WHY DOES THE STANDARD DEVIATION FORMULA HAVE "$n-1$" INSTEAD OF "n"?

THEORETICAL APPROACH

First let's prove this with theory. We will need a few concepts and formulas from higher level statistics courses. The expected value, which is basically a more general version of a mean, is a linear operator: $E[aY+b] = aE[Y]+b$, where Y is a random variable and a and b are constants. The variance is a bit different: $Var[aY+b] = a^2 Var[Y]$. The expected value of a random variable squared, AKA its second moment, is $E[Y^2] = Var[Y] + (E[Y])^2$. Finally, we will make use of the fact that for any distribution of Xs, they all have the same properties: for example, $E[X_i] = E[X_j] = \mu$ and $Var[X_i] = Var[X_j] = \sigma^2$ for any i and j.

Let's use S_n^2 to represent the variance using the more intuitive choice of n rather than $n-1$ in the denominator. We want to find its expected value, and hope it turns out to be σ^2, because that's what we're trying to estimate.

$$
\begin{aligned}
E[S_n^2] &= E\left[\tfrac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right] \\
&= \tfrac{1}{n}E\left[\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right], && \text{applying properties of } E[\cdot] \\
nE[S_n^2] &= E\left[\sum_{i=1}^{n}\left(X_i^2-2X_i\bar{X}+\bar{X}^2\right)\right], && \text{mulitplying both sides by } n\text{; expanding the square} \\
&= E\left[\sum_{i=1}^{n}X_i^2-\sum_{i=1}^{n}2X_i\bar{X}+\sum_{i=1}^{n}\bar{X}^2\right], && \text{distributing the summation over the parentheses} \\
&= E\left[\sum_{i=1}^{n}X_i^2-2\bar{X}\sum_{i=1}^{n}X_i+n\bar{X}^2\right], && \text{applying summation properties (note expressions without "}i\text{")} \\
&= E\left[\sum_{i=1}^{n}X_i^2-2\bar{X}\left(n\bar{X}\right)+n\bar{X}^2\right], && \text{applying the formula for } \bar{X} \\
&= E\left[\sum_{i=1}^{n}X_i^2-2n\bar{X}^2+n\bar{X}^2\right], && \text{simplifying} \\
&= E\left[\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right], && \text{simplifying} \\
&= E\left[\sum_{i=1}^{n}X_i^2\right]-E\left[n\bar{X}^2\right], && \text{applying properties of } E[\cdot] \\
&= \sum_{i=1}^{n}E[X_i^2]-nE\left[\bar{X}^2\right], && \text{applying properties of } E[\cdot] \\
&= nE[X^2]-nE\left[\bar{X}^2\right], && \text{using the fact that } X_i \text{ has the same properties for all } i \\
E[S_n^2] &= E[X^2]-E\left[\bar{X}^2\right], && \text{dividing both sides by } n
\end{aligned}
$$

Now we have to deal each of these squared terms on the right-hand side of the equation. We will use the "second moment" formula given previously.

$$
\begin{aligned}
\text{First, } E[X^2] &= Var[X]+(E[X])^2 \\
&= \sigma^2+\mu^2
\end{aligned}
$$

$$
\begin{aligned}
\text{Then, } E\left[\bar{X}^2\right] &= Var[\bar{X}] + (E[\bar{X}])^2 \\
&= Var\left[\tfrac{1}{n}\sum_{i=1}^{n} X_i\right] + \mu^2, \quad \text{substituting the formula for } \bar{X} \\
&= \tfrac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] + \mu^2, \quad \text{applying properties of } Var[\cdot] \\
&= \tfrac{1}{n^2}\sum_{i=1}^{n} Var\left[X_i\right] + \mu^2, \quad \text{applying properties of } Var[\cdot] \\
&= \tfrac{1}{n^2}\sum_{i=1}^{n} \sigma^2 + \mu^2 \\
&= \tfrac{1}{n^2}\left(n\sigma^2\right) + \mu^2 \\
&= \tfrac{1}{n}\sigma^2 + \mu^2, \quad \text{simplifying}
\end{aligned}
$$

We can now plug these back into the equation we were originally working on.

$$
\begin{aligned}
E\left[S_n^2\right] &= E\left[X^2\right] - E\left[\bar{X}^2\right] \\
&= \sigma^2 + \mu^2 - \left(\tfrac{1}{n}\sigma^2 + \mu^2\right) \\
&= \left(1 - \tfrac{1}{n}\right)\sigma^2 + \mu^2 - \mu^2, \quad \text{regrouping} \\
&= \tfrac{n-1}{n}\sigma^2, \quad \text{simplifying}
\end{aligned}
$$

Uh-oh! It looks like S_n^2 doesn't actually estimate σ^2, it estimates $\frac{n-1}{n}\sigma^2$. Let's rearrange to isolate just σ^2, since that's what we're actually interested in.

$$
\begin{aligned}
E\left[S_n^2\right] &= \tfrac{n-1}{n}\sigma^2 \\
\tfrac{n}{n-1}E\left[S_n^2\right] &= \sigma^2 \\
E\left[\tfrac{n}{n-1}S_n^2\right] &= \sigma^2 \\
E\left[\tfrac{n}{n-1}\tfrac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] &= \sigma^2 \\
E\left[\tfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] &= \sigma^2
\end{aligned}
$$

And finally we have the familiar formula for sample variance, with "$n-1$", giving an unbiased estimate of the population variance.

PRACTICAL APPROACH

Next, let's look at a specific example. Although the theoretical proof above is all we need, it is not very satisfying. By using an example, it helps convince ourselves psychologically. But keep in mind that one example (even many examples!) is always weaker than a mathematical proof, since there could be some contradictory example we didn't think of.

0	2	4

$\mu = 2$ $\sigma^2 = \frac{8}{3}$

There are 9 possible samples of 2 cards

List of all possible samples of size $n=2$	sample average $\bar{x} = \frac{\sum x}{n}$	sample variance $s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$
(0,0)	$\frac{0+0}{2} = 0$	$\frac{(0-0)^2+(0-0)^2}{1} = 0$
(0,2)	$\frac{0+2}{2} = 1$	$\frac{(0-1)^2+(2-1)^2}{1} = 2$
(0,4)	$\frac{0+4}{2} = 2$	$\frac{(0-2)^2+(4-2)^2}{1} = 8$
(2,0)	$\frac{2+0}{2} = 1$	$\frac{(2-1)^2+(0-1)^2}{1} = 2$
(2,2)	$\frac{2+2}{2} = 2$	$\frac{(2-2)^2+(2-2)^2}{1} = 0$
(2,4)	$\frac{2+4}{2} = 3$	$\frac{(2-3)^2+(4-3)^2}{1} = 2$
(4,0)	$\frac{4+0}{2} = 2$	$\frac{(4-2)^2+(0-2)^2}{1} = 8$
(4,2)	$\frac{4+2}{2} = 3$	$\frac{(4-3)^2+(2-3)^2}{1} = 2$
(4,4)	$\frac{4+4}{2} = 4$	$\frac{(4-4)^2+(4-4)^2}{1} = 0$

Average of all $\bar{x}$ sample averages $\frac{0+1+2+1+2+3+2+3+4}{9 \text{ samples}} = \frac{18}{9} = 2$

$\Rightarrow$ (average of all $\bar{x}$) $= \mu$

Average of all s^2 sample variances $\frac{0+2+8+2+0+2+8+2+0}{9 \text{ samples}} = \frac{24}{9} = \frac{8}{3}$

$\Rightarrow$ (average of all s^2) $= \sigma^2$

What if we used $\frac{\sum (x-\bar{x})^2}{n}$ instead?

List of all possible samples of size $n=2$	sample average $\bar{x}=\frac{\sum x}{n}$	$\frac{\sum (x-\bar{x})^2}{n}$
$(0,0)$	$\frac{0+0}{2}=0$	$\frac{(0-0)^2+(0-0)^2}{2}=0$
$(0,2)$	$\frac{0+2}{2}=1$	$\frac{(0-1)^2+(2-1)^2}{2}=1$
$(0,4)$	$\frac{0+4}{2}=2$	$\frac{(0-2)^2+(4-2)^2}{2}=4$
$(2,0)$	$\frac{2+0}{2}=1$	$\frac{(2-1)^2+(0-1)^2}{2}=1$
$(2,2)$	$\frac{2+2}{2}=2$	$\frac{(2-2)^2+(2-2)^2}{2}=0$
$(2,4)$	$\frac{2+4}{2}=3$	$\frac{(2-3)^2+(4-3)^2}{2}=1$
$(4,0)$	$\frac{4+0}{2}=2$	$\frac{(4-2)^2+(0-2)^2}{2}=4$
$(4,2)$	$\frac{4+2}{2}=3$	$\frac{(4-3)^2+(2-3)^2}{2}=1$
$(4,4)$	$\frac{4+4}{2}=4$	$\frac{(4-4)^2+(4-4)^2}{2}=0$

Average of all $\left(\frac{\sum (x-\bar{x})^2}{n}\right)$ for all samples:

$$\frac{0+1+4+1+0+1+4+1+0}{9 \text{ samples}} = \frac{12}{9} = \frac{4}{3}$$

But this average $\frac{4}{3}$ is not equal to $\sigma^2 = \frac{8}{3}$