UNIT 3 · BLOCK 1

Central Tendency

One number that stands in for the whole group.

Why this matters

When someone says "the average student sleeps 7 hours," they just took 200 numbers and squeezed them into **one**. That single squeezed-down number is the most-used idea in all of statistics — it shows up in every headline, every grade report, every salary survey. This unit is about finding the number that best represents a whole pile of data, knowing **which** of three candidates to trust, and spotting when an "average" is quietly lying to you. The math is just **adding, dividing, and lining things up**.

By the end you can…

- ✓ Define **mean**, **median**, and **mode** in plain English and compute each by hand.
- ✓ Calculate all three in Google Sheets with one function each.
- ✓ Pick the **right** one for the shape of the data and its **scale of measurement**.
- ✓ Explain why a few extreme scores can make the "average" lie — and which measure resists them.

1. The big idea — one number for the whole pile

Central tendency is a single value that summarizes where a batch of scores *piles up* — the data's center of gravity. It answers the most natural question you can ask of any dataset: "So… what's typical?" There are three classic ways to answer, and the most important thing to know up front is that **they don't always agree**. When they disagree, the gap between them is itself information — it tells you the shape of your data.

Measure	Plain English	One-line rule	Symbol
Mean	The arithmetic average	Add them up, divide by how many	$\bar{x}$ / μ
Median	The middle score	Line them up, find the middle	Mdn
Mode	The most common score	The one that shows up most	Mo

We write the mean of a *sample* as $\bar{x}$ ("x-bar") and the mean of a whole *population* as μ ("mu"). Same recipe, different label — a distinction that matters more in later units. For now, treat them as the same idea: the balance point of your numbers.

2. The mean (the "average")

Add all the scores, then divide by how many there are. That's the whole thing. Five students sleep **5, 6, 7, 7, 10** hours:

Mean = (5 + 6 + 7 + 7 + 10) ÷ 5 = 35 ÷ 5 = **7 hours**.

In stats shorthand, the recipe is written $\bar{x} = \Sigma x \div n$, where the funny Σ ("sigma") just means "add up all the x's" and n is the count. So $\Sigma x = 35$, $n = 5$, and $\bar{x} = 7$. Don't let the symbols spook you — you've been doing this since grade school.

The balancing point

Picture the scores as equal weights sitting on a see-saw laid along a number line. The mean is the exact spot where the see-saw **balances**. This is why the mean uses *every* value — move any score and the balance point shifts. It's also the mean's one great weakness: **a single extreme score drags it**. The student who slept 10 pulled the average up; one who slept 24 would pull it up hard.

That "uses every value" property is exactly why the mean is the workhorse of statistics — almost every test you'll meet later (z-scores, t-tests, ANOVA) is built on it. But it also means the mean is only honest when no single score is wildly out of line with the rest.

CHECK **Four quiz scores are 8, 9, 9, and 10. What is the mean?**

9. Add them: 8 + 9 + 9 + 10 = 36. Then divide by how many, *n* = 4: 36 ÷ 4 = 9. (36 is the sum, Σx — a tempting wrong answer if you forget the divide step.)

3. The median (the "middle")

Put the scores **in order**, then find the one sitting dead center. In order, **5, 6, 7, 7, 10** → the middle (3rd of 5) is **7**. The median splits the data into two equal halves: half the scores fall at or below it, half at or above.

Two small wrinkles handle every case:

- **Odd count** — there's a single middle value; read it straight off. (5 scores → the 3rd.)

- **Even count** — there's no single middle, so **average the middle two**. For 4, 6, 8, 10 the middle two are 6 and 8, so the median is (6 + 8) ÷ 2 = **7**.

> **The median's superpower**
>
> The median only cares about **position**, not size. Change that 10 to 24 — or even 1,000 — and the middle value *doesn't budge*. Because it ignores how extreme the extremes are, the median is **resistant to outliers**. That's why it's the honest "typical" value whenever data is stretched out by a few unusual scores.

CHECK **Six commute times (sorted) are 8, 12, 15, 19, 24, 71 minutes. What is the median?**

17. With an even count (6), average the middle two — the 3rd and 4th values, 15 and 19: (15 + 19) ÷ 2 = 17. Notice the lone 71-minute commute didn't drag the median (it *would* drag the mean, which is 24.8 — option B).

4. The mode (the "most common")

The mode is simply the score that appears **most often**. In **5, 6, 7, 7, 10**, the mode is **7** — it shows up twice while everything else appears once. To find it, you just count repeats; no arithmetic at all.

The mode has one property the other two lack: it's the **only** measure that works for **categories**. You can't add up *Freshman* and *Senior* and divide by two, and you can't line up *Omnivore*, *Vegetarian*, and *Vegan* to find a "middle" — but you *can* ask which one appears most. In our class dataset the modal `class_year` is *Freshman* and the modal `diet` is *Omnivore*.

Mode quirks (all normal)

A dataset can have **two modes** (bimodal — two values tie for most common), **several** (multimodal), or **no useful mode at all** when every value appears exactly once. Spreadsheets disagree about ties: older `=MODE` reports just one. None of this is an error — it's the nature of "most common."

CHECK **You want a single "typical" value for the `diet` column (Omnivore, Vegetarian, Vegan). Which measure can you even use?**

Mode. Diet is a *nominal* category — there are no numbers to add (ruling out the mean) and no fixed order to find a middle of (ruling out the median). Only the mode — the most common label — makes sense for categories.

5. Outliers & skew — when the average lies

Here's the single most important lesson of the unit, and it's easiest to *see*. Start with a tidy little set, compute all three centers, then drop in one monster value and watch what happens. Predict each step before you reveal it.

Worked example — one outlier, three very different reactions

Our set is **4, 5, 5, 6, 10**. We'll find the mean, median, and mode, then change that 10 into a 100 and recompute.

STEP 1 — MEAN OF THE ORIGINAL SET

Add: 4 + 5 + 5 + 6 + 10 = `30`. Divide by `n = 5`: 30 ÷ 5 = `6`. Mean = 6.

STEP 2 — MEDIAN OF THE ORIGINAL SET

Already in order: 4, 5, **5**, 6, 10. The middle (3rd of 5) is `5`. Median = 5.

STEP 3 — MODE OF THE ORIGINAL SET

The value 5 appears twice, everything else once. Mode = `5`. So far the three are close: 6, 5, 5.

STEP 4 — NOW MAKE IT AN OUTLIER: CHANGE 10 → 100

New set: 4, 5, 5, 6, **100**. This single edit is the whole experiment.

STEP 5 — RECOMPUTE THE MEAN

Add: 4 + 5 + 5 + 6 + 100 = `120`. Divide by 5: 120 ÷ 5 = `24`. The mean leaped from 6 to **24** — it's now bigger than four of the five scores!

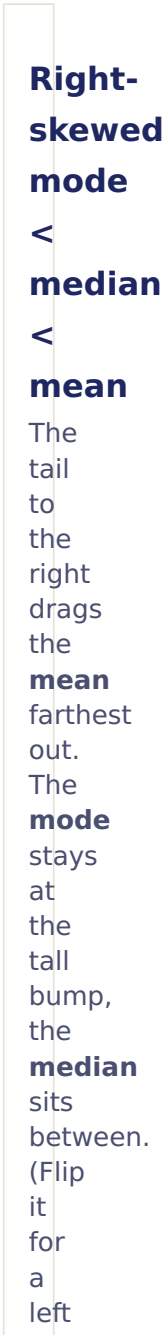
STEP 6 — RECOMPUTE THE MEDIAN & MODE

The order is still 4, 5, **5**, 6, 100, so the middle is still `5`. The mode is still `5`. Both barely noticed.

THE PUNCHLINE

> One outlier dragged the **mean** from 6 to 24 but left the **median** (5) and **mode** (5) untouched. When a few extreme scores stretch your data, the mean chases them and the median holds steady.

That stretched-out shape has a name from Unit 2: **skew**. When a long tail pulls the data to one side, the three centers spread out in a *predictable order*. The mean, chasing the tail, ends up farthest out; the mode (the tallest bump) stays put; the median sits between them. Here's the classic right-skewed picture:



skew:
mean
<
median
<
mode.)

Remember the order

In a **right**-skew the alphabet helps: **mean, median, mode** line up from the tail back toward the bump (mode < median < mean as you read up the number line). In a **left**-skew it reverses. And when the data is symmetric, all three sit on top of each other — that's why a mean and median that nearly match signal a balanced distribution.

Stat in the Wild — the $95,000 salary

A college brags that its graduates earn an **average** starting salary of **$95,000**. Sounds amazing — until you learn that *two* grads went to Wall Street and made $400K each, while everyone else makes about $42K. The **mean** got yanked upward by two outliers. The **median** grad earns $42K — the honest number. (This is exactly the 10→100 trick from the worked example, dressed in a suit.)

The lesson: whenever data is **skewed** — income, home prices, follower counts, response times — the **median** tells the truth and the **mean** oversells. Always ask: *which average, and why that one?* Reputable outlets report *median* household income for exactly this reason.

6. So which one should you use?

Two questions decide it: **What's the shape?** and **What's the scale of measurement?** (the NOIR ladder from Unit 1). Here's the whole decision in two tables.

Situation	Best center	Why
Roughly **symmetric**, numeric data	**Mean**	Uses every value; feeds later tests. Mean ≈ median anyway.
Skewed or has outliers (income, prices, times)	**Median**	Resists the extremes; reports the true "typical."
Categorical data (groups, labels)	**Mode**	The only one that works — there's nothing to add or order.
You want the **single most frequent** value	**Mode**	"Most common shoe size," "peak ordering hour."

And here's how the **scale of measurement** sets a ceiling on which centers are even *allowed* — a measure needs the math its scale supports:

Scale	Example	Mode?	Median?	Mean?
Nominal (labels)	diet, gender	✔	✘	✘
Ordinal (ranked)	class year, Likert agree	✔	✔	✱
Interval (equal steps, no true 0)	temperature °F	✔	✔	✔
Ratio (true zero)	commute min, sleep hrs	✔	✔	✔

✱ Means of ordinal data (e.g., averaging a 1–5 Likert scale) are extremely common in psychology but technically assume the steps are equal — a debate you'll hear about for the rest of your research life. Mode and median are always safe on ordinal data; the mean is the contested one.

> **The 30-second rule**
>
> Numbers and roughly symmetric? → **mean**. Numbers but skewed/outliers? → **median**. Categories? → **mode**. When in doubt, report the mean *and* the median — if they're close, use the mean; if they're far apart, your data is skewed and the median is the honest one.

Try it now — the Descriptive Statistics calculator

Paste a list of numbers (or hit **Load class data** to drop in real commute minutes), then read the mean, median, and mode straight off — the same three numbers you'd get from the Sheets formulas below. Load the class commute data and watch the mean sit *above* the median: that's the right-skew you just learned to read.

The typical animal life

Central tendency reframes hard-to-feel numbers into a single graspable one. Each year roughly **83 billion land animals** are slaughtered for food worldwide — a number too big for the mind to hold. But express it as a center: the **typical broiler chicken** (about 90% of those animals) lives only about **42 days** before slaughter, against a natural lifespan of several years. One median-style number — "about six weeks" — does more moral work than the billions ever could.

Our class dataset carries the quieter version. Run the three measures on `animal_compassion` by `diet` and the *centers* separate cleanly: omnivores cluster lower, vegans higher, vegetarians between — a

preview of the group comparisons we'll formalize in later blocks. Zoom out and the framing sharpens further: animal agriculture uses **83% of the world's farmland** but provides only **18% of our calories** (Poore & Nemecek, 2018). A single well-chosen "average" can carry real weight — and choosing it *honestly* is the skill this unit teaches.

In Google Sheets

You won't compute these by hand on exam day — Sheets does it in one function each. Open the **SJNY Student Life** dataset and try these on the `stress` column (column J):

```
=AVERAGE(J2:J201)     → the mean stress score   (≈ 20.1)
=MEDIAN(J2:J201)      → the middle stress score (≈ 19)
=MODE(J2:J201)        → the most common score   (= 18)
```

Because the mean (≈20.1) and median (≈19) sit close together, stress is roughly **symmetric** — the mean is a fine summary here. Now try the same three on `commute_minutes` (column W) and you'll see the gap open up:

```
=AVERAGE(W2:W201)     → mean commute   (≈ 22.8 min)
=MEDIAN(W2:W201)      → median commute (= 20.5 min)
```

The mean sits above the median because a handful of brutal 70-minute-plus commutes pull it up — **right-skewed**, so the median (20.5) is the typical commuter. Two notes: newer Sheets uses `=MODE.SNGL` for one mode and `=MODE.MULT` when several tie, and `=AVERAGE` politely ignores blank cells (good) but counts a stray 0 (be careful). When you want a quick read on shape, drop both `=AVERAGE` and `=MEDIAN` side by side — far apart means skew.

Watch out for

Outliers drag the mean, not the median — one weird score changes everything for the mean and almost nothing for the median. **Round at the very end**, never mid-calculation. **You can't average a category** (there's no "mean gender") — use the mode. **The mode can tie** or barely exist — that's normal, not an error. And when someone quotes "the average," **ask which one** — mean and median can tell very different stories about the same data.

Key terms

Central tendency · **Mean ($\bar{x}$ / μ)** · **Median** · **Mode** · **Σ (sum)** · **Balancing point** · **Outlier** · **Skew** · **Resistant measure** · **Bimodal / multimodal** · **Scale of measurement**

PRACTICE — NEVER GRADED

Check yourself

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

Sigma says: Before you pick an "average," glance at the shape. If a few extreme scores are stretching the data out, the *median* is the honest one. When mean and median sit close together, the mean is fine — and for categories, the *mode* is your only option.