

A Brief Introduction to the Median

Su Jianlin

March 31, 2026

I recently revisited the concept of the median, and while it is still fresh in my mind, I would like to write down the key points.

When performing outlier removal or clipping, we often need a “baseline.” For example, for a pile of non-negative data, we may consider any value more than 50 times the baseline to be an outlier. So how should this baseline be chosen? A commonly used statistic is the mean; however, the mean is easily “dragged off course” by outliers, so using it as the baseline may bias the result toward the outliers and thereby miss some findings. In that case, we can consider choosing the median as the baseline.

1 Basic Properties

For one-dimensional data points $x_1, x_2, \dots, x_n$, their mean is defined as

$$\text{mean}(x_1, x_2, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i \quad (1)$$

Since all of the data participate directly in the computation of the mean, as soon as a few points are extremely large, the mean grows accordingly, which then disturbs the judgment of outliers.

The idea of the median is to find a “neutral party” among the data to serve as the baseline. Concretely, it seeks a dividing point such that the amount of data greater than or equal to the point matches the amount less than or equal to it, which is why it is also called the “Quantile-50%.” If n is odd, the median equals the $(n+1)/2$ -th largest number of the batch; if n is even, then any number between the $n/2$ -th largest and the $(n/2+1)$ -th largest can be regarded as the median, and in general we take the average of these two.

The robustness of the median is already evident from its definition: as long as the magnitude relation between each number and the median is unchanged, the median will not change. Going one step further, we can introduce the “breakdown point” to characterize this ability; it is “the minimum proportion of outliers that makes the result tend to infinity.” For the mean, the answer is approximately 0, because a single point tending to infinity suffices to make the mean tend to infinity; for the median, however, it is 50%, because more than half of the numbers must tend to infinity before the median tends to infinity.

The drawback of the median is that it is more complicated to compute than the mean, since it requires at least some degree of sorting of the data, which is particularly unfriendly for distributed computing scenarios. Fortunately, most scenarios do not require a highly accurate median, so there is some room for maneuver; for instance, one can first compute medians locally and then compute one global mean or median, which can markedly reduce communication.

2 The Optimization Perspective

Interestingly, from the optimization perspective we can unify the mean and the median:

$$\text{mean}(x_1, x_2, \dots, x_n) = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^n (x_i - \mu)^2 \quad (2)$$

$$\text{median}(x_1, x_2, \dots, x_n) = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^n |x_i - \mu| \quad (3)$$

The proof for mean is fairly simple and is left to the reader; here we briefly sketch the proof for median. Without loss of generality, assume that the x_i are already sorted, i.e., $x_1 \leq x_2 \leq \dots \leq x_n$. Let $f(\mu) = \sum_{i=1}^n |x_i - \mu|$; differentiating directly gives

$$f'(\mu) = \sum_{i=1}^n \operatorname{sign}(\mu - x_i) = \#\{x_i < \mu\} - \#\{x_i > \mu\} \quad (4)$$

where the notation $\#$ denotes the number of elements satisfying the condition. To locate the minimum, we want the derivative to be as close to 0 as possible; that is, we want the number of x_i greater than μ to be as equal as possible to the number less than μ , which already accords with the idea of the median. A more concrete discussion requires a case analysis:

$$f'(\mu) = \begin{cases} \left\{ \begin{array}{l} \underbrace{\#\{x_i < \mu\}}_{\leq k} - \underbrace{\#\{x_i > \mu\}}_{\geq k+1} < 0, \quad \mu < x_{k+1} \\ \underbrace{\#\{x_i < \mu\}}_{\geq k+1} - \underbrace{\#\{x_i > \mu\}}_{\leq k} > 0, \quad \mu > x_{k+1} \end{array} \right\} & n = 2k + 1 \\ \left\{ \begin{array}{l} \underbrace{\#\{x_i < \mu\}}_{\leq k-1} - \underbrace{\#\{x_i > \mu\}}_{\geq k+1} < 0, \quad \mu < x_k \\ \underbrace{\#\{x_i < \mu\}}_{\geq k+1} - \underbrace{\#\{x_i > \mu\}}_{\leq k-1} > 0, \quad \mu > x_{k+1} \end{array} \right\} & n = 2k \end{cases} \quad (5)$$

That is, when n is the odd number $2k+1$, both $\mu < x_{k+1}$ and $\mu > x_{k+1}$ increase $f(\mu)$, so $\mu^* = x_{k+1}$; when n is the even number $2k$, both $\mu < x_k$ and $\mu > x_{k+1}$ increase $f(\mu)$, so $\mu^* \in [x_k, x_{k+1}]$. One can check that as long as μ^* lies within this interval, $f(\mu^*)$ stays constant, so every number in this interval qualifies as μ^* . In summary, μ^* agrees exactly with the definition of the median.

3 High-Dimensional Spaces

From the optimization perspective, we can also understand why the median withstands the interference of outliers better than the mean: if there is a particularly large x_i , the loss it brings to the mean is $(x_i - \mu)^2$, whereas for the median it is $|x_i - \mu|$; in general $(x_i - \mu)^2 \gg |x_i - \mu|$, so the mean leans more toward the outliers, since only in that way can the loss be reduced as much as possible.

Moreover, the optimization perspective has another advantage: it generalizes readily to high-dimensional spaces. We know that the notion of the mean is easy to generalize to high dimensions—for a batch of vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, the mean vector is simply $\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$. The notion of the median, however, is not easy to generalize directly, because it relies on sorting, and it is hard to define a good order for vector data.

Under the optimization perspective, however, this generalization is quite natural:

$$\text{mean}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \underset{\boldsymbol{\mu}}{\operatorname{argmin}} \sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|_2^2 \quad (6)$$

$$\text{median}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \underset{\boldsymbol{\mu}}{\operatorname{argmin}} \sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|_2 \quad (7)$$

where $\|\cdot\|_2$ is the Euclidean distance. It is easy to prove that the mean vector defined in this way is exactly $\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$, in agreement with the empirical definition. As for the median vector, we also call it the “geometric median.” From the objective function one can see that if $n = 3$, it is precisely the classical [Fermat point](#); hence in many cases we simply call the median vector the Fermat point.

Regrettably, the median vector admits no closed-form solution; it is usually computed based on the following Weiszfeld iteration:

$$\boldsymbol{\mu}_{t+1} = \frac{\sum_{i=1}^n \mathbf{x}_i / \|\mathbf{x}_i - \boldsymbol{\mu}_t\|_2}{\sum_{i=1}^n 1 / \|\mathbf{x}_i - \boldsymbol{\mu}_t\|_2} \quad (8)$$

4 Further Generalizations

Evidently, we can consider an even more general formulation:

$$\text{average}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n; \alpha) = \underset{\boldsymbol{\mu}}{\operatorname{argmin}} \sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|_2^\alpha \quad (9)$$

Let $f(\boldsymbol{\mu}) = \sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|_2^\alpha$; then

$$\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu}) = \alpha \sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|_2^{\alpha-2} (\mathbf{x}_i - \boldsymbol{\mu}) \quad (10)$$

Setting $\nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu}) = \mathbf{0}$, the equation to be solved can be written as

$$\boldsymbol{\mu} = \frac{\sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|_2^{\alpha-2} \mathbf{x}_i}{\sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|_2^{\alpha-2}} \quad (11)$$

Substituting $\boldsymbol{\mu}_t$ into the right-hand side and denoting the result by $\boldsymbol{\mu}_{t+1}$ yields the fixed-point iteration method. Plugging in $\alpha = 2$ recovers the mean vector, and plugging in $\alpha = 1$ recovers the Weiszfeld iteration. Strictly speaking, one would still need to prove convergence and uniqueness; the details are rather involved, and we omit them here.

In addition, there is a less commonly used form of the median vector in high-dimensional spaces, in which the Euclidean norm is replaced by the L_1 norm:

$$\underset{\boldsymbol{\mu}}{\operatorname{argmin}} \sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|_1 \quad (12)$$

This is called the “coordinate-wise median,” because it is essentially the one-dimensional median taken component by component. It is fairly simple to compute, but since it has no obvious geometric meaning, it is used in relatively fewer scenarios.

5 Summary

This article has briefly summarized the concept and properties of the median, as well as its extensions to high-dimensional spaces.

When reposting, please include the address of this article: <https://kexue.fm/archives/11693>

For more detailed information about reposting, please refer to: [Scientific Spaces FAQ](#)