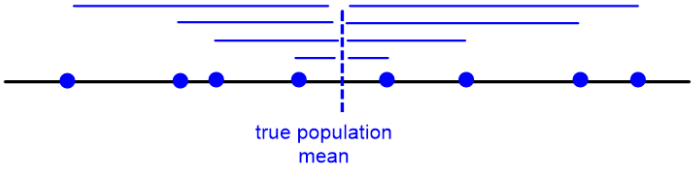
AP Statistics – Why n-1 instead of n in standard deviation calculation?

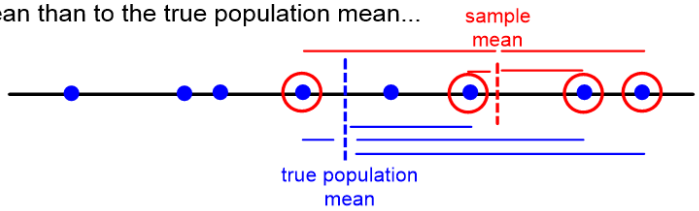
Why is standard deviation from a sample calculated using n-1 instead of n?

Full explanation requires a formal mathematical proof, but we can consider a few things that provide some intuitive insight.

If we use n, the true variance of the population will always be underestimated when calculated from a sample:



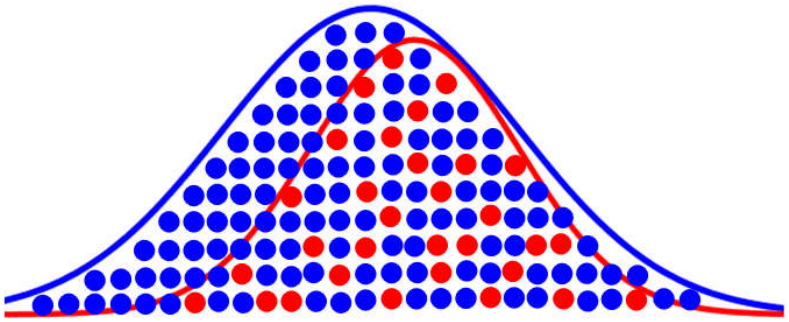
The data in any sample of the population will always be closer to the sample mean than to the true population mean...



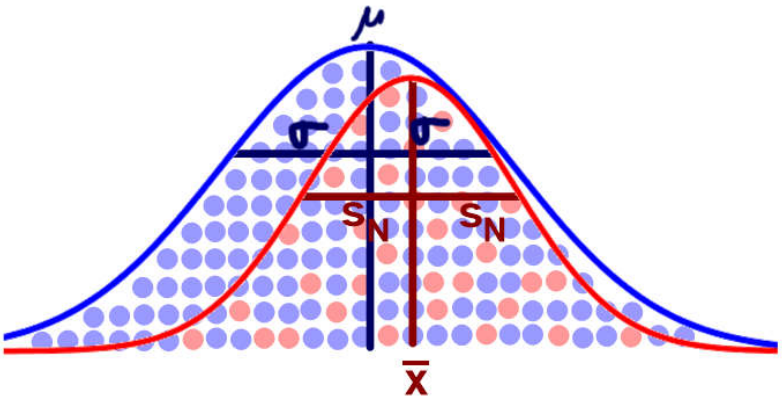
...this is called a **biased estimate** and means the calculated variance (and standard deviation) will always underestimate the true population value.

This is a little easier to see if we assume Normal populations (although it is true for all populations):

If we select a sample (red dots) from a population (blue dots), that sample may have a mean that is slightly different than the true population mean (exaggerated here):



Then the mean and standard deviation of the sample would be different from that of the population, and the standard deviation of the sample would always be smaller than that of the population:



In general, Greek alphabet symbols are used to denote ***parameters*** from populations (or models of populations), and English alphabet symbols are used to denote ***statistics*** computed from samples.

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

σ is the true population standard deviation of a population of N elements.

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

s_N is the uncorrected sample standard deviation for a sample of n out of N.

While s_N is the correct standard deviation of the sample, we know that it will always underestimate the standard deviation of the population. We use the symbol s to denote the corrected sample deviation (which is increased to correct this underestimation):

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

Why do we divide by n-1 instead of n in the formula? It can be shown in formal proof that this exactly corrects for the underestimation, but we can get an intuitive sense of why by looking at the formula itself:

In the formula for standard deviation, the value of the mean is used in calculating the distance for each data value to be squared:

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

This mean was computed from the data sample, so it represents the equivalent of one extra data point of information. Another way to think of it is if we knew all the data values in a sample except one, but we knew the mean, we could reconstruct the final data value:

10 12 13 14 15 22 32 44 45 47 47 48 ( )

If we know $\bar{x}$ = 30.6923 then the final data value must be 50

So using the mean in the calculation removes one ***'degree of freedom'*** from the data list. It is as if we really have n-1 instead of n data values which can vary.

If we divide by n in the standard deviation formula, the value will be too high. Dividing by n-1 instead raises the estimate.

In fact, if we use a sample of the population, but divide by n-1, the result can be shown to be exactly equal to the standard deviation for the entire population.

The factor $\frac{n}{n-1}$ is called **Bessel's correction**.

If you want to research the formal proof showing why this exactly corrects for the underestimation, Wikipedia's article on 'Bessel's correction' provides 3 different proofs. (Proof #3 is the most intuitive and easiest to follow)