Confidence Intervals

Gov 2001: Quantitative Social Science Methods I

Scott Cunningham

Harvard University

Spring 2026

Welcome back — this week we stay in asymptotics

Before break: CLT, standard errors, consistency

This week:

- **Today**: Confidence intervals — the payoff of everything so far
- **Wednesday**: Practice problems together

Next week: Hypothesis testing

Review: the CLT tells us the shape of $\bar{Y}_n$

From Week 7 — the result we'll use all day

If $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} (\mu, \sigma^2)$, then:

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

In words: no matter what the population looks like, the standardized sample mean is approximately normal for large n

Why normal? Each Y_i carries its own idiosyncratic shape — skewness, bumps, gaps — but averaging washes those out, because they point in different directions across observations and cancel, leaving only symmetric variation around μ

Review: the standard error measures precision

From Week 7

Standard error of $\bar{Y}_n$:

$$\boxed{\text{SE}(\bar{Y}_n) = \frac{\sigma}{\sqrt{n}}}$$

Problem: we don't know σ

Solution: plug in $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$

This works because $\hat{\sigma} \xrightarrow{p} \sigma$ (consistency)

Review: Slutsky says replacing σ with $\hat{\sigma}$ is valid

From Week 7 — the step that makes the CLT usable

We know: $\dfrac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1)$

We want: to replace σ (unknown) with $\hat{\sigma}$ (computed from data)

Slutsky's theorem says: if $\hat{\sigma} \xrightarrow{p} \sigma$, then

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

Plugging in a consistent estimator doesn't change the limiting distribution

This is the quantity we'll rearrange to build confidence intervals

Everything we've built leads to one place

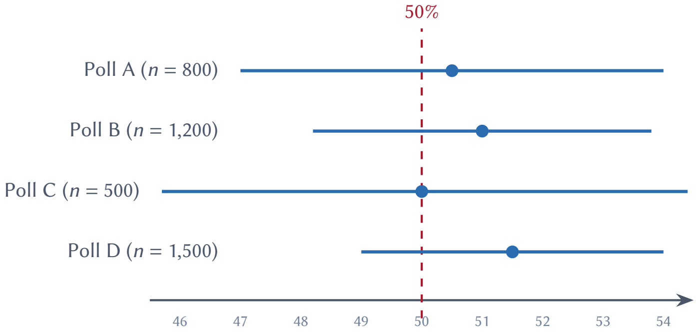
The logical chain from probability to inference

LLN: $\bar{Y}_n \xrightarrow{p} \mu$ → CLT: $\frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$ → **CI**: $\bar{Y} \pm z_{\alpha/2} \cdot \widehat{\text{SE}}$

where — *what shape* — *how precise*

Polls always come with a margin of error

November 2024 Pennsylvania polling



Without uncertainty, point estimates are dangerous

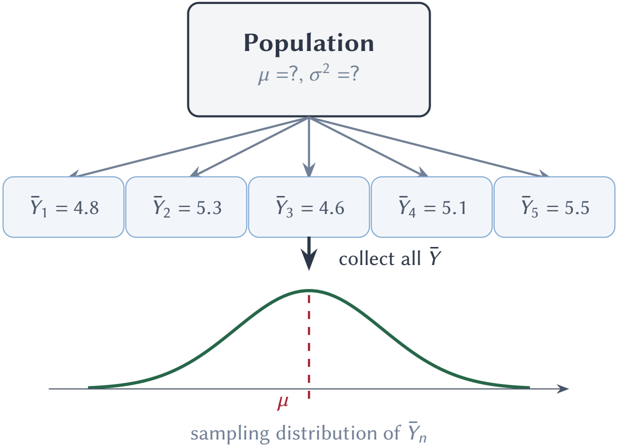
Claim	Point estimate	95% CI
Candidate leads by 2 pts	$\hat{p} = 0.52$	$[0.489,\ 0.551]$
GOTV raises turnout 5 pp	$\hat{\tau} = 0.05$	$[-0.01,\ 0.11]$
Job training raises wages $2k	$\hat{\beta} = 2000$	$[800,\ 3200]$

Today's roadmap

1. **What is a confidence interval?**
   A range of values for the population parameter that are compatible with the data
2. **What is it not?**
   Not a probability statement about μ — it's a statement about the *procedure*
3. **How do we calculate one?**
   CLT + Slutsky $\longrightarrow$ $\bar{Y} \pm z_{\alpha/2} \times \widehat{\text{SE}}$
4. **What does it mean?**
   If we repeated the sampling, 95% of intervals built this way would contain μ
5. **How should you use it in practice?**
   Report it alongside every point estimate — it tells your reader what the data can and cannot rule out

Every sample gives a different answer

The sampling distribution of $\bar{Y}_n$



Variability is the problem — the CLT turns it into a solution

Why not just report $\bar{Y}_n$? Because it isn't μ — it's one realization from a distribution of possible sample means, and a different sample would give a different answer

What we need: to say how far $\bar{Y}_n$ could plausibly be from μ

What we need to know: how much does $\bar{Y}_n$ vary, and in what shape?

The CLT answers both: the variation is $\sigma/\sqrt{n}$, and the shape is normal

That's enough to build an interval

The CLT gives us a quantity we can rearrange

From the CLT:

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

(this is the z-score — it measures how many SEs our sample mean is from the truth)

Slutsky lets us plug in $\hat{\sigma} \xrightarrow{p} \sigma$:

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

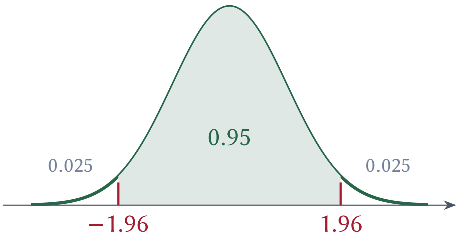
(same thing, but now everything is computable except μ)

This quantity has relatives — we'll meet them soon

Quantity	Random?	What it does	When?
$\widehat{Z}_n = \dfrac{\bar{Y}_n - \mu}{\hat{\sigma}/\sqrt{n}}$	Yes	Build CIs	**Today**
Test statistic: plug in μ_0 for μ	Yes	Test claims about μ	Next week
p-value: tail area of test statistic	Yes	Quantify evidence	Next week

Same machinery, different questions. Today we ask "where is μ?" Next week we ask "is μ equal to some claimed value?"

A standard normal concentrates 95% of its mass between ± 1.96



Since $\widehat{Z}_n \approx N(0, 1)$ by the CLT, this 95% area applies to our quantity directly:

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

Rearrange to isolate μ

Three algebraic steps

Start: $\mathbb{P}\left(-1.96 < \frac{\bar{Y}-\mu}{\hat{\sigma}/\sqrt{n}} < 1.96\right) \approx 0.95$

Multiply through by $\hat{\sigma}/\sqrt{n}$:

$$\mathbb{P}\left(-1.96\,\frac{\hat{\sigma}}{\sqrt{n}} < \bar{Y} - \mu < 1.96\,\frac{\hat{\sigma}}{\sqrt{n}}\right) \approx 0.95$$

Subtract $\bar{Y}$, multiply by -1:

$$\mathbb{P}\left(\bar{Y} - 1.96\,\widehat{\mathsf{SE}} < \mu < \bar{Y} + 1.96\,\widehat{\mathsf{SE}}\right) \approx 0.95$$

The confidence interval formula

General formula for level $1 - \alpha$:

$$\bar{Y} \ \pm \ z_{\alpha/2} \times \widehat{\mathrm{SE}}$$

*$z_{\alpha/2}$ is the **critical value** — the cutoff that leaves $\alpha/2$ probability in each tail*

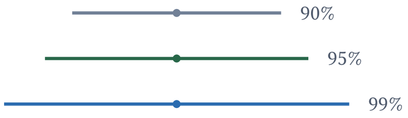
$$\boxed{95\%\ \mathrm{CI}: \quad \bar{Y} \pm 1.96 \times \widehat{\mathrm{SE}}}$$

$$99\%\ \mathrm{CI}: \quad \bar{Y} \pm 2.576 \times \widehat{\mathrm{SE}} \qquad 90\%\ \mathrm{CI}: \quad \bar{Y} \pm 1.645 \times \widehat{\mathrm{SE}}$$

Higher confidence $\rightarrow$ larger critical value $\rightarrow$ wider interval. The tradeoff is precision for coverage.

Higher confidence requires a wider net

Confidence level	α	$z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.960
99%	0.01	2.576



Markov's inequality bounds tail probabilities using only the mean

Review from before break — you will use this on Problem Set 6

Where we saw this: Week 6, asymptotics — a tool for bounding probabilities without knowing the full distribution

Why we need it now: The CLT-based CI assumes normality (large n). Markov and Chebyshev give bounds that work *without* that assumption — a different route to uncertainty quantification

$$\text{If } Y \geq 0 \text{ with mean } \mathbb{E}[Y] = \mu : \quad \mathbb{P}(Y \geq a) \leq \frac{\mu}{a}$$

You only need the mean. No variance, no normality, no large n.

Markov in practice: bounding hospital ER bills

Setup: ER bills are non-negative with $\mathbb{E}[Y] = \$800$

Question	Bound	Calculation
$\mathbb{P}(Y \geq \$2{,}000)$	≤ 0.40	$800/2000$
$\mathbb{P}(Y \geq \$4{,}000)$	≤ 0.20	$800/4000$
$\mathbb{P}(Y \geq \$8{,}000)$	≤ 0.10	$800/8000$

These bounds are loose — but they're guaranteed. No assumptions about the shape of the distribution.

Chebyshev adds the variance for a tighter bound

Setup: Y with $\mathbb{E}[Y] = \mu$, $\text{Var}(Y) = \sigma^2$

$$\mathbb{P}(|Y - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

where k is any distance from the mean — "how far from μ are we asking about?"

Same ER example: $\mathbb{E}[Y] = \$800$, $\text{SD}(Y) = \$400$

$\mathbb{P}(|Y - 800| \geq 1{,}200) \leq \frac{400^2}{1200^2} = 0.111$

$\mathbb{P}(Y \geq 2{,}000) \leq 0.111 \quad$ vs. Markov: ≤ 0.40

Key: more information (variance) $\longrightarrow$ tighter bound

Chebyshev bounds tails even tighter — and that means we can build a CI

From bounding probabilities to bounding μ

Chebyshev bounds how far $\bar{Y}$ can fall from μ:

$$\mathbb{P}(|\bar{Y} - \mu| \geq k) \leq \frac{\mathsf{Var}(\bar{Y})}{k^2}$$

Why this matters for CIs: this is already a statement about $|\bar{Y} - \mu|$ — the distance between our estimate and the truth. If we can bound that distance, we can build an interval around $\bar{Y}$ that traps μ

From Chebyshev to a confidence interval in three steps

1. **Choose** how much tail probability you'll tolerate (α)
2. **Solve** for how far $\bar{Y}$ could plausibly be from μ
3. **That distance is the margin of error** — no normality required

This is the same logic as the CLT-based CI: bound the distance between $\bar{Y}$ and μ, then build an interval. The only difference is *what tool* gives us the bound.

The Chebyshev CI: wider but assumption-free

Set the bound equal to α and solve for k: $\quad \dfrac{\mathrm{Var}(\bar{Y})}{k^2} = \alpha \quad \Longrightarrow \quad k = \dfrac{\sigma}{\sqrt{n\alpha}}$

$$\text{Chebyshev } (1-\alpha) \text{ CI}: \quad \bar{Y} \pm \frac{\hat{\sigma}}{\sqrt{\alpha \cdot n}}$$

At 95%: multiplier $= 1/\sqrt{0.05} = 4.47$ vs. CLT's 1.96

What this means: you need ±4.47 standard errors to guarantee 95% coverage when you don't know the shape. Under the normal, ±1.96 is enough because the bell curve's tails die off fast

Less information about the distribution → wider interval. The normal assumption is powerful precisely because it rules out fat tails.

The CLT gives a tighter CI by using more information

CLT CI (assumes large n, so $\bar{Y}$ is approximately normal):

$$\text{CLT } (1-\alpha) \text{ CI}: \quad \bar{Y} \pm z_{\alpha/2} \times \frac{\hat{\sigma}}{\sqrt{n}}$$

At 95%: multiplier = 1.96

Method	**Multiplier (95%)**	**Assumption**
Chebyshev	4.47	Finite variance
CLT	1.96	Large n

So why not always use the CLT?

You should — when n is large enough for the normal approximation to be accurate

Chebyshev is the guarantee that works even when it isn't

- $n = 1{,}000$ poll? Use the CLT. The normal approximation is excellent
- $n = 15$ with a skewed distribution? Chebyshev is safer — the CLT may not have kicked in yet
- In practice, most applied work uses the CLT because samples are large

The real lesson: assumptions buy you precision. The CLT bets that n is large enough — and that bet usually pays off ($4.47 \to 1.96$).

How large is “large n”?

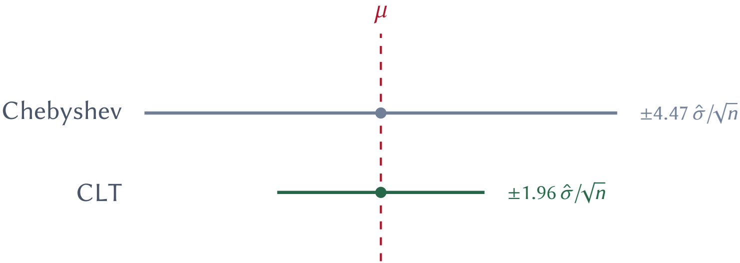
It depends on how non-normal the population is

Population shape	**n needed**	**Example**
Symmetric, light-tailed	10–15	Uniform
Moderately skewed	30–50	Exponential
Heavily skewed / heavy-tailed	100+	Log earnings, Pareto
Rare events ($p \approx 0.01$)	500+	Bernoulli(0.01)

The textbook rule: $n \geq 30$ is a useful shorthand, not a law

The real driver is **skewness**. More skewed populations need larger n for the CLT to kick in. The Berry-Esseen theorem makes this precise: approximation error is $O(1/\sqrt{n})$, with a constant that grows with the third moment.

Chebyshev vs CLT: same center, very different width



Tradeoff: Chebyshev needs only finite variance; CLT needs large n

The most common misinterpretation in statistics

WRONG: "There is a 95% probability that μ is in this interval."

Correct: 95% of intervals constructed this way contain μ

What's fixed?	μ — a number, not random
What's random?	The interval (depends on the sample)
After computing:	It either contains μ or it doesn't

This misinterpretation is so common it prompted a professional reckoning

2016: The American Statistical Association issues its first-ever statement on p-values — six principles clarifying what they do and don't mean

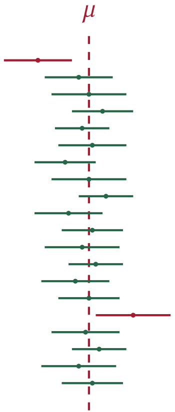
2019: Over 800 statisticians sign a call in *The American Statistician* to retire "statistical significance" as a binary threshold

The problem isn't the tools — it's that researchers routinely misinterpret them, and the misinterpretation has become the de facto interpretation in published work

This is why we drill the correct interpretation. Getting CIs and p-values right is an active problem in the profession — not a settled one.

Twenty samples, twenty intervals — most catch the truth

Each horizontal line is one CI from a fresh i.i.d. sample



Green: contains μ. Red: misses.

The CI is a net, not a target

What's fixed?	μ (a number, not random)
What's random?	The interval ($\bar{Y} \pm 1.96 \times \widehat{\mathsf{SE}}$)
Before sampling:	95% chance the net lands on μ
After sampling:	It either caught μ or it didn't

So which CI do you have?

You computed one interval from one sample. You can't know whether it's one of the 19 that caught μ or the 1 that missed

That discomfort is real — and it's the honest answer. A CI is not a guarantee about *this* interval. It's a guarantee about the *procedure*

Why it's still useful: you are betting on a method that works 95% of the time. You can't verify it in any single case — but repeatedly using a method with 95% coverage is a rational way to navigate uncertainty

It's like wearing a seatbelt. You can't know whether *this* drive is the one where it matters. But using the procedure is still the right call.

Putting it into words: returns to a college degree

Setup: $\hat{\beta} = 15\%$, SE $= 7.5\%$, 95% CI: $[7.5\%, \ 22.5\%]$

Say this:

- "Our estimate of the return to college is 15%. If we repeated this study with a new i.i.d. sample, we'd expect different estimates — the SE of 7.5% tells us how much they'd typically vary"
- "The 95% CI of $[7.5\%, 22.5\%]$ says: values between 7.5% and 22.5% are *consistent with* our data. Values far outside — like 2% or 30% — are not. This is the range of plausible values for the true return, given this sample"
- "If we repeated this procedure many times, 95% of the resulting intervals would contain the true return — that's what justifies calling it a 95% CI"

Putting it into words: what *not* to say

Same setup: $\hat{\beta} = 15\%, \quad \text{SE} = 7.5\%, \quad 95\%\ \text{CI: } [7.5\%,\ 22.5\%]$

Don't say this:

- × "There is a 95% chance the true return is between 7.5% and 22.5%"
- × "The return to college is definitely 15%"
- × "The return is somewhere between 7.5% and 22.5%"

The first sounds right but treats μ as random. The second ignores uncertainty entirely. The third drops the "95%" and states it as a certainty. All three are published regularly.

What are we estimating when we poll 1,000 voters?

Concept: identifying the population parameter

The population: all likely voters in this electorate

The parameter: p = true proportion who support the candidate

The sample: $n = 1{,}000$ voters drawn i.i.d. from this population

The estimate: $\hat{p} = 0.52$ — this number depends on *which* 1,000 voters we happened to draw

The question: how far could $\hat{p}$ plausibly be from p?

Example: polling margin of error

Calculate: SE $\rightarrow$ CI

Setup: Poll of $n = 1{,}000$ likely voters, $\hat{p} = 0.52$

Step 1 — Standard error:

$$\widehat{\mathsf{SE}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.52 \times 0.48}{1000}} = 0.0158$$

Step 2 — 95% CI:

$$0.52 \pm 1.96 \times 0.0158 = [0.489,\ 0.551]$$

Margin of error: $1.96 \times 0.0158 \approx \pm 3.1$ percentage points

Interpreting the polling CI

Interpret: what can we say?

$$[0.489,\ 0.551]$$

The words:

"If we were to repeatedly draw i.i.d. samples of 1,000 voters from this population and compute $\hat{p} \pm 1.96 \times \widehat{\text{SE}}$ each time, 95% of those intervals would contain the true population proportion p."

What's fixed?	p (the true proportion)
What's random?	The interval (it moves with each sample)
This particular interval?	Either contains p or doesn't — we don't know

What are we estimating in a GOTV experiment?

Concept: the population treatment effect

The population: all voters who could have received this mailer

The parameter: $\tau = p_T - p_C$ = the true difference in turnout rates between treatment and control populations

The samples: $n_T = 38{,}218$ treated, $n_C = 191{,}243$ control — drawn i.i.d. from their respective populations

The estimate: $\hat{\tau} = 0.081$ — a function of this particular sample

Example: GOTV experiment

Gerber, Green, and Larimer (2008)

Setup: $n_T = 38{,}218$ (treatment), $n_C = 191{,}243$ (control)
$\hat{p}_T = 0.378, \quad \hat{p}_C = 0.297, \quad \hat{\tau} = 0.081$

Step 1 — SE of the difference:

$$\widehat{\mathsf{SE}}(\hat{\tau}) = \sqrt{\frac{0.378 \times 0.622}{38{,}218} + \frac{0.297 \times 0.703}{191{,}243}} = 0.00264$$

Step 2 — 95% CI:

$$0.081 \pm 1.96 \times 0.00264 = [0.076,\ 0.086]$$

Interpreting the GOTV CI

Interpret: what can we say?

$$[0.076,\ 0.086]$$

The words:

"If we were to repeatedly draw i.i.d. samples of this size from the treatment and control populations and compute $\hat{\tau} \pm 1.96 \times \widehat{\mathsf{SE}}$ each time, 95% of those intervals would contain the true population difference in turnout τ."

Notice: the interval is narrow (±0.5 percentage points) because *n* is enormous

Notice: the entire interval is far from zero — this effect is precisely estimated *and* substantively large

Example: a smaller experiment tells a different story

What if we had less data?

Similar effect size, but $n_T = 500$, $n_C = 500$

$\hat{p}_T = 0.65, \quad \hat{p}_C = 0.60, \quad \hat{\tau} = 0.05$

SE of the difference:

$$\widehat{\mathsf{SE}}(\hat{\tau}) = \sqrt{\frac{0.65 \times 0.35}{500} + \frac{0.60 \times 0.40}{500}} = 0.0305$$

95% CI:

$$0.05 \pm 1.96 \times 0.0305 = [-0.010,\ 0.110]$$

Same formula, very different conclusions

Interpret: sample size changes the story

	Large experiment	Small experiment
$\hat{\tau}$	0.081	0.05
$\widehat{\mathsf{SE}}$	0.00264	0.0305
95% CI	$[0.076,\ 0.086]$	$[-0.010,\ 0.110]$
Excludes 0?	Yes	No

The lesson: an effect can be real but undetectable with small n

The CI doesn't say "the effect is zero" — it says "we can't tell with this much data"

What are we estimating in the canvassing experiment?

Concept: treatment effect on attitudes

The population: all people who could have been canvassed in this way

The parameter: $\tau = \mu_T - \mu_C$ = the true difference in mean attitude scores between treatment and control populations

The samples: $n_T \approx 1{,}100$ treated, $n_C \approx 1{,}050$ control — i.i.d. draws from their respective populations

The estimate: $\hat{\tau} = \bar{Y}_T - \bar{Y}_C$ — varies across samples

The CI will tell us: which values of τ are compatible with the data we observed

CIs in the wild: canvassing and same-sex marriage attitudes

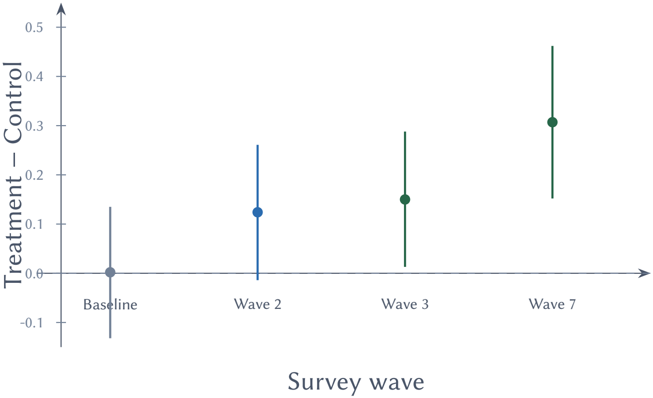
Broockman and Kalla (2016, *Science*)

Background: Door-to-door canvassing experiment

- Canvassers had conversations encouraging perspective-taking
- Treatment: same-sex marriage script ($n_T \approx 1{,}100$)
- Control: no contact ($n_C \approx 1{,}050$)
- Outcome: support for same-sex marriage (1–5 scale)

Five survey waves: baseline, then follow-ups over months

The CIs tell a nuanced story — not every wave is significant



Gray: CI includes zero Blue: borderline Green: excludes zero

Computing the CI: Wave 3 (3 days after canvassing)

Calculate: SE → CI

Wave 3: $n_T = 1{,}058$, $n_C = 1{,}055$

$\bar{Y}_T = 3.102, \quad \bar{Y}_C = 2.952, \quad \hat{\tau} = 0.150$

Step 1 — Standard error:

$$\widehat{\mathsf{SE}}(\hat{\tau}) = \sqrt{\frac{s_T^2}{n_T} + \frac{s_C^2}{n_C}} = 0.070$$

Step 2 — 95% CI:

$$0.150 \pm 1.96 \times 0.070 = [0.013,\ 0.288]$$

Computing the CI: Wave 7 (9 months after canvassing)

Calculate: SE $\rightarrow$ CI

Wave 7: $n_T = 793$, $n_C = 735$

$\bar{Y}_T = 3.333, \quad \bar{Y}_C = 3.026, \quad \hat{\tau} = 0.307$

Step 1 — Standard error:

$$\widehat{\mathsf{SE}}(\hat{\tau}) = \sqrt{\frac{s_T^2}{n_T} + \frac{s_C^2}{n_C}} = 0.079$$

Step 2 — 95% CI:

$$0.307 \pm 1.96 \times 0.079 = [0.152,\ 0.462]$$

Interpreting the Broockman and Kalla CIs

Interpret: what can we say?

Wave 3 — CI: $[0.013,\ 0.288]$

"If we were to repeatedly draw i.i.d. samples of $\approx 1{,}050$ treated and control individuals from these populations and compute $\hat{\tau} \pm 1.96 \times \widehat{\text{SE}}$ each time, 95% of those intervals would contain the true population difference in attitudes τ."

Wave 7 — CI: $[0.152,\ 0.462]$

Same interpretation. But the entire interval is far from zero — and the point estimate *grew* over time.

The CIs tell us: the effect is real, it persists, and it may even strengthen

A note on where the randomness lives

An aside for the intellectually curious

We just said: "if we repeatedly sampled..."

But these 2,113 people aren't drawn from a super-population — they are the actual participants. The randomness came from the coin flip that assigned treatment.

	Our framework	**Design-based**
Randomness	Sampling from P	Treatment assignment
Outcomes	Random draws	Fixed numbers
Target	Population $\mu_T - \mu_C$	Finite-pop. EATT
CI formula	*the same*	

Rambachan and Roth (2025, *JASA*) formalize this. The formula you just learned is more robust than the derivation suggests. For this course: we stay with sampling.

Three levers control the width of a confidence interval

$$\text{Width} = 2 \times z_{\alpha/2} \times \frac{\hat{\sigma}}{\sqrt{n}}$$

Confidence level

$z_{\alpha/2}$

Higher $\longrightarrow$ wider

Variability

$\hat{\sigma}$

Noisier $\longrightarrow$ wider

Sample size

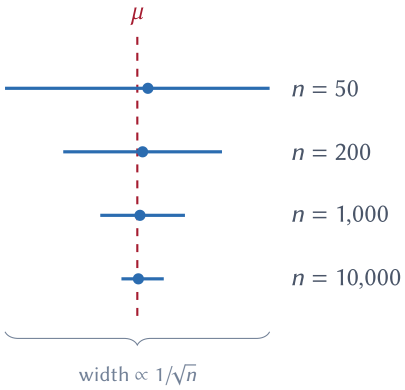
n (via $1/\sqrt{n}$)

More data

$\longrightarrow$ narrower

Visualizing the sample size effect

Same population, same μ, different n



How much data do you need?

Sample size planning

Goal: 95% CI with margin of error $\leq m$

Requirement: $1.96 \times \sigma/\sqrt{n} \leq m$

$$\boxed{n \geq \left(\frac{1.96\,\sigma}{m}\right)^2}$$

Polling example: $\pm 3\%$ margin of error, worst-case $\sigma = 0.5$:

$$n \geq \left(\frac{1.96 \times 0.5}{0.03}\right)^2 = 1{,}067$$

The $\sqrt{n}$ rule makes large effects cheap and small effects expensive

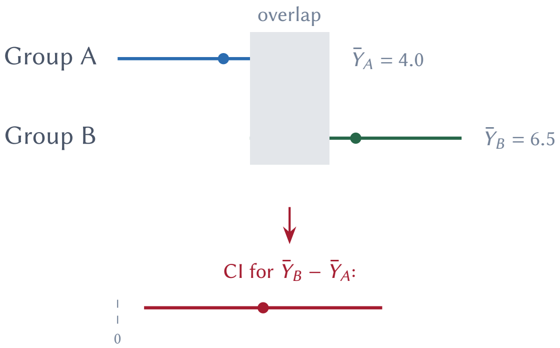
Desired MoE	Required n	Context
$\pm 10\%$	96	Pilot study
$\pm 5\%$	384	Departmental survey
$\pm 3\%$	1,067	National poll
$\pm 1\%$	9,604	Large-scale experiment
$\pm 0.5\%$	38,416	Census-scale

Three common mistakes about confidence intervals

Mistake	Correct
"95% probability $\mu \in [a, b]$"	95% of intervals contain μ
"Overlapping CIs $\Rightarrow$ no difference"	CIs can overlap even when the difference is significant
"Wider CI = worse estimate"	Width reflects n and σ, not estimator quality

Overlapping CIs do not mean "no difference"

A surprisingly common error in published research



Why does overlap mislead? The SE of a difference is not what you think

Each CI uses its own SE:

$$\text{CI}_A : \quad \bar{Y}_A \pm 1.96 \times \frac{\hat{\sigma}_A}{\sqrt{n_A}} \qquad \text{CI}_B : \quad \bar{Y}_B \pm 1.96 \times \frac{\hat{\sigma}_B}{\sqrt{n_B}}$$

But the CI for the difference has a *different* SE:

$$\text{SE}(\bar{Y}_B - \bar{Y}_A) = \sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}$$

*This is **smaller** than what you'd eyeball from two separate CIs, because* $\sqrt{a^2 + b^2} < a + b$

So two intervals can overlap slightly, yet the CI for their difference still excludes zero.

Always build the CI for the quantity you care about

The rule: if your question is about a *difference*, build a CI for the difference — don't compare separate CIs by eye

Example: $\bar{Y}_A = 4.0$, $\bar{Y}_B = 6.5$, $\text{SE}_A = \text{SE}_B = 1.0$, $n_A = n_B = 100$

- Separate CIs: $[2.04, 5.96]$ and $[4.54, 8.46]$ — they overlap
- $\text{SE}(\bar{Y}_B - \bar{Y}_A) = \sqrt{1^2 + 1^2} = 1.414$
- CI for difference: $2.5 \pm 1.96 \times 1.414 = [-0.27,\ 5.27]$

Overlap does not mean "no difference." Schenker & Gentleman (2001, *The American Statistician*) showed this formally — and it remains one of the most common errors in applied work.

What confidence intervals can and cannot tell you

CIs can:

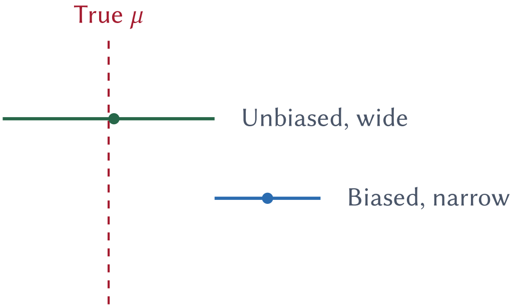
- Quantify sampling uncertainty around an estimate
- Show which values of μ are compatible with the data
- Guide sample size decisions before data collection
- Compare groups via the CI for their *difference*

CIs cannot:

- Fix a biased estimator (garbage in, precise garbage out)
- Account for non-sampling errors (measurement, selection)
- Tell you whether an effect is *practically* important
- Replace thinking about *identification*

A narrow CI around a biased estimate is worse than useless

Precision ≠ accuracy



Sometimes we care about a *function* of μ, not μ itself

Review from before break

Problem: We know how to build a CI for μ. But what if we want a CI for some transformation $g(\mu)$?

Examples:

- Average income is μ, but you want a CI for $\log(\mu)$ (log scale for skewed data)
- Bernoulli probability is p, but you want a CI for $\frac{p}{1-p}$ (the odds)
- Treatment effect is μ, but you want a CI for μ^2 (squared effect size)

The difficulty: the CLT tells us the distribution of $\bar{Y}$, not of $g(\bar{Y})$

We need a way to "transfer" the CLT through a function. That's the delta method.

The delta method: a linear approximation transfers the CLT

Plain language first, then the formula

In words: near μ, the function g is approximately a straight line. A straight line just rescales the normal — it stays normal, with a new variance determined by the slope $g'(\mu)$

Formally: if $\sqrt{n}(\bar{Y} - \mu) \xrightarrow{d} N(0, \sigma^2)$ and $g'(\mu) \neq 0$:

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

The slope $g'(\mu)$ stretches or compresses the variance — steep slope means more variance, flat slope means less

CI for $g(\mu)$:

$$g(\bar{Y}) \ \pm\ z_{\alpha/2} \times |g'(\bar{Y})| \times \widehat{\mathrm{SE}}(\bar{Y})$$

Same recipe as before: estimate ± critical value × SE. The SE just gets scaled by $|g'|$.

The full toolkit: five results, one inference machine

LLN: $\bar{Y}_n \xrightarrow{p} \mu$

CMT: $g(\bar{Y}_n) \xrightarrow{p} g(\mu)$

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

Slutsky: plug in $\hat{\sigma} \xrightarrow{p} \sigma$

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

Confidence Intervals

Wednesday: we practice all of this together

On Wednesday, we will work through problems together in class

Bring: pen, paper, calculator (or laptop with R)

Pset 6 due: Tuesday Mar 31 | **Next week**: Hypothesis testing