

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.

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.

QNTLDEF=1						
Weighted average at x_{np} where x_o is taken to be x_i				$np = j + g$	$y = (1 - g)x_j + gx_{j+1}$	

QNTLDEF=1	n	p	np	j	g	y
Q1	9	0.25	2.25	2	0.25	$(1 - 0.25) \cdot 2 + 0.25 \cdot (2 + 1) = 2.25$
Q3	9	0.75	6.75	6	0.75	$(1 - 0.75) \cdot 6 + 0.75 \cdot (6 + 1) = 6.75$

QNTLDEF=2						
Observation numbered closest to np where i is the integer part of $np + \frac{1}{2}$				$np = j + g$	$y = x_j$ if $g \neq \frac{1}{2}$ $y = x_j$ if $g = \frac{1}{2}$ and j is even $y = x_j + 1$ if $g = \frac{1}{2}$ and j is odd	

QNTLDEF=2	n	p	np	j	g	y
Q1	9	0.25	2.25	2	0.25	$\text{Int}(2.25 + 0.5) = 2$
Q3	9	0.75	6.75	6	0.75	$\text{Int}(6.75 + 0.5) = 7$

QNTLDEF=3						
Empirical distribution function				$np = j + g$	$y = x_j$ if $g = 0$ $y = x_{j+1}$ if $g > 0$	

QNTLDEF=3	n	p	np	j	g	y
Q1	9	0.25	2.25	2	0.25	$(2 + 1) = 3$
Q3	9	0.75	6.75	6	0.75	$(6 + 1) = 7$

QNTLDEF=4						
Weighted average aimed at x (n+1) p where x_{n+1} is taken to be x_n				$(n+1)p = j + g$	$y = (1 - g)x_j + gx_{j+1}$	

QNTLDEF=4	n	p	(n+1)p	j	g	y
Q1	9	0.25	2.5	2	0.5	$(1 - 0.5) \cdot 2 + 0.5 \cdot (2 + 1) = 2.5$
Q3	9	0.75	7.5	7	0.5	$(1 - 0.5) \cdot 7 + 0.5 \cdot (7 + 1) = 7.5$

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 $g > 0$	

QNTLDEF=5	n	p	np	j	g	y
Q1	9	0.25	2.25	2	0.25	$(2 + 1) = 3$
Q3	9	0.75	6.75	6	0.75	$(6 + 1) = 7$

SAS EXAMPLES AND OUTPUT

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 }

PROC MEANS AND PROC REPORT		
QNTLDEF=1		
<pre>proc means data = dataset1 qntldef=1 n median q1 q3; var value; run;</pre>		<pre>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;</pre>
Median	Lower Quartile	Upper Quartile
4.5000000	2.2500000	6.7500000

QNTLDEF=2		
<pre>proc means data = dataset1 qntldef=2 n median q1 q3; var value; run;</pre>		<pre>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;</pre>
Median	Lower Quartile	Upper Quartile
4.0000000	2.0000000	7.0000000

QNTLDEF=3		
<pre>proc means data = dataset1 qntldef=3 n median q1 q3; var value; run;</pre>		<pre>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;</pre>
Median	Lower Quartile	Upper Quartile
5.0000000	3.0000000	7.0000000

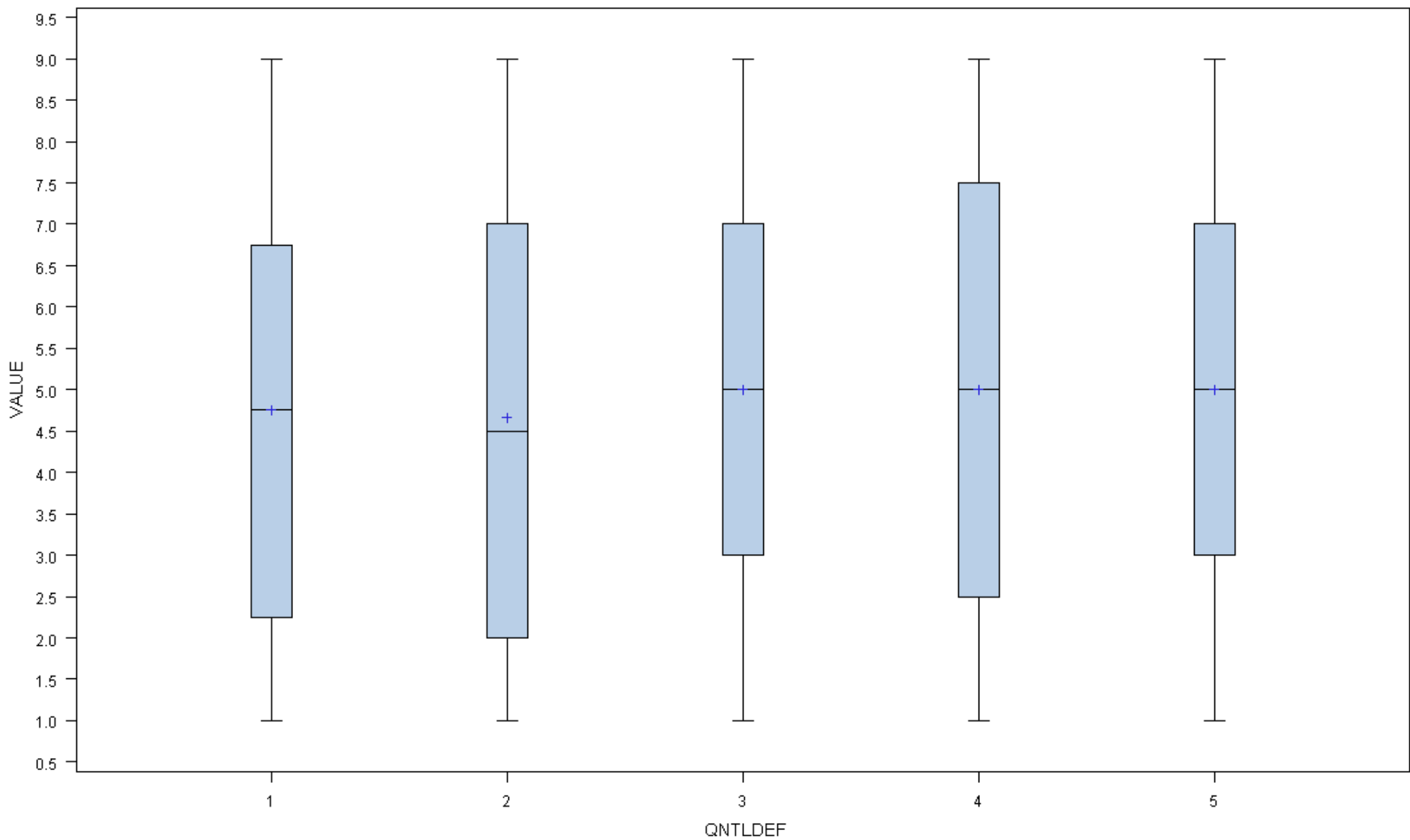
QNTLDEF=4		
<pre>proc means data = dataset1 qntldef=4 n median q1 q3; var value; run;</pre>		<pre>proc report data = dataset1 nowd qntldef=4; column value value1 value2; define value / median format=8.2 'Median'; define value1 / q1 format=8.2 'Q1'; define value2 / q3 format=8.2 'Q3'; run;</pre>
Median	Lower Quartile	Upper Quartile
5.0000000	2.5000000	7.5000000

QNTLDEF=5		
<pre>proc means data = dataset1 qntldef=5 n median q1 q3; var value; run;</pre>		<pre>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;</pre>
Median	Lower Quartile	Upper Quartile
5.0000000	3.0000000	7.0000000

PROC BOXPLOT

```
proc boxplot data = example;
  plot variable1 * variable2 / pctldef = quartile;
run;
```

STATISTICS					
N	9	9	9	9	9
Mean	5	5	5	5	5
Min	1	1	1	1	1
Max	9	9	9	9	9
Q1	2.25	2	3	2.5	3
Q2	4.5	4	5	5	5
Q3	6.75	7	7	7.5	7



OTHER STATISTICAL SOFTWARE EXAMPLES AND OUTPUT

Microsoft Office Excel

Microsoft Excel - Book1				
	A	B	C	D
1	Data set			
2	1			
3	2			
4	3			
5	4			
6	5			
7	6			
8	7			
9	8			
10	9			
11	Statistic	Formula	Result	
12	Median	=MEDIAN(A2:A10)	5	
13	Q1	=QUARTILE(A2:A10,1)	3	
14		=PERCENTILE(A2:A10,0.25)	3	
15	Q3	=QUARTILE(A2:A10,3)	7	
16		=PERCENTILE(A2:A10,0.75)	7	
17				
18				

R Statistical Software (v2.15.0)

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

```
RGui (32-bit) - [R Console]
R version 2.15.0 (2012-03-30)
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-pc-mingw32/i386 (32-bit)

> x=c(1,2,3,4,5,6,7,8,9)
> quantile(x,type=1)
 0% 25% 50% 75% 100%
 1  3  5  7  9
> quantile(x,type=2)
 0% 25% 50% 75% 100%
 1  3  5  7  9
> quantile(x,type=3)
 0% 25% 50% 75% 100%
 1  2  4  7  9
> quantile(x,type=4)
 0% 25% 50% 75% 100%
1.00 2.25 4.50 6.75 9.00
> quantile(x,type=5)
 0% 25% 50% 75% 100%
1.00 2.75 5.00 7.25 9.00
> quantile(x,type=6)
 0% 25% 50% 75% 100%
1.0 2.5 5.0 7.5 9.0
> quantile(x,type=7)
 0% 25% 50% 75% 100%
 1  3  5  7  9
> quantile(x,type=8)
 0% 25% 50% 75% 100%
1.000000 2.666667 5.000000 7.333333 9.000000
> quantile(x,type=9)
 0% 25% 50% 75% 100%
1.0000 2.6875 5.0000 7.3125 9.0000
>
```

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.

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