



Variance & Standard Deviation

Key Info

It would be useful to have a measure of scatter that has the following properties:

1. The measure should be proportional to the scatter of the data (small when the data are clustered together, and large when the data are widely scattered).
2. The measure should be independent of the number of values in the data set (otherwise, simply by taking more measurements the value would increase even if the scatter of the measurements was not increasing).
3. The measure should be independent of the mean (since now we are only interested in the spread of the data, not its central tendency).

Both the **variance** and the **standard deviation** meet these three criteria for normally-distributed (symmetric, "bell-curve") data sets.

The variance (σ^2) is a measure of how far each value in the data set is from the mean. Here is how it is defined:

1. Subtract the mean from each value in the data. This gives you a measure of the distance of each value from the mean.
2. Square each of these distances (so that they are all positive values), and add all of the squares together.
3. Divide the sum of the squares by the number of values in the data set.

The standard deviation (σ) is simply the (positive) square root of the variance.

The Summation Operator

In order to write the equation that defines the variance, it is simplest to use the **summation operator**, Σ . The summation operator is just a shorthand way to write, "Take the sum of a set of numbers." As an example, we'll show how we would use the summation operator to write the equation for calculating the mean value of data set 1. We'll start by assigning each number to variable, X_1 - X_6 , like this:

data set 1	
variable	value
X_1	3
X_2	4
X_3	4
X_4	5
X_5	6
X_6	8

Think of the variable (X) as the measured quantity from your experiment—like number of leaves per plant—and think of the subscript as indicating the trial number (1–6). To calculate the average number of leaves per plant, we first have to add up the values from each of the six trials. Using the summation operator, we'd write it like this:

$$\sum_{i=1}^6 X_i$$

which is equivalent to:

$$X_1 + X_2 + X_3 + X_4 + X_5 + X_6$$

or:

$$3 + 4 + 4 + 5 + 6 + 8 = 30$$

Obviously the sum is a lot more compact to write with the summation operator. Here is the equation for calculating the mean, μ_x , of our data set using the summation operator:

$$\mu_x = \frac{\sum_{i=1}^6 X_i}{6} = \frac{30}{6} = 5$$

The general equation for calculating the mean, μ , of a set of numbers, $X_1 - X_N$, would be written like this:

$$\mu = \frac{\sum_{i=1}^N X_i}{N} = \frac{\sum X}{N}$$

Sometimes, for simplicity, the subscripts are left out, as we did on the right, above. Doing away with the subscripts makes the equations less cluttered, but it is still understood that you are adding up all the values of X .

The Equation Defining Variance

Now that you know how the summation operator works, you can understand the equation that defines the variance:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

The variance (σ^2), is defined as the sum of the squared distances of each term in the distribution from the mean (μ), divided by the number of terms in the distribution (N).

There's a more efficient way to calculate the standard deviation for a group of numbers, shown in the following equation:

$$\sigma^2 = \frac{\sum X^2}{N} - \mu^2$$

You take the sum of the squares of the terms in the distribution, and divide by the number of terms in the distribution (N). From this, you subtract the square of the mean (μ^2). It's a lot less work to calculate the standard deviation this way.

It's easy to prove to yourself that the two equations are equivalent. Start with the definition for the variance (Equation 1, below). Expand the expression for squaring the distance of a term from the mean (Equation 2, below).

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} \quad (1)$$

$$= \frac{\sum (X^2 - 2\mu X + \mu^2)}{N} \quad (2)$$

$$= \frac{\sum X^2}{N} - \frac{2\mu \sum X}{N} + \frac{N\mu^2}{N} \quad (3)$$

$$= \frac{\sum X^2}{N} - 2\mu^2 + \mu^2 \quad (4)$$

$$= \frac{\sum X^2}{N} - \mu^2 \quad (5)$$

Now separate the individual terms of the equation (the summation operator distributes over the terms in parentheses, see Equation 3, above). In the final term, the sum of μ^2/N , taken N times, is just $N\mu^2/N$.

Next, we can simplify the second and third terms in Equation 3. In the second term, you can see that $\sum X/N$ is just another way of writing μ , the average of the terms. So the second term simplifies to $-2\mu^2$ (compare Equations 3 and 4, above). In the third term, N/N is equal to 1, so the third term simplifies to μ^2 (compare Equations 3 and 4, above).

Finally, from Equation 4, you can see that the second and third terms can be combined, giving us the result we were trying to prove in Equation 5.

As an example, let's go back to the two distributions we started our discussion with:

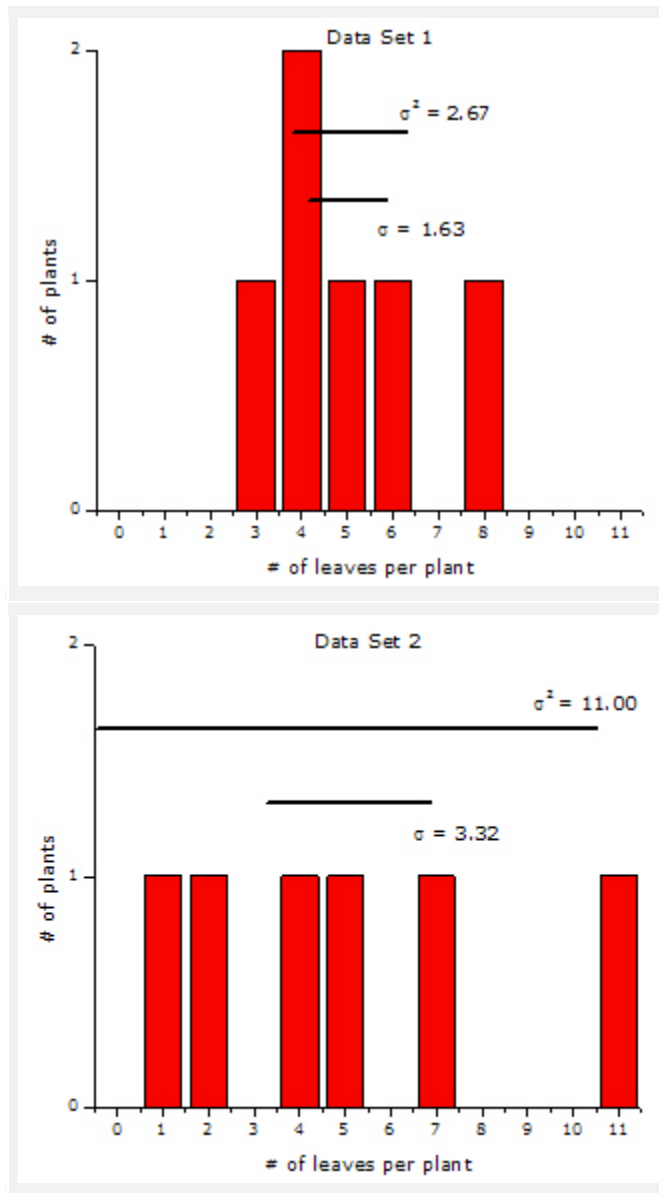
data set 1: 3, 4, 4, 5, 6, 8
 data set 2: 1, 2, 4, 5, 7, 11 .

What are the variance and standard deviation of each data set?

We'll construct a table to calculate the values. You can use a similar table to find the variance and standard deviation for results from your experiments.

data set	N	$\sum X$	$\sum X^2$	μ	μ^2	σ^2	σ
1	6	30	166	5	25	2.67	1.63
2	6	30	216	5	25	11.00	3.32

Although both data sets have the same mean ($\mu = 5$), the variance (σ^2) of the second data set, 11.00, is a little more than *four times* the variance of the first data set, 2.67. The standard deviation (σ) is the square root of the variance, so the standard deviation of the second data set, 3.32, is just over *two times* the standard deviation of the first data set, 1.63.



The variance and the standard deviation give us a numerical measure of the scatter of a data set. These measures are useful for making comparisons between data sets that go beyond simple visual impressions.