

Sample Standard Deviation Explained

Gábor Szijártó

Statistics is a must have for any ML/DL engineer and data scientist!

In this article I will cover a simple yet interesting equation: **Sample Standard Deviation**

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

This is the equation for the **unbiased sample variance**.

Why is it called 'unbiased' and why do we have $n - 1$ term in the denominator?

These are the questions that will be answered.

Bessel's Correction

A commonly used intuitive explanation is the [Bessel's correction](#).

“While there are n independent observations in the sample, there are only $n - 1$ independent residuals, as they sum to 0”

It states that we need to use $n-1$ as there are only $n-1$ independent variables, because the expected value is calculated based on the samples. In case we have $n-1$ degree of freedom, then it makes sense to divide by $n - 1$ instead of n .

Easy to grab explanation, but I feel it kind of confusing and misses the most important point, the real reason behind the need for correction!

Why do I think it can be confusing?

Let's calculate the standard deviation of a uniform discrete random variable like 6 sided fair dice.

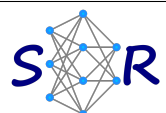
$$\sigma^2 = \mathbb{V}[X] = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} \quad (2)$$

The expected value is not independent from the possible values as it can't be by its definition, yet we are dividing by N instead of $N - 1$.

$$\mu = \mathbb{E}[X] = \sum_{i=1}^6 \frac{1}{6} x_i = 3.5$$

$$\mathbb{V}[X] = \frac{\sum_{i=1}^6 (x_i - \mu)^2}{6} \approx 2.9166$$

It can be confusing.



Understanding the Need for Correction

Understanding the essence behind the correction will immediately reveal why we need to divide by n in some applications and by $n - 1$ in others.

The major difference between sampling from a population and knowing all elements of it or the distribution itself, is that in the latter case we know the exact value of μ , while in case of sampling we have only an estimate of it.

To be able to calculate the variance we must know the real μ , not just an estimation. The error of estimated expected value induces a bias into the variance calculations!

μ = real expected value of the population

$\bar{X} = \mathbb{E}[X]$ = calculated expected value value based on sample

$$\mathbb{V}[X] = \mathbb{E}[(X - \mu)^2] \neq \mathbb{E}[(X - \bar{X})^2]$$

We divide $n - 1$ instead of n to get an unbiased estimation for the population's variance! Can we prove it? Yes, only a small idea needs to be applied.



$$X_i - \mu = (\bar{X} - \mu) + (X_i - \bar{X}) \quad (3)$$

Lets substitute equation (3) into (2).

$$\begin{aligned} \sigma^2 = \mathbb{V}[X] &= \mathbb{E}[(X - \mu)^2] = \mathbb{E}\left[\left((\bar{X} - \mu) + (X - \bar{X})\right)^2\right] \\ &= \mathbb{E}[(\bar{X} - \mu)^2] + 2\mathbb{E}[(\bar{X} - \mu)(X - \bar{X})] + \mathbb{E}[(X - \bar{X})^2] \\ &= \mathbb{V}[\bar{X}] + 2(\bar{X} - \mu)\mathbb{E}[X - \bar{X}] + \mathbb{E}[(X - \bar{X})^2] \end{aligned}$$

Evaluating the terms separately:

$$\mathbb{V}[\bar{X}] = \mathbb{V}\left[\frac{\sum X}{n}\right] = \frac{1}{n^2} \mathbb{V}[\sum X] = \frac{1}{n^2} \sum \mathbb{V}[X] = \frac{\mathbb{V}[X]}{n} = \frac{\sigma^2}{n} \quad (4)$$

The assumption that variables are **independent** was used here!

This is true only in the case we are sampling with replacement, so the same element is allowed to be sampled multiple times!

$$\mathbb{E}[X - \bar{X}] = \mathbb{E}[X] - \bar{X} = 0 \quad (5)$$

This makes the whole middle term zero!

$$\mathbb{E}[(X - \bar{X})^2] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sigma_s^2 \quad (6)$$

By the definition of variance.

In the calculations above basic properties of the expected value and variance were used.



Substitute back the results of (4, 5, 6)!

$$\sigma^2 = \frac{\sigma^2}{n} + 0 + \sigma_s^2$$

$$\sigma^2 = \frac{n}{n-1} \sigma_s^2$$

$\frac{n}{n-1}$ is the correction term applied to get an unbiased estimator for population variance.

This results in the well known equation:

$$\sigma^2 = \frac{n}{n-1} \sigma_s^2 = \frac{n}{n-1} \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^2}{\mathbf{n} - \mathbf{1}} \quad (7)$$

Since $\frac{n}{n-1}$ is greater than 1, we can conclude that σ_s **underestimates** the true sample variance. As the number of samples increases, σ_s will converge to σ .