# **Variability, Standard Deviation and Bias**

### Variability

There is more variability if your data is distributed far from the mean, less if the data points are close to it!

You can indeed have the same mean but high, or low, variability! As you can see when you run the code below, the values for x and y are not the same, yet the red line representing their respective means, is the same!



Figure 13: code for variability

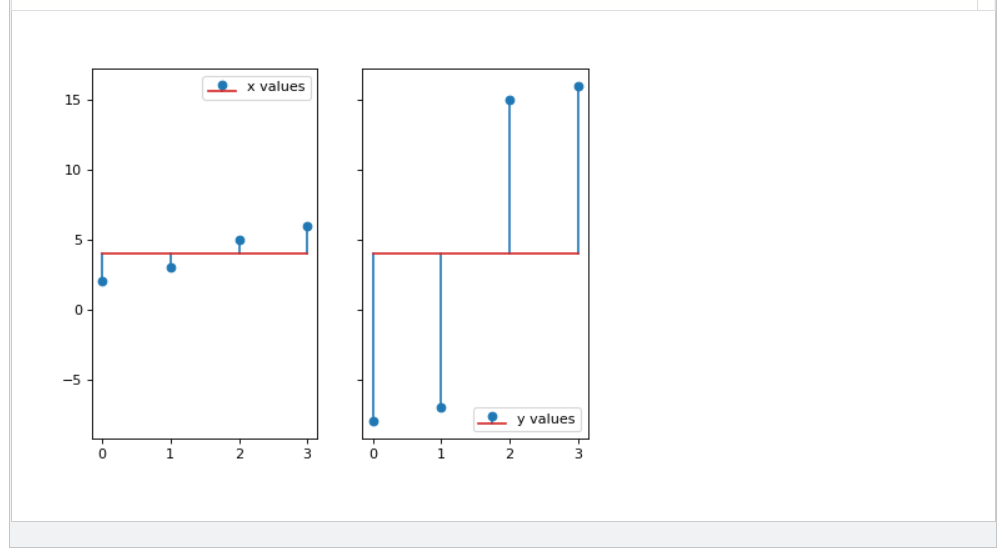
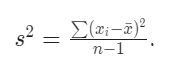


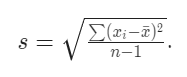
Figure 14: Output for variability problem

Remember this important idea: "the average of deviations is always zero as the sum of the deviations is always zero". And think about why that is. If you look at the above two charts, that may help you figure it out. Intuitively, if you are trying to measure the *variability* of your data around the mean, then by adding all these numbers you will necessarily find zero! (Simply because the mean is, you could say, the middle of all these deviations, where they cancel each other out).

We then use the square of the values to obtain more information: when you square a number, it always becomes positive! The cancellation of all deviations has been avoided. Now our deviations can tell us more about our data, and we can go on calculating the variance:



When we are done with our calculation, and have computed the standard deviation, it is as if we had reached full circle, as we are now taking the square root in our equation, after having squared everything in the first place:



### Don't mix up the letters!

As before:

* when we talk about the **population**, the variance will be denoted by the Greek letter *σ^*2, and we divide by *N*;
* however, when we are dealing with a **sample**, then we use the letter *s^*2 to denote the sample variance and we divide by *n*−1.
* there are other versions where the sample variance is divided by *n* but we do not use this version.

### The bias is back!

Some of you noticed that interesting n-1*n*−1 appearing in the division. As mentioned in the previous reading, this is called [Bessel's correction](https://en.wikipedia.org/wiki/Bessel's_correction). The mathematics for this is not included in this MOOC, but feel free to dig deeper into this, or post on the forum, if you feel like knowing more about it! The main idea to take away is that when you try to estimate the **variance** and/or the **standard deviation** of a population using a **random** **sample**, the result will be **biased**. This bias, which, by the way, carries no value judgement at all, but only indicates that if you repeat this technique an infinite number of times with different random samples, you will not, as you would hope, approximate the population values, but end up with something else, that can be calculated precisely. The division by n-1*n*−1 arises directly from these calculation (if you are hungry for the full details, [you can read them here](https://en.wikipedia.org/wiki/Bias_of_an_estimator#Sample_variance)). Mathematicians, then, having found out what exactly that bias is, deduced what method should be used to correct this, which is precisely to use n-1*n*−1 instead of n*n* in the division!

### Attention!

This will pop up again in this course, but it is good to mention it straight away: the need for this division by n-1*n*−1 only arises when calculating the ***variance*** and ***standard deviation***! If you only estimate the **mean** of a population using a sample, the division by n*n* is the only valid one.