Inquiry concerning StatsBase.jl weighted quantile calculation.

Sam Albert

April 26, 2021

1 Weigthed sample quantile characterized by an optimality condition

Following the discussion on Wikipedia, a sample quantile may be characterized as a solution to the optimization problem,

$$\arg\min_{q\in\mathbf{R}} \sum_{i=1}^{N} \rho_\tau(y_i - q), \tag{1}$$

where ρ_τ is the tilted absolute value function defined by

$$\rho_\tau(y) = \begin{cases} \tau y & \text{if } y \geq 0, \\ (\tau - 1)y & \text{otherwise.} \end{cases}. \tag{2}$$

I believe the appropriate characterization of a weighted sample quantile would be

$$\arg\min_{q\in\mathbf{R}} \sum_{i=1}^{N} w_i \rho_\tau(y_i - q). \tag{3}$$

For example, this characterization results in equivalence between the weighted sample median (corresponding to $\tau = 0.5$) and the minimum weighted absolute deviation as one encounters in fitting a weighted Laplace distribution. To characterize the solutions, first note for the forward and backward derivatives,

$$d_+\rho_\tau(y) = \begin{cases} \tau & \text{if } y \geq 0, \\ \tau - 1 & \text{otherwise.} \end{cases}, \tag{4}$$

$$d_-\rho_\tau(y) = \begin{cases} -\tau & \text{if } y > 0, \\ 1 - \tau & \text{otherwise.} \end{cases}. \tag{5}$$

Letting

$$f(q) = \sum_{i=1}^{N} w_i \rho_\tau(y_i - q), \tag{6}$$

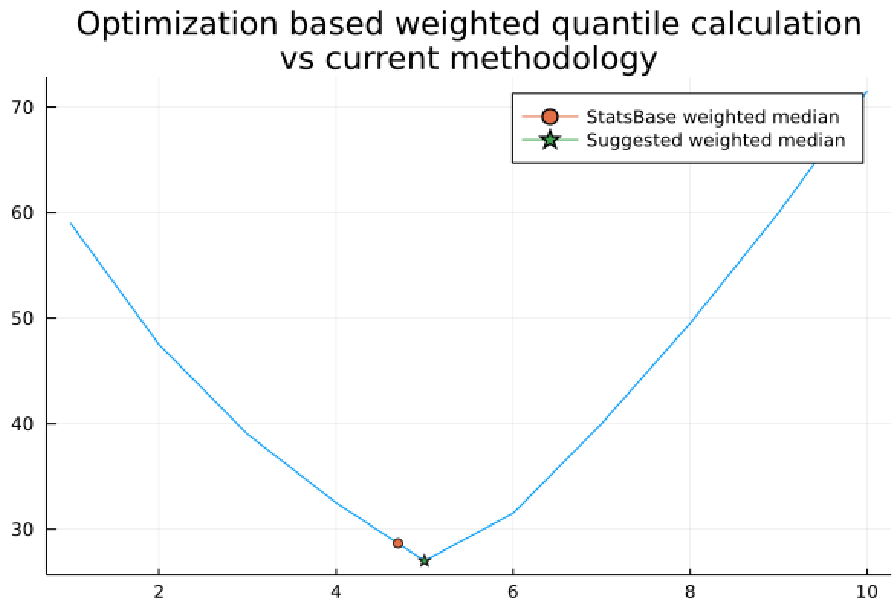

Figure 1: The current weighted median calculation is not optimal for the suggested criterion.

we characterize the minimum using nonnegativity of the directional derivatives (i.e., $d_+ f(q^*) \geq 0, d_- f(q^*) \geq 0$). In our case,

$$0 \leq d_+ f(q^*) = \sum_{y_i \geq q^*} w_i \tau + \sum_{y_i < q^*} w_i (\tau - 1), \tag{7}$$

$$0 \leq d_- f(q^*) = \sum_{y_i > q^*} w_i (-\tau) + \sum_{y_i \leq q^*} w_i (1 - \tau), \tag{8}$$

which simplify to

$$\sum_{y_i < q^*} w_i \leq \tau \sum_{i=1}^{N} w_i, \tag{9}$$

$$\sum_{y_i \leq q^*} w_i \geq \tau \sum_{i=1}^{N} w_i. \tag{10}$$

In Figure 1 we visualize the difference

2 Comparison vs StatsBase with weights = 1

In the case with all weights = 1, the solution generated here might not match that generated by StatsBase. Note the solution of the optimization problem is

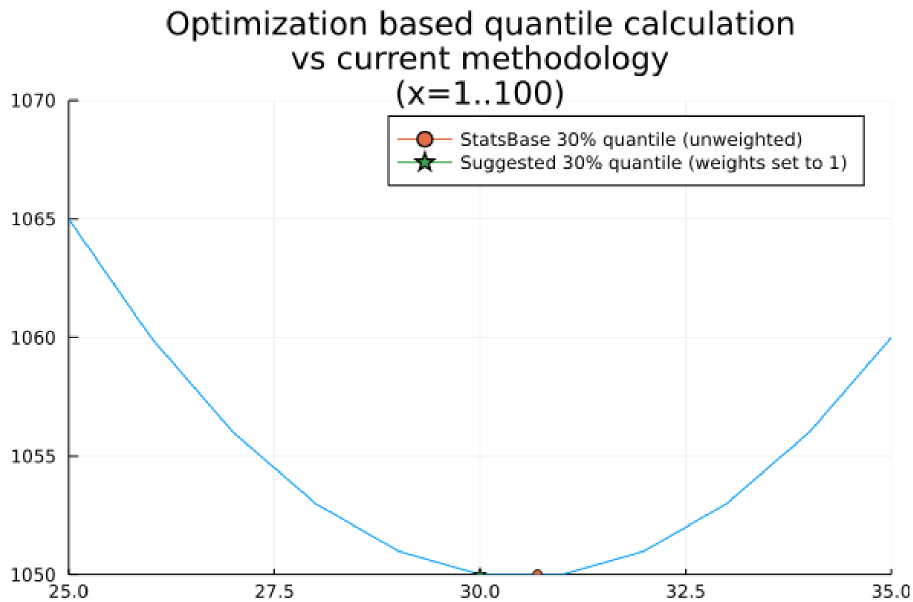

Figure 2: Calculating the 30% quantile of the numbers 1 through 100..

not necessarily unique. For example, given a sample with numbers $1, \dots, 100$, StatsBase.quantile(x, 0.3) generated 30.7 while my suggested code generated 30.0. Both suggested values satisfy (3). This is visualized in Figure 2.

In the case with numbers $1, \dots, 101$, the solution to the optimization problem is unique and the values match, as illustrated in Figure 3.

My personal opinion on this is, in the case of nonuniqueness, any optimal value should be satisfactory for most applications. However, I could see adding to the code a selection criterion in the case of nonuniqueness as an optional parameter. Thoughts?

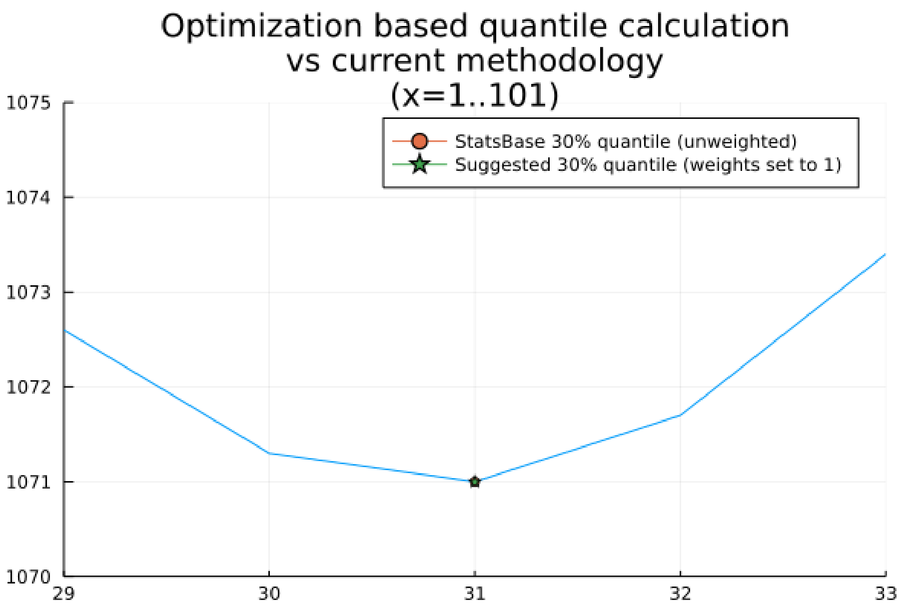

Figure 3: Calculating the 30% quantile of the numbers 1 through 101..