# WHAT IS "STANDARD" ABOUT THE STANDARD DEVIATION?

FLORENCE NEWBERGER, ALAN M. SAFER, AND SALEEM WATSON

ABSTRACT. The choice of the formula for standard deviation is explained in elementary statistics textbooks in various ways. We give an explanation for this formula by representing the data as a vector in $\mathbb{R}^n$ and considering its distance from a central tendency vector. In this setting the "standard" formula represents a shortest distance in the standard metric. We also show that different metrics lead to different measures of central tendency.

## 1. INTRODUCTION

Most courses on data analysis begin, logically enough, with the study of measures of central tendency, such as the mean, median, and mode. This is naturally followed by the study of measures of dispersion – the variance and the standard deviation [4]. But then something unexpected happens – to calculate the variance, instead of averaging the differences between each data point and the mean, we average the squares of these differences. And then, as if to 'undo' the unwanted effect of squaring, this is followed by taking the square root to arrive at the standard deviation.

$$\sigma = \left( \frac{1}{N} \sum_{i=1}^{N} (x_i - m)^2 \right)^{1/2}. \tag{1}$$

But one may ask: why not simply average the distances between each data point and the mean, without squaring? The explanation given in most textbooks is that squaring avoids cancelation. Indeed, the reader can easily verify that for any set of data the expression in Equation (1) is zero when the "square" is eliminated. But this still does not answer the question of why we should square. One can argue that a more natural way to avoid cancelation is to use absolute value

$$\partial = \frac{1}{N} \sum_{i=1}^{N} |x_i - m|. \tag{2}$$

This last expression clearly measures the average variation of the data from the mean. Several explanations are given for not using this measure of variation, chief among them is the algebraic difficulties raised in using absolute value [4, p. 102]. But the choice of Formula (1) for the standard deviation is not an arbitrary one, but rather comes from the way we measure distance in $\mathbb{R}^n$, as we now show. (The term 'standard deviation' was first used by K. Pearson in an 1894 article [1].)

## 2. Turning the Question Around

Observe that Equations (1) and (2) are metrics on $\mathbb{R}^n$. If we consider the data to be the vector $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ and the measure of central tendency to be the constant vector $\mathbf{m} = (m, m, \ldots, m)$, then (1) gives the distance between $\mathbf{x}$ and $\mathbf{m}$ in the $\ell^2$ metric and (2) gives the distance in the $\ell^1$ metric. (These metrics are actually constant multiples of the usual $\ell^2$ and $\ell^1$ metrics, the constants being $N^{-1/2}$ and $N^{-1}$, respectively [2, p. 11]). So now we can ask:

> If we measure distance in some metric, what value of $m$ minimizes the distance between the vector $\mathbf{m}$ and the data vector $\mathbf{x}$?

That value of $m$ would be the proper choice for the central tendency of the data, because it is the number "closest" to all the data points (in that metric).

## 3. The Standard Metric

Suppose we use the standard metric to measure the distance between the data vector $x$ and the measure of central tendency $m$. What is the appropriate choice for the central tendency $m$? To answer this question we need to find the value of $m$ that minimizes the distance between $\mathbf{x}$ and $\mathbf{m}$. Equivalently, we need to find the value of $m$ that minimizes

$$v(m) = \sum_{i=1}^{N} (x_i - m)^2.$$

Taking the derivative with respect to $m$ and setting the result equal to zero we have

$$v'(m) = \sum_{i=1}^{N} 2(x_i - m)(-1) = 0.$$

So $m = \sum_{i=1}^{N} x_i/N$. In other words, $m$ is the mean of the data.

## 4. The $\ell^1$ Metric

If we use the $\ell^1$ metric to measure distance, what is the appropriate choice for the central tendency $m$? To answer this question we want the choice of $m$ that minimizes

$$\delta(m) = \sum_{i=1}^{N} |x_i - m| = \sum_{m \le x_i} (x_i - m) + \sum_{x_i \le m} (m - x_i).$$

Taking the derivative with respect to $m$ and setting it equal to zero we have

$$\delta'(m) = \sum_{m \le x_i} (-1) + \sum_{x_i \le m} (+1) = 0.$$

This means that $m$ is any number that is in the middle of the data, that is, $m$ is a median. See also [3].

## 5. A Measure of Central Tendency for the Hamming Metric

Let's consider the Hamming metric on $\mathbb{R}^n$ [2, p. 9]. The Hamming distance between two vectors in $\mathbb{R}^n$ is the number of coordinates where the two vectors differ. In particular, the distance between a data vector $\mathbf{x}$ and a central tendency vector $\mathbf{m}$ in the Hamming metric is

$$d(x, m) = \big|\{i | x_i \neq m, 1 \le i \le N\}\big|.$$

(Here $|X|$ denotes the cardinality of a set $X$.) It is clear that the value of $m$ that minimizes the Hamming distance between $\mathbf{x}$ and $\mathbf{m}$ is the mode, the value that occurs most often in the data.

## 6. A Measure of Central Tendency for the $\ell^\infty$ Metric

Of course there are many metrics on $\mathbb{R}^n$, and each has a corresponding measure of central tendency. Let's see what measure of central tendency corresponds to the $\ell^\infty$ metric [2, p. 6]. The distance between a data vector $\mathbf{x}$ and a central tendency vector $\mathbf{m}$ in the $\ell^\infty$ metric is

$$d(x, m) = \max\{|x_i - m| : 1 \le i \le N\}.$$

It is clear that $d(\mathbf{x}, \mathbf{m}) = |x_i - m|$ where $x_i$ is either the largest or smallest data point. Denoting these data points by $a$ and $b$, respectively, we can express this distance as a function of $m$:

$$\omega(m) = \max\{a - m, m - b\}.$$

A graph of $\omega$ as a function of $m$ is shown in Figure 1.

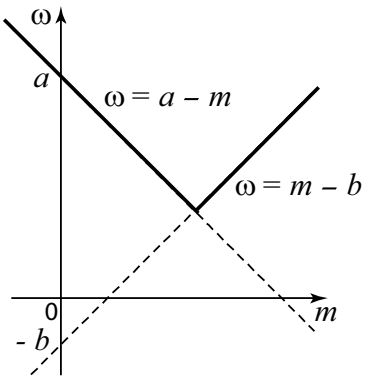


Figure 1. Graph of $\omega(m) = \max\{a - m, m - b\}$.

From Figure 1 it is clear that the minimum value of $\omega(m)$ occurs at the point of intersection of the two lines, at $m = (a+b)/2$. In other words, the measure of central tendency associated with the $\ell^\infty$ metric is the "average of the largest and smallest data points." This measure of central tendency is used by statisticians and is called the *midrange*.

## 7. Conclusion

To measure the distance between a data vector $\mathbf{x}$ and a central tendency vector $\mathbf{m}$ we can use different metrics on $\mathbb{R}^n$. If we use the *standard* metric (the Euclidean metric or $\ell^2$ metric) the distance between the data vector $\mathbf{x}$ and $\mathbf{m}$ is the *standard* deviation. In this case, the value of $m$ that minimizes the distance between $\mathbf{x}$ and $\mathbf{m}$ is the mean of the data. Other metrics induce other well-known choices for the central tendency vector $\mathbf{m}$.

## References

[1] H. A. David, *First occurence of common terms in mathematical statistics*, The American Statistician, **49** (1995), 121–133.
[2] E. Kreyszig, *Introductory Functional Analysis with Applications*, Wiley, New York, 1978.
[3] N. Schwertman, A. Gilks, and J. Cameron, *A simple noncalculus proof that the median minimizes the sum of the absolute deviations*, American Statistician, **44** (1990), 38–39.
[4] M. F. Triola, *Elementary Statistics*, Pearson, New York, 2007.

MSC2000: 62-07, 62F10

Department of Mathematics and Statistics, California State University, Long Beach, Long Beach, CA 90840-1001
*E-mail address*: fnewberg@sculb.edu

Department of Mathematics and Statistics, California State University, Long Beach, Long Beach, CA 90840-1001
*E-mail address*: asafer@csulb.edu

Department of Mathematics and Statistics, California State University, Long Beach, Long Beach, CA 90840-1001
*E-mail address*: saleem@csulb.edu