

SP08



Oh Quartile, Where Art Thou?

David Franklin
TheProgrammersCabin.com

Why look into this?

- Prompted by the question “Why is my first quartile number different from yours?” Who was right?
- Calculation of Quartile is discussed from the point of view of a SAS Programmer - no comment will be made on the merits or demerits of each definition or which definition is best for calculating the statistic.

School Book Definition

- “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”
- first quartile is 25% of the data (Q1)
- second quartile is 50% of the data (Q2, also called the Median)
- third quartile is 75% of the data (Q3)

An Example

- Data: 1, 2, 3, 4, 5, 6, 7 and 8
- SAS Method 5 (default) = 2.5
- SAS Method 4 = 2.25
- Excel = 2.75

In terms of calculation there is no one standard method for quartile

SAS Methods (I)

- 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

SAS Methods (II)

- Definition 2, Closest Observation

If $g=0.5$ and j is even then $y=x_j$, 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 X_n

$$y = (1-g)*x_j + g*x_{j+1}$$

where $n/4=j+g$ for the LQ and $3n/4=j+g$ for the UQ

Back to the Example

- 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: $Q1=2$ $Q3=6$

- In this presentation we will look at five other methods
- **SAS code is in the paper to calculate these other five methods (so no need to scribble notes frantically)**

The Classic Method

- Developed by Tukey, its aim is to find the quartiles of a set of data with little or no calculation. 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.

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 $(3n+2)/4=j+g$

Moore and McCabe

- An adaptation to the Tukey method that does not include sample median in the quartile calculation in both cases where there is an even or odd number of observations.

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 $(3n+2)/4=j+g$

Hyndman and Fan

- 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

Freund and Perles

- Used by Excel (version 2003), S-Plus, among others

$$y = (1-g)*x_j + g*x_{j+1}$$

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

Mendenhall and Sincich

- Variation on the Closest Observation method (related to Definition 2 in SAS):

LQ: if $g < 0.5$ then $y = x_j$, else $y = x_{j+1}$

UQ: if $g \leq 0.5$ then $y = x_j$, else $y = x_{j+1}$

The Results

Method	LQ	UQ
Empirical Distribution, Averaging (SAS Defn. 5)	2.5	6.5
Weighted Average at X(n+1) (SAS Defn. 4)	2.25	6.75
Empirical Distribution Function (SAS Defn. 3)	2	6
Closest Observation (SAS Defn. 2)	2	6
Weighted Average at Xn (SAS Defn. 1)	2	6
Classic Method (Tukey)	2.5	6.5
Moore and McCabe	2.5	6.5
Hyndman and Fan	2.5	6.5
Freund and Perles	2.75	6.25
Mendenhall and Sincich	2	6

Conclusion

- There are different calculation methods for the quartile - this presentation has only touched on 10!
- All methods have their place and the selection of which method to use does depend on the knowledge and experience of a statistician.
- **You should know how your software calculates a statistic before blindly reporting the result.**

Contact Information



David Franklin

TheProgrammersCabin.com

Litchfield, NH

603-275-6809

dfranklin@TheProgrammersCabin.com

<http://www.TheProgrammersCabin.com>

@ThePgrmsCabin