Why Divide Sample Variance by n-1

It is well known that the formula to calculate the variance of a sample dataset is the sum of squares divided by the sample size less one. I remember being asked by a friend: "Why minus one?" To my great shame as a statistics major, I could give neither a rigorous nor intuitive answer. This essay serves to regain some lost pride and is aimed at those with a basic understanding of descriptive statistics.

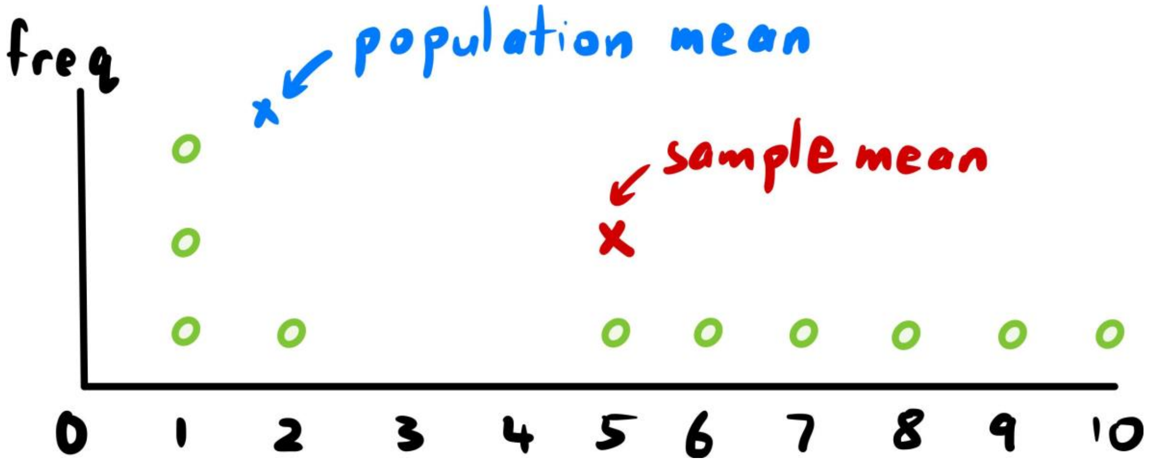
We shall take an empirical approach, which is a fancy way of saying we work with observations rather than pure theory. Suppose we have a random generator of numbers from 1 to 10. We use it to generate 10 numbers and obtain the set (7 8 1 5 6 9 10 1 2 1).

Our goal is to describe this dataset with a single number. Let us take the average to obtain the sample mean = 5. **The sample mean is called a statistic and is a measure of central tendency.** In simple terms, it tells us where the center of the data is. Median and Mode are other statistics that measure central tendency.

Now suppose we want to measure dispersion, or how spread out the data is from the center. The sum of squared deviations from the sample mean is a good indicator of this and illustrated below.

$$\text{Sum of Squares} = (7-5)^2 + (8-5)^2 + \ldots + (1-5)^2 = 112$$

The next logical step would be to average it out by dividing by 10. However, think about this for a moment. We would be introducing bias to our statistic!



If our data was represented on a number line, the sample mean would be in the center. Hence the sum of squares with respect to the sample mean would be the smallest possible sum, and this is possible to prove theoretically! If the true mean of our number generator is not 5, our sum of squares would be larger. **We will be underestimating the variance if we divide the sum of squares by 10.** We need to divide by a lesser value to compensate, and that value turns out to be one less the sample size. In our example, we need to divide by 9.

Sadly, I am not going to delve into why we divide by n-1 instead of say n-2. Good resources such as Khan Academy explain it well from a theoretical point of view. An explanation based on degrees of freedom which I have found lacking in literature is in this video. Kudos for making it this far, and I hope you now have a rough idea on how to answer the title of this essay.