

Lecture 3 - Design: Identification and Beyond

Topics in Econometrics - M2 ENS Lyon

Vincent Bagilet

2026-09-23

Your favorite paper

- Anyone wants to share theirs?
- What is it **about**?
- What did you particularly **like** about this paper?
- Anything else worth **sharing** with the others?

Steps of an Econometrics Analysis

- **Design:** decisions of data collection and measurement
 - eg, decisions related to sample size and ensuring exogeneity of the treatment
- **Modeling:** define statistical models
 - In between design and analysis
- **Analysis:** estimation and questions of statistical inference
 - eg standard errors, hypothesis tests, and estimator properties

Design in Economics

- Typically in economics, we are interested in estimating **causal** effects
- In (non-experimental) economics, design presented in this lexicographic order:
 1. Identification
 2. Unbiasedness
 3. Minimum variance
 4. Robustness to misspecification somewhere in the mix
- Design includes **identification but not only**
- These steps **interact** with one another

Design and Identification

Identification strategy

- It describes how to use observational data to recover the effect of interest
 - It defines **a source of identifying variation** in the treatment
 - It specifies the **assumptions** under which this variation identifies the causal effect
- Types of experiments:
 - A **true experiment**: treatment is randomly assigned by the researcher
 - A **natural experiment**: treatment is *plausibly* randomly assigned by “nature”
 - A **quasi-experiment**: treatment is close-to (but not) randomly assigned

Potential outcomes framework

- Let's denote $D_i \in \{0, 1\}$, the treatment status, Y_i , the realized outcome, Y^0 and Y^1 the potential outcomes

Individual Treatment Effects (TEs)	$Y_i^1 - Y_i^0, \forall i$	<i>What we would ideally estimate</i>
Average Treatment Effects (ATE)	$\mathbb{E}[Y_i^1 - Y_i^0]$	<i>What we reasonably want to estimate</i>
Average Treatment Effects on the Treated (ATT)	$\mathbb{E}[Y_i^1 - Y_i^0 D_i = 1]$	<i>What we reasonably want to estimate</i>
Conditional Average Treatment Effects (CATE)	$\mathbb{E}[Y_i^1 - Y_i^0 X_i]$	<i>What we reasonably want to estimate</i>
Difference in average observed outcomes	$\mathbb{E}[Y_i D_i = 1, X_i] - \mathbb{E}[Y_i D_i = 0, X_i]$	<i>What we can estimate</i>

SUTVA

- Stable unit treatment value assumption (SUTVA):
 - The potential outcome of one individual **does not depend on the treatment status of other individuals**
- Each unit has only 2 potential outcomes: Y_i^0, Y_i^1
- Assumes no spillover effects
- Assumes no general equilibrium effects
- Often not realistic in economics

Selection bias

$$\underbrace{\mathbb{E}[Y_i | D_i = 1, X_i]} - \underbrace{\mathbb{E}[Y_i | D_i = 0, X_i]} =$$

Difference in average observed outcomes

$$\underbrace{\mathbb{E}[Y_i^1 - Y_i^0 | D_i = 1, X_i]} + \underbrace{\mathbb{E}[Y_i^0 | D_i = 1, X_i] - \mathbb{E}[Y_i^0 | D_i = 0, X_i]}$$

CATT

Selection Bias

- Goal: eliminate this selection bias to be able to say something about the quantity of interest (the ATT)
- **Selection bias**: average difference in Y_i^0 between the treated and untreated
- Assumptions regarding the assignment mechanism (**identifying assumptions**) can help eliminate it

Assumed assignment mechanisms

- **Random assignment** (eg experiments)
 - Treatment independent of potential outcomes $\Rightarrow$ no selection bias in expectation
 - It is the *Independence Assumption (IA)*: $(Y_i^0, Y_i^1) \perp D_i$
- **Selection on observables**
 - Random assignment conditional on some pre-treatment characteristic X
 - It is the *Conditional Independence Assumption (CIA)*: $(Y_0, Y_1) \perp D_i | X_i$
 - Compare outcomes within each stratum of X_i
- **Selection on unobservables**
 - Need other identification strategies to eliminate selection bias
 - Will still assume some other independence assumptions

Identifying assumptions

- Can recover an unbiased estimator of a causal effect iff an identifying/independence assumption holds:
 - **IA**: $(Y_i^0, Y_i^1) \perp D_i \Rightarrow$ **can estimate the ATT**
 - No IA but **CIA**: $(Y_i^0, Y_i^1) \perp D_i | X_i \Rightarrow$ **can estimate the CATT** (ATT in each stratum)
 - No CIA but $\exists$ a relevant **instrument** Z_i that is an exogenous source of variation in D_i :
 $(Y_i^0, Y_i^1) \perp Z_i | X_i, Z_i \perp D_i | X_i \Rightarrow$ **can estimate a LATE**
- We always need an identification strategy that convinces us that an IA holds

Summary

- Goal: identifying **causal** effects
- *ie* a difference between two potential outcomes
- But, we cannot observe the potential outcomes
- We only see the differences in observed outcomes
- If (C)IA holds, we can estimate an unbiased ATT
 - Randomized Control Trial (RCT), the gold standard
- But (C)IA rarely holds $\Rightarrow$ need an **identification strategy** to eliminate selection bias

Common identification methods

- **Randomized experiments (RCT)**
 - Randomization of treatment D
- **Difference-in-differences (DiD), event studies, synthetic control methods (SCM)**
 - Research designs that assume or construct parallel trends
- **Instrumental variables (IV) or regression discontinuity (RD)**
 - An instrument or discontinuity induces exogenous variation in treatment status
- **Matching estimators:**
 - Strategies solely based on matching are much less credible
 - But matching can complement natural or quasi-experimental design

Abadie et al. (2020)

- Can anyone **summarize** the idea?
- What do you get out of it?
- What did you think of it?

Design-based identification

- Randomness comes from treatment assignment
- Identification rest on how this treatment was assigned
- Credibility comes from knowledge of the assignment process
- Other identification “types”:
 - **Model-based**: randomness comes from how outcomes are generated (from the errors, population and assignment fixed)
 - **Structural**: based on economic theory, vastly used in industrial organisation

Design Beyond Identification: Low Power

The importance of design

- We want to **have an accurate measure of the quantity of interest**
- → **research design**:
 - Data collection and measurement choices to allow us to draw credible conclusions
 - Includes the causal identification strategy, but not only
- A great strategy useless if a poor design prevents us **detecting** the effect
- For instance if statistical power is low

Statistical power

- Power is a key **implication of design choices**
- *Definition:*
 - Probability of rejecting the null (often of no effect) when it is false:

$$\text{Power} = 1 - \text{rate of Type II error}$$

- Roughly the probability of detecting an effect when there is one
- Power is a function of design: poor designs can lead to low statistical power

Why is low power problematic?

- We want to be able to detect an effect if there is one (that is large enough to be relevant)
- Because costly to run a study for “nothing”
- In RCT, typical threshold for power: 80%
- In observational settings, why not run a study with say 20% power?
- Because low statistical power $\Rightarrow$ exaggeration

Illustration of the exaggeration and power issues

The effect found in the initial study (in red)

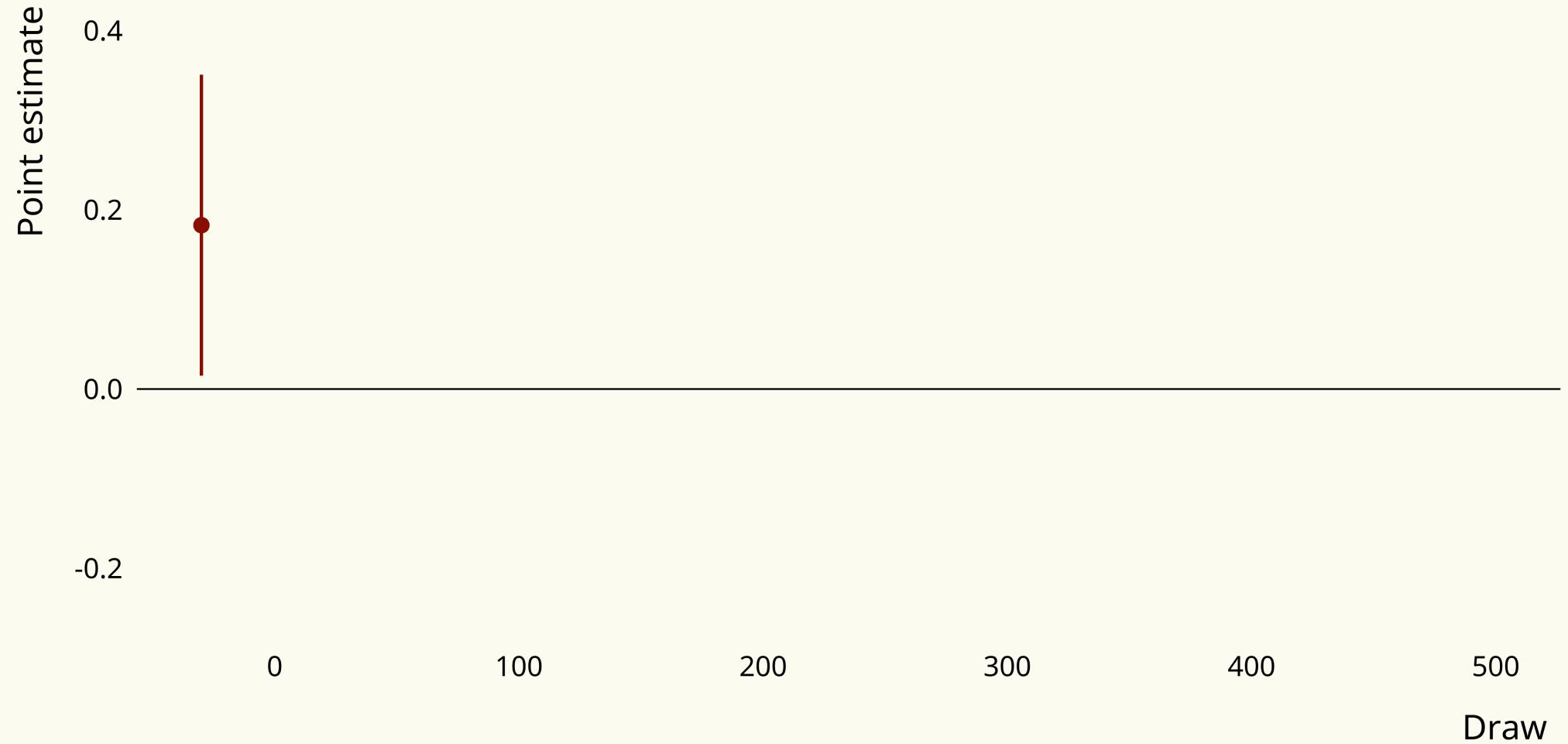


Illustration of the exaggeration and power issues

The effect found in the replication (in blue)

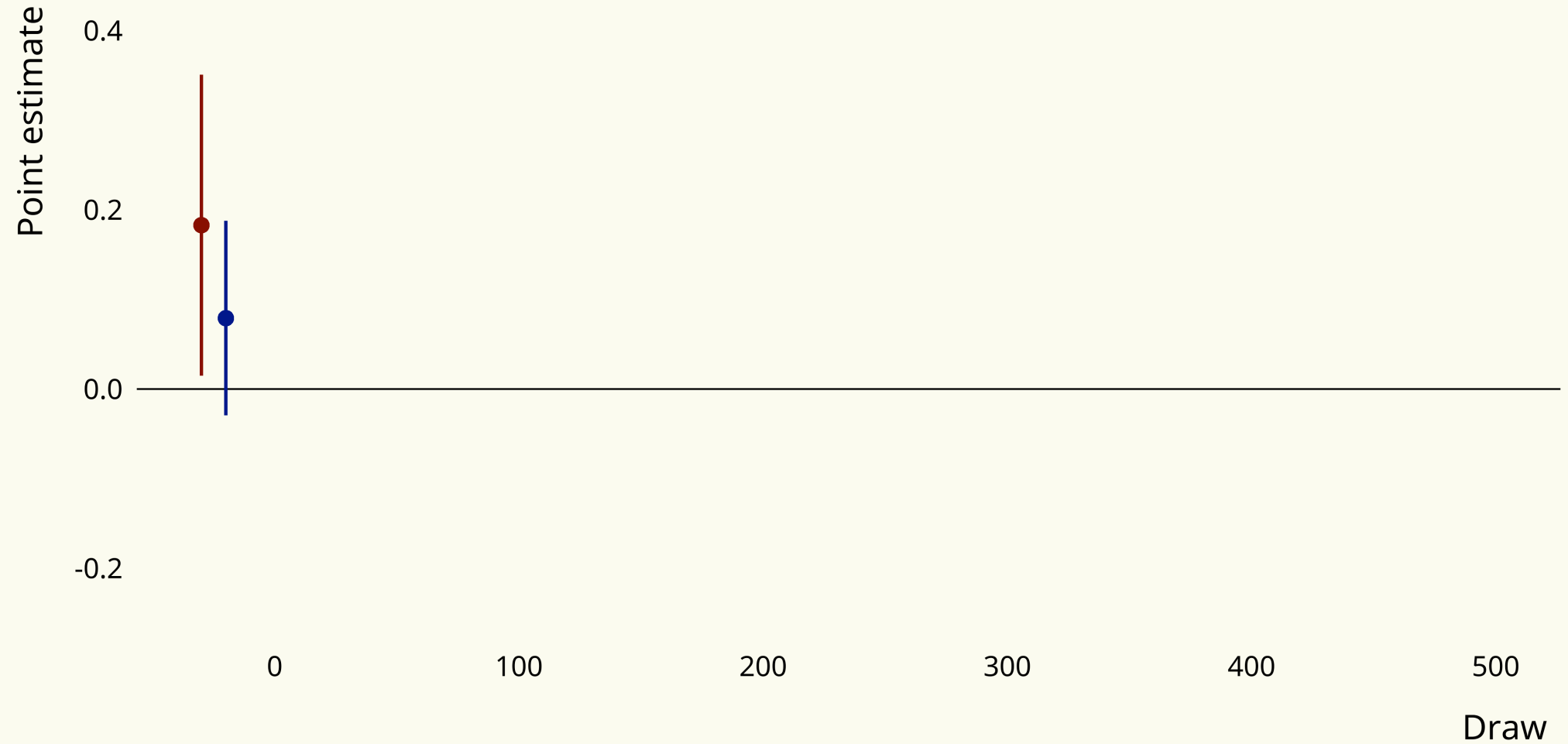


Illustration of the exaggeration and power issues

The effect found in the replication but assuming the initial design (in gray)

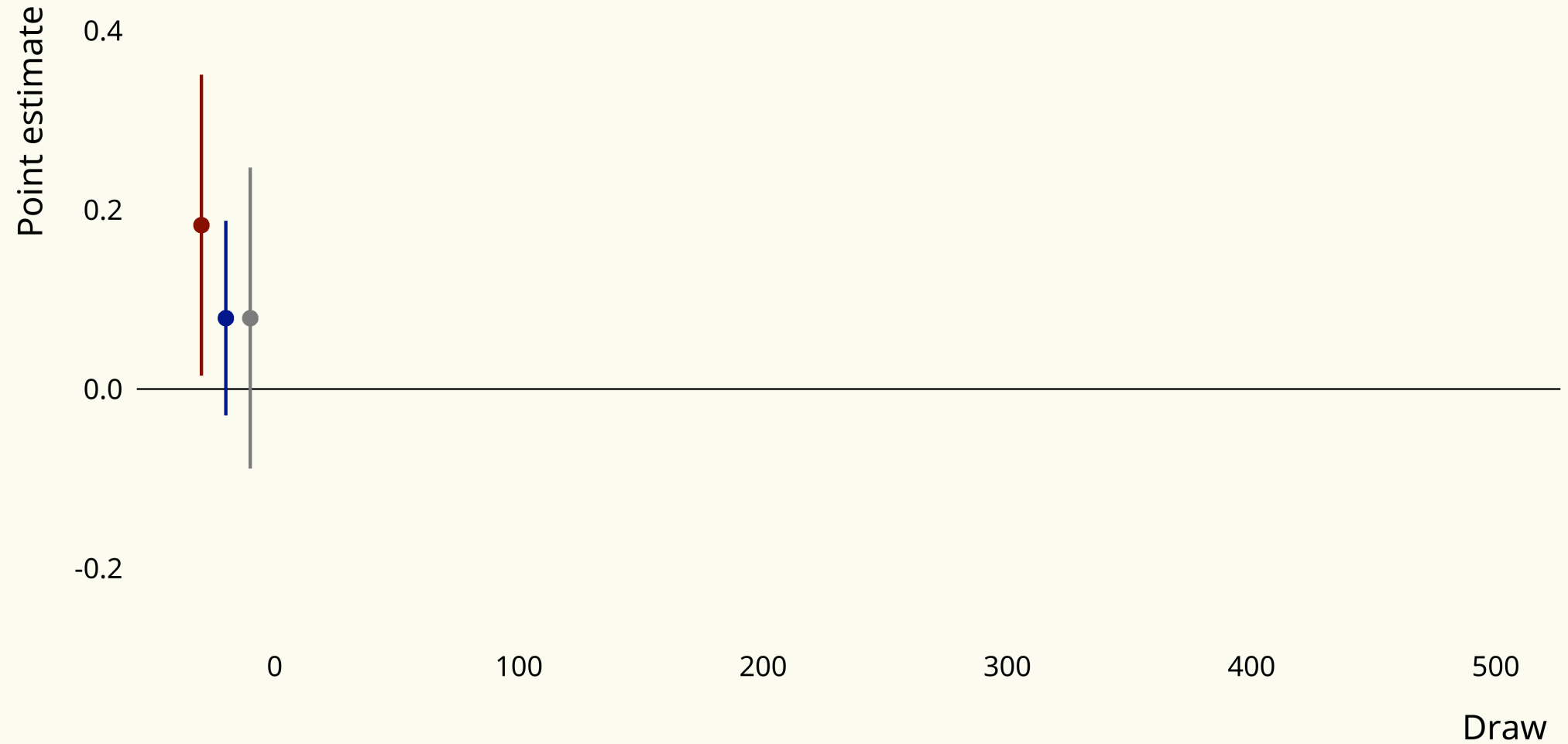


Illustration of the exaggeration and power issues

500 draws of an estimator $\sim N(\text{Effect size in replication, std err in original study})$

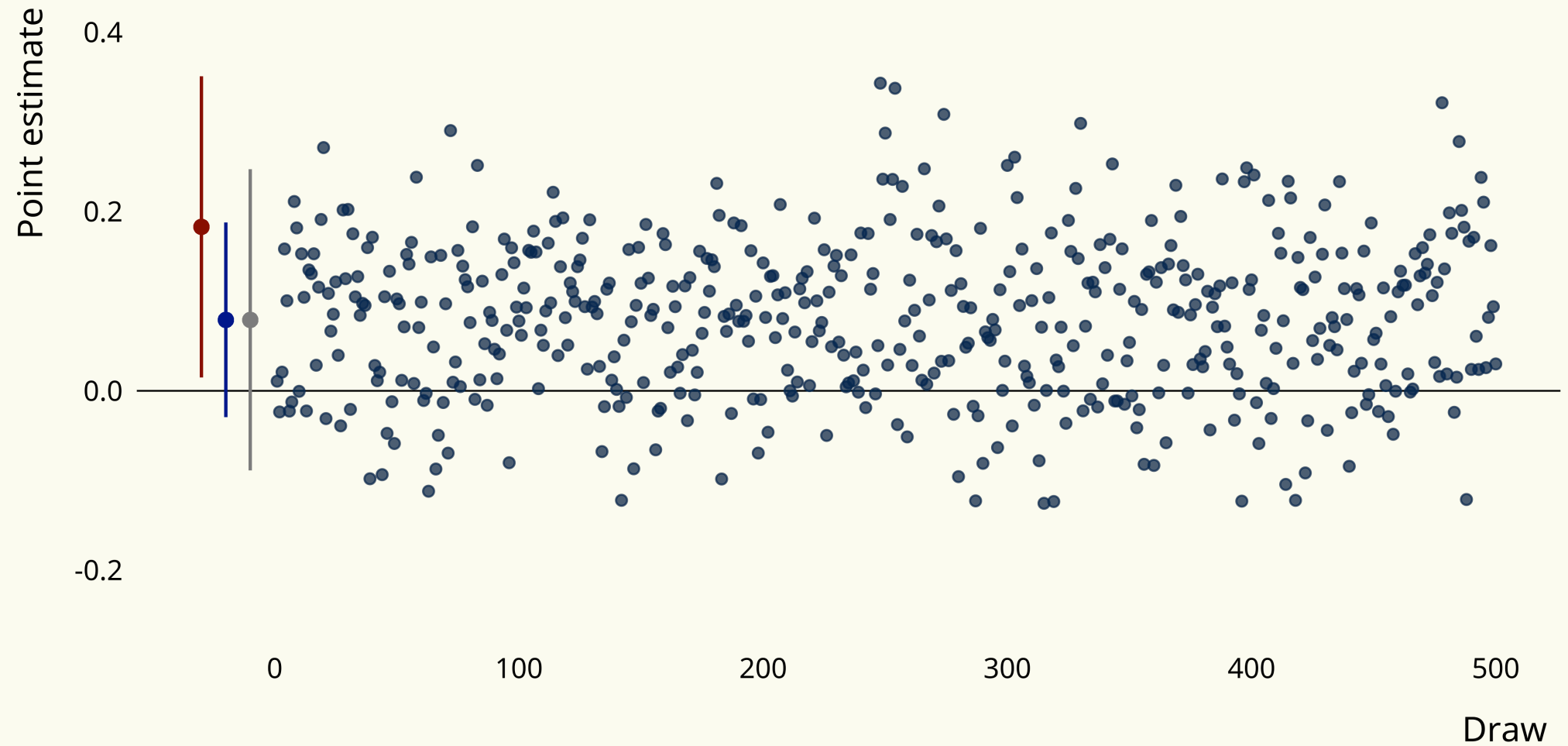
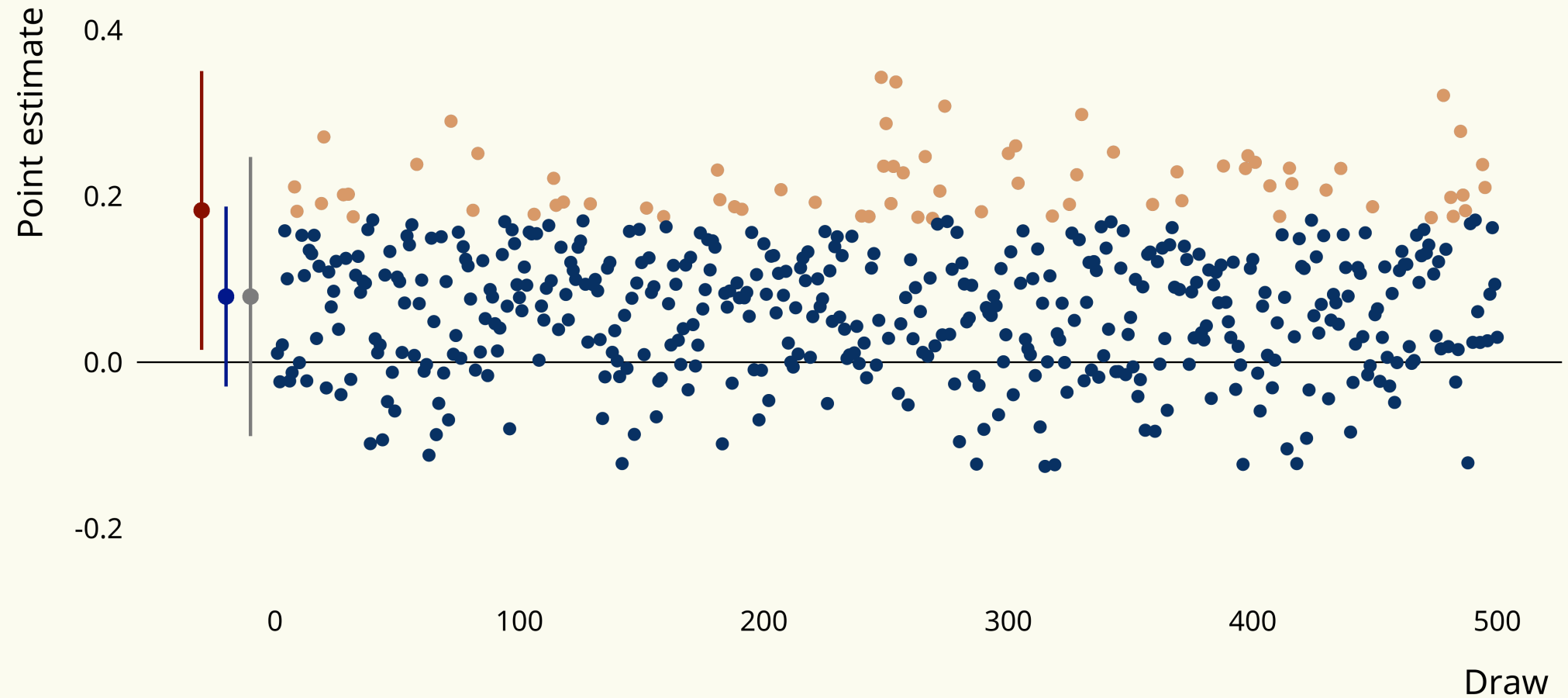


Illustration of the exaggeration and power issues

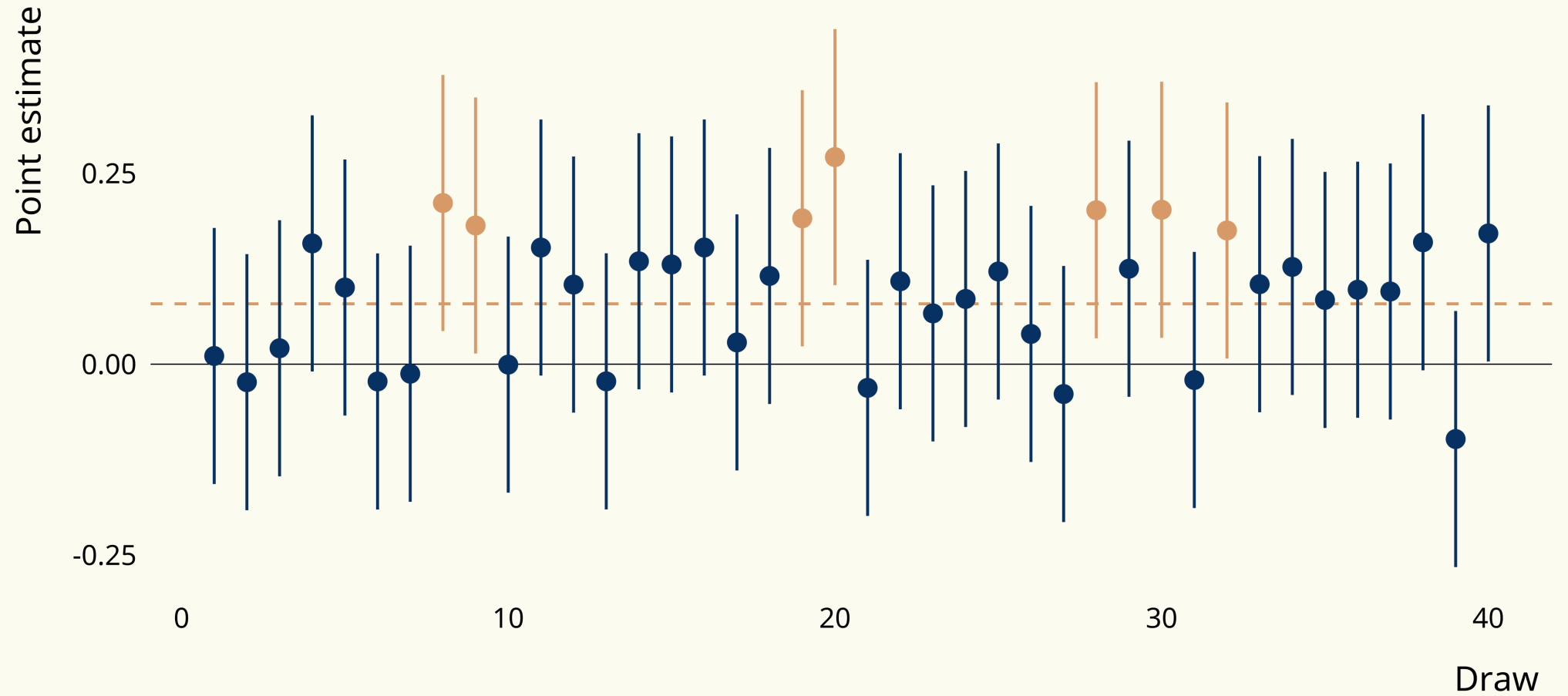
500 draws of an estimator $\sim N(\text{Effect size in replication, std err in original study})$

significant • Non significant • Significant



Draws from the distribution of an estimator

● Non significant ● Significant



The dashed line represents the "true" effect

Exaggeration: definition and main drivers

- Definition:

$$E = \frac{\mathbb{E}[|\hat{\beta}| \text{ signif}]}{|\beta_1|} = \frac{\mathbb{E}[|\hat{\beta}| \beta_1, \sigma, |\hat{\beta}| > z_\alpha \sigma]}{|\beta_1|}$$

- Exaggeration ↘ with statistical power and thus:
 - ↘ with precision
 - ↘ with effect size
- **When power is low, significant estimates from an unbiased estimator ALWAYS exaggerate the true effect**
- There are also less straightforward drivers (*we are going to discuss them later today*)

Two ingredients for exaggeration

- **Significant results are favored**
 - Evidence of a significance filter in economics
 - (Rosenthal 1979; Andrews and Kasy 2019; Abadie 2020; Brodeur et al. 2016; Brodeur, Cook, and Heyes 2020)
- **Low statistical power**
 - Median power in economics: 18%
 - (Ioannidis, Stanley, and Doucouliagos 2017; Ferraro and Shukla 2020)

Why are significant results favored?

- Editorial process favors significant results for publication
- In a way, that makes sense if a non-significant result reflects a poor research question $\Rightarrow$ importance of theory
- But, might also be that the effect is **difficult to capture**
- **File drawer problem**: tend to give up projects more when results are non-significant (put them away in a drawer)
- **Forking paths**: we make many choices when implementing a study and they may be more likely to lead to a significant outcome

Exaggeration matters in actual settings

- In economics, nearly **80% of estimates are exaggerated by a factor of 2** (Ioannidis, Stanley, and Doucouliagos 2017)
- Not all designs suffer from exaggeration
- But exaggeration is likely substantial in many studies



Solution, straightforward?

- Have a large enough **sample size** and we're good?
- Not so simple!
- Other aspects than sample size affect power (and interact multiplicatively):
 - Effect size
 - Proportion of treated
 - Number of shocks
 - Measurement error
 - Strength of the instrument
 - Count of the outcome

Multiple Goals of a Study and Design Implications

ATE But Not Only

- Often, goal of an econometrics study: estimate the ATE (*Does the treatment work?*)
- But also, *where and when does it work?*:
 - Capture **heterogeneity**: treatment effect varies across time and individuals
 - Often consider the effect on **multiple outcomes**
 - **Extrapolate**

Implications of Multiple Goals

- They have **intertwined implications** for how we approach design
- Not possible to have high power for everything
- Goals can be **competing**
- Can take action at the design stage, acknowledging these multiple goals

Heterogeneity

- Treatment effect rarely homogeneous
- The phrase “**Average** Treatment Effect” implicitly acknowledges this
- Variation across individuals, time, space, etc
- There are therefore potential confounders:
 - Need to adjust for such variables
 - Measure them

Heterogeneity

Interactions

- An usual approach to account for heterogeneity is to use interactions
- To measure interactions, we **need 16 times the sample size**:
 - The estimates has twice the s.e. of the main effect
 - Reasonable to assume that interaction have half the magnitude of the main effect
 - Thus Signal to Noise Ratio ($SNR = \frac{\text{True effect}}{\text{s.e.}}$) is 4 times smaller for interaction
 - Thus need $4^2 = 16$ times the sample size

Heterogeneity

Two-Ways Fixed-Effects (TWFE)

- Issues:
 - When treatment effect heterogenous (in time or across groups)
 - Treated units in the control group
 - Negative weights
- The literature addressed it as a analysis problem: proposed alternative estimators
- But can see it as **non-modeled heterogeneity**

Multiple Outcomes

- Rough approximation of the median number of estimates per paper in top economics journals: 19
- Bonferroni correction:
 - Change the significance level to $\frac{\alpha}{\text{Number of hypotheses tested}}$
- Underlines that need more power $\Rightarrow$ need to take that into account

Extrapolation

- External validity
- When increase the sample size, often changes the underlying estimand
 - eg, increasing sample size by increasing the time frame
 - or the spatial frame
- Increasing sample size not always a silver bullet

Improving and Assessing Design

Structural solutions

- Without publication bias this issue disappears
- Abandoning the 5% significance threshold
- Interpretation of CI's width to embrace uncertainty
- Replication of studies with similar designs

Improving design in individual studies

- **Increased sample size**
 - Both nominal and effective
- **Increased effect size**
 - Focus on units with the largest effect
 - Increase take-up of the treatment
- **Decreased inferential uncertainty**
 - More pre-treatment information
 - Better measurement of outcomes
- **Weave empirical models with substantive theory**
 - Adjust the research question
 - Measure intermediate outcomes

Assessing design

- Use simple **design calculations**
 - *Will my design allow me to detect an effect of magnitude m ?*
- **Simulations** (you now got that hopefully)
 - *Same + what happen if some of my hypotheses do not hold?*
- **Retrodesign calculations**
 - *Would my design allow me to detect a smaller effect than the one I got?*

Design calculations

- Goal: choose a design that would yield an adequate statistical power
- Compute the expected power, in this setting, as a function of design and in particular sample size
- Find the necessary sample size
- Before implementing the analysis
- Common practice in experimental economics, much less in observational settings

Ingredients for design calculations

- Statistical power is a function of true effect size and s.e. of the estimator
 - Strictly increasing with true effect sizes
 - Strictly decreasing with s.e. of the estimator
 - Slightly complex closed form
- Need to **hypothesize a s.e. and a true effect size**

Hypothesizing a standard error

- s.e. unknown before the analysis
- Basically boil the analysis down to a difference of average outcome between treatment and controls

$$se_{\bar{y}_t - \bar{y}_c} = \sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}$$

- σ_T^2 and σ_C^2 variance of the outcome for the treatment and control group respectively (after partialing out controls)
- Assuming $\sigma_T^2 = \sigma_C^2 = \sigma^2$ and for $p_T = \frac{n_T}{n}$, this simplifies to $se_{\bar{y}_t - \bar{y}_c} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{1}{p_T(1-p_T)}}$

Hypothesizing effect sizes

1. Consider the proportion of affected individuals
2. Consider a **range of effects** (make several assumptions)
 - Derived from the literature
 - Based on theory
 - Consider what could be reasonable deviations from these effects
3. Multiply the fraction of non-zero effect with the hypothesize effects
 - Help think about reasonable effect sizes and ways to focus on larger effects or reduce s.e.

Retrodesign calculations

- Once an estimate has been obtained
- Ask the question **would my design allow me to detect a smaller effect** (of magnitude m)?
- Need the standard error of your estimate and an hypothetical true effect size (m)
- One line of `r` code: `retrodesign::retrodesign(m, se)`
- Run it for a range of values

Calibrating simulations

Why calibrating?

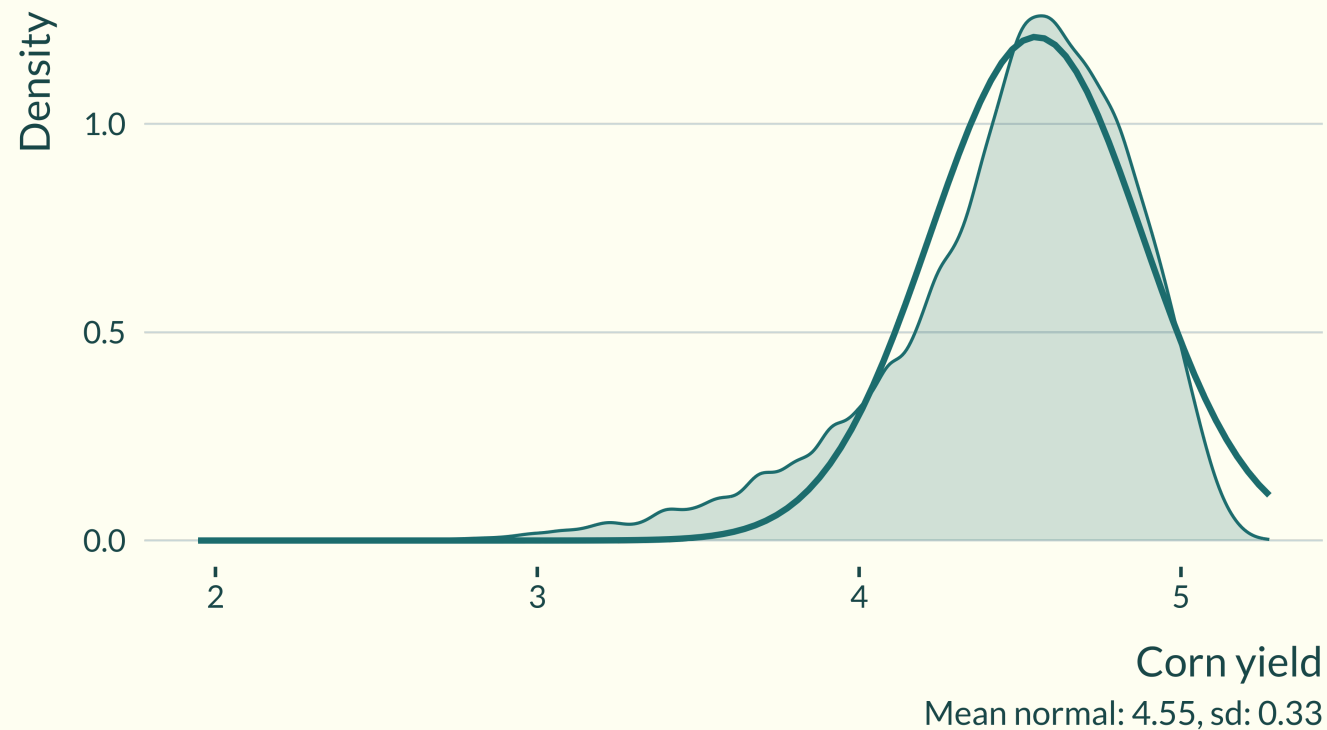
- So far, we considered very simple simulations, with “naive” distributions
- Calibrating can help make simulations more realistic
- But simulations will never be truly realistic
- Yet can still allow to run some sort of **robustness check** on the ability of your design to retrieve the effects of interest
- Also allows you to **think** about the DGP, your identification strategy, and so on

Fake data simulations

Distributions of the variables

- Emulate the distribution of variables in existing data sets

Distribution of corn yields
Tentative normal fit



Fake data simulations

Relationships between variables

- Read the literature
- Get a sense of **typical effect sizes and of relationships** between variables
- Make assumptions on those relationships. Acknowledge them.
- Complexify later if needed. You choose when you stop.
-  Varying parameters values might change the distribution of some “variables”
 - eg of the error term
 - Difficult to work *ceteris paribus*

Real data simulations

General approach

- Start from an existing data set
- Not yours. At least not the subset you are interested in
- Try to pick a subset where there is not already a treatment effect
- Define a treatment allocation mechanism
- **Add an artificial treatment effect** to the outcome variable in your initial data set, eg

$$Y_i(1) = Y_i(0) + \beta_i T_i$$

- Run your analysis and try to recover it

Real data simulations

Complexifying

- There is only one artificial aspect in such simulations: the treatment
- We can play on only 2 components:
 - **Who is treated?** *Treatment allocation*
 - Everyone
 - Only a subset of the population
 - **How?** *Treatment effect*
 - Homogenous
 - Heterogenous but random
 - Some specific correlation structure

Summary

Take-away messages

- **Design matters:**
 - In terms of identification
 - And beyond
 - Even after a significant estimate has been obtained
- When power is low, significant estimates from an unbiased estimator are always far from the true effect
- Might have important **implications for policy making**



References

- Abadie, Alberto. 2020. "Statistical Nonsignificance in Empirical Economics." *American Economic Review: Insights* 2 (2): 193–208. <https://doi.org/10.1257/aeri.20190252>.
- Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2020. "Sampling-Based Versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88 (1): 265–96. <https://doi.org/10.3982/ECTA12675>.
- Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–94. <https://doi.org/10.1257/aer.20180310>.
- Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–60. <https://doi.org/10.1257/aer.20190687>.
- Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32. <https://doi.org/10.1257/app.20150044>.
- Ferraro, Paul J., and Pallavi Shukla. 2020. "Feature—Is a Replicability Crisis on the Horizon for Environmental and Resource Economics?" *Review of Environmental Economics and Policy* 14 (2): 339–51. <https://doi.org/10.1093/reep/reaa011>.
- Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos. 2017. "The Power of Bias in Economics Research." *The Economic Journal* 127 (605): F236–65. <https://doi.org/10.1111/eoj.12461>.
- Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin* 86 (3): 638–41. <https://doi.org/10.1037/0033-2909.86.3.638>.

Thank you!