The Central Limit Theorem

Gov 2001: Quantitative Social Science Methods I

Scott Cunningham

Harvard University

Spring 2026

Today's Reading

Required

- **Aronow & Miller**, §3.2.3–3.2.4: CLT, convergence (pp. 99–110)
- **Blackwell**, Ch. 3: Asymptotics (continue)

The CLT is the most important theorem in statistics. It justifies everything we do with confidence intervals and hypothesis tests.

Where We Are

Monday: Law of Large Numbers

- $\bar{Y} \xrightarrow{p} \mu$ (sample mean converges to population mean)
- Tells us *where* the sampling distribution is centered

Today: Central Limit Theorem

- What is the *shape* of the sampling distribution?
- How can we quantify uncertainty about our estimates?

Answer: For large n, the sampling distribution is approximately **normal**.

A Remarkable Fact

Consider: You sample from a population with *any* distribution.

- Uniform, exponential, binomial, weird multimodal…anything

Compute the sample mean $\bar{Y}$.

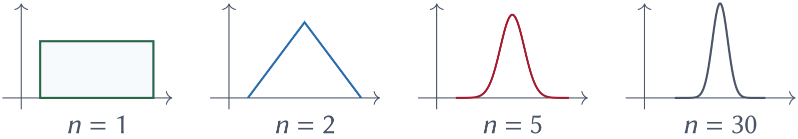
The CLT says: For large n, $\bar{Y}$ is approximately normal.

No matter what the original distribution looks like!

This is why the normal distribution appears everywhere in statistics.

Visual Intuition: Averaging Makes Things Normal

Starting distribution: Uniform on [0, 1]



As n increases: The distribution of $\bar{Y}$ becomes more and more bell-shaped.

Simulating the CLT in R: Setup

Let's see the CLT in action with a simulation.

```r
# Load packages
library(ggplot2)

# Population parameters (Uniform distribution)
pop_mean <- 0.5     # E[X] for Uniform(0,1)
pop_var <- 1/12     # Var(X) for Uniform(0,1)

# Simulation settings
n_sims <- 10000     # Number of samples to draw
```

Key idea: We'll draw many samples, compute each mean, and look at the distribution of those means.

Simulating the CLT: The Core Loop

For each sample size, draw 10,000 samples and compute means:

```
# Function to simulate sampling distribution
simulate_clt <- function(n, n_sims = 10000) {
  # Draw n_sims samples, each of size n
  # Compute mean of each sample
  sample_means <- replicate(n_sims, mean(runif(n)))
  return(sample_means)
}

# Try different sample sizes
n_values <- c(1, 2, 5, 30)
results <- lapply(n_values, simulate_clt)
```

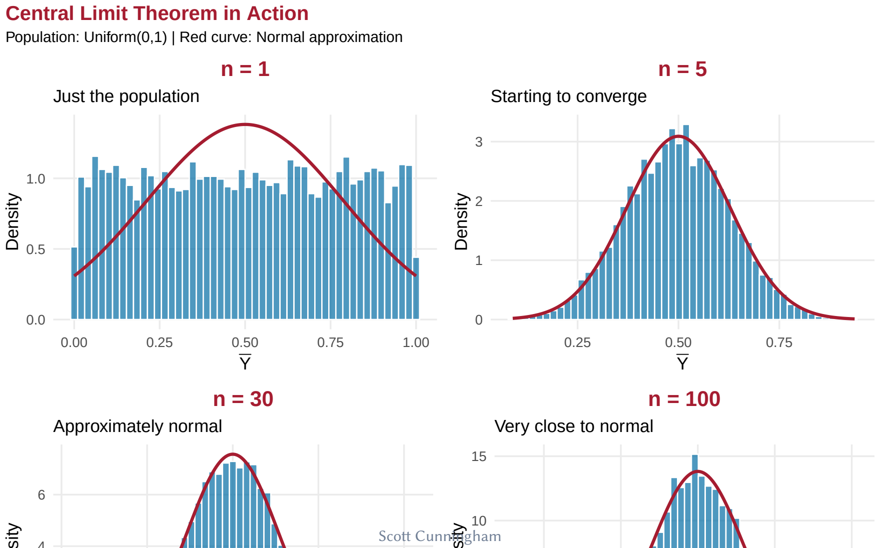
`replicate()` runs the expression `n_sims` times and collects results.

Visualizing the Results

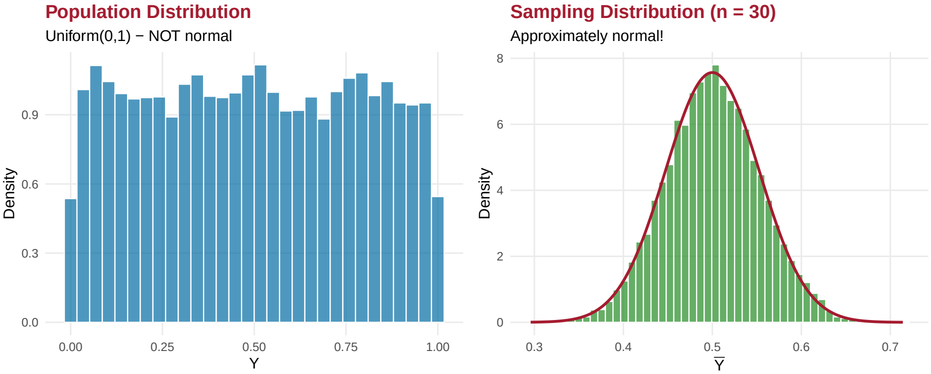
```
# Plot histogram with normal overlay
ggplot(data.frame(xbar = sample_means), aes(x = xbar)) +
  geom_histogram(aes(y = after_stat(density)),
                 bins = 50, fill = "steelblue") +
  stat_function(fun = dnorm,
                args = list(mean = pop_mean,
                            sd = sqrt(pop_var/n)),
                color = "red", linewidth = 1.2) +
  labs(x = "Sample Mean", y = "Density",
       title = paste("n =", n))
```

The red curve: What the CLT predicts—$N(\mu, \sigma^2/n)$.

The CLT in Action



Population vs. Sampling Distribution



Left: The population is uniform (flat). **Right**: The sampling distribution of $\bar{Y}$ (with $n = 30$) is approximately normal.

The Central Limit Theorem

Central Limit Theorem (CLT)

Let $Y_1, Y_2, \ldots$ be I.I.D. with $\mathbb{E}[Y_i] = \mu$ and $\text{Var}(Y_i) = \sigma^2 < \infty$.
Then:

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1)$$

as $n \longrightarrow \infty$.

Equivalent statement:

$$\sqrt{n}(\bar{Y} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

The $\xrightarrow{d}$ means "converges in distribution"—the CDF approaches the normal CDF.

What the CLT Says (Practically)

For large n:

$$\bar{Y} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Or equivalently: $\bar{Y}$ is approximately:

- Centered at μ
- With standard deviation $\sigma/\sqrt{n}$
- And normal (bell-shaped)

This lets us make probability statements about $\bar{Y}$**!**

Example: Presidential Approval Survey

Setup: Feeling thermometer (0–100 scale) toward the president.

- Population: $\mu = 45$, $\sigma = 30$
- Sample: $n = 900$ respondents

By CLT: $\bar{Y} \approx N\left(45, \frac{30^2}{900}\right) = N(45, 1)$

Standard error: $\text{SE} = 30/\sqrt{900} = 1$

Question: What's the probability $\bar{Y}$ is within 2 points of μ?

$$\begin{aligned}\Pr(|\bar{Y} - 45| < 2) &= \Pr\left(\left|\frac{\bar{Y} - 45}{1}\right| < 2\right) \\ &\approx \Pr(|Z| < 2) \approx 0.95\end{aligned}$$

There's a 95% chance the sample mean is within 2 points of the truth.

How Large is "Large Enough"?

The CLT is asymptotic—it's exact only as $n \to \infty$.

In practice: How big does n need to be for the approximation to work?

Rules of thumb (rough heuristics, not guarantees):

- If the population is symmetric: $n \geq 20$ usually fine
- If the population is moderately skewed: $n \geq 30$
- If the population is heavily skewed: $n \geq 50$ or more
- For proportions near 0 or 1: need larger n

A&M simulations show even $n = 100$ can give poor coverage for some distributions.

Why Does the CLT Work? (Intuition)

The magic of averaging:

- Each Y_i deviates from μ by some random amount
- When we average many independent deviations, extremes cancel out
- Positive and negative deviations offset each other
- What remains is tightly concentrated around μ

The shape: Why specifically *normal*?

- The normal is the unique distribution that is "stable" under averaging
- Average of normals is normal; average of anything converges to normal

The formal proof uses characteristic functions (see A&M for references).

Preview: Confidence Intervals

The CLT enables inference:

Since $\bar{Y} \approx N(\mu, \sigma^2/n)$:

$$\Pr\left(-1.96 < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \approx 0.95$$

Rearranging:

$$\Pr\left(\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

This is a 95% confidence interval!

$$\text{CI}: \quad \bar{Y} \pm 1.96 \times \text{SE}$$

We'll formalize this next week.

Slutsky's Theorem: Why Estimated SEs Work

Problem: The CI formula uses σ, but we don't know σ!

Solution: Estimate it with $\hat{\sigma}$. But why is this valid?

Slutsky's Theorem (A&M Theorem 3.2.25)

If $T_n \xrightarrow{d} T$ and $S_n \xrightarrow{p} c$, then:

- $T_n + S_n \xrightarrow{d} T + c$
- $T_n \cdot S_n \xrightarrow{d} c \cdot T$
- $T_n / S_n \xrightarrow{d} T/c \quad (\text{if } c \neq 0)$

Application: $\hat{\sigma} \xrightarrow{p} \sigma$ by LLN, so:

$$\frac{\bar{Y} - \mu}{\hat{\sigma}/\sqrt{n}} \xrightarrow{d} N(0, 1)$$

Replacing σ with $\hat{\sigma}$ doesn't change the asymptotic distribution

The Delta Method (Brief)

What if we care about $g(\mu)$, not just μ?

Delta Method

If $\sqrt{n}(\bar{Y} - \mu) \xrightarrow{d} N(0, \sigma^2)$ and g is differentiable at μ:

$$\sqrt{n}(g(\bar{Y}) - g(\mu)) \xrightarrow{d} N(0, [g'(\mu)]^2\sigma^2)$$

In practice:

$$g(\bar{Y}) \approx N\left(g(\mu), [g'(\mu)]^2\frac{\sigma^2}{n}\right)$$

Transformations of asymptotically normal estimators are also asymptotically normal.

Delta Method Example

Setup: Estimating the odds ratio.

Let p = probability of event, $\hat{p}$ = sample proportion.

We want to estimate the **log odds**: $\theta = \log\left(\frac{p}{1-p}\right)$

By CLT: $\sqrt{n}(\hat{p} - p) \xrightarrow{d} N(0, p(1-p))$

Let $g(p) = \log(p/(1-p))$. Then $g'(p) = \frac{1}{p(1-p)}$.

By Delta Method:

$$\sqrt{n}(g(\hat{p}) - g(p)) \xrightarrow{d} N\left(0, \frac{1}{p(1-p)}\right)$$

This gives us standard errors for log odds ratios in logistic regression.

CLT for Sums

Sometimes we work with sums, not averages:

Let $S_n = \sum_{i=1}^{n} Y_i$. Then:

- $\mathbb{E}[S_n] = n\mu$
- $\text{Var}(S_n) = n\sigma^2$
- $\text{SD}(S_n) = \sqrt{n}\sigma$

CLT for sums:

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} N(0, 1)$$

Or: $S_n \approx N(n\mu, n\sigma^2)$

This is just a rescaled version of the CLT for means.

Example: Polling Margin of Error

Setup: Poll of $n = 1{,}000$ voters. True support $p = 0.52$.

By CLT: $\hat{p} \approx N\left(0.52, \frac{0.52 \times 0.48}{1000}\right) = N(0.52, 0.00025)$

Standard error: $\text{SE} = \sqrt{0.00025} = 0.0158$

95% interval: $0.52 \pm 1.96 \times 0.0158 = [0.489, 0.551]$

Interpretation: 95% of polls would give $\hat{p}$ in this range.

The "margin of error" reported in polls is typically $1.96 \times \text{SE} \approx 3\%$.

Special Case: Normal Approximation to Binomial

If $X \sim \text{Binomial}(n, p)$, then $X = \sum_{i=1}^{n} Y_i$ where $Y_i \sim \text{Bernoulli}(p)$.

By CLT:

$$X \approx N(np, np(1-p))$$

for large n.

Rule of thumb: Approximation is good if $np \geq 5$ and $n(1-p) \geq 5$.

Example: Flip a fair coin 100 times. What's $\Pr(X \geq 60)$?
$X \approx N(50, 25)$, so:

$$\Pr(X \geq 60) \approx \Pr\left(Z \geq \frac{60-50}{5}\right) = \Pr(Z \geq 2) \approx 0.023$$

When the CLT Doesn't Apply

The CLT requires:

- I.I.D. observations
- Finite variance: $\sigma^2 < \infty$

The CLT fails if:

- **Not I.I.D.**: Time series, clustered data, dependent observations
- **Infinite variance**: Heavy-tailed distributions (Cauchy, Pareto with $\alpha \leq 2$)
- n **too small**: Approximation isn't accurate yet

Extensions exist: CLT variants for dependent data, bootstrap methods for small samples.

Blackwell's Take (Chapter 3)

From Blackwell:

> *"The CLT tells us that regression coefficients are approximately normally distributed in large samples. This is why we can construct confidence intervals and perform hypothesis tests using the normal distribution."*

The connection:

- OLS coefficients are (complicated) averages
- Averages are approximately normal by CLT
- Therefore, OLS coefficients are approximately normal
- This justifies t-tests and confidence intervals for regression

Key Takeaways

1. **The CLT**: Sample means are approximately normal for large n

$$\bar{Y} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

2. **Regardless** of the original distribution (as long as $\sigma^2 < \infty$)
3. **The standardized version**: $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \approx N(0, 1)$
4. **Delta method**: Transformations preserve asymptotic normality
5. **This enables inference**: Confidence intervals, hypothesis tests

Next week: Estimation—bias, variance, consistency, and confidence intervals.

Looking Ahead

Week 6: Estimation and Properties of Estimators

- Estimand vs. estimator vs. estimate
- Bias and variance
- Mean squared error = Bias^2 + Variance
- Consistency
- Confidence intervals (finally!)

Reading:

- A&M §3.2.3 and §3.3.1 (estimation, confidence intervals)
- Blackwell Ch. 2 (model-based inference)