Power and Bootstrap

Gov 2001: Quantitative Social Science Methods I

Scott Cunningham

Harvard University

Spring 2026

Today's reading

Required

- **Aronow & Miller**, §3.3.3: Power (pp. 138–142)
- **Aronow & Miller**, §3.4.3: Bootstrap (pp. 145–150)
- **Blackwell**, Ch. 4 (finish)

Last probability lecture. Regression starts next week.

Monday we learned to test — but testing can go wrong

Two kinds of mistakes

Monday: compute $t = (\hat{\theta} - \theta_0)/\widehat{\text{SE}}$ with $\theta_0 = 0$, get p, reject or fail to reject

But what if we're wrong?

- We rejected H_0 — but maybe the effect isn't real (false positive)
- We failed to reject — but maybe the effect is real and we missed it (false negative)

Today: formalizing these two mistakes and what determines how often they happen

The two-by-two table of errors

	H_0 True	H_0 False
Reject H_0	**Type I Error**	**Correct**
Fail to Reject	**Correct**	**Type II Error**

- **Type I**: false positive — convicting an innocent person
- **Type II**: false negative — letting a guilty person go free

Type I error rate is α — the significance level we already chose

Type I Error Rate

$$\alpha = \mathbb{P}(\text{Reject } H_0 \mid H_0 \text{ true})$$

When we set $\alpha = 0.05$, we accepted a 5% chance of rejecting a true null

This is controlled by design — we chose it on Monday

Type II error is β — the probability of missing a real effect

Type II Error Rate

$$\beta = \mathbb{P}(\text{Fail to reject } H_0 \mid H_0 \text{ false})$$

β is **not** controlled directly

It depends on: how big the effect is, how large the sample is, how noisy the data is

Power $= 1 - \beta =$ **probability of detecting a real effect**

Power

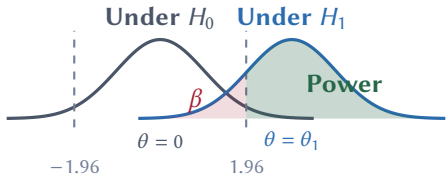
$$\text{Power} = 1 - \beta = \mathbb{P}(\text{Reject } H_0 \mid H_0 \text{ false})$$

Higher power is better — we want to find effects that are really there

Convention: target 80% power (detect the effect 4 out of 5 times)

Two distributions tell the whole story

The key visual for understanding power



Left: distribution of t if H_0 is true **Right**: distribution of t if H_1 is true

Power = green area β = red area

Power is about how much the two distributions overlap

- **Large effect** → H_1 distribution far from H_0 → little overlap → **high power**
- **Small effect** → distributions overlap a lot → low power
- **Large** n → both distributions are narrow → less overlap → **high power**
- **Small** n → wide distributions → more overlap → low power

Four factors that determine power

Power increases when:

1. **Effect size is larger**: easier to detect big effects
2. **Sample size is larger**: smaller SE, narrower distributions
3. **Variance is smaller**: less noise, clearer signal
4. α **is larger**: more willing to reject $\rightarrow$ more rejections

The tradeoff: increasing α increases power but also Type I error. We typically fix $\alpha = 0.05$ and increase n.

What if GGL had only 500 per group?

Neighbors arm: $\tau = 0.081$, binary outcome

SE with $n = 500$ per group:

$$\text{SE} = \sqrt{\frac{0.378 \times 0.622}{500} + \frac{0.297 \times 0.703}{500}} \approx 0.029$$

Reject if $|\hat{\tau}| > 1.96 \times 0.029 = 0.057$

Power: $\mathbb{P}(\hat{\tau} > 0.057 \mid \tau = 0.081)$

$$= \mathbb{P}\left(Z > \frac{0.057 - 0.081}{0.029}\right) = \mathbb{P}(Z > -0.83) = 0.80$$

80% power — we'd detect this effect most of the time, even with $n = 500$

Same sample, smaller effect: Civic Duty arm

$\tau = 0.018$ with $n = 500$ per group

Same SE ≈ 0.029. Same rejection threshold: $|\hat{\tau}| > 0.057$

Power: $\mathbb{P}(\hat{\tau} > 0.057 \mid \tau = 0.018)$

$$= \mathbb{P}\left(Z > \frac{0.057 - 0.018}{0.029}\right) = \mathbb{P}(Z > 1.34) = 0.09$$

Only 9% power — we'd miss this effect 91% of the time

Same sample size, same test. But the small effect is nearly undetectable.

GGL needed 344,000 households for a reason

Arm	τ	Power at $n = 500$	Power at $n = 38{,}000$
Neighbors	8.1 pp	80%	$\approx$ 100%
Self	4.8 pp	40%	$\approx$ 100%
Hawthorne	2.5 pp	14%	$\approx$ 100%
Civic Duty	1.8 pp	9%	$\approx$ 100%

With large enough n, even small effects become detectable

The actual GGL sample sizes made every arm well-powered

Sample size planning: how large a study do you need?

Before you collect data: calculate required n for adequate power

$$n_{\text{per group}} = \left(\frac{(z_{\alpha/2} + z_{\beta}) \cdot \sigma}{\delta} \right)^2$$

where δ = minimum effect you want to detect, $z_{\beta} = 0.84$ for 80% power

GGL example: detect a 3 pp effect, $\sigma \approx 0.49$ (conservative: $\sqrt{0.5 \times 0.5}$), 80% power:

$$n = \left(\frac{(1.96 + 0.84) \times 0.49}{0.03} \right)^2 \approx 2{,}090 \text{ per group}$$

Pre-study power analysis: the workflow

1. **Expected effect size**: from prior literature or minimum meaningful effect
2. **Expected variability**: from prior studies or pilot data
3. **Target power**: usually 80% (sometimes 90%)
4. **Alpha level**: usually 0.05
5. **Calculate required *n***: using formulas or simulation

You must specify the effect size BEFORE seeing data

This is why pre-registration matters — it forces you to commit to these choices

Most published studies fall far short of 80% power

Median power in social science: ~35% (Button et al., 2013)

- Small effects + limited samples = low power
- GOTV effects (~2–3 pp) need $n \approx 5{,}000+$ for 80% power
- Survey experiments with many conditions: power drops rapidly

Underpowered studies inflate published effect sizes

- Real effects go undetected — Type II error
- Published effects are inflated: only the lucky large estimates clear $p < 0.05$
- Contributes to replication crisis across social science

Winner's curse: published estimates are biased upward when studies are underpowered

Summary: errors and power

Concept	Definition	Typical value
Type I Error (α)	$\mathbb{P}(\text{reject } H_0 \mid H_0 \text{ true})$	0.05
Type II Error (β)	$\mathbb{P}(\text{fail to reject} \mid H_0 \text{ false})$	0.20
Power	$1 - \beta$	0.80

Power depends on: effect size, sample size, variance, α

We control α by choice. We increase power by increasing n.

Part II: The Bootstrap

Inference when formulas aren’t available

Everything so far assumed we know the SE formula

But what if we don't?

For means: SE = $\hat{\sigma}/\sqrt{n}$ — clean, simple

But what about:

- The **median**? No closed-form SE
- A **ratio** of two means? Complicated formula
- A **complex estimator** from matching or weighting?

We need a general-purpose tool for getting standard errors

Efron (1979) gave us a general-purpose tool for standard errors

- **Paper**: "Bootstrap Methods: Another Look at the Jackknife" (1979)
- **The name**: Baron Munchausen pulled himself out of a swamp by his own bootstraps
- **The insight**: use the data itself to learn about the data
- **Scale**: 200,000+ citations since 1979

Treat the sample as a stand-in for the population

Problem: we want the sampling distribution of $\hat{\theta}$, but we only have one sample

Insight: treat the sample as a stand-in for the population

Logic: sample $\approx$ population $\implies$ resampling from sample $\approx$ resampling from population

The bootstrap procedure: resample, compute, repeat

Original sample: $Y_1, Y_2, \ldots, Y_n$

For $b = 1, 2, \ldots, B$:

1. Draw n observations **with replacement** from $(Y_1, \ldots, Y_n)$
2. Compute $\hat{\theta}^{*b}$ on this bootstrap sample

Result: $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$

Use this distribution to:

- Estimate SE: $\widehat{\text{SE}} = \text{SD}(\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B})$
- Construct CI: use 2.5th and 97.5th percentiles

Why "with replacement"?

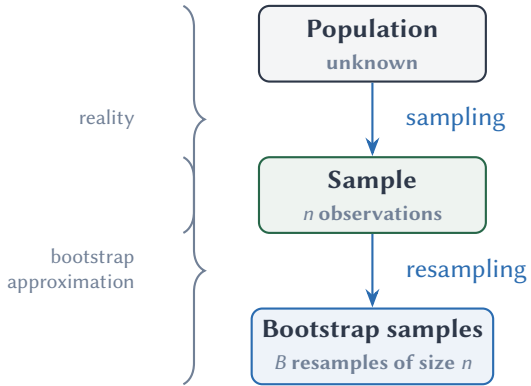
Without it, you just get the same sample

Original: {A, B, C, D, E}

Without replacement: {C, A, E, B, D} — just a reordering
Every resample is identical in composition. No new information.

With replacement: {A, A, C, C, E} — A and C appear twice; B and D absent
Creates genuine variation across bootstrap samples

Bootstrap resampling mirrors the original sampling process



Bootstrap logic: sample replaces the unknown population

Two methods for bootstrap confidence intervals

1. Percentile method (simplest):

$$\text{CI} = \left[\hat{\theta}^*_{(0.025)},\ \hat{\theta}^*_{(0.975)}\right]$$

Use the 2.5th and 97.5th percentiles of the bootstrap distribution

2. Normal approximation:

$$\text{CI} = \hat{\theta} \pm z_{\alpha/2} \times \widehat{\text{SE}}_{\text{boot}}$$

The percentile method is more robust to skewness

Bootstrap works for smooth estimators; fails for extremes

Works well for:

- Means, medians, quantiles
- Regression coefficients
- Most "smooth" functions of the data

Can fail for:

- Extremes (max, min) — not smooth
- Very small samples
- Non-i.i.d. data (need modified versions, e.g., block bootstrap)
- Parameters on the boundary (e.g., variance = 0)

Rule: consistent + asymptotically normal $\Rightarrow$ bootstrap works (same smoothness condition as delta method)

Bootstrap in R

```r
# Original statistic
theta_hat <- median(data)

# Bootstrap
B <- 10000
theta_boot <- numeric(B)
for (b in 1:B) {
  boot_sample <- sample(data, replace = TRUE)
  theta_boot[b] <- median(boot_sample)
}

# SE and CI
se_boot <- sd(theta_boot)
ci_boot <- quantile(theta_boot, c(0.025, 0.975))
```

Key takeaways

1. **Type I error** = false positive; controlled by α
2. **Type II error** = false negative; related to power
3. **Power** = $1 - \beta$ = probability of detecting a real effect
4. **Plan sample size** for 80% power *before* collecting data
5. **Bootstrap** = resample with replacement to approximate the sampling distribution
6. **Bootstrap CI**: use percentiles of the bootstrap distribution

Next week: regression

We have the full toolkit: estimation, CIs, tests, power, bootstrap

Now: apply it all to the workhorse of empirical social science

- What is regression estimating? (the BLP)
- OLS as the sample analog
- Every coefficient comes with a t-statistic and p-value — same framework

Reading: Blackwell Ch. 5; A&M §2.2.4; MHE 3.1