Proofs of some elementary facts on the weighted nearest rank

Vasilis. Anagnostopoulos

April 20, 2021

Abstract

In this note we prove two remarkable properties of the nearest rank when weights are employed. We use elementary properties to provide proofs and show that nearest rank is competitive to other percentile methods in simplicity. The properties also prove its consistency both when the weights are integral or not. It supports the material presented in a article about weighted medians in medium.

As explained in The nearest rank method, nearest rank is a good compromise in finding the percentile and asymptotically they are the same. The definition also extends naturally for the weighted case we are interested in.

To this end, lets craft a "math friendly definition" that will be convenient enough to help us prove theorems. Suppose we are given a series of weighted values, say $\{(x_1, w_1), \ldots, (x_m, w_m)\}$. Since this series could model prices (x) and quantities (w) it is natural to assume that the weights are positive. Also we make the assumption that the series is sorted in increasing prices so that cumulative distribution function can be defined. To summarize, the constraints on this series take the form:

$$w_i \geq 0 \qquad \forall i \qquad (1)$$

$$w_i \geq w_{i-1} \qquad \forall i \geq 1 \qquad (2)$$

Now we are ready to give the definitions.

Definition 0.1 (Nearest Rank)**.** We define the weighted q-nearest rank index as

$$m_q = \min\{i : \frac{\sum_{j=1}^{i} w_j}{\sum_{j=1}^{m} w_j} \geq q\} \qquad (3)$$

and the weighted q-nearest rank is

$$x_{m_q}. \qquad (4)$$

Remark. It is an easy task to prove that the definition given here and the one in Wikipedia are equivalent. Just match the definitions for the case where $w_i = 1 \; \forall i$.

We will find out that these definitions, even though they deviate somehow from the common median (q-percentile), they satisfy some very desirable properties.

1 The invariant property of nearest rank for integral weights

In this section we prove the first interesting property of nearest rank. If the weights are integral and we "explode" the sequence then, the nearest rank calculated by definition 0.1 is the same whether we apply it on the original or the exploded sequence. Consider as above a sequence of integral weighted samples satisfying (1) and (2). Its exploded version is just the same sequence with every value appearing as many times as its weight. In other words we replace this sequence

$$\{(x_1, w_1), (x_2, w_2), \ldots, (x_m, w_m)\} \tag{5}$$

with

$$\{\underbrace{(x_1, 1), \ldots, (x_1, 1)}_{w_1}, \underbrace{(x_2, 1), \ldots, (x_2, 1)}_{w_2}, \ldots, \underbrace{(x_m, 1), \ldots, (x_m, 1)}_{w_m}\} \tag{6}$$

or equivalently

$$\{\underbrace{x_1, \ldots, x_1}_{w_1}, \underbrace{x_2, \ldots, x_2}_{w_2}, \ldots, \underbrace{x_m, \ldots, x_m}_{w_m}\} \tag{7}$$

The "conditions" (1) and (2) are satisfied for the exploded sequence too.

Theorem 1.1. Exploded and unexploded sequences have the same nearest rank.

Proof. Smart counting is at the core of this theorem. Suppose the calculation on the "exploded sequence gives an index n_q. Suppose that the x value at this position, has index m_q' in the $1, \ldots, m$ interval. What we are trying to prove is that

$$m_q' = m_q. \tag{8}$$

Please, take your time to understand why. Now, by definition for the exploded sequence

$$\frac{\sum_{j=1}^{n_q} 1}{\sum_{j=1}^{m} w_j} \geq q \tag{9}$$

and consequently

$$\frac{\sum_{j=1}^{m'_q} w_j}{\sum_{j=1}^{m} w_j} \geq q \Rightarrow m'_q \geq m_q \tag{10}$$

by the minimality of m_q. On the other hand, again by definition of the exploded sequence, the condition

$$\frac{\sum_{j=1}^{m_q} w_j}{\sum_{j=1}^{m} w_j} \geq q \tag{11}$$

implies

$$\sum_{j=1}^{n_q} 1 \leq \sum_{j=1}^{m_q} w_j. \tag{12}$$

As a matter of fact we have the margin to increase n_q until we change x. This means

$$m'_q \leq m_q \tag{13}$$

and the theorem is proved. $\qquad\square$

2 Nearest rank is accessible through samples

While the previous theorem was more or less expected the next property is unexpected and makes it super useful for non-integral weights. Suppose we would like to find the nearest rank of the following sequence satisfying (1) and (2)

$$\{(x_1, w_1), (x_2, w_2), \dots, (x_m, w_m)\} \tag{14}$$

with nearest rank x_{m_q}. We add just another constraint now. the weights are positive probabilities. The weights now, sum to 1

$$\sum_{i=1}^{m} w_i = 1 \tag{15}$$

Suppose now we create another sequence

$$\{(x_1, w'_1), (x_2, w'_2), \dots, (x_m, w'_m)\} \tag{16}$$

3

with weights

$$w_i' = [Nw_i].\tag{17}$$

In the equation above the reader can identify $[.]$ as the integer part. The hope is that in the limit of infinite N the previous theorem applies.

By the properties of the integer part

$$[x] \leq x < [x] + 1\tag{18}$$

and consequently (as the original weights sum to 1)

$$N - m < \sum_{i=1}^{m}[Nw_i] \leq N.$$

The new weights now do not sum to N, possibly something less but the error is bounded by m. In order to correct this shortcoming we add 1's to the first w_i' until we get to N. This amounts to selecting a decreasing sequence of $0/1$ numbers, say k_i such that

$$w_i'' = [Nw_i] + k_i.\tag{19}$$

sums to N. Let the nearest rank of this sequence with weights w'' be $x_{m(N,q)}$.

We also need a very technical assumption that will help convergence.

Definition 2.1 (Admissible q). We say that q is admissible for the weights when

$$q \neq \sum_{i=1}^{k} w_i \text{ for all } k.\tag{20}$$

We can state our theorem.

Theorem 2.1. Non integral nearest rank is accessible through integral computations when weights are probabilities and the q is admissible for them. In other words

$$\lim_{N\to\infty} \frac{x_{m(N,q)}}{N} = x_{m_q}\tag{21}$$

Proof. Smart counting is crucial here too. We also make use of 18.

By definition

$$\sum_{i=1}^{m_q} w_i \geq q \Rightarrow \tag{22}$$

$$\sum_{i=1}^{m_q} (1 + [Nw_i]) \geq Nq \Rightarrow \tag{23}$$

$$m + \sum_{i=1}^{m_q} (k_i + [Nw_i]) \geq Nq \Rightarrow \tag{24}$$

$$\frac{\sum_{i=1}^{m_q} w_i''}{\sum_{i=1}^{m} w_i''} \geq q - \frac{m}{N}. \tag{25}$$

In a similar way,

$$\frac{\sum_{i=1}^{m(N,q)} w_i''}{\sum_{i=1}^{m} w_i''} \geq q \Rightarrow \tag{26}$$

$$\sum_{i=1}^{m(N,q)} (k_i + [Nw_i]) \geq Nq \Rightarrow \tag{27}$$

$$m + \sum_{i=1}^{m(N,q)} Nw_i \geq Nq \Rightarrow \tag{28}$$

$$\sum_{i=1}^{m(N,q)} w_i \geq q - \frac{m}{N}. \tag{29}$$

We cannot have an infinite N such that $m(N, q) < m_q$. This can be see from 27 because then

$$\sum_{i=1}^{m_q-1} w_i \geq \sum_{i=1}^{m(N,q)} w_i \geq q - \frac{m}{N}. \tag{30}$$

By taking the limits

$$\sum_{i=1}^{m_q-1} w_i \geq q \tag{31}$$

which contradicts the minimality of m_q. Consequently from a N_0 onwards

$$m(N, q) \geq m_q. \tag{32}$$

Also we can see from 23 that

$$m(N, q) > m_q + 1. \tag{33}$$

cannot happen for infinite N . So, from a N_0 onwards

$$\text{either } m(N, q) = m_q \text{ or } m(N, q) = m_q + 1. \tag{34}$$

This is where admissibility comes into play. Suppose that $m(N, q) = m_q + 1$, again for infinite N. By the minimality of $m(N, q)$ and 23 we have

$$q - \frac{m}{N} \leq \frac{\sum_{i=1}^{m_q)} w_i''}{\sum_{i=1}^{m} w_i''} < q. \tag{35}$$

By taking the limits

$$\frac{\sum_{i=1}^{m_q)} w_i''}{\sum_{i=1}^{m} w_i''} = q \tag{36}$$

which contradicts the admissibility. $\qquad\square$

The technical assumption come into play in order to select one of the two possible cases. Otherwise the integral median would oscillate. Take for example the value 3 and 4 with weights $\frac{1}{2}$ respectively. It is not difficult to see that

$$m(N, \frac{1}{2}) = \left\{ \begin{array}{ll} 3 & \text{for } N = \text{ even} \\ 4 & \text{for } N = \text{ odd} \end{array} \right\}$$

are the two possible cases. No convergence can happen.