

Statistical Significance Makes our **Mission: Impossible**

StanCon @ Wash U

Mariel Finucane

June 23, 2023



Mathematica's mission

To **improve public well-being** by
advancing evidence-based decision
making for global impact



Pitfalls of p -values



AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • [www.twitter.com/AmstatNews](https://twitter.com/AmstatNews)

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND P -VALUES

*Provides Principles to Improve the Conduct and Interpretation of Quantitative
Science*

March 7, 2016

Pitfalls of p -values

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 1–19: Editorial
<https://doi.org/10.1080/00031305.2019.1583913>

EDITORIAL

Moving to a World Beyond “ $p < 0.05$ ”

Pitfalls of p -values

Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

When was the last time you heard a seminar speaker claim there was 'no difference' between two groups because the difference was 'statistically non-significant'?

If your experience matches ours, there's a good chance that this happened at the last talk you attended. We hope that at least someone in the audience was perplexed if, as frequently happens, a plot or table showed that there actually was a difference.

How do statistics so often lead scientists to deny differences that those not educated in statistics can plainly see? For several generations, researchers have been warned that a statistically non-significant result does not 'prove' the null hypothesis (the hypothesis that there is no difference between groups or no effect of a treatment on some measured outcome)¹. Nor do statistically significant results 'prove' some other hypothesis. Such misconceptions have famously warped the

literature with overstated claims and, less famously, led to claims of conflicts between studies where none exists.

We have some proposals to keep scientists from falling prey to these misconceptions.

PERVASIVE PROBLEM

Let's be clear about what must stop: we should never conclude there is 'no difference' or 'no association' just because a P value is larger than a threshold such as 0.05 ►

© 2019 Springer Nature Limited. All rights reserved.

21 MARCH 2019 | VOL 567 | NATURE | 305



The Washington Post

Democracy Dies in Darkness

Scores fall coast to coast, especially in math, under pandemic's toll



By [Laura Meckler](#)

Updated October 24, 2022 at 2:57 p.m. EDT | Published October 24, 2022 at 12:01 a.m. EDT

<https://www.washingtonpost.com/education/2022/10/24/pandemic-learning-loss-naep-tests/>

... “Most of the 26 large city school districts... saw **no change**, which qualifies as a bright spot given the overall results.”

Take-home points

- / Evaluators need to be able to answer the question: "*What is the probability that the policy worked?*"
- / Under the null hypothesis significance testing framework, the correct answer is: "*We don't know.*"
- / Bayesian methods can provide an answer...
- / ... But they require hard-headed prior evidence.

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P < 0.10$ LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥ 0.1	

<https://xkcd.com/>



Motivating example

/ **We evaluated the impact on voting of receiving an offer of admission to Democracy Prep.**

/ **Research questions:**

1. What is the probability that Democracy Prep increased voting?
2. By at least 10 percentage points?





p -values cannot answer our research question

- / We estimated that DP increased voting by 24 percentage points, more than doubling the expected voting rates of its students, $p < 0.05$.
- / Correct interpretation: If DP had zero effect on voting, there would be a < 5 percent chance of estimating an impact of 24 percentage points or greater.
- / This does not tell us the probability that DP increased voting.

Take-home points

- / Evaluators need to be able to answer the question: "*What is the probability that the policy worked?*"
- / **Under the null hypothesis significance testing framework, the correct answer is: "*We don't know.*"**
- / Bayesian methods can provide an answer...
- / ... But they require hard-headed prior evidence.

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P < 0.10$ LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥ 0.1	

<https://xkcd.com/>



Imagine that federal grants fund 100 locally developed programs

/ The truth (unknown to policymaker):

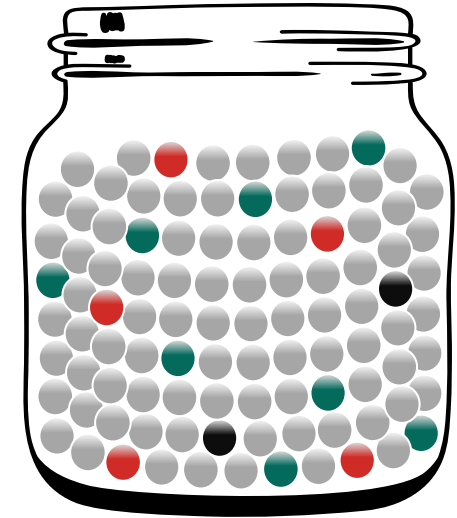
- 10 program have a meaningful impact
- 90 programs have no effect

/ 1 program is evaluated via RCT:

- Statistical testing with $\alpha = 0.05$
- 80% power to detect a meaningful impact

$p < 0.05 \neq \text{"Eureka!"}$

- / **8 Green**: significant and truly effective
- / **2 Black**: insignificant but effective
- / **5 Red**: significant but ineffective
- / **85 Grey**: insignificant and truly ineffective



Probability a significant result is a false positive =

$$\frac{5 \text{ Red}}{8 \text{ Green} + 5 \text{ Red}} = 38\%$$



38% \neq 5%: What went wrong?

- / Remember, in the example, just 10 percent of programs truly had meaningful effects.
- / P(significant result is a false positive) would be closer to the expected 5 percent if half of all programs had meaningful effects.
- / But if nothing works then P(significant result is a false positive) = 100%.

To calculate the probability a policy had a meaningful effect
we need to know
how often similar interventions have had meaningful effects.

Take-home points

- / Evaluators need to be able to answer the question: "*What is the probability that the policy worked?*"
- / Under the null hypothesis significance testing framework, the correct answer is: "*We don't know.*"
- / **Bayesian methods can provide an answer...**
- / ... But they require hard-headed prior evidence.

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P < 0.10$ LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥ 0.1	

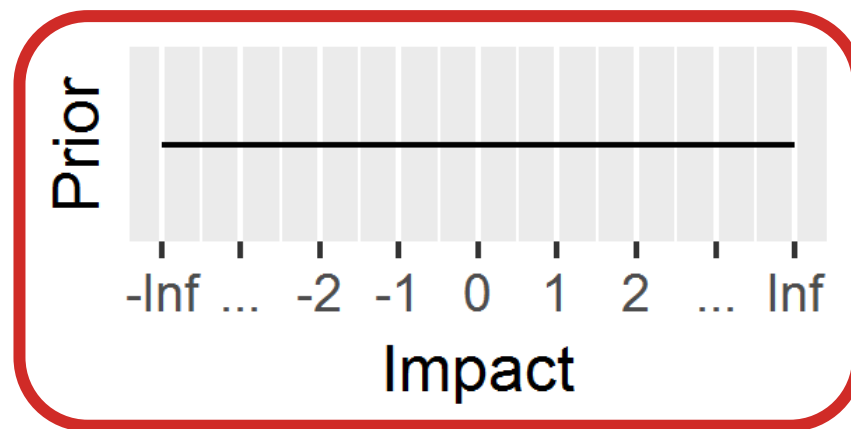
<https://xkcd.com/>



Frequentist vs. Bayesian: What's random and what's fixed?

Frequentist	Fixed: True impact Random: Data	“If the true impact is zero, there is a 5 percent chance of estimating an impact of the observed magnitude or larger”
Bayesian	Fixed: Data Random: True impact	“Given our impact estimate, there is a 64% probability that the intervention increased voting by 10 percentage points or more”

No to the flat prior



- / The flat prior was seen as objective and used to be very popular
- / We reject the flat prior because it has no basis in evidence
- / “The general problem I have with noninformatively-derived Bayesian probabilities is that they tend to be too strong”
 - <https://statmodeling.stat.columbia.edu/2015/05/01/general-problem-noninformatively-derived-bayesian-probabilities-tend-strong/>
- / The flat prior sanctifies misinterpretation of the p -value

Take-home points

- / Evaluators need to be able to answer the question: "*What is the probability that the policy worked?*"
- / Under the null hypothesis significance testing framework, the correct answer is: "*We don't know.*"
- / Bayesian methods can provide an answer...
- / ... But they require hard-headed prior evidence.

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥0.1	


<https://xkcd.com/>



Plausible Priors Precede Persuasive Posteriors

- / **No to prior belief**
- / **Yes to hard-headed prior evidence**
- / **Appropriate for high-stakes impact evaluations**
- / **Specifying an evidence-based prior requires answering two questions:**
 - What broader population of similar evaluations is your evaluation a member of?
 - What is the distribution of impacts in that population?

hard·head·ed

/ˌhɑːdˈhedəd/ 

adjective

adjective: **hard-headed**

practical and realistic; not sentimental.



Rules of thumb for picking prior evidence

1. If we make the infeasible perfect the enemy of the feasible good, then the worst can win
2. Too broad is better than too narrow
3. If the mean effect size in your prior evidence base is far away from 0, something might be wrong
4. When we set *beliefs* aside and focus on *evidence*, picking the prior becomes easier



A hard-headed prior for the impact of Democracy Prep

/ Among 29 published estimates of impacts of educational interventions targeting civic outcomes:

- 90% were positive
- Only 10% were greater than 10 percentage points (pp)

/ Is the DP impact estimate (24 pp) driven by noise or signal?

- **Noise** more likely the wider the CI
- **Signal** more likely the more often similar interventions have had meaningful impacts

Study

DP - Voted in 2016 election

Private school vouchers

Ever registered to vote
Voted in 2008 general election
Voted in 2010 general election
Voted in 2012 general election
Voted in 2008, 2010, or 2012 general election

Catholic schooling

Currently registered to vote (HS&B)
Voted in any election in past year (HS&B)
Voted in 1988 presidential election (HS&B)
Currently registered to vote (NELS88)
Voted in past 2 years (NELS88)
Voted in 1996 presidential election (NELS88)

Additional year of education

Voted in most recent presidential election

Additional year of education (UK)

Voted in most recent general election

High school graduation

Voted in the current year
Voted in November election

College entrance

Currently registered to vote
Voted in any election in past year
Voted in 1988 presidential election

American Government/Civics course (1 semester)

Voted in 1992 presidential election
Voted in 1993-1994 state/local elections
Voted in 1996 presidential election
Voted in any election from 1998-2000
Voted in 2004 presidential election
Voted in any election from 2004-2006

American Government/Civics course (2 semesters)

Voted in 1992 presidential election
Voted in 1993-1994 state/local elections
Voted in 1996 presidential election
Voted in any election from 1998-2000
Voted in 2004 presidential election

0 10 pp
Impact estimate and 95% CI



Which evidence best informs decision making?

Null hypothesis sig. test

- / We estimated that DP increased voting by 24 percentage points, $p < 0.05$
- / #highfive
- / But...
 - To what extent is this estimate driven by noise?
 - Would it replicate?

Bayesian posterior prob.

- / We estimate a 95% chance that DP increased voting...
- / ... but only a 64% chance that DP increased voting by 10 percentage points or more



References

Deke, J. & Finucane, M. (2019). Moving Beyond Statistical Significance: The BASIE (BAyesian Interpretation of Estimates) Framework for Interpreting Findings from Impact Evaluations. <https://www.acf.hhs.gov/opre/resource/moving-beyond-statistical-significance-the-basie-framework-for-interpreting-findings-from-impact-evaluations>

Gelman, A., & Carlin, J. (2017). Some natural solutions to the p -value communication problem—and why they won't work. *Journal of the American Statistical Association*, 112 (519), 899–901.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist*, 97 (4), 310–316.

Gill, B., Whitesell, E. R., Corcoran, S. P., Tilley, C., Finucane, M., & Potamites, L. (2020). Can charter schools boost civic participation? The impact of democracy prep public schools on voting behavior. *American Political Science Review*, 114(4), 1386-1392.

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

Greenland, S., & Poole, C. (2013). Living with p values: Resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology*, 24 (1), 62–68.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can generally only be understood in the context of the likelihood. *Entropy*, 19(555), 1-13.

www.andrewgelman.com

[“Hidden dangers of noninformative priors”](#) Nov 21, 2013

[“Interpreting posterior probabilities in the context of weakly informative priors”](#) June 28, 2015

[“The general problem I have with noninformatively-derived Bayesian probabilities is that they tend to be too strong”](#) May 1, 2015

[“What are some situations in which the classical approach gives worse results than a Bayesian approach?”](#) Nov 13, 2013

[“What is the “true prior distribution”? A hard-nosed answer”](#) April 23, 2016



“We view much of the recent history of Bayesian inference as a set of converging messages from many directions... pointing toward the benefits of including real, subject-matter-specific, prior information.” Gelman et al. 2017

Appendix:

Meta-regression of estimates from the literature

Impact estimates from the literature...

- Are **noisy**
- Likely suffer from **publication bias**
- Are **not independent**

So our Bayesian meta-regression...

$$\hat{\theta}_i \sim N(\theta_i + \beta s_i, s_i^2)$$

$$\theta_i \sim N(\alpha_{j[i]}, \sigma^2)$$

$$\alpha_j \sim N(\mu, \tau^2)$$