Lecture 3 - Design Beyond Identification

Topics in Econometrics

Vincent Bagilet

2025-09-23

Housekeeping

- **Replication games**:
  - October 9, all day, mandatory
  - Make groups and register, quickly
  - Pre-game meeting at 1pm to explain how it will take place
  - I will grade your assignments
- **Project proposal**:
  - Groups?
  - Thought about your subject?

Summary from last week(s)

- *Goal of the class*: develop a better understanding and **intuition** of how applied econometric analyses work **under the hood**
- Last week, learned how to implement simulations:
  - To understand econometric concepts
  - To design a study
  - Run tests and checks
  - Use as a rhetorical tool

Steps of the simulation approach

1. Define a DGP and the distribution of variables
2. Set parameters values (`baseline_param`)
3. Generate a data set (`generate_data()`)
4. Estimate the effect in the generated data set (`run_estim()`)
5. Repeat many times (`compute_sim()` and `pmap()`)
6. Compute the measure of interest
7. Change parameters values (potentially)
8. Complexify the DGP
9. Repeat

Design Matters

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

- 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

The importance of design

- We want to have an accurate measure of the quantity of interest
- For that, need to have a causal identification strategy
- But useless if the design is poor in other dimensions and prevents us from even detecting the effect
- Statistical power will be central here

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

Low power and exaggeration

Illustration of the exaggeration and power issues

The effect found in the initial study (in red)

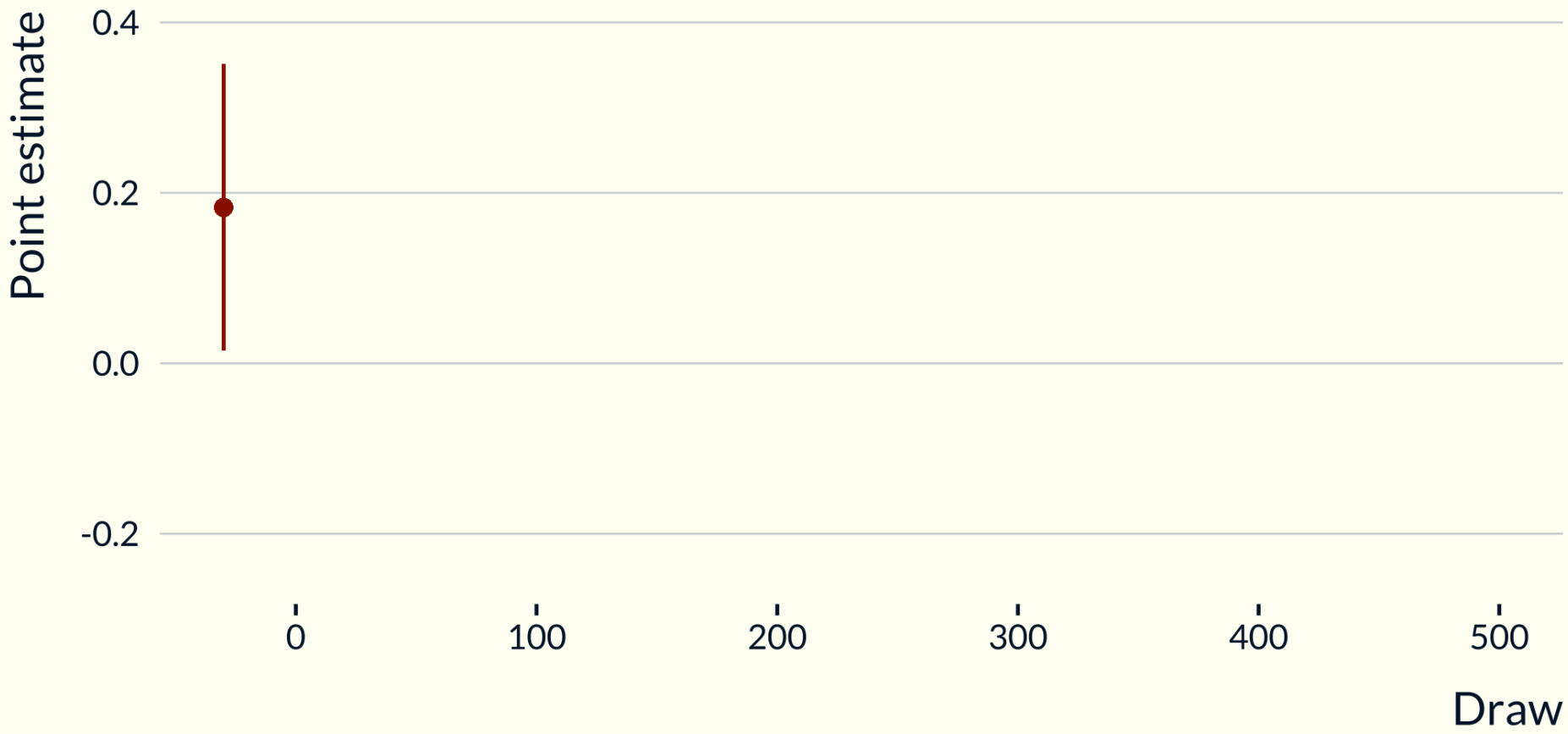


Illustration of the exaggeration and power issues

The effect found in the replication (in blue)

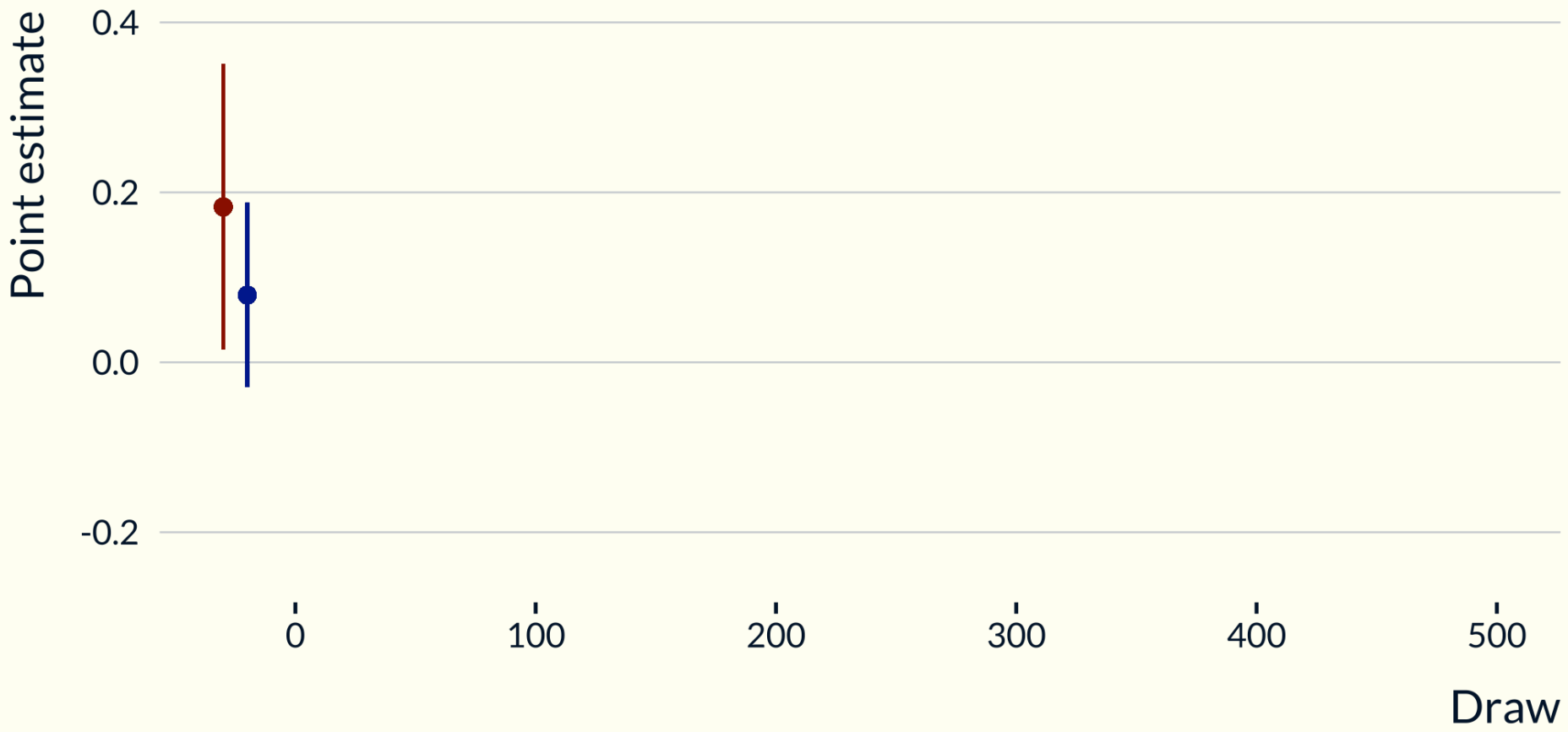


Illustration of the exaggeration and power issues

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

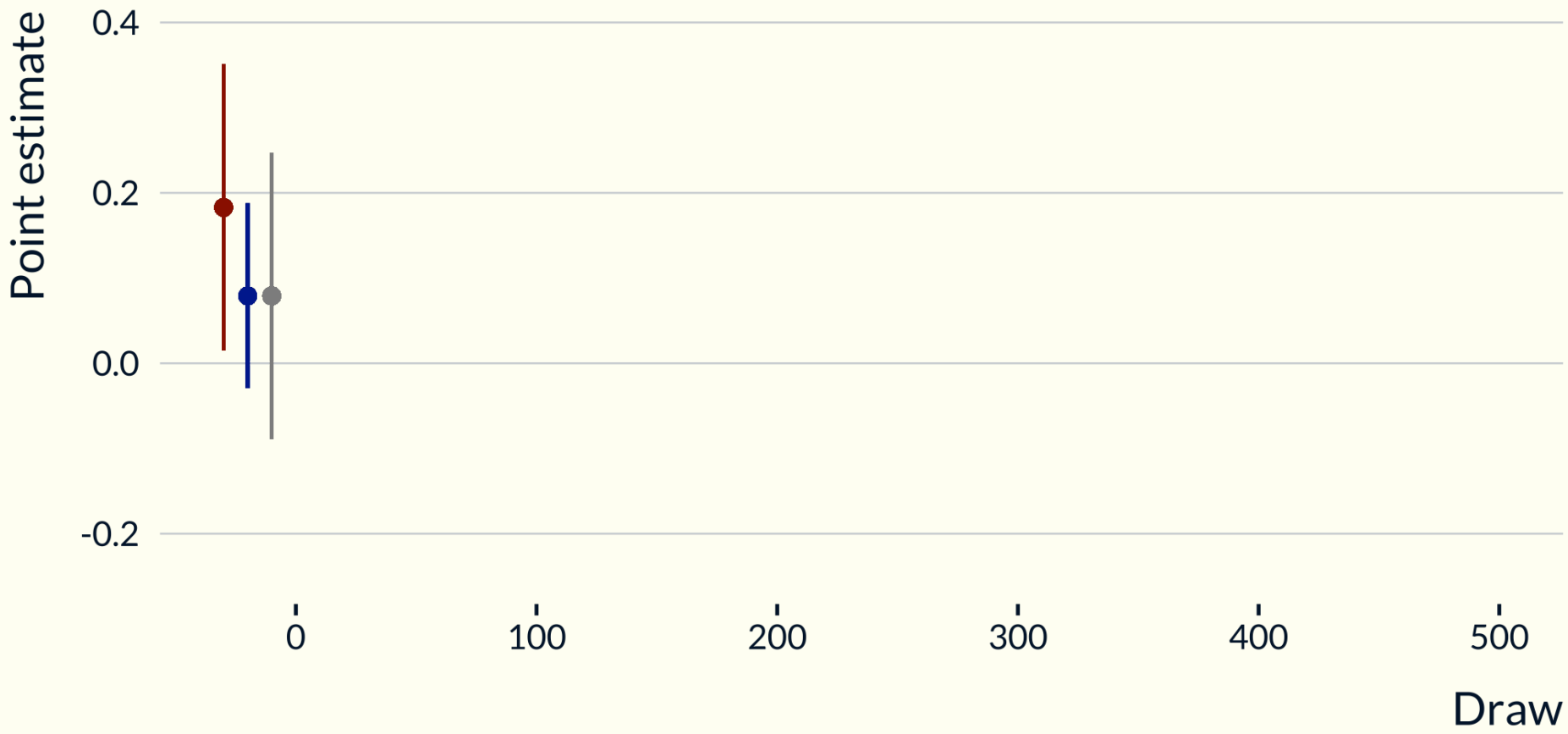


Illustration of the exaggeration and power issues

500 draws of an estimator ~ N(Effect size in replication, std err in original study)

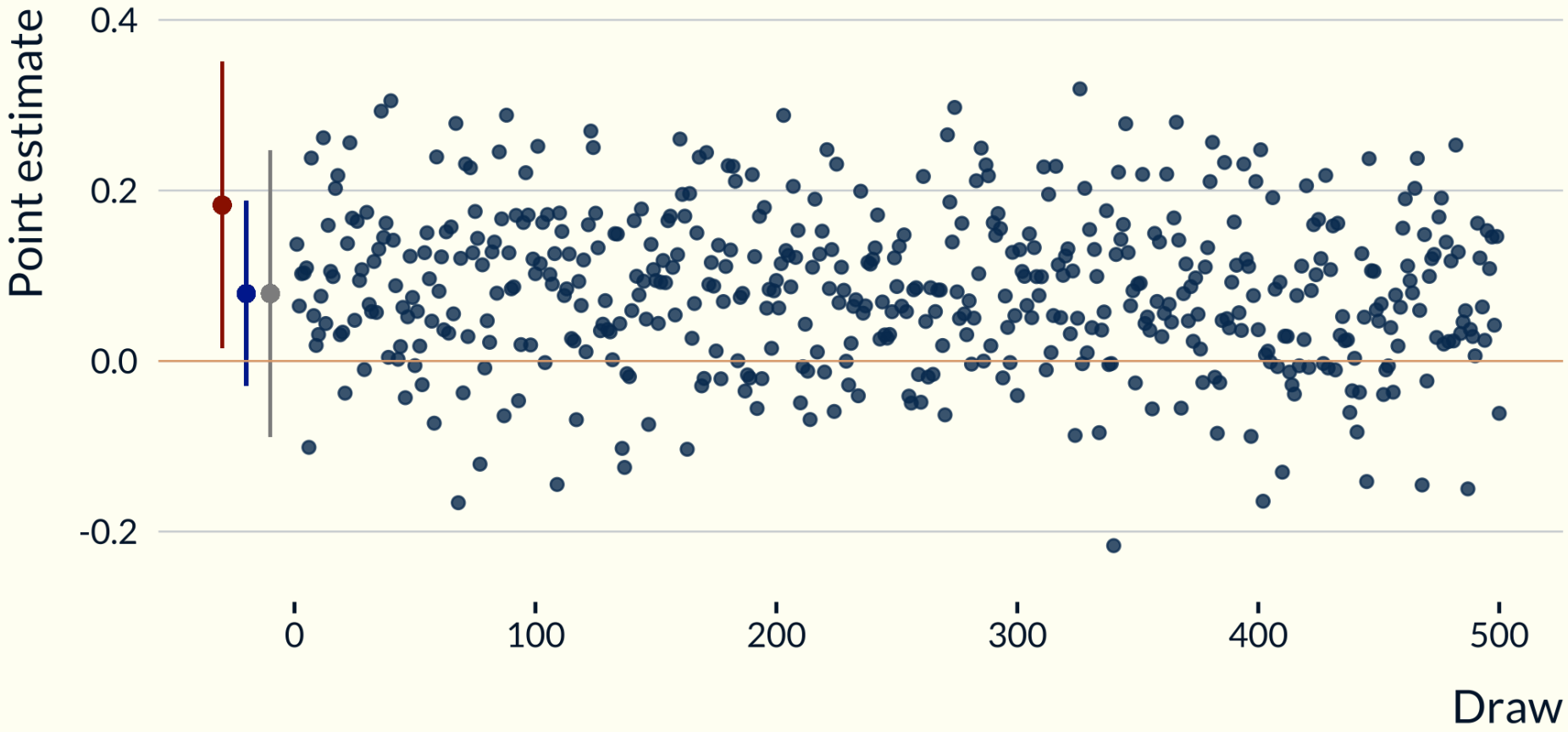
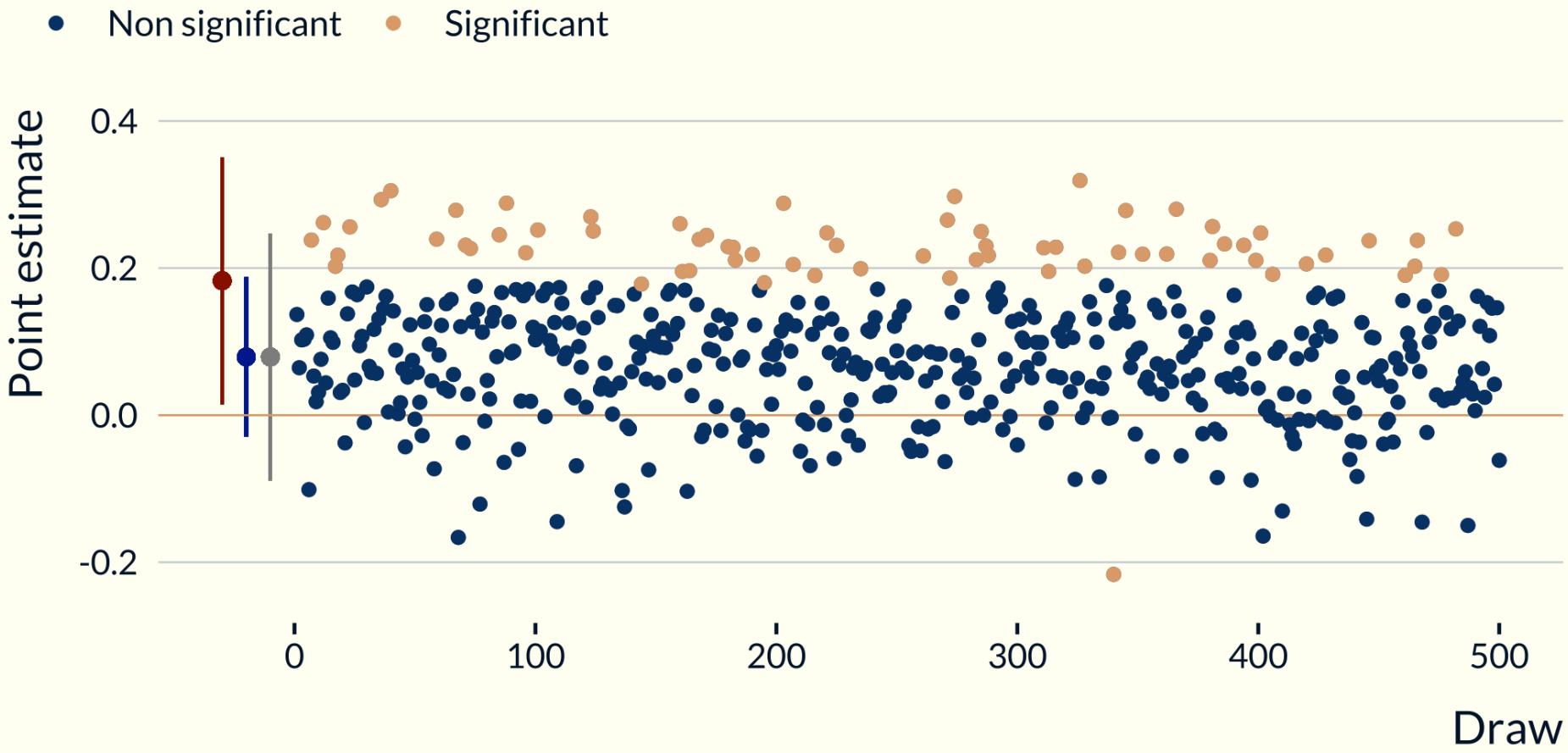
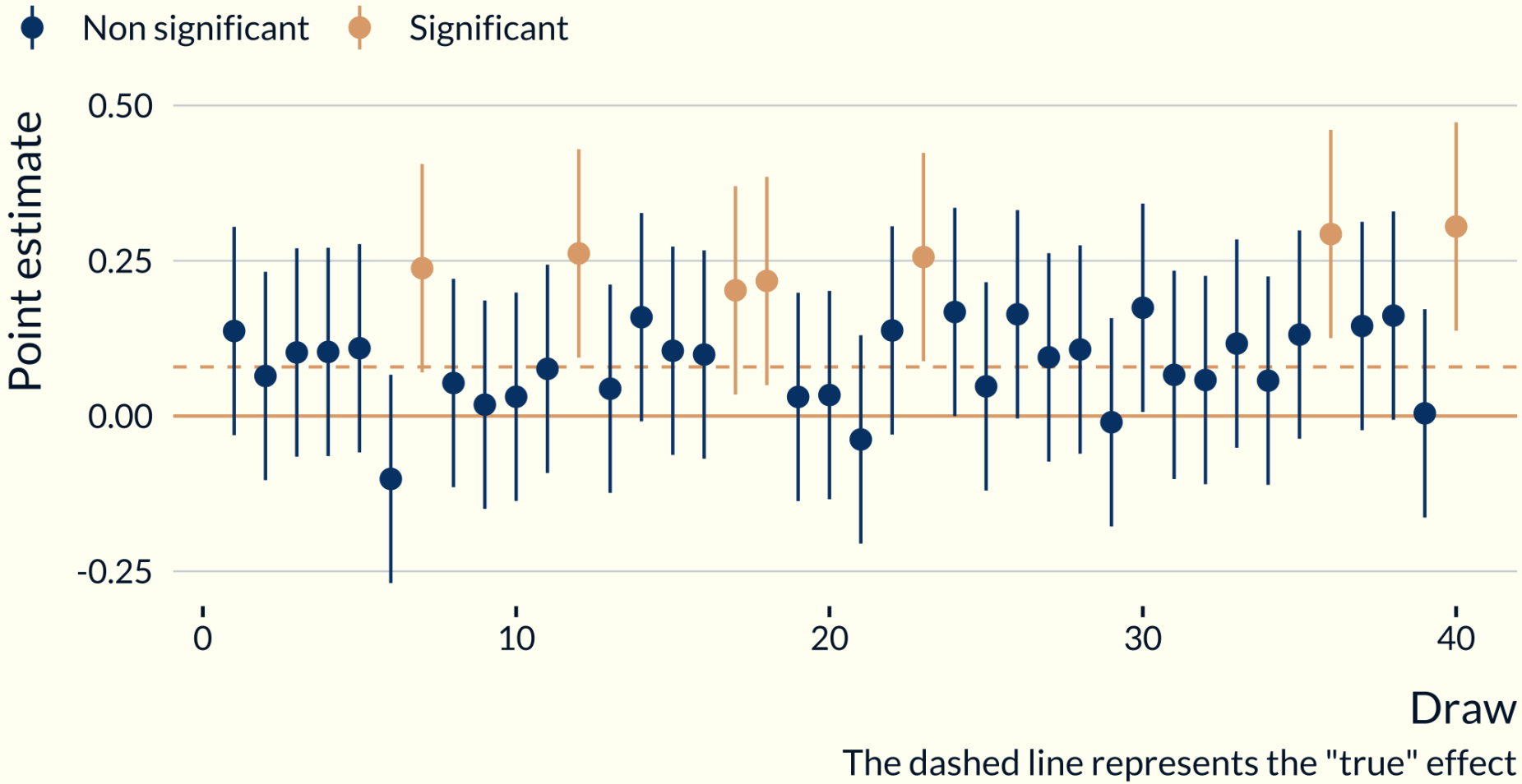


Illustration of the exaggeration and power issues

500 draws of an estimator ~ N(Effect size in replication, std err in original study)



Draws from the distribution of an estimator



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*)

Economics faces the two ingredients for exagg.

1. **Significant results are favored**
    - Evidence of a significance filter in economics
    - (Rosenthal 1979, Andrews and Kasy 2019, Abadie 2020, Brodeur et al. 2016, 2020)
2. **Low statistical power**
    - Median power in economics: 18%
    - (Ioannidis et al. 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 et al. 2017)
- Not all designs suffer from exaggeration
- But exaggeration is likely substantial in many studies

Design Beyond Identification, Straightforward?

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

Multiple Goals

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 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 $\left(SNR = \frac{\text{True effect}}{\text{s.e.}}\right)$ 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: 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

Modeling affects the effective design

- Controlling and FEs partial out variation
- OLS estimator can be seen as a weighted average of individual treatment effects with $w_i = (Di - \mathbb{E}[D_i|X_i])^2$
- Observations for which treatment is well explained by covariates do not contribute to the estimation
- **Modifies the effective sample** $\Rightarrow$ can be different from nominal sample
- Can create power and exaggeration issues

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

Approach to improved design fall into four categories:

- **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

Asssessing 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

Necessary 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

- 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

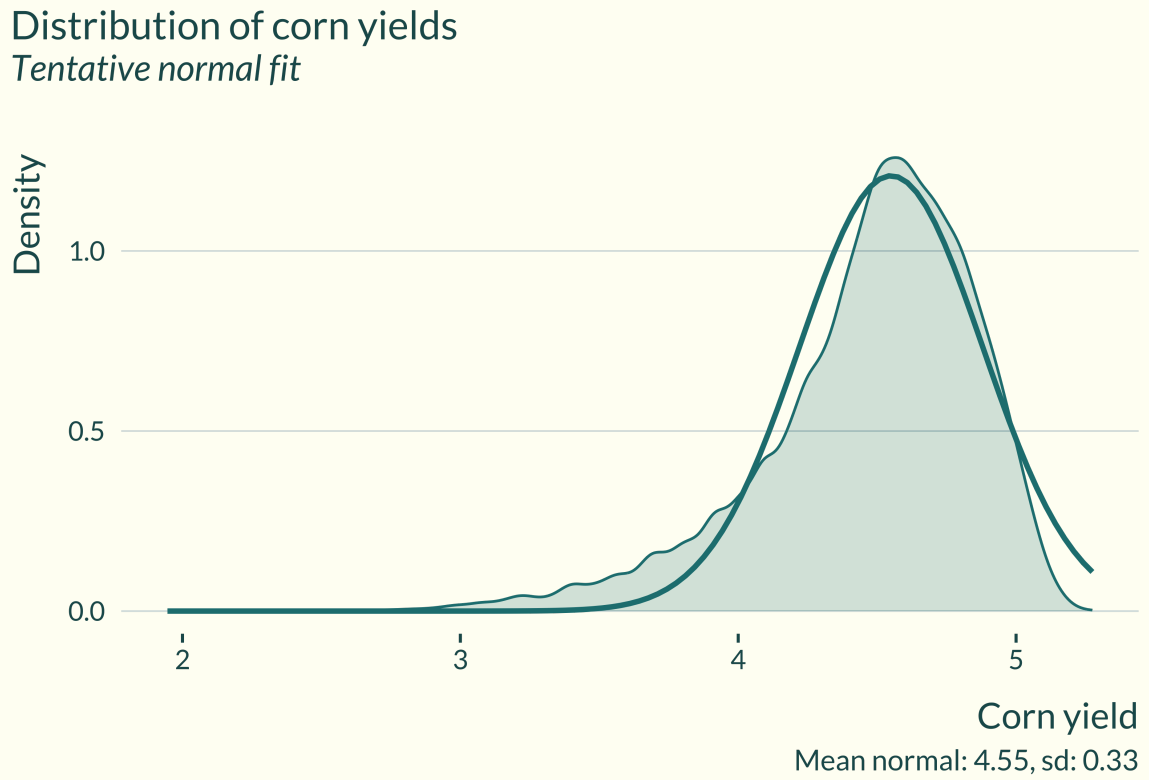
Why calibrating?

- 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



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:
  - Beyond identification
  - 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

Thank you!