Statistical Challenges for Economists

What Breaks When the Assumptions Do

The Assumptions So Far

Everything so far lived in a clean world. Each deck made an assumption, used it, and moved on.

Assumption	Made in	Broken in
The sample is representative	Deck 01 — *"we will simply assume"*	§4
Observations are i.i.d.	Deck 02 — the LLN and CLT both open with it	§3
We know the distribution family	Deck 04 — every likelihood	§1
There is one true θ	Decks 02–04	§2
You ran one test	Deck 05	§5
You observe X	everywhere	§6

This is not a list of complaints

The tools work. They have preconditions — and economics is where those preconditions bind.

One Number, Six Ways to Break It

Every section of this lecture makes the same claim about a different mechanism:

Your nominal 95% is not 95%.

- Six mechanisms — misspecification, dependence over time, dependence across units, selection, forking paths, measurement error
- One shared symptom, and it is measurable
- Every demo today prints the same line: `nominal 95% → actual XX%`

> **Keep score**
>
> By the end we will have six rows. You should be able to recite the table, not just the mechanisms.

1 · There Is No True DGP

All Models Are Wrong

Maximum likelihood asks: *which* θ makes the observed data most probable — **under the family I assumed**?

- In physics that family is sometimes actually right
- In economics it never is. Nobody believes earnings are exactly lognormal, or that the effect of schooling is exactly linear
- So what is the MLE converging to?

It still converges

Not to nothing, and not to the truth: to the **pseudo-true parameter** — the member of your wrong model class that is closest to reality in Kullback–Leibler divergence.

Your estimator is consistent. For the wrong thing.

Misspecification, Live

Data curve; the model is a straight line. Watch what the estimator does — and what its standard error claims.

Curvature

0.6

Heteroskedasticity

0.8

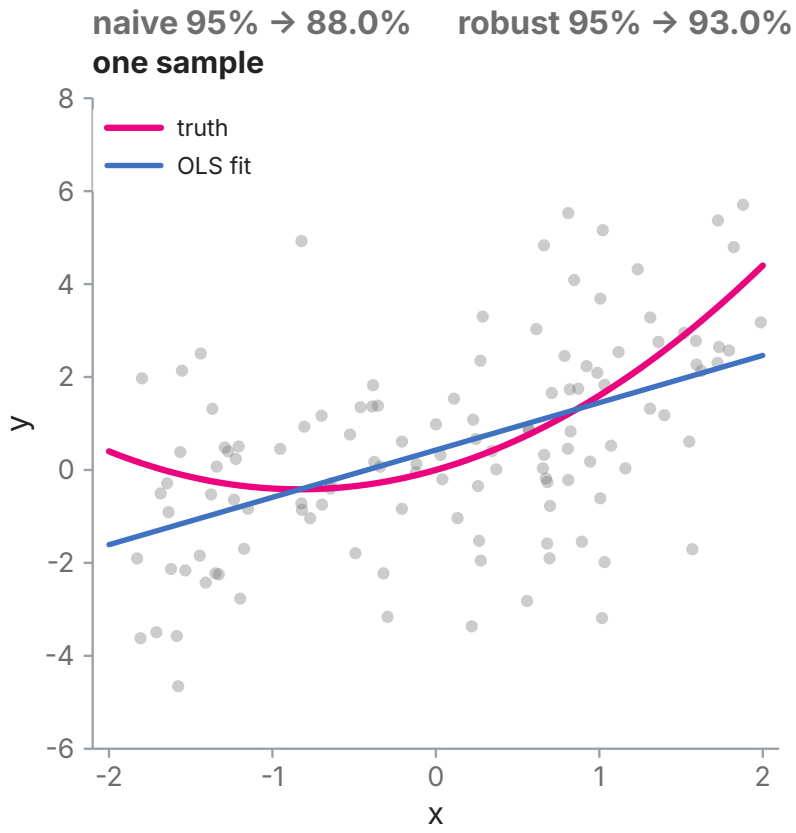
n

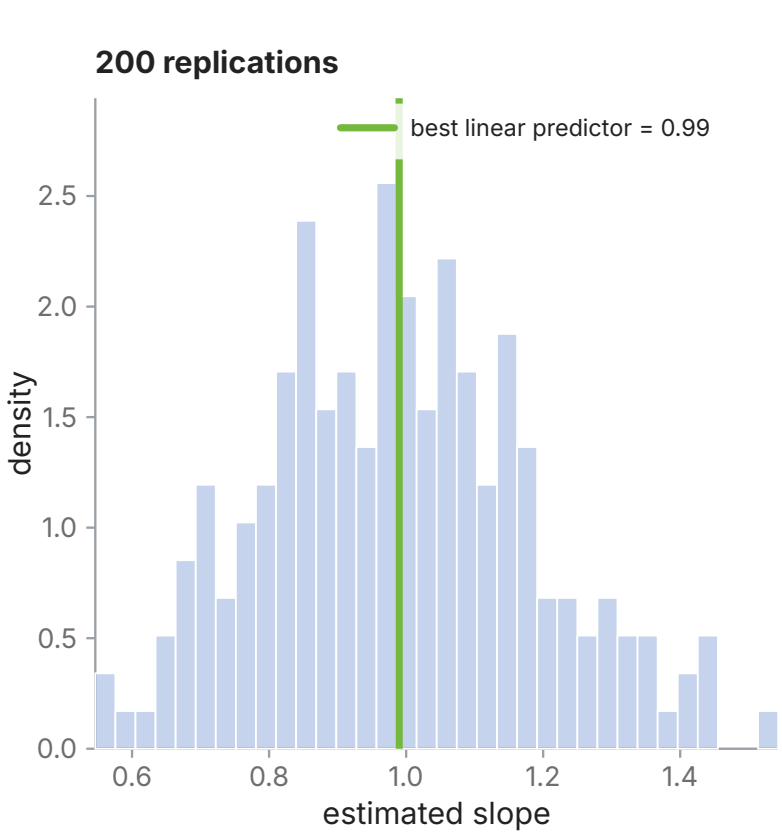
120

Run 200 more

The green line is the **best linear predictor** — the pseudo-true parameter. The estimator finds it, tightly and confidently.

Robust standard errors repair the *coverage*. They do not repair the *question*.





What That Showed

- The estimator does **not** wander. It converges — tightly, confidently — on the best linear approximation, which is not the truth
- Turn heteroskedasticity up and the **naive** interval drops to ~85% coverage; the **robust** one holds near 94%
- And it holds only for the pseudo-true parameter, not for anything structural

Scoreboard, row 1

Correct specification broken · nominal 95% → **actual ~85–88%** with naive standard errors.

Robust standard errors fix the *coverage*. Nothing fixes the fact that you are now precisely estimating the answer to a different question.

The Sandwich

$$\mathbf{Var}(\hat{\beta}) = \underbrace{(X'X)^{-1}}_{\text{bread}} \underbrace{\left(\sum_i e_i^2 x_i x_i'\right)}_{\text{meat}} \underbrace{(X'X)^{-1}}_{\text{bread}}$$

- When the model is right, bread and meat cancel and you recover the textbook $\sigma^2(X'X)^{-1}$
- When it is wrong, only the sandwich survives

The statistical point

The estimator and its standard error fail **independently** — and the standard error fails first. That is why a wrong model can look so convincing: the point estimate is stable and the interval around it is too narrow.

The econometrics course derives this. Here it is enough to know why it exists.

2 · Heterogeneous Populations

Whose Effect Is It?

Decks 02–04 wrote "θ" as though there were one. Suppose instead every unit i has its own effect τ_i.

- **ATE** — the average over everyone. What a randomised experiment estimates
- **ATT** — the average over the treated. What a programme evaluation usually wants
- **LATE** — the average over *compliers*: those an instrument actually moved

> **These differ, and the difference is not noise**
>
> If the people who benefit most are also the most likely to enrol, ATT $>$ ATE by construction — and no sample size closes the gap, because they are **different quantities**.

The estimand is a choice you make. It is not a fact about the data.

An Average That Describes Nobody

Suppose a training programme helps half its participants by 1.0 and harms the other half by 1.0.

- ATE $= 0$
- The confidence interval is tight around zero
- The p-value is 0.98
- **The programme is life-changing for everybody in it**

What gets reported

"No statistically significant effect."

What is true

A large effect on everyone, in two directions.

The fix is not statistical

Report the *distribution*: quantile treatment effects, subgroup analysis, variance of outcomes. An average without a distribution is a mean without a variance — and we would never accept that.

Simpson's Paradox, Live

Every group trends **down**. The pooled data trends **up**. Nothing is wrong with the data.

Group separation

1.8

Within-group slope

-0.6

Groups

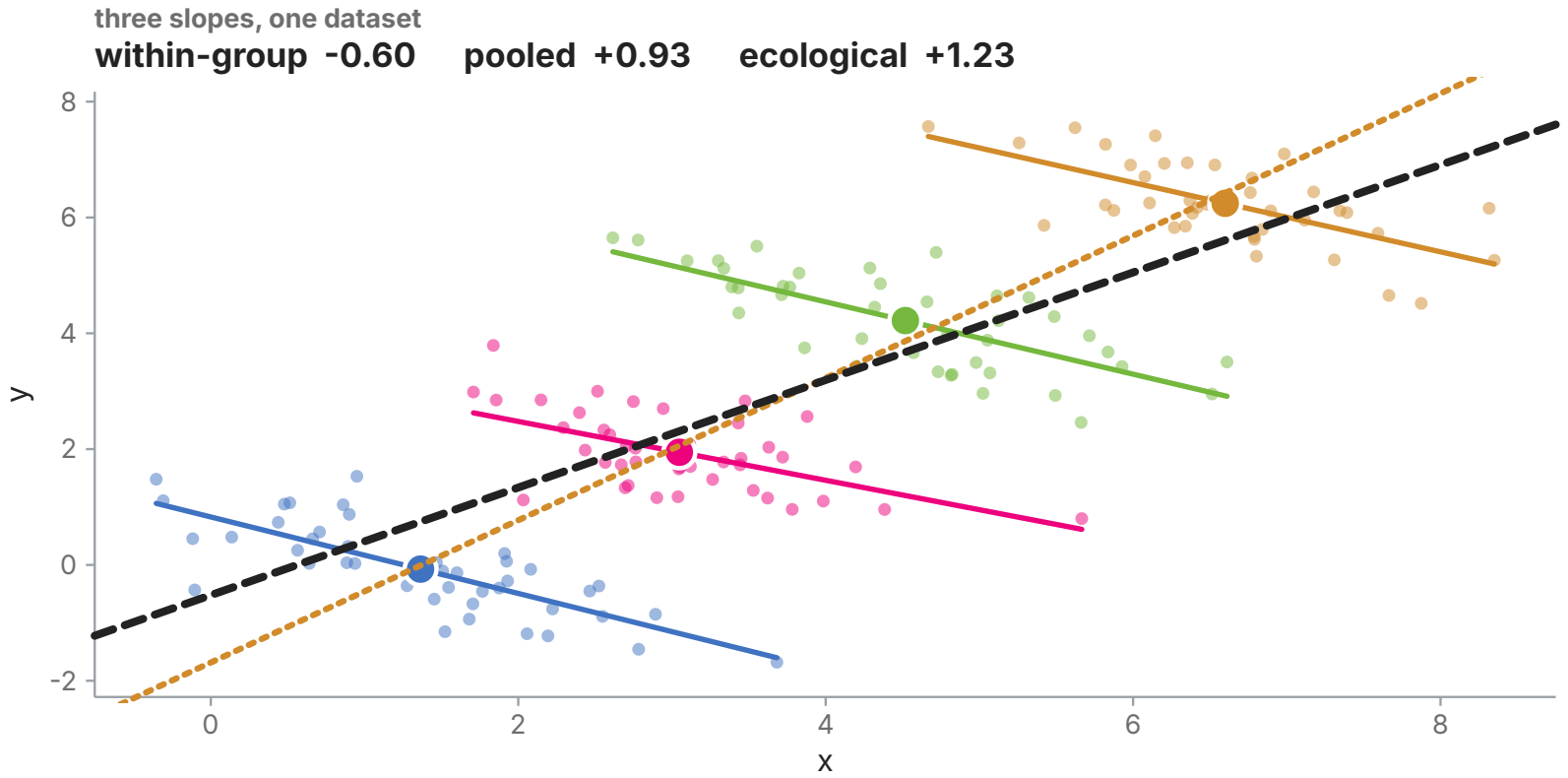
4

Show

Both

Slide separation to zero and the paradox dissolves. Slide it up and the pooled line flips sign — while **no individual data point moves**.

"Group means" fits a line through the centroids: the *ecological* regression, a third answer again.



Three Slopes, One Dataset

- **Within-group** — negative. The relationship for any actual person
- **Pooled** — positive. What you get by ignoring the groups
- **Ecological** — steeper still. What you get from group averages, which is what published aggregate data usually *is*

The ecological fallacy

Regressions on regional or national averages do not estimate the individual-level relationship. They estimate the between-group relationship, which can have the opposite sign.

Historically the original sin of cross-country growth regressions. Currently alive and well in anything estimated on regional aggregates.

3 · Dependence

The i.i.d. Assumption Was Doing a Lot of Work

Both of deck 02's theorems open the same way:

Laws of Large Numbers · Central Limit Theorem

"Let $X_1, X_2, \ldots$ be **i.i.d.** with $E[X_i] = \mu$ …"

Economic data essentially never is:

- **Over time** — GDP this quarter looks like GDP last quarter
- **Across units** — households in a village share a harvest, a school, a labour market

The good news first

Under stationarity and mixing, the averages **still converge**. The LLN survives.

What does not survive is $\mathbf{Var}(\bar{X}) = \sigma^2/n$.

The Dependence Engine, Live

The same failure, twice — once indexed by time, once by group.

Structure

Autocorrelated in time

ρ

0.8

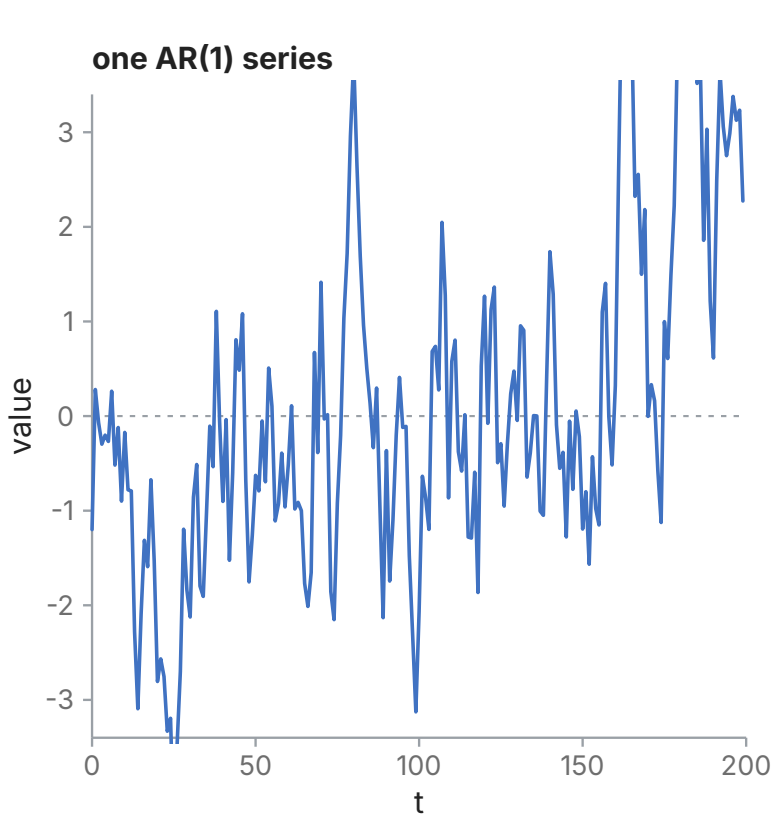
Cluster size m

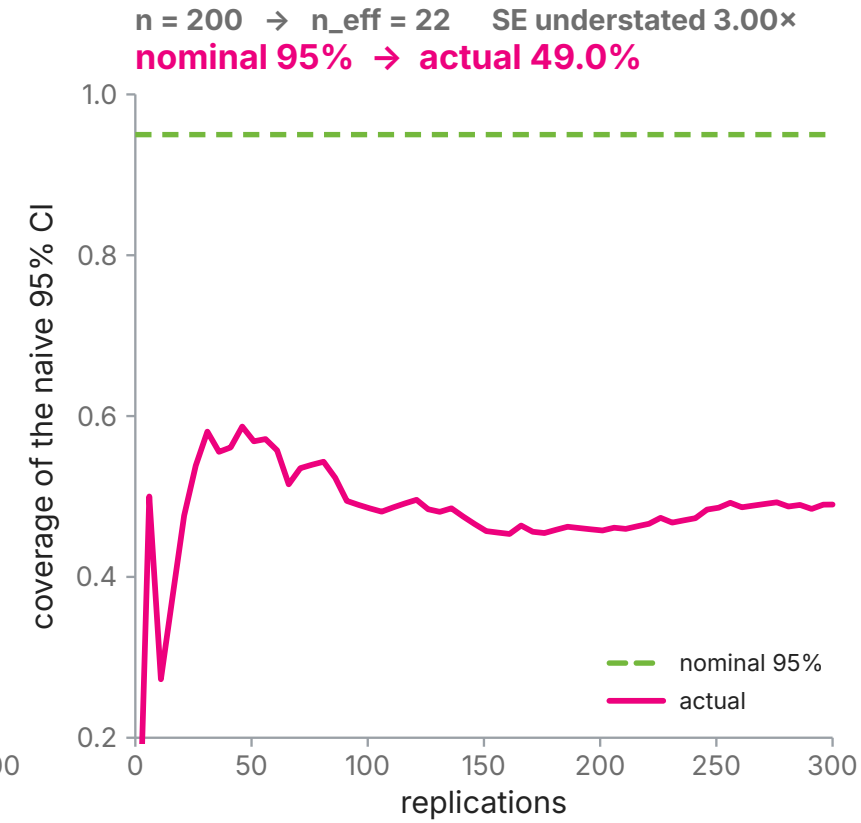
30

Run 300 more

$n_{\text{eff}} = n\frac{1-\rho}{1+\rho}$ in time; $n_{\text{eff}} = \frac{n}{1+(m-1)\rho}$ in clusters.

Both are n divided by a dependence inflation factor. The data look fine either way — it is the standard error that is lying.





Clusters Are Autocorrelation, Reindexed

The Moulton factor

$$\frac{\mathrm{SE}_{\text{true}}}{\mathrm{SE}_{\text{naive}}} = \sqrt{1 + (m-1)\rho}$$

With clusters of $m = 30$ and an intra-cluster correlation of only $\rho = 0.1$, that is **1.97**.

What that does to a result

A reported $t = 2.1$ — comfortably significant, publishable —

is really $t = 1.06$.

Nothing.

- $\rho = 0.1$ is small enough that nobody would notice it in the data
- Policies vary at the state, district or school level, but data arrive per person
- “Cluster your standard errors” is not folklore or fashion. It is this arithmetic

Scoreboard, rows 2 and 3

Independence over time broken · 95% → **~49%** · across units · 95% → **~66%**

Spurious Regression, Live

Two series. Nothing whatsoever connects them. How often do we “find” a relationship?

Series

Random walks

Length

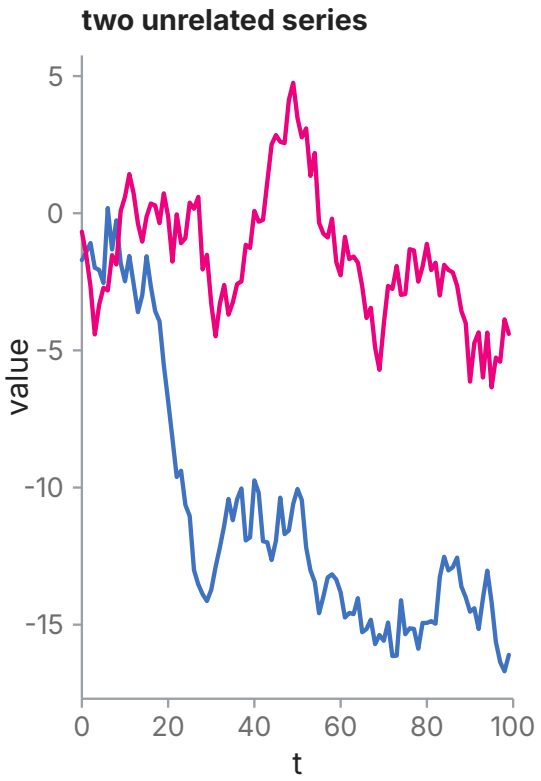
100

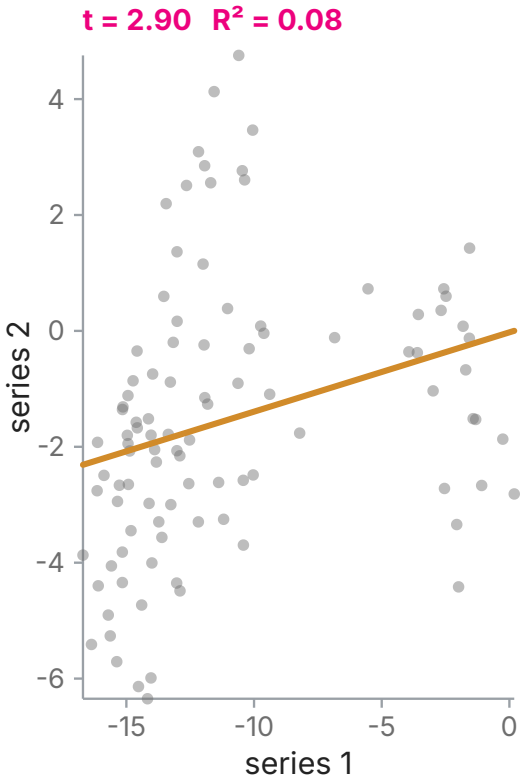
Draw 200 more

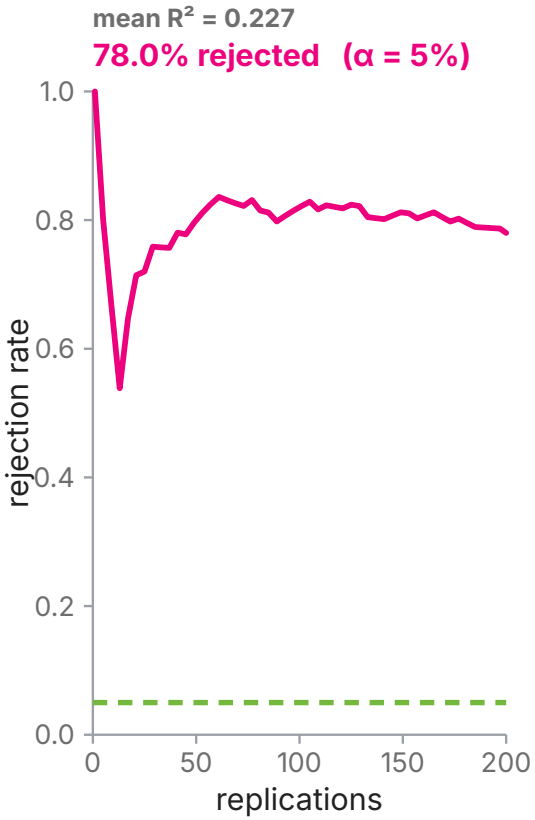
i.i.d. is the control: deck 05 promised 5%, and you get 5%.

Random walks — trending, non-stationary, and utterly unrelated — reject about three quarters of the time, often with a high R^2.

Differencing restores the guarantee.







What the Tally Says

- **i.i.d. noise** — 5%. Exactly what deck 05 promised. The guarantee works
- **Random walks** — about **three quarters** (≈77%), with a mean R^2 around 0.2
- **Differenced** — back to 5%

> **Read that again**
>
> Two series with nothing to do with each other are “significantly related” about three quarters of the time — and produce a respectable R^2 while doing it.

Most macroeconomic series are trending and non-stationary. A literature could be built out of this, and to some extent was.

Parameters That Move

Even with the right model and independent data, there may be no fixed θ to estimate.

- **Structural breaks** — a rolling estimate that jumps in 2008, and again in 2020
- Estimating over the whole sample averages regimes that no longer exist
- The “parameter” is a weighted average of several worlds

The Lucas critique, as a statistical statement

Acting on an estimated relationship can **change** that relationship, because the agents generating your data are optimising against the policy you inferred from it.

No other science has quite this problem. It is the strongest possible version of “there is no fixed true DGP” — the DGP reads your paper.

4 · Selection

Two Questions, Not One

Internal validity

Is the estimate right **for this sample**?

Threatened by: confounding, selection, bad controls, measurement.

An RCT buys you this.

External validity

Does it transfer **to anyone else**?

Threatened by: unusual populations, unusual doses, unusual times, general equilibrium.

An RCT buys you nothing here.

They fail independently

A perfectly executed experiment on 400 university students in one city can be internally flawless and externally useless. Both facts can be true at once, and neither standard error mentions the second.

The Selection Engine, Live

X and Y are **independent**. True slope: zero. Three ways to make a relationship appear.

Mechanism

Select on the outcome Y

Severity

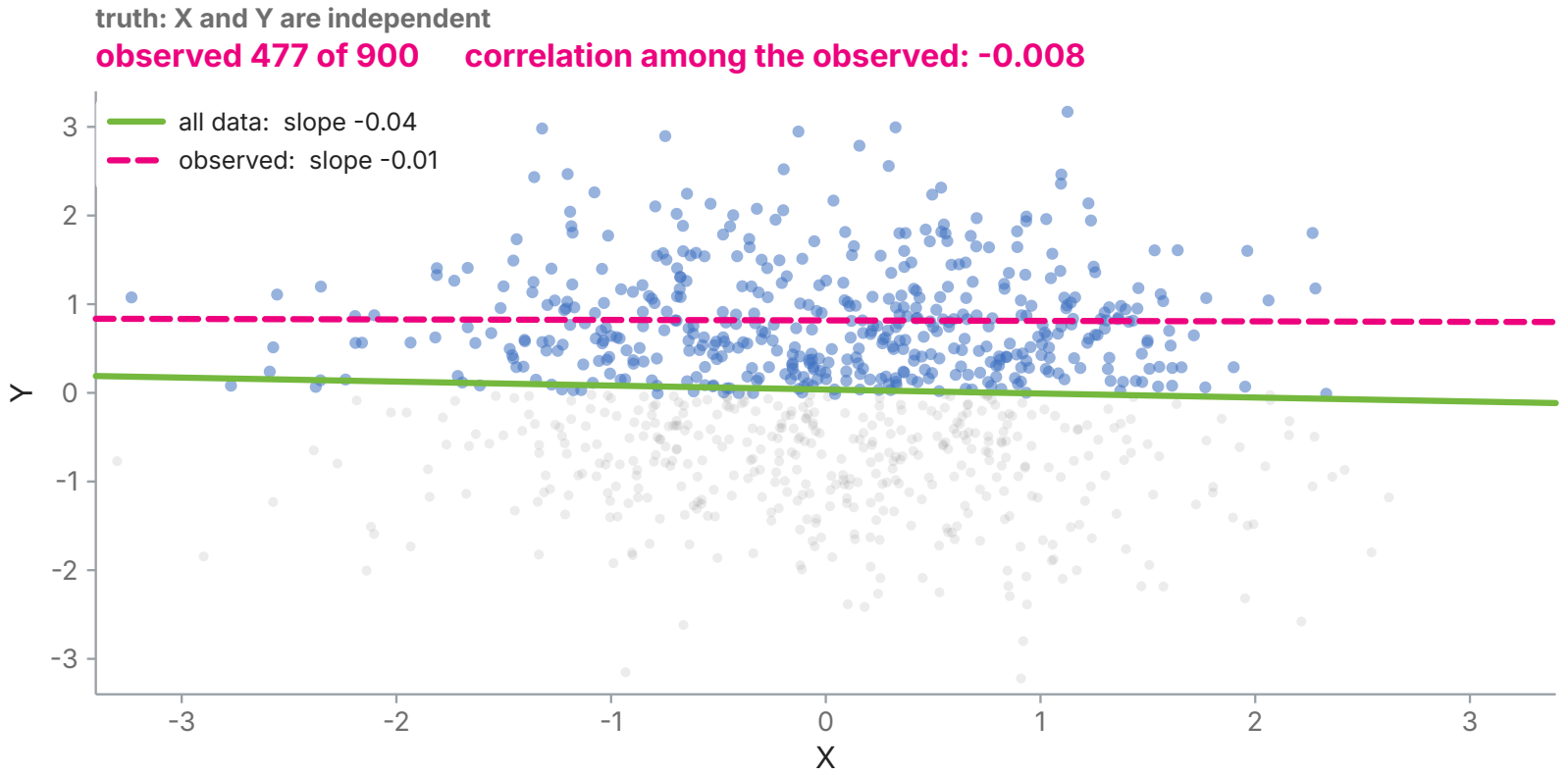
0.55

n

900

Select on Y — you keep the top slice. **Collider** — you condition on $C = X + Y + \varepsilon$, something *caused by both*. **Survivorship** — units with low Y vanish; nothing was conditioned on, the rows were never written.

Modes 1 and 3 look almost identical and are completely different operations.



Bad Controls

Conditioning can *create* dependence

X and Y independent. $C = X + Y + \varepsilon$ — caused by both. Condition on C, and X and Y become strongly **negatively** correlated.

- Among people with the same C, a high X *implies* a low Y — arithmetic, not causation
- This is the exact counterexample to the instinct: *"add more controls to be safe"*
- Controlling for occupation when estimating the gender wage gap. Controlling for anything measured **after** treatment

Set the band to full width

The induced correlation vanishes. It is the **conditioning**, not the variable, that does the damage.

Berkson, in the Wild

The same mechanism, in data nobody deliberately selected:

- Among **admitted** students, test scores and grades correlate negatively
- Among **hired** workers, credentials and ability correlate negatively
- Among **funded** startups, team quality and idea quality correlate negatively

Each admission rule is a collider: you get in by being good enough on the *sum*. Nothing causal is happening in any of them.

Survivorship is different

Funds that closed, firms that failed, workers who left the panel, countries that stopped reporting.

Nothing was conditioned on. **The rows were never written.** The picture looks the same; there is no analysis step to point at.

More Data, Same Bias

Deck 01 closed with: *“Sample size fixes noise. It never fixes bias.”*

- You are about to be handed scanner data, payroll records, platform logs, an administrative register
- $n = 2{,}000{,}000$. The instinct is that the question is settled
- It is not, and the reason is arithmetic

$$\bar{X}_n - \bar{X}_N = \underbrace{\rho_{R,X}}_{\text{data defect}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\text{sampling fraction}} \times \underbrace{\sigma_X}_{\text{variability}}$$

Only the middle term improves as n grows — and when n/N is small it is the smallest of the three.

The Big-Data Paradox, Live

How large a *random* sample would carry the same information as your huge non-random one?

$\log_{10}$ n

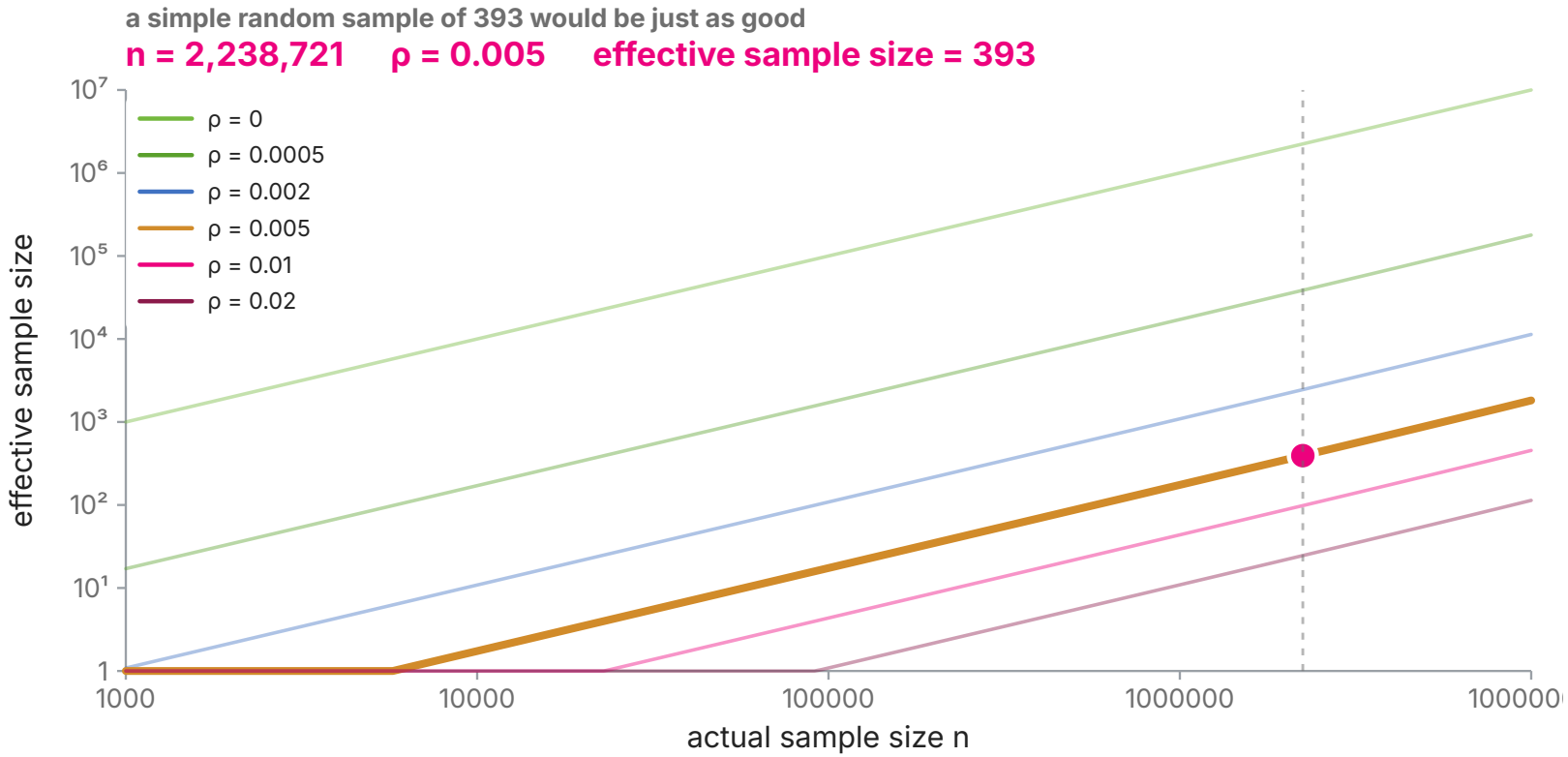
6.35

Data defect ρ

0.005

$\rho_{R,X}$ is the correlation between *being in the sample* and *the outcome*. It is far too small to detect and it dominates everything.

Set it to zero: the effective size jumps to n. Raise n with ρ fixed: almost nothing happens.



Two Point Three Million, and Worth Four Hundred

- A selection correlation of $\rho = 0.005$ — **one two-hundredth** — is undetectable by any diagnostic you would run
- It reduces 2.3 million respondents to an effective **402**
- Set $\rho = 0$ and the same n is worth its full 2.3 million

Scoreboard, row 4

Representative sample broken · the interval is **not merely wrong, it is confidently wrong**, and it gets narrower as n grows.

What to do

A small probability sample beats a huge convenience sample. Weight toward known population margins. Report who is *missing*, not just how many you have.

Regression to the Mean, Live

Select the worst performers. Treat them. Measure again. With **no treatment effect at all**.

Pre–post correlation r

0.6

Selected group

Bottom 10%

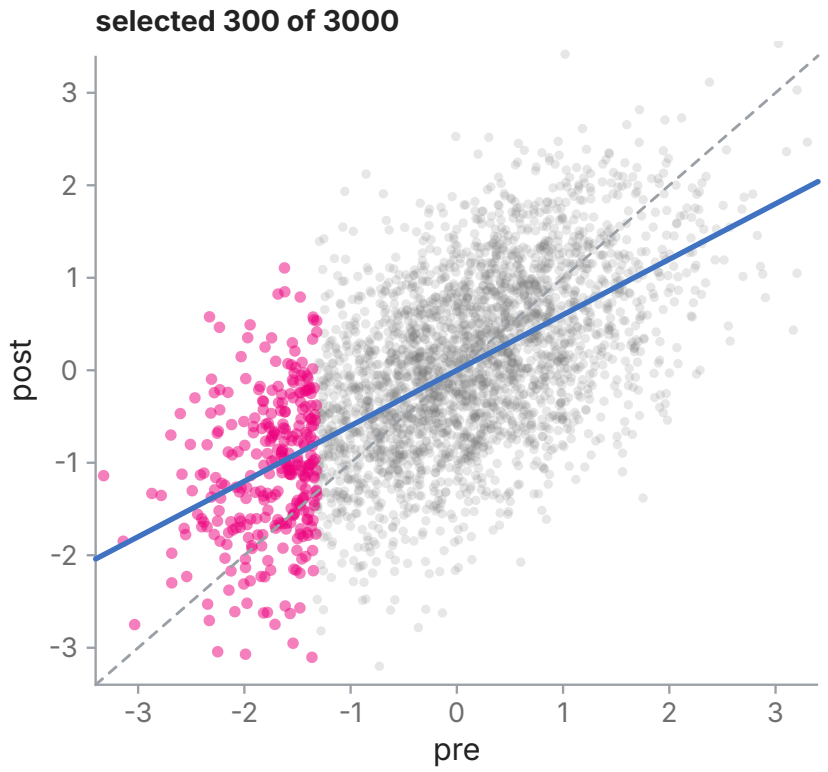
True effect τ

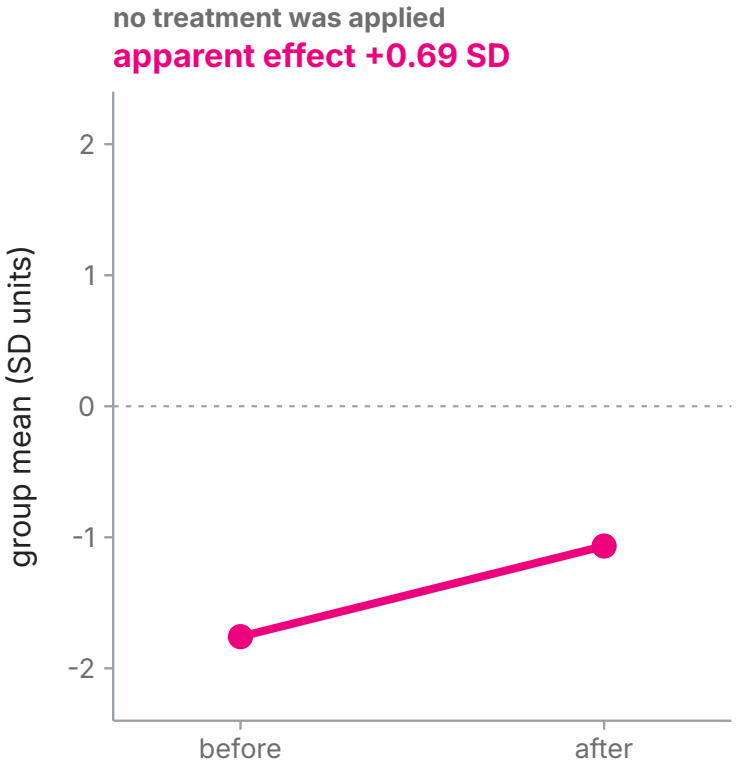
0

Show untreated control

The selected group improves because it was selected on a **noisy** measure — part of why they scored low was bad luck, and luck does not repeat.

Turn on the control group: it improves by the same amount, and the *difference* recovers τ.





Why This One Matters for Policy

With $r = 0.6$ and the bottom decile: the group moves from -1.74 to -1.06 standard deviations. An apparent improvement of **0.68 SD**, from nothing.

- Policy targets the tail **by design** — struggling schools, shrinking regions, the long-term unemployed
- Every one of those is a bottom-decile selection on a noisy measure
- The naive before/after evaluation reports the same 0.68 SD of nothing

Ashenfelter's dip

Workers enrol in training programmes exactly when their earnings have dipped. Their earnings then recover — with or without the programme.

Economics has known about this since the 1970s and still publishes it.

External Validity

An estimate is an average over **the people you studied, where you studied them, at the dose you used, in the year you did it**.

- A LATE is narrower still: an average over *compliers* — those the instrument moved
- Nobody can point to the compliers. They are defined by a counterfactual
- Microcredit in rural Bangladesh does not price credit for Austrian SMEs

No standard error speaks to this

Transportability is a scientific judgement about mechanisms and populations, not a statistical quantity. Your interval can be perfect and the extrapolation still worthless.

5 · Researcher Degrees of Freedom

The Garden of Forking Paths

You did not run twenty tests. You ran one.

But before you ran it you chose:

- the sample window · the outlier rule · the deflator · which controls · the functional form · the clustering level · how to handle missing values

The effective number of tests is not 1

Every one of those choices was defensible, and several were made *after* seeing the data. Deck 05's multiple-testing correction adjusts for the tests you ran — not for the ones you could have run.

Economics Got There First

This critique is often imported from psychology's replication crisis. It did not start there.

Leamer (1983)

"Let's Take the Con Out of Econometrics"

Extreme-bounds analysis: report the **range** of coefficients over all defensible specifications, not the one you liked.

Sala-i-Martin (1997)

"I Just Ran Two Million Regressions"

If a cross-country growth result depends on which four controls you picked, it is not a result.

The specification curve is a modern name for a forty-year-old economic idea.

The Specification Curve, Live

2^k defensible specifications, all run on the same data. Start with **no true effect**.

Binary choices k

6

True effect

0

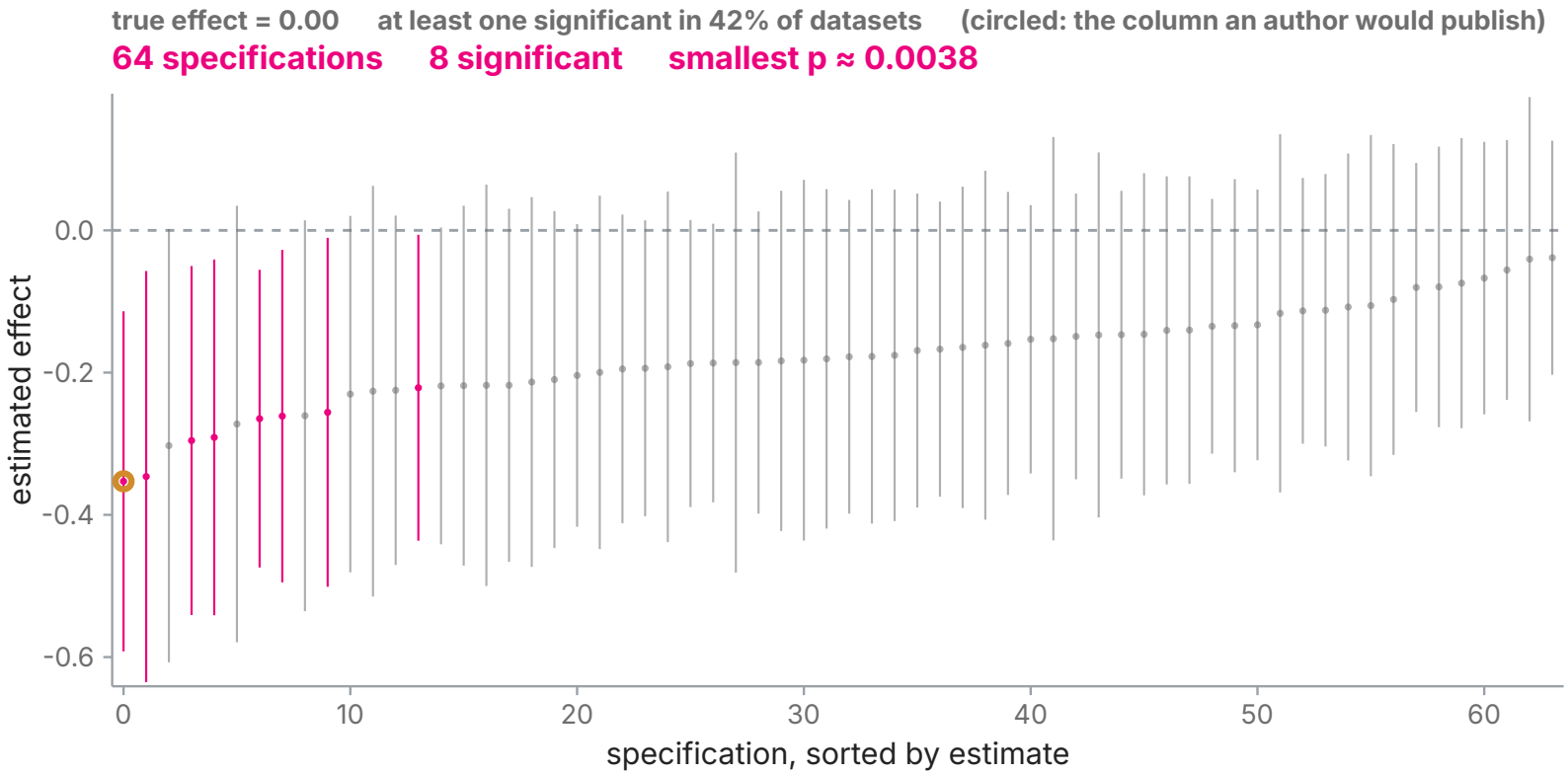
n

150

New dataset

Each specification adds or drops a control that is pure noise — every one of them defensible, none of them right.

At $k = 0$ — one specification — you get the 5% the test promises. By $k = 6$ something significant is available in roughly **four datasets in ten**, with no effect there at all.



What the Curve Is Actually For

With **no true effect at all**, how often is *some* specification significant?

Defensible choices k	Specifications	At least one significant
0	1	6% — the test working as advertised
3	8	~16%
5	32	~30%
6	64	**~42%**
7	128	~55%

- The circled specification — the one an author would report — looks like a finding
- Raise the true effect and the whole curve lifts above zero and stays there

This is the fix, not just the diagnosis

A robust result is one that survives the whole curve. **Report the curve**, not the column.

The Significance Filter, Live

What happens to the estimates that *survive* — either a significance threshold, or a "pick the best pilot" competition.

Selection rule

Keep if p < 0.05

Power

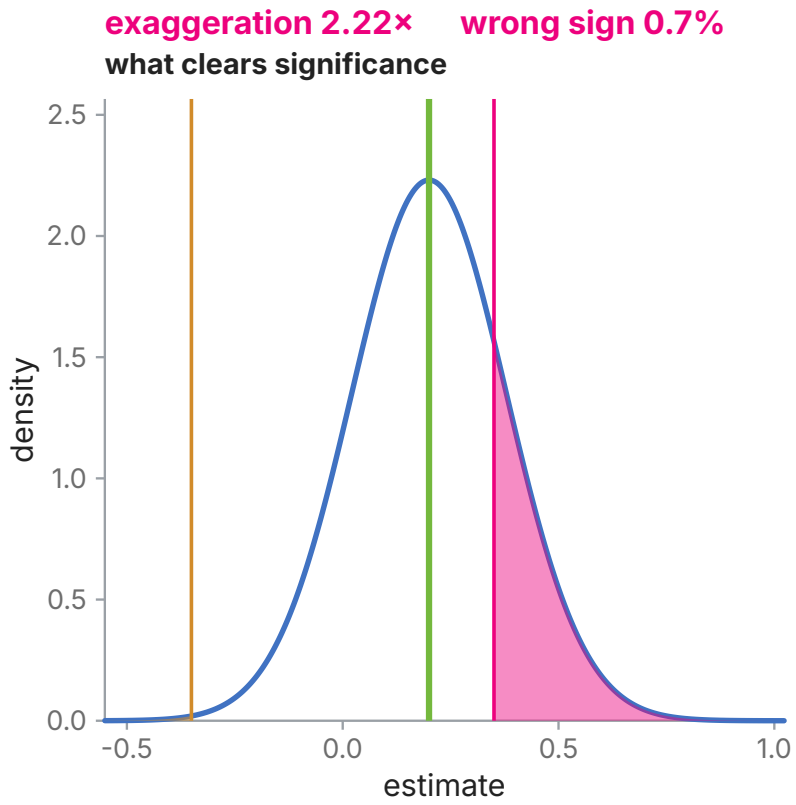
0.2

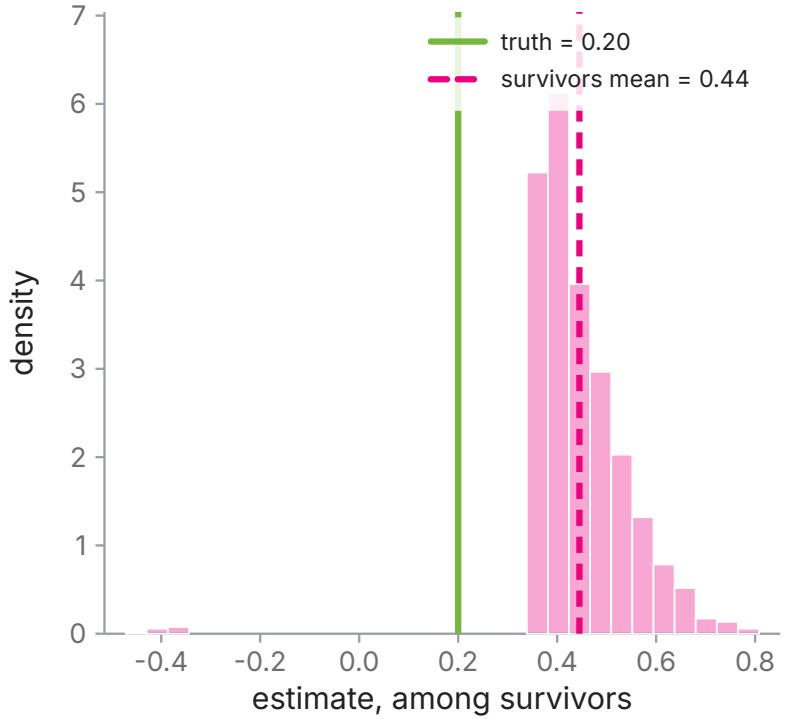
Pilots k

12

Both modes do the same thing: **condition on a maximum**.

Push power to 0.9 and the exaggeration nearly vanishes — which is the argument for planning sample size before you start.





Underpowered Studies Do Not Merely Miss

Move the power slider and watch the two readouts:

Power	Exaggeration	Wrong sign
80%	1.1×	0%
20%	2.2×	0.7%
10%	3.4×	3.5%
6%	**4.4×**	**13%**

- Well-powered studies barely suffer. Underpowered ones are not merely noisier — they are **biased upward, conditional on publication**
- Below about 10% power a meaningful share of published estimates have the wrong sign

The winner's curse version

Pick the best-performing pilot of twelve and scale it nationally. It will underperform by an amount you can compute in advance — a decision economists are actually asked to make.

Publication Bias

If only $p < 0.05$ survives the journey to print, the literature is a non-random sample of the research.

- A funnel plot with a hole in it: small studies with small effects never appear
- The meta-analytic average is biased away from zero — even when **every individual study was honest**
- Nobody committed misconduct. The filter did all of it

What helps

Pre-registration and pre-analysis plans · registered reports · publishing null results · reporting the specification curve · replication.

6 · Measurement

Attenuation Bias, Live

You do not observe X. You observe $X + u$.

Measurement noise σ_u

0.76

$\log_{10}$ n

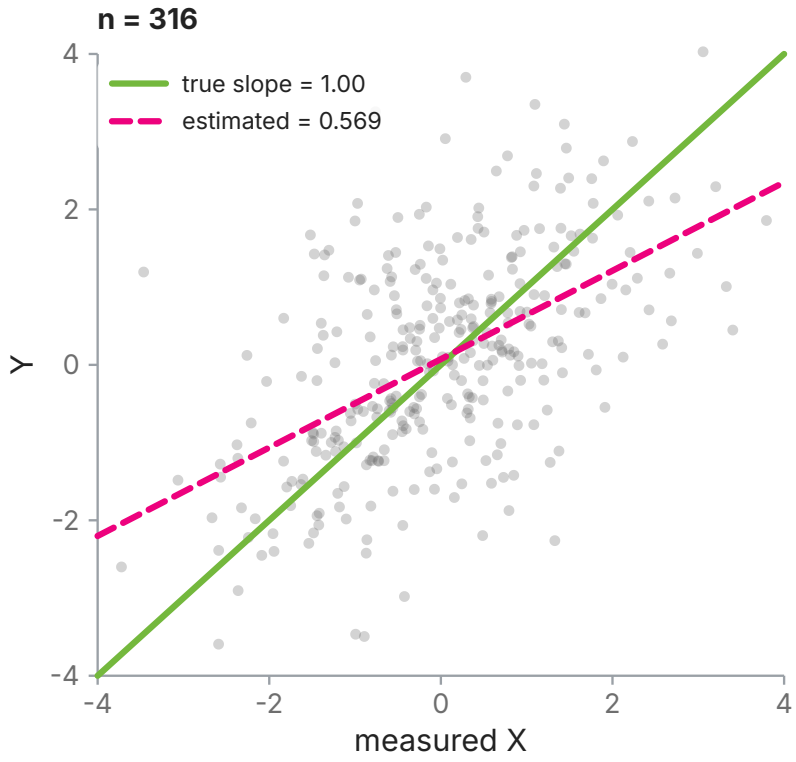
2.5

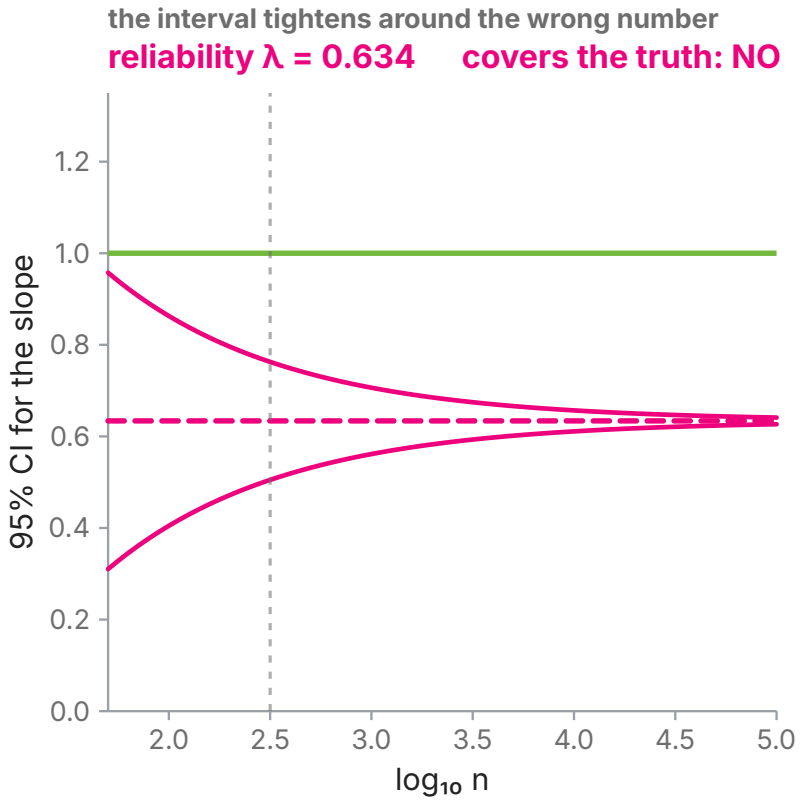
Error type

Classical (independent)

Classical error shrinks the slope by the reliability ratio $\lambda = \frac{\sigma_X^2}{\sigma_X^2+\sigma_u^2}$ — always **toward zero**.

Non-classical error, correlated with the truth, can push the estimate either way — including past the true slope.





More Data, More Confidence, No Less Wrong

- Classical measurement error shrinks the slope by $\lambda = \sigma_X^2/(\sigma_X^2 + \sigma_u^2)$
- The estimator is **consistent** — for $\beta\lambda$, not β
- As n grows the interval narrows around the wrong value and excludes the truth with rising confidence

Scoreboard, row 6

Perfect measurement broken · nominal 95% → $\longrightarrow 0\%$ as n grows.

This is the purest illustration in the course of deck 01's closing line: sample size fixes noise, and never bias.

Where the Noise Comes From

- **GDP revisions** — first releases differ from final figures by amounts comparable to the effects being studied
- **Self-reports** — income, hours, consumption, hours worked last week
- **Recall error** — anything asked about the past
- **Imputation** — owner-equivalent rent, informal-sector output
- **Seasonal adjustment** — a model, applied before you see the data

Non-classical is the normal case

Classical error is the *optimistic* assumption: independent of the truth, so the bias direction is known. Real error correlates with the truth — high earners under-report more — and then the bias can go either way.

Missing Data

	Mechanism	Can you ignore it?
MCAR	missing for reasons unrelated to anything	yes, you just have less data
MAR	missing depends on **observed** variables	yes, if you model them
MNAR	missing depends on the **unobserved** value itself	no

The distinction is untestable

Nothing in the observed data can tell MAR from MNAR — the evidence you would need is exactly what is missing. It has to be argued substantively.

Panel attrition is rarely MCAR: people who lose their job leave the survey.

7 · What to Do About It

The Scoreboard, Complete

Assumption broken	§	Nominal	Actual
Correct specification	1	95%	~88% (naive SEs)
Independence over time	3	95%	**~49%**
Independence across clusters	3	95%	**~66%**
Representative sample	4	95%	2.3M rows worth 402
One specification	5	95%	**~42%** find *something* at $k = 6$
Perfect measurement	6	95%	$\to 0\%$ as n grows

One claim, six mechanisms

Not one of these is fixed by collecting more data. Four of them get **worse**.

Simulate Your Own Estimator

Every demo in this lecture was the same thing: **a Monte Carlo of an estimator under assumptions you control**.

You can do this for your own work, and it takes an afternoon:

(1) Simulate data that looks like yours — same n, same clustering, same missingness

(2) Set the true parameter to something you choose, often **zero**

(3) Run your actual procedure, a thousand times

(4) Count how often the interval covers, and how often you reject

If it does not cover at 95%, your standard errors are wrong

This is the single most useful diagnostic in applied work, it needs no theory beyond this course, and almost nobody does it.

Design Beats Correction

Before the data

- Randomise if you possibly can
- Pre-register the analysis
- Compute power honestly, and walk away from hopeless designs
- Decide the specification in advance

After the data

- Robust and clustered standard errors as the **default**
- Placebo and falsification tests
- Bounds and sensitivity analysis
- Out-of-sample validation
- Publish code and data

The ordering is the point

No amount of post-hoc correction rescues a design that could not have answered the question. Correction is triage; design is treatment.

Report Honestly

- **Effect sizes and intervals**, not stars — deck 05 said statistical significance is not practical significance, and with administrative data that stops being a caution and becomes the operating condition
- **A range you can defend** beats a point estimate you cannot
- **The whole specification curve**, not the best column
- **Who your estimate applies to** — the population, the period, the dose
- **What is missing**, not just what you have

Key Takeaways

(1) The tools from decks 01–06 are correct. They have **preconditions**

(2) Economics is where those preconditions bind — every single one of them

(3) The shared symptom is measurable: **your nominal 95% is not 95%**

(4) More data fixes noise. It never fixes bias, and it narrows the interval around whatever you are converging to

(5) Simulation tells you which of these applies to *your* data

The closing frame

Statistics gives you valid answers to precisely posed questions under stated assumptions.

The assumptions are the economics.

Getting them right is not a technicality that precedes the real work — it *is* the real work.