Standard Deviation

Consider a population of size N and a sample (subset) of the population of size $n < N$. We draw the values for a feature of the entire population (or we sample) for an attribute, say age.

Now, the standard deviation is as follows for the **population**. Given $a_i, i \in [1, N]$ is the set of N ages, first find the average $\bar{a}_i$

$$\bar{a}_i = \sum_{i=1}^{N} a_i / N \tag{1}$$

Then *stdev* is

$$stdev = \sqrt{\frac{\Sigma_{i=1}^{N} (a_i - \bar{a}_i)^2}{N}} \tag{2}$$

Now, if I **sample n** examples from N, the *stdev* is calculated as:

$$\bar{a}_i = \sum_{i=1}^{n} a_i / n \tag{3}$$

Then *stdev* is

$$stdev = \sqrt{\frac{\Sigma_{i=1}^{n} (a_i - \bar{a}_i)^2}{n-1}} \tag{4}$$

What is different?

What is the effect of dividing by $n - 1$? Why might this be reasonable?

The effect is you get a higher standard deviation. It is reasonable because you only have a sample or subset of the data which may not be enough to get a tight estimate of the true standard deviation.