iNeuron Assigment (Why n-1?)

Population dataset: - contains tons of data i.e. millions or even billions of data

Sample dataset: - By using the sampling technique we draw samples from the population dataset and put them in the sample dataset which we consider for analysis and to draw some inferences from it.

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

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

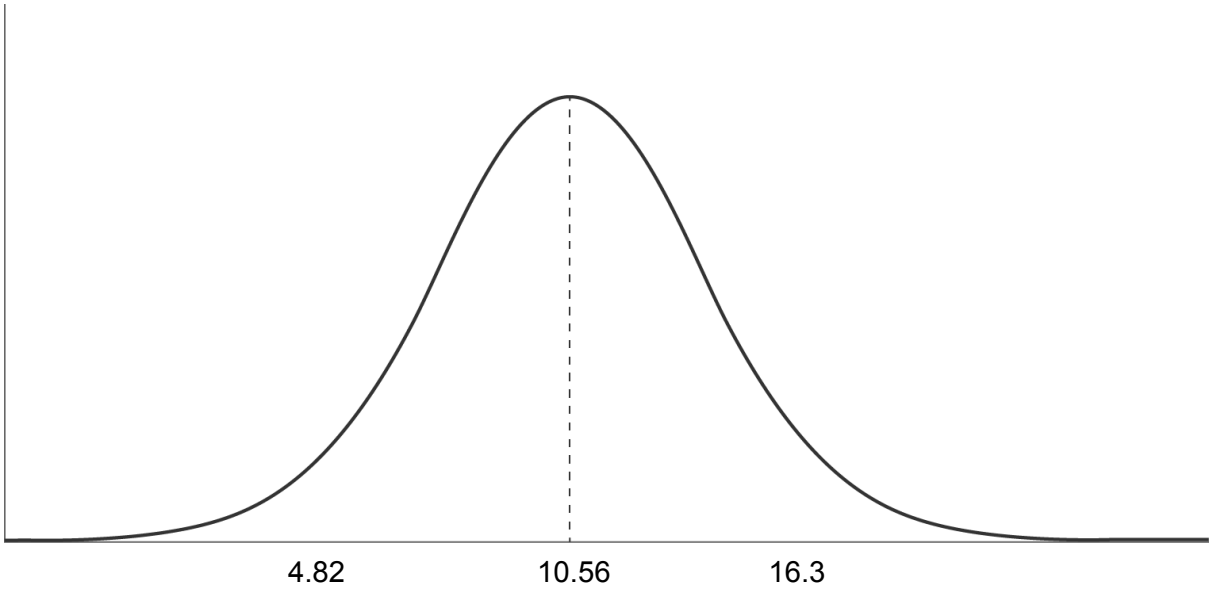
If we use n in the denominator to calculate sample variance then we are underestimating the true population variance

The inferences which will be drawn from the sample variance should be approximately relevant to the population. Thus, research has proved that, by taking n-1 it will give the nearest inferences.

Lets take an example

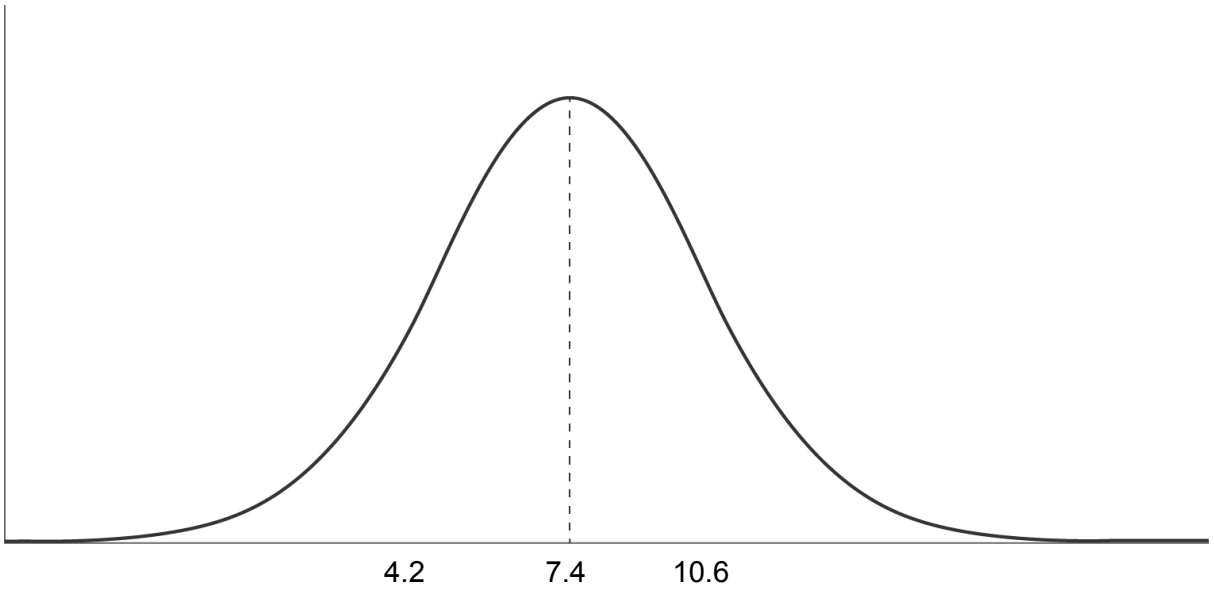
If x is a population dataset
x = {2,5,6,8,10,11,15,18,20}

Population mean = 10.56
Population Standard Deviation = 5.74
Population Variance = 32.91



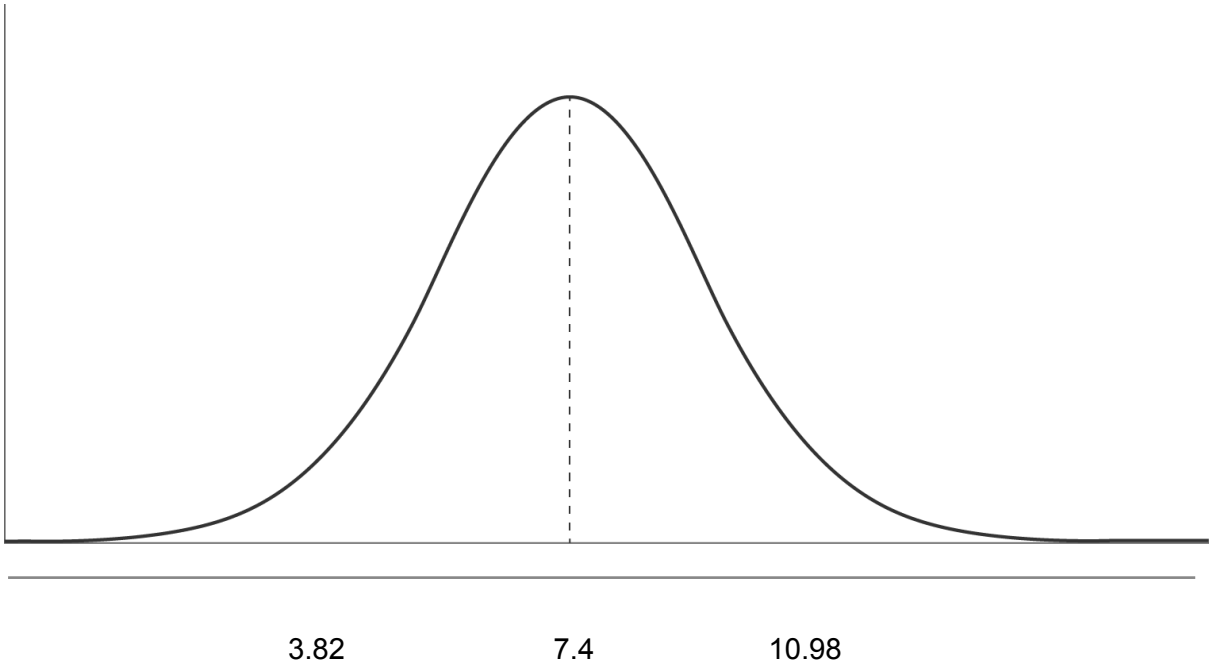
If we take n in the denominator of sample dataset, let's see the result
If y is the sample dataset
Y = {2,6,8,10,11}

Sample mean = 7.4
Sample Standard Deviation = 3.2
Sample Variance = 10.24



Lets take n-1

Sample mean = 7.4
Sample Standard Deviation = 3.58
Sample Variance = 12.8



Thus, by the above example its proved that when we take n-1 we get closer inferences to the population dataset which helps to analyse the data more effectively. In other words, the reason we use n-1 rather than n is so that the sample variance will be what is called an unbiased estimator of the population variance