Chapter 3: Asymptotic Statistics

Jonathan Roth

Mathematical Econometrics I
Brown University
Fall 2023

Outline

1. Overview

2. LLN, CLT, and CMT

3. Putting Asymptotics into Practice

Motivation

- We've seen how we can test hypotheses about population means using information from the sample mean $\hat{\mu}$ when it is **normally distributed** with a known variance

- This situation arises when we know that $Y_i \sim \mathrm{N}(\mu, \sigma^2)$ with known σ

- But this situation is rare... how do we "do inference" more generally?

- Fortunately, the assumption of normally distributed sample means turns out to be a good **approximation** when samples are large

- What we mean by a "good approximation" is formalized by asymptotic statistics, which considers the distribution of $\hat{\mu}$ in the limit as $N \to \infty$

Overview of Important Results

- The **Law of Large Numbers** (LLN) says that when N is large, $\hat{\mu}$ is close to μ with very high probability

- The **Central Limit Theorem** (CLT) says that when N is large, the distribution of $\hat{\mu}$ is approximately normally distributed with mean μ and variance σ^2/n

- The **Continuous Mapping Theorem** says that when N is large, continuous functions of $\hat{\mu}$, say $g(\hat{\mu})$, are also close to $g(\mu)$

Outline

1. Overview ✓

2. LLN, CLT, and CMT

3. Putting Asymptotics into Practice

Convergence in Probability

- Intuitively, a random variable X_N **converges in probability** to x if the probability that X_N is "close to" x is almost 1 when N is large

- Formally, we say X_N converges in probability to x, $X_n \to_p x$ or $plim X_n = x$, if for all $\varepsilon > 0$,

$$P(|X_N - x| > \varepsilon) \to 0$$

- If $X_n \to_p x$ for a constant x, we say X_n is *consistent* for x

- Typically x is a constant, although we will sometimes also say $X_N \to X$ for X a random variable (using the same definition as above)

Convergence in Probability (Cont.)

- Useful fact: if $E[(X_N - x)^2] \to 0$, then $X_N \to_p x$
- **Proof** (you won't be responsible for this):
  By the law of iterated expectations,

$$\begin{aligned} E[(X_N - x)^2] =& P(|X_n - x| > \varepsilon) E[(X_N - x)^2 || X_n - x| > \varepsilon] + \\ & P(|X_n - x| \leq \varepsilon) E[(X_N - x)^2 || X_n - x| \leq \varepsilon] \\ \geq & P(|X_n - x| > \varepsilon)\varepsilon^2 + 0 \end{aligned}$$

  This implies that

$$P(|X_N - x| > \varepsilon) \leq E[(X_N - x)^2]/\varepsilon^2 \text{ (Chebychev's Inequality)}$$

  Hence, $E[(X_N - x)^2] \to 0$ implies $P(|X_N - x| > \varepsilon) \to 0$

Law of Large Numbers

- **Law of Large Numbers**. Suppose that $Y_1, ..., Y_N$ are drawn *iid* from a distribution with $Var(Y_i) = \sigma^2 < \infty$. Then

$$\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i \rightarrow_p \mu = E[Y_i]$$

- In words: as the sample gets large, the sample mean will be close to the population mean with high probability.

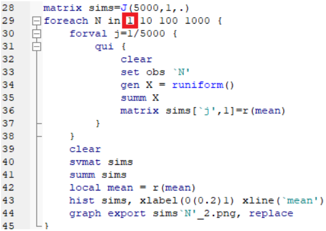
- **Proof:** We saw last chapter that $E[\hat{\mu}_N] = \mu$ and $Var(\hat{\mu}_N) = \sigma^2/N$. Thus,

$$Var(\hat{\mu}_N) = E[(\hat{\mu}_N - \mu)^2] = \sigma^2/N \rightarrow 0$$

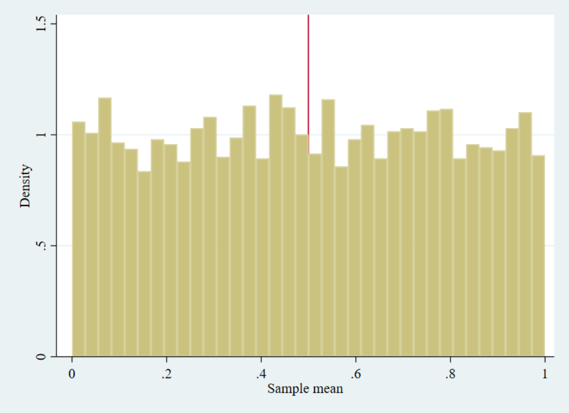
  Hence, $\hat{\mu}_N \rightarrow_p \mu$ by our "useful fact".

Laws of Large Numbers Illustration

Distribution and mean of $\frac{1}{N}\sum_i Z_i$ when $Z_i \sim \mathrm{U}(0,1)$, $\mathbf{N=1}$

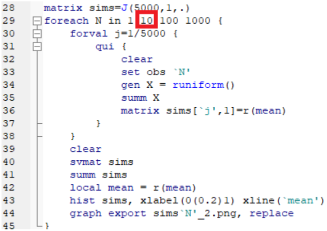


```
matrix sims=J(5000,1,.)
foreach N in 1 10 100 1000 {
    forval j=1/5000 {
        qui {
            clear
            set obs `N'
            gen X = runiform()
            summ X
            matrix sims[`j',1]=r(mean)
        }
    }
    clear
    svmat sims
    summ sims
    local mean = r(mean)
    hist sims, xlabel(0(0.2)1) xline(`mean')
    graph export sims`N'_2.png, replace
}
```

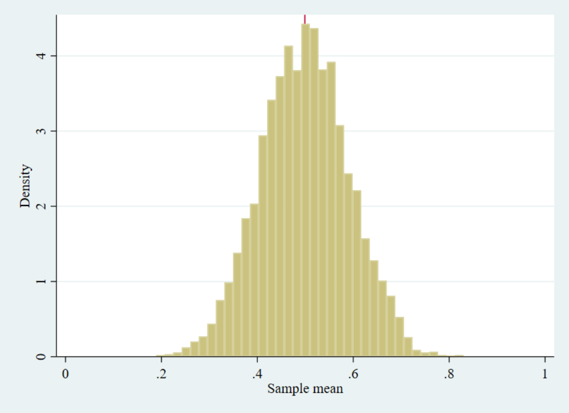


Laws of Large Numbers Illustration

Distribution and mean of $\frac{1}{N}\sum_i Z_i$ when $Z_i \sim \mathrm{U}(0,1)$, $\mathbf{N = 10}$

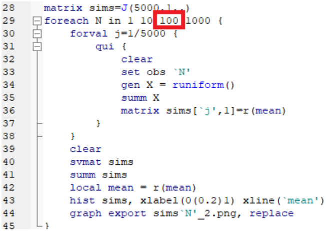


```
matrix sims=J(5000,1,.)
foreach N in 1 10 100 1000 {
    forval j=1/5000 {
        qui {
            clear
            set obs `N'
            gen X = runiform()
            summ X
            matrix sims[`j',1]=r(mean)
        }
    }
    clear
    svmat sims
    summ sims
    local mean = r(mean)
    hist sims, xlabel(0(0.2)1) xline(`mean')
    graph export sims`N'_2.png, replace
}
```

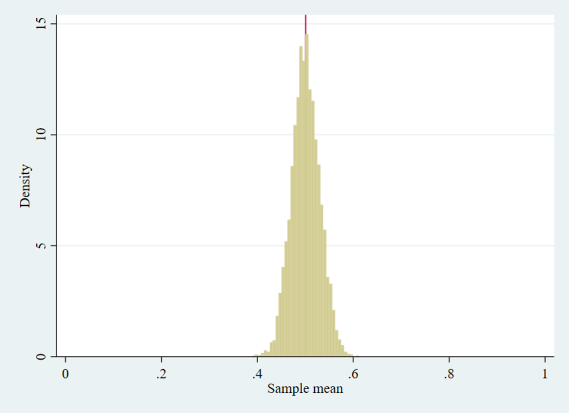


Laws of Large Numbers Illustration

Distribution and mean of $\frac{1}{N}\sum_i Z_i$ when $Z_i \sim \mathrm{U}(0,1)$, $\mathbf{N = 100}$

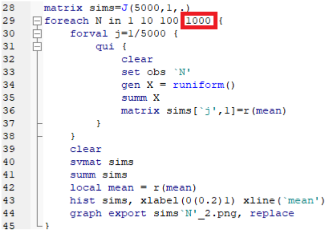


```
matrix sims=J(5000,1,.)
foreach N in 1 10 100 1000 {
    forval j=1/5000 {
        qui {
            clear
            set obs `N'
            gen X = runiform()
            summ X
            matrix sims[`j',1]=r(mean)
        }
    }
    clear
    svmat sims
    summ sims
    local mean = r(mean)
    hist sims, xlabel(0(0.2)1) xline(`mean')
    graph export sims`N'_2.png, replace
}
```

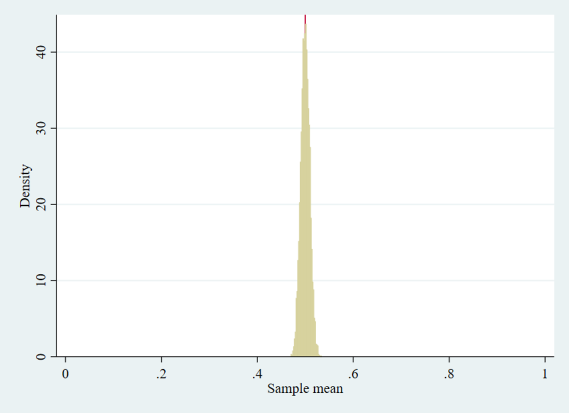


Laws of Large Numbers Illustration

Distribution and mean of $\frac{1}{N}\sum_i Z_i$ when $Z_i \sim \mathrm{U}(0,1)$, $\mathbf{N = 1000}$



```
matrix sims=J(5000,1,.)
foreach N in 1 10 100 1000 {
    forval j=1/5000 {
        qui {
            clear
            set obs `N'
            gen X = runiform()
            summ X
            matrix sims[`j',1]=r(mean)
        }
    }
    clear
    svmat sims
    summ sims
    local mean = r(mean)
    hist sims, xlabel(0(0.2)1) xline(`mean')
    graph export sims`N'_2.png, replace
}
```



Convergence in Distribution

- You might have noticed that the distribution of $\hat{\mu}$ in the simulations looks close to a normal distribution as N gets large

- The notion of **convergence in distribution** formalizes what it means for one distribution to be close to another distribution

- Definition: We say that X_N converges in distribution to a continuously distributed variable X, denoted $X_n \rightarrow_d X$ or $X_n \Rightarrow X$, if the CDF of X_N converges (pointwise) to the CDF of X,

$$F_{X_N}(x) \rightarrow F_X(x) \text{ for all } x$$

Central Limit Theorem

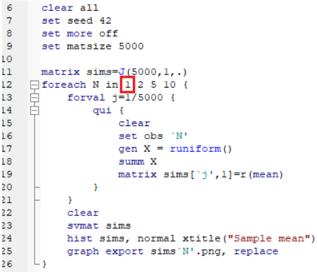
- **The Central Limit Theorem (CLT)** formalizes the sense in which sample means are approximately normally distributed in large samples
- Theorem: Suppose that $Y_1, ..., Y_N$ are drawn *iid* from a distribution with mean $\mu = E[Y_i]$ and variance $Var(Y_i) = \sigma^2 < \infty$. Then the sample mean $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ satisfies

$$\sqrt{N}(\hat{\mu} - \mu) \rightarrow_d N(0, \sigma^2)$$

- In words, the theorem says the following:
  1. We can start with any distribution Y_i, possibly non-normal
  2. If we take the average of the $Y_1, ..., Y_N$ in a sample sufficiently large, the distribution of $\hat{\mu} = \frac{1}{N}\sum_i Y_i$ is (approximately) normal!

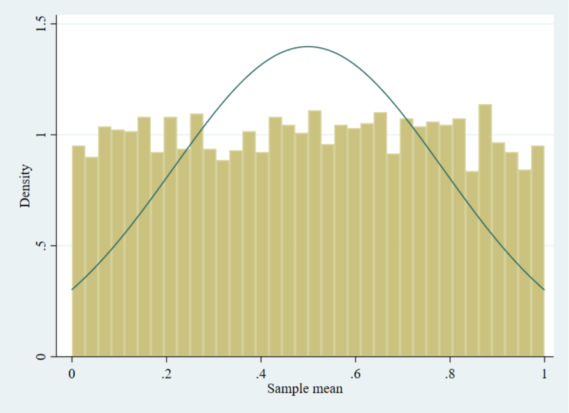
CLT Illustration

Distributions of $\hat{\mu} = \frac{1}{N}\sum_i X_i$ vs. $\mathrm{N}(E[\hat{\mu}], Var(\hat{\mu}))$: $X_i \sim \mathrm{U}(0,1)$, $\mathbf{N = 1}$



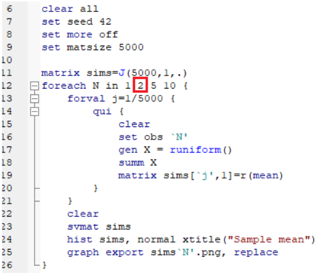
```
clear all
set seed 42
set more off
set matsize 5000

matrix sims=J(5000,1,.)
foreach N in 1 2 5 10 {
    forval j=1/5000 {
        qui {
            clear
            set obs `N'
            gen X = runiform()
            summ X
            matrix sims[`j',1]=r(mean)
        }
    }
    clear
    svmat sims
    hist sims, normal xtitle("Sample mean")
    graph export sims`N'.png, replace
}
```



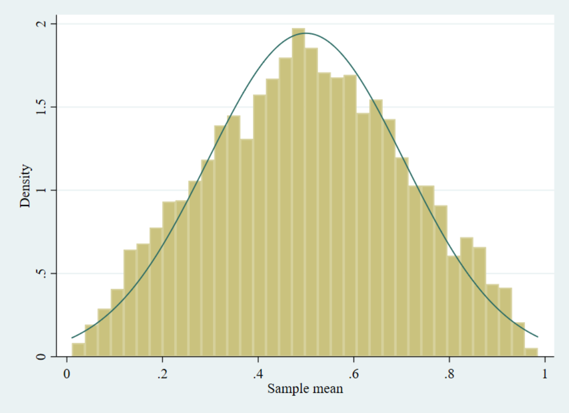
CLT Illustration

Distributions of $\hat{\mu} = \frac{1}{N}\sum_i X_i$ vs. $\mathrm{N}(E[\hat{\mu}], Var(\hat{\mu}))$: $X_i \sim \mathrm{U}(0,1)$, $\mathbf{N = 2}$



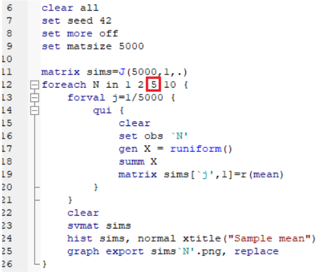
```
clear all
set seed 42
set more off
set matsize 5000

matrix sims=J(5000,1,.)
foreach N in 1 2 5 10 {
    forval j=1/5000 {
        qui {
            clear
            set obs `N'
            gen X = runiform()
            summ X
            matrix sims[`j',1]=r(mean)
        }
    }
    clear
    svmat sims
    hist sims, normal xtitle("Sample mean")
    graph export sims`N'.png, replace
}
```



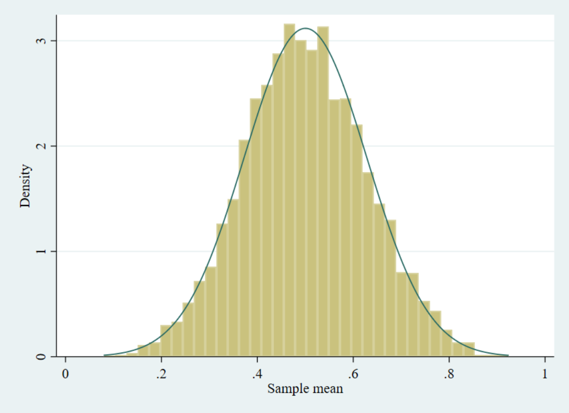
CLT Illustration

Distributions of $\hat{\mu} = \frac{1}{N}\sum_i X_i$ vs. $\mathrm{N}(E[\hat{\mu}], Var(\hat{\mu}))$: $X_i \sim \mathrm{U}(0,1)$, $\mathbf{N = 5}$



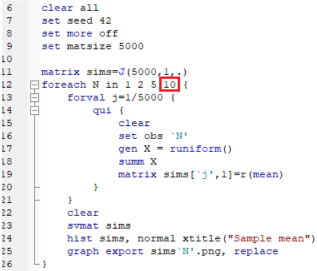
```
clear all
set seed 42
set more off
set matsize 5000

matrix sims=J(5000,1,.)
foreach N in 1 2 5 10 {
    forval j=1/5000 {
        qui {
            clear
            set obs `N'
            gen X = runiform()
            summ X
            matrix sims[`j',1]=r(mean)
        }
    }
    clear
    svmat sims
    hist sims, normal xtitle("Sample mean")
    graph export sims`N'.png, replace
}
```



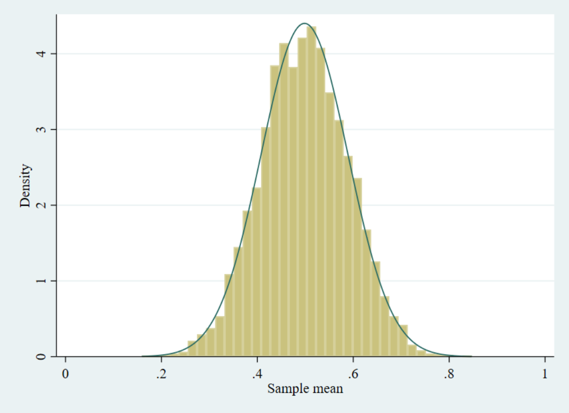
CLT Illustration

Distributions of $\hat{\mu} = \frac{1}{N}\sum_i X_i$ vs. $\mathrm{N}(E[\hat{\mu}], Var(\hat{\mu}))$: $X_i \sim \mathrm{U}(0,1)$, $\mathbf{N} = \mathbf{10}$

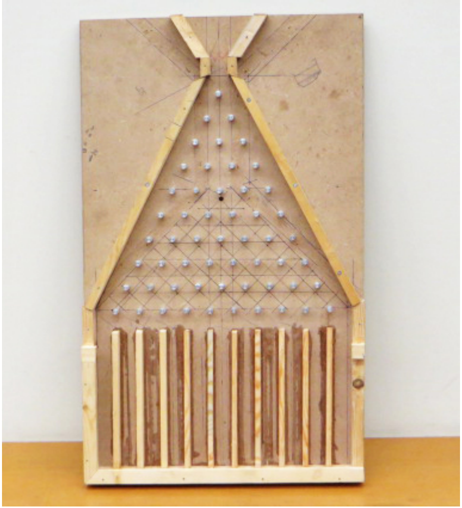


```
clear all
set seed 42
set more off
set matsize 5000

matrix sims=J(5000,1,.)
foreach N in 1 2 5 10 {
    forval j=1/5000 {
        qui {
            clear
            set obs `N'
            gen X = runiform()
            summ X
            matrix sims[`j',1]=r(mean)
        }
    }
    clear
    svmat sims
    hist sims, normal xtitle("Sample mean")
    graph export sims`N'.png, replace
}
```



CLT Illustration II



https://www.youtube.com/watch?v=EvHiee7gs9Y

Multivariate Versions

- The results we've discussed extend naturally to the multivariate case
- For a vector $\mathbf{X_N} \in \mathbb{R}^k$, we say $\mathbf{X_N} \rightarrow_p \mathbf{x}$ if each component of $\mathbf{X_N}$ converges in probability to each component of $\mathbf{x}$.
- **LLN**: For $\hat{\boldsymbol{\mu}}_N$, the sample mean of *iid* vectors $\mathbf{Y_1}, ... \mathbf{Y_N}$ with mean $\boldsymbol{\mu}$ and finite variance, $\hat{\boldsymbol{\mu}}_N \rightarrow_p \boldsymbol{\mu}$
- For a vector $\mathbf{X_N} \in \mathbb{R}^k$, we say $\mathbf{X_N} \rightarrow_d \mathbf{X}$ for $\mathbf{X}$ continuously distributed if $F_{\mathbf{X_N}}(\mathbf{x}) \rightarrow F_{\mathbf{X}}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^k$.
- **CLT**: For $\hat{\boldsymbol{\mu}}_N$, the sample mean of *iid* vectors $\mathbf{Y_1}, ... \mathbf{Y_N}$ with mean $\boldsymbol{\mu}$ and finite variance $\boldsymbol{\Sigma}$, $\sqrt{N}(\hat{\boldsymbol{\mu}}_N - \boldsymbol{\mu}) \rightarrow_d \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$

Continuous Mapping Theorem

- Sometimes we are interested in functions of sample means (e.g., the t-statistic is a function of $\hat{\mu}$ and σ).

- The **continuous mapping theorem** (CMT) tells us about continuous functions of random variables that converge in distribution/probability

- Theorem: suppose $g(\cdot)$ is a continuous function

  If $X_N \rightarrow_p X$, then $g(X_N) \rightarrow_p g(X)$

  If $X_N \rightarrow_d X$, then $g(X_N) \rightarrow_d g(X)$

  Multivariate versions here too: If $\mathbf{X_N} \rightarrow_p \mathbf{X}$, then $g(\mathbf{X_N}) \rightarrow_p g(\mathbf{X})$ and if $\mathbf{X_N} \rightarrow_d \mathbf{X}$, then $g(\mathbf{X_N}) \rightarrow_d g(\mathbf{X})$

Convergence of Sample Variance

- One useful application of the CMT is to show convergence in probability of the sample variance

- Let $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{\mu})^2$ be the sample variance of Y_i.

- Claim: if $Y_1, ..., Y_N$ are *iid* and $Var(Y_i^2)$ is finite, then $\hat{\sigma}^2 \rightarrow_p \sigma^2 = Var(Y_i)$.

- Proof:
  We can write the sample variance as $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \hat{\mu}^2$.
  First term: by the LLN, $\frac{1}{N}\sum_{i=1}^{N} Y_i^2 \rightarrow_p E[Y_i^2]$.
  Second term: by the LLN, $\hat{\mu} \rightarrow_p \mu = E[Y_i]$. Thus, by the CMT, $\hat{\mu}^2 \rightarrow_p E[Y_i]^2$.
  Thus, by the CMT again, $\frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \hat{\mu}^2 \rightarrow_p E[Y_i^2] - E[Y_i]^2 = \sigma^2$.

Slutsky's Lemma

- **Slutsky's lemma** (sometimes Slutsky's theorem) summarizes a few special cases of the CMT that are very useful.
- Suppose that $X_N \to_p c$ for a constant c, and $Y_N \to_d Y$. Then:
- $X_N + Y_N \to_d c + Y$.
- $X_n Y_n \to_d cY$.
- If $c \neq 0$, then $Y_n / X_n \to_d Y/c$.
- Analogous versions apply for vector-valued random variables.

Asymptotic Hypothesis Testing

- Recall that when $Y_i \sim \mathrm{N}(\mu, \sigma^2)$, we showed that the t-statistic $\hat{t} = \frac{\hat{\mu}-\mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$ under $H_0 : \mu = \mu_0$.

- Thus, when $Y_i \sim \mathrm{N}(\mu, \sigma^2)$, we had that $Pr(|\hat{t}| > 1.96) = 0.05$ under the null.

- Now, suppose that Y_i is not normally distributed and we don't know its variance.

- By CLT, $\sqrt{N}(\hat{\mu} - \mu_0) \rightarrow_d N(0, \sigma^2)$.
  By CMT and LLN (as shown above), $\hat{\sigma} \rightarrow_p \sigma$.

- Thus, by Slutsky's lemma, $\hat{t} = \dfrac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}} \rightarrow_d N(0,1)$.

- Hence, asymptotically $Pr(|\hat{t}| > 1.96) \rightarrow 0.05$, even though Y_i is not normal and $\hat{\sigma}$ is estimated! We can hypothesis test just like before.

Asymptotic Confidence Intervals

- Similarly, when Y_i was normal w/ σ known, we showed the confidence interval $\hat{\mu} \pm 1.96\sigma/\sqrt{N}$ contained the true μ 95% of the time

- Analogously, when Y_i is non-normal with unknown variance, $\hat{\mu} \pm 1.96\hat{\sigma}/\sqrt{N}$ contains the true μ with probability approaching 95% as N grows large.

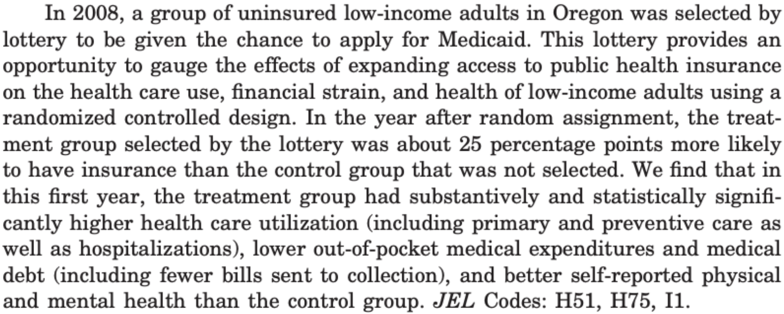
Outline

1. Overview ✓

2. LLN, CLT, and CMT ✓

3. Putting Asymptotics into Practice

Example – Oregon Health Insurance Experiment



In 2008, a group of uninsured low-income adults in Oregon was selected by lottery to be given the chance to apply for Medicaid. This lottery provides an opportunity to gauge the effects of expanding access to public health insurance on the health care use, financial strain, and health of low-income adults using a randomized controlled design. In the year after random assignment, the treatment group selected by the lottery was about 25 percentage points more likely to have insurance than the control group that was not selected. We find that in this first year, the treatment group had substantively and statistically significantly higher health care utilization (including primary and preventive care as well as hospitalizations), lower out-of-pocket medical expenditures and medical debt (including fewer bills sent to collection), and better self-reported physical and mental health than the control group. *JEL* Codes: H51, H75, I1.

Sample Means for Depression Outcome

	Control Group	Treated Group
Mean	0.329	0.306
SD	0.470	0.461
N	10426	13315

- Say we want a CI for the population mean in the control group
- We have

$$\hat{\mu} \pm 1.96 \times \hat{\sigma}/\sqrt{N} = 0.329 \pm 1.96 \times 0.470/\sqrt{10426} = [0.319, 0.338]$$

- What about for the treated group?

$$\hat{\mu} \pm 1.96 \times \hat{\sigma}/\sqrt{N} = 0.306 \pm 1.96 \times 0.461/\sqrt{13315} = [0.298, 0.313]$$

Hypothesis Testing for Experiments

- Suppose we have an experiment so that $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$. We showed previously that the average treatment effect is given by

$$\tau = E[Y_i(1) - Y_i(0)] = E[Y_i | D_i = 1] - E[Y_i | D_i = 0].$$

- Suppose we have an *iid* sample of (Y_i, D_i). How can we test the null $H_0 : \tau = \tau_0$?

- Let $\bar{Y}_1 = \frac{1}{N_1} \sum_{i:D_i=1} Y_i$ be the sample mean for the treated group. Let $\bar{Y}_0 = \frac{1}{N_0} \sum_{i:D_i=0} Y_i$ be the sample mean for the control group.

  By the CLT,

$$\begin{pmatrix} \sqrt{N_1} \left(\bar{Y}_1 - E[Y_i(1)] \right) \\ \sqrt{N_0} \left(\bar{Y}_0 - E[Y_i(0)] \right) \end{pmatrix} \rightarrow_d \mathrm{N} \left(\mathbf{0}, \begin{pmatrix} Var(Y_i(1)) & 0 \\ 0 & Var(Y_i(0)) \end{pmatrix} \right).$$

Hypothesis Testing for Experiments (continued)

- Note that $\frac{N_1}{N} = \frac{1}{N}\sum_i D_i \to_p E[D_i]$ by the LLN. Similarly, $\frac{N_0}{N} \to_p E[1 - D_i]$

- Hence, applying the continuous mapping theorem,

$$\begin{aligned}\sqrt{N}(\bar{Y}_1 - E[Y_i(1)]) &= (1/\sqrt{N_1/N}) \cdot \sqrt{N_1}(\bar{Y}_1 - E[Y_i(1)]) \\ &\to_d (1/\sqrt{E[D_i]}) \cdot \mathrm{N}(0, Var(Y_i(1))) \\ &= \mathrm{N}\left(0, \frac{1}{E[D_i]} Var(Y_i(1))\right)\end{aligned}$$

- Applying similar steps for $\bar{Y}_0$, we obtain that

$$\sqrt{N}\begin{pmatrix} \bar{Y}_1 - E[Y_i(1)] \\ \bar{Y}_0 - E[Y_i(0)] \end{pmatrix} \to_d \mathrm{N}\left(0, \begin{pmatrix} \frac{1}{E[D_i]} Var(Y_i(1)) & 0 \\ 0 & \frac{1}{1-E[D_i]} Var(Y_i(0)) \end{pmatrix}\right)$$

Hypothesis Testing for Experiments (continued)

- We just showed that

$$\sqrt{N}\begin{pmatrix} \bar{Y}_1 - E[Y_i(1)] \\ \bar{Y}_0 - E[Y_i(0)] \end{pmatrix} \rightarrow_d \mathrm{N}\left(0, \begin{pmatrix} \frac{1}{E[D_i]} Var(Y_i(1)) & 0 \\ 0 & \frac{1}{1-E[D_i]} Var(Y_i(0)) \end{pmatrix}\right)$$

- Applying the CMT,

$$\sqrt{N}(\bar{Y}_1 - \bar{Y}_0 - E[Y_i(1) - Y_i(0)]) \rightarrow_d N(0, \sigma^2),$$

where $\sigma^2 = \frac{1}{E[D_i]} Var(Y_i(1)) + \frac{1}{E[1-D_i]} Var(Y_i(0))$

- We can thus form a 95% confidence interval for $\tau = E[Y_i(1) - Y_i(0)]$,

$$\bar{Y}_1 - \bar{Y}_0 \pm 1.96\hat{\sigma}/\sqrt{N},$$

where $\hat{\sigma}^2 = \frac{N}{N_1}\hat{\sigma}_1^2 + \frac{N}{N_0}\hat{\sigma}_0^2$, where $\hat{\sigma}_d^2$ is the sample variance for treatment group $d \in \{0, 1\}$

Sample Means for Depression Outcome (Again)

	Control Group	Treated Group
Mean	0.329	0.306
SD	0.470	0.461
N	10426	13315

- Our point estimate of the treatment effect is $\hat{\tau} = 0.306 - 0.329 = -0.023$.
- Our CI for the treatment effect is:

$$\hat{\tau} \pm 1.96 \times \sqrt{\frac{1}{N_1}\hat{\sigma}_1^2 + \frac{1}{N_0}\hat{\sigma}_0^2} =$$

$$-0.023 \pm 1.96 \times \sqrt{\frac{1}{13315}0.461^2 + \frac{1}{10426}0.470^2}$$

$$= [-0.035, -0.001]$$

Hypothesis Testing under Unconfoundedness

- Recall that under unconfoundedness, $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))|X_i$, we have

$$\underbrace{E[Y_i(1) - Y_i(0)|X_i = x]}_{CATE(x)} = E[Y_i|D_i = 1, X_i = x] - E[Y_i|D_i = 0, X_i = x]$$

  That is, within each value of X_i, it's as if we have an experiment.

- By the same logic as for experiments, we have that

$$\sqrt{N_x}(\bar{Y}_{1,x} - \bar{Y}_{0,x} - E[Y_i(1) - Y_i(0)|X_i = x]) \to_d N(0, \sigma_x^2),$$

  where $N_x = |i : X_i = x|$ and
  $\sigma_x^2 = \frac{1}{E[D_i|X_i = x]} Var(Y_i(1)|X_i = x) + \frac{1}{E[1 - D_i|X_i = x]} Var(Y_i(0)|X_i = x)$.

- So we can also do hyptothesis testing on $CATE(x)$ when N_x is large.

- By averaging $CATE(x)$, we can do hypothesis testing / form CIs for ATE.

The Challenge of Continuous x

- We've shown thus far how we can estimate $CATE(x)$ when the number of observations with $X_i = x$ is large.

- This works great when X_i is binary (e.g. an indicator for college) or takes on a small number of discrete values (e.g. 50 states).

- But what about when X_i is continuous?

- For example, if X_i is income, then to estimate $CATE(50,351)$, the theory we have says we need a large number of treated and control units both with income \$50,351. In most datasets, we won't have very many people with exactly this income.

- We thus need a different way of estimating conditional means when X_i is continuously distributed.

- The next part of the course will focus on achieving this take using linear regression as an approximation to the CEF.