UNIT 1 · BLOCK 1

Intro & Variables

What statistics really is, and the kinds of data we'll work with all semester.

Why this matters

Statistics has a scary reputation, but here's the secret — it's really about **thinking and deciding**, and the math is just a little 8th-grade arithmetic. This first unit is about getting comfortable: what stats is (and isn't), the words you'll hear all semester, and the one dataset we'll use every week. By the end you'll have it open and running — and you'll never again do math on a column without first asking *what kind of number am I holding?*

By the end you can...

- ✓ Say what statistics **is** and **isn't** — and tell **descriptive** from **inferential**.
- ✓ Tell a **population** from a **sample** (and a parameter from a statistic).
- ✓ Sort any variable into **categorical vs quantitative** and name its **scale of measurement**.
- ✓ Open the **SJNY Student Life** dataset and find your way around Google Sheets.

1. The big idea — stats is thinking, not just math

Statistics is **not** a branch of pure math.

It uses math, yes — but its real job is to take a messy pile of numbers and help you **think clearly and decide** something. "Do students who sleep more get better grades?" That's a stats question, and the arithmetic is the easy part. If math has scared you off before, take a breath: the hard part here is the *thinking*, and you already do that every day.

Here's the shape of it: every statistical question is really a **decision under uncertainty**. You have *data* (raw numbers), you turn it into *information* (a summary, a comparison, a pattern), and you use that to make a *decision* or a claim. Nearly every finding you'll read in psychology — "therapy reduces anxiety," "sleep predicts GPA," "this drug beats placebo" — is exactly this move: numbers from a small group, turned into a careful claim about a bigger one.

Statistics does two big jobs:

Type	Plain English	What it answers	Example
Descriptive	*Summarize* the data you have	"What does this group look like?"	"Our 200 students sleep 6.8 hours on average."
Inferential	*Generalize* beyond the data	"What's true of the bigger group?"	"Students in general probably sleep about 7."

Most of **Block 1** is **descriptive** — describing the group in front of us with averages, spreads, and pictures. Later in the course we move to **inferential**, where we use a small group to make a smart, hedged guess about a much bigger one. Same data, two very different jobs.

CHECK

A news site reports that the *average age of its 1,200 survey respondents is 41*. Is that descriptive or inferential?

Descriptive. They're summarizing the 1,200 people they actually measured — no leap to a bigger group. It would become *inferential* the moment they used those 1,200 to claim something about all readers.

2. Population vs sample

A **population** is *everyone* you care about — say, all students at the university. A **sample** is the smaller group you actually measured — like our 200 students. Two more words travel with these:

- A number describing the **whole population** is a **parameter**.
- A number describing your **sample** is a **statistic**.

Why bother with a sample at all? Because measuring the whole population is usually impossible, too slow, or too expensive — and sometimes the population is hypothetical (all *future* patients who might take a drug). So we measure a manageable slice instead. The catch: a sample is only useful if it's **representative** — if it resembles the population in the ways that matter. A **biased** sample (say, only students who volunteer for an 8 a.m. study) can mislead you no matter how large it is. Bigger is not the same as fairer.

One tiny wrinkle to expect every time: a sample statistic won't exactly equal the population parameter — that gap is **sampling error**, and taming it is what the whole back half of the course is about.

Notation you'll meet later (no need to memorize yet)

There's a handy convention: numbers about a **population** usually wear **Greek** letters — μ (mu) for its mean, σ (sigma) for its standard deviation. Numbers about a **sample** wear **Latin** ones — $\bar{x}$ ("x-bar") for its mean, s for its standard deviation. Quick mnemonic: **Greek = population/parameter, Latin = sample/statistic**.

The whole game in one sentence

We almost never get the whole population, so we work with a **sample** and a **statistic**, then use **inferential** stats to make a smart guess about the **parameter**. That's it — that's the entire arc of the course.

CHECK **A polling firm surveys 1,000 voters to predict how all 5 million voters in the state will vote. The 1,000 surveyed are the…**

Sample. The 5 million voters are the population (everyone they care about); the 1,000 actually measured are the sample. Their predicted percentage is a *statistic* used to guess the population *parameter*.

3. Variables: categorical vs quantitative

A **variable** is just a thing that *varies* from person to person — like gender, sleep, or GPA. (If it doesn't vary — say everyone in your study is a sophomore — then for your study it's a **constant**, not a variable.) Variables come in two flavors:

Type	What it does	In our dataset
Categorical *(qualitative)*	Puts people in **groups** — the value is a label, not a quantity	`gender` (B), `commuter` (D)
Quantitative	An actual **number** you can do math on	`sleep_hours` (F), `gpa` (K)

Quantitative numbers come in two styles:

- **Discrete** — countable, in whole steps. It answers "*how many?*" (number of siblings, `words_recalled` in column R). You can't recall 7.5 words.
- **Continuous** — can land anywhere on a scale, decimals included. It answers "*how much?*" (`commute_minutes` in column W, `sleep_hours`). The only limit is how precisely your instrument measures.

One subtlety worth knowing: the *same* real-world thing can be measured either way. "Caffeine" could be a category (coffee drinker vs not) or a quantity (mg per day). How you *measure* it — not the thing itself — decides the type, and that decision controls every statistic that follows.

CHECK **"Number of siblings a student has" is which kind of quantitative variable?**

Discrete. Siblings come in whole counts — 0, 1, 2, 3 — with no values in between. Nobody has 2.4 siblings. Contrast with `sleep_hours`, which can be 6.8 (continuous).

4. The 4 scales of measurement

This is the big one. Every variable sits on one of four scales, and **the scale decides which statistics you're allowed to use.** They build on each other, from least to most information:

Scale	What it is	In our dataset
Nominal	Unordered **categories** — just names	`gender` (B), `commuter` (D), `pay_condition` (O)
Ordinal	Categories with an **order**, but uneven gaps	`class_year` (C), `processing_level` (Q)
Interval	Numbers, **equal gaps**, but **no true zero**	scale scores: `stress` (J), `wellbeing` (M), `self_compassion` (N)
Ratio	Numbers with a **true zero** (zero means none)	`sleep_hours` (F), `gpa` (K), `commute_minutes` (W)

The reason this matters so much is that the scale sets a **menu** of moves you're allowed to make. Each step up the ladder unlocks more:

Scale	You can...	Center you may report
Nominal	count, compare frequencies, find the most common category	**Mode**
Ordinal	...also rank and find the middle position	**Median** (+ mode)
Interval	...also add/subtract and average; differences are meaningful	**Mean** (+ median, mode)
Ratio	...also form true ratios ("twice as much") because zero is real	**Mean** + ratios/ CV

Quick test for the top two

Is there a true zero? With `commute_minutes`, 0 means *no commute at all* — and 40 minutes is truly twice 20. That's **ratio**. With a `stress` scale, a 0 doesn't mean "no stress exists," and a 30 isn't literally "twice as stressed" as a 15 — the gaps are equal but the zero is fake. That's **interval**.

The classic interval trap is temperature. 20°C is not "twice as hot" as 10°C, because 0°C doesn't mean "no temperature" — it's just where someone decided to put the zero. Equal gaps (each degree is the same size), fake zero → interval. Height and weight, by contrast, have real zeros, so they're ratio: 100 kg really is twice 50 kg.

Worked example — is `commute_minutes` interval or ratio?

Walk the same checklist you'll use on the exam. Reveal one step at a time and try to predict each before you click.

STEP 1 — NUMBER OR LABEL?

It's a real count of minutes you can add and subtract, so it's **quantitative**, not categorical. That already rules out nominal and ordinal.

STEP 2 — EQUAL GAPS?

The jump from 10→20 minutes is the same size as 50→60. Equal intervals everywhere — so it's at least **interval**.

STEP 3 — TRUE ZERO?

0 minutes means **no commute at all** — genuine "none," not an arbitrary marker. That's the signature of a **ratio** scale.

STEP 4 — RATIOS MEANINGFUL?

Because the zero is real, "40 is twice 20" is a true statement. **Verdict: ratio.**

Why it matters: you can find the most common `gender` (a mode), but you can't take its *average* — there's no "mean gender." **Scale first, statistic second.** We'll lean on this all semester.

The 8th-grade math refresher

Don't panic. Here's *everything* you'll actually need — and it really is just middle-school arithmetic dressed up in Greek letters.

> **The whole toolkit**
>
> **Order of operations (PEMDAS):** Parentheses, Exponents, Multiply/Divide, Add/Subtract. So `2 + 3 × 4 = 14`, not 20 — the × happens first.
>
> **Negative numbers:** subtracting a bigger number gives a negative. `5 − 8 = −3`. Two negatives multiplied turn positive: `−2 × −2 = 4`.
>
> **Rounding to 2 decimals:** look at the 3rd decimal. `3.146 → 3.15`; `3.142 → 3.14`. Round at the **end**, not mid-calculation.
>
> **Squaring & square roots:** squaring means times-itself — `4² = 16`. A square root undoes it — `√16 = 4`. In Sheets that's `=SQRT(16)`.
>
> **Summation Σ:** that big Greek letter just says **"add them all up."** `Σx` = add up all the x's. Nothing more.

Σ shows up constantly because almost every formula starts by adding things together. The very first one you'll meet is the **mean** (the average), and it's just two of these tools stacked:

> **Your first formula**
>
> **mean** = `x̄ = Σx ÷ n` — add up all the values (**Σx**), then divide by how many there are (**n**). That's the entire thing. Try it below.

Google Sheets basics

We do everything in Sheets, in the lab. Four things to know:

- **Columns are letters** (A, B, C…) and **rows are numbers** (1, 2, 3…).
- A **cell** is one box, named by its column + row — like `J2` (column J, row 2).
- A **range** is many cells in a line — like `J2:J201` (all 200 students' stress scores).
- **Every formula starts with `=`.** Type a formula and Sheets does the math for you.

In our dataset the **data lives in rows 2–201** (row 1 is the column headings), so a full column of data is always `2:201`. A handful of functions will carry you through Block 1:

```
=AVERAGE(F2:F201)     → the mean of all 200 sleep_hours values
=SUM(G2:G201)         → Σx: adds a whole column
=COUNT(F2:F201)       → n: how many numbers are in the range
=COUNTIF(C2:C201,"Senior")  → counts how many Seniors (categorical!)
=SQRT(16)             → 4   (Sheets does square roots too)
```

Notice the payoff of Unit 1's big lesson right here: `AVERAGE` and `SUM` only make sense on *quantitative* columns. Point them at `gender` and you'll get nonsense — for a categorical column you reach for `COUNTIF` instead. **The scale you decided on tells you which formula is even legal.**

CHECK **You want to know how many students in the dataset are commuters (`commuter` = "Yes"/"No"). Which formula fits?**

=COUNTIF. `commuter` is categorical (nominal), so averaging is meaningless — there's no "mean Yes/No." You *count* categories. `AVERAGE` is for quantitative columns like `sleep_hours`.

A number that needs a scale

Here's a number to keep in your back pocket all semester: roughly **83 billion land animals** are farmed for food each year — about **2,600 every second**, and **around 90% of them are chickens**. That's a *count* of a *categorical* variable (species), which means the right summary is a frequency or a percentage, never an "average species."

One more figure shows why the *kind* of number matters: meat and dairy use **83% of the world's farmland** but provide only **18% of our calories** (Poore & Nemecek, 2018). Those are percentages — ratio-scale numbers with a true zero — so they're fair to compare directly. The skill this whole unit is building is exactly that: before you do *any* math, ask what kind of number you're holding. We'll graph and argue about data like this all term, because statistics is most worth learning when the numbers are about something that matters.

Stat in the Wild — numbers that aren't really numbers

Your **ZIP code** is written with digits. So is your **phone number** and a player's **jersey number**. But try to find the **average ZIP code** of your class — it's nonsense. ZIP 10301 isn't "more" than 07024, and the average of two phone numbers doesn't call anyone.

The lesson: these are **nominal** labels that just happen to be printed as digits. Before you do *any* math on a column, ask "*is this a real number, or just a label in disguise?*" That single question saves you from a lot of nonsense.

Watch out for

Digits ≠ quantity — ZIP codes, IDs, and pay-condition codes are categories even when they look numeric. There's **no true zero on interval scales**, so you can't say one stress score is "twice" another. **You can't average a category** (no "mean gender") — use a count or the most-common value. **Round at the end**, and remember every **formula needs the** `=` or Sheets just shows your text.

Key terms

Statistics · Descriptive · Inferential · Population · Sample · Parameter · Statistic · Representative sample · Sampling error · Variable · Constant · Categorical · Quantitative · Discrete · Continuous · Nominal · Ordinal · Interval · Ratio · True zero · Mode · Median · Mean · Cell · Range

PRACTICE — NEVER GRADED

Check yourself

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

Sigma says: When a column of numbers shows up, don't reach for the calculator yet — ask "*is this a real quantity, or a label in disguise?*" Decide the scale first; the right statistic falls out of that one question.