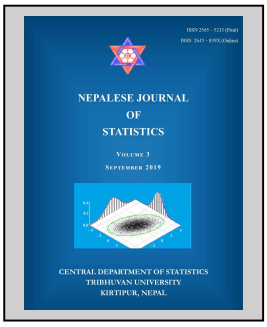
Have You *Seen* the Standard Deviation?

Jyotirmoy Sarkar [1] and Mamunur Rashid [2*]

Submitted: 28 August 2018; Accepted: 19 March 2019

Published online: 16 September 2019
DOI: https://doi.org/10.3126/njs.v3i0.25574



ABSTRACT

Background: Sarkar and Rashid (2016a) introduced a geometric way to visualize the mean based on either the empirical cumulative distribution function of raw data, or the cumulative histogram of tabular data.

Objective: Here, we extend the geometric method to visualize measures of spread such as the mean deviation, the root mean squared deviation and the standard deviation of similar data.

Materials and Methods: We utilized elementary high school geometric method and the graph of a quadratic transformation.

Results: We obtain concrete depictions of various measures of spread.

Conclusion: We anticipate such visualizations will help readers understand, distinguish and remember these concepts.

Keywords: Cumulative histogram, dot plot, empirical cumulative distribution function, histogram, mean, standard deviation.

Address correspondence to the author: Department of Mathematical Sciences, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA,

E-mail: jsarkar@iupui.edu [1]; Department of Mathematics, DePauw University, Greencastle, IN 46135, USA, E-mail: mrashid@depauw.edu [2*] (Corresponding author)

INTRODUCTION

The (arithmetic) mean is a common measure of center, and the standard deviation (SD) of spread, of a set of values of a quantitative variable. These are basic concepts in all quantitative disciplines, and they appear frequently in everyday life applications (Pollatsek, Lima & Well, 1981; Lesser, Wagler & Abormegah, 2014). While most users understand the mean reasonably well,

the SD remains a challenging concept for many, even after they have learned its definition and mastered its computation.

Sarkar and Rashid (2016a) gave a new way to visualize the mean, which we shall review here briefly. On this foundation, we will build the geometric methods to visualize measures of spread: the primary objective is to visualize the SD; additionally, we will visualize the mean deviation (MD) and the root mean squared deviation (RMSD) which serve as lower bounds for the SD. Such visualizations will not only help students and users of statistics understand the concepts better, but also help them distinguish these concepts and remember them with ease. As the old Chinese proverb (Confucius, n.d.) says:

"*I hear and I forget. I see and I remember. I do and I understand.*" — Confucius.

The mean of a set of *n* numbers $\{x_1, x_2, \dots, x_n\}$ is defined by: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (1)

The absolute deviations from the mean are $d_i = |x_i - \bar{x}|$, for $i = 1, 2, \dots, n$; and the MD of $\{x_1, x_2, \dots, x_n\}$ from the mean is the mean of the deviations $\{d_1, d_2, \dots, d_n\}$; and is defined by

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| \quad (2)$$

The SD of $\{x_1, x_2, \dots, x_n\}$ is defined by: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (3)

Often, a student is inclined to think of the SD as "the average distance of the numbers from their center." Unfortunately, this description, though perfect for the MD, is erroneous for the SD! In an attempt to rectify the error, a teacher may offer a better explanation of the SD as "the root mean squared deviation (RMSD) from the mean." However, when translated into an algebraic expression, the RMSD becomes: $\tilde{s} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (4)

This is *different* from the SD in (3), because the denominators under the radical signs differ. Expression (4) is also known as the population SD (see, for example, Sarkar and Rashid (2016b)) when the set of values $\{x_1, x_2, \dots, x_n\}$ is considered as an entire population. The correct interpretation of the SD is somewhat convoluted: It is "the positive root *mean* squared deviation (RMSD) from the mean," where only the positive root is admissible and the first *mean* involves a division by $(n-1)$. For an explanation of why it is so, we refer the reader to Martin (2003), and Sarkar and Rashid (2016b). Here we simply mention that the knowledge of the complete original data $\{x_1, x_2, \dots, x_n\}$ is equivalent to the knowledge of the transformed data
$\{\bar{x}, (x_1 - \bar{x}), (x_2 - \bar{x}), \dots, (x_{n-1} - \bar{x})\}$ since $(x_n - \bar{x})$ can be recovered from the fact that
$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_{n-1} - \bar{x}) + (x_n - \bar{x}) = 0$
This is an easy consequence of (1). Recall that s^2 is called the variance and $\tilde{s}^2$ is called the mean squared deviation (MSD). The algebraic definitions (2)–(4) and their verbal interpretations may suffice to teach students how to compute the three measures of spread—the MD, the RMSD and the SD. But these algebraic definitions fall short of giving students deep insight or complete understanding. We contend that geometric visualizations of these measures of spread will go a

long way to help students understand each concept better, distinguish one concept from another, and remember them all with ease. Unlike in Sarkar and Rashid (2016b), where three-dimensional solids of revolution are used to visualize the RMSD and the SD, here we only use plane geometry, thereby making the method more accessible to a wider readership.

We illustrate the visualizations of the mean, the MD, the RMSD and the SD of a set of numbers in the context of two examples taken from Sarkar and Rashid (2016a), which only depicts the mean. Example 1 uses a small set of raw data; and Example 2 involves a moderately sized data displayed as a histogram. We draw all figures using the freeware R.

MEAN, MD, RMSD AND SD FROM RAW DATA

Example 1. The number (x) of College Spirit Sweatshirts sold by the Campus Center Store each day during a week are: 23, 13, 19, 15, 20; which after sorting become: 13, 15, 19, 20, 23. Using formulas (1)–(4), we compute $\bar{x} = 18$, MD= 3.2, RMSD=$\tilde{s} = \sqrt{12.8} \approx 3.58$ and SD=$s = 4$.

We recall from Sarkar and Rashid (2016a) the geometric way to see the mean. Let $y = F(x)$ denote the empirical cumulative distribution function (ECDF) of the data; that is, $F(x) = N(x)/n$, where $N(x)$ is the number of terms among the set of n numbers that are *no more than* x. The ECDF is a step function in which, starting from 0 on the left extreme, the steps are raised at each of the given numbers by a height of $1/n$, thereby ending at height 1 on the right extreme. See, for example, Rice (2007). Fig. 1 depicts the ECDF of the data in Example 1.

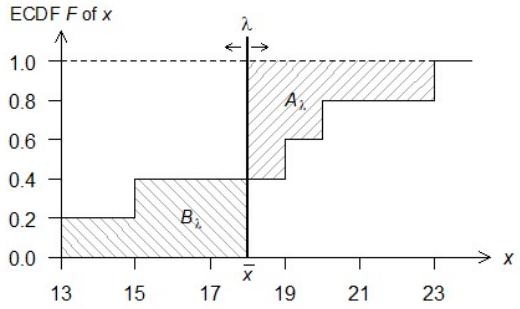


Fig. 1. The ECDF $F(x)$ of the data given in Example 1.

The vertical line, when placed at the mean $x = \bar{x}$, equalizes the shaded areas to the left and to the right.

Next, we contemplate a sliding vertical line λ that intersects the graph of $y = F(x)$, and consider two regions: (i) the shaded region A_λ to the right of the vertical line above the ECDF and below the horizontal line $y = 1$; and (ii) the shaded region B_λ to the left of the vertical line, below the ECDF and above the horizontal axis $y = 0$. These two shaded regions have equal

areas if and only if the sliding vertical line λ comes to rest exactly at $x = \bar{x}$. The reason is as follows: If we interchange the horizontal and the vertical axes in Fig. 1, then the total area under the inverse mapping $x = F^{-1}(y)$, over the interval $y \in [0, 1]$, equals

$\int_0^1 F^{-1}(y)\, dy = \int_0^{1/n} F^{-1}(y)\, dy + \cdots + \int_{1-1/n}^1 F^{-1}(y)\, dy = \frac{x_1}{n} + \cdots + \frac{x_n}{n} = \bar{x},$

which is also the area of the single rectangle $[0,1] \times [0, \bar{x}]$.

To see the three measures of spread—the MD, the RMSD and the SD—from the ECDF of the raw data, we use Algorithm 1, which we illustrate in Fig. 2 for the data in Example 1.

Algorithm 1. (To see the MD, the RMSD and the SD, starting from the ECDF)

1) On the ECDF *F* of the raw data, draw the mean vertical line $x = \bar{x}$. Reflect the portion of the ECDF to the left of the vertical line $x = \bar{x}$, about that line, with the reflection falling on the right side.
2) Let *G* be the resultant graph to the right of the vertical line $x = \bar{x}$ consisting of both the reflected part and the part of *F* that was already to the right of the vertical axis. See Remark 1 below for an interpretation of graph *G*.
3) At the top edge of graph *G*, introduce a new horizontal axis $d = |x - \bar{x}|$ representing the (absolute) deviation from the mean. The vertical axis v is the line $d = 0$. Thus, the graph *G* now resides in the fourth quadrant between $v = -1$ and $v = 0$. Draw the graph *Q* of $v = d^2$ in the first quadrant.
4) Find the mean vertical line $d = \bar{d}$ of *G* so that the areas of shaded regions to its two sides are equal. Then $\bar{d}$ is the mean deviation (MD) of the data. In fact, $\bar{d}$ is also the total area of the two shaded regions $A_{\bar{x}}$ and $B_{\bar{x}}$ in Fig. 1.
5) Extend the steps of *G* all the way left to reach the vertical axis $d = 0$, thereby obtaining a collection of rectangles G that are left aligned at $d = 0$, contiguously stacked between $v = -1$ and $v = 0$, and of varying widths.
6) Consider any member rectangle in G. Call its vertices $(0, -u), (0, -l), (d, -l), (d, -u)$. To these vertices apply the transformation $(d, v) \rightarrow (v, d^2)$, so that they move to new vertices $(-u, 0), (-l, 0), (-l, d^2), (-u, d^2)$ respectively. First draw a vertical line through $(d, -u)$ until it meets *Q* at (d, d^2); then turning left, draw a horizontal line $v = d^2$. (See the light, solid, directed lines.) These new vertices form a transformed rectangle. When all members of G have been so transformed, call the collection of all such transformed rectangles H. The rectangles in H are bottom aligned at $v = 0$; they are contiguously arranged between $d = -1$ and $d = 0$; and they have varying heights resembling a histogram. Call their graph *H*. See Remark 1 for its interpretation.
7) Find the mean horizontal line $v = \bar{v}$ of *H* so that the shaded areas of two regions above and below it are equal. Indeed, $\bar{v} = \tilde{s}^2$ is the MSD, and its square root is the RMSD.
8) Join O and J= $(-1 + 1/n, \bar{v})$ by a line and extend it to meet the vertical line $d = -1$ at

W= $(-1, w = \bar{v}.n/(n-1))$. Indeed, $w = \bar{v}.n/(n-1) = s^2$ is the variance, and its positive square root is the SD.

9) The horizontal lines through J and W intersect the quadratic curve Q at $R = (\tilde{s}, \tilde{s}^2)$ and $S = (s, s^2)$ respectively. Dropping vertical lines from R and S, we see the RMSD $\tilde{s}$ and the SD s on the horizontal d-axis. (See the dark, dotted/solid, directed lines.)

Remark 1: The graph *G* represents almost the ECDF of the deviations $d_i = |x_i - \bar{x}|$, except that the rectangles are not sorted from the narrowest at the bottom to the widest at the top. Indeed, this sorting is optional, and to keep matters simple we choose to forgo sorting. Likewise, the graph *H* (viewed after a clockwise right-angled rotation) represents almost the ECDF of the squared deviations $d_i^2 = (x_i - \bar{x})^2$, except the rectangles are not sorted.

Thus, using Algorithm 1, we actually *see* the MD, the RMSD and the SD of the raw data. Furthermore, we *see* that the SD is bigger than the RMSD, which is at least as big as the MD. As numerical evidence of these claims, note that

$\bar{x} = 18,\ \text{area}(A_{\bar{x}}) = \frac{(5+2+1)}{5} = 1.6,\ \text{area}(B_{\bar{x}}) = \frac{(5+3)}{5} = 1.6,\ \bar{d} = 3.2, \tilde{s}^2 = \frac{64}{5} = 12.8,\ s^2 = 16.$

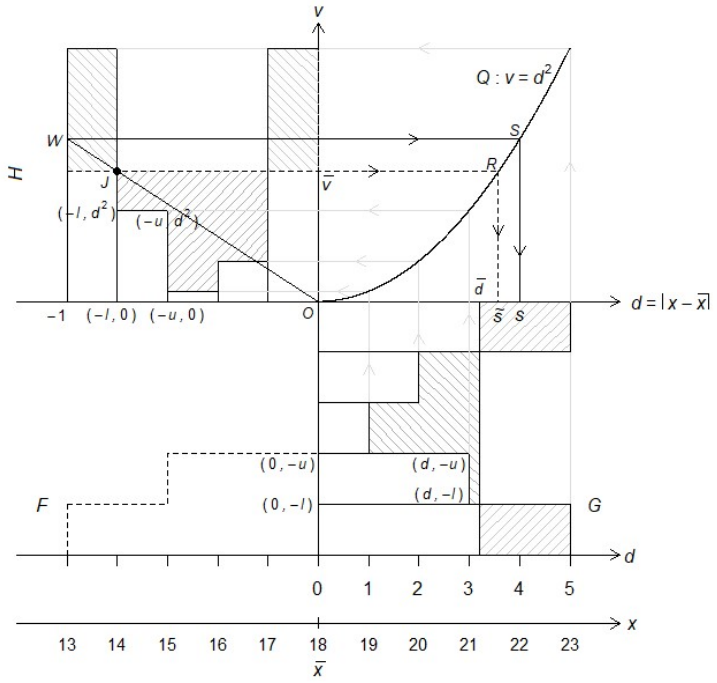


Fig. 2. Visualizing the mean $\bar{x}$, the MD $\bar{d}$, the RMSD $\tilde{s}$ and the SD s by applying Algorithm 1 to the ECDF of the raw data in Example 1.

MEAN, MD, RMSD AND SD FROM HISTOGRAM DATA

Example 2. The scores (y) of 25 students in STAT 101 exam are given below after sorting.

Table 1. The (sorted) scores of 25 students in STAT 101 exam.

96	94	93	90	88
88	88	85	85	85
82	80	77	75	75
75	72	72	71	66
65	65	55	50	48

Source: Sarkar and Rashid (2016a)

For the raw data in Table 1, Formulas (1)–(4), yield mean=$\bar{y}_R = 76.8$, MD=$\bar{d}_R$=10.608, RMSD=$\tilde{s}_R \approx 12.9985$ and SD=$s_R \approx 13.2665$.

Here, we write subscript R to denote raw data. When the number of terms in the raw dataset is moderate to large, it may be laborious to construct the ECDF; and hard to approximate the value of the mean by drawing a vertical line that equalizes the areas to its left and right. We oftentimes summarize the raw data in the form of a tabular data and draw a histogram as shown in Fig. 3(a). The histogram forfeits a little bit of precision, but it helps users comprehend the data better and approximate the value of the mean much more quickly. For instance, the histogram in Fig. 3(a) shows that there are 3 scores in [60, 70), or more precisely in the bin (59.5, 69.5); but their exact values {65, 65, 66} are lost. Sarkar and Rashid (2016a) converts the histogram in Fig. 3(a) into a cumulative histogram (CH) shown in Fig. 3(b). The CH is the graph of $y = F(x)$, where $F(x)$ is the area under the histogram to the left of x, expressed as a fraction of the total area under the histogram. In fact, since the histogram is a non-negative, step function, the corresponding CH $y = F(x)$ is a piecewise linear, non-decreasing function.

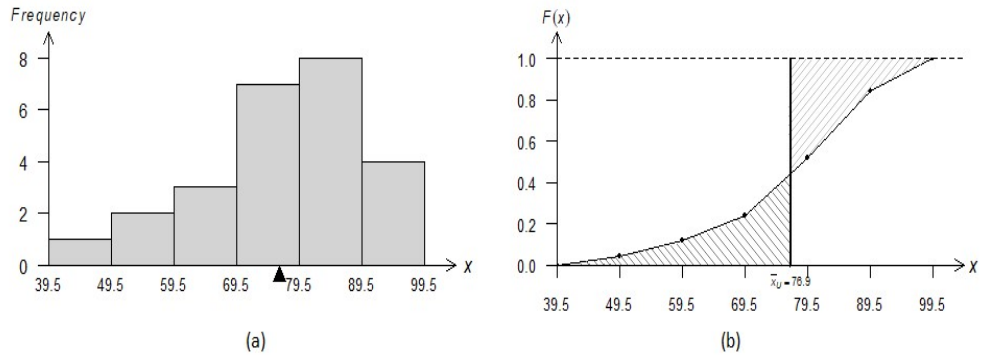


Fig. 3. (a) Histogram of the data in Example 2, with the mean shown as a fulcrum at $\bar{x}_U$=76.9. (b) The cumulative histogram $y = F(x)$ with the mean shown as a vertical line at $x = \bar{x}_U$ which equalizes the areas of the shaded regions to the left and to the right.

These figures are taken from Sarkar and Rashid (2016a). In order to visualize the mean from the CH $y = F(x)$, simply place a vertical line that equalizes the areas of the shaded regions to the left and to the right of this line, just as we did for the ECDF of the raw data in Fig. 1. We denote the mean by $\bar{x}_U$, where the subscript *U* refers to the assumption that the data values within a bin are uniformly spread out within that bin. Hence, $\bar{x}_U$ may somewhat differ from $\bar{x}_R$. To obtain the MD, the RMSD and the SD from the CH, we use Algorithm 2 (which almost mimics Algorithm 1 of the previous section, except Steps 5 and 6 differ since there are no rectangles here). We illustrate Algorithm 2 in Fig. 4 for the data in Example 2.

Algorithm 2. (To construct the mean, the MD, the RMSD, and the SD, starting from the CH)

1) On the CH *F* of the tabular data, draw the mean vertical line $x = \bar{x}_U$. Reflect the portion of the CH to the left of the vertical line at the mean, about that line, with the reflection falling on the right side.
2) Denote by *G* the resultant graph to the right of the vertical line at the mean consisting of both the reflected part and the part of *F* that was already to the right of the vertical axis. From Graph *G* one can obtain (but it is not necessary to do so) the ECDF of the deviation $d = |x - \bar{x}_U|$ by taking the vertical difference between the existing part of *F* and the reflected part.
3) At the top edge of graph *G*, introduce a new horizontal axis $d = |x - \bar{x}_U|$ representing the (absolute) deviation from the mean. The vertical axis v is $d = 0$. Thus, *G* resides in the fourth quadrant between $v = -1$ and $v = 0$. Draw the graph *Q* of $v = d^2$ in the first quadrant.
4) Find the mean vertical line $d = \bar{d}$ of *G* so that the areas of regions to its two sides are equal. Then $\bar{d}$ is the mean deviation (MD) of the histogram data. Indeed, $\bar{d}$ is also the total area of the two shaded regions in Fig. 3(b).
5) Consider any typical pair of point $(0, -u), (d, -u)$ on graph *G*. Join them by a horizontal line segment. Apply the transformation $(d, v) \to (v, d^2)$, so that these points move to new points $(-u, 0), (-u, d^2)$ respectively: Draw a vertical line through $(d, -u)$ until it meets *Q* at (d, d^2); then turning left, draw a horizontal line $v = d^2$. (See the light, solid, directed lines.) Join these transformed points vertically. Repeat the process for several values of u.
6) Join freehand (not linearly, but by using *quadratic* curves) the top ends of these newly drawn vertical line segments successively. Call the resulting graph *H*.
7) Find the mean horizontal line $v = \bar{v}$ of *H* so that the areas of the two shaded regions above and below it are equal. Indeed, $\bar{v}$ is the MSD, and its square root is the RMSD.
8) Join O and J= $(-1 + 1/n, \bar{v})$ by a line and extend it to meet the vertical line $d = -1$ at W= $(-1, w = \bar{v}.n/(n-1))$. Indeed, $w = \bar{v}.n/(n+1) = s^2$ is the variance, and its positive square root is the SD.
9) The horizontal lines through J and W intersect the quadratic curve *Q* at $R = (\tilde{s}, \tilde{s}^2)$ and $S = (s, s^2)$ respectively. Dropping vertical lines from R and S, we see the RMSD $\tilde{s}$ and the SD s on the horizontal d axis. (See the dark, dotted/solid, directed lines.)

Applying Algorithm 2 to the CH data in Example 2, we actually see the MD, the RMSD and the SD of the CH data, where $A_{\bar{x}_U} = 5.34264 = B_{\bar{x}_U}$, $\bar{d} = 10.68528 \approx \bar{d}_R = 10.6080$, $\tilde{s} = 13.32173 \approx \tilde{s}_R = 12.9985$, $s = 13.59644 \approx s_R = 13.2665$.

Remark 2: In practice, Step 5 of Algorithm 2 can be performed only for a finite number of points on *G*. Since *G* is piecewise linear and continuous, we recommend choosing these points at the vertices on *G* and at some additional (one or two) intermediate points between each successive pairs of vertices. Alternatively, one may choose ten or 20 equally-spaced values of u.

In view of Remark 2, we note that Algorithm 2 is not an exact method. But, for all practical purposes, the method is sufficiently good to approximate $\bar{v}$ (or the MSD) and hence the RMSD, the variance and the SD. Furthermore, since the sample size for a histogram is typically large, there is hardly any difference between $\bar{v}$ and $w = \bar{v}\, n/(n-1)$ (or the variance). Thus, the RMSD and the SD are approximately equal (and the approximate equality gets better and better as *n* becomes larger and larger); and they exceed the MD. Therefore, the MD, which is easier to calculate, serves as a lower bound for the SD.

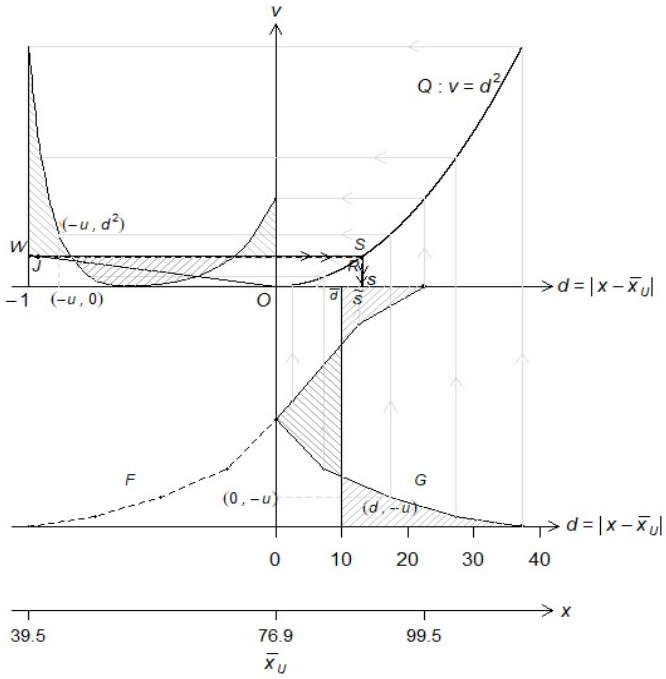


Fig. 4. Visualizing the mean $\bar{x}_U$, the MD $\bar{d}$, the RMSD $\tilde{s}$ and the SD s by applying Algorithm 2 to the cumulative histogram in Example 2.

CONCLUSION

From Sarkar and Rashid (2016a), we borrow the geometric visualization of the mean (a measure of center) of a set of numbers (or of a histogram) by superimposing a vertical line on to the ECDF (or the cumulative histogram) in order to equalize the areas of regions to the left and to the right of the vertical line. Thereafter, we provide elementary geometric visualizations for all three measures of spread namely the MD, the RMSD and the SD based on the ECDF (or the cumulative histogram). We hope that such vivid representations will not only reinforce the correct interpretations of these concepts, but also help students distinguish them and remember them with ease. To ensure that students really do understand the notions well, following Confucius, we urge them to draw Fig. 2 based on their own favorite raw data. We can extend the notions developed here to interpret the mean, the MD, the RMSD and the SD of a random variable which can be either discrete or continuous. Such a geometric point of view is also beneficial in understanding many useful properties of random variables and their transforms.

CONFLICT OF INTEREST

The authors declared that there is no conflict of interest.

ACKNOWLEDGEMENTS

We thank our colleagues and students for a lively discussion at the weekly seminar. We also thank the editor and two anonymous referees for their guidance in correcting errors and omissions.

REFERENCES

Confucius. (n.d.). Chinese Proverb. Retrieved from https://www.brainyquote.com/quotes/confucius_136802

Lesser, L., Wagler, A., & Abormegah, P. (2014). Finding a Happy Median: Another Balance Representation for Measures of Center. *Journal of Statistics Education*, *22*(3), 1–27.

Martin, M. A. (2003). "It's like… you know": The use of analogies and heuristic in teaching introductory statistical methods. *Journal of Statistics Education*, *11*(2). doi:10.1080/10691898.2003.11910705

Pollatsek, A., Lima, S., & Well, A. D. (1981). Concept or Computation: Students' Understanding of the Mean. *Educational Studies in Mathematics*, *12*(2), 191-204. doi: 10.1007/BF00305621

R Core Team (2017). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org/

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis* (3rd ed.). Boston, MA: Brooks/Cole, Cengage Learning.

Sarkar, J., & Rashid, M. (2016a). A geometric view of the mean of a set of numbers. *Teaching Statistics: An International Journal for Teachers*, *38*(3), 77-82. doi:10.1111/test.12101

Sarkar, J., & Rashid, M. (2016b). Visualizing mean, median, mean deviation and standard deviation of a set of numbers. *The American Statistician, 70*(3), 304-312. doi:10.1080/00031305.2016.1165734

Reference to this paper should be made as follows:
Sarkar, J., & Rashid, M. (2019). Have you *seen* the standard deviation?. *Nep. J. Stat., 3,* 1-10.