Hypothesis Testing

Gov 2001: Quantitative Social Science Methods I

Scott Cunningham

Harvard University

Spring 2026

Hypothesis testing: from plausible ranges to yes-or-no decisions

Required

- **Aronow & Miller**, §3.3.2–3.3.3 (pp. 130–142)
- **Blackwell**, Ch. 4 (pp. 79–97)

This is the last new probability material before regression.

Last week we asked: what values are plausible?

The confidence interval framework

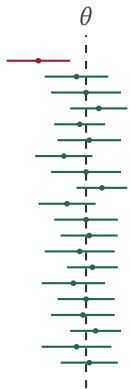
Monday: we built the 95% CI

$$\bar{Y}_n \pm 1.96 \times \widehat{\text{SE}}$$

The interpretation: if we repeated this ***procedure*** many times, 95% of the resulting intervals would contain μ

- The 95% is a property of the ***procedure***, not of any one interval
- This specific interval either contains μ or it doesn't — we just don't know which

Your interval either caught θ or it didn't — the 95% doesn't tell you which



Green: contains θ **Red**: misses — you cannot tell which color yours is from the interval alone

The CI tells you what the data cannot rule out — not where θ lives

What the CI licenses	What it does not license
95% of intervals from this procedure contain θ	$\mathbb{P}(\theta \in [a, b]) = 0.95$ (θ is fixed, not random)
Values inside: those θ_0 the data would *not* reject	"θ probably lives in here" (that is Bayesian language)
We used a reliable procedure	We know this interval caught θ

"Consistent with the data" = not ruled out by the test at this level — nothing more, nothing less

Today we flip the question

From "what's plausible?" to "is *this* plausible?"

Confidence interval: "What values of θ are consistent with my data?"

Hypothesis test: "Is my data consistent with *this specific value* of θ?"

Same machinery — CLT, standard errors, the sampling distribution. Just packaged differently.

Hypothesis testing follows the logic of a criminal trial

Trial	Hypothesis test
Defendant is innocent	H_0: no effect
Prosecution argues guilty	H_1: there is an effect
Evidence presented	Data collected
Beyond reasonable doubt?	Is $p < \alpha$?
Verdict: guilty / not guilty	Reject H_0 / fail to reject

The key asymmetry: we assume innocence until proven guilty

- We never prove H_0 true — we only fail to reject it
- **"Not guilty" ≠ "innocent"** — insufficient evidence to convict
- **"Fail to reject H_0" ≠ "no effect"** — insufficient evidence to reject

The fallacy: H_0 true $\Rightarrow$ likely fail to reject. We fail to reject. $\therefore$ H_0 true. **Invalid.**

- The effect may be real — you may simply lack power to detect it
- **"No evidence of an effect" ≠ "evidence of no effect"**

This asymmetry is the most important idea in hypothesis testing

The null hypothesis is the claim we put on trial

Definitions

- **Null hypothesis** (H_0): the default claim (usually "no effect")
- **Alternative hypothesis** (H_1): what we believe if we reject H_0

Two-sided (standard): $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$

"Two-sided" means a surprise in either direction counts as evidence

- **"Sided"** refers to which direction would surprise you
- **Two-sided**: a big effect in *either* direction — positive or negative — counts as evidence against H_0
- You are not committing in advance to the sign of the effect
- **One-sided**: you've decided before seeing data that only one direction could matter — a much stronger assumption, and usually hard to defend

The experiment: social pressure and voter turnout

Gerber, Green, and Larimer (2008), *APSR*

Question: can social pressure increase voter turnout?

Design: 344,084 Michigan households randomly assigned to one of five groups before the August 2006 primary

- **Control**: no mailing
- **Civic Duty**: "Do your civic duty — vote!"
- **Hawthorne**: "You are being studied"
- **Self**: shows your own voting record
- **Neighbors**: shows your neighbors' voting records

Outcome: whether the person voted (binary)

Are these turnout differences real or just noise?

Group	Turnout	Difference from control	*n*
Control	29.7%	—	191,243
Civic Duty	31.5%	+1.8 pp	38,218
Hawthorne	32.2%	+2.5 pp	38,204
Self	34.5%	+4.8 pp	38,218
Neighbors	37.8%	+8.1 pp	38,201

Every treatment group has higher turnout than control

But could these differences be sampling variability?

We need a formal framework to answer this

For each treatment arm, we can state the hypotheses

Parameter: τ_k = average treatment effect for arm k

Neighbors arm:

- $H_0 : \tau_{\text{Neighbors}} = 0$ (no effect on turnout)
- $H_1 : \tau_{\text{Neighbors}} \neq 0$ (some effect)

Civic Duty arm:

- $H_0 : \tau_{\text{Civic}} = 0$
- $H_1 : \tau_{\text{Civic}} \neq 0$

Same hypotheses, very different effect sizes — we'll see this matters

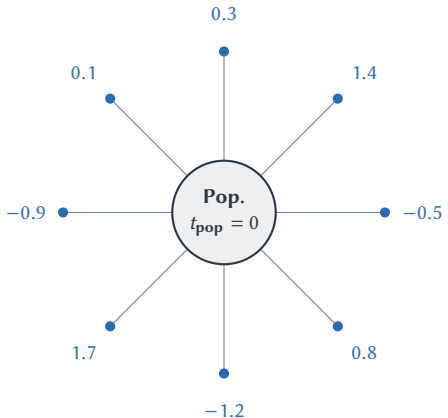
The t-statistic is a sample estimate of a population concept

Sample	Population
$\hat{t} = \dfrac{\hat{\theta} - \theta_0}{\widehat{\mathrm{SE}}(\hat{\theta})}$	$t_{\text{pop}} = \dfrac{\theta - \theta_0}{\sigma_\theta}$

- Under H_0: $t_{\text{pop}} = 0$ — a fixed population number
- Just as $\hat{\theta}$ estimates θ, our $\hat{t}$ estimates t_{pop}
- In the sample: one draw $\longrightarrow$ one observed $\hat{t}$

Even with all data in the world, you could calculate this — it would equal zero

Each sample gives one $\hat{t}$ — the wheel shows where they all land



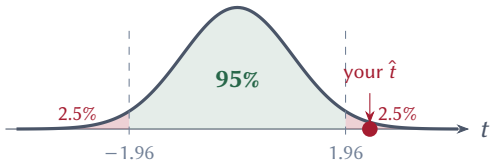
Each spoke is one sample's $\hat{t}$ — they scatter around the population value of zero

LLN centers the wheel; CLT shapes the distribution of spokes

- **LLN**: average $\hat{t}$ across samples $\xrightarrow{p} t_{\text{pop}} = 0$ — the wheel is right on average
- **CLT**: distribution of $\hat{t}$ across samples $\xrightarrow{d} N(0, 1)$ — the spokes follow a normal curve
- **Therefore**: 95% of all sample t-statistics fall within ± 1.96

We know the shape of the wheel before we collect any data

Judging a t: place it on the normal and see where it lands



- Reject H_0 when $\hat{t}$ falls in the tail: $|\hat{t}| > 1.96$
- $N(0, 1)$ has infinite tails — no value is impossible under H_0
- **We never say "true" or "false" — we say where on this distribution your $\hat{t}$ falls**

Computing the test statistic: Neighbors arm

$$\hat{\tau} = 0.378 - 0.297 = 0.081 \quad (8.1 \text{ pp})$$

$$\widehat{\text{SE}}(\hat{\tau}) = \sqrt{\frac{0.378 \times 0.622}{38{,}201} + \frac{0.297 \times 0.703}{191{,}243}} \approx 0.0026$$

$$t = \frac{0.081 - 0}{0.0026} = 31.2$$

31 standard errors from zero — astronomically far

Computing the test statistic: Civic Duty arm

$\hat{\tau} = 0.315 - 0.297 = 0.018$ (1.8 pp)

$\widehat{\mathrm{SE}}(\hat{\tau}) \approx 0.0026$ (similar sample sizes)

$$t = \frac{0.018 - 0}{0.0026} = 6.9$$

7× **smaller than Neighbors** — but far from zero

Both significant — but are the implications the same?

How far is "far enough"?

We need a decision rule

We have two numbers: $t = 31.2$ and $t = 6.9$

The question: how do we decide whether $|t|$ is large enough to reject H_0?

Enter the p-value

The p-value answers: how surprising is our data under H_0?

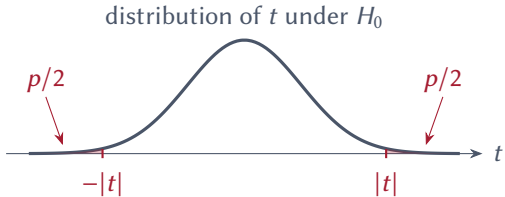
In words: if there were *truly* no effect, what is the probability of seeing a test statistic this large or larger?

P-value

$$p = \mathbb{P}(|T| \geq |t| \mid H_0 \text{ true})$$

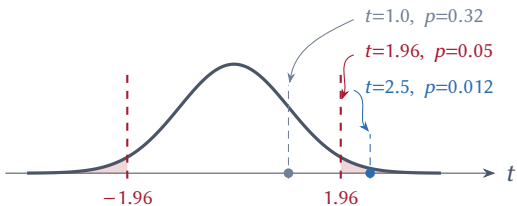
- Small p: data unlikely under H_0 $\rightarrow$ evidence against H_0
- Large p: data consistent with H_0 $\rightarrow$ no evidence against it

Visualizing the p-value



- **P-value** = total shaded area = $\mathbb{P}(|T| \geq |t| \mid H_0)$
- **Why $p/2$ on each side?** Two-sided test: a large t in *either* direction is evidence against H_0
- $N(0, 1)$ is symmetric $\Rightarrow$ each tail contributes exactly half: $p = p/2 + p/2$

The p-value moves as your t moves along the sampling distribution



- As $|t|$ grows, the tail area shrinks — the data become more surprising under H_0
- The 1.96 threshold **(red)** is always our reference: $p = 0.05$

Small p-value = surprising data = evidence against H_0

- $p = 0.50$: data perfectly consistent with H_0
- $p = 0.10$: mildly unusual, but not compelling
- $p = 0.01$: quite surprising under H_0
- $p < 0.001$: extremely surprising — strong evidence against H_0

The p-value is a continuous measure of evidence, not a binary switch

The ASA issued its first-ever statement on p-values in 2016

- Wasserstein & Lazar (2016), *American Statistician* — unprecedented
- Not: p-values are mathematically wrong
- But: systematic misuse as a binary pass/fail gate on scientific claims
- 2019 follow-up: "retire the phrase *statistical significance*"

CIs are just as misunderstood — both suffer from the same root cause

The modal researcher does not know what a p-value actually measures

What most researchers believe

- $p < 0.05$ means the result is real
- Small p means H_0 is probably false
- p measures the probability of a mistake

What it actually measures

- Where $\hat{t}$ falls on the sampling distribution under H_0
- $\hat{t}$ is itself an estimate of a population quantity ($= 0$ under H_0)
- Nothing about whether H_0 is true

Imprecision at the researcher level propagates into the scientific record at scale

The p-value measures; it does not decide

Positive

Where $\hat{t}$ falls on the sampling distribution under H_0 — a fact about the data

Normative

How extreme must $\hat{t}$ be before we act as if H_0 is false — a judgment

The math stops at the p-value — to act, we need a rule

Type I error convicts the innocent; Type II error frees the guilty

Type I Error

- *Trial:* Convict the innocent
- *Test:* Reject H_0 when H_0 is true
- False positive; probability = α

Type II Error

- *Trial:* Acquit the guilty
- *Test:* Fail to reject H_0 when H_0 is false
- False negative; probability = β

Criminal justice sets α low: "beyond reasonable doubt" — but the historical record suggests the effective α has not been equal across defendants

The threshold α is a subjective choice, not a mathematical fact

- Nothing in the math says 0.05 is "the" cutoff
- Fisher chose it as a convenient round number
- Different fields set different standards: genetics uses 5×10^{-8}; physics uses 5σ
- The choice encodes your tolerance for false positives (Type I errors)

α is normative — a human judgment about what counts as convincing evidence

The decision rule: reject if $p < \alpha$

Convention

Choose a **significance level** α (typically 0.05):

- If $p < \alpha$: **reject** H_0
- If $p \geq \alpha$: **fail to reject** H_0

Why 0.05? Tradition (Fisher), not law

Other fields use different thresholds: particle physics requires 5σ ($p < 3 \times 10^{-7}$)

"Fail to reject" is not the same as "accept H_0"

Back to the trial: "not guilty" ≠ "innocent"

When $p \geq 0.05$:

- We are NOT saying the effect is zero
- We are saying the evidence is not strong enough to rule it out
- The effect might exist — we just can't detect it with this data

This distinction matters enormously in policy: "no evidence of harm" is not "evidence of no harm"

Both GGL treatment arms are statistically significant

Neighbors: $t = 31.2$, $p \approx 0$ $\longrightarrow$ Reject H_0

Civic Duty: $t = 6.9$, $p \approx 0$ $\longrightarrow$ Reject H_0

Both are "statistically significant" at any conventional level

Statistical significance alone does not tell us which effect matters more

Equivalent approach: compare $|t|$ to a critical value

For two-sided test at $\alpha = 0.05$:

Reject H_0 if $\lvert t \rvert > 1.96$

$p < 0.05 \iff |t| > 1.96$

- $p < 0.05 \iff |t| > 1.96$ — **always**
- 1.96: same critical value as your 95% CI formula

The test and the CI are the same thing

The duality that connects Week 8 to today

Reject $H_0: \theta = \theta_0$ at $\alpha = 0.05$ **if and only if** θ_0 is **outside** the 95% CI

- CI = the set of all θ_0 values you would *not* reject
- θ_0 inside CI $\rightarrow |t| < 1.96 \rightarrow$ fail to reject
- θ_0 outside CI $\rightarrow |t| > 1.96 \rightarrow$ reject

Visualizing the duality

- Outside CI $\longrightarrow$ reject Inside CI $\longrightarrow$ fail to reject

Applying the duality: Neighbors arm

95% CI: $0.081 \pm 1.96 \times 0.0026 = [0.076,\ 0.086]$

Is 0 **in this interval?** No — not even close

$\rightarrow$ Reject $H_0 : \tau = 0$

Same conclusion as the test. Same machinery.

CIs are more informative: they tell you the *range* of plausible values, not just yes/no

Statistical significance is not practical significance

Statistical significance: $p < 0.05$

- The effect is probably not exactly zero
- Says nothing about whether the effect is *large* or *important*

Practical significance: is the effect big enough to matter?

- A 0.1 pp increase in turnout might be statistically significant with $n = 1{,}000{,}000$
- But is it worth the cost of a mailing campaign?

Both effects are significant — but they are not the same

GGL: Neighbors vs. Civic Duty

Arm	$\hat{\tau}$	t	Policy implication
Neighbors	8.1 pp	31.2	massive behavioral change
Civic Duty	1.8 pp	6.9	real but modest

- Both: $p \approx 0$ — but the policy implications are not the same

Always report effect sizes and CIs, not just p-values

Three things p-values do NOT mean

WRONG: "$p = 0.03$ means there's a 3% chance H_0 is true"

Right: 3% chance of data this extreme *if* H_0 were true

WRONG: "$p = 0.06$ means there's no effect"

Right: evidence not strong enough by convention — the effect might still exist

WRONG: "$p = 0.001$ means the effect is large"

Right: small p can come from small effects + large n

Testing n hypotheses at $\alpha = 0.05$ guarantees false positives

$$\mathbb{P}(\geq 1 \text{ false positive}) = 1 - 0.95^n$$

n	Effective false positive rate
1	5%
5	23%
10	40%
20	64%

- Stated $\alpha = 0.05$ throughout
- Nothing in the math changes
- Only the number of tests changes

Stated $\alpha = 0.05$; effective α grows with every additional test

P-hacking and publication bias are the same error at different scales

P-hacking (researcher)

- Run specs until $p < 0.05$, then stop
- Elastic α disguised as rigor

Publication bias (journal)

- Publish $p < 0.05$; file-drawer the rest
- Same selection, institutional scale

The scientific record = Type I errors curated to look like discoveries

GGL reported all 4 treatment arms — not just the largest

One-sided vs. two-sided tests

Two-sided (standard): $H_1 : \theta \neq 0$

- Reject if estimate is far from 0 in *either* direction
- P-value uses both tails

One-sided: $H_1 : \theta > 0$

- Only reject if estimate is positive and large
- P-value uses one tail (half as large)

Rule: use one-sided only if you would ignore evidence in the other direction

The full testing procedure is six steps you already know

1. **State hypotheses**: H_0 and H_1
2. **Choose significance level**: usually $\alpha = 0.05$
3. **Compute test statistic**: $t = (\hat{\theta} - \theta_0)/\widehat{\text{SE}}$
4. **Find p-value**: $p = \mathbb{P}(|T| \geq |t| \mid H_0)$
5. **Make decision**: reject H_0 if $p < \alpha$
6. **Interpret in context**: report effect size and CI

Key takeaways

1. **Hypothesis testing** asks: is data consistent with H_0?
2. **Test statistic** = $(\hat{\theta} - \theta_0)/\widehat{\text{SE}}$ — same CLT machinery
3. **P-value** = probability of data as extreme, if H_0 true
4. **Reject** H_0 if $p < \alpha$ (equivalently, if $|t| > 1.96$)
5. **Duality**: reject H_0 at 5% $\Leftrightarrow$ θ_0 outside 95% CI
6. **Statistical $\neq$ practical significance** — always report effect sizes

The test is only as good as the estimand you chose to test

A caveat that applies to everything we just did

Everything we've built today assumes $\hat{\theta}$ is a credible estimate of the right θ

- **LLN + CLT**: $\hat{\theta}$ centers on *your* θ and is bell-shaped around it
- **But**: you can run a perfect t-test on the wrong quantity
- Selected sample, bad comparison group, wrong outcome measure $\longrightarrow$ $p < 0.001$ anyway
- The test presupposes valid design — it cannot rescue a bad one

Rejecting H_0 is only meaningful if you set up the estimation correctly first

Kosuke Imai's classes will go deeper on identification — what it takes for $\hat{\theta}$ to be the right θ

$p < 0.001$ on ventilators → mortality does not mean ventilators kill

Selection bias + hypothesis testing = confident wrong answer

Data: regress mortality on ventilator use in hospitals. Find: **positive** coefficient, $p < 0.001$

What the test did correctly:

- $\hat{\beta} \xrightarrow{p} \beta_{\text{pop}}$ and t-stat $\approx N(0,1)$ under H_0 — valid
- Quantified: this estimate almost certainly did not arise by chance

What the test cannot see:

- Ventilated patients are already near death — selection bias
- β_{pop} is a biased estimand for the causal effect
- The sampling distribution converges, precisely, to the *wrong* target

A hypothesis test is not a truth machine — it is a measurement tool pointed at whatever estimand you gave it

Substituting p-values for causal assumptions is common. Kosuke Imai's classes are about not making that mistake.

Wednesday: power and bootstrap

Preview: both GGL effects were significant. But what if the sample had been smaller?

- Type I and Type II errors
- Power: probability of detecting a real effect
- Sample size planning
- The bootstrap: inference without formulas

Reading: A&M §3.3.3 and §3.4.3; Blackwell Ch. 4 (finish)