UNIT 4 · BLOCK 1

↔ Variability

How spread out is everyone?

Why this matters

Last unit you found the **center** of a pile of data. But the center only tells half the story. Two groups can have the **exact same average** and still feel completely different — one tight and consistent, the other all over the place. This unit is about measuring that *spread*. Like central tendency, the math is just **subtracting, squaring, adding, and dividing** — and once you see *why* each step happens, standard deviation stops being a mystery formula and becomes a sentence you can read out loud.

By the end you can…

- ✓ Explain why two groups with the **same mean** can still be very different.
- ✓ Define **range**, **variance**, and **standard deviation** in plain English — and say what each one captures.
- ✓ Walk through a **standard deviation by hand**: deviations → squares → sum of squares → variance → square root.
- ✓ Tell the **population** formula (σ, ÷ N) from the **sample** formula (s, ÷ n − 1) and know which to use.
- ✓

Compute all three in Google Sheets with one function each, and read SD as "the typical distance from the mean."

1. The big idea — center isn't enough

The **mean** tells you where the data piles up. **Variability** tells you how *tightly* it piles up. Picture two stats classes that both average 80 on a test. In Class A everyone scored between 78 and 82 — rock steady. In Class B half the class got 60 and half got 100 — same average, totally different vibe. The average alone can't tell them apart. Variability can. Whenever someone hands you an average, the smart follow-up question is always the same: *how spread out is it?*

There are three workhorse measures of spread, and they build on each other. Here's the whole unit in one table — what each captures, the units it comes out in, and how badly a single wild value (an **outlier**) can throw it off:

Measure	What it captures	Units	Sensitive to outliers?
Range	Distance from lowest to highest score (uses only 2 numbers)	Original units	**Very** — one extreme score sets it
Variance	Average squared distance from the mean (uses every score)	*Squared* units	Yes — squaring magnifies far-out scores
Standard deviation	The typical distance from the mean (the square root of variance)	Original units	Yes, but reads in plain units

The one-line rules

Range = max − min. **Variance** = square the deviations, then average them. **Standard deviation** = the square root of the variance. The rest of this unit is just unpacking *why* each step is there.

2. The range — quick but fragile

The simplest measure of spread. Subtract the smallest score from the largest: **Range = max − min.**

For our **stress** scores (column J), the highest is **38** and the lowest is **4**, so the range is **38 − 4 = 34**. Fast and easy — but notice what it ignores. It looks at exactly **two** numbers and throws away everyone in between. Change all 198 middle scores however you like and, as long as the top and bottom stay put, the range never budges. That's why one unusually high or low score — an **outlier** — can blow it wide open and make a tight group look wildly spread out. The range is a fine quick gut-check, but it's no foundation for serious work. For that, we want a measure that listens to *every* score.

CHECK **A class's quiz scores are 7, 8, 8, 9, and one student who scored 0. Compared with the standard deviation, the range here is…**

Heavily distorted by the 0. The range is just max − min = 9 − 0 = 9, so that single outlier *sets* the bottom and dominates the whole measure. The range only ever looks at two numbers, which is exactly why it's so fragile.

3. Deviations — how far is each score from the mean?

To use *all* the scores, start by asking how far each one sits from the mean. That distance — score minus mean, written **$X - \bar{x}$** — is a **deviation**. A positive deviation means the score is above average; a negative one means below; and a deviation of zero means the score lands exactly on the mean. Take five simple numbers: **2, 4, 6, 8, 10**. Their mean is **6**. Now subtract 6 from each:

Score (X)	Minus mean ($X - \bar{x}$)	Deviation
2	2 − 6	−4
4	4 − 6	−2
6	6 − 6	0
8	8 − 6	2
10	10 − 6	4

The snag — deviations always sum to zero

Add those deviations up — −4 + −2 + 0 + 2 + 4 — and you get **0**. This isn't a coincidence: the negatives and positives **always cancel out**, for *any* data set, because the mean is the exact balance point of the data. So the average raw deviation is useless as a measure of spread — it's zero every single time. We need a fix that stops the cancelling.

CHECK **A student computes the deviations from the mean for a data set and adds them all up. They get a number close to −18. The most likely explanation is…**

They made an arithmetic mistake. Raw deviations from the mean *always* sum to exactly zero, no matter how spread out the data are. A nonzero total means the mean was computed wrong or a subtraction slipped — it's a handy built-in error check.

4. Variance — square the deviations, then average

The fix is to **square** each deviation before adding. Squaring turns every number positive (a negative times a negative is positive), so nothing cancels — and as a bonus it gives extra weight to scores that sit far from the mean, which is exactly what we want a spread measure to notice:

Deviation $(X - \bar{x})$	Squared $(X - \bar{x})^2$
−4	16
−2	4
0	0
2	4
4	16

Add the squares: 16 + 4 + 0 + 4 + 16 = **40**. That total has a name you'll see everywhere from here on: the **sum of squares** (often written **SS**) — the sum of the squared deviations. It's the raw fuel for variance, standard deviation, and nearly every test in the back half of

this course. Now we average it. We have 5 numbers, but for a **sample** we divide by **n − 1 = 4** (the reason is the next section):

> **Variance**
>
> **Variance = SS ÷ (n − 1) = 40 ÷ 4 = 10.** It's the *average squared distance* from the mean. The logic is airtight — it uses every score and never collapses to zero — but there's a catch: the units are **squared**. If the scores were in points, the variance is in "points squared," which nobody can picture. That awkwardness is exactly why we take one more step.

5. Standard deviation — back into real units

Variance lives in squared units, so we **undo the squaring**: take the **square root** of the variance and you land back in the data's original units. **Standard deviation = √10 ≈ 3.16.**

> **What SD actually means**
>
> Now it says something plain: **on average, the scores sit about 3.16 away from the mean.** That's the whole point of standard deviation — it's **the typical distance from the mean**, in the data's own units. A *small* SD means everyone clusters tightly near the average; a *large* SD means the scores are scattered far and wide. Variance and SD always rank groups the same way (bigger variance, bigger SD) — but SD is the one you *interpret and report*, because it speaks in the units you actually care about.

So the assembly line, start to finish, is: **deviations → square them → sum of squares (SS) → divide (variance) → square root (SD)**. Five steps, and you've turned a column of numbers into a single, readable measure of spread. Let's run that line on a fresh set of numbers, one step at a time:

Worked example — standard deviation of {2, 4, 4, 6, 9} by hand

Treat these five scores as a **sample**. Predict each step before revealing it.

STEP 1 — FIND THE MEAN

Add them and divide by n = 5: (2 + 4 + 4 + 6 + 9) ÷ 5 = 25 ÷ 5 = `5`. Every deviation is measured from this.

STEP 2 — GET EACH DEVIATION (X − X̄)

2 − 5 = −3, 4 − 5 = −1, 4 − 5 = −1, 6 − 5 = 1, 9 − 5 = 4. Quick check: they sum to `0` ✔ (−3 − 1 − 1 + 1 + 4).

STEP 3 — SQUARE EACH DEVIATION

$(-3)^2 = 9$, $(-1)^2 = 1$, $(-1)^2 = 1$, $1^2 = 1$, 4^2 = `16`. Now nothing is negative.

STEP 4 — SUM THE SQUARES (SS)

9 + 1 + 1 + 1 + 16 = `28`. That's the sum of squares.

STEP 5 — DIVIDE BY N − 1 FOR THE VARIANCE

n − 1 = 4, so variance = 28 ÷ 4 = `7`. (Average squared distance — still in squared units.)

STEP 6 — SQUARE-ROOT FOR THE STANDARD DEVIATION

SD = √7 ≈ `2.65`. Read it: a typical score sits about **2.65** away from the mean of 5.

CHECK

A set of reaction times has a variance of 144 milliseconds-squared. What is the standard deviation, and in what units?

12 milliseconds. SD is the square root of the variance (√144 = 12), and taking the root undoes the squaring — so SD lands back in the *original* units (milliseconds), not squared units. That's exactly why we report SD instead of variance.

6. Population vs. sample — why divide by *n − 1*?

You may have noticed we divided by **n − 1**, not n. Here's the reason, in plain terms. If you have the *entire* group you care about — the whole **population** — you divide the sum of squares by **N**, and the result is the population variance (σ^2, "sigma squared"); its square root is the population standard deviation (σ, "sigma"). But in real research we almost never have the whole population. We have a **sample** — a slice of a bigger group — and a sample's scores tend to cluster a little *closer* to their own mean than the full population's scores do to the true mean. Divide by n and you'd systematically *underestimate* the real spread. Dividing by **n − 1** instead nudges the answer up just enough to correct that bias. You don't need to memorize the proof — just the rule.

	Population	Sample
You have…	Every member of the group	A subset standing in for the group
SD symbol	σ (sigma)	s
Variance symbol	σ^2	s^2
Divide SS by…	N	**n − 1**
Sheets function	`=STDEV.P` / `=VAR.P`	`=STDEV.S` / `=VAR.S`

The rule to keep

Sample data → use n − 1 (the `.S` functions). That's almost always your situation in this course, because your data are a sample, not the whole world. The bias correction only matters much for small samples; with hundreds of scores, n and n − 1 give nearly the same answer — but using the right one is still the correct habit.

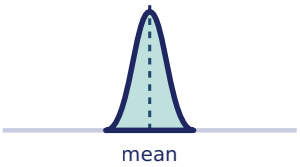
CHECK

You survey 50 students from your university to estimate the spread of stress in *the whole student body*. When you compute the standard deviation, you should divide the sum of squares by…

n − 1 (= 49). Those 50 students are a *sample* standing in for the larger population, so you use the sample formula with n − 1 to correct the slight underestimate of spread. You'd divide by N only if those 50 students *were* the entire group you cared about.

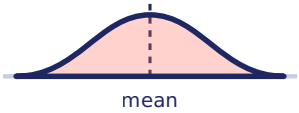
7. Picture it — same center, different spread

This is the whole unit in one image. Both curves below are centered on the **same mean** — but one is tall and narrow and the other is short and wide. The mean can't tell them apart; the standard deviation is the *only* thing that does.



Small SD

Tall and narrow — scores cluster tightly near the mean. Consistent, predictable.



Large SD

Short and wide — same center, but scores scatter far and wide. Inconsistent, surprising.

Both piles balance on the same point, so they share a mean. But the coral one spreads its scores across a much wider stretch of the number line — a larger typical distance from the center, and therefore a larger standard deviation. Tall-and-narrow is a small SD; short-and-wide is a large SD. Hold that picture; it's the intuition behind every spread number you'll ever read.

Worked example — on OUR data

Open the **SJNY Student Life** dataset. We want the spread of **stress** scores (column J). In any empty cell:

```
=MAX(J2:J201)-MIN(J2:J201)   → the range (38 − 4 = 34)
=VAR.S(J2:J201)              → the sample variance (≈ 34.86)
=STDEV.S(J2:J201)            → the sample standard deviation (≈ 5.9)
```

So stress has a mean of about **20.1** and a standard deviation of about **5.9** — meaning a typical student's stress sits roughly **6 points** away from the average, in either direction. Both `=VAR.S` and `=STDEV.S` use the **n − 1** (sample) version, which is what you want for our data.

Try it now — the Descriptive Statistics calculator

The same calculator from Unit 3 also reports **range, variance, and standard deviation**. Paste numbers or hit **Load class data**, and watch the spread measures appear right alongside the center. Try pasting a tight set (like 4, 5, 5, 6) and then a scattered one (like 0, 5, 5, 50) with a similar mean — and see the SD jump.

> **Stat in the Wild — same average, different life**
>
> San Diego and Phoenix can post nearly the **same average yearly temperature** — somewhere around a pleasant 70°F. But living there feels nothing alike. San Diego is mild almost every single day: small standard deviation. The Phoenix desert swings from brutal 110°F afternoons to chilly nights: **huge** standard deviation. Same mean, wildly different experience.
>
> **The lesson:** the average tells you the center, but the **standard deviation** tells you what you'd actually *feel* day to day. If a friend only told you the average temperature, you'd pack completely wrong for the desert. Whenever someone reports an average, ask the follow-up: *how spread out is it?*

In Google Sheets

You won't crunch the sum of squares by hand on exam day — Sheets does it in one cell each. Open the **SJNY Student Life** dataset and try these on the **stress** column:

```
=MAX(J2:J201)-MIN(J2:J201)   → the range (max − min)
=VAR.S(J2:J201)              → the sample variance (÷ n − 1)
=STDEV.S(J2:J201)            → the sample standard deviation (÷ n − 1)
```

The `.S` suffix means "sample" — these divide by **n − 1**, which is what you want for our data. (There are `.P` twins — `=VAR.P` and `=STDEV.P` — that divide by N; use those only when your numbers *are* the whole population.) Want to confirm the engine isn't hiding

anything? Build the SD by hand in columns: put each deviation in one column with `=J2-AVERAGE(J$2:J$201)`, square it in the next, sum that column for SS, divide by `COUNT()-1`, and take `=SQRT()`. You'll land on the same number `=STDEV.S` gives you — proof the shortcut is just the five-step recipe in disguise.

Watch out for

Range uses only two numbers — one outlier can make it lie; variance and SD use everyone. **Deviations always sum to zero**, which is why we square them first. **Variance is in squared units; SD is in normal units** — SD is the one you interpret and report. And **use the sample versions** (`=VAR.S`, `=STDEV.S`, dividing by n − 1) — that's right for sample data, which is nearly always what you have.

Key terms

Variability · **Range** · **Outlier** · **Deviation** · **Sum of squares (SS)** · **Variance** · **Standard deviation** · **Population (σ, ÷ N)** · **Sample (s, ÷ n − 1)**

PRACTICE — NEVER GRADED

Check yourself

Ten quick true/false items with instant explanations. (Different from your Canvas quiz — real practice.)

Sigma says: Read a standard deviation as a sentence — "a typical score sits about *this far* from the average." If you can fill in that blank in the data's own units, you understand SD better than most.