

Part 3.2 | The Central Limit Theorem

Overview

- Recap: data is a sample from an unknown population; if we knew the distribution, we could answer everything; but we only observe the sample, not the population.
 - Classroom simulation with dice builds intuition for why sample means behave differently from individual observations
 - Increasing sample size ($n = 1, 2, 3, 30$) reveals the bell curve emerging
 - The reveal: sample means follow a normal distribution centered on μ with standard deviation σ/\sqrt{n}
 - Works for any population shape — demonstrated with a skewed distribution
 - Brief derivation of why the standard error is σ/\sqrt{n}
 - This is perhaps the most underappreciated idea in modern science
-

Opening — Where We Left Off

Last time we established a fundamental tension. Data is a sample drawn from an unknown population. We care about the population — which county sleeps longer, what's the true average — but all we see is the sample. If we knew the population's probability function, we could answer any question exactly. But in practice, we never know it.

So all we have are three numbers: n (the sample size), \bar{x} (the sample mean), and S (the sample standard deviation). How do we learn about the population from just these?

Quick Recap — Known vs Unknown

[Stage direction: show two panels side by side. Left panel: a known normal distribution with calculations — " $P(X < 5) = 0.07$," shaded areas, exact answers. Right panel: a question mark where the distribution should be, with only $n = 50$, $\bar{x} = 7.24$, $S = 1.48$ listed below. Label left "Known Distribution" and right "Unknown Distribution."]

When we know the distribution, we can do everything — exact probabilities, exact intervals, exact answers. When we don't, we're stuck with sample statistics that we know aren't quite right. The sample mean isn't the population mean. The sample SD isn't the population SD.

But here's the question for today: even though we don't know the population distribution, is there *anything* we can know about how \bar{x} behaves? The answer is yes, and it's remarkable.

Exercise 3.2 | Simulation: $n = 1$

Let's start with something simple. Everyone in the class is going to help me collect data. Open your exercise notebook and roll a die once. Give me the result.

Exercise 3.2: each student generates a single die roll.

[Stage direction: show a histogram of many die rolls (one per student). The distribution should be roughly uniform — each face appears about equally often, though with noise.]

What does this distribution look like? It's roughly uniform. Each number comes up about the same number of times. That's exactly what we'd expect from a fair die — the probability of each face is $1/6$. Nothing surprising here.

Exercise 3.2 | Simulation: $n = 2$

Now roll twice and calculate the *mean* of your two rolls. Give me that number.

You're each giving me a different number. Why? Each of you drew a different sample! Even though you're all rolling the same die, your two rolls happened to come out differently.

[Stage direction: show a table of a few students' rolls — Student 1 got (3, 5), mean = 4.0; Student 2 got (1, 6), mean = 3.5; Student 3 got (2, 2), mean = 2.0. Then show the histogram of all students' sample means.]

What do you notice about the distribution of these sample means?

[Stage direction: show the histogram of $n = 2$ sample means. It should bunch up in the middle — a pyramid or triangular shape. The extremes (means near 1 or near 6) are rare.]

They bunch in the middle! The means pile up around 3.5 and thin out at the extremes. Why? Think about it — how many ways can you get a mean of 1? One: you have to roll (1, 1). How many ways can you get a mean of 3.5? Many: (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1). If you've ever played games with dice like Settlers of Catan, this is exactly why rolling a 7 with two dice is much more likely than rolling a 2 or a 12. The same logic applies to means.

Exercise 3.2 | Simulation: $n = 3$

Now roll three times and give me the mean.

[Stage direction: show the histogram of $n = 3$ sample means. The bunching is even more pronounced. There's clear curvature — the shape is starting to look bell-like rather than triangular.]

What do you notice compared to $n = 2$? There's curvature now. The edges aren't straight anymore — the distribution is rounding into a curve. And the shape is tighter around the same center as in the two previous samples. The means don't spread as far from the center as they did with $n = 2$.

Exercise 3.2 | Simulation: $n = 30$

Now roll 30 times. Actually, instead of collecting all those numbers a fourth time, I'm just going to simulate rolling 30 dice 1000 times.

[Stage direction: show individual students' samples — they each have 30 numbers, each looks somewhat different. Show a few side by side to emphasize that every sample is unique.]

Look at the variation in your individual samples. Some of you have lots of high numbers, some have lots of low numbers. Every sample looks different. But what about the means?

[Stage direction: show the histogram of $n = 30$ sample means. Very bell-shaped, very tight — clustered closely around 3.5. Then overlay a red curve on top of the histogram.]

The distribution of sample means is very bell-shaped and very tight. Almost all the means are between 3.0 and 4.0.

I'm going to trace out the distribution with a mathematical function in red. What is it?

Central Limit Theorem

That red curve is a **normal distribution** centered on the population mean ($\mu = 3.5$ for a fair die) with a standard deviation of σ/\sqrt{n} .

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad (1)$$

This is the main result of the **central limit theorem**. If you take a sample of n observations from *any* distribution and compute the mean, the distribution of that mean across many samples will be approximately normal, centered on the population mean, with standard deviation equal to the population standard deviation over the square root of the sample size. With very few exceptions, it doesn't matter what the population looks like — uniform, skewed, bimodal. The sample means converge to a normal distribution.

[Stage direction: show the formula prominently. Below it, show the progression from $n = 1$ to $n = 30$ in a row of four small histograms, each with its red normal overlay. The fit improves dramatically from left to right.]

There are three important ideas here.

First, the CLT tells us that the **sampling distribution** will be normal. Even though we don't know the population distribution, the CLT tells us the exact shape of the sample means! Remember we can answer many kinds of questions about a distribution when we know it's shape, so this puts us on some solid footing.

Second, the CLT tells us the center of the **sampling distribution** is the same as the population distribution. So not only do we know the shape, but we know the center gives us information about the population.

Third, the CLT tells us that the variability in the **sampling distribution** gets smaller as the sample size gets larger. This quantity σ/\sqrt{n} has a name — the **standard error**. It tells us how much variability to expect in the sample mean. Larger samples produce less variability.

The Standard Error: σ/\sqrt{n}

Ok but where does \sqrt{n} come from? To show you all the details would take us a bit beyond the core scope of this class, but here's the brief derivation.

We know each observation x_i is drawn independently from a population with variance σ^2 . One of the properties of variance is that the variance of a sum of independent variables is the sum of their variances. So the variance of the sum of n observations is:

$$Var(x_1 + x_2 + \cdots + x_n) = n\sigma^2 \quad (2)$$

To find the variance of the sample mean, the sum of all the data divided by n , you get the its variance gets divided by n^2 :

$$Var\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \quad (3)$$

Then we take the square root of the variance to get the standard deviation of the sample mean.

$$SD(\bar{x}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \quad (4)$$

This is what we call the standard error. As n grows, the standard error shrinks — but it shrinks at the rate of \sqrt{n} , not n . To cut the standard error in half, you need four times as many observations, not twice as many. This is worth remembering.

(Nearly) Any Distribution

The die example used a uniform distribution — every face was equally likely. But the CLT works for other shapes. Let's try a skewed distribution, the chi-squared distribution.

Exercise 3.2 | Chi-Square

Imagine a population where most values are small but a few are very large — something like income, or wait times, or insurance claims.

[Stage direction: show a 4-panel figure, stacked vertically, each panel showing the distribution of sample means for a different sample size from a chi-squared ($df = 3$) population:]

- Panel 1 ($n = 1$): The histogram looks *very skewed right* — just like the population itself. No red curve overlay (or a very poor fit).
- Panel 2 ($n = 5$): Less skewed, starting to look more symmetric. A red normal curve overlaid — it's getting closer but doesn't fit perfectly.
- Panel 3 ($n = 30$): Much more bell-shaped. The red curve fits quite well.
- Panel 4 ($n = 1000$): Very tight, very bell-shaped. The red curve matches almost exactly.]

With $n = 1$, the sample means just *are* the population — skewed, with a long right tail. With $n = 5$, the skew is already diminishing. By $n = 30$, it looks normal. By $n = 1000$, it's indistinguishable from the red normal curve.

This is the power of the CLT. The population can be (nearly) *anything* and the sample means will still converge to a normal distribution.

Key Properties

Three things to notice about the sampling distribution of \bar{x} :

1. **Centered on μ .** The distribution of sample means is centered on the true population mean. The sample mean is neither systematically too high nor too low — it's an unbiased estimator of μ .
2. **Gets tighter with larger n .** As the sample size increases, the standard error σ/\sqrt{n} decreases. Larger samples produce more precise estimates. With $n = 30$, the sample means are clustered much more tightly than with $n = 5$.
3. **Shape approaches normal regardless of the population.** Whether the population is uniform, skewed, bimodal, or something else entirely, the distribution of \bar{x} approaches a normal. The CLT is about the *mean*, not the individual observations.

Assumptions

The CLT isn't all magic. There are a few conditions:

- The observations need to be **independent** — each observation is drawn separately, without being influenced by the others. When we talk about models of timeseries data in Part 4, we'll show that the CLT doesn't work well in that context because timeseries data is rarely **independent**.

- The observations need to be **identically distributed** — they all come from the same population.
- If the population is very non-normal (extremely skewed, heavy-tailed), you may need a **larger** n for the approximation to work well. For most practical purposes, $n = 30$ is often sufficient, but this is general wisdom, not a theorem.

And the CLT says the distribution *approaches* a normal — it doesn't say it equals one exactly. With very small samples from non-normal populations, the approximation can be poor.

What We've Achieved

Let's step back and appreciate what just happened.

We started with an unknown population. We don't know μ . We don't know σ . We don't know the shape of the distribution. All we see is a sample.

And yet, we now know *exactly* how the sample mean behaves. We know it follows an approximately normal distribution. We know it's centered on the population mean. We know how much it varies from sample to sample. We know the formula for its spread.

This is among the most powerful ideas in modern science, not just in empirical economics. It's what makes polling, clinical trials, quality control, physical measurements, and empirical economics possible. We can learn about things we can't observe using things we can.

Looking Ahead

But knowing the sampling distribution is only half the story. We know \bar{x} is approximately $N(\mu, \sigma/\sqrt{n})$. So how do we actually *use* this? How close is our observed \bar{x} to the true μ ? Can we build an interval around \bar{x} that we're confident contains μ ? Can we test whether μ equals some specific value?

Next time, we'll answer all of these. The CLT gives us the distribution. Part 3.3 shows us how to use it — for confidence intervals, for hypothesis tests, and for quantifying how surprised we should be by any claim about the population.