Why divide by (n – 1) instead of by n
when we are calculating the sample standard deviation?

To answer this question, we will talk about the **sample variance $\mathbf{s^2}$**
The sample variance $\mathbf{s^2}$ is the square of the sample standard deviation **s**.
It is the "sample standard deviation BEFORE taking the square root" in the final step of the calculation by hand.
The sample variance $\mathbf{s^2}$ is easier to work with in the examples on pages 3 and 4 because it does not have square roots.

The POPULATION VARIANCE σ^2 is a PARAMETER of the population.

The SAMPLE VARIANCE $\mathbf{s^2}$ is a STATISTIC of the sample.

We use the sample statistic to estimate the population parameter
The sample variance $\mathbf{s^2}$ is an estimate of the population variance σ^2

Suppose we have a population with **N** individuals or items.
Suppose that we want to take samples of size **n** individuals or items from that population

IF we could list all possible samples of **n** items that could be selected from the population of **N** items, then we could find the sample variance for each possible sample.

We would want the following to be true:

> **We would want the average of the sample variances for all possible samples to equal the population variance.**

It seems like a logical property and a reasonable thing to happen.
This is called "**unbiased**"

When we divide by (n −1) when calculating the sample variance, then it turns out that
the average of the sample variances for all possible samples is equal the population variance.
So the sample variance is what we call an unbiased estimate of the population variance.

If instead we were to divide by n (rather than n −1) when calculating the sample variance,
then the average for all possible samples would NOT equal the population variance.
Dividing by n does not give an "unbiased" estimate of the population standard deviation.

Dividing by n−1 satisfies this property of being "unbiased", but dividing by n does not.
Therefore we prefer to divide by n-1 when calculating the sample variance.

The examples on the next 3 pages help explain this:
Page 2 starts with a **population of N = 3** items, and also contains more explanation
Page 3 looks at **samples of size n = 2** selected from the population of N = 3 items,
and shows that **dividing by (n −1) gives an unbiased estimate of σ^2**
Page 4 looks at **samples of size n = 2** selected from the population of N = 3 items,
and shows that **dividing by n gives a biased estimate of σ^2**

The example is not a mathematical proof that this is always true. But it is always true.
If you want to see a rigorous mathematical proof, you can find it in books about mathematical statistics, (generally calculus based statistics books) which are beyond the scope of this course.

WHY DIVIDE BY (n-1) INSTEAD OF BY n?

Suppose you have a bag with 3 cards in it
The cards are numbered 0, 2 and 4

$$\underbrace{\boxed{0}\ \boxed{2}\ \boxed{4}}_{\text{Population of all } N=3 \text{ cards in bag}}$$

Population Mean: $\mu = \frac{0+2+4}{3} = \frac{6}{3} = 2$

Population Variance: $\sigma^2 = \frac{(0-2)^2+(2-2)^2+(4-2)^2}{3} = \frac{4+0+4}{3} = \frac{8}{3}$

An important property of a sample statistic that estimates a population parameter is that if you evaluate the sample statistic for every possible sample and average them all, the average of the sample statistics should equal the population parameter.

We want: average of $\left(\begin{array}{c}\text{all possible}\\ \text{sample}\\ \text{variances}\end{array}\right)$ = population variance

This is called unbiased

When we divide by (n-1) in the sample variance s^2, then s^2 is an unbiased estimate of the population variance σ^2 Average of $\left(\begin{array}{c}s^2 \text{ for}\\ \text{all possible}\\ \text{samples}\end{array}\right) = \sigma^2$

When we divide by n in the sample variance it is not an unbiased estimate of the population variance σ^2.

This is why using $\sum \frac{(x-\bar{x})^2}{n-1}$ is better than using $\sum \frac{(x-\bar{x})^2}{n}$

$\boxed{0}$ $\boxed{2}$ $\boxed{4}$ $\mu = 2$ $\sigma^2 = \frac{8}{3}$

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 $\bar{x}$) = μ

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) = σ^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 $\sigma^2 = \frac{8}{3}$