Part 3.1 | Data vs the Population

Overview

- We shift from describing data to asking questions about things we can't observe
- Sleep data motivates the distinction between data questions and population questions
- Measures of centrality: mode, median, mean — why the mean wins (tennis court analogy: where to stand)
- Measures of dispersion: build up from range → mean deviation → MAD → variance → SD, each fixing the previous (tennis court analogy: how far you'll run)
- Random variable vocabulary: probability function, observation, sample
- If we knew the population distribution, we could answer anything (mathematically or by simulation) — but we never do
- What can we know about the population from just a sample?

The Shift

We've spent Parts 1 and 2 learning to understand data. We can summarize it, visualize it, transform it, and describe relationships between variables. We ended Part 2 with the Starbucks question — Bogo 10 customers spend more than Bogo 5 customers on average, but the distributions overlap. Is the difference real, or just noise?

We couldn't answer that question with the tools we had. And the reason we couldn't is that we were asking questions about *the data itself*. Today we're going to make a subtle but important shift. We're going to start asking questions about things we can't directly observe.

Sleep Data — The Data Question

Here's some data on the sleep patterns of two groups of people. Which group sleeps longer?

[Stage direction: show two overlapping strip plots or histograms of Group A and Group B sleep hours. Group A has higher mean (~7.2 hrs) but wider spread. Group B has lower mean (~6.8 hrs) but tighter distribution. The overlap between the two is substantial.]

Let's start with the obvious. Can we tell from looking at these distributions?

Not immediately — they overlap. Some people in Group B sleep longer than most people in Group A. And some people in Group A barely sleep at all. So which group sleeps longer? We need to be more precise about what "longer" means. Let's compare their centers.

Measures of Centrality

We need a single number that captures the "center" of each group's distribution. Think of it this way: imagine you're on a tennis court and you're about to play me. You know the distribution of where my shots land — some go left, some go right. Where should you stand? You want the spot that minimizes how far you have to run on average. That's what a measure of centrality does — it finds the best single point to summarize a distribution.

There are several options:

Mode: The most frequent value. It tells you the single most likely outcome, but it ignores everything else about the distribution. Two very different distributions can have the same mode. It's not responsive enough.

Median: The middle value when you sort the data. It has nice properties — it's robust to outliers, for instance. You can actually build a whole branch of statistics around medians. But medians are harder to compute and harder to work with mathematically, so they're used less often in practice.

Mean: The average value.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

The mean is the workhorse of statistics. It's simple to compute, it's "smooth" in a mathematical sense (small changes in data produce small changes in the mean), and — as we'll see — it connects beautifully to measures of variability. It's the point that minimizes the total squared distance to all observations. For our tennis court analogy, standing at the mean minimizes the average squared distance you'd have to cover.

Calculate the mean for both groups.

[Stage direction: show the same two distributions with vertical dashed lines at each group's mean. Group A mean ≈ 7.2, Group B mean ≈ 6.8.]

Group A sleeps longer on average. But look at those distributions again. Some people in Group B sleep longer than most people in Group A. The means are different, but the groups are far from separated. Why? Because Group A has more *variability* — the observations are more spread out. We need a way to measure that spread.

Measures of Dispersion

We know the centers of the two groups. But how spread out is each group's data? Thinking back to the tennis court: you're standing at the mean, but how much *running* will you have to do? If my shots are tightly clustered, you won't run not much. If they're spread all over the court, you'll run a lot. We need a number that captures this idea. Let's build up a measure of spread step by step.

Range: The difference between the largest and smallest value. This is simple, but it only uses two data points. It doesn't respond to changes near the middle of the distribution. If one outlier moves, the range changes dramatically. But if any of the other data moves, range doesn't change at all. It ignores most of the information we have. Instead lets think about how far I have to run if I'm standing at the mean.

Mean deviation: You're standing at the mean. Then for each shot, you measure the distance you'd have to run, which is the distance between the mean and the datapoint itself: $x_i - \bar{x}$. Then you take the average of all those distances.

$$\frac{1}{n}\sum(x_i - \bar{x}) \tag{2}$$

This seems natural — how far do you run on average? But there's a problem. Try computing it. Some shots go left (negative distance), some go right (positive distance), and they cancel out perfectly. The mean deviation is always exactly zero. That's not useful.

[Stage direction: show a number line with a few data points and their mean. Draw arrows from each point to the mean. Some point left, some point right. Label the positive and negative distances. Note that they sum to zero.]

Mean absolute deviation: Instead, lets ignore the direction, just measuring how *far* you run. Running left one unit or right one unit are measured as the same deviation. Lets do this by take the absolute value first, then the average.

$$\frac{1}{n}\sum|x_i - \bar{x}| \tag{3}$$

Now the cancellation problem is gone. This is a perfectly reasonable measure of spread, and it has an intuitive interpretation: the average distance you'd run per shot. But mathematically, absolute values are awkward — they create a "kink" at zero that makes them harder to work with in calculus, proofs, and derivations. We want something that handles the sign problem and is also mathematically convenient.

Variance: Square the deviations instead of taking absolute values.

$$Var = \frac{1}{n}\sum(x_i - \bar{x})^2 \tag{4}$$

Squaring accomplishes two things: it makes all deviations positive (since any number squared is non-negative), and it's smooth and differentiable — no kink. The downside is that the units are squared — if we're measuring hours, the variance is in hours-squared, which isn't interpretable.

Standard deviation: Take the square root of the variance.

$$S_X = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2} \tag{5}$$

Now we have a measure that handles negative deviations, is mathematically nice, and has interpretable units. The standard deviation is roughly the average distance each observation sits from the mean — back in the same units as the original data.

[Stage direction: show formulas appearing progressively, each one building on and fixing the previous. Keep all five visible at the end so students can see the progression: range → mean deviation (fails: always zero) → MAD (works but mathematically awkward) → variance (nice math, bad units) → SD (nice math, good units).]

Group A has a larger standard deviation than Group B. That's why the distributions overlap so much — Group A has a higher average but more variability, so its observations are more spread out.

The Shift — The Population Question

We've just described 100 people, 50 in Group A sampled from **Allegheny County** and 50 in Group B sampled from **Butler County**. But I don't particularly care about these 100 people specifically. What I'm interested in is the differences in sleep time not for those in the sample I just drew but for the people living in these two counties *as a population*. And that population is *not* in my sample.

This is a subtlely different question. Instead of asking "which *sample group* sleeps longer?" — a question about the data — we're now asking "which *county* sleeps longer?" That's a question about the **population**.

[Stage direction: show a diagram with two large clouds labeled "County A Population" and "County B Population," each with a small subset of dots highlighted and an arrow pointing down to the sample data we've been looking at.]

Why is this different? The data is just 50 people we happened to sample. If we sampled again tomorrow, we'd get 50 different people with slightly different sleep patterns. The sample might look a little different. But the county hasn't changed.

And "the county" here doesn't just mean the people currently living there. Even if we surveyed every resident today, tomorrow brings new residents, new sleep schedules. The population is a theoretical concept — an infinite pool of possible observations. It's a **random variable**.

Vocabulary

Think of a random variable like a deck of cards.

The **probability function** tells us which cards are in the deck — what values are possible and how likely each one is. For sleep time, the probability function describes how sleep hours are distributed across the entire population.

An **observation** is the card you drew — one person's sleep time.

A **sample** is the record of cards you've drawn — all 50 sleep times we collected.

Data is a sample drawn from a random variable.

[Stage direction: show a visual analogy — a deck of cards on the left labeled "Random Variable (Population)," a single face-up card labeled "Observation," and a row of face-up cards labeled "Sample (Data)." An arrow from the deck to the sample reads "draw."]

We'll use "population" and "random variable" interchangeably — both refer to the process generating our data. The key insight is that the sample statistics we compute ($\bar{x}$, S) are properties of the *sample*, not the *population*. The population has its own mean (μ) and its own standard deviation (σ), and these are what we actually care about.

Known Distributions — What We Could Do

Here's a thought experiment. Suppose I told you I actually made up this data on a computer. I *know* County A's probability function exactly:

$$x_i \sim N(\mu = 7.2, \sigma = 1.5) \tag{6}$$

This says: sleep times in County A follow a normal distribution centered at 7.2 hours with a standard deviation of 1.5 hours. If we know the distribution, we can answer *any* probability question about the population.

Question 1: What proportion of County A sleeps less than 5 hours?

[Stage direction: show a normal curve centered at 7.2 with σ = 1.5. Shade the area to the left of 5. The shaded region is small — roughly 7%.]

We compute `stats.norm.cdf(5, loc=7.2, scale=1.5)` and get about 0.07. Seven percent of the population sleeps less than 5 hours.

Question 2: What proportion sleeps more than 9 hours?

[Stage direction: show the same normal curve, now shading the area to the right of 9. Similar small region — roughly 12%.]

We compute `1 - stats.norm.cdf(9, loc=7.2, scale=1.5)` and get about 0.12.

Question 3: Where does the middle 92% of the population fall?

[Stage direction: show the normal curve with symmetric shading covering the central 92%, leaving 4% in each tail. Mark the lower and upper bounds.]

We compute `stats.norm.ppf(0.04, loc=7.2, scale=1.5)` for the lower bound and `stats.norm.ppf(0.96, loc=7.2, scale=1.5)` for the upper bound. The middle 92% of County A sleeps between about 4.6 and 9.8 hours.

When we know the probability function, we can calculate everything exactly. Any question about the population has a precise answer.

Sometimes the math gets complicated. If the distribution is very complex — like the election models Nate Silver built at FiveThirtyEight — you can't always solve things with a formula. In those cases, we *simulate*: draw thousands of observations from the distribution and compute whatever we need from the simulated data. The random variable captures every possible way the world could go and each draw is one particular way the world went. With enough draws, the simulated answers converge to the true answers. This idea — that repeated sampling from a distribution reveals its properties — is going to become central to much of what we do in Part 3.

Many Distribution Shapes

And random variables don't have to be normal. They come in all shapes — skewed, bimodal, uniform, heavy-tailed, discrete, continuous.

[Stage direction: show a gallery of 6–8 different probability distributions — normal, uniform, exponential, chi-squared, bimodal mixture, Poisson, beta. Label each one. The point is that the "shape of the deck" varies widely.]

Each one of these describes a different kind of random process. The probability function tells us which values are likely, which are rare, and how spread out the outcomes are. If we knew the function, we'd know everything.

PDFs and CDFs

Every continuous random variable has two key functions that describe it.

The **probability density function (PDF)** gives the height of the curve at each value — $f(x)$. It tells us where observations are more or less likely to fall. Two properties define a valid PDF: $f(x) \geq 0$ (probabilities are never negative) and $\int f(x)\,dx = 1$ (total probability sums to one).

[Stage direction: show a normal curve (N(7.2, 1.5)) with a red dashed vertical line at x = 6.0 showing the height f(6.0). The y-axis is unlabeled (no tick marks). The height of the curve at that point is the PDF value.]

The **cumulative distribution function (CDF)** gives the area under $f(x)$ up to each value — $F(x) = P(X \leq x)$. Instead of asking "how dense is the distribution here?", the CDF asks "what fraction of the population falls below this point?" Three properties: $F(x)$ is non-decreasing (more area accumulates as x increases), $F(-\infty) = 0$ and $F(\infty) = 1$ (ranges from 0 to 1), and $F(x)$ gives probability directly — $P(X \leq 5) = F(5)$.

[Stage direction: show the same normal curve with the area under the curve shaded red up to x = 6.0. A red dashed vertical line marks x = 6.0. The shaded area represents F(6.0) = P(X ≤ 6.0).]

These two functions are how we answer probability questions. The PDF tells us where observations concentrate, and the CDF tells us what proportion falls below any threshold. All three exercises that follow use these functions.

The Cliff-Hanger

But the problem is that in practice we *never* know the probability function. I made up the County A example — I generated the data from a known distribution. In real research, we don't have that luxury. We just see the sample.

And the sample statistics aren't the population parameters:

$$\bar{x} \neq \mu \tag{7}$$

$$S \neq \sigma \tag{8}$$

The sample mean $\bar{x}$ is our best guess of μ, and the sample standard deviation S is our best guess of σ. But they're just guesses. If we drew a different sample, we'd get different values for both.

So what can we do? How do we make claims about the population when all we see is the sample?

Looking Ahead

This is the fundamental tension that the rest of Part 3 resolves:

- **Part 3.2** — The Central Limit Theorem tells us the *distribution* of the sample mean, even when we know nothing about the population.
- **Part 3.3** — We use that distribution to build confidence intervals and test hypotheses — quantifying how surprised we should be by any claim about μ.
- **Part 3.4** — Hypothesis testing turns out to be the simplest case of a much more powerful framework — the general linear model — which sets up everything in Part 4.

We can answer questions about an unknown population using just a sample. Next time, we'll start building the tools to resolve this tension.