# Explaining the difference between population variance and sample variance

Many laboratory analysts cannot see the logic of the difference in the equations used for population variance and sample variance which lead to their respective standard deviation. Perhaps the following explanations will clear the doubt.

When we have a set of data on a population or sample, we can calculate its mean (or average) easily. The measures of dispersion or *deviation* of these data from the mean value are the *variance* and *standard deviation* which describe how much the individual value in a data set vary from the mean or average value.

For example, given a set of $N$ values ( $x_1, x_2, \ldots x_i, \ldots, x_N$ ) in a population, we have a mean of $\mu = \sum_{i=1}^{N} x_i$ and the sum of all deviations $D = \sum_{i=1}^{N} (x_i - \mu)$. But this sum of all deviations returns a value of zero because the difference of each value $x_i$ and the mean would give either positive or negative deviation.

This fact can be illustrated clearly by the following hypothetical set of data in a population, (*1.0, 2.0, 3.0, 4.0, 5.0* ). The mean value is 3.0 and the sum of deviation

$$D = (1.0-3.0)+(2.0-3.0)+(3.0-3.0)+(4.0-3.0)+(5.0-3.0) = 0$$

Hence, the sum of deviations does not tell us much about how good or how bad is the spread of these results from the mean value. Since a square of the deviation always give a positive result, the sum of squares of deviations should exhibit the extent of the data spread. If we were to have taken $N$ samples from the population for such measurements, then averaging the sum of deviation by dividing this sum with the number of samples which are independently drawn from the population would give us the population variance $\sigma^2$, viz

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

and, the population standard deviation is of course the square root of variance, i.e. $\sigma$. For the population set of data (*1.0, 2.0, 3.0, 4.0, 5.0* ), the population variance $\sigma^2=2.0$ and hence the population standard deviation $\sigma = 1.4$.

In the discussion of sample variance $s^2$ and sample standard deviation $s$ for a set of $n$ repeated measured values found in a sample, we have a sample mean of $\bar{x} = \sum_{i=1}^{n} x_i$ and the sum of deviation $D = \sum_{i=1}^{n} (x_i - \bar{x})$. The sample variance $s^2$ is equal to the sum of deviation divided by a *degree of freedom* ( $n$-$1$ ) instead of $n$ number of repeated values. Why? This is because the sample mean value $\bar{x}$ would no longer be the same if any one of the $n$ repeated measured values in the sample was altered. In other words, the $n$ repeated measurements in a sample are not full independent to each other but have $n - 1$ degrees of freedom for us to estimate its sample variance. The sample variance therefore is expressed as

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

If the repeated values of a sample are (*1.0, 2.0, 3.0, 4.0, 5.0* ), the sample variance $s^2$=2.5 and its sample standard deviation $s$ = 1.6. .

Another observation is that if the $n$ becomes very large, (say, $n$ > 30) the difference between s and $\sigma$ is getting so small that can be neglected without serious consequence.

: