9-6-2022

SD Prism: Visualizing the Standard Deviation as the Size of a Prism Using R/RStudio

Hieu Nguyen
DePauw University

Tom Nguyen
DePauw University

Mamunur Rashid
DePauw University, mrashid@depauw.edu

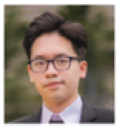
Jyotirmoy Sarkar
Indiana University Purdue University Indianapolis

Recommended Citation

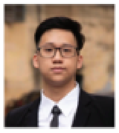
Nguyen, Hieu; Nguyen, Tom; Rashid, Mamunur; and Sarkar, Jyotirmoy, "SD Prism: Visualizing the Standard Deviation as the Size of a Prism Using R/RStudio" (2022). *Student Research*. 47.
https://scholarship.depauw.edu/studentresearchother/47

SD Prism: Visualizing the Standard Deviation as the Size of a Prism Using R/RStudio

Authors



Hieu Nguyen
Department of Computer Science
DePauw University
ORCiD: 0000-0002-4354-2374



Tom Nguyen
Department of Computer Science
DePauw University
ORCiD: 0000-0002-3360-6741

Mamunur Rashid
Department of Mathematical Science
DePauw University
ORCiD: 0000-0001-8759-3803

Jyotirmoy Sarkar
Department of Mathematical Science
Indiana University-Purdue University Indianapolis
ORCiD: 0000-0001-5002-5845

Abstract

SD Prism is a graphical R package used for visualizing the standard de-viation of a data set. Given a raw data set, the standard deviation (SD) is defined as the square-root of the sample variance. Sarkar and Rashid (2016) interpret the sample SD as the square-root of twice the mean square of all pairwise half deviations between any two sample observations. This inter-pretation leads to a geometric visualization of the sample SD, and a more elementary explanation as to why the denominator in the sample variance is *one less* than the sample size. In this article we will explain step by step how to understand it mathematically and how the package implements the methodology to visualize the SD

Keywords: cumulative distribution function, mean, right isosceles triangle, right prism, spread, variance.

1 Introduction

Two most frequently used summary measures in probability and statistics are the mean and the standard deviation (SD). For a randomly selected sample, we view the sample mean as a measure of center and the sample SD as a measure of spread (Lesser, Wagler, and Abormegah, 2014).

Suppose that we have a sample of n numbers $x_1, x_2, \ldots, x_n$. The sample mean (denoted by $\bar{x}$) is defined as the sum of the values divided by the number of values; that is,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

The sample variance (denoted by s^2) is the 'mean' squared deviation of the sample observations from the sample mean $\bar{x}$, and is defined by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2}$$

which can be rewritten as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i^2 - 2\bar{x} \cdot n\bar{x} + n\bar{x}^2 \right] = \frac{1}{n-1} \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right]$$

The sample SD (denoted by s) is the positive square-root of the sample variance given by

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1} \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right]} \tag{3}$$

The expression of the sample SD s is more complicated than that of the sample mean $\bar{x}$. Some people might hastily interpreted the sample SD as a 'typical distance' of each observation from the sample mean, resulting in a clash with the concept of mean (absolute) deviation MAD=$(1/n)\sum|x_i - \bar{x}|$. The accurate reading of the sample SD as the square-root of the 'mean' squared deviation (RMSD) from the sample mean does not explain why the denominator in (2) is $(n-1)$ instead of the often anticipated n. One explanation goes as follows: We choose the denominator in (2) to be $(n-1)$ in order to eliminate the bias involved in estimating the population variance or reduce the bias in estimating the population SD if the denominator were chosen to be n. Here is a second way to think about it: In the data, there are n pieces of information available. To estimate the center of the data, one piece of information is used up. Thereafter, the n differences $(x_i - \bar{x})$ are not all free; they satisfy a linear constraint: their sum equals zero. Accordingly, there remain $(n-1)$ linearly independent pieces of information among the n differences to estimate the spread or the standard deviation. However, neither explanation makes complete sense to the beginner who has not yet learned the notion of estimation and bias or the notion of "degrees of freedom." This paper provides yet another justification that is purely algebraic and is comprehensible to beginning statistics students.

An Equivalent Expression for the Sample Variance

We begin with the definition of deviation between two numbers a and b given by $|a-b|$. Likewise, the deviation of each number a and b from their average $(a+b)/2$, is given by the half-deviation between them, or $|a-b|/2$. To combine all $\binom{n}{2} = n(n-1)/2$ pairwise deviations (PD) (or pairwise half-deviations, PHD) into one single measure of spread of a random sample of size n > 2, we can calculate either the mean or the mean square of these pairwise deviations (or half-deviations). For example, the mean of pairwise deviations (MPD), the mean of pairwise half-deviations (MPHD), the mean square of pairwise deviations (MS-PD), and the mean square of pairwise half-deviations (MS-PHD) are given below.

$$\begin{aligned} \text{MPD} &= \sum_{i=1}^{n}\sum_{j>i}^{n} |x_i - x_j| / \binom{n}{2}, \\ \text{MPHD} &= \sum_{i=1}^{n}\sum_{j>i}^{n} \frac{|x_i - x_j|}{2} / \binom{n}{2}, \\ \text{MS-PD} &= \sum_{i=1}^{n}\sum_{j>i}^{n} (x_i - x_j)^2 / \binom{n}{2}, \\ \text{MS-PHD} &= \sum_{i=1}^{n}\sum_{j>i}^{n} \left(\frac{x_i - x_j}{2}\right)^2 / \binom{n}{2} \end{aligned} \tag{4}$$

It is easy to verify that MPHD = MPD / 2 and MS-PHD = MS-PD / 4. It turns out that MS-PD is exactly twice the sample variance. Hence, the square root of MS-PD (RMS-PD) is $\sqrt{2}$ times the sample SD; that is, RMS-PD= $\sqrt{2}$ s. Equivalently, MS-PHD is half the sample variance as stated below.

Proposition 1. The MS-PHD given in (4) equals half the sample variance given in (2); that is, MS-PHD equals

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \left(\frac{x_i - x_j}{2}\right)^2 / \binom{n}{2} = \frac{s^2}{2} \tag{5}$$

See the proof in Sarkar and Rashid (2016).

We can see that the sample variance is twice the MS-PHD, and the sample SD is the square-root of twice the MS-PHD. Moreover, it is clear that there are

$\binom{n}{2} = n(n-1)/2$ possible pairwise half-deviations on the left hand side of (5) (though not all are distinct). This proof of Proposition 1 shows that the numerator on the left hand side of (5) also has a factor n which cancels with that in the denominator, leaving the other factor $(n-1)$ in the denominator. Thus, it explains why the denominator in (2) is $(n-1)$.

Geometric Visualization of the Sample SD

Proposition 1 lends itself to the creation of a geometric object that represents the sample SD, allowing it to be visualized. The construction is described in five steps. But first, let us explain a typical term on the left side of equation (5). Suppose that we have a right isosceles triangle (RIT) whose hypotenuse equals any typical PD, say $h = |x_i - x_j|$. Then the area of such an RIT is $(\frac{h}{2})^2 = \left(\frac{|x_i - x_j|}{2}\right)^2$. See Fig. 1(a). When this RIT is translated in the three-dimensional space orthogonal to itself through a distance of $w = 1/\binom{n}{2}$, it generates a right prism whose *volume* equals $(\frac{h}{2})^2 w = \left(\frac{|x_i - x_j|}{2}\right)^2 / \binom{n}{2}$, a typical term on the left side of equation (5).

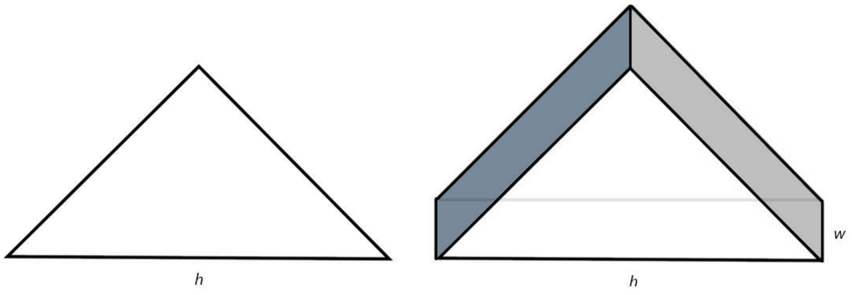


Figure 1: (a) A right isosceles triangle, and (b) a prism generated by translating it

Steps to construct the sample SD geometrically

We document the steps to construct the sample SD. The corresponding figures are shown in the next section as we describe the R package SDPrism2D.

Step 0 (Preliminary preparation): Given the dot plot of x, showing the random sample of *n* values, draw a vertical reference line *l* given by $x = x_{min}$ below the dot plot of *x*; draw also a horizontal reference line (parallel to the *x*-axis) to right of *l* with all other x-values marked on it and surrounded by tiny circles. From each of these marked points draw a line with slope 1 (or the so-called 45° line) that intersects *l* at some point, from which draw a horizontal line to the right of *l* (parallel to the reference line). Mark and surround with a tiny circle every point of intersection so generated which are not on *l*. Then counting row by row, there will be exactly $(n-1)+(n-2)+...+2+1 = \binom{n}{2}$ marked points, each representing a PD in the order of its distance from *l*.

Step 1 (Dot plot of PD): Projecting the marked points (surrounded by tiny circles) vertically up, we construct the dot plot of all PD's, which we depict just above the given dot plot of x. Note that even though the marked points are distinct, some of them may be vertically aligned. Hence, these $\binom{n}{2}$ PD's need not be distinct.

Step 2 (CDF of PD): Suppose that the distinct values of the PD's are $0 < d_1 < d_2 < ... < d_k$ with associated frequencies $f_1, f_2, ..f_k$. Clearly then $f_1 + f_2 + .. + f_k = \binom{n}{2}$. Draw the cumulative distribution function (CDF) of the PD's: It is a step function of the form $y = G(d)$, which begins at height $y = 0$ for $d \leq 0$, has jumps of magnitude $f_k/\binom{n}{2}$ occurring at d_k for $k = 1, 2, ..., K$ and ends at height $y = 1$ for $d \geq d_K$.

Step 3 (Erect right prisms): If we superimpose on the graph of the CDF of PD horizontal lines $y = t/\binom{n}{2}$ for $t = 1, 2, \ldots, \binom{n}{2}$, stretching from l to the CDF, we see $\binom{n}{2}$ rectangles corresponding to the $\binom{n}{2}$ PD's. Each rectangle has width given by a PD and height $1/\binom{n}{2}$. Using each of these $\binom{n}{2}$ rectangles as the xy-face, we erect a right prism of y-thickness $1/\binom{n}{2}$ and xz-cross-section given by an RIT whose hypotenuse equals the width of the rectangle (or the corresponding PD). Then the total volume V_+ of all $\binom{n}{2}$ such right prisms is precisely the left hand side of expression (5), or the MS-PHD. Therefore, in view of Proposition 1, the total volume V_+ also equals half the sample variance. Optionally, rectangles that are equally wide (representing duplicate PD's) may be joined together by removing the internal horizontal lines.

Step 4 (Search for a composite right prism) : Consider a single composite right prism having y-thickness 1 and a xz-cross-section in the shape of an RIT. Allow the size of this RIT to vary by changing the x-size until we find a suitable size such that the volume of the single composite right prism equals V_+. The intermediate value theorem guarantees the existence and uniqueness of this composite prism. Then each leg of the RIT of the composite prism equals the sample SD, and the hypotenuse of that RIT is $\sqrt{2}$ times the sample SD, or the RMS-PD.

2 The SDPrism2D Package

The SDPrism2D function admits three arguments, with only one mandatory argument:

- `data` the user must input the data set, usually a vector of values.

- `hlim` the height limit for the prism figure, which is 4 by default if not specified by the user. If the upper parts of one or more prisms are cut off, the user can increase hlim. If there is a big white space above the largest prism, the user can decrease hlim.

- `xyscale` the ratio of scales between the x-axis and the y-axis, which is 4 by default (that is, the physical spread of the x-axis equals 4 times that of the y-axis) if not specified. the user can change it as desired.

For illustration, we use a sample data set with 5 values: 10, 18, 23, 30, 36. Create an empty vector for all pairwise deviation (PD). A pair of nested 'for' loops is used to calculate all the PD's of the data set and append them to the vector.

```
pd <- c()
cnt = 1
for (i in data[1:length(data)-1]){
  for (j in data[0:-cnt]){
    pd <- append(pd, abs(j-i))
  }
  cnt <- cnt + 1
}
```

Open a new window to display 4 rows containing one plot each.

```
dev.new(width=6,height=8)
par(mfrow=c(4,1), mai = c(.15,2,0,2))
```

Plot the data set on a horizontal axis.

```
stripchart(data,method="stack",offset=.2,pch=19,frame.plot
= FALSE,xaxt="n", cex=1.4)
```

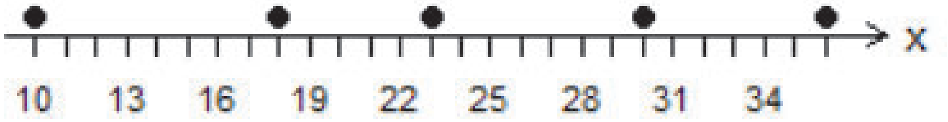


Figure 2: A dotplot of the data set

Plot all pair-wise deviations on a horizontal axis.

```
stripchart(pd,method="stack",offset=.2,pch=19,frame.plot
= FALSE,xaxt="n",cex = 1.4,xlim=c(0,max(pd)))
```

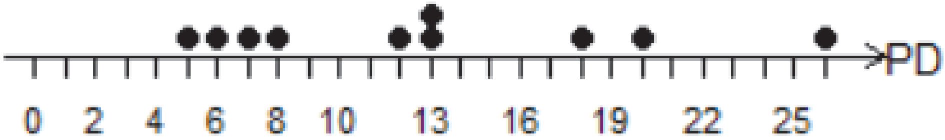


Figure 3: A dotplot of the pairwise differences (PD)

Sort all the values in pd in ascending orders. Create a vector, named F, of cumulative sums of frequencies in f.

```
pd<-sort(pd)
n<-length(pd)
f<-rep(1/n,n)
F<-cumsum(f)
```

Draw the Cumulative Distribution Function (CDF) of all PD's. Draw the vertical line that equalizes the areas of the colored regions to its left and right and represents the mean of all the PD's (MPD).

```
plot(pd, F, type = "s", xlab = "", ylab = "", frame.plot
= FALSE, las = 1, xaxs = "r", yaxs = "r", xlim = c(0,max(pd)),
ylim = c(0,hlim), xaxt = "n", yaxt = "n")
m<-mean(pd)
segments(m, 0, m, 1.1,lty = 1, lwd = 2, xpd = T)
```

Define a function 'rprism' needed to draw a right isosceles prism.

```
rprism <- function(hyp, xstart, y, t, ratio) {
  x1<-hyp/2
  x2<-x1/ratio
  segments(xstart , y, hyp, y, lty = 1, lwd = 1.75)
  segments(hyp, y, hyp, 0, lty = 2, lwd = 1, col = "gray")
```

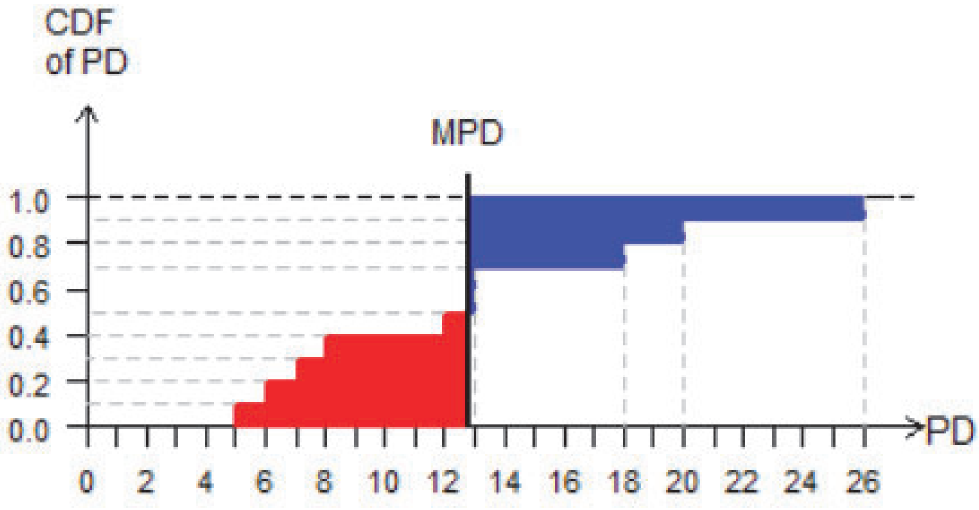


Figure 4: The Cumulative Distribution Function of the PD's together with the mean PD vertical line that equalizes the areas of the colored regions

```
  segments(hyp, y, x1, y+x2, lty=1, lwd = 2)
  segments(x1, y+x2, 0, y, lty = 1, lwd = 2)
  segments(0, y, 0, y+t, lty = 1, lwd = 1.75)
  segments(0, y+t, x1, y+x2+t,lty = 1, lwd = 1.75)
  for(i in c(seq(0,x1,.01))) {
  segments(i, y+i/ratio+.015, i, y+i/ratio+t-.015, lty =
1, col = "#778899")
  }
  for(i in c(seq(x1,hyp,.01))){
  segments(i, y+(hyp-i)/ratio+.015, i, y+(hyp-i)/ratio+t-.015,
lty = 1, col = "#C0C0C0")
  }
  segments(x1, y+x2+t, hyp, y+t, lty = 1, lwd = 1.75)
  segments(hyp, y+t, hyp, y, lty = 1, lwd = 1.75)
  segments(x1, y+x2, x1, y+x2+t, lty = 1, lwd = 1.75)
}
```

Draw all right prisms with each PD being the hypotenuse of the corresponding right triangular surface and with thickness $1/\binom{n}{2}$, using the inner function rprism.

```
yinit <- 0
prev <- 0
for (i in pd) {
  rprism(i,prev,yinit,1/n,r)
  yinit <- yinit + 1/n
  prev = i
}
```

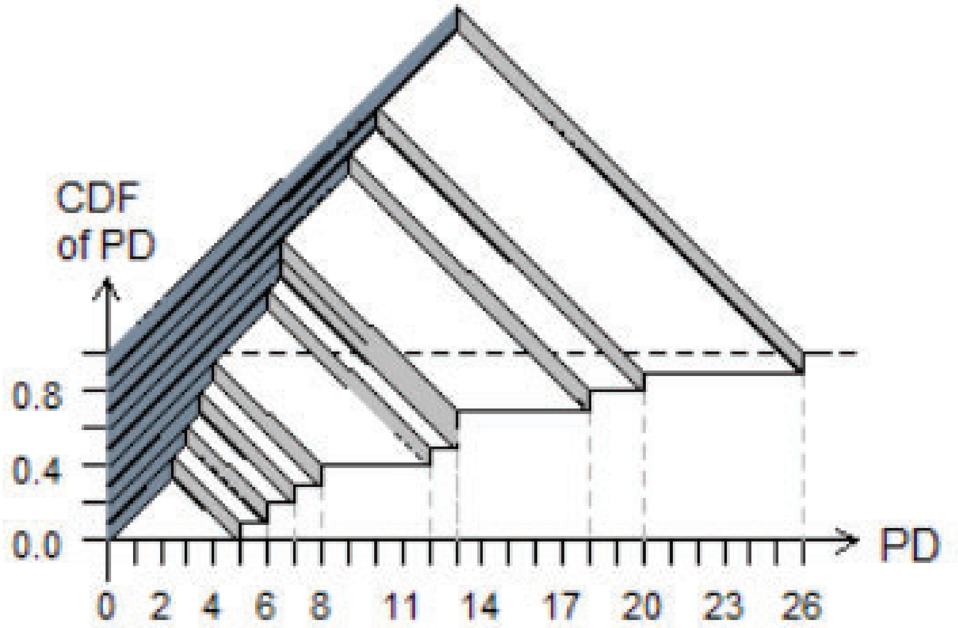


Figure 5: On the rectangles to the left of the CDF shown in Figure 4, erect right isosceles prisms

Let the total volume of these prisms be V_+. We want to construct a single right isosceles prism with thickness 1 and sufficiently large so that its volume equals V_+. Then each leg of the right isosceles triangle equals the standard deviation (sd) and the hypotenuse equals $\sqrt{2}$ times the sd. The construction is done by using the inner function rmspdprism. This function is very similar to the rprism function, except the thickness is always 1.

```
spd <- 0
for(i in pd) {
  spd <- spd + i * i
}
mspd <- spd/n
rmspd<-sqrt(mspd)
rmspdprism(rmspd,r)
```

The final product should look like below:

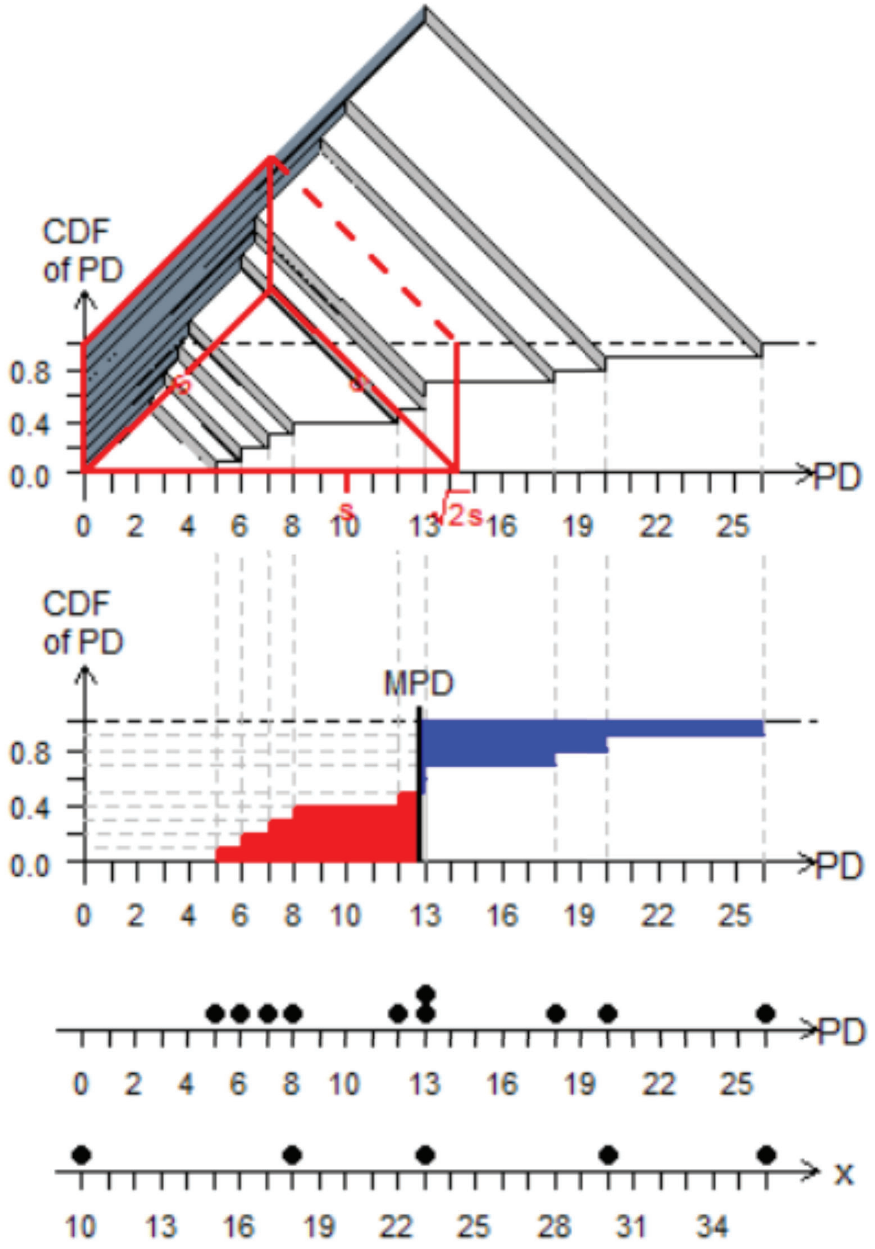


Figure 6: The two equal legs of the single composite right isosceles prism (shown in red) represent the standard deviation of the data

3 The SDPrism3D Package

The SDPrism3D function enhances the output of the SDPrism2D function, and produces a 3D object which the user can rotate and visualize from any desired direction. It requires a third-party software called XQuart (freely accessible from https://www.xquartz.org/) to generate the 3D object. It also invokes two R packages `rgl` and `plot3d`. The function takes three arguments, with only one mandatory argument:

- `data` the user must input the data set, usually a vector of values.

- `PriCol`: The color for the individual prisms. Users can use a number or a word to specify the color they desire. The default color is lightblue.

- `sdCol`: The color for the composite prism whose size represents the Standard Deviation. Users can use a number or a word to specify the color they desire. The default color is red.

For the first step, the function will take the data and create a vector of all pairwise deviations.

```
for (i in 1 :  (length(data)-1) {
   for (j in (i+1) :  length(data)) {
     add <- data[i] - data[j]
     PD <- append(PD, abs(add))
   }
}
PD <- sort(PD)
```

We calculate the coordinates of the vertices in order to draw the prisms. To begin, we create empty vectors corresponding to each dimension of the prisms.

```
x <- c()
y <- c()
z <- c()
```

We also need two supporting variables.

`curStep`: for the jump

`thelwd`: for the thickness of the lines graphing the prisms

The second step calculates all coordinates of vertices of the prisms and stores them into corresponding vectors.

```
for (r in 1:length(PD)) {
   len <- PD[r]
   hyp <- len/2

   xAdd <- c(0,len,len,0,hyp,hyp)
   x <- c(x,xAdd)

   yAdd <- c(curStep/length(PD), curStep/length(PD),
         (curStep+1)/length(PD), (curStep+1)/length(PD),
         curStep/length(PD), (curStep+1)/length(PD))
   y <- c(y,yAdd)
   zAdd <- c(0,0,0,0,hyp,hyp)
   z <- c(z,zAdd)

   curStep <- curStep + 1
}
```

Next, we calculate the coordinates of the composite prism representing the SD and store them at the end of the vectors.

```
x <- c(x, 0, sqrt(2)*SD, sqrt(2)*SD, 0, sqrt(2)*SD/2,
sqrt(2)*SD/2)
y <- c(0, 0, 1, 1, 0, 1)
z <- c(z, 0, 0, 0, 0, sqrt(2)*SD/2, sqrt(2)*SD/2)
```

The fourth step is to adjust meaningful scales for the dimensions of the interactive plot. In the 3D environment, every dimension is proportional to one another. This might distort the 3D object. To solve that, we ensure that all dimensions spread equally: The *x*- direction spreads from 0 to the largest PD (which is the last element of the PD vector after sorting); the *y*-direction goes from 0 to 1, and the *z*-direction is similar to the *x*-direction.

```
xdim <- PD[length(PD)]
x <- c(x,xdim)
ydim <- 1
y <- c(y,ydim)
```

```
zdim <- xdim
z <- c(z,zdim)
```

The final step is to plot the graph in 3D environment.

```
s1 <- plot3d(x,y,z)
```

We draw the prisms by connecting the corresponding coordinates.

```
for(c in 1 :  (length(PD) + 1)) {
   if (c == length(PD) + 1) {
      PriCol <- sdCol
      thelwd = 4
      segments3d(c(x[4],x[1]), c(y[4],y[1]), c(z[4],z[1]),
      lwd = thelwd, col = PriCol)
   }
   segments3d(c(x[1],x[2]), c(y[1],y[2]), c(z[1],z[2]),
   lwd = thelwd, col = PriCol)
   segments3d(c(x[2],x[3]), c(y[2],y[3]), c(z[2],z[3]),
   lwd = thelwd, col = PriCol)
   segments3d(c(x[3],x[4]), c(y[3],y[4]), c(z[3],z[4]),
   lwd = thelwd, col = PriCol)
   segments3d(c(x[1],x[5]), c(y[1],y[5]), c(z[1],z[5]),
   lwd = thelwd, col = PriCol)
   segments3d(c(x[5],x[2]), c(y[5],y[2]), c(z[5],z[2]),
   lwd = thelwd, col = PriCol)
   segments3d(c(x[5],x[6]), c(y[5],y[6]), c(z[5],z[6]),
   lwd = thelwd, col = PriCol)
   segments3d(c(x[6],x[3]), c(y[6],y[3]), c(z[6],z[3]),
   lwd = thelwd, col = PriCol)
   segments3d(c(x[6],x[4]), c(y[6],y[4]), c(z[6],z[4]),
   lwd = thelwd, col = PriCol)

   if (c < length(PD) + 1) {
      polygon3d(c(x[1], x[5], x[6], x[4]),
      c(y[1], y[5], y[6], y[4]), c(z[1], z[5], z[6], z[4]))
      polygon3d(c(x[5], x[2], x[3], x[6]),
      c(y[5], y[2], y[3], y[6]), c(z[5], z[2], z[3], z[6]))
   }
   x <- [-c(1:6)]
   y <- [-c(1:6)]
   z <- [-c(1:6)]
}
```

Here is our finished product:

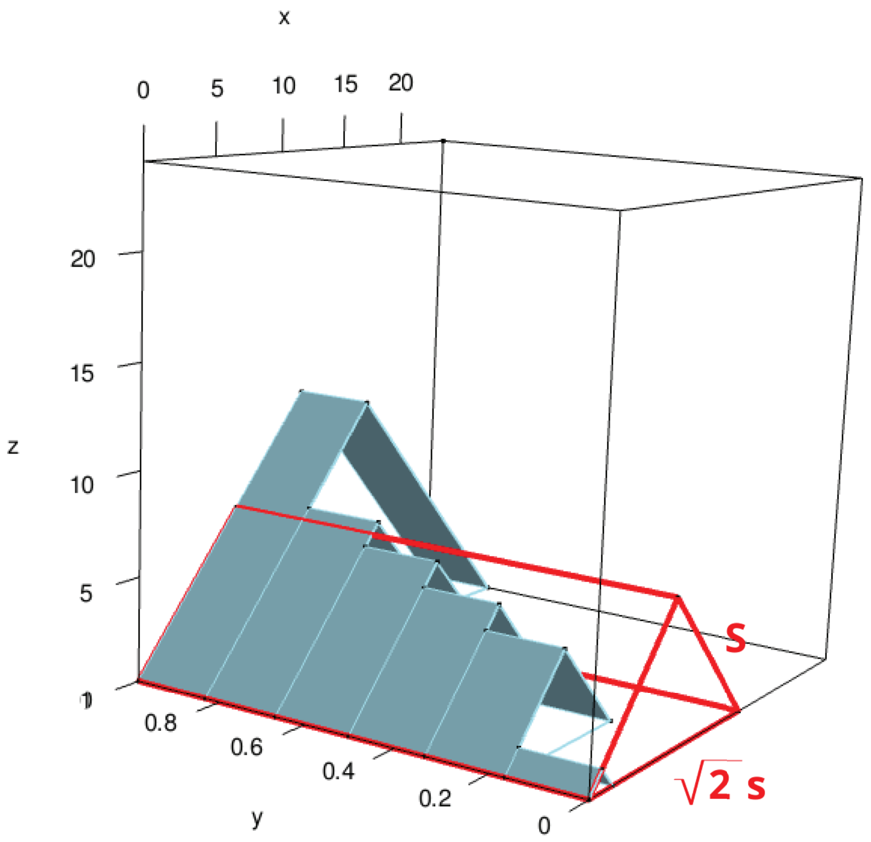


Figure 7: 3D view of the family of prisms and composite prism presenting the standard deviation

Acknowledgments

We give special thanks to J. William Asher and Melanie J. Norton Endowed Fund in the Sciences for sponsoring our project. Thanks to Science Research Fellow and Student-Faculty Summer Research collaborator for this opportunity. The student authors thank the faculty authors for their guidance and support.

References

[1] Lesser, L.M., Wagler, A.E., Abormegah, P. (2014). Finding a Happy Median: Another Balance Representation for Measures of Center. *Journal of Statistics Education*, 22(3), 1-27. Retrieved from www.amstat.org/publications/jse/v22n3/lesser.pdf

[2] R Core Team (2021). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org/

[3] Sarkar, J. Rashid, M. (2016). 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