Hypothesis Testing

Gov 2001: Quantitative Social Science Methods I

Scott Cunningham

Harvard University

Spring 2026

Today's Reading

Required

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

Note: This is the last new material before the midterm!

Two Frameworks for Inference

Confidence Intervals (last time):

- Start with data, construct range of plausible values
- "What values of θ are consistent with my data?"

Hypothesis Testing (today):

- Start with a claim, ask if data provide evidence against it
- "Is my data consistent with this specific value of θ?"

They're connected: Testing $H_0 : \theta = \theta_0$ at level α is equivalent to checking if θ_0 is in the $(1 - \alpha)$ CI.

The Logic of Hypothesis Testing

Analogy: A criminal trial.

- **Null hypothesis** (H_0): Defendant is innocent
- **Alternative** (H_1): Defendant is guilty
- **Evidence**: The data
- **Decision**: Reject H_0 (guilty) or fail to reject (not guilty)

Key asymmetry:

- We assume innocence until proven guilty
- "Not guilty" $\neq$ "innocent"—just insufficient evidence
- Burden of proof is on the prosecution (the alternative)

Null and Alternative Hypotheses

Definitions

- **Null hypothesis** (H_0): The claim we're testing (usually "no effect")
- **Alternative hypothesis** (H_1 or H_a): What we believe if H_0 is false

Political science examples:

- GOTV intervention: H_0: treatment effect $= 0$
- UN peacekeeping: H_0: no effect on conflict duration
- Campaign spending: H_0: $\beta_{\text{spending}} = 0$ on vote share

Two-sided tests are more common: we test $\neq$ rather than $>$ or $<$.

Example: Testing a Treatment Effect

Research question: Does a get-out-the-vote intervention increase turnout?

Parameter: τ = average treatment effect on turnout

Hypotheses:

- $H_0 : \tau = 0$ (no effect)
- $H_1 : \tau \neq 0$ (some effect, positive or negative)

Data: Treatment group mean = 0.65, Control group mean = 0.60
Estimate: $\hat{\tau} = 0.05$ (5 percentage point increase)

Question: Is this 5pp difference real, or could it be sampling variability?

The Test Statistic

Test Statistic

A **test statistic** measures how far the estimate is from the null hypothesis value, in standard error units:

$$t = \frac{\hat{\theta} - \theta_0}{\text{SE}(\hat{\theta})}$$

Under H_0: If $\theta = \theta_0$, then $t \approx N(0, 1)$ by CLT.

Intuition:

- Large $|t| \Rightarrow$ estimate far from $H_0 \Rightarrow$ evidence against H_0
- Small $|t| \Rightarrow$ estimate consistent with H_0

Example: Computing the Test Statistic

Setup: $\hat{\tau} = 0.05$, $\text{SE}(\hat{\tau}) = 0.02$, $H_0 : \tau = 0$

Test statistic:

$$t = \frac{0.05 - 0}{0.02} = 2.5$$

Interpretation: The estimate is 2.5 standard errors away from zero.

Question: Is 2.5 "far enough" to reject H_0?
We need a decision rule. Enter the p-value.

The P-Value

Definition: P-Value

The **p-value** is the probability of observing a test statistic *at least as extreme* as the one we got, *assuming H_0 is true.*

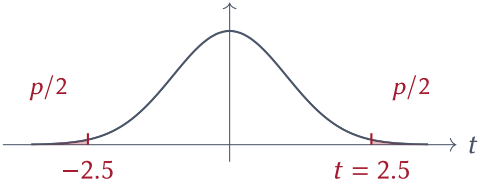
For two-sided test:

$$p = \Pr(|T| \geq |t| \mid H_0) = 2 \times \Pr(T \geq |t|)$$

Intuition: How "surprising" is our data under H_0?

- Small p-value $\Rightarrow$ data unlikely under H_0 $\Rightarrow$ evidence against H_0
- Large p-value $\Rightarrow$ data consistent with H_0

Visualizing the P-Value



P-value = shaded area = probability of getting $|t| \geq 2.5$ under H_0

For $t = 2.5$: $p = 2 \times \Pr(Z > 2.5) \approx 0.012$

The Decision Rule

Decision Rule

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

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

Our example: $p = 0.012 < 0.05$
Conclusion: Reject H_0. The treatment effect is statistically significant.

Important: "Fail to reject" $\neq$ "accept H_0"
We're saying the evidence isn't strong enough, not that H_0 is true.

Equivalence with Critical Values

Alternative approach: Compare $|t|$ to a critical value.

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

- Critical value: $z_{0.025} = 1.96$
- Reject H_0 if $|t| > 1.96$

Our example: $|t| = 2.5 > 1.96 \Rightarrow$ Reject H_0

The two approaches are equivalent:

- $p < 0.05 \Leftrightarrow |t| > 1.96$
- Both lead to the same decision

Connection to Confidence Intervals

Key insight: The test and CI use the same information.

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

Our example:

- 95% CI for τ: $0.05 \pm 1.96 \times 0.02 = [0.011, 0.089]$
- Is 0 in this interval? No!
- Therefore: Reject $H_0 : \tau = 0$

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

Statistical vs. Practical Significance

Critical distinction:

Statistical significance: $p < 0.05$

- The effect is unlikely to be exactly zero
- Says nothing about whether the effect is *large* or *important*

Practical significance: Is the effect big enough to matter?

- A 0.1 percentage point increase in turnout might be statistically significant with $n = 1{,}000{,}000$
- But is it meaningful for policy?

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

Common P-Value Mistakes

Wrong: "$p = 0.03$ means there's a 3% chance H_0 is true."
Right: $p = 0.03$ means there's a 3% chance of data this extreme *if* H_0 were true.

Wrong: "$p = 0.06$ means there's no effect."
Right: $p = 0.06$ means the evidence isn't quite strong enough by conventional standards. The effect might still exist.

Wrong: "$p = 0.001$ means the effect is large."
Right: Small p-values can come from small effects + large samples.

Caution: Multiple Testing

If you test many hypotheses, some will be "significant" by chance.

At $\alpha = 0.05$: You expect 1 false positive per 20 true null hypotheses.

P-hacking: Trying many specifications until finding $p < 0.05$

- Inflates false positive rate beyond stated α
- Contributes to replication failures

Best practice: Pre-register your hypothesis, report all tests, focus on effect sizes.

One-Sided vs. Two-Sided Tests

Two-sided (most common):

- $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$
- Reject if estimate is far from 0 in *either* direction
- P-value uses both tails

One-sided:

- $H_0 : \mu \leq 0$ vs. $H_1 : \mu > 0$
- Only reject if estimate is positive and large
- P-value uses one tail (half as large)

Rule: Use one-sided only if you'd ignore evidence in the other direction. Usually, use two-sided.

Summary: Hypothesis Testing Steps

1. **State hypotheses**: H_0 and H_1
2. **Choose significance level**: Usually $\alpha = 0.05$
3. **Compute test statistic**: $t = (\hat{\theta} - \theta_0)/\text{SE}$
4. **Find p-value**: $p = \Pr(|T| \geq |t| \mid H_0)$
5. **Make decision**: Reject H_0 if $p < \alpha$
6. **Interpret**: In context, with effect sizes!

Key Takeaways

1. **Hypothesis testing** asks: Is data consistent with H_0?
2. **P-value** = probability of data as extreme, if H_0 true
3. **Reject** H_0 if $p < \alpha$ (typically 0.05)
4. **Equivalent**: Reject if $|t| >$ critical value, or if θ_0 outside CI
5. **Statistical $\neq$ practical significance**
6. **Report effect sizes and CIs**, not just p-values

Next: Type I/II errors, power, and bootstrap.

Looking Ahead

Wednesday: Power and Bootstrap

- Type I error (false positive): Reject H_0 when true
- Type II error (false negative): Fail to reject when false
- Power: Probability of detecting a real effect
- Bootstrap: Inference when CLT doesn't apply

Then: MIDTERM EXAM covering Weeks 1–7!

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