Sampling and Central Limit Theorem

Lesson 09

Digital Finance

Outline

Why Sampling?

Population: Entire group of interest

- All stocks, all customers, all transactions
- Often too large to measure completely

Sample: Subset we actually observe

- Use sample to infer about population
- Key question: How reliable are our conclusions?

Statistics bridges the gap from sample to population.

Random Sampling

Simple random sample:

- Every member has equal chance of selection
- Selections are independent

IID assumption:

- $X_1, X_2, \ldots, X_n$ are independent
- All have same distribution
- "Independent and Identically Distributed"

IID is a strong assumption but enables powerful theory.

Sample Statistics

Sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Key insight: These are random variables!

- Different samples give different values
- Have their own distributions

Statistics vary from sample to sample.

Sampling Distribution of $\bar{X}$

Key concept: If you took many samples and computed each sample mean, those means form their own distribution – the "sampling distribution."

Properties of sample means:

- Center: The average of all possible sample means $= \mu$ (population mean)
- Spread: Sample means vary less than individual observations
- Formula: $\text{SE}(\bar{X}) = \sigma/\sqrt{n}$

Key insight: Double the sample size $\rightarrow$ reduce error by $\sqrt{2} \approx 1.4$

Larger samples give more precise estimates (less spread in sampling distribution).

Law of Large Numbers

As $n \to \infty$:

$$\bar{X}_n \to \mu \quad \text{(in probability)}$$

Meaning:

- Sample mean converges to population mean
- Larger samples are more accurate
- Foundation of Monte Carlo methods

Casino example: House always wins in the long run

Averages stabilize as we collect more data.

Central Limit Theorem

The most important theorem in statistics!
Plain English version:

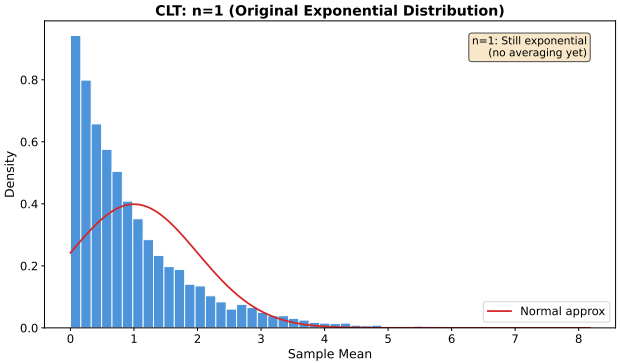
- Take a sample and compute the mean
- The distribution of sample means is approximately normal
- This works *no matter what shape* the original data has!

Why this matters:

- Skewed data, uniform data, any data – sample means become bell-shaped
- This is why the normal distribution appears everywhere
- Larger n = better approximation

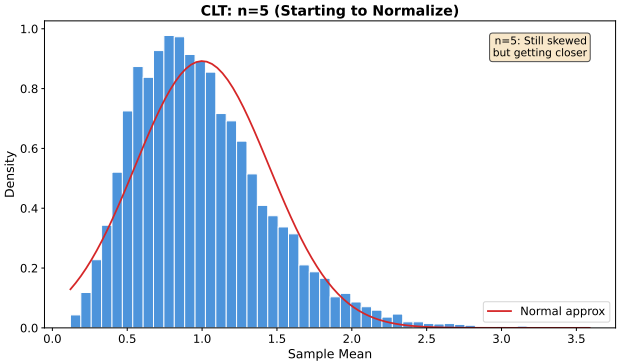
CLT lets us use normal-based methods even when data isn't normal.

CLT: n=1 (Original Distribution)



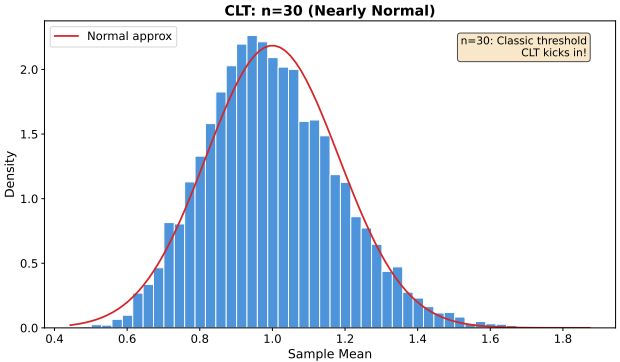
With n=1, we see the original exponential distribution.

CLT: n=5 (Starting to Normalize)



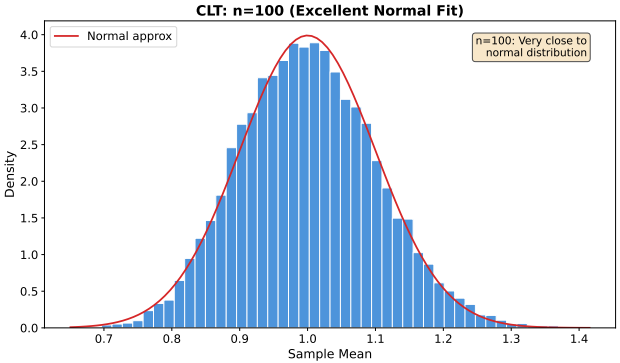
With n=5, sample means start approaching normality.

CLT: n=30 (Classic Threshold)



n=30: The classic CLT threshold. Sample means are nearly normal.

CLT: n=100 (Excellent Fit)



With larger n, the normal approximation becomes excellent.

When Does CLT Apply?

Rule of thumb: $n \geq 30$ usually sufficient (common heuristic)

Depends on original distribution:

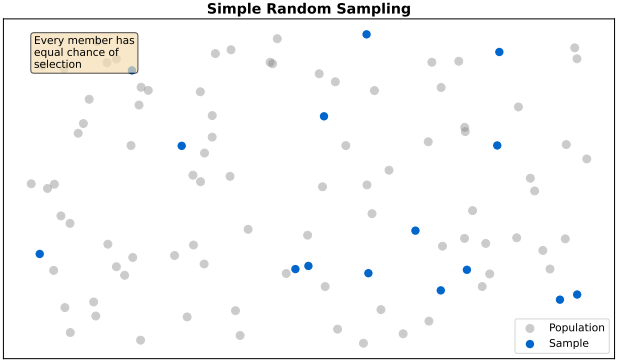
- Symmetric: CLT kicks in quickly
- Skewed: Need larger n
- Heavy tails: May need $n > 100$

Finance implication:

- Portfolio returns (many assets) tend toward normal
- Justifies many financial models

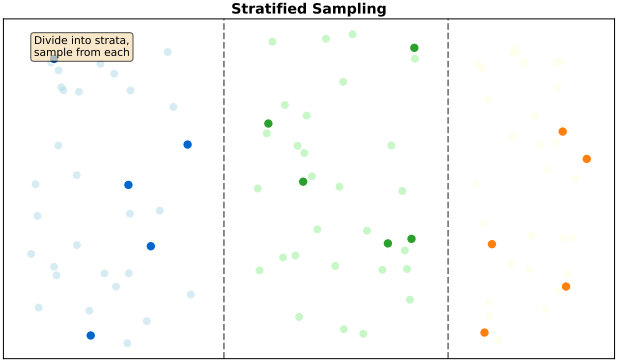
CLT is why the normal distribution is so prevalent.

Simple Random Sampling



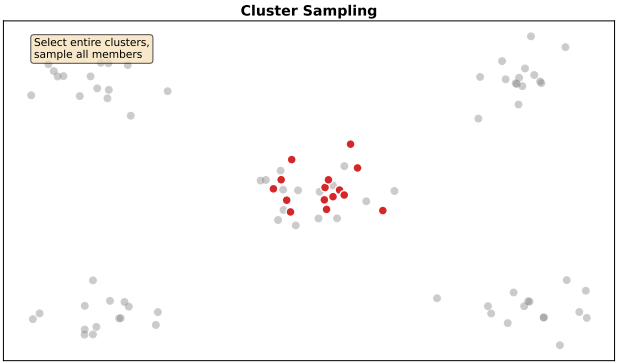
Every member has equal probability of being selected.

Stratified Sampling



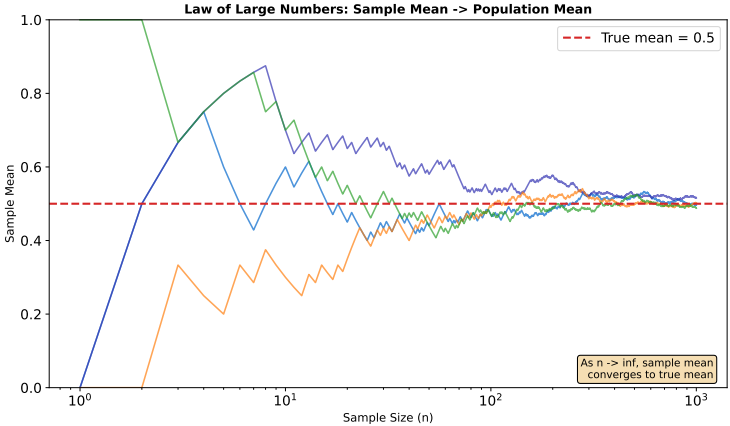
Divide into strata, then sample from each group proportionally.

Cluster Sampling



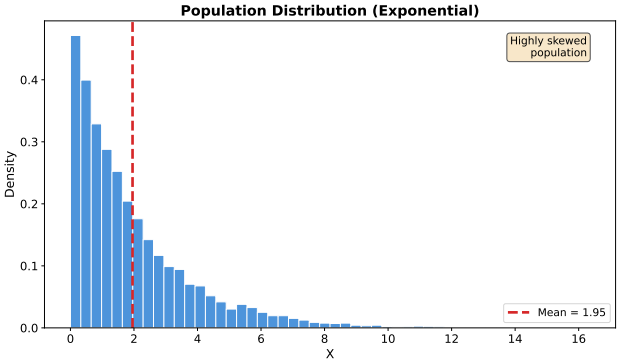
Select entire clusters, then measure all members within.

Law of Large Numbers



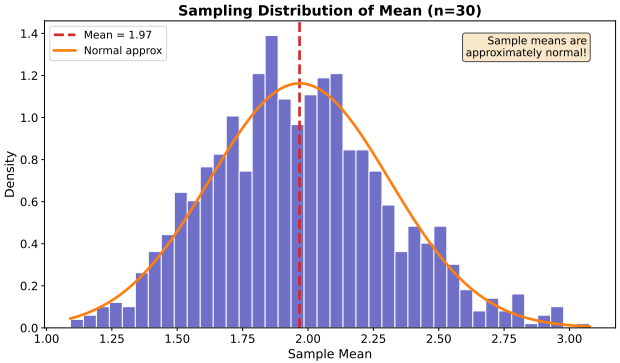
Sample mean converges to population mean.

Population Distribution



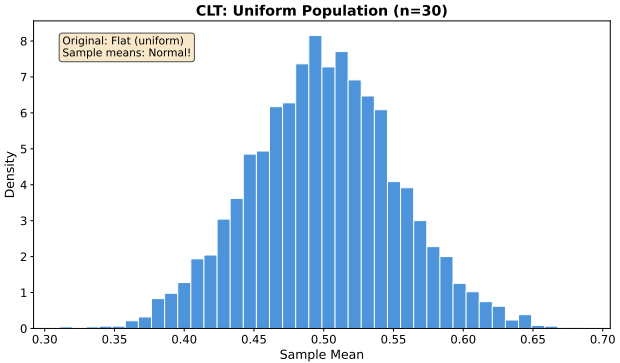
Population distribution (exponential): highly skewed.

Sampling Distribution of Mean



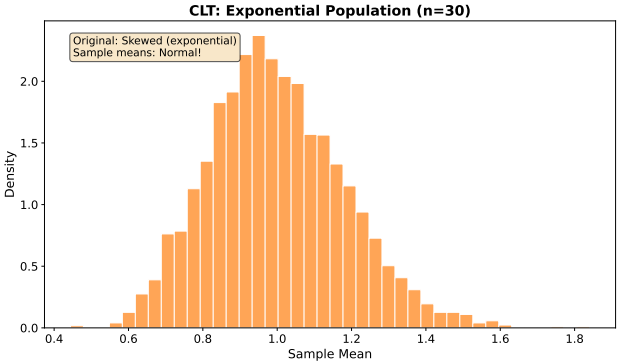
Sampling distribution of the mean (n=30): approximately normal!

CLT: Uniform Population



Original: flat (uniform). Sample means: normal!

CLT: Exponential Population



Original: skewed (exponential). Sample means: normal!

Summary

Sampling:

- Use samples to learn about populations
- Sample statistics are random variables

Law of Large Numbers:

- $\bar{X} \to \mu$ as $n \to \infty$
- Guarantees convergence

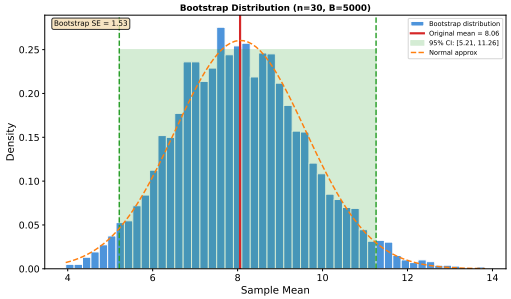
Central Limit Theorem:

- $\bar{X}$ is approximately normal for large n
- Foundation for inference

Next lesson: Point Estimation

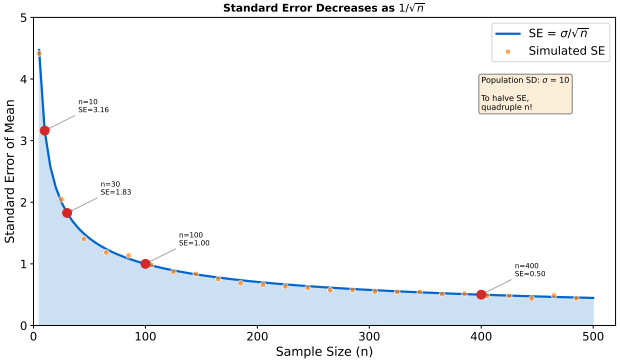
Bootstrap Distribution

Bootstrap idea: Resample from your data (with replacement) many times to estimate the sampling distribution.



Bootstrap: pretend your sample IS the population, resample from it to approximate variability.

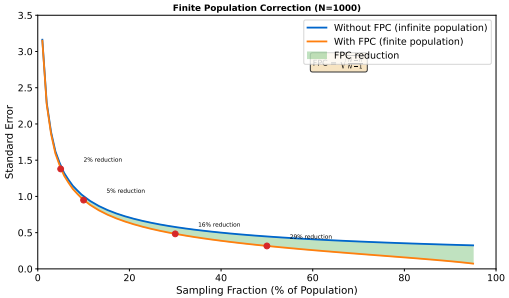
Standard Error vs Sample Size



Standard error decreases proportional to 1/sqrt(n).

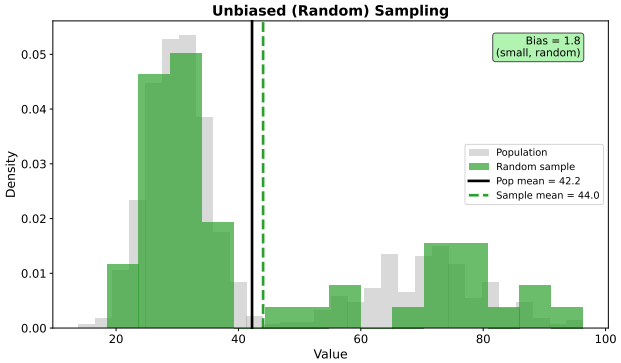
Finite Population Correction

When it matters: If you sample a large fraction of the population (e.g., 10%+), standard SE formulas overstate uncertainty.



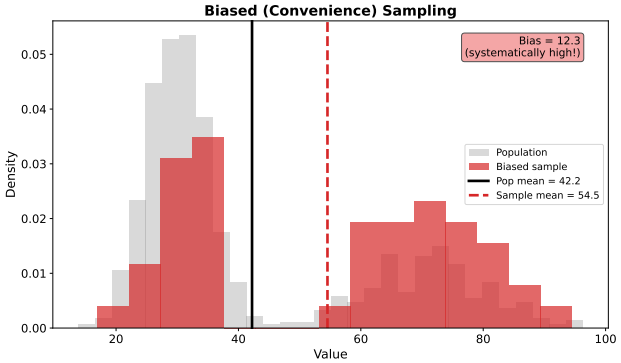
FPC factor: $\sqrt{(N-n)/(N-1)}$. **Reduces SE when sample is large relative to population.**

Unbiased (Random) Sampling



Random sampling: sample mean is close to population mean.

Biased (Convenience) Sampling



Biased sampling: systematically overestimates the population mean.