Lecture 16

Hypothesis testing for proportions
P-values, what's the problem?

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

Review: Key Points

Chapter 19: Confidence Interval for Proportions

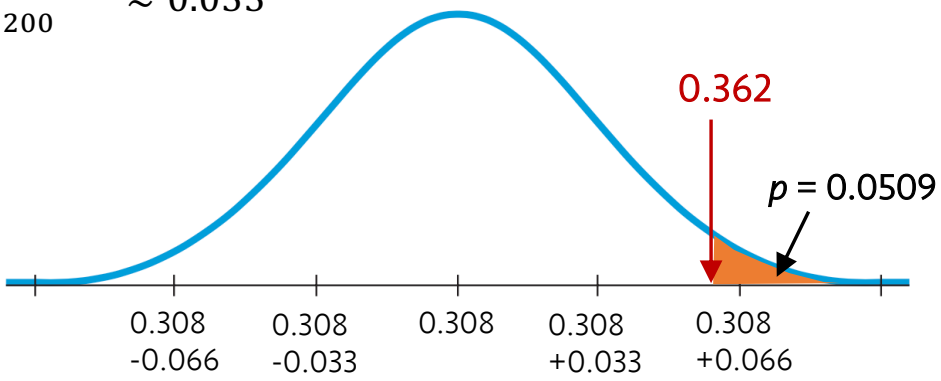
- **Standard error:** standard deviation of a sampling distribution
- For a sample proportion, $\hat{p}$, the standard error is $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$
- 95% confidence interval for a proportion, $\hat{p} \pm 1.96\ SE(\hat{p})$
- General form: *Estimate* $\pm$ *Margin of Error (ME)*
- Critical value, z^* = the *number* of SEs (e.g., 2 in 2SEs)
- Every confidence interval is a balance between certainty and precision.
- You can choose your sample size based on confidence interval.

Hypotheses

- **Hypothesis**: questions like, has the Facebook users who update their status daily increased since last month?

- **Null hypothesis:** null because it assumes no changes, thus p = 30.8%
- **Alternative hypothesis**: H_A: $p > 30.8\%$

- We observed a new $\hat{p}$ from 200 respondents.
- Based on the null hypothesis, $SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.308 \times 0.692}{200}} \approx 0.033$
- Let's say the observed $\hat{p}$ = 36.2%
- Then, $z = \frac{0.362 - 0.308}{0.033} = 1.6364$
- p = 0.0509 (one-tail)



A Trial as a "Null Hypothesis Statistical Test" (NHST)

- It's the logic of jury trials.
  - The null hypothesis is that the defendant is *innocent*
  - Judge the evidence
  - Juries ask "Could these evidence plausibly have happened by chance if the defendant were in fact *innocent*?"
  - Make a decision

- In hypothesis testing:
  - We quantify *"how surprising the evidence would be if the null hypothesis were true."*

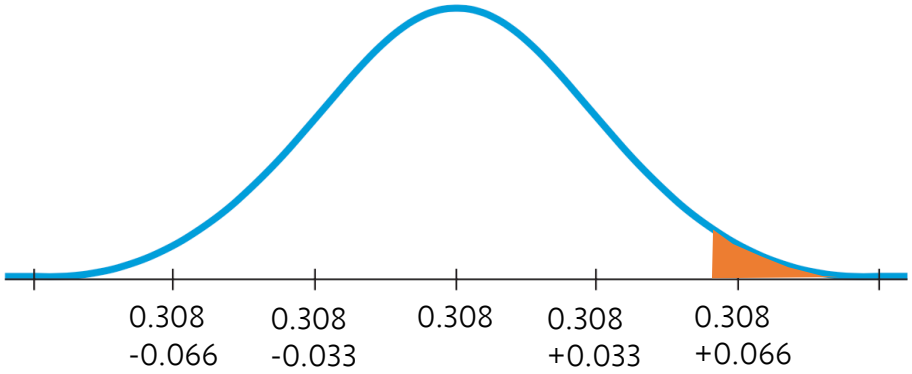
P-values

- "Are the data surprising given the null hypothesis?"
- The probability of seeing the observed data given that the null hypothesis is true: P-value.
  - P-value is small enough: "we are very surprised"
  - P-value is high: no surprise, no reason to reject the null hypothesis.
  - BUT we haven't proven that the null hypothesis is true. What we can say is that it doesn't appear to be false.
  - Formally, we can say: "Fail to reject" the null hypothesis.

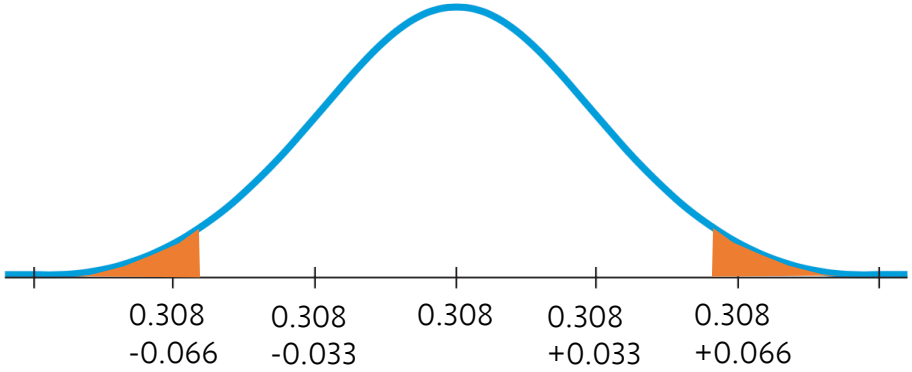
- Same in a trial
  - H_0 = innocent defendant
  - If we fail to reject H_0, the most we can say is the defendant is "*not guilty*", rather than the defendant is "*innocent*".
  - What would happen we do the opposite? I.e., H_0: "the defendant is guilty", then need to prove the defendant is innocent?

Two-sided, one-sided alternative

- H_0: p = 30.8%
- H_A (or H_1): $p > 30.8\%$
- One-sided alternative (one-tailed)



- H_0: p = 30.8%
- H_A (or H_1): $p \neq 30.8\%$
- Two-sided alternative (two-tailed)



P-values and decisions

- Hypothesis test is useful when we must make a decision. Guilty or not guilty?
- But decision can be often arbitrary. E.g., How small the P-value has to be?
- It is highly context-dependent.

- When you report your decision about rejecting (or failing to reject) the hypothesis, always provide the p-value, and also confidence interval whenever possible.

Recent discussions about p-values

https://www.nature.com/news/scientific-method-statistical-errors-1.14700

STATISTICAL ERRORS

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

BY REGINA NUZZO

For a brief moment in 2010, Matt Motyl was on the brink of scientific glory: he had discovered that extremists quite literally see the world in black and white.

The results were "plain as day", recalls Motyl, a psychology PhD student at the University of Virginia in Charlottesville. Data from a study of nearly 2,000 people seemed to show that

It turned out that the problem was not in the data or in Motyl's analyses. It lay in the surprisingly slippery nature of the *P* value, which is neither as reliable nor as objective as most scientists assume. "*P* values are not doing their job, because they can't," says Stephen Ziliak, an economist at Roosevelt University in Chicago, Illinois, and a frequent critic of the way statis-

Goodman, a physician and statistician at Stanford. "Then 'laws' handed down from God are no longer handed down from God. They're actually handed down to us by ourselves, through the methodology we adopt."

OUT OF CONTEXT

P values have always had critics. In their almost

DALE EDWIN MURRAY

Vox

What a nerdy debate about p-values shows about science — and how to fix it

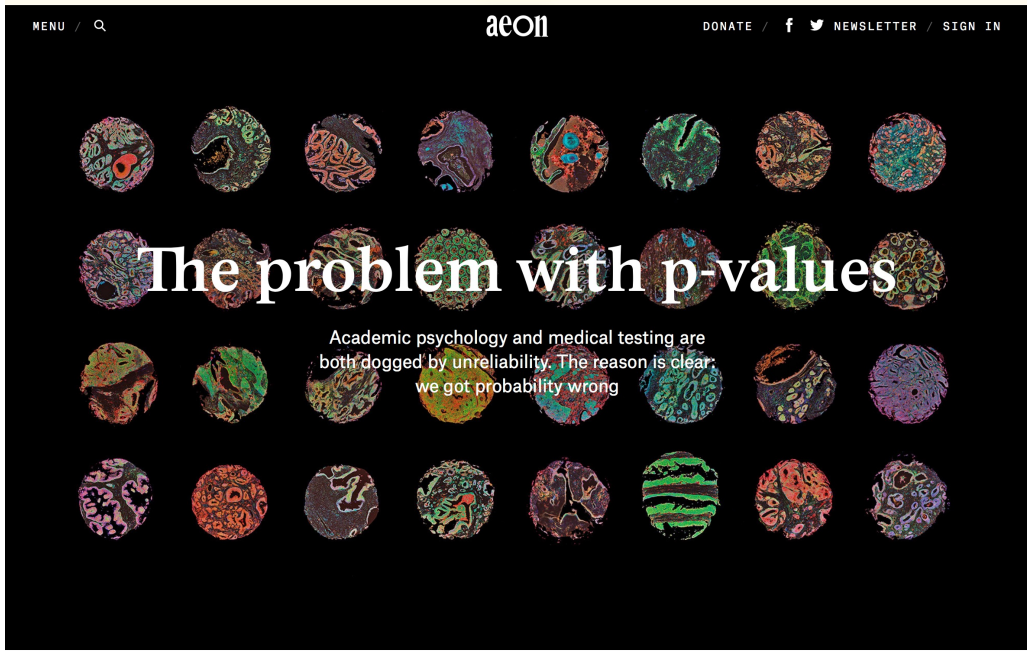
The case for, and against, redefining "statistical significance."

Updated by Brian Resnick on July 31, 2017 12:00 pm

TWEET SHARE

https://www.vox.com/platform/amp/science-and-health/2017/7/31/16021654/p-values-statistical-significance-redefine-0005

Recent discussions about p-values



False positives; numerous microscopic cancerous and non-cancerous human tissue samples. *Photo courtesy Wellcome Images*

David Colquhoun is a professor of pharmacology at University College

The aim of science is to establish facts, as accurately as possible. It is therefore crucially important to determine whether an observed

https://aeon.co/essays/it-s-time-for-science-to-abandon-the-term-statistically-significant

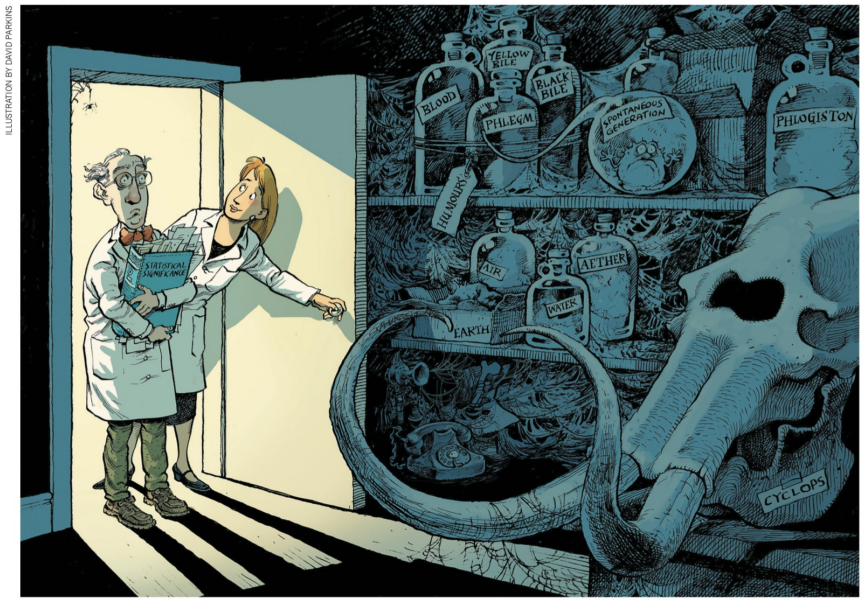
COMMENT

EVOLUTION Cooperation and conflict from ants and chimps to us p.308

HISTORY To fight denial, study Galileo and Arendt p.309

CHEMISTRY Three more unsung women — of astatine discovery p.311

PUBLISHING As well as ORCID ID and English, list authors in their own script p.311



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

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

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.

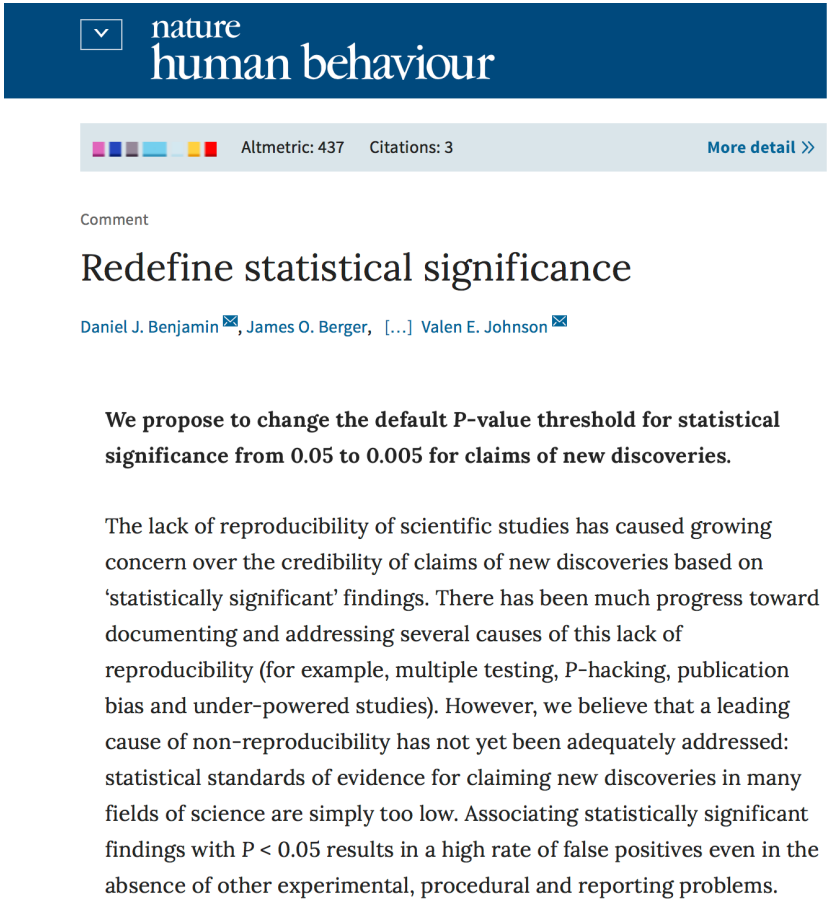
We have some proposals to keep scientists from falling prey to these misconceptions.

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 ▶

https://www.nature.com/articles/d41586-019-00857-9

Recent discussions about p-values



nature human behaviour

Altmetric: 437 Citations: 3 More detail »

Comment

Redefine statistical significance

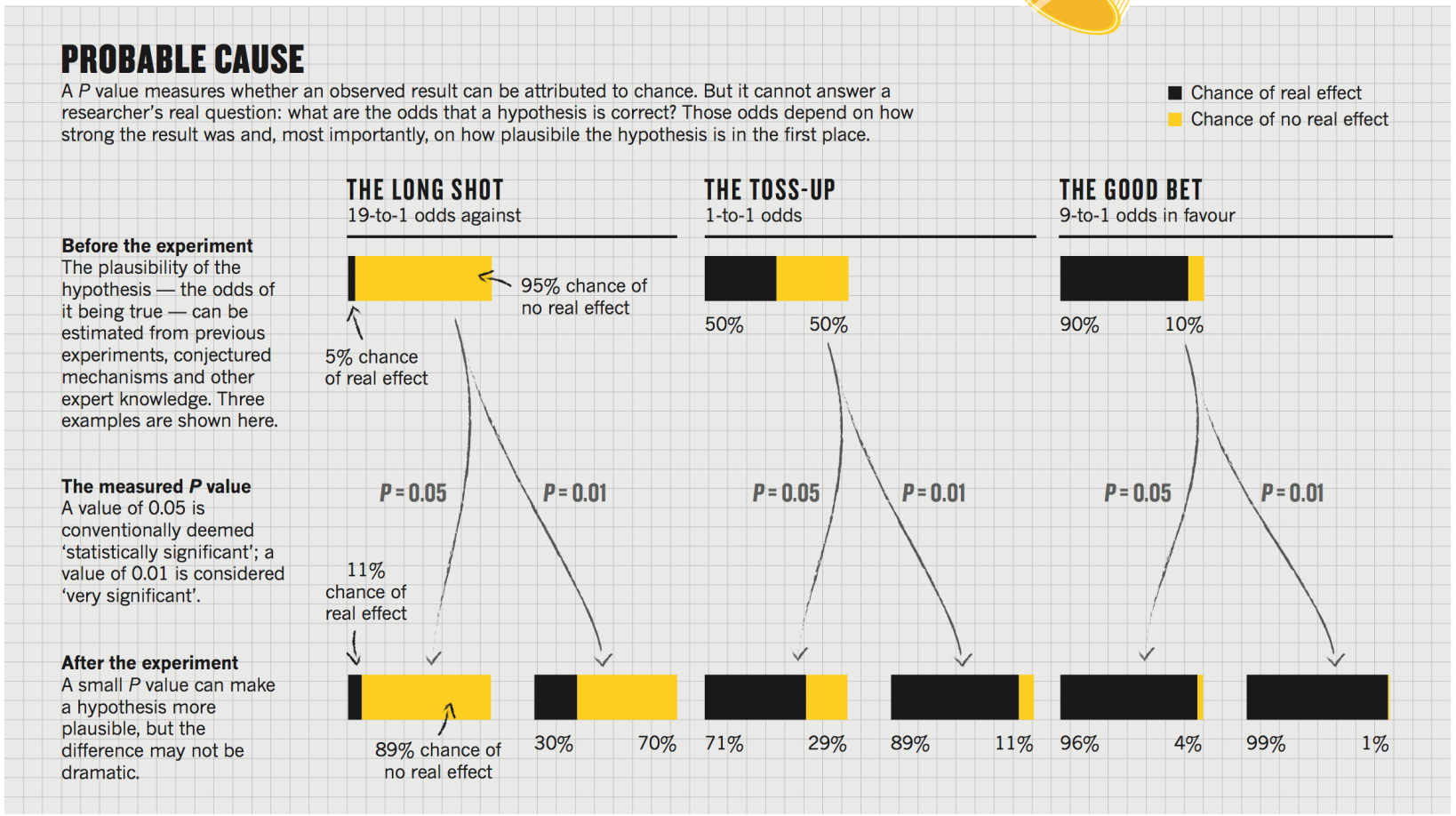
Daniel J. Benjamin, James O. Berger, […] Valen E. Johnson

We propose to change the default P-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

The lack of reproducibility of scientific studies has caused growing concern over the credibility of claims of new discoveries based on 'statistically significant' findings. There has been much progress toward documenting and addressing several causes of this lack of reproducibility (for example, multiple testing, P-hacking, publication bias and under-powered studies). However, we believe that a leading cause of non-reproducibility has not yet been adequately addressed: statistical standards of evidence for claiming new discoveries in many fields of science are simply too low. Associating statistically significant findings with $P < 0.05$ results in a high rate of false positives even in the absence of other experimental, procedural and reporting problems.

https://www.nature.com/articles/s41562-017-0189-z

"Small P value can make a hypothesis more plausible, but the difference may not be dramatic."



"How much of an effect is there?"

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio &

News & Comment > News > 2017 > November > Article

NATURE | NEWS

Online daters do better in the marriage stakes

Those who first find each other through the Internet are more likely to stay hitched.

Regina Nuzzo

03 June 2013

Rights & Permissions

Couples in the United States who meet online seem to enjoy at least as much marital bliss as those who meet in more traditional venues, according to the results of an online survey of more than 19,000 people funded by online dating service eHarmony.

Jake Wyman/Corbis

Couples who met in online venues — ranging from dating services to chat rooms — had slightly better outcomes in their marital life than those who met in other ways.

The survey's participants consisted of people who married between 2005 and 2012. About 35% reported that they had met their spouse online, more than through introductions by friends, work and school combined. The study revealed that people who used this method to meet their spouses were slightly older, wealthier, more educated and more likely to be employed than those who went with tradition[1].

- Study of more than 19,000 people
- Those who meet their spouses online are less likely to divorce ($p < 0.002$) and more likely to have high marital satisfaction ($p < 0.001$) than those who meet offline.
- The divorce rate from 7.67% down to 5.96%; Happiness score from 5.48 to 5.64 on a 7-point scale
- "Seductive certainty of significance"

"P-hacking"

"THE *P* VALUE WAS NEVER MEANT TO BE USED THE WAY IT'S USED TODAY."

FOOLING OURSELVES

HUMANS ARE REMARKABLY GOOD AT SELF-DECEPTION. BUT GROWING CONCERN ABOUT REPRODUCIBILITY IS DRIVING MANY RESEARCHERS TO SEEK WAYS TO FIGHT THEIR OWN WORST INSTINCTS.

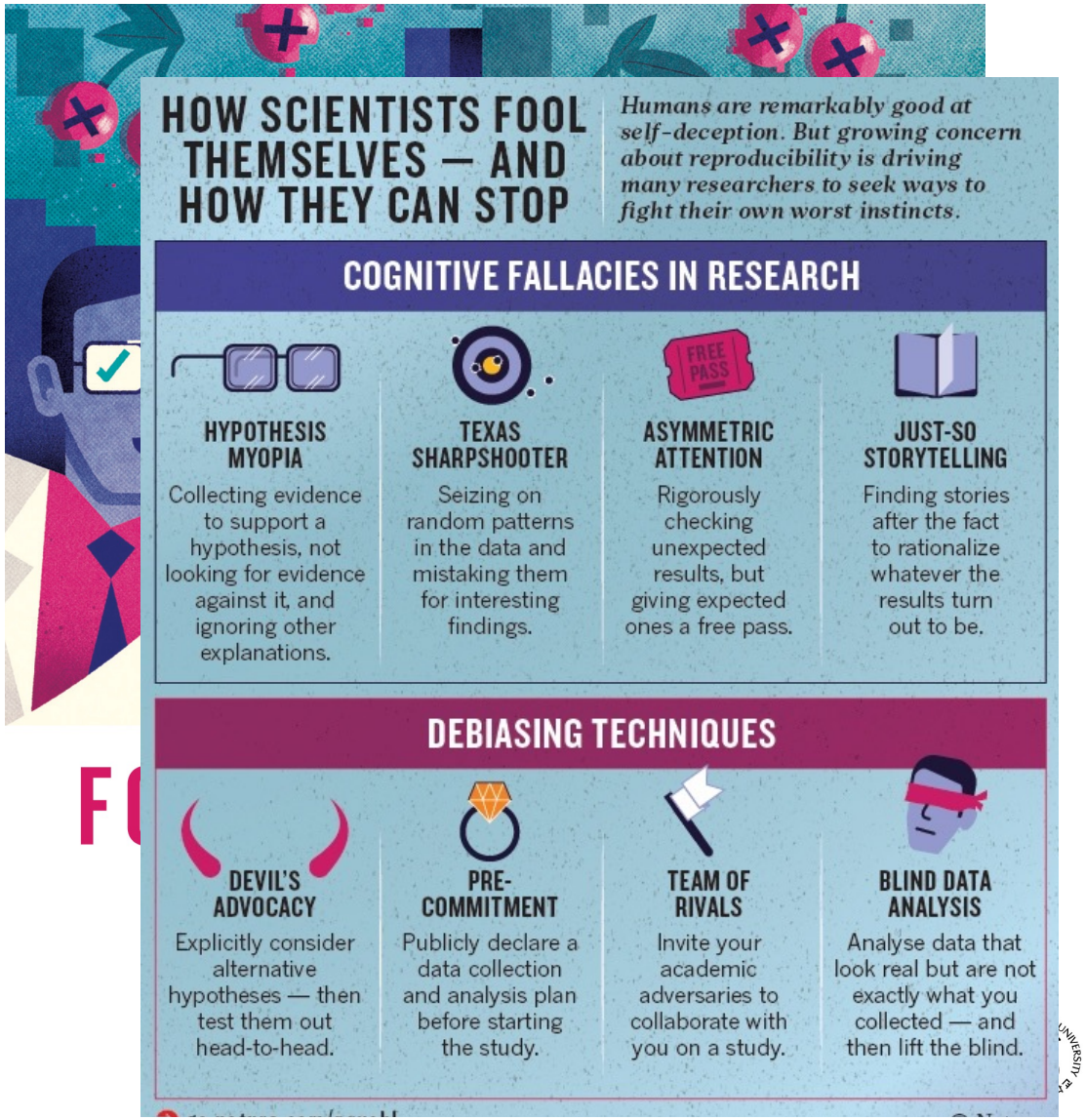
BY REGINA NUZZO

https://www.nature.com/polopoly_fs/1.18517!/menu/main/topColumns/topLeftColumn/pdf/526182a.pdf

"P-hacking"

"THE *P* VALUE WAS NEVER MEANT TO BE USED THE WAY IT'S USED TODAY."

https://www.nature.com/polopoly_fs/1.18517!/menu/main/topColumns/topLeftColumn/pdf/526182a.pdf



COMMENT

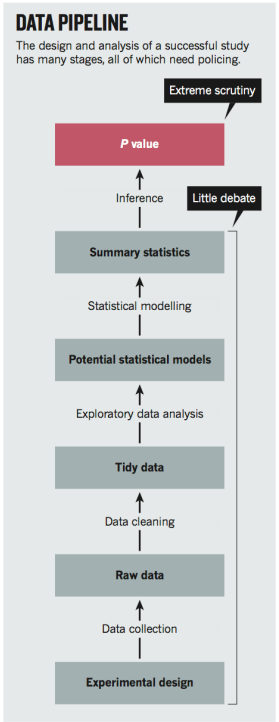
P values are just the tip of the iceberg

Ridding science of shoddy statistics will require scrutiny of every step, not merely the last one, say **Jeffrey T. Leek** and **Roger D. Peng**.

There is no statistic more maligned than the *P* value. Hundreds of papers and blogposts have been written about what some statisticians deride as 'null hypothesis significance testing' (NHST; see, for example, go.nature.com/pfvgqe). NHST deems whether the results of a data analysis are important on the basis of whether a summary statistic (such as a *P* value) has crossed a threshold. Given the discourse, it is no surprise that some hailed as a victory the banning of NHST methods (and all of statistical inference) in the journal *Basic and Applied Social Psychology* in February[1].

Such a ban will in fact have scant effect on the quality of published science. There are many stages to the design and analysis of a successful study (see 'Data pipeline'). The last of these steps is the calculation of an inferential statistic such as a *P* value, and the application of a 'decision rule' to it (for example, $P < 0.05$). In practice, decisions that are made earlier in data analysis have a much greater impact on results — from experimental design to batch effects, lack of adjustment for confounding factors, or simple measurement error. Arbitrary levels of statistical significance can be achieved by changing the ways in which data are cleaned, summarized or modelled[2].

P values are an easy target: being widely used, they are widely abused. But, in practice, deregulating statistical significance opens the door to even more ways to game statistics — intentionally or unintentionally — to get a result. Replacing *P* values with Bayes factors or another statistic is ultimately about choosing a different trade-off of true positives and false positives. Arguing about the *P* value is like focusing on a single misspelling, rather than on the faulty logic of a sentence.



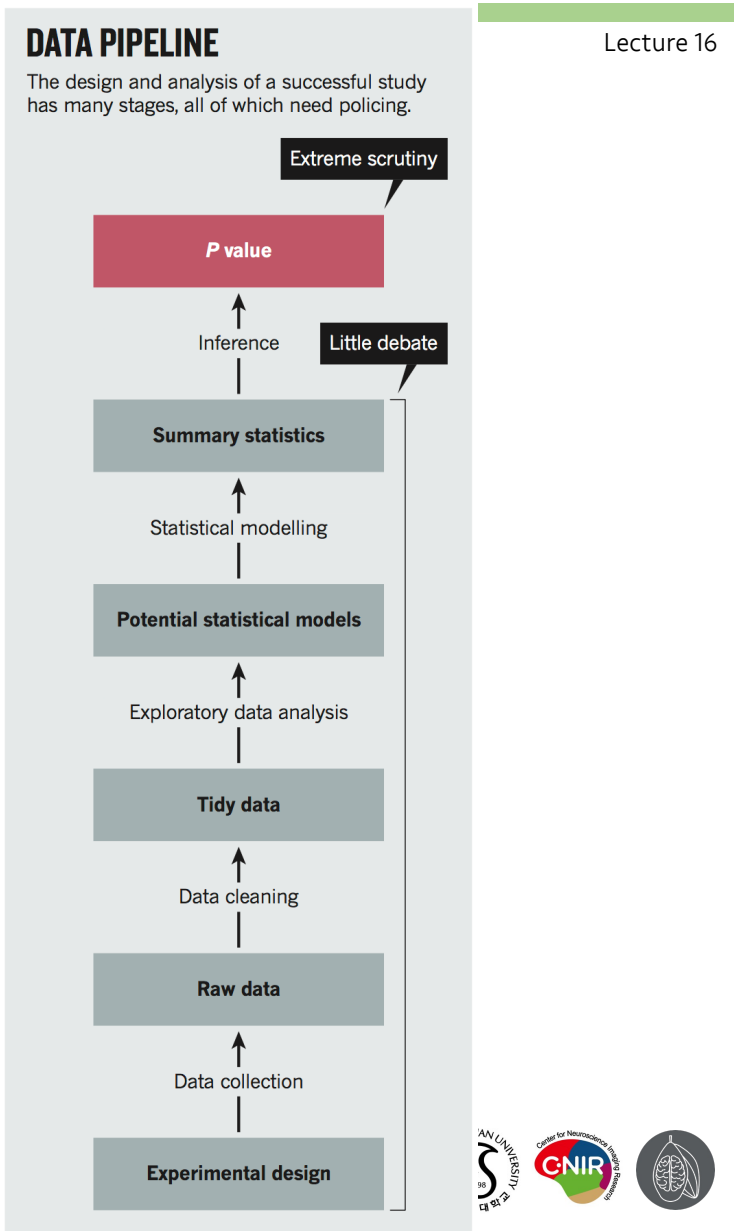
designed to address this crisis. For analysis is taught through an apprenticeship model, and different disciplines develop their own analysis subcultures. Decisions are based on cultural conventions in specific communities rather than on empirical evidence. For example, economists call data measured over time 'panel data', to which they frequently apply mixed-effects models. Biomedical scientists refer to the same type of data structure as 'longitudinal data', and often go at it with generalized estimating equations.

Statistical research largely focuses on mathematical statistics, to the exclusion of the behaviour and processes involved in data analysis. To solve this deeper problem, we must study how people perform data analysis in the real world. What sets them up for success, and what for failure? Controlled experiments have been done in visualization[3] and risk interpretation[4] to evaluate how humans perceive and interact with data and statistics. More recently, we and others have been studying the entire analysis pipeline. We found, for example, that recently trained data analysts do not know how to infer *P* values from plots of data[5], but they can learn to do so with practice.

The ultimate goal is evidence-based data analysis[6]. This is analogous to evidence-based medicine, in which physicians are encouraged to use only treatments for which efficacy has been proved in controlled trials. Statisticians and the people they teach and collaborate with need to stop arguing about *P* values, and prevent the rest of the iceberg from sinking science. ■

Jeffrey T. Leek *and* **Roger D. Peng** *are associate professors of biostatistics at the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, USA.*

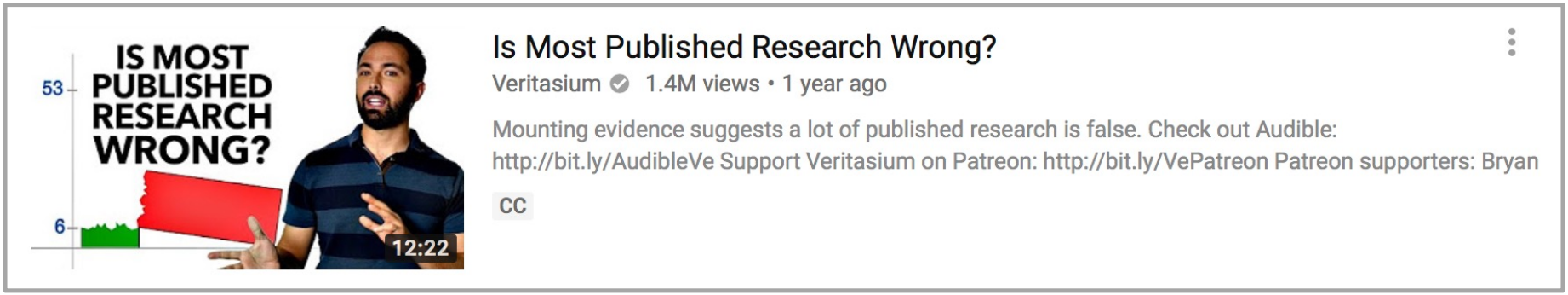
http://www.nature.com/news/statistics-p-values-are-just-the-tip-of-the-iceberg-1.17412



Special section in *Nature* on "irreproducible research"

http://www.nature.com/news/reproducibility-1.17552

Good Youtube video on this topic



https://www.youtube.com/watch?v=42QuXLucH3Q

Key Points

Chapter 20: Testing Hypotheses About Proportions

- **Hypothesis testing:** the logic of jury trial
- **P-value:** the probability of seeing the observed data given that the null hypothesis is true.
- P-value is for decision-making, but it should not be blinded. It's context-dependent.
- Many issues related to P-values: *"Small P value can make a hypothesis more plausible, but the difference may not be dramatic."*
- Problems in misuses of P-value are just the tip of the iceberg.
- Let's not fool ourselves!