Applied Microeconometrics

Lecture 1: Should we trust published research?

Manuel Bagues

August 2022

Lecture Slides

Practicalities

- Schedule:
  - August 29 - September 9: Mondays, Wednesdays and Fridays, 10:00-13:00
  - Location: Economicum, seminar room 1.
  - 20 min break around 10:20
- Class presentations:
  - There will be some short class presentations by students (around 15 min each)
- Problem sets:
  - Problem sets will be available after lectures 2, 4 and 6.
  - You can work in teams of up to 3 people
  - Problem sets will be marked on a pass/fail basis, based on effort.

Practicalities

- Office hours:
  - email me to schedule an appointment - I am available most of the time
- Material
  - Lecture slides: available at the beginning of each lecture
  - Articles cited in each lecture (check links in slides)
  - Recommended reading for each lecture
- Today: Gelman, Andrew and Eric Loken (2013), "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time", mimeo.

Structure of the module

- Main topics
  1. Multiple testing
  2. Statistical power
  3. Recent developments in:
     - Identification based on observables
     - IV
     - DID
     - RDD
- Any questions?

Overview

1. Should we trust published research?
   - 'Let's Take the Con out of Econometrics'
   - Abnormal distribution of p-values
   - Reproducibility crisis
2. Why do findings do not replicate?
   - Inference problems
   - Multiple testing
3. Possible solutions
   - Adjust standard errors for multiple testing
   - P<0.005
   - Preregistration
   - Result-blind evaluations
   - Replications

Should we trust published research?

Leamer (AER 1983): 'Let's Take the Con out of Econometrics'

- *'Hardly anyone takes data analysis seriously. Or perhaps more accurately, hardly anyone takes anyone's else data analyses seriously.'*
- Results may vary depending on:
  - Set of controls considered
  - Sample selection
  - Functional form assumptions
  - Distributional assumptions
- *'If you torture the data long enough, Nature will confess'* (attributed to Coase)

- Progress since Leamer's comments:
  - Sensitivity analysis
    - authors show that results are insensitive to the choice of assumptions (Leamer 1983)
  - *'Credibility revolution'*: rise of a design-based approach (see Angrist and Pischke (2010)
  - Data is typically available for replication

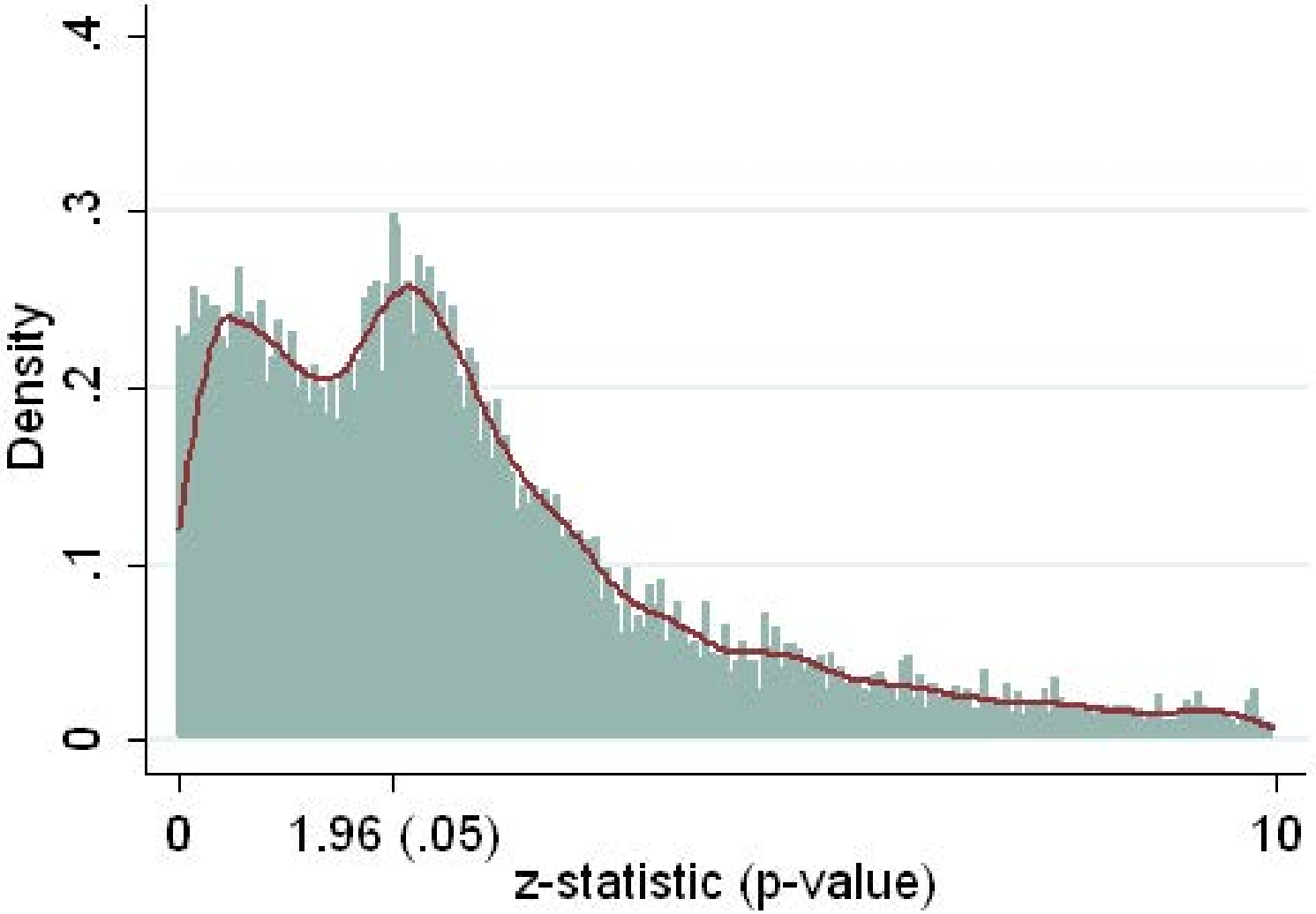
Should we trust published research?

- Despite this progress, we may still want to be cautious:
  - Abnormal distribution of p-values
  - 'Reproducibility crisis': Failure to replicate published research

Abnormal distribution of p-values

- Brodeur, Le, Sangnier and Zylberberg (2012) study the distribution of p-values in articles published between 2005 and 2011 in the main three economics journals (QJE, AER, JPE)
  - 'Star Wars: the Empirics Strike Back'
- It has a very peculiar shape!

Distribution of p-values

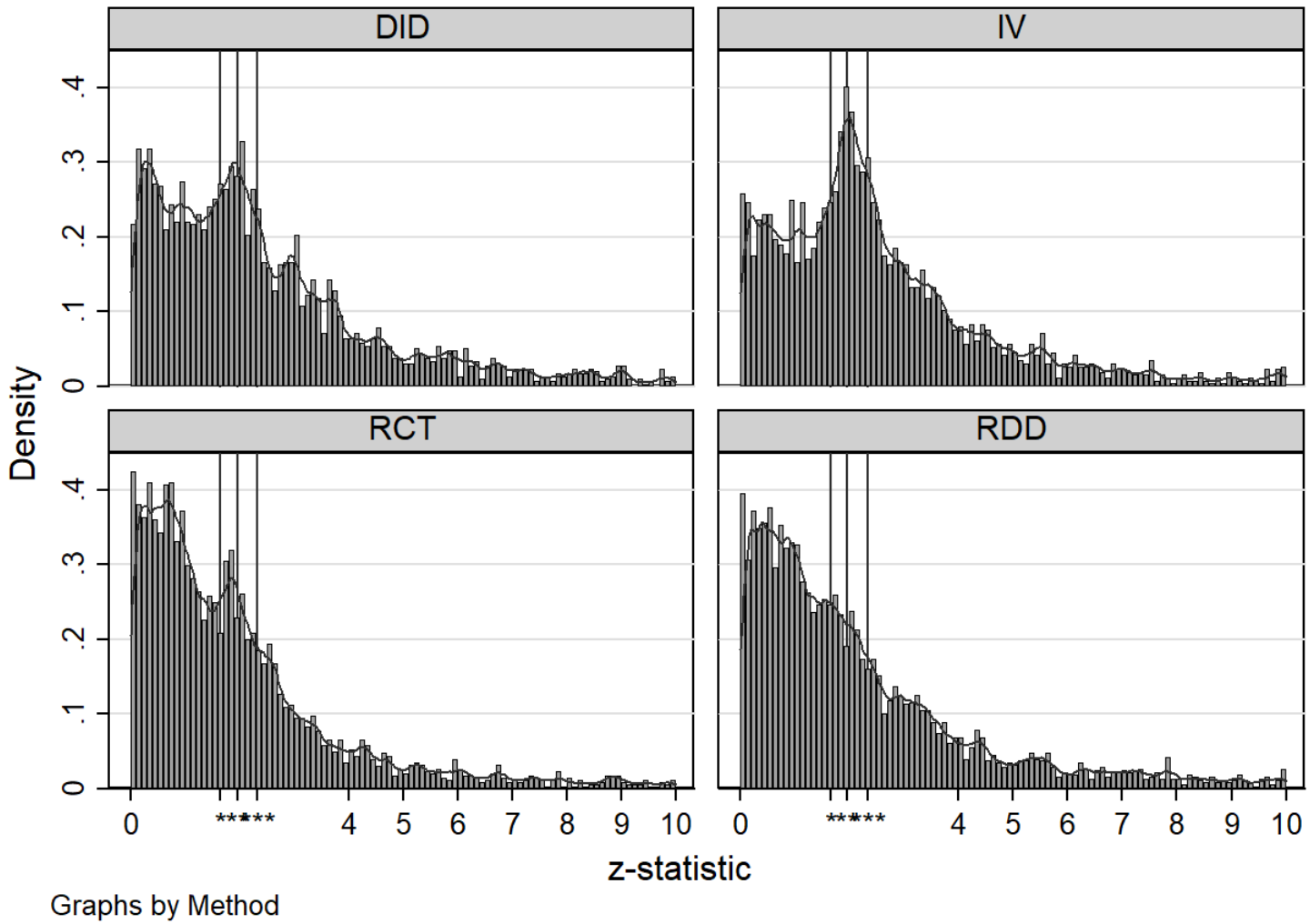


Abnormal distribution of p-values

- More recently, Brodeur et al. (2018) have studied the distribution of p-values in articles published in 25 economics journals
- Four methods:
  - DID
  - IV
  - RCT
  - RDD
- Which one do you think provides more degrees of freedom to researchers?

Distribution of p-values

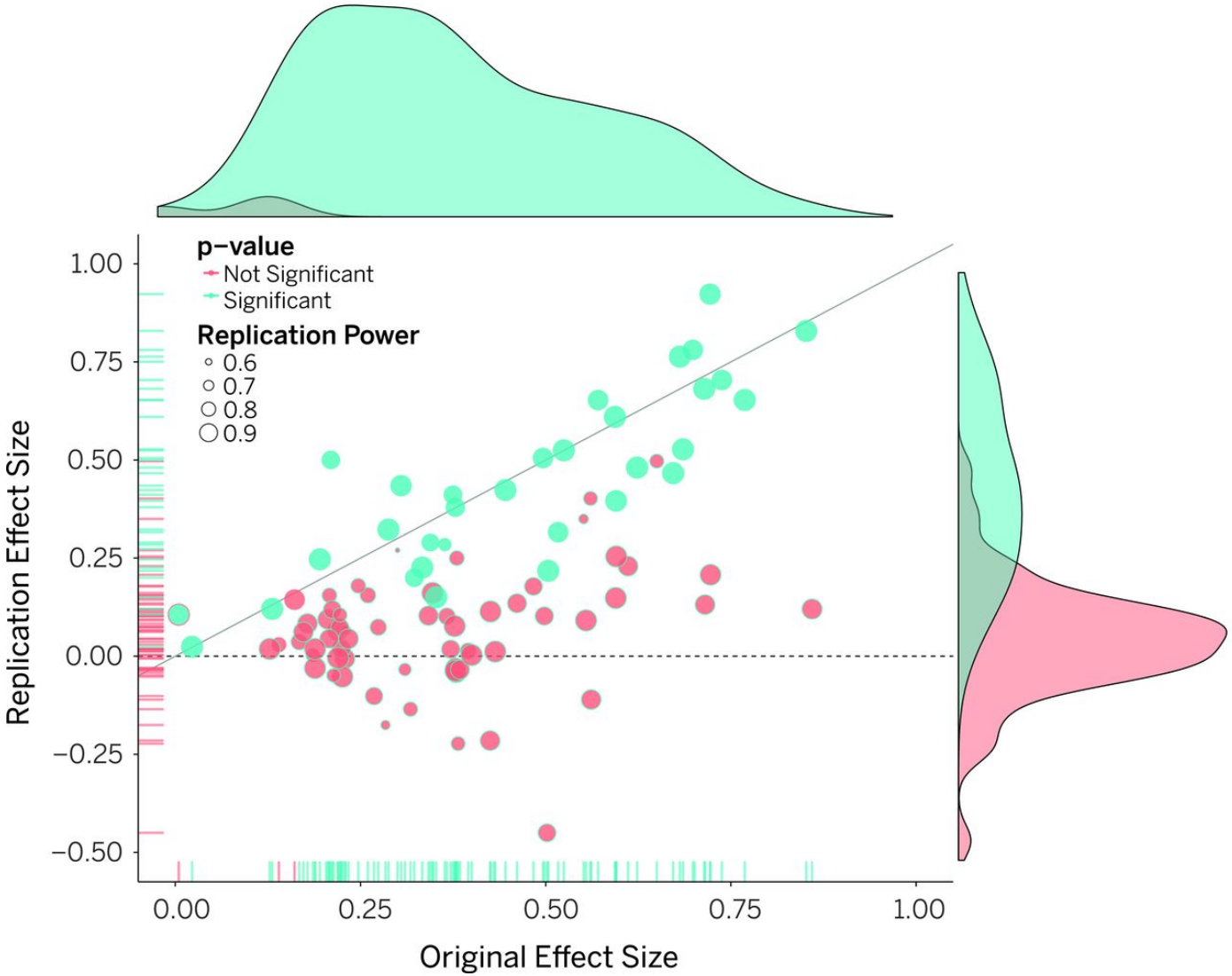
Figure 1: z-Statistics by Method



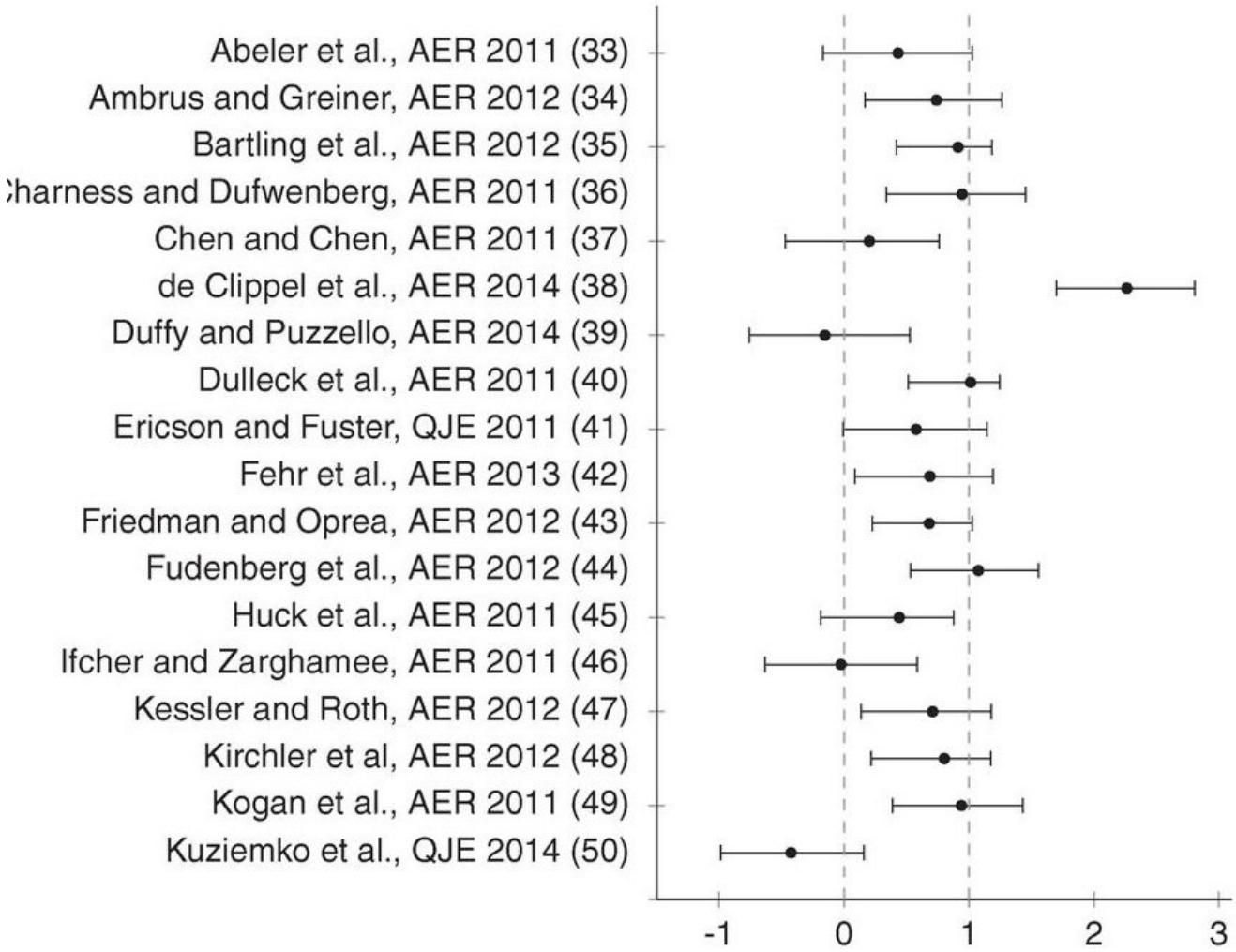
‘Reproducibility crisis’

- Psychology (Open Science Collaboration 2015):
  - Replication of 100 experiments published in 2008 in three high-ranking psychology journals.
  - 97% original studies had significant results (P < .05).
  - Only 36% of replications had significant results.
- Lab experiments in economics (Camerer et al. 2016)
  - 18 papers published in QJE and AER in 2011-2014
  - Replications had 90% power to detect an effect of the size that was originally reported.
  - 11 papers replicated; 7 were not replicated (3 were close misses)

Psychology (Open Science Collaboration 2015)

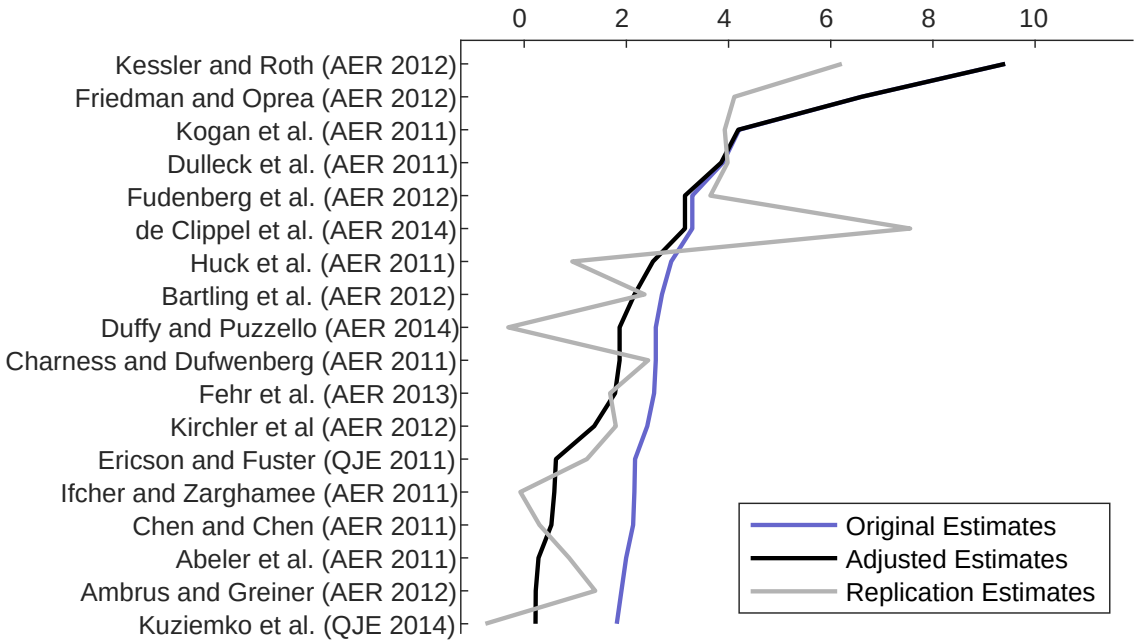


Economics (Camerer et al. 2016)



Panel A: CIs of replication effect sizes

Andrews and Kasy (2018), 'Identification of and correction for publication bias,' use this information to test their model of bias-correction



Why do findings do not replicate?

- Lack of external validity
  - Angus Deaton and Nancy Cartwright's critique of RCTs
  - Eg.: Azmat, Bagues, Cabrales and Iriberri (2019)

- Lack of internal validity
  - False positives

Publication bias → false positives

- Publication bias: journals (and therefore authors) have a preference for papers with statistically significant results (Chopra, Haaland, Stegmann and Roth (2022)
- From a scientific perspective this is awkward: a research question is interesting precisely because we do not know the answer
- Precise zero results vs. uninformative imprecise null results
  - Impact of lottery wins on health (Cesarini et al. 2016)
  - Impact of using masks on Covid infections ((Bundgaard et al. 2020)

Cesarini et al. 2016

- Abstract: 'We use administrative data on Swedish lottery players to estimate the causal impact of substantial wealth shocks on players' own health and their children's health and developmental outcomes. Our estimation sample is large, virtually free of attrition, and allows us to control for the factors conditional on which the prizes were randomly assigned. In adults, we find no evidence that wealth impacts mortality or health care utilization, with the possible exception of a small reduction in the consumption of mental health drugs. Our estimates allow us to rule out effects on 10-year mortality one sixth as large as the cross-sectional wealth-mortality gradient.'

Bundgaard et al. 2020

- Sumary: A randomized controlled trial (RCT) was conducted in Denmark in April and May 2020 to analyze the impact of mask recommendation on SARS-CoV-2 infections. The RCT assessed whether recommending surgical mask use outside the home reduces wearers' risk for SARS-CoV-2 infection in a setting where masks were uncommon and not among recommended public health measures. In the control group, participants were just encouraged to follow social distancing measures but there was no mask recommendation. In the treatment group, in addition to a recommendation for social distancing measures, participants were encouraged to wear a mask when outside the home among other persons together with a supply of 50 surgical masks and instructions for proper use. Around 4862 participants completed the study. Infection with SARS-CoV-2 occurred in 1.8% of participants in the treatment group, compared to 2.1% in the control group. The between-group difference was equal to -0.3 percentage points, with standard error equal to 0.4 percentage points.

Sources of false positives

- Academic fraud
- Inference problems
- Multiple testing

Inference

The estimation of standard errors is often non-trivial

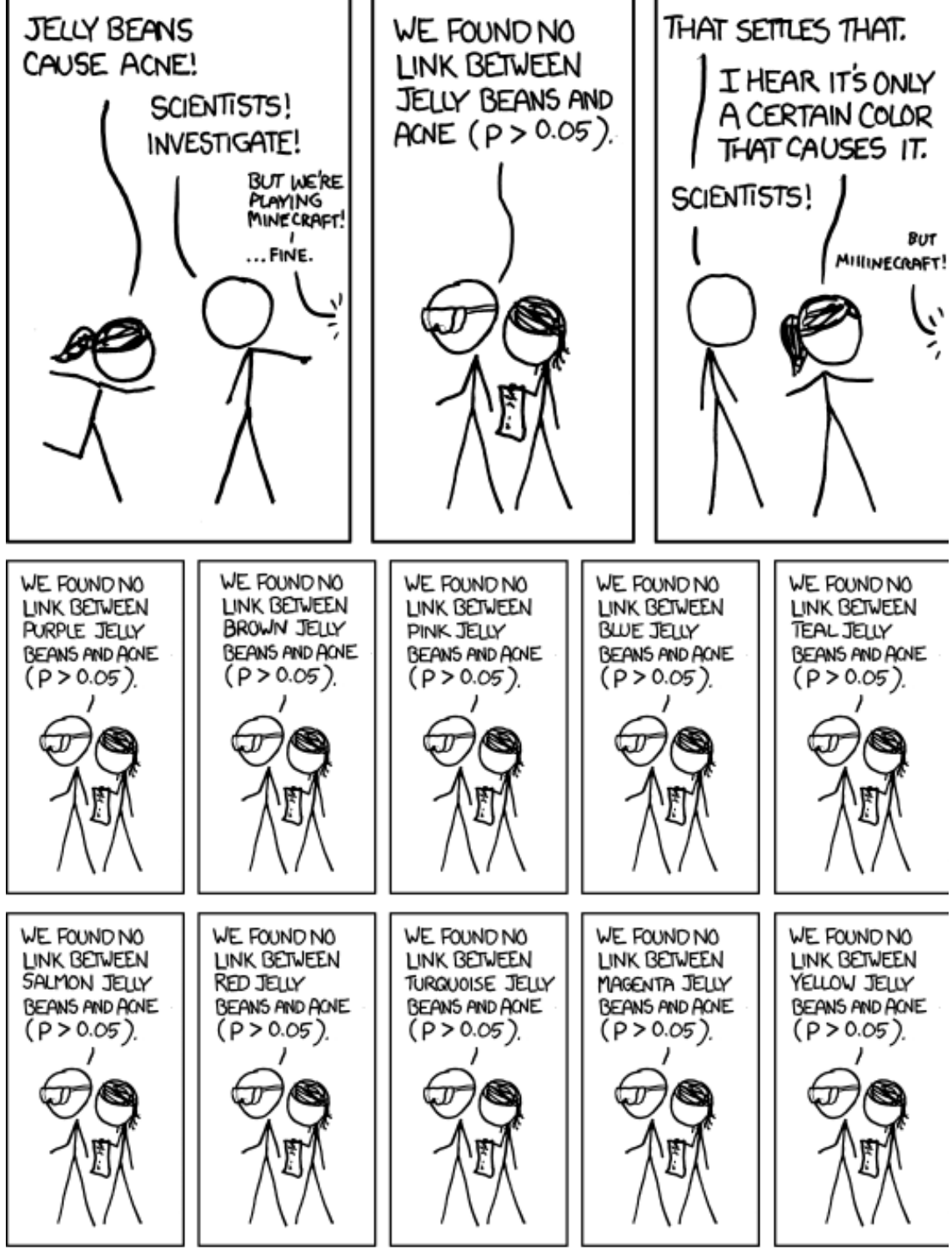
- Clustering:
  - Bertrand, Duflo and Mullainathan (2003), 'How much should we trust difference-in-differences estimates'
  - Abadie, Athey, Imbens and Wooldridge (2017), 'When Should You Adjust Standard Errors for Clustering?'

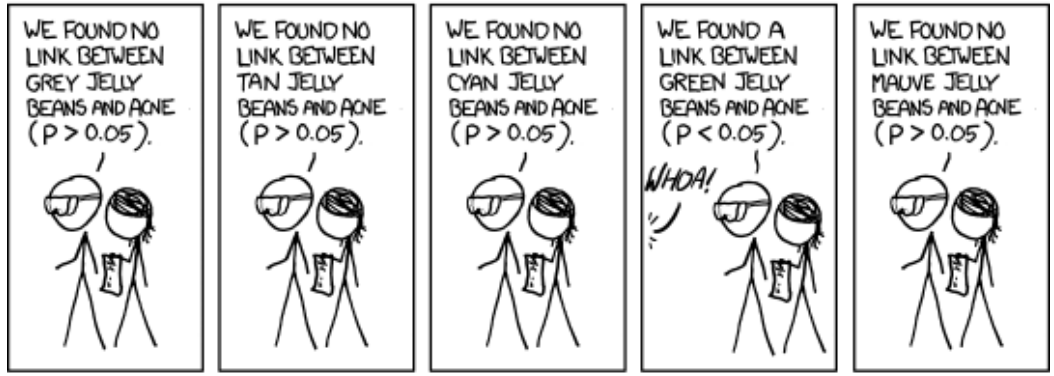
- Small sample properties of standard errors estimates:
  - Alwyn Young (2019), 'Hannelling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results'
  - Alwyn Young (2022), 'Consistency Without Inference: Instrumental Variables in Practical Application.'

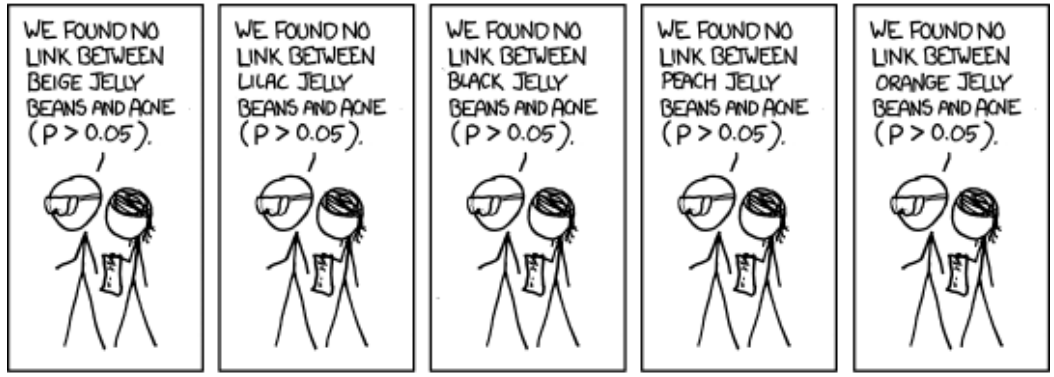
Multiple testing

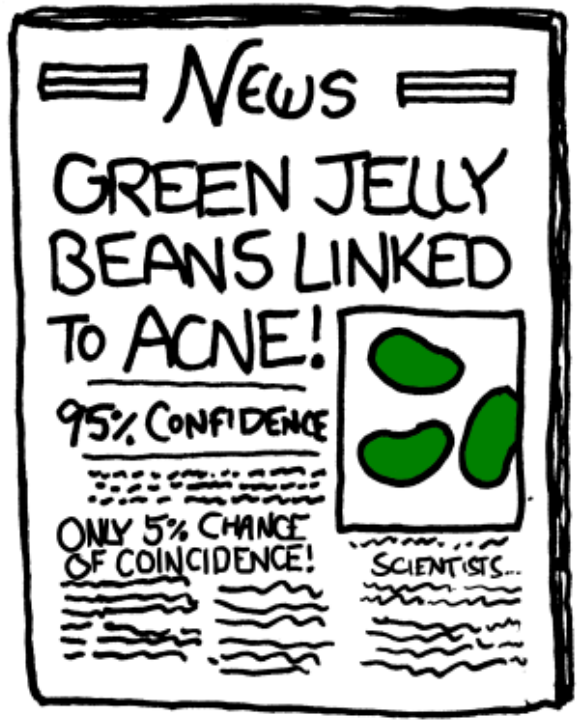
- Three ways in which multiple testing may become a problem:
  1. jointly identifying treatment effects for a set of outcomes
  2. estimating heterogeneous treatment effects through subgroup analysis
  3. conducting hypothesis testing for multiple treatment conditions.

- Example: we run an RCT to estimate what is the impact of taking this module on your future academic performance. We can maximize the probability of getting at least one significant estimate if we:
  1. classify students in N different groups: according to gender, nationality, previous grades...
  2. test the main hypothesis for N additional courses.
  3. consider N additional outcome variables: weight, self-reported happiness, marital status...


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Multiple testing

- Two possible sources:
  1. P-hacking
     - deliberately try all possible combinations, and select the 'successful' one
  2. Unconscious process
     - Gelman and Loken (2013) refer to the "The Garden of Forking Paths" to describe the near infinite number of choices facing researchers in cleaning and analyzing data.
     - Gelman argues that scientists can make false discoveries when they do not pre-specify a data analysis plan and instead choose 'one analysis for the particular data they saw.'

- Gelman and Loken (2013) discuss several academic papers:
  1. fat arms and political attitudes
  2. extrasensory perception
  3. pink shirts and peak fertility
  4. menstrual cycle and vote intentions
- Let us discuss the last case.

Durante, Arsena, and Griskevicius (2013)

The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle

Abstract:

Building on theory suggesting that political and religious orientation are linked to reproductive goals, we tested how fertility influenced women's politics, religiosity, and voting in the 2012 U.S. presidential election. In two studies with large and diverse samples, ovulation had drastically different effects on single versus married women. **Ovulation led single women to become more liberal, less religious, and more likely to vote for Barack Obama. In contrast, ovulation led married women to become more conservative, more religious, and more likely to vote for Mitt Romney.** (...) Overall, the ovulatory cycle not only influences women's politics, but appears to do so differently for single versus married women.

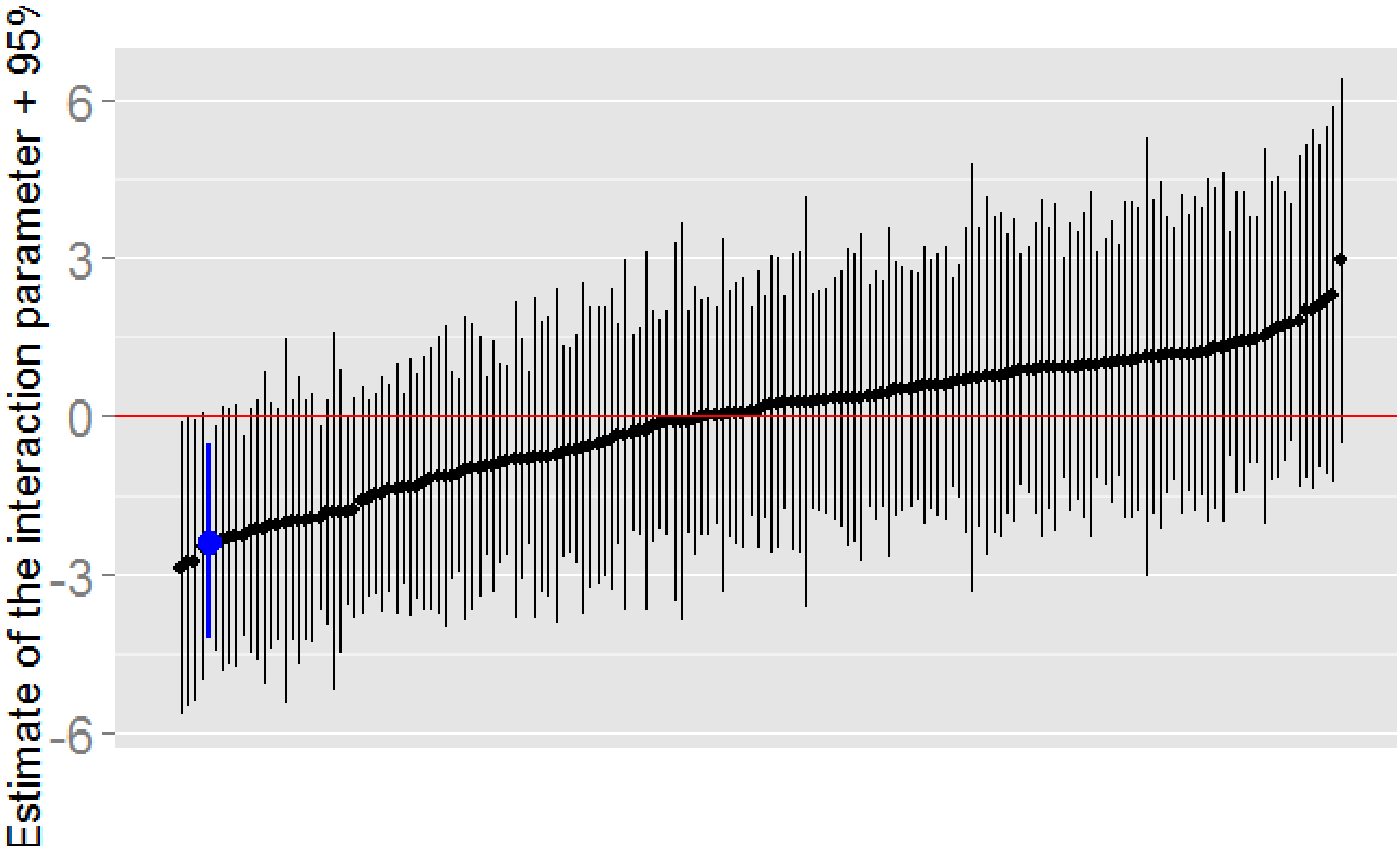
- The magnitude of the effect is very large, in the range of a 20 percentage point difference.

Choices!

1. Exclusion criteria based on cycle length (3 options)
2. Exclusion criteria based on "How sure are you?" response (2)
3. Cycle day assessment (3)
4. Fertility assessment (4)
5. Relationship status assessment (3)

168 possibilities (after excluding some contradictory combinations)

Living in the multiverse



Gelman and Loken (2013)

- "In this garden of forking paths, whatever route you take seems predetermined, but that's because the choices are done implicitly. The researchers are not trying multiple tests to see which has the best p-value; rather, they are using their scientific common sense to formulate their hypotheses in reasonable way, given the data they have. The mistake is in thinking that, if the particular path that was chosen yields statistical significance, that this is strong evidence in favor of the hypothesis."
- "In all the cases we have discussed, the published analysis has a story that is consistent with the scientific hypotheses that motivated the work, but other data patterns would naturally have led to different data analyses (for example, a focus on main effects rather than interactions, or a different choice of data subsets to compare) which equally could have been used to support the research hypotheses."

Zinovyeva and Bagues (IZA working paper 2011)

Does Gender Matter for Academic Promotion? Evidence from a Randomized Natural Experiment

- Another example of how researchers can get confused by the Garden of Forking Paths, this time from my own work.
- Bagues and Zinovyeva (IZA working paper 2011) study whether a higher presence of women in scientific committees increases the chances of success of female candidates.
- The identification strategy exploits the random assignment mechanism in place between 2002 and 2006 in all academic disciplines in Spain to select the members of promotion committees.
- The database includes around 30,000 applicants in all fields and in two categories, Associate and Full Professor.

Zinovyeva and Bagues (IZA working paper 2011)

Does Gender Matter for Academic Promotion? Evidence from a Randomized Natural Experiment

- Overall, the gender composition of committees has no significant impact on the chances of success of female applicants.
- However, we do find a significant impact when we examine how the effect varies depending on the type of position at stake: full vs associate professorship
- We emphasize this finding in the abstract and propose a theoretical explanation:
  - We find that a larger proportion of female evaluators increases the chances of success of female applicants to full professor positions. Conversely, when committee members decide on promotions to associate professor positions, a larger share of female evaluators is associated with fewer successful female applicants. The evidence is consistent with the existence of ambivalent sexism.

Zinovyeva and Bagues (IZA working paper 2011)

Does Gender Matter for Academic Promotion? Evidence from a Randomized Natural Experiment

- In the paper we report a number of additional heterogeneity analysis (you can think about a good theoretical justification for each on of these analyses)
  - By field (social sciences and humanities vs science)
  - By size of the field (small vs large fields)
  - By degree of feminization of the field (above and below median)
- None of these analysis yields any significant result.
- Taking into account all the regressions that have been run, how confident should we be about the statistical significance of the particular result that we emphasize in the abstract?

Bagues, Sylos-Labini and Zinovyeva (American Economic Review 2017)

Does the Gender Composition of Scientific Committees Matter?

- A few years later similar data became available for Italy, a country which is similar to Spain in many dimensions. For instance, the share of women in different academic positions is almost identical.
- The Italian dataset includes information on around 70,000 evaluations in all fields.
- Again, evaluators are randomly selected from the pool of eligible evaluators.
- We decided to integrate the results from Spain and from Italy in a new paper.

Bagues, Sylos-Labini and Zinovyeva (American Economic Review 2017)

Does the Gender Composition of Scientific Committees Matter?

Table 7: Heterogeneity analysis

	1	2	3	4
	Italy		**Spain**	
Discipline	SSH	STEMM	SSH	STEMM
	-0.116**	-0.128***	-0.026	0.004
	(0.054)	(0.034)	(0.038)	(0.041)
Feminization of field	$\geq$ *median*	$<$ *median*	$\geq$ *median*	$<$ *median*
	-0.149***	-0.072	-0.018	-0.016
	(0.042)	(0.057)	(0.040)	(0.037)
Level of promotion	FP	AP	FP	AP
	-0.111*	-0.138***	0.120**	-0.072**
	(0.059)	(0.038)	(0.054)	(0.032)

Bagues, Sylos-Labini and Zinovyeva (American Economic Review 2017)

Does the Gender Composition of Scientific Committees Matter?

Let me ask you a tricky question:

- Taking into account the evidence reported in the previous slide, should a policy-maker whose objective is to increase the share of women in Full Professor positions introduce gender quotas in exams to Full Professor positions in Spain?

Similar problems affect also other disciplines

- Example: does Hydroxychloroquine treatment work for COVID-19?
  - Gautret et al. (International Journal of Antimicrobial Agent, March 2020): It reduces viral load reduction by 57.5 p.p.
  - Mehra et al. (Lancet, May 2020 - retracted): It increases mortality by 22%-45%
- Many stars in both studies :-)
- Why do results differ?
  - Differences in the implementation of the treatment
    - size of the dosis, timing...
    - The economist as a 'plumber' $\rightarrow$ details matter! (Duflo 2017)
  - External validity: French genes are special
  - Internal validity:
    - Empirical strategy not valid (in both cases the identification relies on observables)
    - Publication bias and multiple testing
    - Publication and academic fraud

Possible solutions to the proliferation of false positives

Possible solutions to the proliferation of false positives

1. Adjust standard errors for multiple testing
2. Redefine statistical significance: P<0.005 (Benjamin et al. 2017)
3. Pre-analysis planning and hypothesis registries (Gelman 2013, Coffman and Niederle 2015)
4. Result-blind evaluations
5. Independent replication (Gelman 2013, Maniadis, Tufano and List 2014)

Adjust standard errors for multiple testing

- We may want to adjust the standard errors taking into account all the different tests that you have conducted
- Simplest method: Bonferroni
  - Multiply p-values by the number of hypothesis testing
- Bonferroni typically too conservative, it implicitly assumes that we consider independent outcomes
- A bit more advanced: Anderson 2008
- List, Shaikh and Xu (2016) also provide a correction procedure that controls for the familywise error. The stata command is available here: `https://github.com/seidelj/mht`

Redefine statistical significance: P<0.005

- Proposal supported by 50+ scientists, statisticians and economists (Camerer et al. 2016)
- Advantages?
- Disadvantages? (type II errors, inequality...)

Preregistration

- Credibly fixed plan of how a researcher will collect and analyze data, which is submitted before a project begins
- Hypothesis registries
  - Database of all projects attempted; the goal of this promising mechanism is to alleviate the 'file drawer' problem
- In some disciplines preregistration is becoming the rule
  - E.g.: clinical trials
- In social sciences is becoming increasingly common:
  - American Economic Association's registry for RCTs: https://www.socialscienceregistry.org/
  - To this day (8 Jan 2020): 3165 studies in 142 countries.
  - Required

Result-blind' evaluations

- 'Result-blind' evaluations (e.g. Journal of Economic Development, JPE Micro)
  - Assess papers based on relevance of question, interest of the theoretical framework, empirical strategy, accuracy... (but not on stars)

Hold-up samples

- Potential drawback
  - It might be sometimes difficult to think carefully about an analysis before you have the data
  - E.g.: Bagues and Roth (2022)
- Possible solution
  - Use a hold-up sample
  - E.g.: ask the data provider to release only a random sample of the dataset

Replications

- In the context of experiments, there are at least three levels at which replications can take place (Levitt and List 2009):
  1. Reanalyzing the original data
  2. Implementing an experiment under a similar protocol to the original experiment
  3. New research design with the purpose of testing the hypothesis of the first study

Replications

- "Economists treat replication the way teenagers treat chastity, as an ideal to be professed but not to be practiced." - Dan Hamermesh
- Incentives to conduct replications are weak, the current incentive system in sciences emphasizes innovation to the detriment of verification.
- Journals tend to be reluctant to to publish replications.
  - original journal may not want to admit that they have made a mistake
  - moreover, replications tend to attract few citations
  - interesting exception: Journal of Applied Econometrics replication section
- But even if replications were published in top journals, a replication typically does not help to signal your skills or your originality

- To a large extent, this task is being conducted by PhD students as part of their coursework.
- Some PhD students have uncovered coding mistakes in very influential papers:
  - Reinhart and Rogoff (2010) excel coding error
  - Foote and Goetz's (QJE 2008) replication of Donohue and Levitt's (2001)
    - See also Donohue and Levitt's (2020)
- In other occasions it seems that it was the garden of forking paths which might have misguided some researchers:
  - Jesse Rothstein's (AER 2007) replication of Hoxby (2000)
  - David Albouy's (AER 2012) replication of Acemoglu, Johnson, and Robinson (2001)

- The above replications were based mainly on the material provided by the authors
- The incentive problem is more problematic when the replication requires a big investment (new RCT).
- The Bill and Melinda Gates Foundation launched the International Initiative for Impact Evaluation (3ie)
    - Includes funding to replicate the 20 most relevant studies in Development Economics

- Posible ways to increase the incentives to replicate?

- Posible ways to increase the incentives to replicate?
  - Butera and List (NBER 2017), 'An Economic Approach to Alleviate the Crises of Confidence in Science: With an Application to the Public Goods Game'
    - Final version: Butera, Grossman, Houser, List and Villeval (NBER 2020)
    - Unpublished as of August 2022, it has received 11 citations.
  - Coffman, Niederle and Wilson (2017), 'A Proposal to Organize and Promote Replications'

- Recommended reading for next lecture:
  - Gelman, Andrew and David Weakliem (2009), "Of beauty, sex, and power", American Scientist 97(4), 310-316
  - Maniadis, Zacharias, Fabio Tufano and John List (2014), "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects," American Economic Review 104(1), 277-290.
  - Danzer and Lavy (2018), "Paid Parental Leave and Children's Schooling's Outcomes," Economic Journal, Vol. 128(608), 81-117.

What about you?

1. Name
2. Program and Year (e.g. 2nd year PhD at Aalto)
3. Research interests (e.g. what is the main question that you are addressing in your current research)