Quartile Conundrum

Patricia Guldin, Merck Research Labs, Merck & Co., Inc., Upper Gwynedd, PA
Liping Zhang, Merck Research Labs, Merck & Co., Inc., Upper Gwynedd, PA

ABSTRACT

Calculating quartiles is a common task that can be accomplished in multiple ways, but not all calculations return the same results for the same data. There are many definitions of a quartile as well as multiple descriptive terms (percentile, quartile). The challenge for a programmer is in determining which code to use. This paper will offer, to both new and seasoned programmers, the awareness that understanding what you are being asked to provide regarding quartiles is just as important as knowing how to provide it.

INTRODUCTION

Have you ever had to calculate quartiles and didn't know where to start? You may not be aware that there are many definitions of a quartile, that several methods can be used to compute them, and that the different methods may not provide the same results.

The first and third quartiles are often referred to as the 25th and 75th percentiles and the median is often referred to as the 2nd quartile or 50th percentile.

The SAS® definition of quartiles and percentiles states that: "Percentiles, including quantiles, quartiles, and the median, are useful for a detailed study of a distribution. For a set of measurements arranged in order of magnitude, the pth percentile is the value that has p percent of the measurements below it and (100-p) percent above it. The median is the 50th percentile." "The upper quartile of a distribution is the value below which 75 percent of the measurements fall (the 75th percentile). Twenty-five percent of the measurements fall below the lower quartile value."

This paper will focus on the ways SAS calculates quartiles and the code that should be used to provide quartiles. SAS calculates quartiles the same way it calculates percentiles so we sometimes use these terms interchangeably.

METHODS OF CALCULATING QUARTILES

There are many ways of calculating percentiles/quartiles and different software may use different methods. However, this paper will only focus on 5 methods that are used in the base SAS software. In SAS, there are two approaches for estimating quartiles: the one–pass approach and the order-statistics approach. Users can specify one of the 5 methods for calculating quartiles for the order-statistics approach and only method 5 for one- pass approach. The SAS PCTLDEF= option specifies the method that the SAS procedure uses to compute quartiles/percentiles.

Let $n$be the number of non-missing values for a variable, and let $x_1, x_2, \ldots, x_n$represent the ordered values of the variable such that $x_1$is the smallest value, $x_2$is next smallest value, and $x_n$is the largest value. For the tth percentile, let $p = t/100$, p between 0 and 1. Then define $j$as the integer part of $np$and $g$as the fractional part of np or $(n+1)p$, so that

$$\begin{array}{ll} np = j + g & \text{when PCTLDEF} = 1, 2, 3, \text{or } 5 \\ (n+1)p = j + g & \text{when PCTLDEF} = 4 \end{array}$$

Here, PCTLDEF= specifies the method that the procedure uses to compute the 100*pth percentile, as shown in the table that follows.

When you use the WEIGHT statement, the 100*pth percentile is computed as

$$y = \begin{cases} \frac{1}{2}(x_i + x_{i+1}) & \text{if } \sum_{j=1}^{i} w_j = pW \\ x_{i+1} & \text{if } \sum_{j=1}^{i} w_j < pW < \sum_{j=1}^{i+1} w_j \end{cases}$$

where w_j is the weight associated with x_i and $W = \sum_{i=1}^{n} w_i$ is the sum of the weights. When the observations have identical weights, the weighted percentiles are the same as the unweighted percentiles with PCTLDEF=5.

Mathematical Methods for Computing Percentile Statistics			
PCTLDEF=	**Description**	**Formula**	
1	**weighted average at x_{np}**	$y = (1-g)\,x_j + g x_{j+1}$	
		where x_0 is taken to be x_1	
2	**observation numbered closest to np**	$y = x_i$	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
		where j is the integer part of $np + \frac{1}{2}$	
3	**empirical distribution function**	$y = x_i$	if $g = 0$
		$y = x_{j+1}$	if $g > 0$
4	**weighted average aimed at $x_{(n+1)p}$**	$y = (1-g)\,x_j + g x_{j+1}$	
		Where x_{n+1} is taken to be x_n	
5	**empirical distribution function with averaging**	$y = \frac{1}{2}\left(x_j + x_{j+1}\right)$	if $g = 0$
		$y = x_{j+1}$	if $g > 0$

SAS PROCEDURES TO CALCULATE QUARTILES

Many SAS procedures can be used to calculate quartiles. e.g., PROC UNIVARIATE, PROC MEANS and PROC STDIZE. In these procedures, two options (QMETHOD/PCTLMTD, QNTLDEF/PCTLDEF) can be used to determine which estimation method (QMETHOD/PCTLMTD) to use and which mathematical definitions / methods (QNTLDEF/ PCTLDEF) to apply.

The following are some examples of calculating quartiles using several SAS procedures.

```
 data a;
    input x;
    cards;
    1
    2
    3
    4
    ;
 run;

/* PROC UNIVARIATE, PROC MEANS and PROC SUMMARY */
 %macro quat(stat, method);
    proc &stat data=a PCTLDEF=&method noprint;
      var x;
      output out=&stat._m&method q1=q1 median=median q3=q3;
    run;
    proc print data=&stat._m&method noobs;
         title "stat=&stat, method=SAS Method &method";
    run;
 %mend quat;

 %quat(means,      5);
 %quat(means,      4);
 %quat(means,      3);
 %quat(means,      2);
 %quat(means,      1);

 %quat(summary,    5);
 %quat(univariate, 5);

/* PROC REPORT and PROC TABULATE */
 ** proc tabulate **;
 proc tabulate data=a QNTLDEF=5;
      var x;
      table x*q1;
 run;

 ** proc report **;
 proc report data=a nowd qmethod=os QNTLDEF=5;
     column x ;
     define x / q1 format=8.2 "Q1" ;
  run;

  proc report data=a nowd QNTLDEF=4 ;
     column x ;
     define x / median format=8.2 "Median" ;
  run;

  proc report data=a nowd;
     column x ;
     define x / q3 format=8.2 "Q3" ;
  run;


/* PROC STDIZE */
 proc stdize data=a outstat=out_stat5
      PCTLMTD=ord_stat PCTLDEF=5 pctlpts=25 50 75;
      var x;
 run;
 proc print data=out_stat5; run;

 proc stdize data=a outstat=out_stat4
      PCTLMTD=ord_stat PCTLDEF=4 pctlpts=25 50 75;
   var x;
 run;
 proc print data=out_stat4; run;
```

Results summary:

SAS procedures	Method	Q1	Median	Q3
All	5	1.5	2.5	3.5
	4	1.5	2.5	3.75
	3	1	2	3
	2	1	2	3
	1	1	2	3

The examples show that different methods may generate different results. However, no matter which procedure you use, it will give you the same results for the same method. For example, using PROC UNIVARIATE, with PCTLDEF=5 will have the same results as using PROC SUMMARY or PROC MEANS.

COMPARISON OF METHODS AND PROCEDURES

In SAS, use QMETHOD to specify the quartile estimation method. Use QNTLDEF or PCTLDEF to specify the mathematical definition used to compute quartiles or percentiles.

There are two options for QMETHOD, OS and P2. OS refers to ordered statistics where all the data is read into memory and sorted by the unique values. This approach is faster than P2 but is very memory intensive and is the default method. When you use ordered statistics you can specify any of the five mathematical methods, defined by QNTLDEF or PCTLDEF. P2 is the one-pass method. It is based on the piecewise-parabolic (P^2) algorithm developed by Jain and Chlamtac (1985). This one-pass algorithm is more efficient for large data sets because it requires a fixed amount of memory. When you use QMETHOD=P2 you must use PCTLDEF (or QNTLDEF) =5. The accuracy is comparable for both methods when you are between 25-75% or between the lower and upper quartiles.

Comparison of SAS Methods and Procedures for Calculating Quartiles

PROC UNIVARIATE	PROC SUMMARY	PROC MEANS	PROC TABULATE	PROC STDIZE	PROC REPORT
- most comprehensive when it comes to providing statistics on numeric data - default report automatically provides the quartiles - more flexible - can specify Q1, Q3, median -can use QMETHOD and QNTLDEF	- displays all summary statistics by default - can specify Q1, Q3, median -can use QMETHOD and QNTLDEF	- automatically calculates summary statistics - uses less memory than PROC UNIVARIATE because it does not automatically calculate quartiles -can use QMETHOD and QNTLDEF	- uses less memory than PROC UNIVARIATE because it does not automatically calculate quartiles -can use QMETHOD and QNTLDEF	-uses PCTLMTD=ORD_STAT or ONEPASS, METHOD=STD, and can use PCTLDEF(same as QNTLDEF)	-quartiles not automatically provided - need to specify q1, q3, median -can use QMETHOD and QNTLDEF

CONCLUSION

This paper provides a summarization of five mathematical definitions for calculating quartiles, two quartile estimation methods, and six procedures that are used in SAS to provide quartiles. Quartiles, although commonly used, have no standard process for calculation. Different methods may provide different results for the same data. The choice on the method that is used depends on the results needed by the statistician. Quartile results should not be reported unless you are aware that different methods may produce different results.

REFERENCES

Franklin, David. “Calculating the Quartile” Proceedings of PharmaSUG 2007.
http://ourworld.compuserve.com/homepages/dfranklinuk/NS07PO08.pdf

Journet, David. “Quartiles: How to calculate them?” http://www.haiweb.org/medicineprices/manual/quartiles_iTSS.pdf.

Langford, Eric. “Quartiles in Elementary Statistics” *Journal of Statistics* Volume 14,Number 3 (2006).
www.amstat.org/publications/jse/v14n3/langford.html.

Bressler, Lynne. “Data Summarization Methods in Base SAS® Procedures”.
http://analytics.ncsu.edu/sesug/1999/054.pdf.

SAS Institute Inc. (2003), *Online User Documentation*, Cary, NC: SAS Institute Inc.
http://support.sas.com/onlinedoc/913/docMainpage.jsp.

SAS Institute Inc. (1999), *SAS Procedures Guide*, Cary , NC: SAS Institute Inc.
http://www.sfu.ca/sasdoc/sashtml/proc/ztatback.htm

Ask Dr. Math, “Defining Quartiles”. http://mathforum.org/library/drmath/view/60969.html.

TRADEMARKS

CONTACT INFORMATION

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

Patricia Guldin
UG 1D-10
Merck Research Labs
Merck Co. & Inc
Upper Gwynedd, PA 19454
(267) 305-8242
patricia_guldin@merck.com

Liping Zhang
UG 1CD-44
Merck Research Labs
Merck Co. & Inc
Upper Gwynedd, PA 19454
(267) 305-7980
liping_zhang@merck.com