Publication Policies for Replicable Research: Addressing the False Publication Rate

Josh Habiger with Ye Liang

Oklahoma State University
Department of Statistics

October 24, 2022

Case Study: Replication Crisis

Hidden Brain: The Scientific Process, NPR

1. Stereotype Susceptibility: Identity Salience and Shifts in Quantitative Performance, *Psychological Science* (1999), Shih et. al.
   - Asian Women reminded of either heritage or gender before math test.
   - Conclusion: Stereotypes can positively or negatively affect performance.
   - Significant impact in Psychology (**760 citations**) until...
2. Replication studies in *Social Pscyhology* 2014 speacial issue
   - Replication Attempt of Stereotype Susceptibility, Gibson et. al.
   - A Second Replication Attempt of Stereotype Susceptibility, Moon and Roeder
3. Conclusion drawn: At least a third of scientists are wrong

Society vs Statisticians

ASA's Statement on p-Values: Context Process and Purpose

- Why: Statisticians / *p*-values getting blamed for replicability crisis in popular media
- When: 2016
- Who: ASA Board and 21 Expert Statisticians
- What:
  - 1 Correct interpretation of *p*-value
  - 5 Common misuses:
    1. $Pr(H_0|Data)$
    2. practical significance
    3. $P < .05$
    4. selective p-value reporting
    5. use of only a *p*-value
- Summary: "It does not tell us what we want to know, and we so desperately want to know what we want to know that, out of desperation, we nevertheless believe that it does" - Cohen (1994)

The Aftermath of ASA Statement

- Q: Is there an "Incompleteness in the foundations of Statistics"?
  - Krants (1999). JASA review of book "What if there were no significance tests"
- A: Answers may vary, *some*[a] new ideas
  - 5000+citations
  - TAS Special issue on *p*-values
- Refocusing on Replicability Crisis
  - The ASA president's task force statement on replicability and statistical significance
  - *Selective Inference: Silent Killer of Replicability* - Benjamini (2020)
    - Nice review of TAS special issue and context
    - Misguided Attack on p-values
    - Replicability crisis: Consequence of "The Industrialization of the Scientific Process"

[a] "After four decades of severe criticism, the ritual of null hypothesis significance testing - the mechanical dichotomous decisions around a sacred 0.05 criterion - still persists" Cohen (1994)

What is Replicability?

Definition

Replicability is the ability of a scientific experiment X to to be repeated to obtain a consistent result $T(X)$

- An experiment is *reproduced* if $T_{rep}(X_{orig}) \approx T_{orig}(X_{orig})$
- An experiment is *replicated* if $T_{rep}(X_{rep}) \approx T_{orig}(X_{orig})$
- **Reproducible experiments need not replicate**: $T(X_{rep}) \neq T(X_{orig})$
- **Focus: Reproducible experiments + replication of "publication" (say $P_{orig} < 0.05$)**

Replication Probabilities for $P < \alpha$

Assume perfect replication: $P_{rep} \stackrel{d}{=} P_{orig}$

- Probability of replicating publication ($P_{orig} < \alpha$):

$$\begin{aligned} \Pr(P_{rep} < \alpha) &= \Pr(H_0) \times \Pr(P_{rep} < \alpha | H_0) + \Pr(H_1) \Pr(P_{rep} < \alpha | H_1) \\ &= \pi_0 \times \alpha + (1 - \pi_0) \times (1 - \beta) \end{aligned}$$

  - $1 - \beta$: Power / Probability of replicating if H_0 false
  - π_0: Probability / Proportion true nulls tested
  - α: Type 1 error / Probability replicating if H_0 true

- Proportion of false positives we're trying to replicate

$$\text{FPR} = \Pr(H_0 | P_{orig} < \alpha) = \frac{\pi_0 \times \alpha}{\pi_0 \times \alpha + (1 - \pi_0) \times (1 - \beta)}$$

Open Science Collaboration Data

- 100 published studies from in mainstream Pscyhology journals in 2008 replicated
- *P*-values for original and replicated studies available for 73
- How many studies will replicate statistical significance?

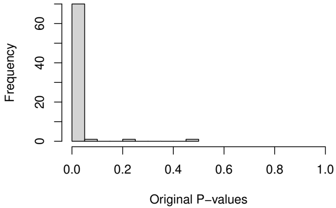


Figure: Histogram of 73 Originally Published p-values

Open Science Collaboration Data

- 100 published studies from in mainstream Pscyhology journals in 2008 replicated
- *P*-values for original and replicated studies available for 73
- How many studies will replicate statistical significance $P < .05$? 27

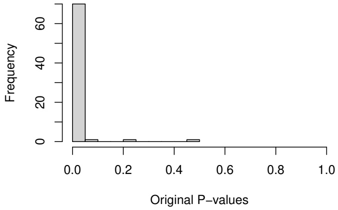


Figure: Histograms of 73 Originally Published and 73 Replicated *p*-values

What Happened?

Q1: Why would only 27 / 70 replicate statistical significance?

A1: On the Reproducibility of Psychological Science, *JASA*, Johnson et. al. (2017).

- About half are False Positives!

$$\widehat{FPR} = \frac{\pi_0 \times \alpha}{\pi_0 \times \alpha + (1 - \pi_0) \times (1 - \beta)} = 0.49$$

- $\hat{\pi}_0 = 0.87$ or 0.93!!!!

Q2: Is this a broader problem?

A2: YES If $P < 0.05$ and π_0 large.

- "Why most published research findings are false" - Ioannidis (2005).
- "The Industrialization of the Scientific Process" in Benjamini (2020)

Proposed Solutions

Some recommendations in the literature:

- Report $P = p$ (not $P < p$) and other summaries - Wasserstein et.al, Summary of TAS Special issue on p-values, TAS (2019)
- Redefine Statistical Significance: $P < 0.005$ - Benjamin + 71, Nature (2017)
- Justify your α - Lakens + 87, Nature (2018)

What most proposed solutions have in common

1. Recognize issue: **False Positive Rate**
2. Recognize cause: π_0 large and/or $(1-\beta)$ small.
3. Recognize solution: Must involve π_0 and/or $(1-\beta)$.

What is missing from the literature? **A FORMAL FRAMEWORK**

Basic Elements

Data/p-value/test stat: $X_1, X_2, ..., X_m$

Hypotheses: $H_1, H_2, ..., H_m$ null hypotheses ($H_i = 0$ or 1)

Basic Model: $X_i \sim f_i(x) = \pi_{0i} f_{0i}(x) + (1 - \pi_{0i}) f_{1i}(x)$

- p-Value Model: $f_i(p) = \pi_{0i} + (1 - \pi_{0i})\gamma_i p^{\gamma_i - 1}$
- Ex: $pow_i = 0.05^{\gamma_i}$ or $\gamma_i = \log(pow_i)/\log(.05)$

Publication Decisions: $\delta_1, \delta_2, ..., \delta_m$ where $\delta_i = 0$ ("do not publish") or 1 ("publish")

- If $\delta_i = 0$, X_i unobservable
- **If $\delta_i = 1$, some information is observable**
- δ_i should depend on X_i but need not
- **Required:** $E[\delta_i]$ is well defined so replicable

The Thesis

Local false discovery rate:

$$lfdr_i = \frac{\pi_{0i} f_{0i}(x_i)}{f_i(x_i)} = \frac{\pi_{0i}}{\pi_{0i} + (1 - \pi_{0i})\gamma_i p_i^{\gamma_i - 1}}$$

False Publication Rate:

$$FPR = E\left[\frac{\sum_i (1 - H_i)\delta_i}{\sum_i \delta_i \vee 1}\right]$$

Thesis

If the **local false discovery rate** is published whenever a study published, then the false publication rate can be addressed.

- $lfdr_i$ is minimally sufficient
- $(P_i, \pi_{0i}, \gamma_i)$ or (P_i, π_{0i}, pow_i) are sufficient
- P_i is not sufficient

FPR Control

Theorem

Let $H_1, H_2, ..., H_m$ be hypotheses of interest from studies $X_1, X_2, ...X_m$. Consider community wide publication policy

$$\delta(\mathrm{x}) = [I(lfdr_1 < t), I(lfdr_2 < t), ..., I(lfdr_m < t)].$$

If $\mathrm{X} = (X_1, X_2, ...X_m)$ is a mutually independent collection then $FPR \leq t$.

- Remark: $FPR << t$
- Corollary: If $t(\alpha)$ chosen such that $\overline{lfdr} < \alpha$ among published $lfdr$ values, then $FPR \leq \alpha$.
  - Not directly practical

FPR Estimation

Theorem

Consider a well defined decision process that results in r published lfdr values. Define estimate

$$\widehat{FPR} = \frac{1}{r}\sum_{i=1}^{r} lfdr_i.$$

Then $E[\widehat{FPR}] = pFPR \approx FPR.$

- Important: Only need to 1. observe *lfdr* *among published studies* and 2. compute an average
- Examples:
  - $\delta_i = I(lfdr_i < t)$
  - $\delta_i = I(lfdr_i < t_i)$ for t_i chosen based on impact, scope, ...
  - $\delta_i = I(p_i < \alpha)$
  - $\delta_i = I(p_i < \alpha_i)$ for α_i chosen based on impact, scope, ...
  - $\delta_i = I(\text{Heads on coin toss})$

- Motivation: Unlikely that π_{0i} and pow_i are known precisely in practice.

Summary of Theorems 2 and 4: Methods are Robust

The *FPR* is controlled and conservatively estimated if

$$E[\overline{lfdr}] \leq E[\overline{lfdr}']$$

for $lfdr_i'$ used in estimation or policy, but $lfdr_i$ is correct.

- Interpretation: Conservative specifications of parameters work but aren't necessary
- Examples:
  - $\pi_0' \geq \max_i \pi_{0i}$
  - $\pi_0' \geq E[\pi_{0i}]$

Revisiting Open Science Collaboration Data

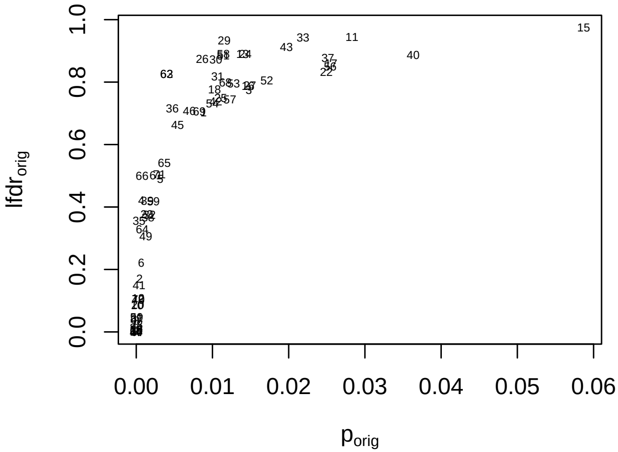
Question: What if the 73 original studies had provided *lfdr*-values? What could we have learned without a replication?

Answer: First some details from Johnson et. al (2017) and Habiger and Liang (2022)

- $H_i : \rho_i = 0$
- $\hat{\pi}_0 = 0.87$ and $\hat{\pi}_0 = 0.93$
- $BF_i = BF(n_i, z_i)$
- $lfdr_i = \frac{\hat{\pi}_0}{\hat{\pi}_0 + (1 - \hat{\pi}_0)BF_i} = \frac{0.93}{0.93 + 0.07 BF_i}$

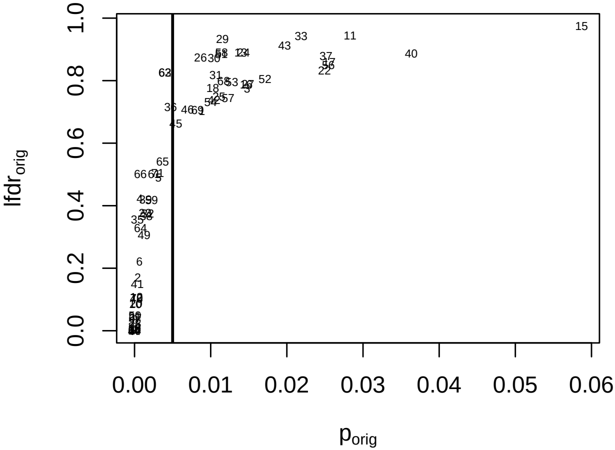
FPR Estimates: Original Publication Rule

$$\widehat{FPR} = \overline{lfdr} = 0.52$$



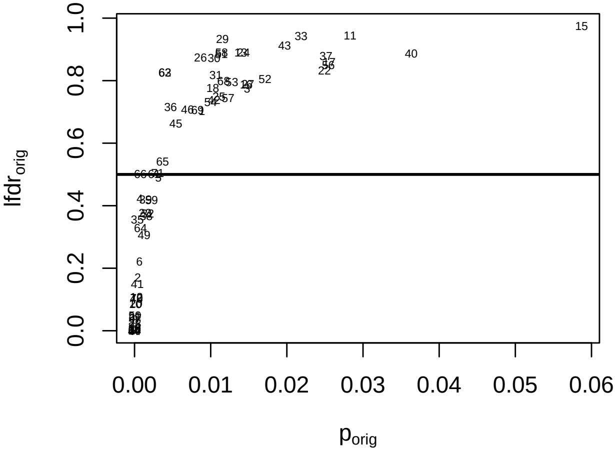
FPR Estimates: Original Publication Rule + $P < 0.005$

$$\widehat{FPR} = \overline{lfdr} = 0.24$$



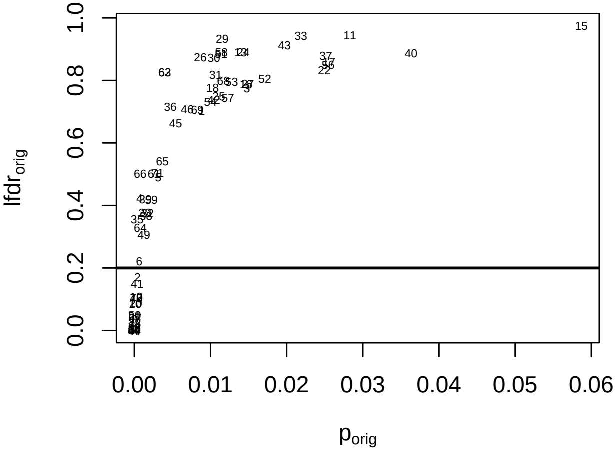
FPR Estimates: Original Publication Rule + lfdr< .5

$$\widehat{FPR} = \overline{lfdr} = 0.16$$



FPR Estimates: Original Publication Rule + lfdr$<$ .2

$$\widehat{FPR} = \overline{lfdr} = 0.05$$



Recap

- Replicability crisis: Due (in part) to high false positive rate attributable to $P < 0.05$ when π_0 large and/or pow_i low.
- Contributions of Habiger and Liang (2022): Formalize proposed solutions
  - (P_i, π_{0i}, pow_i) are sufficient
  - $lfdr_i$ is minimally sufficient
  - P_i is not sufficient
- Illustration: If $lfdr_i$'s are reported then simple solutions / estimators are available.

What's Next?

Most Obvious Limitation: $lfdr_i = lfdr(X_i, \pi_0, pow)$

- Patience - think decades
- Next steps
  - Methodological ($\hat{\pi}_0$, $\hat{\gamma}_i$)
  - Broad dissemination
  - Publication policy δ
- Marketing
- Resilience

References

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E. J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. A. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, and V. E. Johnson (2017, September). Redefine statistical significance. *Nature Human Behaviour 2*(1), 6–10.

Benjamini, Y. (2020, Dec). Selective inference: The silent killer of replicability. *Harvard Data Science Review 2*(4). https://hdsr.mitpress.mit.edu/pub/l39rpgyc.

Habiger, J. and Y. Liang (2022). Publication policies for replicable research and the community-wide false discovery rate. *The American Statistician 76*(2), 131–141.

Ioannidis, J. P. (2005, Aug). Why most published research findings are false. *PLoS Med. 2*(8), e124.

Johnson, V. E., R. D. Payne, T. Wang, A. Asher, and S. Mandal (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association 112*(517), 1–10. PMID: 29861517.

Lakens, D., F. G. Adolfi, C. Albers, F. Anvari, M. Apps, S. Argamon, T. Baguley, R. Becker, S. D. Benning, D. Bradford, E. M. Buchanan, A. R. Caldwell, B. Calster, R. Carlsson, S. chin Chen, B. Chung, L. J. Colling, G. Collins, Z. Crook, E. S. Cross, S. Daniels, H. Danielsson, L. DeBruine, D. J. Dunleavy, B. Earp, M. I. Feist, J. D. Ferrell, J. G. Field, N. W. Fox, A. Friesen, C. Gomes, M. Gonzalez-Marquez, J. Grange, A. Grieve, R. Guggenberger, J. Grist, A.-L. Harmelen, F. Hasselman, K. D. Hochard, M. Hoffarth, N. Holmes, M. Ingre, P. Isager, H. Isotalus, C. Johansson, K. Juszczyk, D. Kenny, A. Khalil, B. Konat, J. Lao, E. G. Larsen, G. Lodder, J. Lukavský, C. Madan, D. Manheim, S. R. Martin, A. E. Martin, D. Mayo, R. J. McCarthy, K. McConway, C. McFarland, A. Nio, G. Nilsonne, C. L. Oliveira, J. O. Xivry, S. Parsons, G. Pfuhl, K. Quinn, J. J. Sakon, S. A. Saribay, I. Schneider, M. Selvaraju, Z. Sjoerds, S. G. Smith, T. Smits, J. R. Spies, V. Sreekumar, C. N. Steltenpohl, N. Stenhouse, W. Światkowski, M. A. Vadillo, M. V. Assen, M. Williams, S. E. Williams, D. R. Williams, T. Yarkoni, I. Ziano, and R. A. Zwaan (2018). Justify your alpha. *Nature Human Behaviour 2*, 168–171.

Wasserstein, R. L. and N. A. Lazar (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician 70*(2), 129–133.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar (2019). Moving to a world beyond $p<0.05$. *The American Statistician 73*(sup1), 1–19.