Testing for and Evaluating the Extent of Selective Reporting

Nikolay Kudrin
Queen's University

February 27, 2025

How do we detect selective reporting?

- How can we **test** for (the absence of) selective reporting?
  - ➢ Exploit the *distribution of published results* – t-stats or p-vals
- **Challenges**:
  - ➢ *Composite Null* – p-curve shape depends on power in underlying studies
  - ➢ *Composite Alternative* – many ways to p-hack
- *This paper* derives tests that
  - ➢ **Control Type I error** over the entire (or mildly restricted) null set
  - ➢ More **powerful vs. wider range** of alternatives relative to existing tests

One study (absent selection)

Consider study s: $X_{s,1}, \ldots, X_{s,n_s} \sim$ i.i.d. $\mathcal{N}(\mu_s, \sigma_s^2)$, σ_s is known

- Researchers are testing

$$H_0 : E[X_s] = 0 \quad \text{against} \quad H_1 : E[X_s] \neq 0.$$

- t-statistic

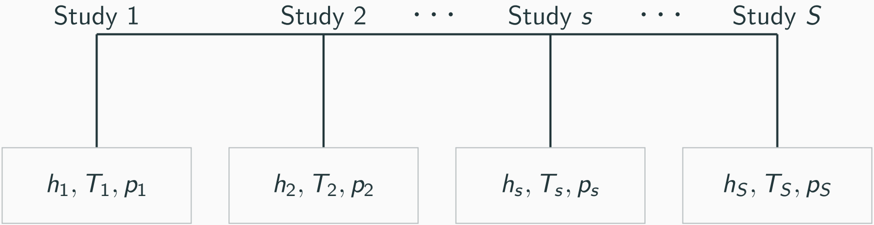
$$T_s = \frac{\sqrt{n_s}\bar{X}_s}{\sigma_s} = \frac{\sqrt{n_s}\mu_s}{\sigma_s} + \frac{\sqrt{n_s}(\bar{X}_s - \mu_s)}{\sigma_s} = \underbrace{h_s}_{\text{(local) alternative/effect}} + \underbrace{W_s}_{\sim\mathcal{N}(0,1)}$$

- What is the power at significance level p?

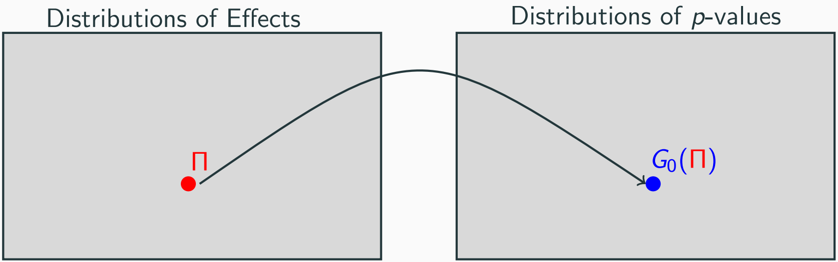
$$\begin{aligned} \beta(p, h_s) &= \Pr(|T_s| > \mathrm{cv}(p) \mid h_s) = \Pr(p_s \leq p \mid h_s) \\ &= 2 - \Phi(\mathrm{cv}(p) - h_s) - \Phi(\mathrm{cv}(p) + h_s) \leftarrow \textbf{known function} \end{aligned}$$

- Immediate generalization to testing problems with *limiting normal experiments* (asy. t-tests)

Literature (absent selection)



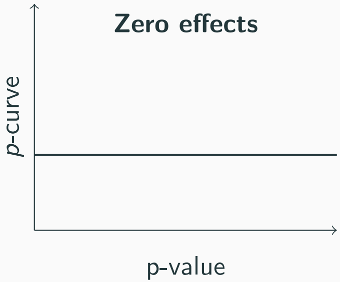
Treat h as *random*: $h \sim \Pi \Rightarrow$ Distribution of p_s: $G_0(p) = \int \beta(p,h)d\Pi(h)$

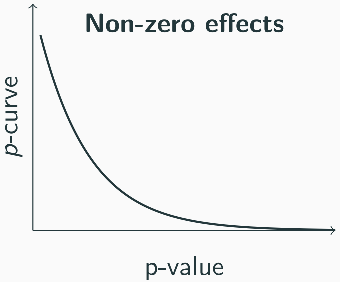


Null: Which Distributions are consistent with 'No Selection'?

The *p*-curve shape depends on the *distribution of effects in the literature*

- or the implied *distribution of power*





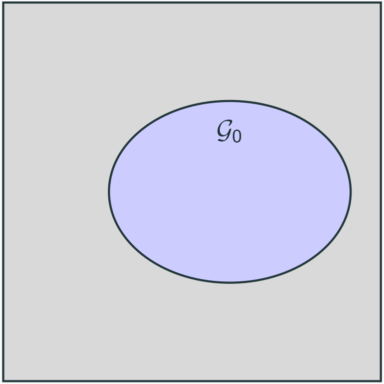
The Null Set

$$\mathcal{G}_0 := \left\{ G_0 \mid G_0(p) = \int_{\mathbb{R}} \beta(p,h) d\Pi(h), \quad \Pi \in \{\text{all probability distributions}\} \right\}$$

No selective reporting:

$H_0 : \; G \in \mathcal{G}_0$

Distributions of p-values



Existing Tests

$$\mathcal{G}_0 := \left\{ G_0 \mid G_0(p) = \int_{\mathbb{R}} \beta(p,h)d\Pi(h), \quad \Pi \in \{\text{all probability distributions}\} \right\}$$

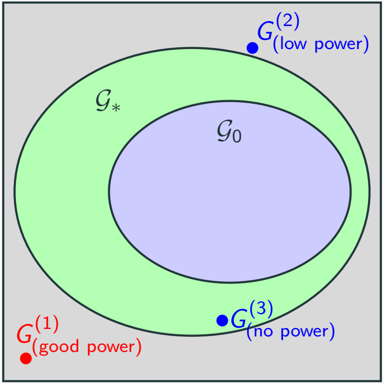
No selective reporting:

$H_0: \ G \in \mathcal{G}_0$

Existing Tests: *testable implications $\mathcal{G}_0 \subset \mathcal{G}_*$:*

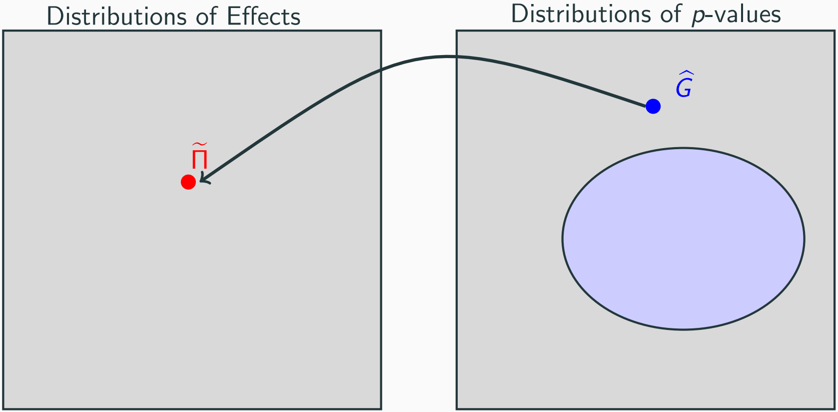
- Continuity
- (Complete) Monotonicity
- Bounds

Distributions of p-values



New (More Powerful) Tests

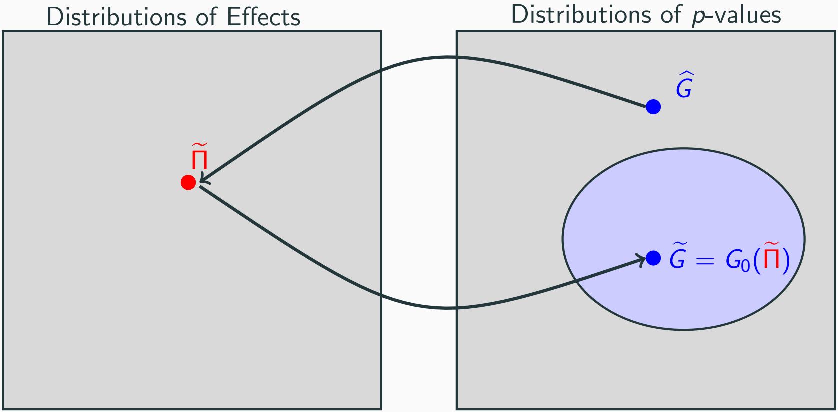
$$\underbrace{T}_{\text{observable}} = \underbrace{h}_{\sim\Pi} + \underbrace{W}_{\sim\mathcal{N}(0,1)} \Rightarrow \overbrace{\underbrace{\varphi_T}_{\text{observable}} = \varphi_{\Pi} \cdot \underbrace{\varphi_W}_{\text{known}}}^{\text{Characteristic functions}}$$



Non-parametric step – requires regularization (*Kernel Deconvolution*)

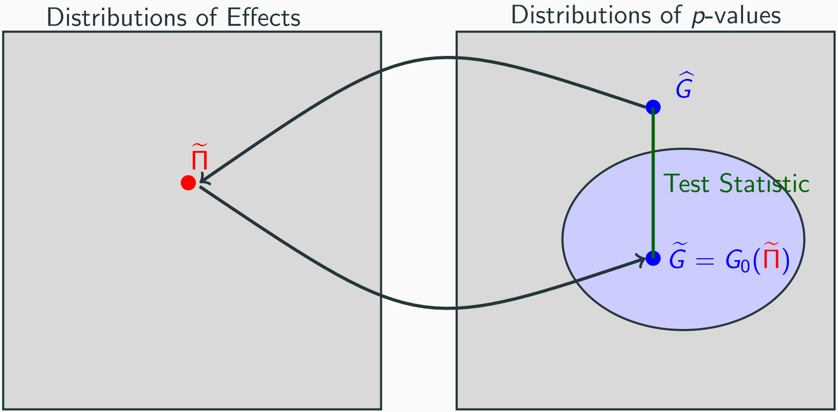
New (More Powerful) Tests

$$\underbrace{T}_{\text{observable}} = \underbrace{h}_{\sim\Pi} + \underbrace{W}_{\sim\mathcal{N}(0,1)} \Rightarrow \overbrace{\underbrace{\varphi_T}_{\text{observable}} = \varphi_{\Pi} \cdot \underbrace{\varphi_W}_{\text{known}}}^{\text{Characteristic functions}}$$



Convenient to bypass estimation of Π and focus on $G_0(\Pi)$

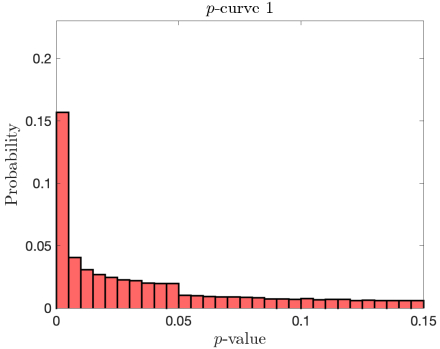
New (More Powerful) Tests

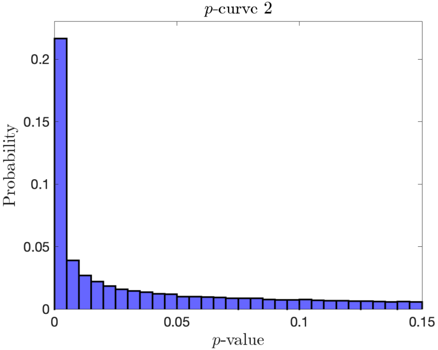


- Kolmogorov-Smirnov (KS) distance: $T_\infty := \left\| \widehat{G} - \widetilde{G} \right\|_\infty$
- Distance between histograms: $\widehat{G}(x_1) - \widetilde{G}(x_1), \ldots, \widehat{G}(x_J) - \widetilde{G}(x_J)$

Alternative: Which Distributions indicate Selective Reporting?

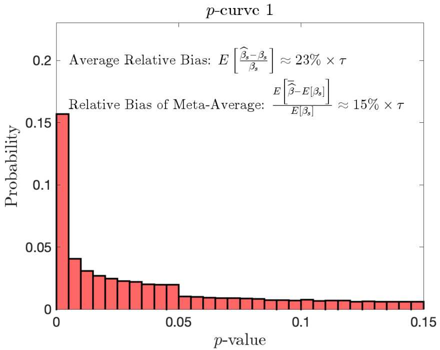
Which literature is p-hacked?

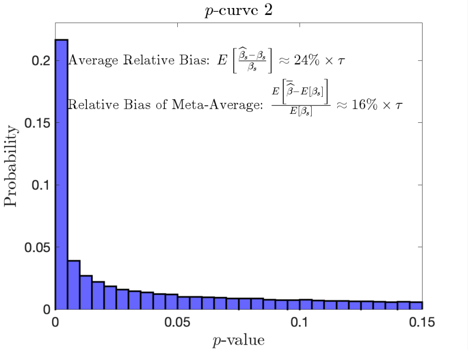


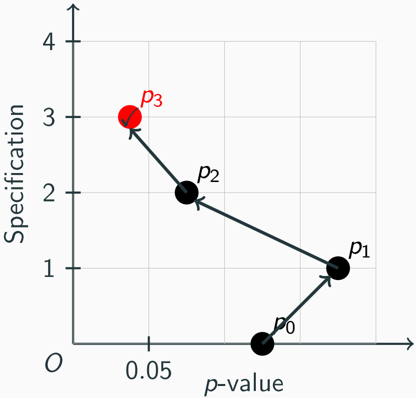


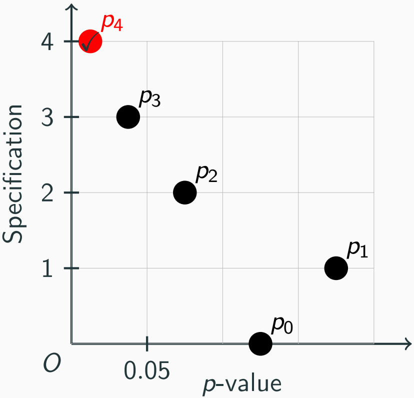
Alternative: Which Distributions indicate Selective Reporting?

Both! Each contains $100 \times \tau\%$ of results reported selectively. (Here $\tau = 0.5$)





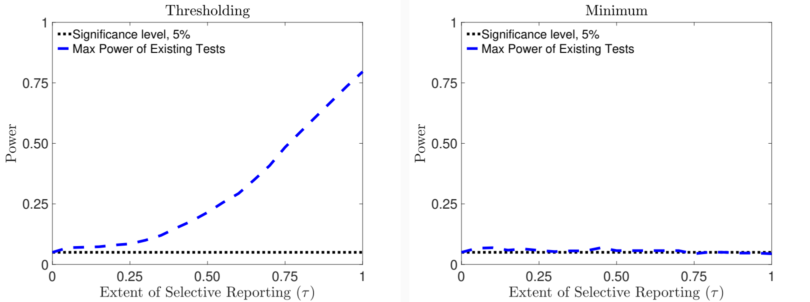




- *Thresholding* and *Minimum* approaches
- *p*-hacking “slow” and “fast” (Data Colada terminology)
- Generate very different distributions

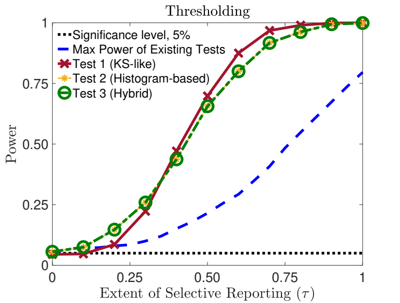
Power Improvements

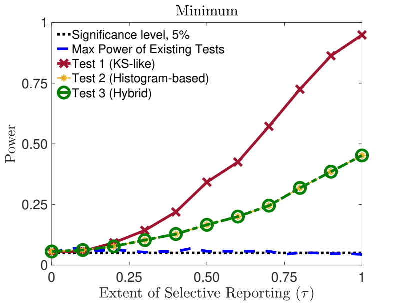
- DGPs tailored to existing dataset of published results
- Two types of p-hacking as before
- Sample size 1000



Power Improvements

- DGPs tailored to existing dataset of published results
- Two types of p-hacking as before
- Sample size 1000





Additionally, provide a **lower bound estimate on τ**