Visualizing Mean, Mean Deviation and Standard Deviation of a Continuous Random Variable

Jyotirmoy Sarkar[1], Mamunur Rashid[2]
[1] Department of Mathematical Sciences, IUPUI, Indianapolis, IN 46202
[2] Department of Mathematics, DePauw University, Greencastle, IN 46135

Abstract
We review recent interpretations of the mean, the mean deviation (MD) and the standard deviation (SD) of a set of numbers. For each quantity, the process begins with the empirical cumulative distribution function (ECDF) or a suitable transformation of it, and then finds the location of a vertical line that renders equal the areas of two regions bounded by the line itself, the (transformed) ECDF and the horizontal line $y = 0$ or $y = 1$. Here, the above interpretations are extended to a continuous random variable. These interpretations help users of statistics refine their intuition, and anticipate the numerical values of the mean, the MD and the SD even before evaluating them using the Calculus.

Key Words: Cumulative Distribution Function, Euclidean Method, Mean Square Deviation, Solid of Revolution

1. Introduction

The mean is the most common measure of location or center, and the standard deviation (SD) is the most common measure of the scale or spread in a dataset or a random variable (RV). These are extensively used summaries of data/variable. While the mean has a familiar depiction as a fulcrum along the horizontal axis which balances the dot plot of the raw data, or the graph of the probability mass function (PMF) of a discrete random variable (DRV), or the graph of the probability density function (PDF) of a continuous random variable (CRV), no such visualization of the SD was available in the literature until recently!

Recently, Sarkar and Rashid (2016 a-d) introduced a vertical line method to visualize the mean of the raw data based on its empirical cumulative distribution function (ECDF). We briefly review the method in Sections 2. In Section 3, we review a visualization of the mean deviation (MD), the mean square deviation (MSD) and the SD using the vertical line method applied to the ECDF of suitably transformed data. Thereafter, in Section 4, we extend the vertical line method of visualizing the MD, the MSD and the SD to a CRV. Section 5 documents a summary and some concluding remarks.

2. Methods to Visualize the Mean of a Given Dataset or a CRV

The mean is the most common measure of center. See Pollatsek et al. (1981) and Lesser et al. (2014). The (arithmetic) mean of a set of n numbers $\{x_1, x_2, x_3, \cdots, x_n\}$ is defined by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1a)$$

For a CRV X with support set S and PDF $f(x)$, the mean is defined as

$$\mu = \int_S x\, f(x)dx \qquad (1b)$$

Traditionally, the mean is visualized as the location of a fulcrum that balances the dot plot of the data, the graph of the PMF of a DRV, or the graph of the PDF of a CRV. See, for example, Watier, et al. (2011). Following that tradition, we depict the fulcrum of the mean in Figure 1 (a-b) based on the data and the CRV in Examples 1-2.

Example 1. The number of cars sold by a dealership on five weekdays are: 7, 4, 8, 3, 9.

Example 2. A dart is thrown at a circular target of radius one, and if and only if the dart hits the target the distance of the point of impact from the center of the circle is recorded. Then the recorded distance is modelled by the PDF $f(x) = 2x,$ for $0 < x < 1$.

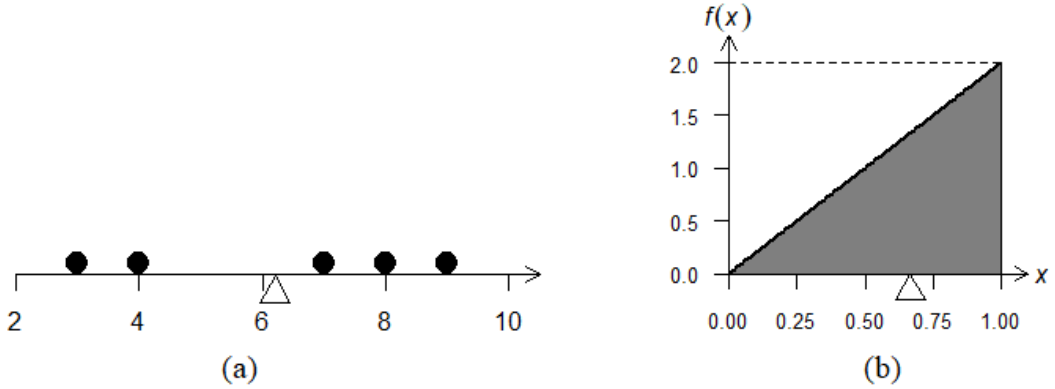


Figure 1. The mean shown as a fulcrum that balances (a) the dot plot of the data in Example 1, and (b) the graph of the PDF in Example 2

An alternative interpretation of the mean of a data set involves the ECDF of the data, which is a step function given by $F_n(x) = N(x)/n$, where $N(x)$ is a count of data values that are no more than x. Similarly, for a CRV, the mean involves the CDF given by $F(x) = \int_{-\infty}^{x} f(u)du$ for all real number x. It is well known that the mean can be obtained directly from the (E)CDF (by algebraic manipulations for the data and by applying Fubini's theorem for a CRV). In fact, using a common symbol F to represent both the ECDF and the CDF, we have

$$\mu = -\int_{-\infty}^{0} F(x)dx + \int_{0}^{\infty}[1 - F(x)]dx \qquad (2a)$$

Although not as widely known, much more is true: For any real number l, the mean is

$$\mu = l - \int_{-\infty}^{l} F(x)dx + \int_{l}^{\infty}[1 - F(x)]dx$$

In particular, μ is the unique solution to l so that

$$\int_{-\infty}^{l} F(x)dx = \int_{l}^{\infty}[1 - F(x)]dx \qquad (2b)$$

In view of (2b), the mean is the location of a vertical line $x = l$ that renders equal the shaded areas of two regions—one to its left and bounded by itself, the horizontal line y=0,

and the (E)CDF, and the other to its right and bounded by itself, the horizontal line y=1, and the (E)CDF. See Figure 2. For a detailed proof see Sarkar and Rashid (2016 a).

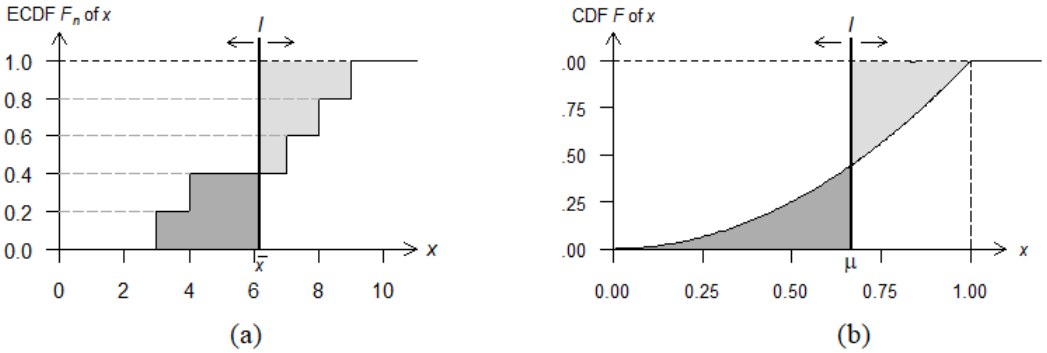


Figure 2. The mean shown as a vertical line $x = l$ that renders equal areas of two shaded regions: (a) for the data in Example 1 and (b) for the CRV in Example 2

3. Visualizing the MD, the MSD and the SD for the Given Dataset

The deviations of the n numbers in the dataset from their mean are $d_i = |x_i - \bar{x}|$. The average of all deviations from the mean is called the MD, and is given by

$$MD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = \frac{1}{n}\sum_{i=1}^{n} d_i \qquad (3)$$

To visualize the MD, we can first construct the ECDF G_n of the deviations. This is done by simply reflecting the portion of F_n to the left of the vertical line $x = \bar{x}$ at the mean, about that line, with the reflection falling to the right side of this line, and then sorting the resultant rectangles of heights 1/n, with the narrowest at the bottom and the widest at the top. However, the sorting of rectangles is ony optional. See Figure 3(a), which implements sorting, and Figure 3(b), which skips sorting. To find the MD, we search for a vertical line $d = \bar{d}$ that equalizes the areas of regions to its left and right that are bounded by the line itself, the horizontal lines y=0 and y=1, and the ECDF G_n, or the reflected ECDF F_n (without sorting).

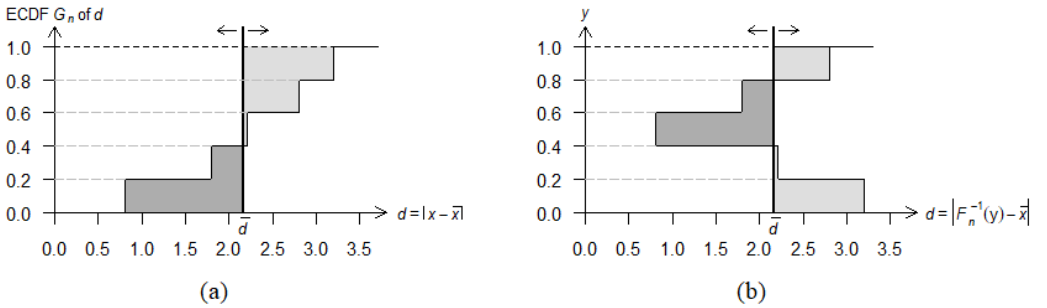


Figure 3. Using either (a) the ECDF G_n of deviations, or (b) the reflected ECDF F_n of the data we obtain the MD for the data in Example 1

Let us next review the geometric visualization of the MSD and the SD for a set of n given numbers. The sample variance of the set of numbers is defined by

$$\text{Variance} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (4a)$$

and the sample MSD is defined by

$$\text{MSD} = \tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (4b)$$

Thus, the sample variance is just a multiple $\alpha = n/(n-1)$ of the sample MSD, since

$$s^2 = \frac{n}{n-1}\tilde{s}^2 = \alpha\,\tilde{s}^2 \qquad (4c)$$

Taking the positive square roots of (4a) and (4b), we obtain the sample SD s and the sample RMSD $\tilde{s}$ respectively. For various interpretations of $\tilde{s}$ and s, see Sarkar and Rashid (2016 a–c).

For a geometric visualization of the MSD, one can construct the ECDF H_n of (scaled) squared deviations as explained below. Proofs are given in Sarkar and Rashid (2016 d).

The ECDF G_n of the deviations form a collection $\boldsymbol{\mathcal{R}}$ of rectangles whose widths equal the deviations and heights equal $1/n$. We transform each rectangle in $\boldsymbol{\mathcal{R}}$ by changing only its width, but keeping it left aligned at d=0 and maintaining its height unaltered as follows: Choose R to be a suitable positive magnitude (for example, let R be the largest deviation from the mean), and fix it. Let d be the width of any one rectangle in $\boldsymbol{\mathcal{R}}$. We construct the third proportional to R and d; that is, we seek a value v such that $R: d = d: v$. Thus, a rectangle of width d changes into a new rectangle of width $v = d^2/R$. Applying this width-transformation to each rectangle in $\boldsymbol{\mathcal{R}}$, using the same R, we obtain the ECDF H_n of the scaled (that is, divided by R) squared deviations. Henceforth, the horizontal axis also represents $v = d^2/R$.

Over H_n, shown in Figure 4(a), we superimpose the vertical line $v = \bar{v}$ that equalizes the areas of the shaded regions to its two sides and bounded by itself, two horizontal lines y=0, y=1 and H_n. Then the vertical line $v = \bar{v}$ represents the scaled MSD given by

$$\bar{v} = \frac{\sigma^2}{R} = \int_0^1 {H_n}^{-1}(y)dy \qquad (5)$$

Finally, to obtain the (unscaled) RMSD, we construct the mean proportional between $\bar{v}$ and R, as explained in the paragraph below, since

$$\sqrt{\bar{v}\,R} = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{\infty}\frac{d_i^2}{R}\right)R} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2} = RMSD = \tilde{s} \qquad (6)$$

Indeed, to construct the mean proportional $\sqrt{ab}$ between a and b (with $a > b > 0$), we draw a right triangle with one leg $(a-b)/2$ and hypotenuse $(a+b)/2$. Then the other leg of that right triangle has length $\sqrt{ab}$. Such a right triangle, showing the mean proportional between $\bar{v}$ and R, is depicted below the horizontal axis in Figure 4(a).

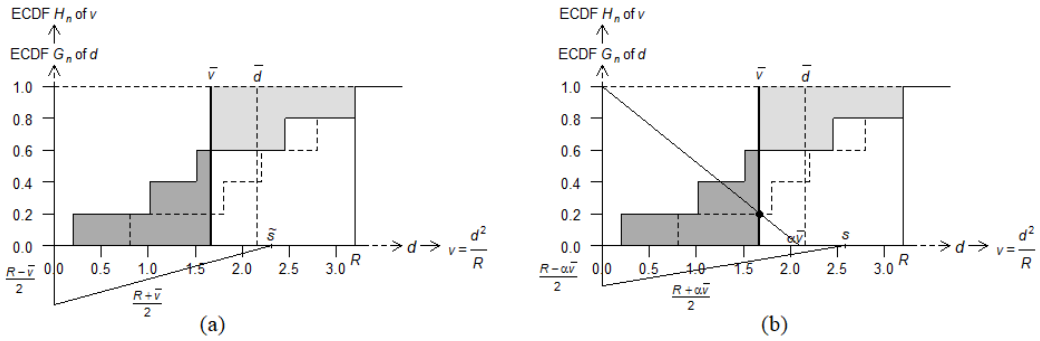


Figure 4. (a) The scaled MSD and the (unscaled) RMSD $\tilde{s} = \sqrt{\bar{v}.R}$ and (b) the scaled MSD and (unscaled) SD $s = \sqrt{\alpha\bar{v}.R}$ for the data in Example 1

Also, starting from Figure 4(a), if we join (0, 1) to $(\bar{v}, 1/n)$ by a line and extend it to meet the horizontal axis, it will do so at a distance $\frac{n\bar{v}}{n-1} = \alpha\bar{v} = \frac{s^2}{R}$ (which is a scaled variance) from the origin. The mean proportional between $\alpha\bar{v}$ and R gives the (unscaled) SD s. See the right triangle below the horizontal axis in Figure 4(b).

Expression (6) justifies why we can choose R to be any arbitrary positive number. Its effect is eliminated in the end, and we obtain the unscaled RMSD $\tilde{s}$ and the unscaled SD s. However, to avoid needing additional space to draw H_n and to ensure precision in drawing, we recommend choosing R to be the largest deviation from the mean. Alternatively, if one chooses R to be the MD, the above described geometric visualization also vividly demonstrates that $s \geq \tilde{s} \geq MD$.

3. Visualizing the MD, the MSD and the SD for a CRV

For a CRV, the CDF F is a strictly increasing, continuous function, and hence it is invertible. The inverse-CDF F^{-1} can be visualized simply by looking at the set of points $\{(x, F(x)): -\infty < x < \infty\}$, since this set is exactly the same as $\{(F^{-1}(y), y): 0 < y < 1\}$. This point of view is advantageous for visualizing the MD, the MSD and the SD of a CRV even when it is difficult to obtain the CDF G of the deviations and H of the scaled squared deviations.

For a CRV, a typical deviation is $d = |x - \mu|$, the MD is defined as

$$\delta = \int_{-\infty}^{\infty} |x - \mu| f(x)\, dx \qquad (7)$$

and the MSD (also called the variance) is defined as

$$\text{MSD} = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx \qquad (8)$$

As done in Figure 3(a)-(b) for a set of numbers, we can also visualize the MD of a CRV either after constructing the CDF G of the deviations or by simply reflecting the CDF F about the mean vertical line $x = \mu$. Since the set $\{(|x - \mu|, F(x)): -\infty < x < \infty\}$ is the same as $\{(|F^{-1}(y) - \mu|, y): 0 < y < 1\}$, we have equivalent expressions for the MD as

$$\delta = \int_0^1 G^{-1}(y)\, dy = \int_0^1 |F^{-1}(y) - \mu|\, dy = 2 \int_{F(0)}^1 [F^{-1}(y) - \mu]\, dy \qquad (9)$$

Likewise, we can visualize the MSD of a CRV either after constructing the CDF H of the scaled squared deviations or after suitably transforming the reflected CDF F. Since the set $\{((x-\mu)^2, F(x)): -\infty < x < \infty\}$ is the same as $\{((F^{-1}(y)-\mu)^2, y): 0 < y < 1\}$, we have equivalent expressions for scaled variance as

$$\bar{v} = \frac{\sigma^2}{R} = \int_0^1 H^{-1}(y)\, dy = \int_0^1 \frac{[F^{-1}(y)-\mu]^2}{R}\, dy \qquad (10)$$

Finally, by taking the mean proportional between $\bar{v}$ and R, we obtain the SD of the CRV.

Recall from Figure 2(b) the CDF of the CRV X in Example 2. Figure 5 shows the MD, the scaled MSD and the SD of X using the vertical line method applied to CDF G of deviations and CDF H of scaled squared deviations. Figure 6 shows the same quantities based on only $F^{-1}(y)$, without constructing the CDF G or CDF H at all!

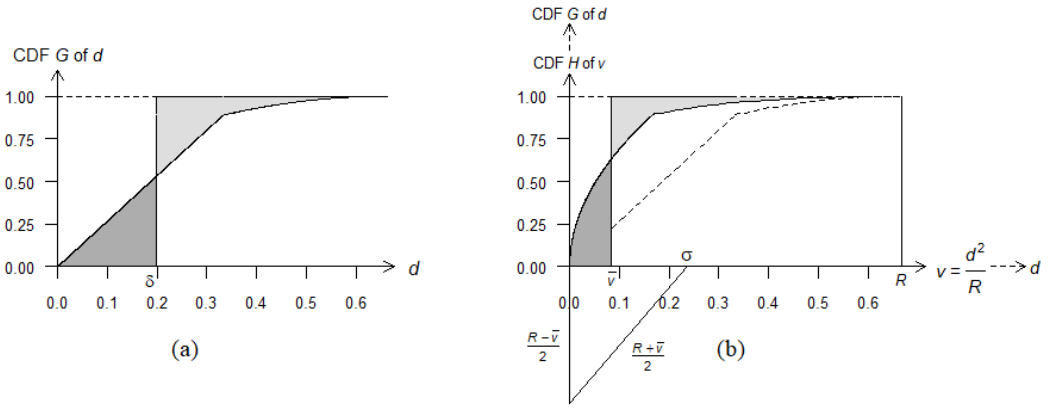


Figure 5. (a) The CDF G of deviations yields the MD, and (b) the CDF H of scaled squared deviations yields the (scaled) MSD= σ^2/R and the (unscaled) SD $\sigma = \sqrt{\bar{v}.R}$ for the CRV in Example 2

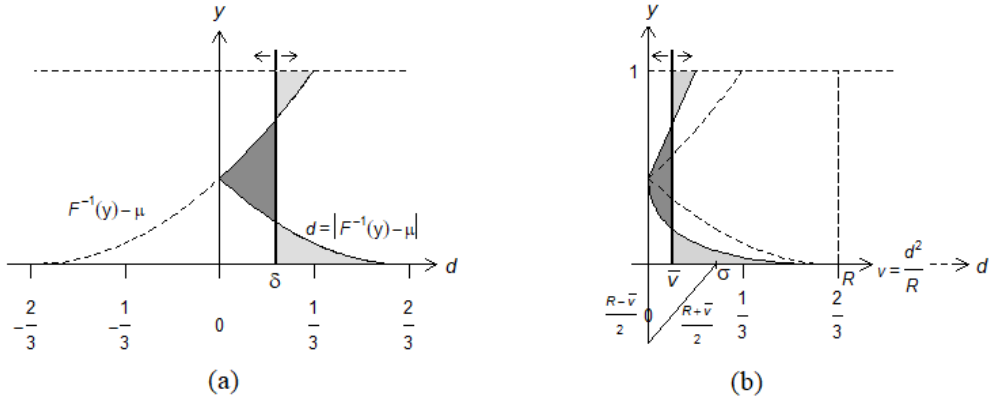


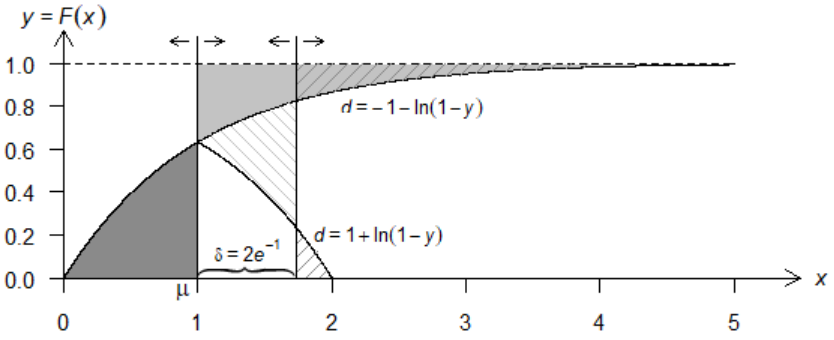
Figure 6. Visualizing (a) the mean and the MD, and (b) the scaled MSD and the (unscaled) SD of the CRV in Example 2, using only F^{-1}, its reflection and its scaled square

4. Discussion

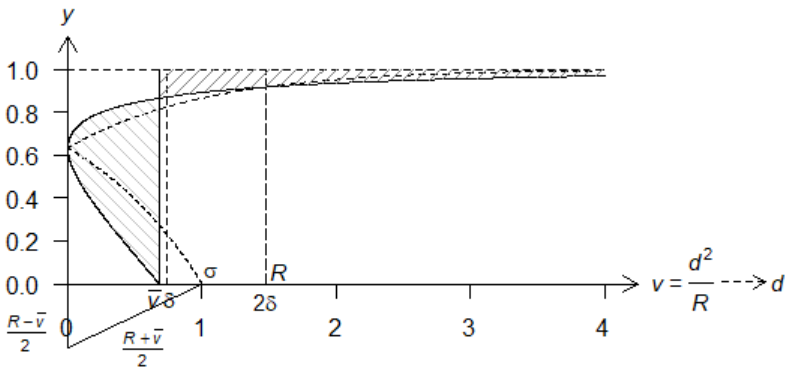
In order to visualize the mean, the MD, the MSD and the SD of a data set, Sarkar and Rashid (2016 d) constructed a vertical line that equalizes the areas of two sets of rectangles. This method utilizes the ECDF F_n of the data, the ECDF G_n of the deviations and the ECDF H_n of the (scaled) squared deviations, all of which are constructible using two-dimensional Euclidean methods. In this paper, we have extended the vertical line method to visualize the mean, the MD and the SD of a CRV on a bounded support. Admittedly, in this case, the CDFs F, G and H are constructed approximately by plotting the functional values against finitely many arguments, and then joining the plotted points freehand. But a statistical software, such as R, can draw these graphs with ease by joining successive points on a finely-gridded scatter plot.

For a CRV on an unbounded support, finding the vertical line that equalizes the areas on its two sides is more challenging, since the area of an unbounded region is difficult to approximate visually. Also, the visualization of the SD requires a wise choice of R. Therefore, we alert the practitioner to take special care in dealing with a CRV on unbounded support. We illustrate the visualization of the mean, the MD and the SD for a CRV on an unbounded support in Example 3, where we have chosen R as twice the MD, although any other choice is permissible.

Example 3. Suppose that a CRV X has the standard exponential(1) distribution. Then its PDF is $f(x) = e^{-x}$, for $0 < x < \infty$; its CDF is $F(x) = 1 - e^{-x}$, for $0 < x < \infty$; and its inverse-CDF is $F^{-1}(y) = -\ln(1-y)$, for $0 < y < 1$.



(a)



(b)

Figure 7. Visualizing (a) the mean and the MD and (b) the SD of an exponential(1) variable, using only the inverse-CDF, its reflection and scaled square

The vertical line method is also useful in demonstrating many other properties of a RV—either discrete or continuous—such as truncation, transformation, skewness, kurtosis, Markov inequality, Chebyshev's inequality, Jensen's inequality, etc.

References

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

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

R Core Team (2013), R: A Language and environment for statistical computing, R-Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Sarkar, J. and Rashid, M. (2016 a), Visualizing Mean, Median, Mean Deviation and Standard Deviation of a Set of Numbers, *The American Statistician*, 70(3), 304-312.

Sarkar, J. and Rashid, M. (2016 b), A Two-Dimensional Visualization of the Standard Deviation, *Journal of Propagation in Probability and Statistics*, 16(1), 13-21.

Sarkar, J. and Rashid, M. (2016 c), Visualizing the Sample Standard Deviation, *Education Research Quarterly*, 40(4), 45-60, 2017.

Sarkar, J. and Rashid, M. (2016 d), Euclidean Plane Geometry Suffices to Visualize the Standard Deviation, *Revision submitted to The College Mathematics Journal*.

Watier, N.N., Lamontagne, C., and Chartier, S. (2011), What does the mean mean? *Journal of Statistics Education,* 19(2).