Reconsidering the p-value

Readings for today

- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "p < 0.05". The American Statistician, 73(S1), 1-19.

Topics

1. Null hypothesis test statistics (NHTSs)

2. ATOM approach

Null hypothesis test statistics

The replication crisis

Problem: Uncontrolled sources of variability can lead to dramatically different findings, even when using identical methods.

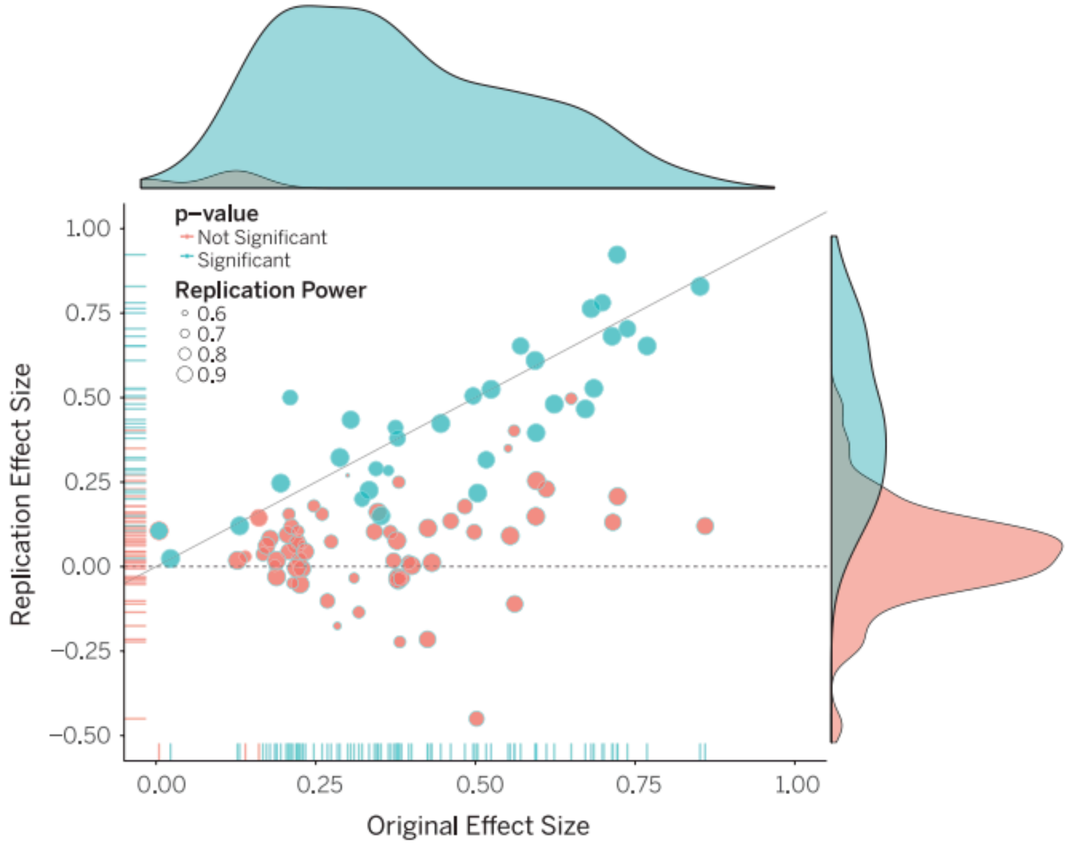
RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

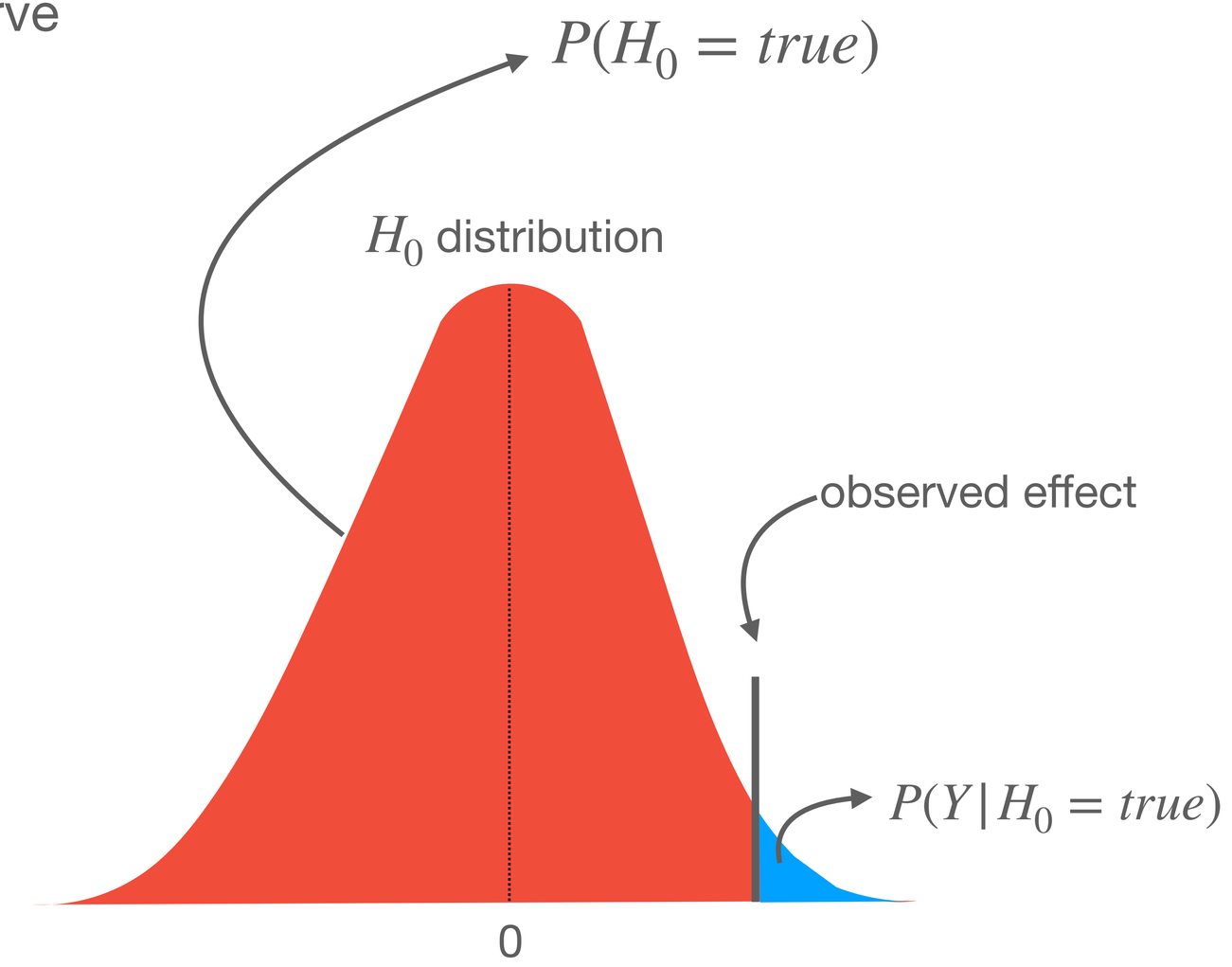
(Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).)



The null hypothesis (H_0)

H_0: The probability that you observe your data if your predicted relationships are not true.

Sometimes “chance”, but not always.



(Wasserstein et al. 2019)

Null hypothesis test statistics (NHTSs)

test statistic

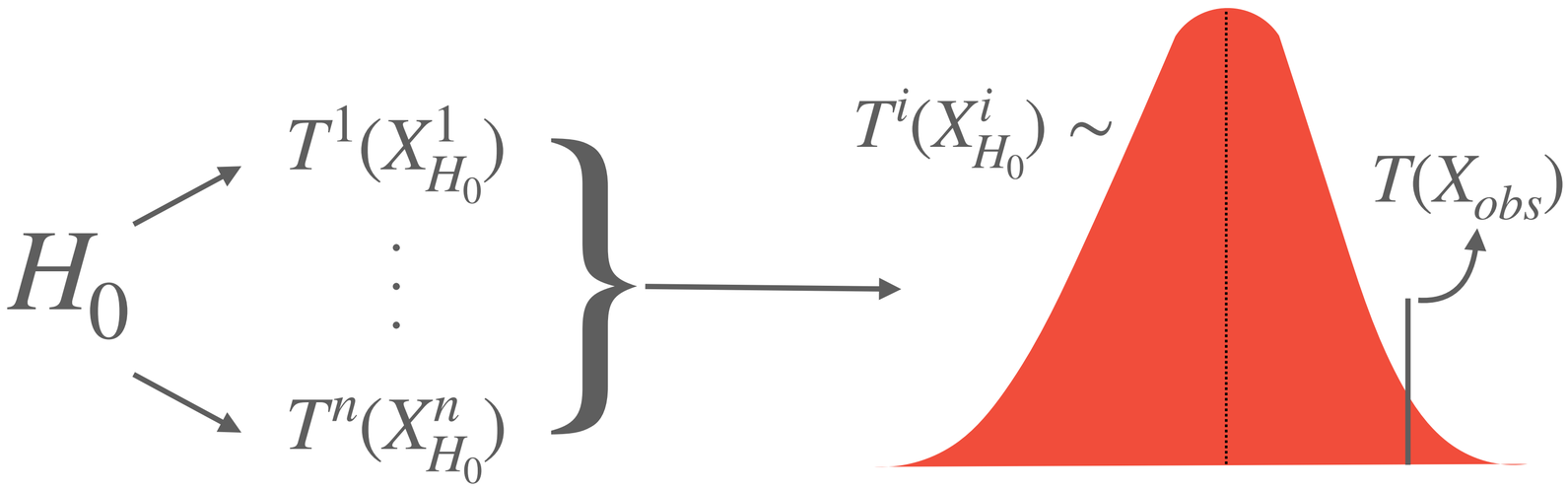
$$p = P(Y \mid H_o = true) = P(T(X_{H_0}) \geq T(X_{obs}))$$

observed data

null hypothesis

predictor variables generated by H_0

predictor variables you observe



(Wasserstein et al. 2019)

Problems with NHTSs

Fundamental limitations:

1. Depend on unobserved data. ← true even for permutation tests.
2. Depend on subjective intentions. ← e.g., p-hacking
3. Do not quantify statistical evidence. ← existence ≠ evidence
4. Poorly understood. ← e.g., p=0.051 is not that different than p=0.049.

(Wasserstein et al. 2019)

Poor practices with NHTSs

NHTSs (e.g., p-values) are simply meant to be a flag to indicate whether results warrant further inspection.

p-values are the beginning, not the end.

<u>Common poor practices:</u>

- Concluding <u>solely</u> on “statistical significance” (e.g., $p < 0.05$)
- Believing that an effect exists simply because of “statistical significance”
- Believing that an effect <u>does not</u> exist because of lack of “statistical significance”
- Believing that the p-value gives you the probability of “chance” alone.

(Wasserstein et al. 2019)

ATOM approach

ATOM

Accept uncertainty

(Be) **T**houghtful

Openness

(Be) **M**odest

First principles for using NHTSs

(Wasserstein et al. 2019)

Accept uncertainty

Acknowledge that uncertainty exists & test for it.

1. Seek <u>better</u> or <u>more comprehensive</u> measures of uncertainty. ← always measure uncertainty
2. Increase focus on <u>data quality</u> and <u>experimental design</u>. ← “poop” in, “poop” out
3. Move away from the false certainty of “statistical significance” ← p-value ≠ truth

(Wasserstein et al. 2019)

(Be) Thoughtful

Develop a holistic approach to inferring from your data.

look at the whole story!

1. Be mindful of data quality at all stages.
2. Consider context & prior evidence.
3. Look ahead to how results predict future outcomes.
4. Be mindful of effect sizes.
5. Consider both mechanisms and possible confounds.
6. Think of your entire toolbox of methods.
7. Effectively communicate your uncertainty in different ways.

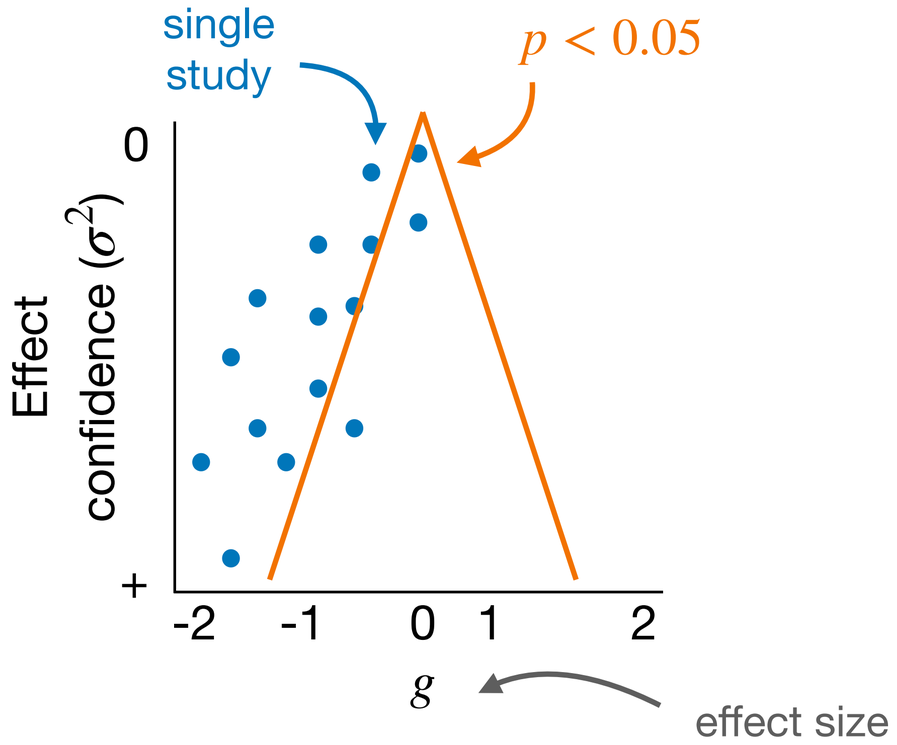
(Wasserstein et al. 2019)

Openness

Adopt “open science” practices

1. Release your code.
2. Release your data.
3. Communicate your methods clearly & completely.
4. Maximize replicability.

Signature of “file drawer” problem



selection bias = $\uparrow \sigma^2$ with $\uparrow g$

(Wasserstein et al. 2019)

(Be) Modest

Being overzealous and inflating conclusions is only effective in the short-term.

1. There is no “perfect study”.
2. Inflation of conclusions reduces replicability.
3. Be a neutral judge of your findings.
    - Temper conclusions
    - Encourage others to replicate

Aggrandizement is a means of impairing confidence in your findings.

(Wasserstein et al. 2019)

ATOM

Accept uncertainty

(Be) **T**houghtful

Openness

(Be) **M**odest

Think of NHSTs as a single tool in an arsenal of many tools that look at your data in different ways.

(Wasserstein et al. 2019)

Take home message

- The p-value is a very useful existential test of your null hypothesis, but should be used in conjunction with many other tests to draw reliable conclusions from your data.