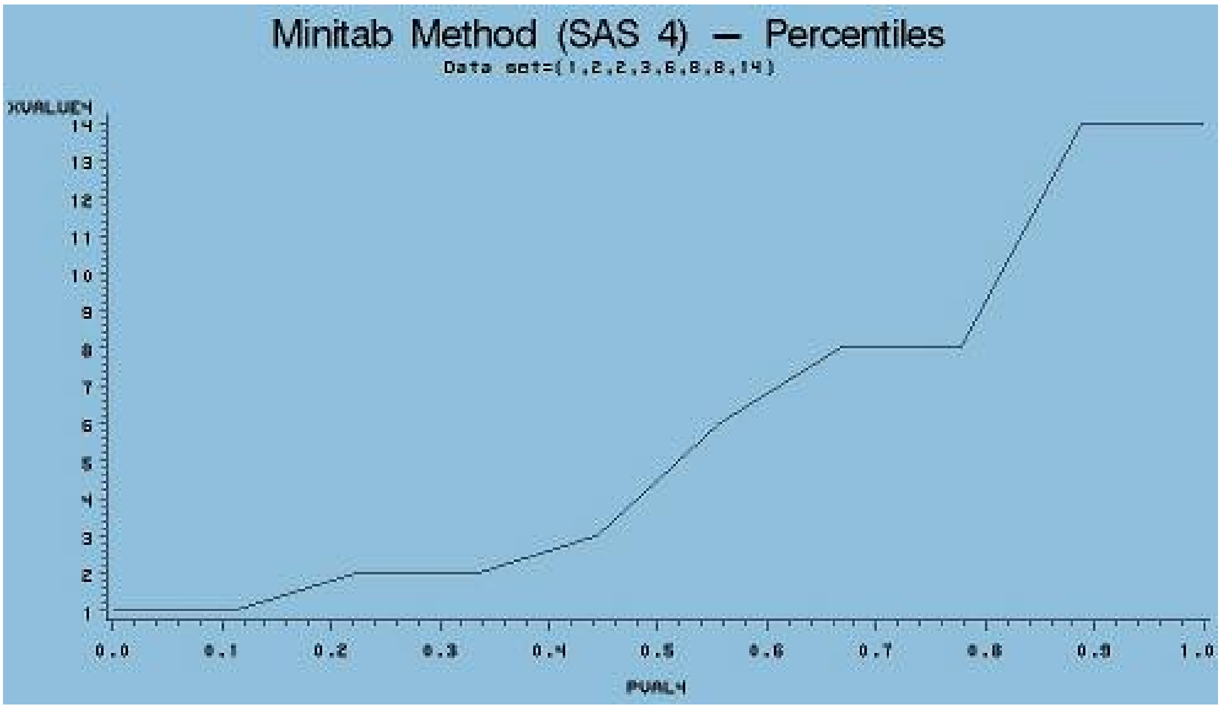
## *2. Example results*

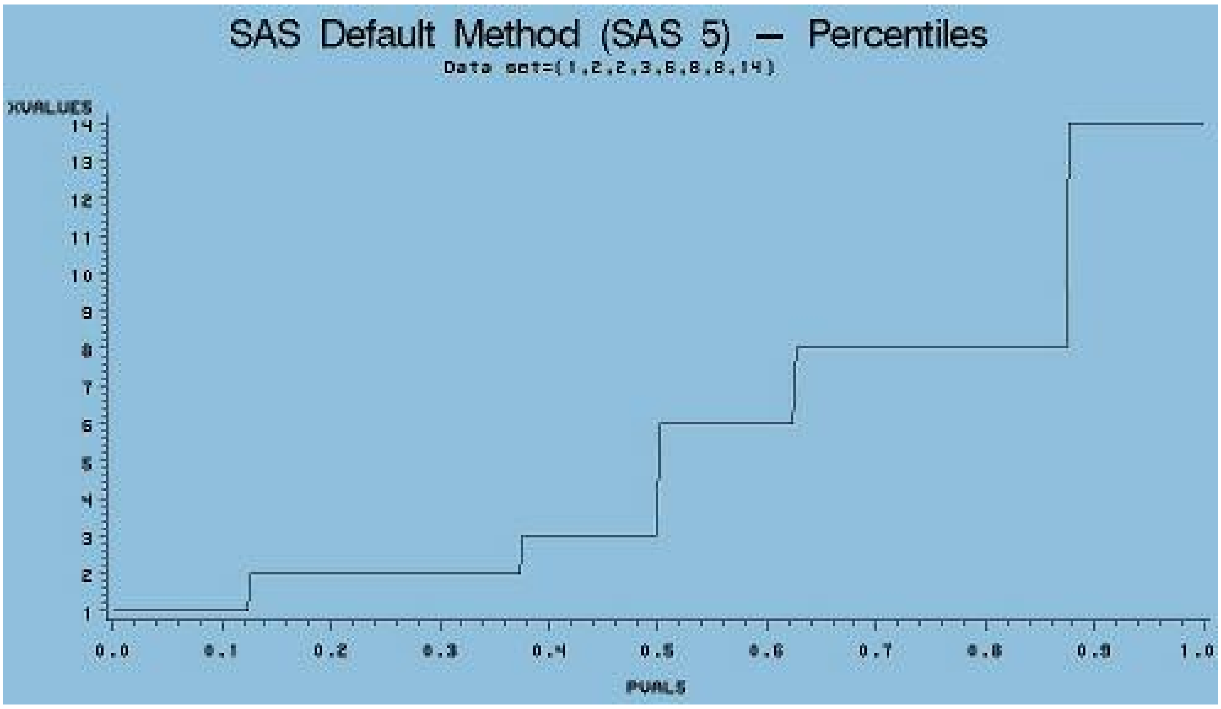
To better illustrate the differences obtained with the different methods, four graphics are copied here. Each of these graphics displays the percentile curve against the p value for the methods discussed. They were created with the data set (1,2,2,3,6,8,8,14).

Please note that for this last example data set, Q1, the median and the Q3 values are the same for all methods (this is because x(2)=x(3) and x(6)=x(7)).


WS Cleveland Method – Percentiles
Data set={1,2,2,3,6,8,8,14}
XVALUE7
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
PVAL7


Excel Method – Percentiles
Data set={1,2,2,3,6,8,8,14}
XVALUE6
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
PVAL6


Minitab Method (SAS 4) — Percentiles
Data set=(1,2,2,3,6,8,8,14)
XVALUE4
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
PVAL4


SAS Default Method (SAS 5) — Percentiles
Data set=(1,2,2,3,6,8,8,14)
XVALUE5
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
PVAL5

## D. Conclusion

### 1. *In summary…*

Although quartiles/percentiles are widely used by scientists, they have to be considered with caution. Indeed, in most of the cases, the values given by a software will constitute only simple indicators of the data distribution. It is only in few cases that one might want to use a specific definition for the quartiles (for example, one of the five different definitions available in SAS).

### 2. *In case of problems*

Please contact your iTSS supporter via your local helpdesk in case you have any problems when using the quartiles/percentiles at your site from any of the statistical software supported by iTSS.
He can be contacted by email at dfj@wpo.nerc.ac.uk