Moving to a World Beyond p<0.05

Nicole Lazar, Allen Schirm,
Ron Wasserstein

Begin at the beginning

What is a p-value?

P-value "clarified"

Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

"That definition is about as clear as mud"

-Christie Aschwanden, lead writer for science, *FiveThirtyEight*

Perhaps this is clearer

---

[4]The simplest general definition of a p-value of a point null hypothesis I know of is as follows. Suppose the null hypothesis is that $\mathbb{P}$ is the probability distribution of the data X, which takes values in the measurable space $\mathcal{X}$. Let $\{R_\alpha\}_{\alpha\in[0,1]}$ be a collection of $\mathbb{P}$-measurable subsets of $\mathcal{X}$ such that (1) $\mathbb{P}(R_\alpha) = \alpha$ and (2) If $\alpha' < \alpha$ then $R_{\alpha'} \subset R_\alpha$. Then the p-value of H_0 for data $X = x$ is $\inf_{\alpha\in[0,1]}\{\alpha : x \in R_\alpha\}$.

(Stark, 2016)

So, what is a p-value?

- We know some stuff
- We want to know more
- We design an experiment to help us
- We collect data from the experiment
- We numerically summarize the results
- Now what do we know?

Numerically summarizing the data

- Summarize the data into a number we call a “statistic”
- Compute a probability from that statistic – that’s the p-value
- If the p-value is small enough, call it “statistically significant”
- What is small enough? Typically, 0.05.

Example

- New treatment compared to placebo to improve TKV (a measure of health in kidney patients, in ml)
- After one year, difference in median TKV (treatment – placebo) = **96.8**
  - 95% confidence interval 10.8 to 182.7
- P-value computed to be 0.027, round off to **0.03**
- **https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002777**

Now home in on 0.03

How did we get *that* number?

To compute the p-value

- We assumed a **bunch** of stuff
- One key assumption: **no difference** between the treatment and the placebo

What does the p-value mean?

- If there is no difference between the treatment and the placebo
- If **everything** else we assumed is also true
- Then the probability that we would observe the difference we found (97 ml), or one even larger, is 0.03

Common misinterpretations of the p-value

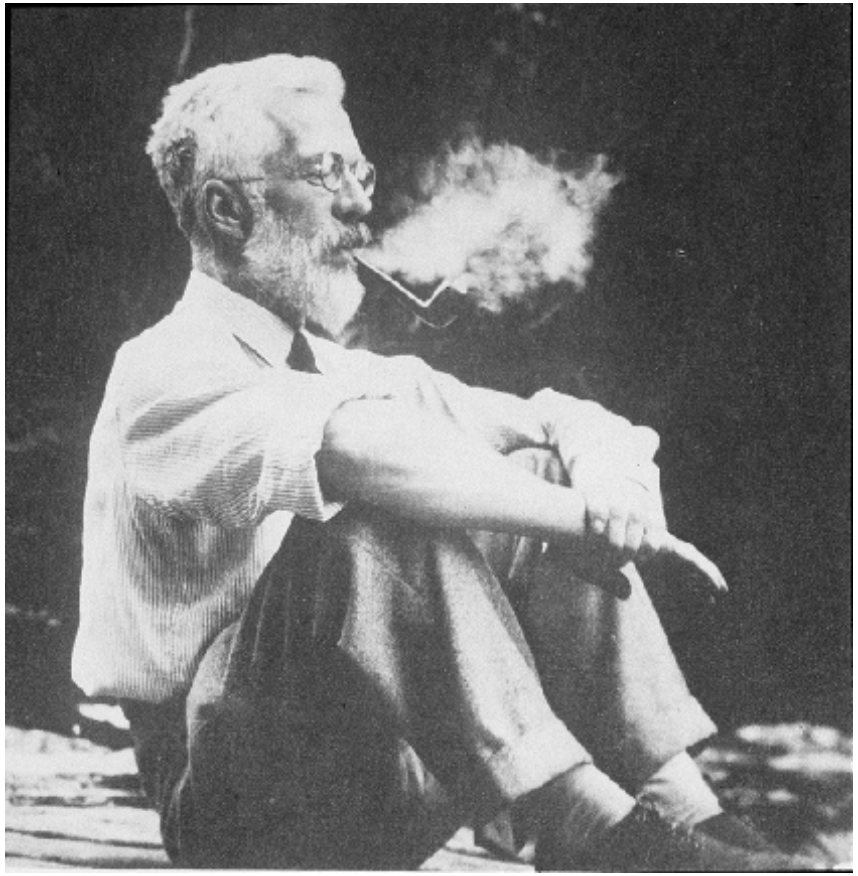
- There is only a 3% chance the placebo was better than the treatment
- There is only a 3% chance of getting the result we did by chance alone
- The probability the null hypothesis is false is 97%
- …

So, what can we conclude?

- We had really bad luck in our sample selection, or
- One or more of our assumptions was wrong, possibly but not necessarily that treatment was no better than placebo (and failure of any assumption affects p-value)

So, our p-value was p=.03

- That's smaller than .05
- So it is…**statistically significant**



R. A. Fisher called such results “significant”

To Fisher, this meant that the result was worth further scrutiny.

sig·nif·i·cant

/sigˈnifikənt/

adjective

1. **sufficiently great or important to be worthy of attention; noteworthy.**
   "a significant increase in sales"
   synonyms: notable, noteworthy, worthy of attention, remarkable, important, of importance, of consequence, signal; More
2. **having a particular meaning; indicative of something.**
   "in times of stress her dreams seemed to her especially significant"


insignificant
unimportant
meaningless


significant increase
significant event
significant other

mole

The amount or sample of a chemical substance that contains as many constitutive particles, e.g., atoms, molecules, ions, electrons, or photons, as there are atoms in 12 grams of carbon-12

"You keep using that word. I don't think that it means what you think it means." – Inigo Montoya

"Just a Theory": 7 Misused Scientific Words, Scientific American, April 2, 2013
https://www.scientificamerican.com/article/just-a-theory-7-misused-science-words/

Word #1

Hypothesis

A proposed explanation **that can be tested**

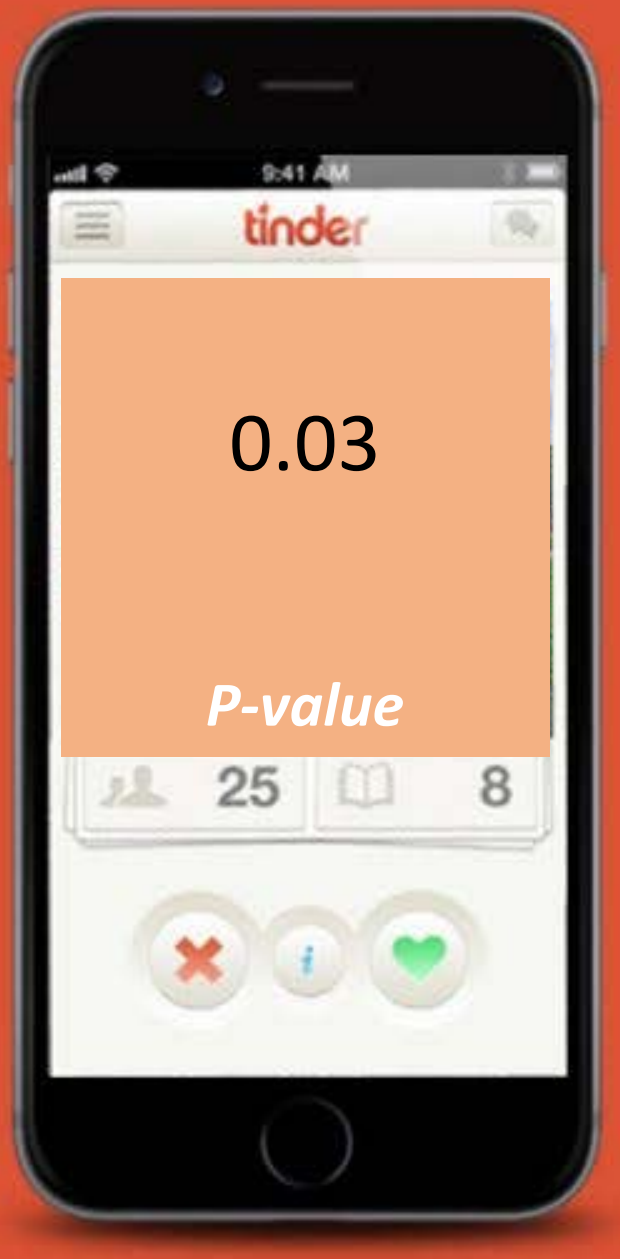
Word #2

Theory

An explanation of some aspect of the natural world that has been **substantiated through repeated experiments or testing**

Word #6

Significant


9:41 AM
tinder
0.03
P-value
25
8

My experimental results are interesting. I should spend more time with them, maybe repeat the experiment. I may be on to something, but it will take time to be sure.

You tiny, beautiful p-value. You are the result that I want to spent the rest of my life with. Let’s publish and get grants together. I love you!

p equal or nearly equal to 0.06

- almost significant
- almost attained significance
- almost significant tendency
- almost became significant
- almost but not quite significant
- almost statistically significant
- almost reached statistical significance
- just barely below the level of significance
- just beyond significance

p equal or nearly equal to 0.08

- a certain trend toward significance
- a definite trend
- a slight tendency toward significance
- a strong trend toward significance
- a trend close to significance
- an expected trend
- approached our criteria of significance
- approaching borderline significance
- approaching, although not reaching, significance

p close to but not less than 0.05

- hovered at nearly a significant level (p=0.058)
- hovers on the brink of significance (p=0.055)
- just about significant (p=0.051)
- just above the margin of significance (p=0.053)
- just at the conventional level of significance (p=0.05001)
- just barely statistically significant (p=0.054)
- just borderline significant (p=0.058)
- just escaped significance (p=0.057)
- just failed significance (p=0.057)

Thanks to Matthew Hankins for these quotes

https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/

A fundamental problem

Generally, we want to be able to conclude something about our hypothesis (H) based on the data (D) we have.

That is, what is the probability that our hypothesis is true based on the data we have observed?

We write that as P(H|D).

A fundamental problem

Unfortunately, a p-value is a probability statement about our data assuming the hypothesis! That is, P(D|H).

No equivalence here

$$P(H|D) \neq P(D|H)$$

The problem, illustrated

What is the probability a person is dead (D) given that the person was hanged (H); that is, in symbol form, what is P(D|H)?

Lacking data, let's make up a number.
P(D|H)=.98 (only 2% hanging survival rate)

The problem, illustrated

Now, let us reverse the question: What is the probability that a person has been hanged (H) given that the person is dead (D); that is, what is P(H|D)?

Let’s say P(H|D)=.0001

(one death in 10,000 by hanging)

Carver, R.P. 1978. The case against statistical testing. *Harvard Educational Review* 48: 378-399.

No one would be likely to make the mistake of substituting the first estimate (.98) for the second (.0001); that is, to accept .98 as the probability that a person has been hanged given that the person is dead.

"Even though this seems to be an unlikely mistake, it is exactly the kind of mistake that is made with the interpretation of statistical significance testing---by analogy, calculated estimates of p(D|H) are interpreted as if they were estimates of p(H|D), when they are clearly not the same."

It is well past time for change

- "It has been widely felt, **probably for thirty years and more**, that significance tests are overemphasized and often misused and that more emphasis should be put on estimation and prediction."
- Cox, D.R. **1986**. Some general aspects of the theory of statistics. *International Statistical Review* 54: 117-126.
- A world of quotes illustrating the long history of concern about this can be viewed at David F. Parkhurst, School of Public and Environmental Affairs, Indiana University
- http://www.indiana.edu/~stigtsts/quotsagn.html

ASA statement articulated six principles

1. **P-values can indicate how incompatible the data are with a specified statistical model.**
2. **P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.**
3. **Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.**

ASA statement articulated six principles

4. **Proper inference requires full reporting and transparency**
5. **A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.**
6. **By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.**

Biggest takeaway from ASA Statement

Bright line thinking is bad for science

"(S)cientists have embraced and even avidly **pursued meaningless differences** solely because they are statistically significant, and have **ignored important effects** because they failed to pass the screen of statistical significance…It is a safe bet that **people have suffered or died** because scientists (and editors, regulators, journalists and others) have used significance tests to interpret results, and have consequently failed to identify the most beneficial courses of action.

(Rothman, supplement to the 2016 ASA statement)


time to move on

The big change

"Moving to a World Beyond p<0.05"
https://amstat.tandfonline.com/doi/full/10.1080/00031305.2019.1583913#.XYjKQ25FxPY

"Scientists rise up against statistical significance"
https://www.nature.com/articles/d41586-019-00857-9

It's time to say farewell to "statistically significant"

Why?

- Significance has lost its meaning
- Bright lines lead to bizarre behavior
- Decades of complaining have done nothing
- “A label of statistical significance adds nothing to what is already conveyed by the value of p; in fact, this dichotomization of p-values makes matters worse.” (TAS editorial)
- File drawer effect

Psychological Bulletin
1979, Vol. 86, No. 3, 638–641

The "File Drawer Problem" and Tolerance for Null Results

Robert Rosenthal
Harvard University

For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the "file drawer problem" is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show non-significant results. Quantitative procedures for computing the tolerance for filed and future null results are reported and illustrated, and the implications are discussed.

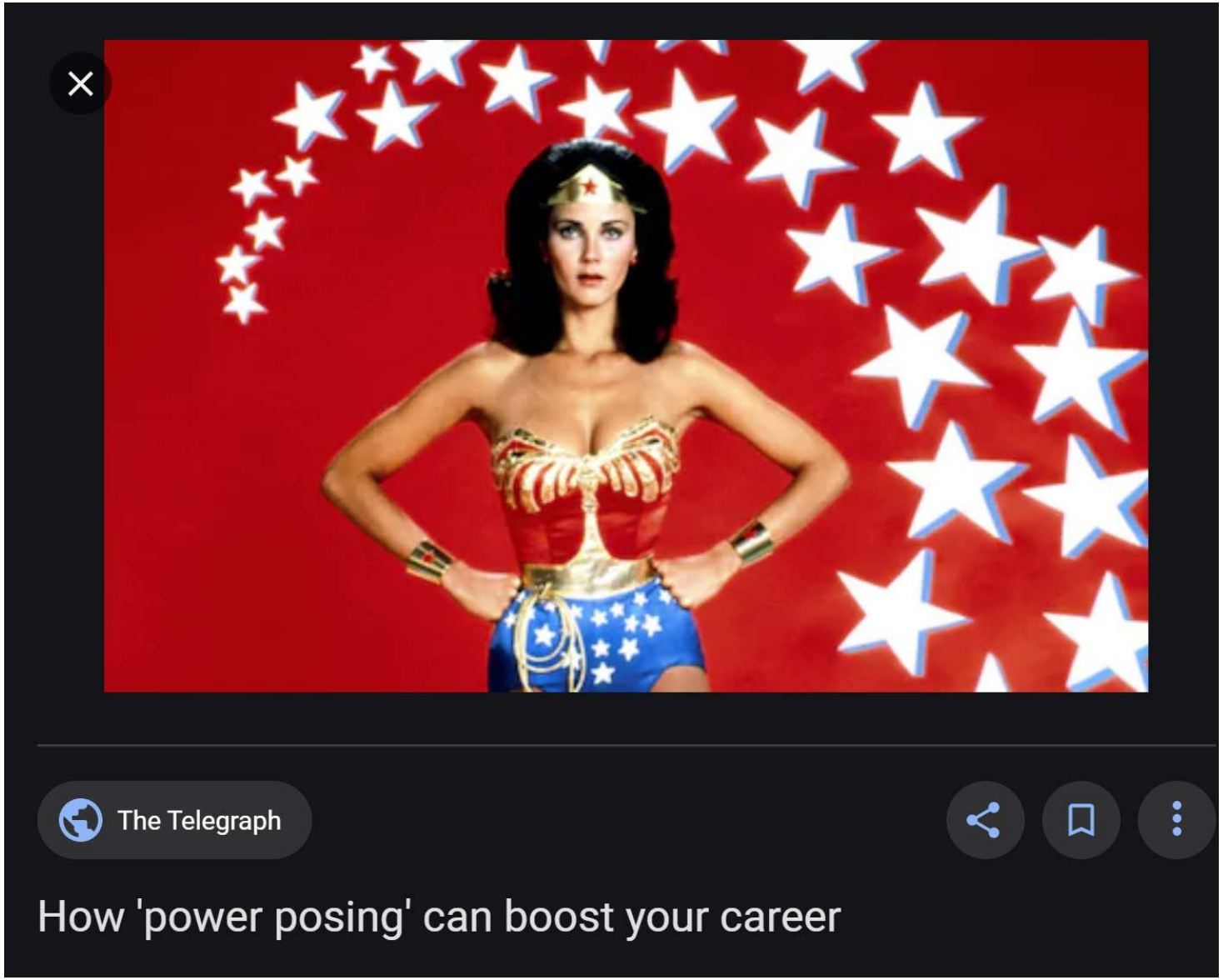
http://datacolada.org/wp-content/uploads/2016/12/3386-Rosenthal-1979-The-file-drawer-problem-and-tolerance-for-null-results.pdf

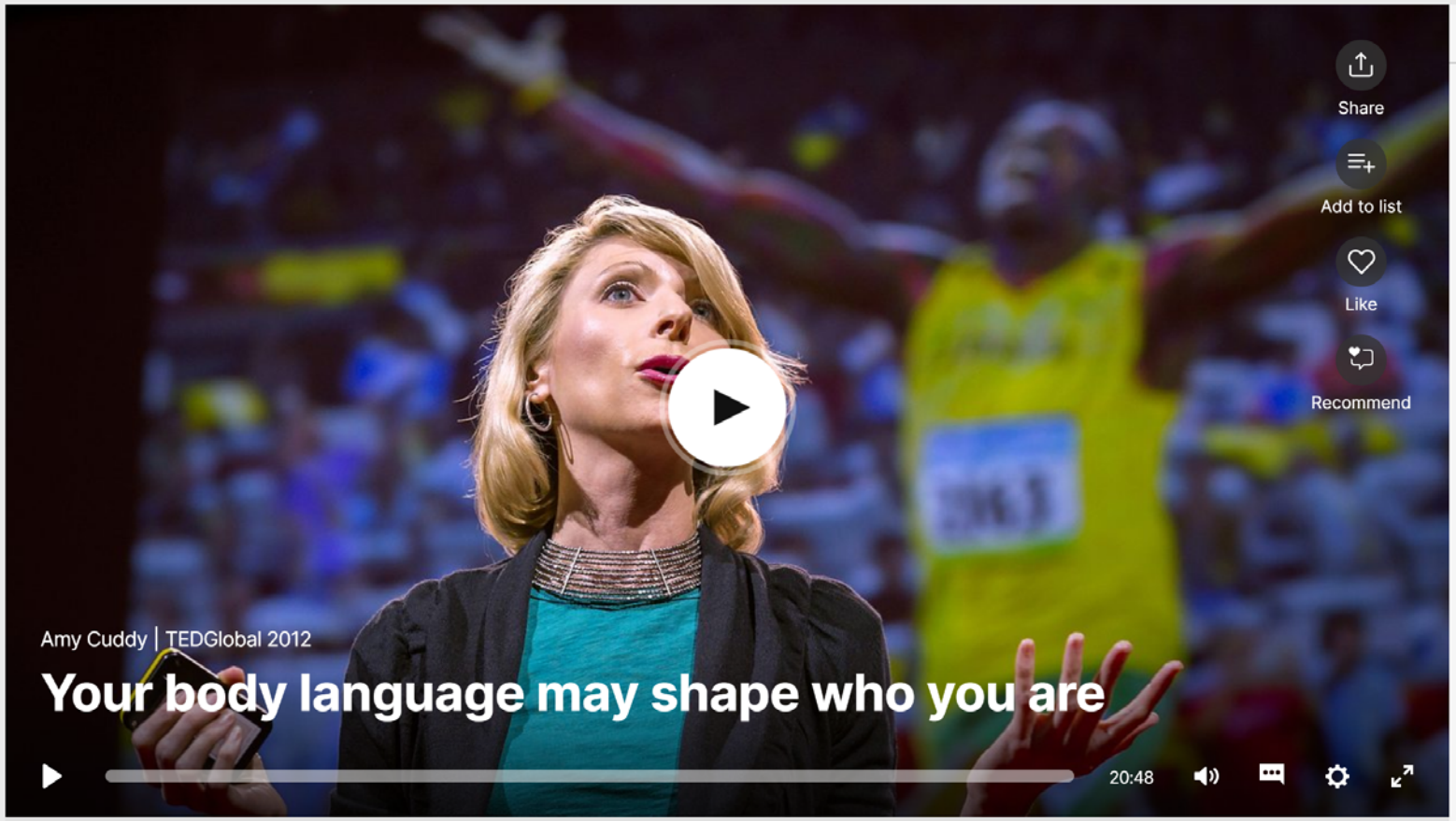
And to be absolutely clear here…

- We are not advocating getting rid of p-values.
- We are not speaking for the American Statistical Association.
  - The 2016 ASA Statement is a statement of the association
  - The 2019 special issue of *The American Statistician* is a publication of the ASA but not an official statement of the ASA.

And why does it matter?

Here are two illustrative examples, one trivial and one very crucial


The Telegraph
How 'power posing' can boost your career



English transcription by Joseph Geni. Reviewed by Morton Bast.

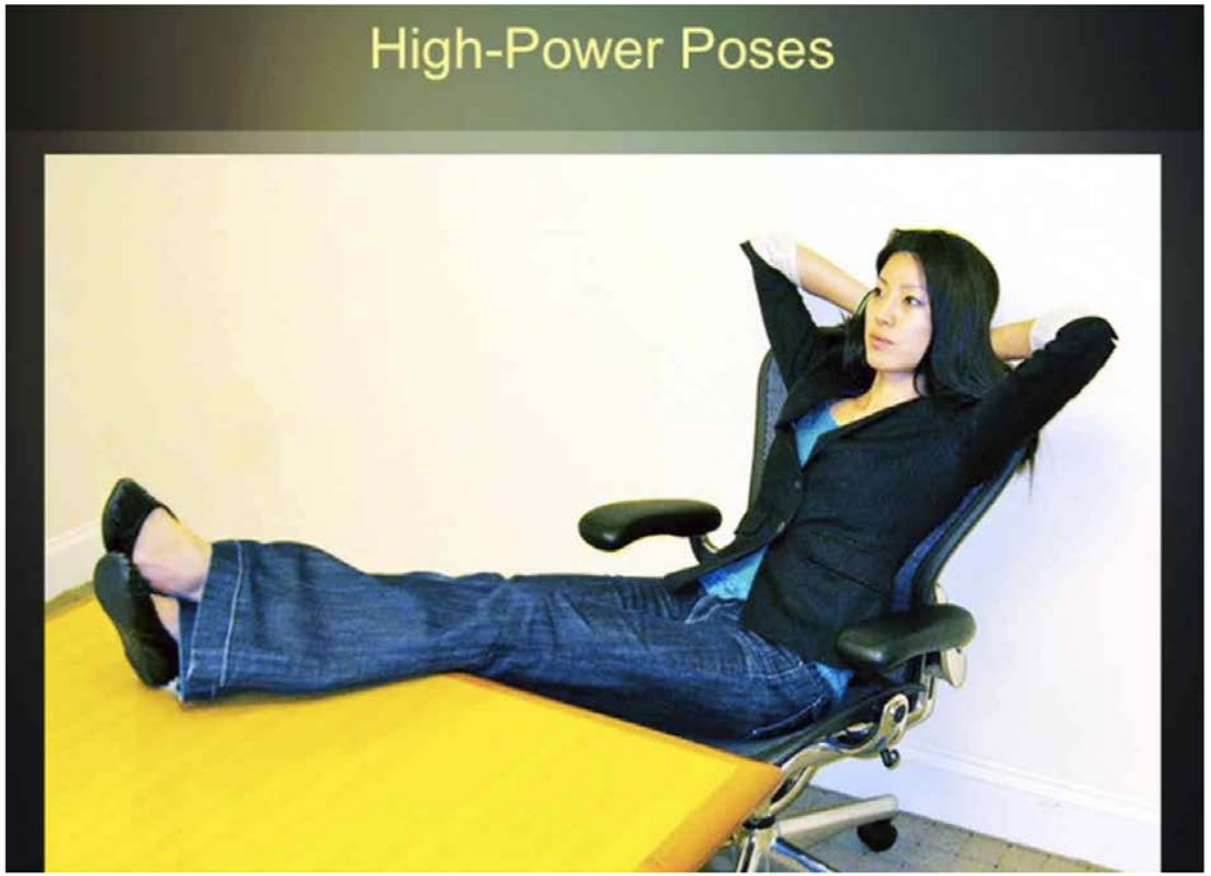
Details About the talk | **Transcript** 51 languages | **Comments (2526)** Join the conversation

(NOTE: Some of the findings presented in this talk have been referenced in an ongoing debate among social scientists about robustness and reproducibility. Read "Criticisms & updates" below for more details as well as Amy

54,502,034 views

“Change your posture for two minutes”

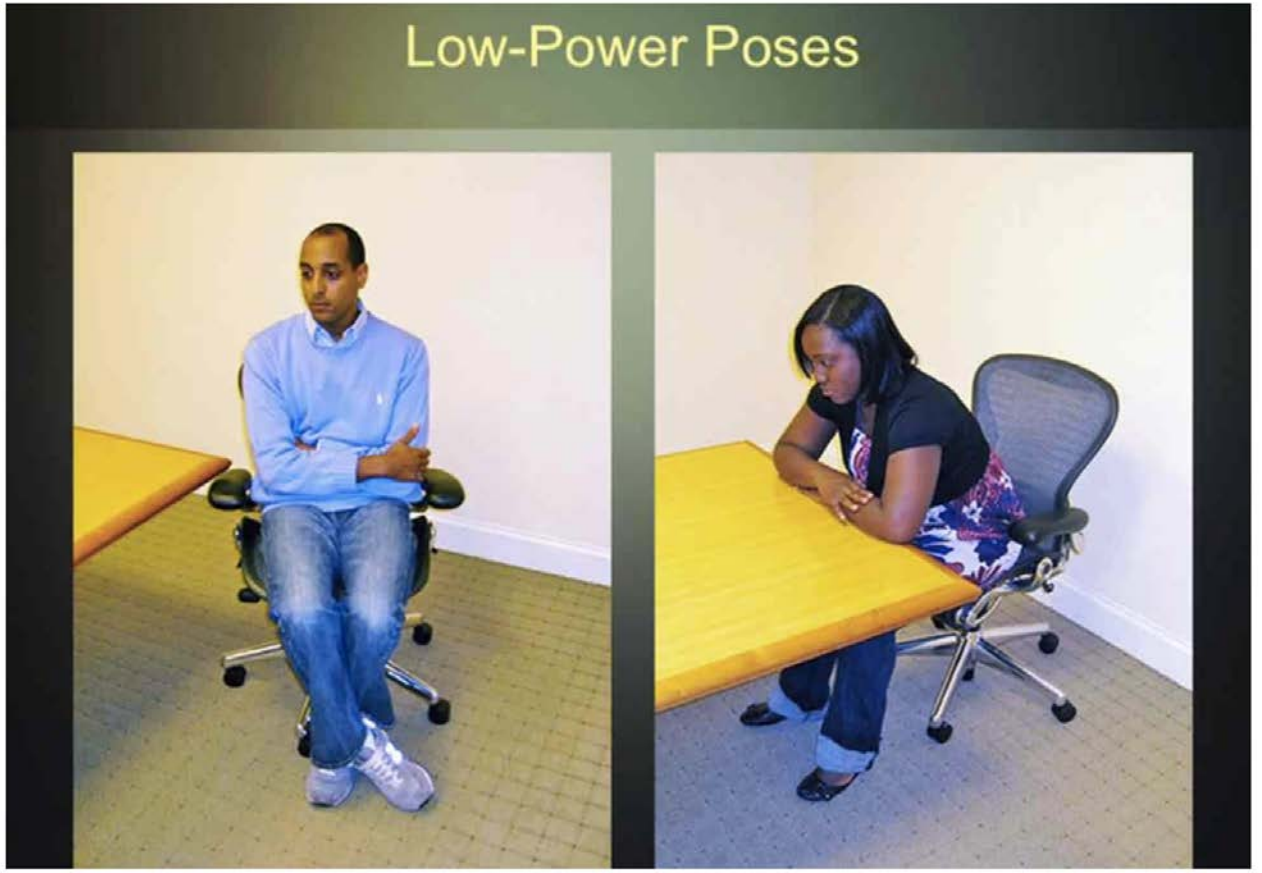
- “When you pretend to feel powerful, are you more likely to actually feel powerful?”
- Found statistically significant changes in testosterone and cortisol levels in people who used high-power poses versus those who used low-power poses
- Low-power posers were more risk averse



Amy Cuddy, TED

Amy Cuddy, TED

Amy Cuddy, TED



Amy Cuddy, TED

Amy Cuddy, TED

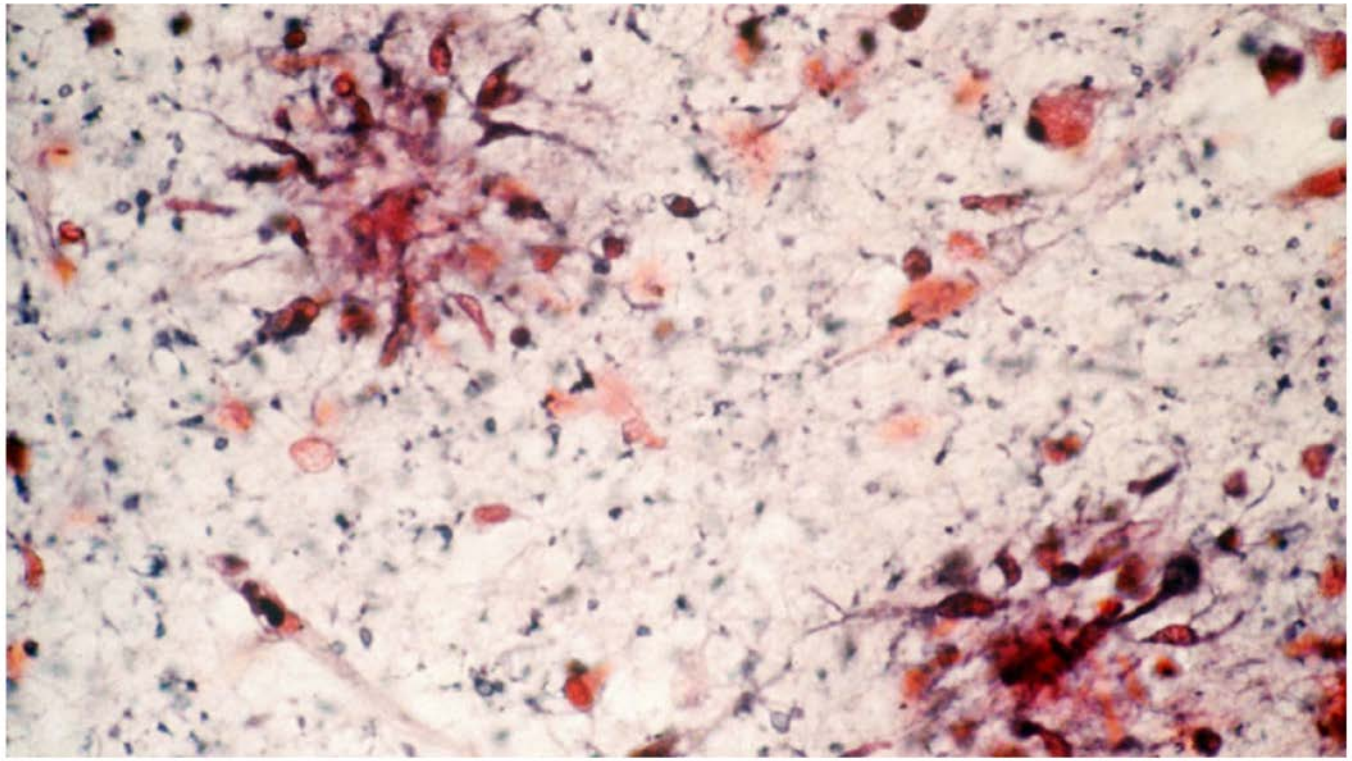
Amy Cuddy, TED

What went wrong?

- The study didn’t reproduce very well
- Too much dependence on statistical significance one of the problems
- “Not bad scientists or bad people – just noisy data.” – Andrew Gelman

Now a much less trivial example

Aducanumab as a treatment for Alzheimer's Disease



By targeting sticky globs of amyloid (red) in the brains of people with Alzheimer's, a new drug may offer a way to slow the disease's spread.

MARTIN M. ROTKER/SCIENCE SOURCE

https://www.sciencenews.org/article/once-scrapped-alzheimers-drug-aducanumab-may-work-after-all

The plot elements

The drug aducanumab, an antibody, has been shown to remove amyloid clusters from the brain.

Such buildup on the brain is associated with Alzheimer's Disease.

The question is whether removal of amyloids would reduce the effects of Alzheimer's

No drug has thus far succeeded in doing this

The plot elements

Biogen stopped two simultaneous clinical trials on the effectiveness of aducanumab last March after futility analysis indicated the study would not likely demonstrate efficacy

However, data continued to come in from the study even though it had been halted

“Between December 2018, when data were cut for the futility analysis, and March 2019, when the trials were discontinued, an additional 179 EMERGE and 139 ENGAGE participants completed 18 months of follow-up”

Howard, R., and Liu, K. Y. (2019), “Questions EMERGE as Biogen claims aducanumab turnaround,” Nature Reviews Neurology, 1–2. https://doi.org/10.1038/s41582-019-0295-9.

The plot elements

A subset analysis was undertaken of those participants who received the full, uninterrupted treatment

In ONE of the two trials, statistical significance was achieved. The higher dose led to 23% less cognitive decline than a placebo after 78 weeks.

The plot thickens

Biogen didn’t actually report the absolute difference for CDR-SB, the primary endpoint (Cognitive Dementia Rating Scale - Sum of Boxes)

They report the p-value and the % change from placebo for the larger datasets collected after the futility analysis. The relative change tells us nothing about the actual change or the clinical significance, so they're basically asking people to rely on the statistical significance for importance.

The plot thickens

Biogen argues that the difference in the results can be explained by a protocol change, but this is based on post hoc subgroup analysis, not the best place to focus on p-values

What can be inferred from the (minimal) information provided is that the effect sizes may not actually meet a threshold of clinical significance.

What to do?

Change is indeed needed, as these examples suggest

But change is never easy…

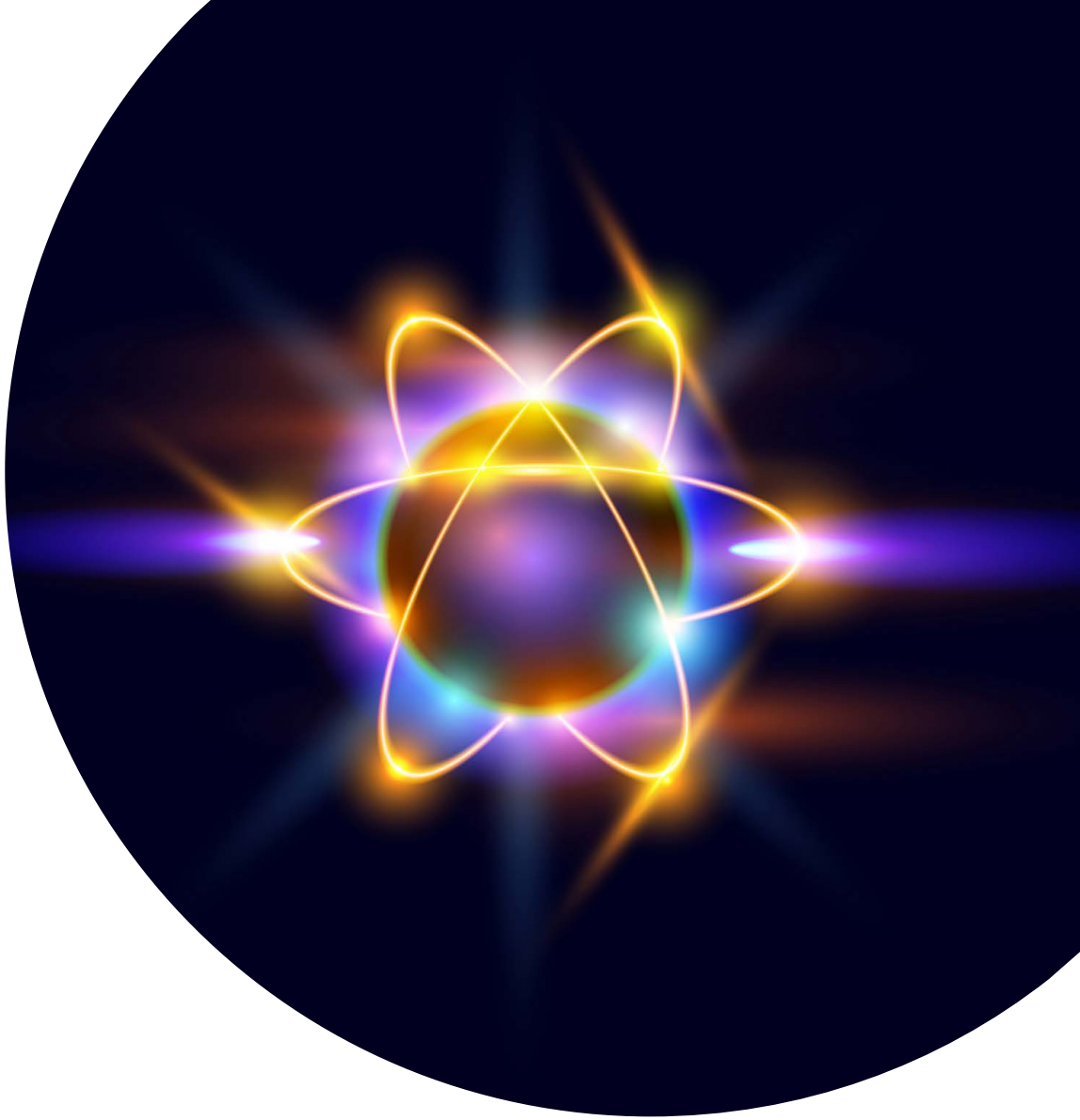
Change won’t be easy

“The basic explanation is neither philosophical nor scientific, but sociologic; everyone uses them. It’s the same reason we can use money. When everyone believes in something’s value, we can use it for real things; money for food, and p-values for knowledge claims, publication, funding, and promotion. It doesn’t matter if the p-value doesn’t mean what people think it means; it becomes valuable because of what it buys.” (Goodman – 2019 (TAS))

What do we do instead?

- If we are telling students (and everyone else) to stop using thresholds to interpret p-values, what should we do?
- Look for some answers in the March 2019 special issue of *The American Statistician* (online and open access)
- We'll talk about a few here, but…
- As you think about moving to a world beyond $p<0.05$, ask yourself: "If this arbitrary threshold had never been created, what would you have to do to get your paper published, your research grant funded, your drug approved, your policy or business recommendation accepted?"

- Accept Uncertainty
- Be Thoughtful
- Be Open
- Be Modest



Accept uncertainty

- Uncertainty exists everywhere in research.
- Statistical methods do not rid data of their uncertainty.
- To accept uncertainty requires that we treat statistical results as being much more incomplete and uncertain than is currently the norm

Thoughtful research:

…considers the scientific context and looks ahead to prospective outcomes

(What magnitudes of differences, odds ratios, or other effect sizes are practically important?)

Thoughtful research:

…considers “related prior evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain…without giving priority to p-values or other purely statistical measures.” (McShane et.al)

Thoughtful researchers:

…use a toolbox of statistical techniques

…consider multiple approaches for solving problems

Potential alternatives

No alternatives have been widely agreed upon. There are interesting ideas out there, as mentioned in the special issue of *The American Statistician*

Potential alternatives

- Shannon Information
- Bayes Factor Bound
- False positive risk
- Analysis of credibility
- Second generation p-values

(in the interest of time, we'll just briefly show the two highlighted ones)

S-Value

- Shannon information or S-value (surprisal) of the test
- $s = -\log_2(p)$
- Evidence against the model supplied by the test, expressed in binary digits of information
- One advantage: no upper bound. Difficult to misinterpret as a hypothesis probability.
- $-\log_2(0.05) = 4.3$
- Interpretation: Values in a 95% confidence interval have at most 4 bits of information against them = evidence against “fairness” of coin tosses provided by obtaining 4 heads in a row.

False positive risk (FPR)

- "Reverse Bayes" approach
- H_0: effect = 0
- Report actual p-value
- Report FPR = $P(H_0|\text{data})$
- Assume prior $P(H_1) = 0.5$ (optimistic/conservative$\Rightarrow$ *minimum* FPR)
- For p = 0.05
  - FPR = 20-30% for $P(H_1) = 0.5$
  - FPR > 70% for $P(H_1) = 0.1$

Be open

- Understand that subjectivity is involved in any statistical analysis.
- “(T)here is essentially no aspect of scientific investigation in which judgment is not required.” (Brownstein et.al)

Be open

Remember that one study is rarely enough. The words “a groundbreaking new study” might be loved by news writers but must be resisted by researchers. Breaking ground is only the first step in building a house. It will be suitable for habitation only after much more hard work.

Be open

Adopt open science practices

- Public pre-registration
- Transparency and completeness in reporting
- Sharing data and code

Be modest

- P-values, confidence intervals, and other statistical measures are all uncertain.
- Encourage others to reproduce your work
- Statistical inference is (or should be) just one part of scientific inference

Be modest about our teaching, too

- These are hard concepts
- We aren’t all geniuses at explaining them
- Students and others won’t grasp them immediately
- Ask them what they understand, go from there, then “rinse and repeat”

Your presenters need to be modest, too

- We don’t have all/most/any answers about best practices for teaching these things
- There is a community working on it, and you can reach them through the ASA’s Statistical Education Section or the CAUSE website https://www.causeweb.org/cause/

Five changes that could be taught relatively easily

- Lead with (focus on) effect sizes and related measures of uncertainty (for instance, interval estimates)
- Focus on the substantive implications of those estimates. (For example, don't focus on whether the interval contains zero, but on whether the interval bounds have qualitatively different practical consequences.)

Five changes that could be taught relatively easily

- Interpret confidence intervals as compatibility intervals (that is, describing how compatible the data are with your hypothesized model)

Example of compatibility interval interpretation

- Study: Covid-19 patients received lopinavir–ritonavir in addition to standard care or standard care alone (randomized trial) (NEJM, March 18, 2020, DOI: 10.1056/NEJMoa2001282)
- Result: Mortality difference at 28 days of –5.8 percentage points, 95% CI (–17.3, 5.7)
- Conclusion: “Mortality at 28 days was similar in the lopinavir–ritonavir group and the standard-care group (19.2% vs. 25.0%). … In hospitalized adult patients with severe Covid-19, no benefit was observed with lopinavir–ritonavir treatment beyond standard care.”

A Trial of Lopinavir–Ritonavir in Adults Hospitalized with Severe Covid-19

Bin Cao, M.D., Yeming Wang, M.D., Danning Wen, M.D., Wen Liu, M.S., Jingli Wang, M.D., Guohui Fan, M.S., Lianguo Ruan, M.D., Bin Song, M.D., Yanping Cai, M.D., Ming Wei, M.D., Xingwang Li, M.D., Jiaan Xia, M.D., et al.

Example of compatibility interval interpretation

A better statement of this result:

"Our estimate of the mortality difference at 28 days was –5.8 percentage points (= 19.2% – 25.0%); thus, adding lopinavir-ritonavir to standard care could result in a clinically large decrease in mortality. However, possible mortality differences that are highly compatible with our data, given our model, ranged from –17.3 (a very large decrease in mortality) to 5.7 (a large increase in mortality). Our trial was small, including only 199 patients, all with severe Covid-19. Further study of this potentially effective treatment is needed."

This result should be discussed in the context of the plausibility of the causal mechanism for a beneficial effect (based on prior evidence), the high consistency of results across different study outcomes, study limitations (including but not limited to the large imprecision of the estimates), potential adverse effects of lopinavir-ritonavir, and other relevant considerations.

Five changes that could be taught relatively easily

- When presenting p-values, present them as continuous values (not categorized into significant or not), and along with the standard p-value (null hypothesis), report p-values for other pre-specified hypotheses.

(One example: instead of assuming no effect, assume the minimum meaningful effect size.)

Five changes that could be taught relatively easily

- Interpret p-values as (uncertain) descriptive measures of compatibility with the model, and recognize that the value of p is impacted not just by the assumption of the null hypothesis, but by the many other assumptions/choices data analysts make

(The Tinder example helps here in indicating not to rush to fall in love with a low p-value)

And one more change, a little harder to teach, but maybe most important

Encourage students not to focus on the statistical measure alone (for example, the p-value) but also to consider

- related prior evidence
- plausibility of mechanism
- study design and data quality
- real world costs and benefits
- novelty of finding
- other factors that vary by research domain

(per McShane et al)

Wrapping up

- It’s time to stop using “statistical significance” as any kind of metric for scientific inference, and teaching it as a foundational concept
- We and many others have written a lot about what “a world beyond P<0.05” should look like
- P-values still have their uses
- You are the front line in making these changes

Thanks. Now we hope to hear from you!

nlazar@stat.uga.edu
allenschirm@gmail.com
ron@amstat.org