Why We Divide by n-1: An Intuitive and Mathematical Guide to Bessel's Correction

Saurabh

September 7, 2025

Abstract

In statistics, we often encounter two formulas for variance: one dividing by 'N' and another by 'n-1'. This document explains why the 'n-1' is used for sample variance. We will build an intuitive understanding of "degrees of freedom" and then walk through the mathematical proof that shows how dividing by 'n-1' gives us a better, unbiased estimate of the true population variance.

1 The Two Faces of Variance

Variance measures the spread or dispersion of data points around their mean. The confusion often starts because there are two distinct formulas, depending on whether you have data for the entire **population** or just a **sample**.

1.1 Population Variance (σ^2)

If you have data for every single member of a group (the entire population), you use the population mean (μ) and divide by the total number of members, N.

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

This formula gives you the exact variance of the population because you have all the information.

1.2 Sample Variance (s^2)

More often, we can't measure the whole population. Instead, we take a smaller sample and use it to *estimate* the population variance. For this, we use the sample mean ($\bar{x}$) and divide by the sample size minus one, $n - 1$.

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

The central question is: **Why subtract one from n?**

2 The Intuitive Reason: Bias and the Sample Mean

When we calculate the sample variance, we don't know the true population mean (μ). We have to use the sample mean ($\bar{x}$) as a stand-in. Herein lies the problem.

The sample mean, $\bar{x}$, is calculated *from the sample itself*. By its very definition, it is the point that minimizes the sum of squared differences for that specific sample. This means the data points in our sample will, on average, be closer to their own mean ($\bar{x}$) than they would be to the true population mean (μ).

Think about it: The sample mean is tailor-made for the sample, while the population mean is not.

As a result, the sum of squared deviations from the sample mean, $\sum(x_i - \bar{x})^2$, will almost always be smaller than it would be if we could use the true population mean. If we were to just divide by n, our estimate of the variance would be systematically too small. It would be a **biased estimator**. To fix this underestimation, we need to make our denominator slightly smaller. Dividing by $n-1$ instead of n inflates our estimate just enough to correct for this bias. This correction is known as **Bessel's Correction**.

2.1 A Concrete Example

Let's imagine a tiny population of three numbers: $\{1, 3, 5\}$.

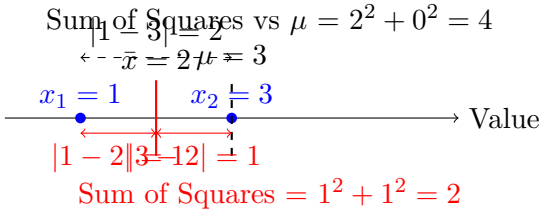
- The true population mean is $\mu = (1+3+5)/3 = 3$.
- The true population variance is $\sigma^2 = \frac{(1-3)^2+(3-3)^2+(5-3)^2}{3} = \frac{4+0+4}{3} \approx 2.67$.

Now, let's take a sample of size $n = 2$. Suppose we draw the sample $\{1, 3\}$.

- The sample mean is $\bar{x} = (1+3)/2 = 2$.
- Let's calculate the sum of squared deviations from our *sample mean*: $\sum(x_i - \bar{x})^2 = (1-2)^2 + (3-2)^2 = 1 + 1 = \mathbf{2}$.
- Now let's see what the sum of squares would have been if we'd used the true *population mean*: $\sum(x_i - \mu)^2 = (1-3)^2 + (3-3)^2 = 4 + 0 = \mathbf{4}$.

Example

Notice that the sum of squares calculated with the sample mean (2) is smaller than the sum of squares calculated with the population mean (4). This happens every time! Using $\bar{x}$ gives us an underestimate of the spread. Dividing by $n-1 = 1$ instead of $n = 2$ corrects our variance estimate for this sample from $2/2 = 1$ to $2/1 = 2$, which is much closer to the true value of 2.67.



3 Degrees of Freedom: The "Free" Variables

The concept of 'n-1' is formally explained by "degrees of freedom".

Definition: Degrees of freedom represent the number of values in a final calculation that are free to vary.

Let's use a simple analogy. Imagine you have three numbers (x_1, x_2, x_3) and you are told their mean is 10.

$$\frac{x_1 + x_2 + x_3}{3} = 10 \implies x_1 + x_2 + x_3 = 30$$

How many of these numbers can you pick freely?

- You can pick x_1 to be anything. Let's say $x_1 = 5$.
- You can pick x_2 to be anything. Let's say $x_2 = 15$.
- Now, can you pick x_3? No. The value of x_3 is now constrained by the mean. It *must* be $30 - 5 - 15 = 10$.

Once the mean was fixed, one value lost its freedom. We started with 3 numbers, but only 2 were free to vary. We had $n - 1 = 3 - 1 = 2$ degrees of freedom.

Connection to Variance: When we calculate the sample variance, we use the sum of deviations from the sample mean: $(x_1 - \bar{x}), (x_2 - \bar{x}), \ldots, (x_n - \bar{x})$. A fundamental property of the mean is that the sum of these deviations is always zero:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Because of this constraint, if you know the first $n - 1$ deviations, the last one is not free. It is fixed. Therefore, when calculating the sample variance, we only have $n - 1$ independent pieces of information about the spread of the data. This is why we average the sum of squares over the degrees of freedom, $n - 1$.

4 The Mathematical Proof

Here we will prove that the expected value of the sample variance formula with $n - 1$ is indeed the true population variance σ^2. In other words, $E[s^2] = \sigma^2$.

Let's start with the sum of squared deviations from the sample mean, $\sum(x_i - \bar{x})^2$. We can cleverly rewrite the term inside the summation:

$$(x_i - \mu) = (x_i - \bar{x}) + (\bar{x} - \mu)$$

Now, let's square both sides and sum over all i:

$$\sum(x_i - \mu)^2 = \sum\left[(x_i - \bar{x}) + (\bar{x} - \mu)\right]^2$$

Expanding the square:

$$\sum(x_i - \mu)^2 = \sum\left[(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - \mu) + (\bar{x} - \mu)^2\right]$$

Distribute the summation:

$$\sum(x_i - \mu)^2 = \sum(x_i - \bar{x})^2 + \sum 2(x_i - \bar{x})(\bar{x} - \mu) + \sum(\bar{x} - \mu)^2$$

The middle term can be rewritten as $2(\bar{x} - \mu)\sum(x_i - \bar{x})$. Since we know $\sum(x_i - \bar{x}) = 0$, the entire middle term becomes zero.

$$\sum(x_i - \mu)^2 = \sum(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

Let's rearrange to isolate the term we are interested in:

$$\sum(x_i - \bar{x})^2 = \sum(x_i - \mu)^2 - n(\bar{x} - \mu)^2$$

Now, we take the expected value ($E[\cdot]$) of the entire equation.

$$E\left[\sum(x_i - \bar{x})^2\right] = E\left[\sum(x_i - \mu)^2\right] - E\left[n(\bar{x} - \mu)^2\right]$$

By definition, $E[(x_i - \mu)^2] = \sigma^2$ and $E[(\bar{x} - \mu)^2]$ is the variance of the sample mean, which is $\frac{\sigma^2}{n}$.

$$E\left[\sum(x_i - \bar{x})^2\right] = \sum E[(x_i - \mu)^2] - nE[(\bar{x} - \mu)^2]$$

$$E\left[\sum(x_i - \bar{x})^2\right] = n\sigma^2 - n\left(\frac{\sigma^2}{n}\right) = n\sigma^2 - \sigma^2$$

The key result of the derivation is:

$$\mathbf{E\left[\sum(x_i - \bar{x})^2\right] = (n-1)\sigma^2}$$

This says that the expected value of the sum of squares is actually $(n-1)$ times the true population variance.

Therefore, to get an unbiased estimator for σ^2, we must divide the sum of squares by $(n-1)$:

$$E[s^2] = E\left[\frac{\sum(x_i - \bar{x})^2}{n-1}\right] = \frac{1}{n-1}E\left[\sum(x_i - \bar{x})^2\right] = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$$

This proves that $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$ is an unbiased estimator of σ^2.

5 Conclusion

We divide the sum of squared differences by $n - 1$ when calculating sample variance for a profound reason. The sample mean, being derived from the sample itself, makes the data appear less spread out than it actually is relative to the true population mean. This introduces a systematic underestimation. The ‘n-1‘, representing the degrees of freedom, is the precise mathematical correction needed to counteract this bias, giving us the most accurate possible estimate of the population variance from our sample.