Statistical Vices and Virtues

CLPS 2908

L7 | Feb 21, 2019

Consider Your Results

Case 1: Most of your studies show the effect, $p < .05$

- 5 of 6 3 of 4 2 of 3 1 of 2
- What to do with the outlier study: report? ignore? try to explain?
- What if the outlier study is first or last?
- Remedy: effect sizes, mini meta-analysis, moderators
  - **New best practice:** ***report everything, somewhere***

Case 2: You run 40 subjects.

- **A.** You get $p = .20$. You run 20 more subjects, $p < .05$.
- **B.** You get $p < .05$. You don't run any more subjects.
  - For A: *Report as two studies, one replicating the first; could do mini meta-analysis*
  - For B: *Replicate*

50% check significance before deciding to run more subjects

John, L. K., Loewenstein, G. & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23, 5234-532.

Origins of "Significance" Threshold

Statistical "significance" (Fisher, 1925)

- .05, a "convenient" cut-off
- Later relativized, but too late… *categorical* thinking had taken hold

The "null hypothesis" — why are really afraid of it

- Two ways of being wrong; and their differential costs
  - α error = (claim | false). *We avoid it like the pest, allowing* ***.05*** *of this error rate.*
  - β error = (deny | true); 1 - β = (declare | true) = statistical power.
    - *We are very tolerant (or ignorant), allowing* ***.20*** *(or often more) of this error rate.*

The reasonable concern:

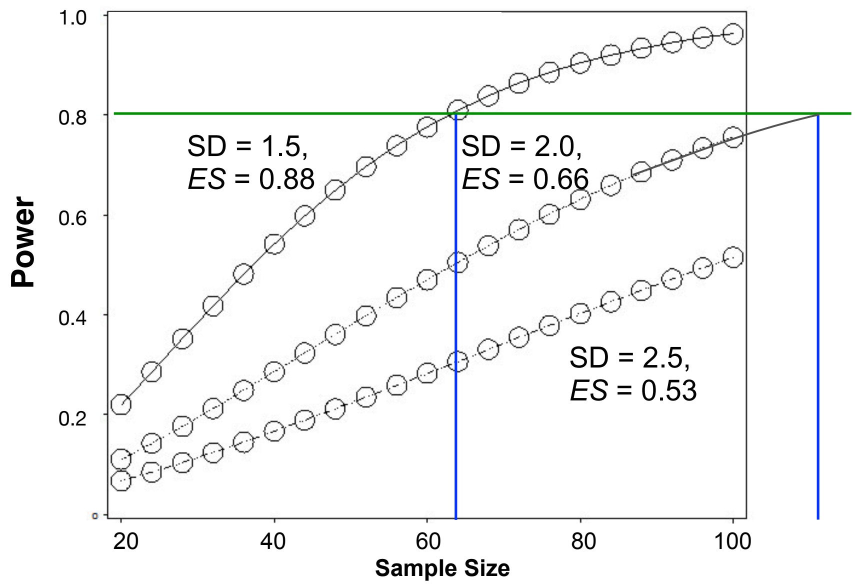
- We shouldn't believe false claims (but should we be ignorant about true claims?)
- Maybe worse: Colquhoun, 2014: 30% of our $p < .05$ results are false claims

Significance Testing Procedure

Problems with standard approach:

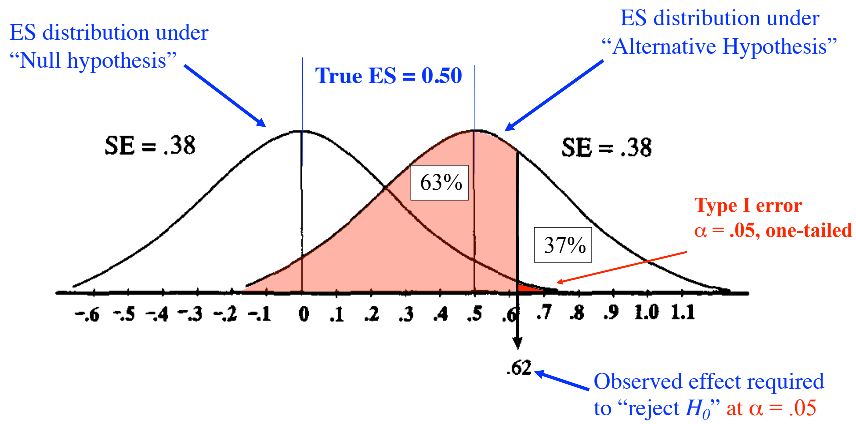
- Dichotomy vs. continuum
- Doesn't actually test hypotheses of interest (e.g., H_1 vs. H_2)
  - We are testing against chance (is H_0 our greatest concern?)
- Ignores effect sizes, what *actually* generalizes
  (**not** a probability of replication but p(data | H_0).
- Power problems, partly due to overly stringent α, especially
  in replications due to effect size sampling errors.
- Accepting/rejecting individual tests *vs.* accumulation
  of results (Rosenthal & Rosnow, Schmidt) ➛ **meta-analysis**

Powerless: Two-sample *t* test, $M_1 = 2.8$, $M_2 = 4.1$



Assume: true ES = .50, s_e(ES) = .38

Data: $N_E = N_C = 15$



Power = 37%. Type II error = 63%

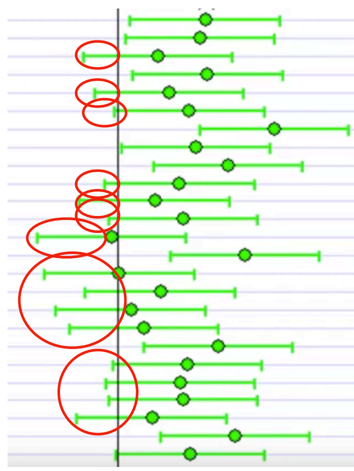
Type I error rate = *actually* 0%.

p-value Roulette

https://www.youtube.com/watch?v=ez4DgdurRPg

Simulating d = .50
at power ~ .50

How many times would you (or your advisor) gamble before you stop if the criterion of winning is $p < .05$?



Amplifiers of *p* obsession

Publication pressure (➛ researcher decisions)
Publication bias (➛ audience judgments)
Little public regulation until recently

Replication crisis has changed things

Lilienfeld, S. (2010) *Scientific American, 303, 18*

Two factors fostering confirmation bias in science:

1. Data show that eminent scientists tend to be more arrogant and confident than other scientists.
   ➟ especially vulnerable to confirmation bias
2. Pressure on scholars to conduct single-hypothesis-driven research programs supported by huge federal grants is a recipe for trouble.
   ➟ highly motivated to disregard or selectively reinterpret negative results that could doom one's career.

p Practices

Past: Search until you find...

- EDA done right and wrong

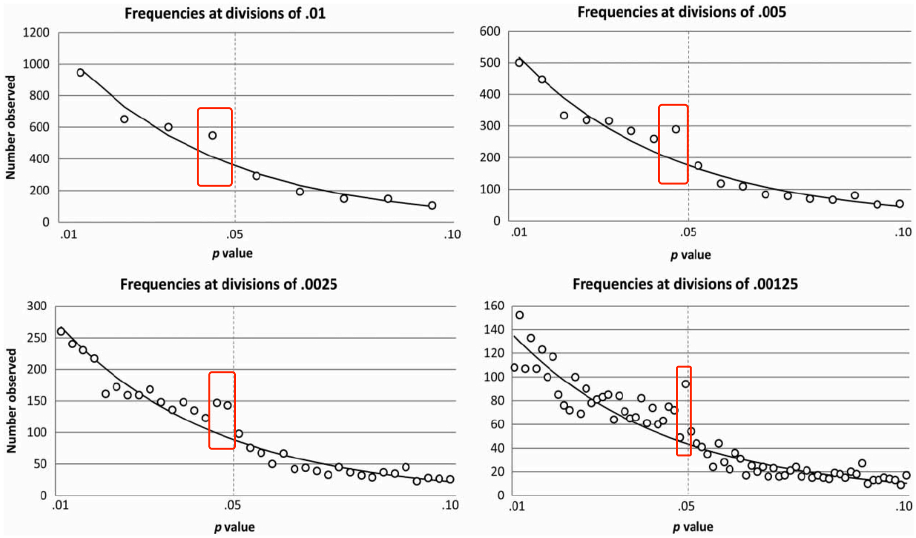
What you needed to reach until recently: $p < .05$, $p < .01$, $p < .001$

- Sometimes incredible efforts to get below the threshold
- **New best practice:** ***report exact p;*** **leave interpretation to body of evidence**

Effect size reporting

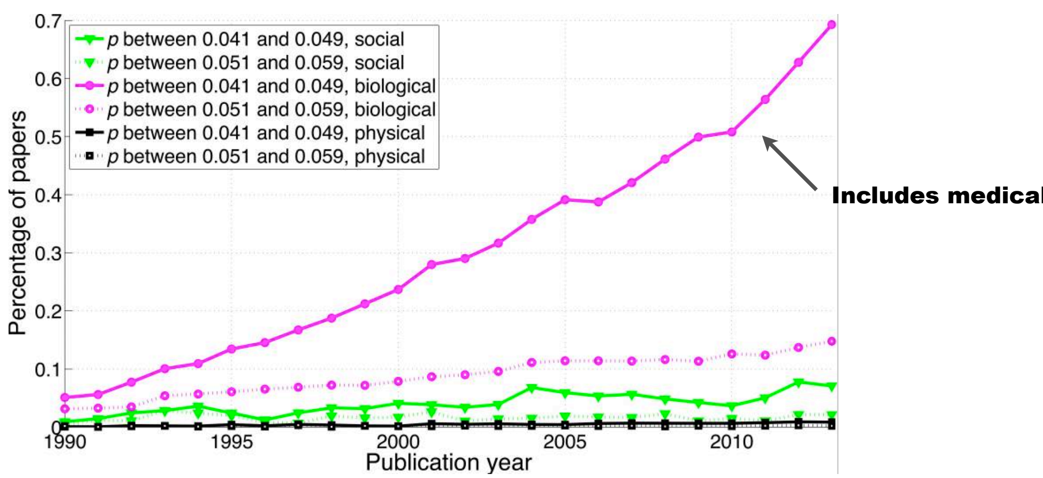
- Past: strategic pactices (yes, when large and $p > .05$; no, when small but $p < .05$)
- **New best practice:** ***always*** **report (typically** ***d*****)**

E. J. Masicampo & Daniel R. Lalande (2012): A peculiar prevalence of *p* values just below .05. *The Quarterly Journal of Experimental Psychology*



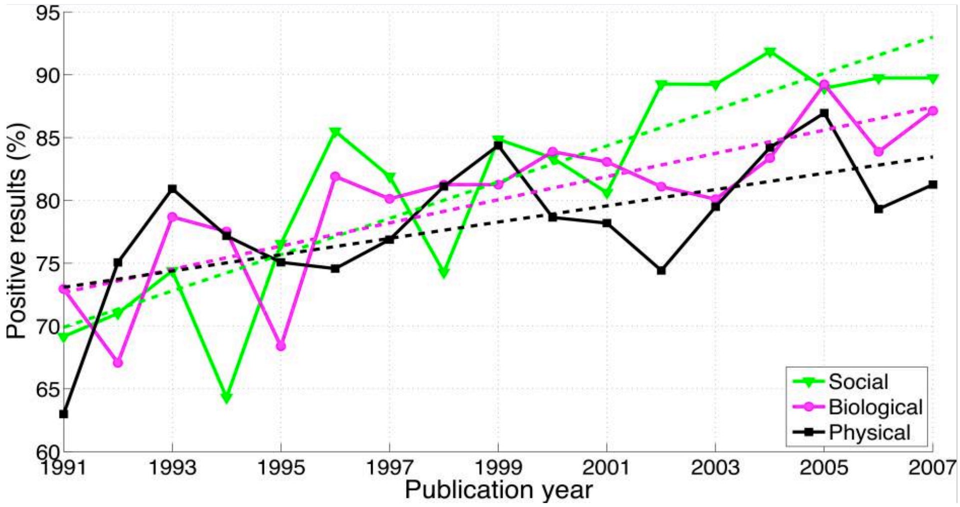
3,627 p values from JEP: General; JPSP; Psychological Science: 12 issues each

Surge of *p*-values between .041 and .049 across time and disciplines



Wicherts, Bakker, and Molenaar (2011) found that researchers were especially unlikely to share their published data for reanalysis if their *p* values were just below .05.

Percentage of positive results in published articles



A Collection of Vices

Two groups, N = 20 each.

Situation A: three *t* tests, one on each of two dependent variables and a third on their average. Pick.

Situation B: one *t* test after collecting *n* = 20 per cell and another after collecting additional *n* = 10 per cell.

Situation C: one *t* test, an analysis of variance adding a gender main effect, and an analysis of covariance with a gender interaction.

	Effective False Positives *if:*		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.