# The Standard Deviation

If we review our measures of the spread of data, we have the range, that is the low and the high values, the quartile points, $Q_0$, $Q_1$, $Q_2$, $Q_3$, and $Q_4$ which include the range, and percentiles. These are all related to the order of the values in the data list. Our measures of central tendency were the mean, the median, and the mode, where the median was the one measure related to the order of values. Our next measure of the spread of the data will be a measure that is related to the mean. That new measure is the standard deviation.

The standard deviation of a population is symbolized via the Greek letter sigma, $\boldsymbol{\sigma}$. The computation behind the standard deviation is really encompassed by its full name, "the root mean squared deviation from the mean". That is, if we have a data list such as A=[4,8,6,13,8,2,8], we first find the mean of that list of values, symbolized as the Greek letter mu, $\boldsymbol{\mu}$. Then we find the deviation of each item in the list from the mean. That entails finding $\mathbf{x}_i - \boldsymbol{\mu}$ for each $\mathbf{x}_i$ in the list. Having done that, we have found the "deviations from the mean". Then we want to find the "squared deviations from the mean". To do this we simply square each of the values that we just found. Now we have a list of squared deviations from the mean and we compute the "mean" of those values by finding their sum and dividing by the number of values. At that point we have the "mean squared deviations from the mean". All that is left to do is to find the square root of that value.

For the data list that we have, A=[4,8,6,13,8,2,8], we could construct the following table:

| i | $\mathbf{x}_i$ | $\mathbf{x}_i - \boldsymbol{\mu}$ | $(\mathbf{x}_i - \boldsymbol{\mu})^2$ |
|---|---|---|---|
| 1 | 4 | -3 | 9 |
| 2 | 8 | 1 | 1 |
| 3 | 6 | -1 | 1 |
| 4 | 13 | 6 | 36 |
| 5 | 8 | 1 | 1 |
| 6 | 2 | -5 | 25 |
| 7 | 8 | 1 | 1 |
| Sum | 49 | | 74 |
| mean | 7 | | 10.571428… |

All that remains is to find the square root of 10.571428 and that is approximately 3.2513733. That is the standard deviation for the data list A.

Mathematically, the computation is written as

$$\sigma = \sqrt{\frac{\sum_1^n (x_i - \mu)^2}{n}}$$

where the $\sum$ character, the upper case Greek letter sigma, indicates the sum of the values from i=1 to i=n.

There are many equivalent formulae

. for the computation of the standard deviation. One such formula is

$$\sigma = \sqrt{\frac{\sum (x_i)^2}{n} - \left(\frac{\sum x_i}{n}\right)^2}$$

This formula is easier to compute because we only need to find the sum of the values and the sum of the squares of the values. Here is a table that does this:

| i | $\mathbf{x}_i$ | $(\mathbf{x}_i)^2$ |
|---|---|---|
| 1 | 4 | 16 |
| 2 | 8 | 64 |
| 3 | 6 | 36 |
| 4 | 13 | 169 |
| 5 | 8 | 64 |
| 6 | 2 | 4 |
| 7 | 8 | 64 |
| sum | 49 | 417 |

Then $417/7-(49/7)^2 = 59.571428-49 = 10.571428$, and if we find the square root of that we have the same 3.2513733 approximation.

The standard deviation is a powerful indicator of how closely the data is packed around the mean. For the data list B=[1,8,2,20,8,2,8] we still have the mean=7, but now the standard deviation is approximately 6.071. For the data list C=[-44,8,-25,100,8,-6,8], which still has the mean=7, the standard deviation is approximately 42.166. Finally, for the data list D=[5,8,5,10,8,6,8], where the mean remains 7, the standard deviation is now approximately 1.8516.

Chebyshev's inequality gives us a direct way to interpret the standard deviation. According to this inequatilty, it is always

the case that the fraction of all the data points in the list within n standard deviations will be $1-1/n^2$. Thus, within 2 standard deviations of the mean there will always be $1-1/2^2$ or 1-1/4 or ¾ of all the data points in the list. Within 3 standard deviations of the mean there will always be $1-1/3^2$ or 1-1/9 or 8/9 of all the data points in the list. Chebyshev's inequality works for any data list. In fact it is impossible to even construct a list of values for which this inequality is not true. In our list A above, A=[4,8,6,13,8,2,8], the mean was 7 and the standard deviation was about 3.2513733. We are therefore certain that at least 75% of the values are within 2 standard deviations of 7, that is between 7-2*3.2513733 and 7+2*3.2513733, or between 0.4972534 and 13.5027466. Indeed, for our 7 values, 6 are in that interval and 6 of the 7 values represents 85.14% of the values.

In populations where measurements are more normally distributed, measurements like height, or IQ, or SAT scores, that is, where we have what is known as a bell-shaped distribution of values, we can be even more restrictive in interpreting the standard deviation. For a normally distributed list of values, we are confident that 68% of the values will be within 1 standard deviation of the mean, 95% of the values will be within 2 standard deviations of the mean, and 99% of the values will be within 3 standard deviations of the mean