Large Sample Theory: WLLN and CLT

Zeyang (Arthur) Yu

Princeton University

October 31, 2024

Outline

Outline

Large Sample Approximation: Motivation

Recall: three canonical problems in mathematical statistics

- Sample: $\{X_i\}_{i=1}^n$ distributed according to P (population *dist.*)
- "Learn" some "features" of P (e.g., a *param.* $\theta(P)$) from the data
  - E.g., $\theta(P) = E(X_i)$
- Provides a "best guess" $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ for $\theta(P)$
- Test a hypothesis about $\theta(P)$
  - E.g., $\theta(P) = E(X_i) = 100$?
- Construct a confidence region for $\theta(P)$

To Solve the Problems, Collect Plenty of Data

1. Weak Law of Large Numbers
   - A tool for finding a good guess for $\theta(P)$
2. Central Limit Theorem
   - A tool for hypothesis testing and creating a confidence region

Outline

Simulations in R

Probability simulations in R: motivation

- Simulations in R is a powerful tool to understand theorems
  - Visualize and calculate under the assumptions in theorems
- Simulations might help us develop new statistical methodologies
  - Simulations might give us some intuitions on how things work

Probability simulations in R

- It can be used to simulate theorem without having real data
- R can generate random numbers that follow certain *dist.*
  - E.g., Bernoulli, Poisson, normal, etc.
- Make up “fake” data set that satisfies assumptions in theorem
  - E.g., if we assume $X_i \sim N(0, 1)$, R can give data follows this *dist.*
  - E.g., if we assume $E(X_i)$ exists, R has plenty of options

Outline

Large Sample Theory: Weak Law of Large Numbers

Weak Law of Large Numbers

Let $\{X_i\}_{i=1}^n$ be a sequence of *iid* random variables on $\mathbb{R}$ with distribution P. Suppose that $E(X_i)$ exists, then:

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} E(X_i).$$

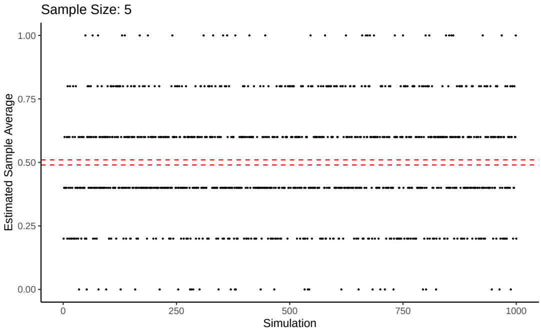
Weak Law of Large Numbers: Remarks

- $\{X_i\}_{i=1}^n$ being *iid*: $P(X_1, \ldots, X_n) = P(X_1) \times \ldots P(X_n) = P^n$
- *iid*: independent and identically distributed

  Across i, there is independence (SRS can give you this)

  All i is from the same population gives identically distributed
- $\xrightarrow{P}$: converge in probability
- Sample size $\uparrow$, sample average being close to $E(X_i)$ *w.* $\uparrow$ *prob.*

Weak Law of Large Numbers: Simulations

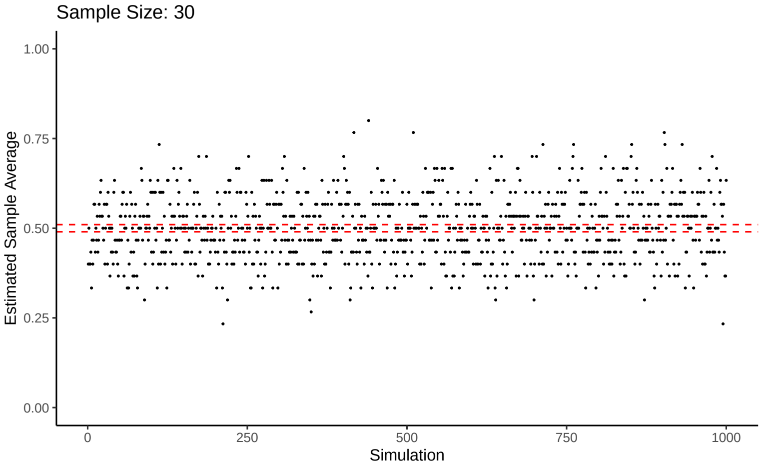


Simulation Setup

- $\{X_i\}_{i=1}^5$ is an *iid* sample from Ber(0.5)

- Each dot: $\frac{\sum_{i=1}^n X_i}{5}$
- 1000 dots: 1000 simulations (draw $\{X_i\}_{i=1}^5$ 1000 times)

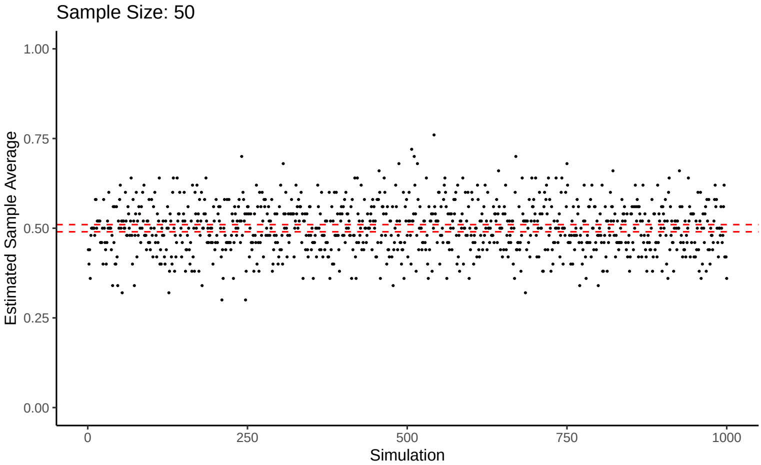
Weak Law of Large Numbers: Simulations (Conti.)



Simulation Setup

- $\{X_i\}_{i=1}^{30}$ is an *iid* sample from Ber(0.5)

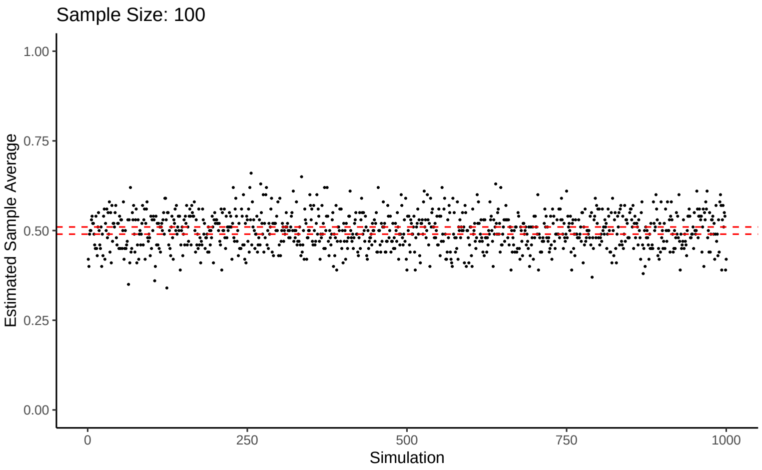
Weak Law of Large Numbers: Simulations (Conti.)



Simulation Setup

- $\{X_i\}_{i=1}^{50}$ is an *iid* sample from Ber(0.5)

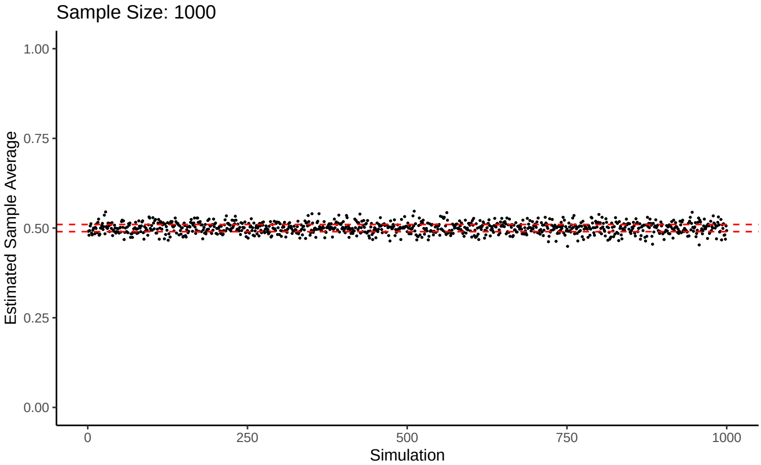
Weak Law of Large Numbers: Simulations (Conti.)



Simulation Setup

- $\{X_i\}_{i=1}^{100}$ is an *iid* sample from Ber(0.5)

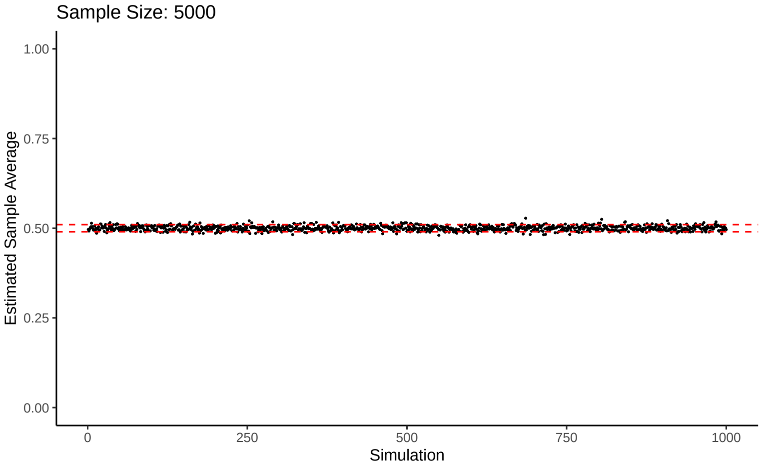
Weak Law of Large Numbers: Simulations (Conti.)



Simulation Setup

- $\{X_i\}_{i=1}^{1000}$ is an *iid* sample from Ber(0.5)

Weak Law of Large Numbers: Simulations (Conti.)



Simulation Setup

- $\{X_i\}_{i=1}^{5000}$ is an *iid* sample from Ber(0.5)

Outline

Large Sample Theory: Central Limit Theorem

Central Limit Theorem

Let $\{X_i\}_{i=1}^{n}$ be a sequence of *iid* random variables on $\mathbb{R}$ with distribution P. Suppose that $\sigma^2(P)$ exists, then:

$$\sqrt{n}\left(\overline{X}_n - E(X_i)\right) \xrightarrow{\mathcal{D}} N\left(0, \sigma^2(P)\right).$$

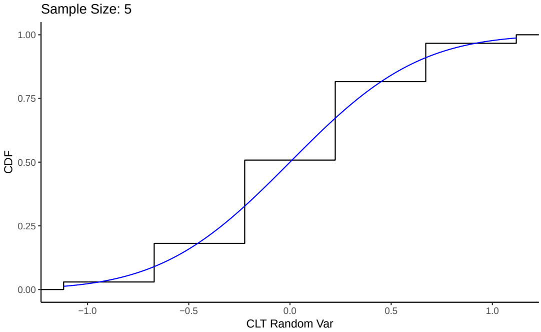
Central Limit Theorem: Remarks

- Recall that $\sigma^2(P)$ is the variance of the random variable:

$$\sigma^2(P) = E\left((X_i - E(X_i))^2\right) = E(X_i^2) - E(X_i)^2$$

  - Notation (P): emphasize that the distribution is P
- $\xrightarrow{\mathcal{D}}$: converge in distribution (subtle point: convergence of CDFs)
- Sample size $\uparrow$, *dist.* of $\sqrt{n}\left(\overline{X}_n - E(X_i)\right)$ goes closer to $N\left(0, \sigma^2(P)\right)$

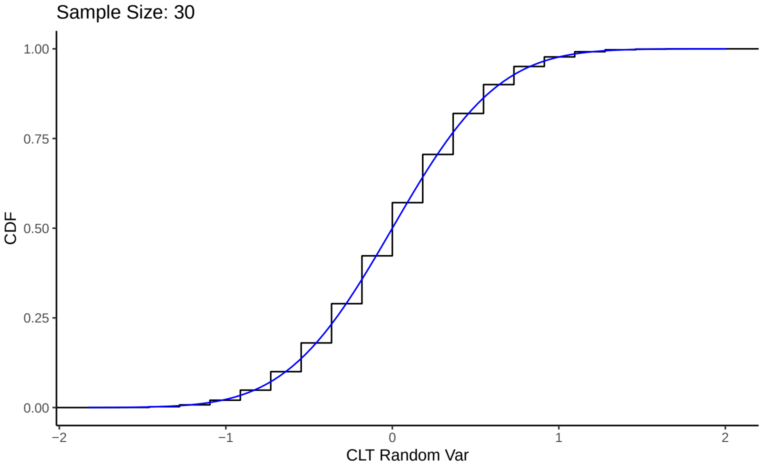
Central Limit Theorem: Simulations



Simulation Setup

- $\{X_i\}_{i=1}^5$ is an *iid* sample from Ber(0.5)
- Therefore, $E(X_i) = 0.5$, $\sigma^2(P) = 0.25$
- 10000 simulations for: calculating $\sqrt{n}\left(\overline{X}_n - E(X_i)\right)$ from $\{X_i\}_{i=1}^5$

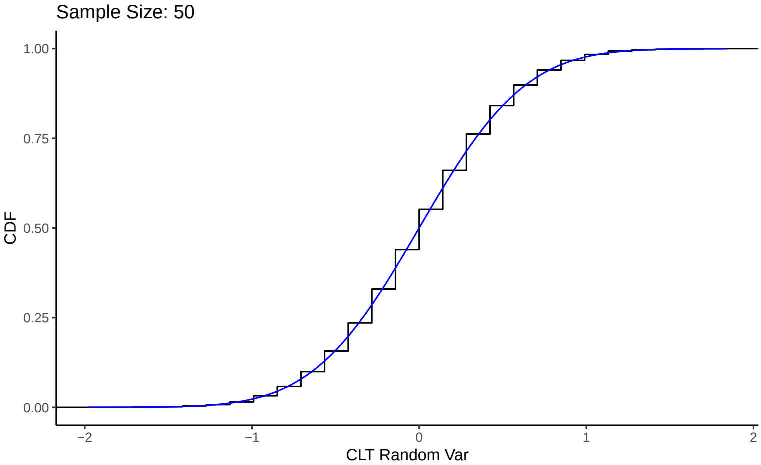
Central Limit Theorem: Simulations (Conti.)



Simulation Setup

- $\{X_i\}_{i=1}^{30}$ is an *iid* sample from Ber(0.5)

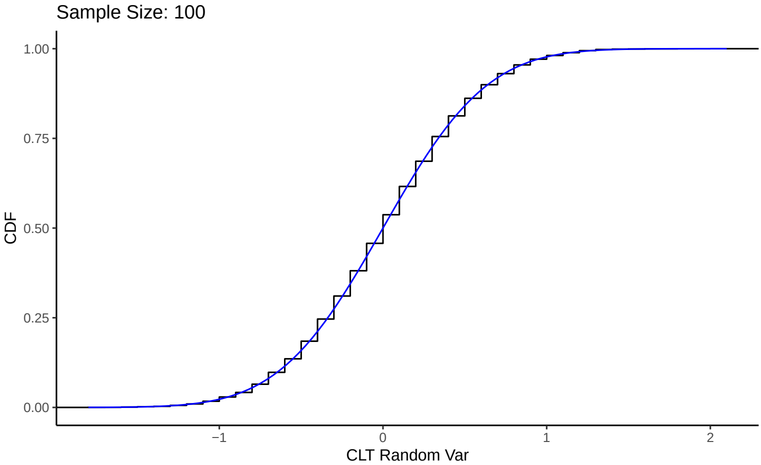
Central Limit Theorem: Simulations (Conti.)



Simulation Setup

- $\{X_i\}_{i=1}^{50}$ is an *iid* sample from Ber(0.5)

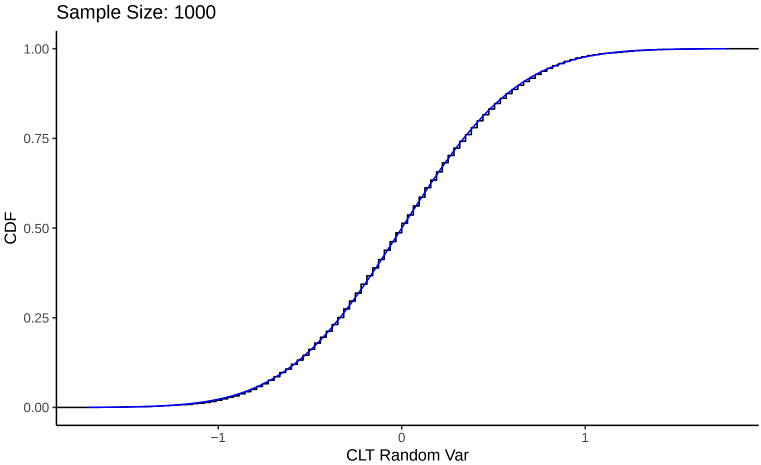
Central Limit Theorem: Simulations (Conti.)



Simulation Setup

- $\{X_i\}_{i=1}^{100}$ is an *iid* sample from Ber(0.5)

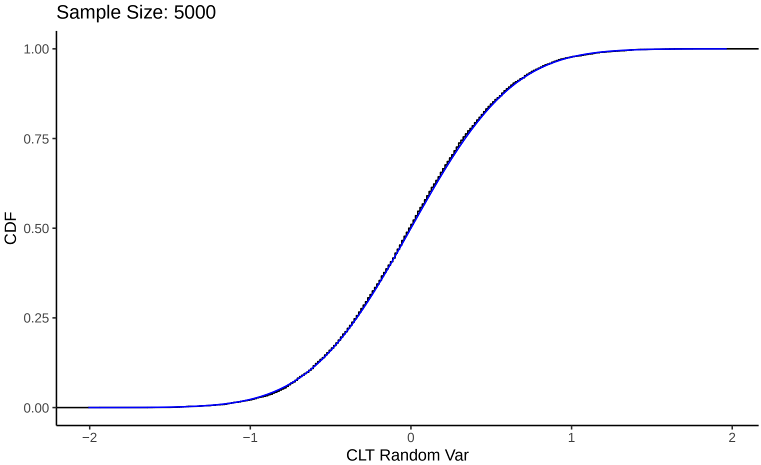
Central Limit Theorem: Simulations (Conti.)



Simulation Setup

- $\{X_i\}_{i=1}^{1000}$ is an *iid* sample from Ber(0.5)

Central Limit Theorem: Simulations (Conti.)



Simulation Setup

- $\{X_i\}_{i=1}^{5000}$ is an *iid* sample from Ber(0.5)

Large Sample Theory: Central Limit Theorem (Conti.)

Two Other Forms of the Central Limit Theorem

1. Rescale by $\sigma(P)$

$$\sqrt{n}\frac{\left(\overline{X}_n - E(X_i)\right)}{\sigma(P)} \xrightarrow{\mathcal{D}} N(0,1)$$

   - $N(0,1)$: standard normal distribution

2. Rescale by estimated $\sigma(P)$

$$\sqrt{n}\frac{\left(\overline{X}_n - E(X_i)\right)}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2} \xrightarrow{\mathcal{D}} N(0,1)$$

   - An estimated $\sigma(P) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$

Learning Goals: WLLN and CLT

Students will be able to:

- Understand R can simulate *r.v.* that follow some distributions
- Understand the statement of the WLLN
  - Know the assumptions in WLLN
  - Explain the WLLN in plain language
- Understand the statement of the CLT
  - Know the assumptions in CLT
  - Explain the CLT in plain language
- Simulate WLLN in R
- Simulate CLT in R
  - Simulate three versions of CLT in R