Calculate Quantiles by Different Methods

Keren Zhou
Institute of Computing Technology

September 2014

1 Introduction

Quantiles divide ordered data into several equal-sized subsets. Mostly, we use the variation of 4-Quantiles method, which results in 4 subsets, to illustrate the distribution of data. Five special points are included in the method. For instance, $Q1$ represents the point where 25% of values are less than it.
Though the definition seems clear, when searching the internet, we might find several methods for computing quantiles values: [1] lists out about 9 methods to calculate quantils. Furthermore, mathematical tools such as Octave, R, and SAS implement methods consist of confusing details for average users[2].
In this article, we describe the particular steps apart from sophisticated jargon, and attach Octave routines to show the usage. We hope it can help readers aware of the distinction between these methods.

2 Background

[3] provide explanations for why there are so many ways to calculate Quantiles. Methods have been existed for nearly 100 years, but there is no standard way. Softwares tend to support previous developed methods, so more and more formulas are included. In this way, we have to do research before judging which method is suitable for the project.

3 Methods

3.1 Preliminary

Octave provides four functions to calculate Quantiles, we adopt the fourth function because we have to point out which method is used:

```
q = quantile(x, p, dim, method)
```

In the above function, x is the set of data we want to analyze; p represents cumulative probability values; dim is the dimension indicator according to x; method indicates which one is used. We initialize a simple example:

```
x = [1; 2; 3; 4];
p = [0 0.25 0.5 0.75 1.0]; %calculate 4-Quantile
dim = 1;
```

By definition in [1], the general way of calculating quantile is as follows:

1. Given n values, the hth value should be the estimated position of Q_p, where h can be got by different ways.
2. If h is an integer, then x_h is the answer.
3. Or some special strategies are applied to get the answer.

Next, we analyze each method step by step.

3.2 Routines

1. **R1** Function:

   ```
   q = quantile(x, p, 1, 1)
   ```

   (a) $h_i = np_i + 1/2$

   $$\begin{aligned} h_1 &= 4 * 0 + 1/2 = 0.5 \\ h_2 &= 4 * 0.25 + 1/2 = 1.5 \\ h_3 &= 4 * 0.5 + 1/2 = 2.5 \\ h_4 &= 4 * 0.75 + 1/2 = 3.5 \\ h_5 &= 4 * 1.0 + 1/2 = 4.5 \end{aligned}$$

   (b) $Qp_i = x_{\lceil h_i - 1/2 \rceil}$

   $$\begin{aligned} Q_0 &= x_1 = 1; \\ Q_{0.25} &= x_1 = 1; \\ Q_{0.5} &= x_2 = 2; \\ Q_{0.75} &= x_3 = 3; \\ Q_1 &= x_4 = 4; \end{aligned}$$

   Notice Q_0 is a special case defined in *empirical distribution function.*

2. **R2** Function:

   ```
   q = quantile(x, p, 1, 2)
   ```

(a) $h_i = np_i + 1/2$

$$
\begin{aligned}
h_1 &= 4 * 0 + 1/2 = 0.5 \\
h_2 &= 4 * 0.25 + 1/2 = 1.5 \\
h_3 &= 4 * 0.5 + 1/2 = 2.5 \\
h_4 &= 4 * 0.75 + 1/2 = 3.5 \\
h_5 &= 4 * 1.0 + 1/2 = 4.5
\end{aligned}
$$

(b) $Qp_i = (x_{\lceil h_i - 1/2 \rceil} + x_{\lfloor h_i + 1/2 \rfloor})/2$

$$
\begin{aligned}
Q_0 &= x_1 = 1; \\
Q_{0.25} &= (x_1 + x_2)/2 = 1.5; \\
Q_{0.5} &= (x_2 + x_3)/2 = 2.5; \\
Q_{0.75} &= (x_3 + x_4)/2 = 3.5; \\
Q_1 &= x_4 = 4;
\end{aligned}
$$

Notice, when $p = 0$ use x_1; when $p = 1$, use x_N.

3. **R3** Function:

```
q = quantile(x, p, 1, 3)
```

(a) $h_i = np_i$

$$
\begin{aligned}
h_1 &= 4 * 0 = 0 \\
h_2 &= 4 * 0.25 = 1 \\
h_3 &= 4 * 0.5 = 2 \\
h_4 &= 4 * 0.75 = 3 \\
h_5 &= 4 * 1.0 = 4
\end{aligned}
$$

(b) $Qp_i = \lfloor x \rceil$

$$
\begin{aligned}
Q_0 &= x_1 = 1; \\
Q_{0.25} &= x_1 = 1; \\
Q_{0.5} &= x_2 = 2; \\
Q_{0.75} &= x_3 = 3; \\
Q_1 &= x_4 = 4;
\end{aligned}
$$

Notice, when $p = 0$ use x_1; when $p = 1$, use x_N.

4. **R4** Function:

```
q = quantile(x, p, 1, 4)
```

(a) $h_i = np_i$

$$
\begin{aligned}
h_1 &= 4 * 0 = 0 \\
h_2 &= 4 * 0.25 = 1 \\
h_3 &= 4 * 0.5 = 2 \\
h_4 &= 4 * 0.75 = 3 \\
h_5 &= 4 * 1.0 = 4
\end{aligned}
$$

(b) $Qp_i = \lfloor x \rfloor + (h_i - \lfloor h_i \rfloor)(x_{\lfloor h_i \rfloor + 1} - x_{\lfloor h_i \rfloor})$

$$
\begin{aligned}
Q_0 &= x_1 = 1; \\
Q_{0.25} &= x_1 = 1; \\
Q_{0.5} &= x_2 = 2; \\
Q_{0.75} &= x_3 = 3; \\
Q_1 &= x_4 = 4;
\end{aligned}
$$

Notice, when$p < 1/N$, use x_1; when $p = 1$, use x_N. Because here each of h_i is an integer, we do not have to adjust the position by interpolation.

Notice that method from R4-R9 use the same interpolation function, therefore we do not list them out.

$R5$	$h_i = np_i + 1/2$
$R6$	$h_i = (n+1) * p_i$
$R7$	$h_i = (n-1) * p_i + 1$
$R8$	$h_i = (n+1/3) * p_i + 1/3$
$R9$	$h_i = (n+1/4) * p_i + 3/8$

4 Canonical Method

Hyndman and Fan (1996)[4] recommend method 8, which was approximately median-unbiased estimates of the quantiles, regardless of the distribution. Matlab and Octave use method 5[5]. R uses method 7[2].

5 Conclusion

Because we mainly see Quantile analysis with samples of small sizes, we can hardly tell which method is better for specific situation. Researchers have to

take different situations into consider before choosing a method. However, there are even cases that neither of the above methods are not suitable. It is reasonable for us to propose an alternative method to calculate Quantiles, since there's no rigid principle.

References

[1] Quantile in wikipedia. *Available on-line at: http://en.wikipedia.org/wiki/Quantile [Accessed: September 24, 2014]*, 2014.

[2] David Journet. Quartiles: How to calculate them. *Available on-line at: http://www. haiweb. org/medicineprices/manual/quartiles_iTSS. pdf.[Accessed: September 29, 2005]*, 1999.

[3] John Thompson. Quantiles, percentiles: Why so many ways to calculate them? *Available on-line at: http://analyse-it.com/blog/2013/2/quantiles-percentiles-why-so-many-ways-to-calculate-them [Accessed: September 24, 2014]*, 2013.

[4] Rob J Hyndman and Yanan Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365, 1996.

[5] Quantile function usage in octave. *Available on-line at: http://octave.sourceforge.net/octave/function/quantile.html [Accessed: September 24, 2014].*