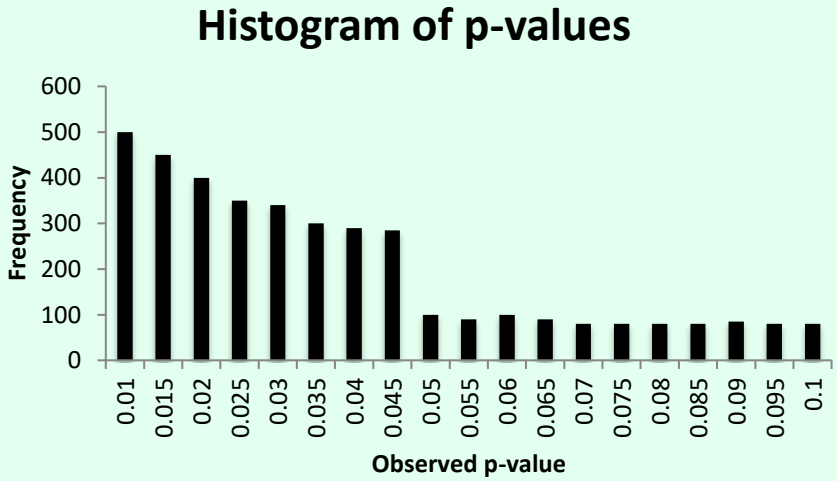
Non-Standard Errors

Background

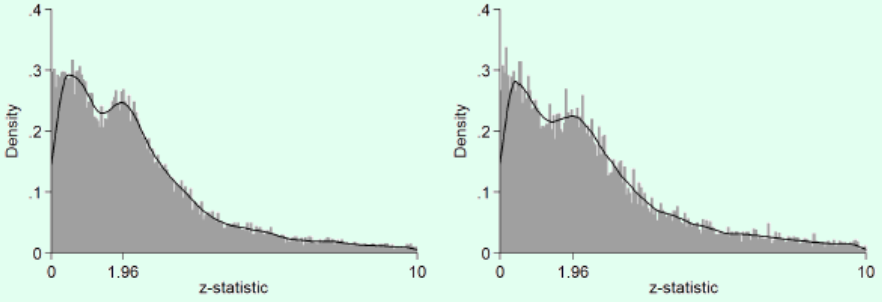
The replication crisis

Open Science Collaboration (2015). "Estimating the Reproducibility of Psychological Science." *Science*, 349(6251); Camerer et al. (2016) "Evaluating replicability of laboratory experiments in economics." *Science;* Camerer et al. (2018) "Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015." *Nature Human Behaviour*

Researcher degrees of freedom

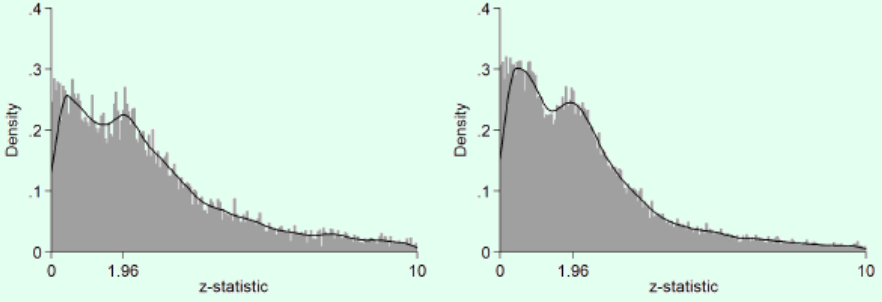


Ioannidis 2005 Why Most Published Research Findings Are False; Simmons, Nelson and Simonsohn 2011 False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant; Gelman and Loken 2013 The Garden of Forking Paths



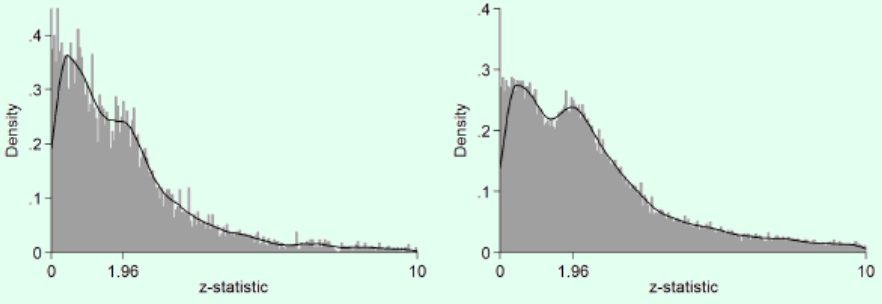
(a) Eye-catchers.

(b) No eye-catchers.



(c) Model.

(d) No model.



(e) Lab. experiments or RCT data.

(f) Other data.

Sources: AER, JPE, and QJE (2005-2011). Distributions are unweighted and plotted using de-rounded statistics. Lines correspond to kernel density estimates.

Brodeur et al 2016 AEJ: AE

Specification Choice in Randomized and Natural Experiments: Lessons from the Regulation SHO Experiment

Bernard S. Black
Northwestern University, Pritzker School of Law and Kellogg School of Management

Hemang Desai
Southern Methodist University, Cox School of Business

Kate Litvak
Northwestern University, Pritzker School of Law

Woongsun Yoo
Central Michigan University, College of Business Administration

Jeff Jiewei Yu
University of Arizona, School of Accountancy

Presidential Address: The Scientific Outlook in Financial Economics

CAMPBELL R. HARVEY*

ABSTRACT

Given the competition for top journal space, there is an incentive to produce "significant" results. With the combination of unreported tests, lack of adjustment for multiple tests, and direct and indirect *p*-hacking, many of the results being published will fail to hold up in the future. In addition, there are basic issues with the interpretation of statistical significance. Increasing thresholds may be necessary, but still may not be sufficient: if the effect being studied is rare, even $t > 3$ will produce a large number of false positives. Here I explore the meaning and limitations of a *p*-value. I offer a simple alternative (the minimum Bayes factor). I present guidelines for a robust, transparent research culture in financial economics. Finally, I offer some thoughts on the importance of risk-taking (from the perspective of authors and editors) to advance our field.

Pre-analysis plans

- Removes the researcher degrees of freedom
- Exploratory analyses can be very interesting!
  - But should not be presented as confirmatory

Pre-analysis plan is one fork

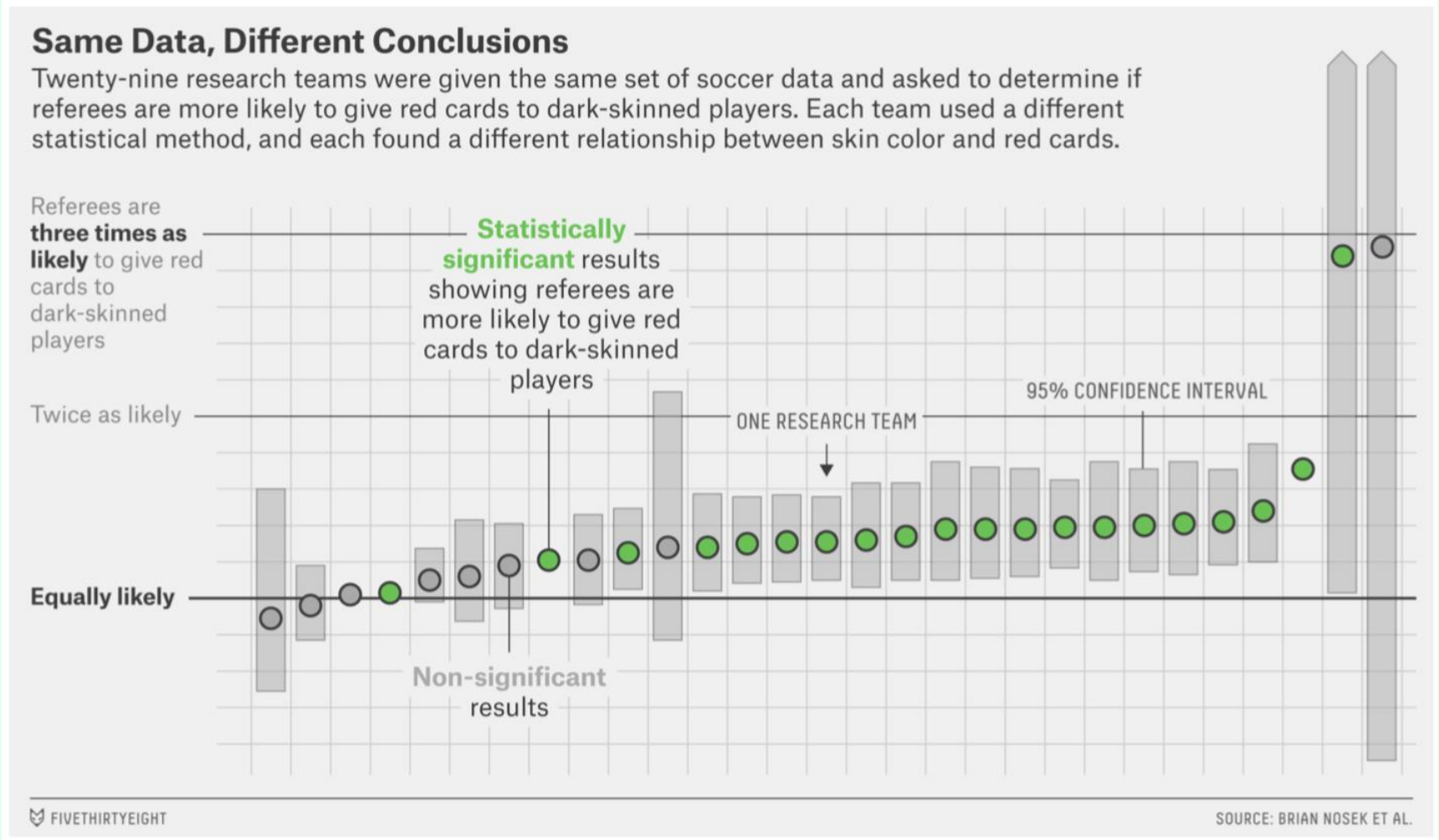
- Pre-analysis plan shows one potential fork in the data
  - With meaningful p-values
- Many forks possible
- Different researchers might choose different forks/pre-analysis plans
  - Pre-analysis plan should not result in systematic bias in effect sizes, but will underestimate the standard error in the statistical test

Many analysts vs multiverse

- What is the natural variation in analyses and results in a given data set?
  - Can be explored with multi-analyst approach

- What is the theoretically justified set of analyses and results in a given data set?
  - Can be explored with multiverse analyses, vibration of effects, specification curve analysis

Multi-analyst approach



20 out of 29 teams find statistically significant positive result

Silberzahn et al. 2018

Multi-analyst approach

- Neuroscience (Botvinik-Nezer et al 2020)
  - 9 hypotheses, 70 teams
- Economics replications (Huntington-Klein et al 2021)
  - 2 hypotheses (previous studies), 7 analysts for each
- Sociology (Breznau et al 2021 and Schweinsberg et al 2021)
  - Breznau et al: 1 hypothesis, 73 teams
  - Schweinsberg et al: 2 hypotheses, 14-15 analysts for each
- Results: Lots of variation

Botvinik-Nezer et al

	Hypothesis description	Fraction of teams reporting a significant result	Median confidence level	Median similarity estimation
#1	Positive parametric effect of gains in the vmPFC (equal indifference group)	0.371	7 (2)	7 (1.5)
#2	Positive parametric effect of gains in the vmPFC (equal range group)	0.214	7 (1.5)	7 (1)
#3	Positive parametric effect of gains in the ventral striatum (equal indifference group)	0.229	6 (1)	7 (1)
#4	Positive parametric effect of gains in the ventral striatum (equal range group)	0.329	6 (1)	7 (1)
#5	Negative parametric effect of losses in the vmPFC (equal indifference group)	0.843	8 (1)	8 (1)
#6	Negative parametric effect of losses in the vmPFC (equal range group)	0.329	7 (1)	7 (1)
#7	Positive parametric effect of losses in the amygdala (equal indifference group)	0.057	7 (1)	8 (1)
#8	Positive parametric effect of losses in the amygdala (equal range group)	0.057	7 (1)	8 (1)
#9	Greater positive response to losses in amygdala for equal range group vs. equal indifference group	0.057	6 (1)	7 (1)