PharmaSUG 2013 - Paper SP08

Oh Quartile, Where Art Thou?

David Franklin, TheProgrammersCabin.com, Litchfield, NH

ABSTRACT

"Why is my first quartile number different from yours?" It was this question that led to a study of methods that can be used for calculating the first and third quartile for a set of data. SAS® provides five methods for calculation of the quartile but other methods exist and are used in other common software packages. In this paper ten methods are discussed from the point of view of a SAS programmer and a macro presented that will calculate first and third quartiles for the five methods not provided in the Base SAS package.

INTRODUCTION

It was a late afternoon discussion that the subject of how a quartile is calculated came up. A colleague was checking some quartile calculations using a well-known spreadsheet package and different results were appearing. So who was right? Had a bug been found in the way SAS calculated quartiles? Rerunning the program selecting the other quartile calculation options did not produce the same answer as the spreadsheet did. So what was going on?

The “school book” definition of Quartile is “any of the three values which divide the sorted data set into four equal parts, so that each part represents 1/4th of the sample“ - the first quartile is 25% of the data denoted usually by Q1, the second quartile is 50% of the data denoted usually by Q2 (also called the Median), and the third quartile is 75% of the data denoted usually by Q3.

Lets look at the following example:

For the data 1, 2, 3, 4, 5, 6, 7 and 8 the following results are calculated for the 25^{th} percentile:
SAS Method 5 (default) = 2.5
SAS Method 4 = 2.25
Excel = 2.75

Unlike the median that has a standard calculation method, there is no one standard for the calculation of the quartile. Base SAS provides five methods but there are others used by different software packages. This paper presents ten definitions that are commonly used, including the five that SAS uses. The discussion will be from the perspective of a SAS programmer so no comment will be made on the merits or demerits of each definition or which definition is best for calculating the statistic.

THE SAS METHODS

There are five methods that SAS uses to calculate the quartile, each of which can be called using various procedure options including the QNTLDEF=value option the MEANS procedure and PCTLDEF=value in the UNIVARIATE procedure. Interestingly SAS uses definition 5 as default so in keeping with SAS the definitions will be presented in reverse order.

Definition 5, Empirical Distribution, Averaging
$y = (x_j - x_{j+1})/2$ if g=0 or $y = x_{j+1}$ if g>0
where n/4=j+g for the LQ and 3n/4=j+g for the UQ

Definition 4, Weighted Average at X(n+1)
$y = (1-g)*x_j + g*x_{j+1}$
where (n+1)/4=j+g for the LQ and 3(n+1)/4=j+g for the UQ

Definition 3, Empirical Distribution Function
$y = x_j$ if g=0, or $y = x_{j+1}$ if g>0
where n/4=j+g for the LQ and 3n/4=j+g for the UQ

Definition 2, Closest Observation
If g=0.5 and j is even then y=xj, else if g=0.5 and j is odd then $y=x_j+1$, else x_j
where (n/4)+0.5=j+g for LQ, (3n/4)+0.5=j+g for UQ

Definition 1, Weighted Average at Xn
$y = (1-g)*x_j + g*x_{j+1}$
where n/4=j+g for the LQ and 3n/4=j+g for the UQ

Only methods 1 and 4 use interpolation to calculate the quartile value.

Using the data 1, 2, 3, 4, 5, 6, 7 and 8 the following first and third quartiles are computed:

Method 5: Q1=2.5 Q3=6.5
Method 4: Q1=2.25 Q3=6.75
Method 3: Q1=2 Q3=6
Method 2: Q1=2 Q3=6
Method 1: Q1=2 Q3=6

THE CLASSIC METHOD

Developed by Tukey, its aim is to find the quartiles of a set of data with little or no calculation. The procedure for finding the quartile is:

1. find the median
2. if an odd number of observations include the sample median value and then find the median of the subset, else if an even number of observations exclude the sample median value then find the median of the subset.

Putting this into a formula looks something like:

LQ: if n is odd, (n+3)/4=j+g; else (n+2)/4=j+g
UQ: if n is odd, (3n+1)/4=j+g; else (3n+2)/4=j+g
$y=x_j$

Using the data 1,2,3,4,5,6,7,8, the sample median is 4.5; the Lower Quartile is from the set {1,2,3,4} so the value is 2.5; and the Upper Quartile is from the set {5,6,7,8} so the value is 6.5.

There is an adaptation to the Tukey method presented by Moore and McCabe that does not include sample median in the quartile calculation in both cases where there is an even or odd number of observations. Putting this into a formula looks something like:

LQ: if n is odd, (n+1)/4=j+g; else (n+2)/4=j+g
UQ: if n is odd, (3n+3)/4=j+g; else (3n+2)/4=j+g
$y=x_j$

OTHER METHODS

Three methods will be discussed here - the first two involve some sort of interpolation while the third does not. There are other methods but these are rarely used.

The first is a method that very few packages use - the only one found with documentation relating to this method is 'R':

$y = (1-g)*x_j + g*x_{j+1}$
where $(n+2)/4=j+g$ for LQ and $(3n+2)/4=j+g$ for UQ

A paper written by Hyndman an Fan identified this method as the only one that met all of the six requirements they were considering for the calculation of a quartile. However this method is one of the least used methods in commercial software.

Some packages, including Excel, after version 2002, and S-Plus, use yet another calculation presented by Freund and Perles:

$y = (1-g)*x_j + g*x_{j+1}$
where $(n+3)/4=j+g$ for LQ and $(3n+1)/4=j+g$ for UQ

Prior to Excel version 2002, Excel used the method:

$y=x_j+g(x_j+1 -x_j)$
where $((n-1)/4)+1=j+g$ for LQ and $(3(n-1)/4)+1=j+g$ for UQ

There is a third method called presented by Mendenhall and Sincich that is a variation on the Closest Observation method (Definition 2 in SAS):

LQ: if $g<0.5$ then $y=x_j$, else $y=x_j+1$
UQ: if $g<=0.5$ then $y=x_j$, else $y=x_j+1$
where $(n+1)/4=j+g$ for the LQ and $(3n+2)/4=j+g$ for the UQ

THE MACRO

The five methods that SAS does not use for the Quartile calculation are calculated in the macro listed in the Appendix. The methods used in the macro are:

- Tukey
- Moore and McCabe
- Hyndman and Fan
- Freund and Perles
- Mendenhall and Sincich

CONCLUSION

Going back to the original question at the start of the paper, “Why is my first quartile number different from yours?“, it can be seen that there are different calculation methods for the quartile - all definitions have their place and the selection of which definition to use does depend on the knowledge and experience of a statistician. For the question itself, SAS Definition 5 was being used for the calculation and the check was being done using Excel - they do not have similar definitions so the results were different. The moral of this story is that you should know how your software calculates a statistic before blindly reporting the result.

REFERENCES

Hyndman, R. J. and Fan, Y. Sample Quantiles in Statistical Packages. Amer. Stat. 50, 361-365, 1996.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: David Franklin
Enterprise: TheProgrammersCabin.com
E-mail: dfranklin@TheProgrammersCabin.com
Web: http://www.TheProgrammersCabin.com
Twitter: ThePgmrsCabin

Appendix - Listing of SAS Macro to Calculated Quartiles not Provided by SAS

```
%macro mkquart(dsin=, /*Dataset where data to be analyzed is stored*/
               var=, /*Variable to be analyzed*/
               dsout= /*Dataset with results*/
               );

    *Find number of non-missing values;
    data _null_;
        retain _k 0;
        set &dsin end=eof;
        if ^missing(&var) then _k+1;
        if eof then call symput('xobs',compress(put(_k,8.)));
    run;

    *Load dataset into an array;
    data &dsout;
        attrib TK_LQ length=8 label='Tukey, First Quartile'
               TK_UQ length=8 label='Tukey, Third Quartile'
               MM_LQ length=8 label='Moore and McCabe, First Quartile'
               MM_UQ length=8 label='Moore and McCabe, Third Quartile'
               HF_LQ length=8 label='Hyndman and Fan, First Quartile'
               HF_UQ length=8 label='Hyndman and Fan, Third Quartile'
               FP_LQ length=8 label='Freund and Perles, First Quartile'
               FP_UQ length=8 label='Freund and Perles, Third Quartile'
               MS_LQ length=8 label='Mendenhall and Sincich, First Quartile'
               MS_UQ length=8 label='Mendenhall and Sincich, Third Quartile';
        set &dsin (where=(^missing(&var))) end=eof;
        array xx{&xobs} _temporary_;
        keep TK_LQ TK_UQ MM_LQ MM_UQ HF_LQ HF_UQ FP_LQ FP_UQ MS_LQ MS_UQ;
        xx{_n_}=&var;
        if eof then do;

            **Tukey;
            if mod(dim(xx),2)=1 then do;
                TK_LQpos=(dim(xx)+3)/4;
                TK_UQpos=(3*dim(xx)+1)/4;
            end;
            else do;
                TK_LQpos=(dim(xx)+2)/4;
                TK_UQpos=(3*dim(xx)+2)/4;
            end;
```

```
            TK_LQ=((1-(TK_LQpos-floor(TK_LQpos)))*xx{floor(TK_LQpos)}) +
                  ((TK_LQpos-floor(TK_LQpos))*xx{floor(TK_LQpos)+1});
            TK_UQ=((1-(TK_UQpos-floor(TK_UQpos)))*xx{floor(TK_UQpos)}) +
                  ((TK_UQpos-floor(TK_UQpos))*xx{floor(TK_UQpos)+1});

            **Moore and McCabe;
            if mod(x,2)=1 then do;
                MM_LQpos=(dim(xx)+1)/4;
                MM_UQpos=(3*dim(xx)+3)/4;
            end;
            else do;
                MM_LQpos=(dim(xx)+2)/4;
                MM_UQpos=(3*dim(xx)+2)/4;
            end;
            MM_LQ=((1-(MM_LQpos-floor(MM_LQpos)))*xx{floor(MM_LQpos)}) +
                  ((MM_LQpos-floor(MM_LQpos))*xx{floor(MM_LQpos)+1});
            MM_UQ=((1-(MM_UQpos-floor(MM_UQpos)))*xx{floor(MM_UQpos)}) +
                  ((MM_UQpos-floor(MM_UQpos))*xx{floor(MM_UQpos)+1});

            **Hyndman and Fan;
            HF_LQpos=(dim(xx)+2)/4;
            HF_LQ=((1-(HF_LQpos-floor(HF_LQpos)))*xx{floor(HF_LQpos)}) +
                  ((HF_LQpos-floor(HF_LQpos))*xx{floor(HF_LQpos)+1});
            HF_UQpos=((3*dim(xx))+2)/4;
            HF_UQ=((1-(HF_UQpos-floor(HF_UQpos)))*xx{floor(HF_UQpos)}) +
                  ((HF_UQpos-floor(HF_UQpos))*xx{floor(HF_UQpos)+1});

            **Freund and Perles;
            FP_LQpos=(dim(xx)+3)/4;
            FP_LQ=((1-(FP_LQpos-floor(FP_LQpos)))*xx{floor(FP_LQpos)}) +
                  ((FP_LQpos-floor(FP_LQpos))*xx{floor(FP_LQpos)+1});
            FP_UQpos=((3*dim(xx))+1)/4;
            FP_UQ=((1-(FP_UQpos-floor(FP_UQpos)))*xx{floor(FP_UQpos)}) +
               ((FP_UQpos-floor(FP_UQpos))*xx{floor(FP_UQpos)+1});

            **Mendenhall and Sincich;
            MS_LQpos=(dim(xx)+1)/4;
            if (MS_LQpos-floor(MS_LQpos))<0.5 then
              MS_LQ=xx{floor(MS_LQpos)};
            else MS_LQ=xx{floor(MS_LQpos)+1};
            MS_UQpos=((3*dim(xx))+2)/4;
            if (MS_UQpos-floor(MS_UQpos))<=0.5 then
              MS_UQ=xx{floor(MS_UQpos)};
            else MS_UQ=xx{floor(MS_UQpos)+1};
        end;
    run;

%mend mkquart;
```