UNIT 8 · BLOCK 3

Probability & the CLT

How one sample tells us about everyone.

Why this matters

Up to now we've just **described** the 200 students in front of us. Starting today, we want to say something about **all** students — people we never measured. We can't survey everyone, so we take a **sample** and trust it to stand in for the whole group. This unit is the bridge from "here's what we found" to "here's what's probably true everywhere." One idea does the heavy lifting: the **Central Limit Theorem**. It's the engine behind every test in the rest of the course.

By the end you can…

- ✓ Say what **probability** means in one sentence — a number from 0 to 1, the long-run relative frequency.
- ✓ Explain why we use a **sample** to learn about a **population**.
- ✓ Describe the **sampling distribution of the mean** and state the **Central Limit Theorem** correctly.
- ✓ Compute a **standard error** with **SE = σ ÷ √n** and explain why it shrinks as *n* grows.

The big idea

We can't measure everyone. So we grab a **random, representative sample**, compute its **mean**, and use that to **estimate** the whole group. The catch: a different sample would give a slightly different mean. The Central Limit Theorem tells us *exactly* how much those means bounce around — and that's what makes guessing about a whole population a science, not a shot in the dark. Here's the cast of characters we'll meet:

Term	Symbol	Plain English
Population	—	Everyone you care about (all college students)
Sample	n	The smaller group you actually measure
Population mean	μ	The true average you're trying to learn
Sample mean	$\bar{x}$	The average of your one sample — your estimate of μ
Sampling distribution	—	All the possible sample means, gathered into one picture
Standard error	SE	How much those sample means spread out

Greek = population, Latin = sample

A small convention that pays off all term: **Greek letters describe the population** (μ for its mean, σ for its standard deviation) and **Latin letters describe your sample** ($\bar{x}$ for its mean, s for its standard deviation). The population values are usually unknown — that's the whole reason we sample.

1. Probability in one breath

Probability is just a number from **0 to 1** that says how likely something is. **0** means it never happens, **1** means it always happens, and **0.5** is a coin-flip. Multiply by 100 if you'd rather speak in percent (0.5 = 50%). Two things can never happen: a probability below 0 or above 1. If your arithmetic ever produces one, you made an error.

Where does that number come from? It's the **long-run relative frequency** — the share of times an event happens if you repeat the situation many, many times. Flip a fair coin **10** times and you might get 7 heads; that's noise. Flip it **1,000** times and you'll be hovering right near **500** — about half. Probability is the value that count *settles toward* as the run gets long. And notice the move: a probability is a **count over a total**, exactly the relative-frequency calculation you did building tables back in Unit 2.

A quick class example

In our dataset, 18 of the 200 students are vegan. So the probability of randomly drawing a vegan student is **18 ÷ 200 = 0.09 = 9%**. Probability is just a count over a total — the same move you made building frequency tables back in Unit 2.

CHECK **A classmate computes a probability and gets 1.4. What should you tell them?**

It's an error. Probability lives on a 0-to-1 scale: 0 = never, 1 = always. Anything above 1 (or below 0) means a count or total got mixed up somewhere. There's no such thing as "140% likely."

2. Why we sample

Nobody can measure every college student in America. It would cost a fortune and take forever. So we take a **sample** — a manageable slice

— as long as it's **random** (everyone has a fair chance of being picked) and **representative** (it looks like the group it came from). Then we use the **sample mean ($\bar{x}$)** to estimate the **population mean (μ)**. A good sample is a small, honest mirror of the whole.

The word "random" is doing real work. Random selection is what protects you from quietly stacking the deck — surveying only your friends, only morning people, only students who answer email. A sample of 50 chosen at random beats a sample of 5,000 that all came from one dorm. **How** you pick matters more than how many you pick.

> **Size can't fix bias**
>
> This is worth saying twice because it costs people exams and costs companies millions: a **big** sample that was gathered the **wrong** way is still biased. More people just gives you a more precise wrong answer. Everything in this unit assumes the sample was drawn fairly to begin with.

3. The sampling distribution of the mean

Here's the move that unlocks everything. Imagine you don't take *one* sample — you take **many**, each of the same size n. For each one, you write down only its **mean**. Sample 1's mean is 20.3. Sample 2's is 19.7. Sample 3's is 20.1. Keep going for thousands of samples.

Those means **pile up into their own distribution** — the **sampling distribution of the mean**. Read that name carefully: it is a distribution of *averages*, not of individual people. Each dot in it is one whole sample's mean. This is the single most-confused idea in the course, so hold onto the picture: you're drawing a histogram where every value plotted is itself an average of n scores.

And it behaves in a beautifully predictable way. Averages don't swing as wildly as individuals — one unusually tall person barely budges the average height of a roomful of people. So the sampling distribution is

narrower than the population, and it's **centered right on μ**. That predictability is what the next idea names and explains.

CHECK **One dot in a sampling distribution of the mean represents…**

One whole sample's average. The sampling distribution is built from *means*, not individuals. Each point is the x̄ of an entire sample of *n* people — that's why it's tighter than the distribution of raw scores.

4. 🎯 The Central Limit Theorem

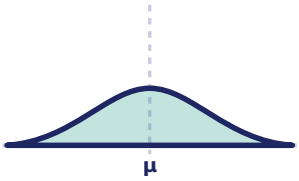
This is the star of the whole course. The **Central Limit Theorem (CLT)** says: **as the sample size *n* grows, the sampling distribution of the mean approaches a normal distribution — no matter what shape the population has.** Spelled out, the distribution of sample means is:

- **Centered** on the true population mean μ — it aims at the right target, so on average x̄ neither over- nor under-shoots,
- **Approximately normal** — that friendly bell shape — even if the population itself was lopsided, and the bigger *n* is, the more perfectly bell-shaped it gets,
- **Much narrower** than the population, because averages don't swing as wildly as individuals (we'll measure exactly how narrow in the next section).

The magic

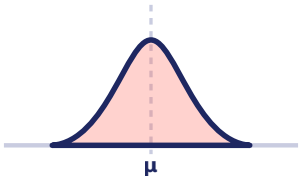
That second point is the payoff. The population can be skewed, jagged, weird — the **sample means still come out bell-shaped.** That's why we can use the normal curve (from Unit 5) for almost any variable. A common rule of thumb: by about **n = 30** the sampling distribution is close enough to normal to use, and the more skewed the population, the larger *n* you'd want. Here's that bell — slide the controls to see how a mean and spread shape it:

Watch what the SD slider does: a **small SD** gives a tall, tight bell; a **big SD** gives a wide, flat one. When you picture the sampling distribution of the mean, imagine that bell getting **narrower and taller** as your sample grows — that's the next idea, and here it is in three pictures:



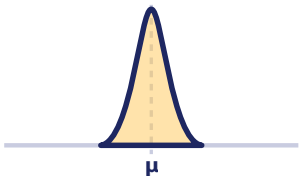
n = 4

Wide and low. Small samples → means scatter a lot. SE is large.



n = 25

Taller and narrower. Bigger samples pull the means in toward μ.



n = 100

Tall and tight. Large samples → means hug μ. SE is small.

All three bells sit over the **same center, μ** — bigger samples don't move the target, they just **tighten the aim**. The width keeps shrinking, and that width has a name and a formula.

5. The standard error = how much the means spread

The spread of that distribution of means has a name: the **standard error (SE)**. It's just the standard deviation of the sample means — how far a typical sample mean lands from μ. And there's a clean formula:

The one formula

SE = σ ÷ √n. Read it slowly. A **bigger n** (a bigger sample) makes the bottom bigger, which makes the **standard error smaller**. A smaller standard error means the sample means cluster **tighter** around the truth — so your estimate is **more reliable**. Big samples give steady answers.

Notice the **√n**, not *n*. That square root is why precision comes at a rising price: to cut the standard error in **half** you don't double the sample, you have to **quadruple** it. Going from n = 25 to n = 100 (4×) only halves the spread. The first hundred people you measure buy you far more precision than the second hundred — a real budgeting fact for anyone who runs studies. Let's watch the formula bite, holding σ fixed at 10:

Worked example — watch SE shrink as *n* grows

Population SD is fixed: σ = 10. We change only the sample size *n* and recompute SE = σ ÷ √n. Predict each result before you reveal it.

STEP 1 — SMALL SAMPLE, N = 4

√4 = 2, so SE = 10 ÷ 2 = 5. With only 4 people per sample, the means scatter widely — typically about 5 units from μ.

STEP 2 — MEDIUM SAMPLE, N = 25

√25 = 5, so SE = 10 ÷ 5 = 2. Jumping to 25 more than halved the spread. The means now land within about 2 of μ.

STEP 3 — LARGE SAMPLE, N = 100

√100 = 10, so SE = 10 ÷ 10 = 1. With 100 per sample the means hug μ — typically within 1 unit.

STEP 4 — READ THE PATTERN

The SD never changed (σ = 10 the whole time), yet SE fell 5 → 2 → 1. **Bigger samples give more precise means.** And note the √n tax: it took 4× the sample (25→100) just to halve the spread (2→1).

CHECK **A population has σ = 20. You take samples of n = 4. What's the standard error?**

10. SE = σ ÷ √n = 20 ÷ √4 = 20 ÷ 2 = 10. Don't divide by *n* (that would give 5) — you divide by the *square root* of *n*. The √ is the whole reason precision improves slowly.

6. Worked example on OUR data — and back to z

Open the **SJNY Student Life** dataset. We want the standard error of `stress` (column J). We already know stress has **SD = 5.90** across all **n = 200** students. So:

> **Plug it in**
>
> **SE = 5.90 ÷ √200 = 5.90 ÷ 14.14 = 0.42**

What does **0.42** mean? If we took sample after sample of 200 students, their mean stress scores would typically land within about **0.42** of the true average. That's a tight, trustworthy estimate — because *n* is big.

And here's the CLT flexing: our `commute_minutes` (column W) is **right-skewed** — a few students have brutal 90-minute commutes that stretch it out. But take samples from it and the **means still come out normal**. The lopsided shape of the raw data doesn't matter. The means behave.

> **This is where z comes back**
>
> Remember the **z-score** from Unit 5: **z = (X − μ) ÷ σ** — how many standard deviations a *score* sits from the mean. For a *sample mean*, swap in the same pieces from the sampling distribution: the center is still μ, but the spread is now the standard error. So **z = (x̄ − μ) ÷ SE**. Because the CLT made the sampling distribution normal, that z drops straight onto the normal curve you already know — and that single substitution is the whole machinery behind every hypothesis test in Block 4.

Play with a sample yourself

Paste any list of numbers below (or load class data) to see its mean and standard deviation. Then try the formula by hand: take the **SD** it

reports, pick a sample size *n*, and compute **SE = SD ÷ √n**. The bigger the *n* you imagine, the smaller that standard error gets.

Sampling in the real world

This unit isn't abstract — it's how every honest claim about a large group gets made. When researchers report that, say, only a **minority of Americans** have ever met a vegan, or that public concern for farmed-animal welfare is rising, they did *not* ask all 330 million people. They drew a **random, representative sample** of perhaps 1,000–2,000 and let the Central Limit Theorem carry the inference from the few to the many. The standard error is what lets them attach a margin to the claim instead of just guessing.

The same logic powers the studies behind this whole site's running thread. A finding like "meat and dairy use 83% of farmland but provide only 18% of calories" (Poore & Nemecek, 2018) rests on a **sample** of roughly 38,000 farms standing in for global agriculture — not a census of every farm on Earth. Whenever you read a number about a population too big to count, a sampling distribution is quietly doing the work underneath.

Stat in the Wild — why bigger samples are trustworthy

You've seen a poll say "1,500 adults, margin of error ±3%." Why 1,500? Because of everything in this unit. A poll of **1,500** people has a **smaller** standard error than a poll of **100** — bigger *n*, smaller SE, tighter estimate. This is the **law of large numbers**: as your sample grows, its average **settles down** and stops jumping around.

Flip a coin 10 times and you might get 7 heads (70%). Flip it 1,000 times and you'll be hovering right near 50%. The average **calms down as n grows.** That's why a national poll of 1,500 beats a quick survey of your 10 friends — and why "we asked more people" really does mean "trust us more" (as long as the asking was random).

In Google Sheets

You don't compute a standard error by hand — Sheets does it. Open the **SJNY Student Life** dataset and try the standard-error formula on `stress` (column J):

```
=STDEV(J2:J201)/SQRT(COUNT(J2:J201))     → the standard error of stress  ≈ 0.42
```

You can also **watch the Central Limit Theorem happen** with two formulas. `=RANDBETWEEN` draws a random whole number; wrap a batch of draws in `=AVERAGE` and you've simulated one sample mean:

```
=AVERAGE(RANDBETWEEN(1,6), RANDBETWEEN(1,6), RANDBETWEEN(1,6), RANDBETWEEN(1,6))
      → the mean of four random dice rolls — one dot in the sampling distribution
```

Fill that formula down 500 rows and you have 500 sample means. Then graph them with **Insert ▸ Chart ▸ Histogram**: even though a single die roll is **flat** (uniform — every face equally likely), the **averages pile into a bell**. That's the CLT, live, in your own spreadsheet. Add more dice to each average and watch the bell get narrower — you've just demonstrated SE = σ ÷ √n by hand.

Watch out for

Standard error is **not** the standard deviation — SD describes how spread out *people* are; SE describes how spread out *sample means* are, and SE is always smaller (you divided by √n). Bigger *n* shrinks the standard error, never the other way around — and you divide by **√n**, not *n*. The CLT is about *means*, not individuals — individual stress scores aren't bell-shaped just because the means are. And a bad sample stays bad: the CLT can't fix a sample that wasn't random or representative. Garbage in, garbage out.

Key terms

- **Probability**
- **Long-run relative frequency**
- **Population (μ, σ)**
- **Sample ($\bar{x}$, s, n)**
- **Sampling distribution of the mean**
- **Central Limit Theorem**
- **Standard error (σ ÷ √n)**
- **z = ($\bar{x}$ − μ) ÷ SE**
- **Law of large numbers**

PRACTICE — NEVER GRADED

Check yourself

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

Sigma says: When the CLT feels slippery, hold onto one line — *the means go normal even when the people don't.* The raw data can be as lopsided as it likes; gather enough sample means and they pile into a bell. That single sentence is the engine for the whole rest of the course.