PhUSE 2012

Paper PP16

Quartiles within SAS

Jorine Putter, Quanticate, Oxford, United Kingdom
Liza Faber, Quanticate, Bloemfontein, South Africa

ABSTRACT

Many times during the reporting of a study, programmers blindly report whichever statistics are generated by default by the specific SAS® procedure, often without reading up on the SAS documentation and thus knowing the default behaviour of the procedure.

Unbeknownst to the programmer, the reported statistics may not accurately reflect what the statistician is expecting to see.

The purpose of this paper is to educate programmers on the different methods for calculating Q1 and Q3, and ensuring the statistician has clearly documented the appropriate method to use.

The paper will explore the different methods used by specific SAS procedures, highlight any differences between the procedures, and also demonstrate how these methods can be set within specific SAS procedures. Comparisons with other software, such as Excel will also be included.

INTRODUCTION

To understand the different methods used to calculate quartiles, you first need to understand what a quartile is. When arranging a data set of values in ascending order and divide them into four roughly equal parts, these four parts are called quartiles.

The upper quartile is the part that contains the highest values, the upper middle quartile is the part that contains the next-highest data values, the lower quartile is the part that contains the lowest data values and the lower middle quartile is the part that contains the next-lowest data values.

Similarly the upper quartile can also be the cut-off value (referred to as Q3 or UQ) between the upper middle quartile values and upper quartile values, and the lower quartile can also be the cut-off value (referred to as Q1 or LQ) between the lower middle quartile and lower quartile values.

Definition according to SAS[1] : ''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.''

The concept is simple; however the execution can be much more complicated.

Throughout this paper, we will look at different statistical methods of determining the Q1 and Q3 cut-off values, making use of examples through SAS and other statistical software packages available.

[1] http://support.sas.com/documentation/cdl/en/lab/60951/HTML/default/viewer.htm#sumstats_percent.htm

SOME STATISTICAL METHODS

Name	***John Tukey***	***Moore & McCabe (M&M)***	***Mendenhall & Sincich (M&S)***	***Minitab***	***Freud & Perles (Excel)***
Method	• Find the median of the entire set • If the number of data points in the entire set is odd, find the median of the lower and upper halves by including the overall median as a data point in both halves • If the number of data points in the entire set is even, find the median of the lower and upper halves by excluding the overall median as a data point in both halves	• Find the median of the entire set • Find the median of the lower and upper halves, excluding the overall median as a data point in both halves	• L = (¼)(n+1), rounded to the nearest integer value (however if L falls exactly halfway between to integers, round up) • U = (¾)(n+1), rounded to the nearest integer value (however if U falls exactly halfway between to integers, round down) Q1 = L^{th} element in the entire set Q3 = U^{th} element in the entire set	• L = (¼)(n+1), instead of rounding to the nearest integer, use linear interpolation between the closest data points • U = (¾)(n+1), instead of rounding to the nearest integer, use linear interpolation between the closest data points Q1 = L^{th} element in the entire set Q3 = U^{th} element in the entire set	• L = (¼)(n+3), instead of rounding to the nearest integer, use linear interpolation between the closest data points • U = (¼)(3n+1), instead of rounding to the nearest integer, use linear interpolation between the closest data points Q1 = L^{th} element in the entire set Q3 = U^{th} element in the entire set
Advantages	• Can be calculated without the aid of calculating machinery, as Q1 and Q3 will always either be one of the data points or exactly halfway between two data points	• Can be calculated without the aid of calculating machinery, as Q1 and Q3 will always either be one of the data points or exactly halfway between two data points	• Can be calculated without the aid of calculating machinery, as Q1 and Q3 will always either be one of the data points or exactly halfway between two data points		
Disadvantages	• This method is not extendable to quintiles or percentiles • They tend to have a high bias as they then tend to vary more and they are not good predictions of quartile values of the entire set	• This method is not extendable to quintiles or percentiles • They tend to have a high bias as they then tend to vary more and they are not good predictions of quartile values of the entire set		• Can only be calculated with the aid of calculating machinery, as Q1 and Q3 will not always either be one of the data points or exactly halfway between two data points	• Can only be calculated with the aid of calculating machinery, as Q1 and Q3 will not always either be one of the data points or exactly halfway between two data points

SAS METHODS

You use the QNTLDEF(PCTLDEF) option to set the method used by the SAS procedure to compute quartiles. The default method used by SAS, is QNTLDEF=5.

Table 1 – Summary of the 5 different methods to calculate quartiles[2]

QNTLDEF=1	Weighted average at x_{np} where x_0 is taken to be x_1	$np = j + g$	$y = (1 - g)x_j + gx_{j+1}$
QNTLDEF=2	Observation numbered closest to np where i is the integer part of np + ½	np = j + g	$y = x_i$ if $g \neq ½$ $y = x_j$ if $g = ½$ and j is even $y = x_j + 1$ if $g = ½$ and j is odd
QNTLDEF=3	Empirical distribution function	np = j + g	$y = x_j$ if $g = 0$ $y = x_j + 1$ if $g > 0$
QNTLDEF=4	Weighted average aimed at x (n+1) p where xn+1 is taken to be xn	(n+1)p = j + g	$y = (1 - g)x_j + gx_{j+1}$
QNTLDEF=5	Empirical distribution function with averaging	$np = j + g$	$y = (x_j + x_{j+1})/2$ if $g = 0$ $y = x_{j+1}$ if $g > 0$

n = number of observations in data set *p = percentile value (e.g 0.25 for 25th percentile)*
j = integer part of np *g = fractional or decimal part of np*

We will now look at two methods, QNTLDEF=4 and QNTLEF=5, as these are mostly used in SAS.
Example data set -- 1, 2, 3, 4, 5

1. METHOD: QNTLDEF=5

Calculating Q1 (25th percentile)
Firstly calculate *np*.
n = 5 and *p* = 0.25. So np = 5*0.25 = 1.25, thus *j* = 1 and *g* = 0.25.
Secondly calculate *y*. By referencing Table 1 and since $g > 0$, $y = x_{j+1}$, as *j*=1, Q1 is the 2nd (1+1) element in the data set.
Answer: Q1 = 2.

Calculating Q3 (75th percentile)
Firstly calculate *np*.
n = 5 and *p* = 0.75. So np = 5*0.75 = 3.75, thus *j* = 3 and *g* = 0.75.
Secondly calculate *y*. By referencing Table 1 and since $g > 0$, $y = x_{j+1}$, as *j*=3, Q3 is the 4th (3+1) element in the data set.
Answer: Q3 = 4.

2. METHOD: QNTLDEF=4

Calculating Q1 (25th percentile)
Firstly calculate (n+1)p.
n = 5 and p = 0.25. So (n+1)p = 6*0.25 = 1.5, thus j = 1 and g = 0.5.
Secondly calculate y. By referencing Table 1, use the formula $y = (1 - g)x_j + gx_{j+1}$, so (1-0.5)*1 + 0.5*2 = 1.5
Answer: Q1 = 1.5

Calculating Q3 (75th percentile)
Firstly calculate (n+1)p.
n = 5 and p = 0.75. So (n+1)p = 6*0.75 = 4.5, thus j = 4 and g = 0.5.
Secondly calculate y. By referencing Table 1, use the formula $y = (1 - g)x_j + gx_{j+1}$, (1-0.5)*4 + 0.5*5 = 4.5
Answer: Q3 = 4.5

PROCEDURES WITHIN SAS WITH THE ABILITY TO CALCULATE Q1 AND Q3

Procedures available in SAS with which you can calculate Q1 and Q3 values:

- Proc Means and Proc Summary
- Proc Univariate
- Proc Boxplot

[2] http://v8doc.sas.com/sashtml/stat/chap18/sect12.htm

- Proc Stdize
- Proc Capability
- Proc Tabulate
- Proc Report

SAS EXAMPLES AND OUTPUT

Following are some examples of quartiles calculated with different SAS procedures. In each case, the same example dataset (dataset1) has been used: $1, 2, 3, 4, 5, 6, 7, 8, 9$

1. PROC MEANS

PROC MEANS – QNTLDEF=1

```
proc means data = dataset1
           qntldef=1
           n median q1 q3;
  var value;
run;
```

N	Median	Lower Quartile	Upper Quartile
9	4.5000000	2.2500000	6.7500000

PROC MEANS -- QNTLDEF=2

```
proc means data = dataset1
           qntldef=2
           n median q1 q3;
  var value;
run;
```

N	Median	Lower Quartile	Upper Quartile
9	4.0000000	2.0000000	7.0000000

PROC MEANS -- QNTLDEF=3

```
proc means data = dataset1
           qntldef=3
           n median q1 q3;
  var value;
run;
```

N	Median	Lower Quartile	Upper Quartile
9	5.0000000	3.0000000	7.0000000

PROC MEANS -- QNTLDEF=4

```
proc means data = dataset1
           qntldef=4
           n median q1 q3;
  var value;
run;
```

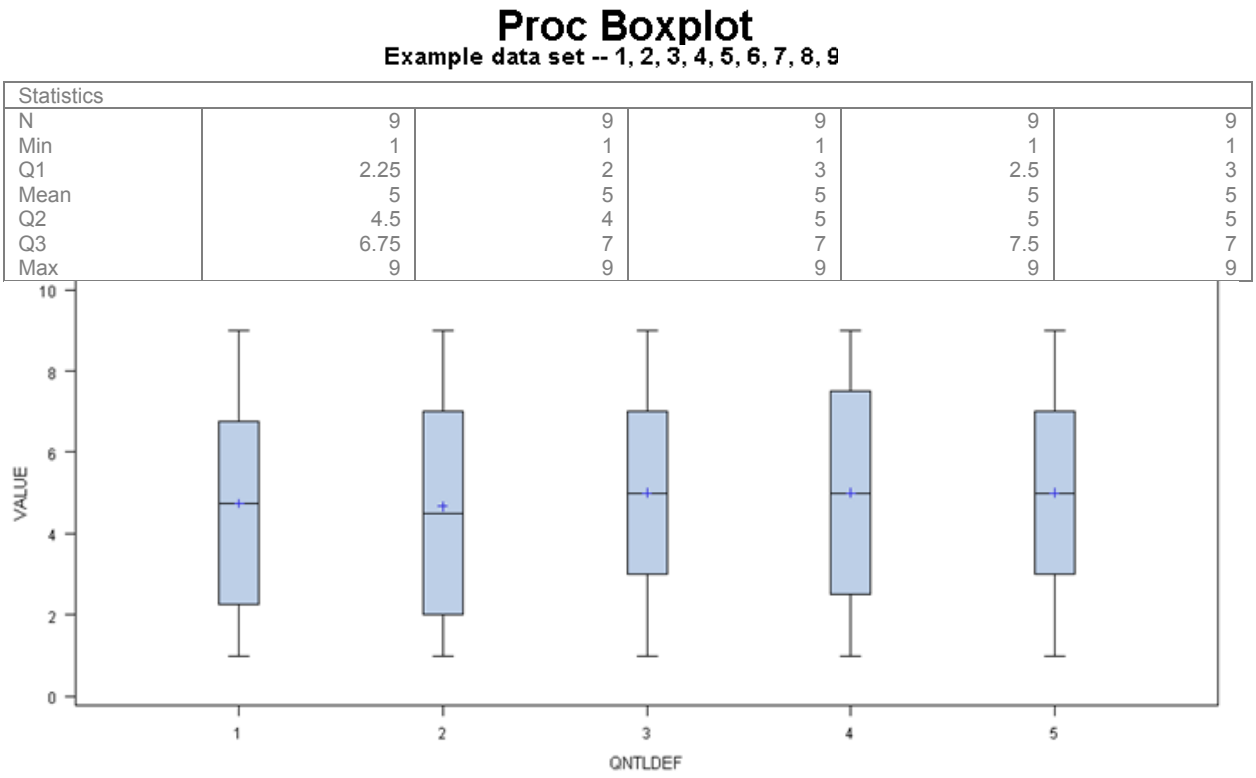
N	Median	Lower Quartile	Upper Quartile
9	5.0000000	2.5000000	7.5000000

PROC MEANS -- QNTLDEF=5 (BY DEFAULT)

```
proc means data = dataset1
          qntldef=5
          n median q1 q3;
  var value;
run;
```

N	Median	Lower Quartile	Upper Quartile
9	5.0000000	3.0000000	7.0000000

2. PROC BOXPLOT



3. PROC REPORT

PROC REPORT -- QNTLDEF=1

```
proc report data = dataset1 nowd qntldef=1;
  column value value1 value2;
  define value  / median format=8.2 'Median';
  define value1 / q1 format=8.2 'Q1';
  define value2 / q3 format=8.2 'Q3';
run;
    Median        Q1        Q3
------------------------------------------------------------------------
      4.50      2.25      6.75
------------------------------------------------------------------------
```

PROC REPORT -- QNTLDEF=2

```
proc report data = dataset1 nowd qntldef=2;
  column value value1 value2;
  define value  / median format=8.2 'Median';
  define value1 / q1 format=8.2 'Q1';
  define value2 / q3 format=8.2 'Q3';
run;
    Median        Q1        Q3
------------------------------------------------------------------------
      4.00      2.00      7.00
------------------------------------------------------------------------
```

PROC REPORT -- QNTLDEF=3

```
proc report data = dataset1 nowd qntldef=3;
  column value value1 value2;
  define value  / median format=8.2 'Median';
  define value1 / q1 format=8.2 'Q1';
  define value2 / q3 format=8.2 'Q3';
run;
    Median        Q1        Q3
------------------------------------------------------------------------
      5.00      3.00      7.00
------------------------------------------------------------------------
```

PROC REPORT -- QNTLDEF=4

```
proc report data = dataset1 nowd qntldef=4;
  column value value1 value2;
  define value  / median format=8.2 'Median';
  define value1 / q1 format=8.2 'Q1';
  define value2 / q3 format=8.2 'Q3';
run;
    Median        Q1        Q3
-------------------------------------------------------------------------
      5.00      2.50      7.50
-------------------------------------------------------------------------
```

PROC REPORT -- QNTLDEF=5 (BY DEFAULT)

```
proc report data = dataset1 nowd qntldef=5;
  column value value1 value2;
  define value  / median format=8.2 'Median';
  define value1 / q1 format=8.2 'Q1';
  define value2 / q3 format=8.2 'Q3';
run;
    Median        Q1        Q3
-------------------------------------------------------------------------
      5.00      3.00      7.00
-------------------------------------------------------------------------
```

Table 2 – Summary of Median, Q1 and Q3 values of the above example

Example data set -- 1, 2, 3, 4, 5, 6, 7, 8, 9

	Median (Q2)	***Q1***	***Q3***
QNTLDEF=1 *(Weighted Average at Xn)*	4.5	*2.25*	*6.75*
QNTLDEF=2 *(Closest Observation)*	4	2	7
QNTLDEF=3 *(Empirical Distribution Function)*	5	3	7
QNTLDEF=4 *(Weighted Average at X(n+1))*	5	2.5	7.5
QNTLDEF=5 *(Empirical Distribution, Averaging)*	5	*3*	*7*

OTHER STATISTICAL SOFTWARE EXAMPLES AND OUTPUT

1. R EXAMPLES AND OUTPUT

R has nine different methods to calculate quartiles. The default method is type 7, which is equivalent to Excel.

Table 3 – R Summary of Median, Q1 and Q3

```
x=c(1,2,3,4,5,6,7,8,9)
quantile(x,type=T), where T is type 1 through 9
```

Type	***Median (Q2)***	***Q1***	***Q3***
Type 1 *– empirical distribution function (equivalent to SAS QNTLDEF = 3)*	5	*3*	*7*
Type 2 *– averaging empirical distribution function (equivalent to SAS QNTLDEF = 5)*	5	3	7
Type 3 – *nearest even order statistic (equivalent to SAS QNTLDEF = 2)*	4	2	7
Type 4 – *linear interpolation of empirical cdf (equivalent to SAS QNTLDEF = 1)*	4.5	2.25	6.75
Type 5 *– piecewise linear function where the knots are the values midway through the steps of the empirical cdf*	5	*2.75*	*7.25*
Type 6 *– used by Minitab and by SPSS (equivalent to SAS QNTLDEF = 4)*	5	*2.5*	*7.5*
Type 7 *– used by S and is the default for R (equivalent to Excel)*	5	*3*	*7*
Type 8 *– are approximately median-unbiased regardless of the distribution of X*	5	*2.27*	*7.33*
Type 9 *– are approximately unbiased for the expected order statistics if X is normally distributed*	5	*2.69*	*7.31*

For further information on R, please see R help files.

2. EXCEL EXAMPLES AND OUTPUT

Table 4 – Excel functions and results for Median, Q1 and Q3

	Function	*Result*
Median	MEDIAN(1,2,3,4,5,6,7,8,9)	5
Q1	PERCENTILE({1,2,3,4,5,6,7,8,9},0.25)	3
Q3	PERCENTILE({1,2,3,4,5,6,7,8,9},0.75)	7

For further information on Excel, please see Excel help files.

CONCLUSION

This paper has demonstrated the different methods employed by SAS to calculate quartiles, as well as the different SAS procedures available to do so. The paper also illustrated that different software packages will produce different results. It is up to the programmer, in conjunction with the statistician to be aware of the different approaches used by SAS, and to ensure that the best approach is used.

REFERENCES

1. Dr. Twe (2002), Reply to "Tom" about quartiles, [http://mathforum.org/library/drmath/view/60969.html].
2. Journet, D. (1999), "Quartiles: How to Calculate Them?", [http://www.haiweb.org/medicineprices/manual/quartiles_iTSS.pdf].
3. Franklin, D. (2007), "Calculating the Quartile (or why are my Quartile answers different?)", [http://www.nesug.org/proceedings/nesug07/po/po08.pdf].
4. Pincon, B. (designed form the R's help page), "Quantile", [http://www.iecn.u-nancy.fr/~pincon/nsp/nsp_manual/manualli159.html].

ACKNOWLEDGMENTS

None

RECOMMENDED READING

None

CONTACT INFORMATION

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

Jorine Putter
Quanticate
Bevan House
Bancroft Court
Bancroft
HITCHIN
Herts
United Kingdom
SG5 1LH
Work Phone: +44 (0)1865 823811
Email: jorine.putter@quanticate.com
Web: http://www.quanticate.com

Liza Faber
Quanticate
Bevan House
Bancroft Court
Bancroft
HITCHIN
Herts
United Kingdom
SG5 1LH
Work Phone: +27 (0)51 4104621
Email: liza.faber@quanticate.com
Web: http://www.quanticate.com