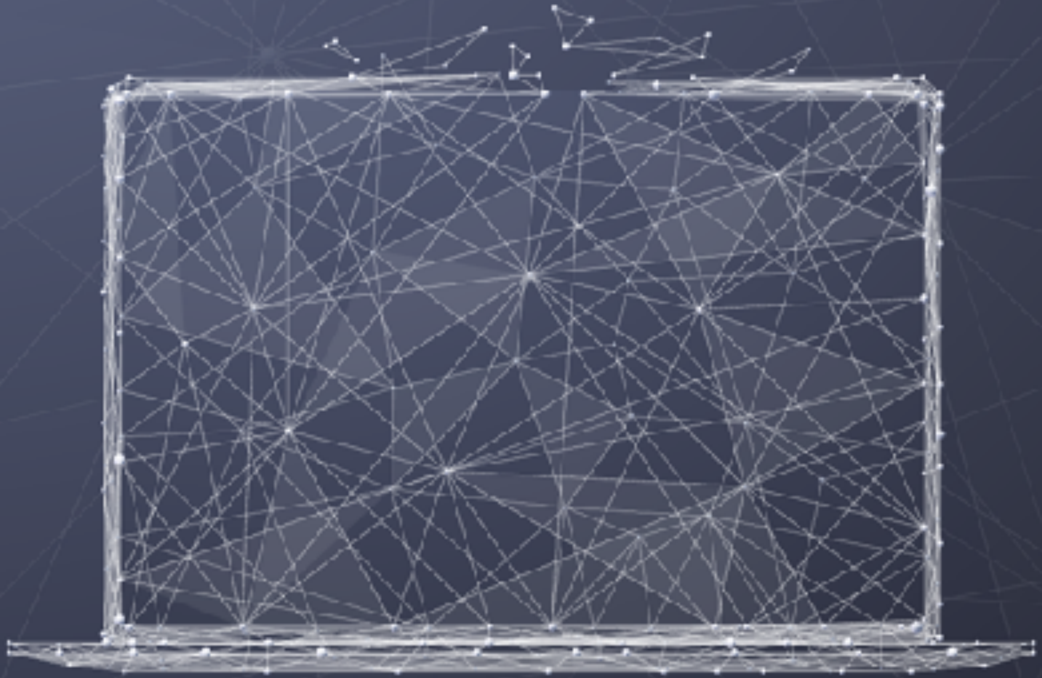
Data Science
Foundations of Decision Making

Hypothesis testing



Purdue University | College of Science

Hypothesis testing

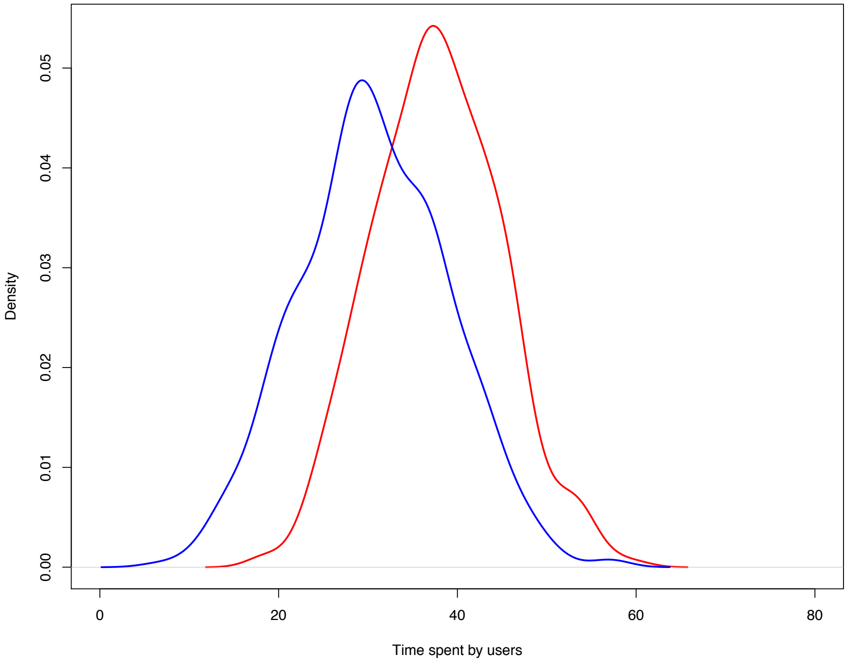
- Using basic statistical techniques, we can devise tests to determine whether observed data gives evidence that some effect “truly” occurs in the real world
- Fundamentally, this evaluates whether things are (likely to be) true about the population (all the data) given a sample (observed data)
- There are many caveats about the precise meaning of these terms, to the point that many people debate the usefulness of hypothesis testing at all
- But, still incredibly common in practice, and important to understand

Basic approach to hypothesis testing

- Posit a null hypothesis H_0 and an alternative hypothesis H_1 (usually just that "H_0 is not true"), e.g.:
  - H_0 : No difference in average user engagement (time spent on site) between design A and design B.
  - H_1 : With new design (B) users spend more time on site, on average
- Collect data x that will be used to test hypothesis, e.g.,
  - Measure user times under both design A (blue) and B (red)
  - Average times: A=30.65s, B=37.7s

Basic approach to hypothesis testing

- Posit a null hypothesis H_0 and an alternative hypothesis H_1 (usually just that “H_0 is not true”), e.g.:
  - H_0 : No difference in average user engagement (time spent on site) between design A and design B.
  - H_1 : With new design (B) users spend more time on site, on average
- Collect data x that will be used to test hypothesis, e.g.,
  - Measure user times under both design A (blue) and B (red)
  - Average times: A=30.65s, B=37.7s



Basic approach to hypothesis testing

- Basic approach: compute the probability of observing the data under the null hypothesis (this is the p-value of the statistical test)

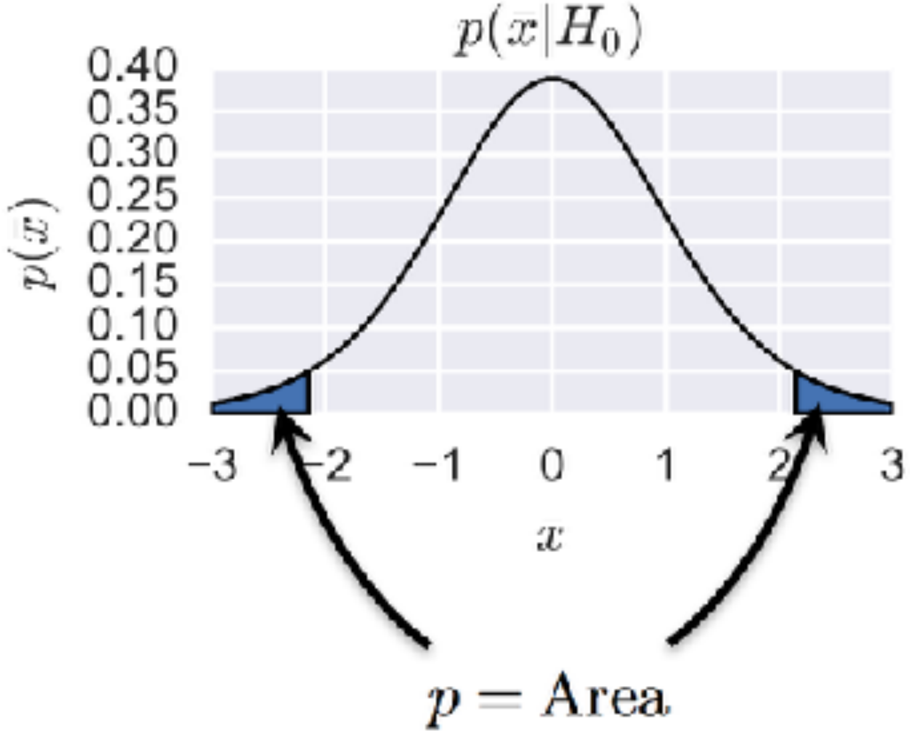
$$p = p(\text{data}|H_0 \text{ is true})$$

- Reject the null hypothesis if the p-value is below the desired significance level (alternatively, just report the p-value itself, which is the lowest significance level we could use to reject hypothesis)

- Important: p-value is p(data | H_0 is true) not p(H_0 not true | data)

Illustration

- Fundamentally this models the distribution $p(\bar{x}\ |H_0)$ and then determines the probability of the observed $\bar{x}$ or a more extreme value



Logic of hypothesis testing

- Effort is made to reject the null hypothesis
- If H_0 is rejected, we tentatively conclude H_1 to be the case
- Otherwise we fail to reject the null (we can never prove the null to be true)
- Role of evidence: Null is assumed true and a contradiction is sought in order to reject it.

Hypothesis testing framework

- Set null (H_0) and alternative (H_1) hypothesis
- Pick significance level (e.g., $\alpha=0.05$)
- Choose test: one vs two sample, one vs two sided, paired or not
- Calculate test statistic and associated p value
- p-value is the probability of observing current data (or more extreme), if the null hypothesis is true
- Make a decision about whether the hypothesis is supported by comparing the p value to the significance level (e.g., if $p<\alpha$ then reject H0)

Hypothesis test example

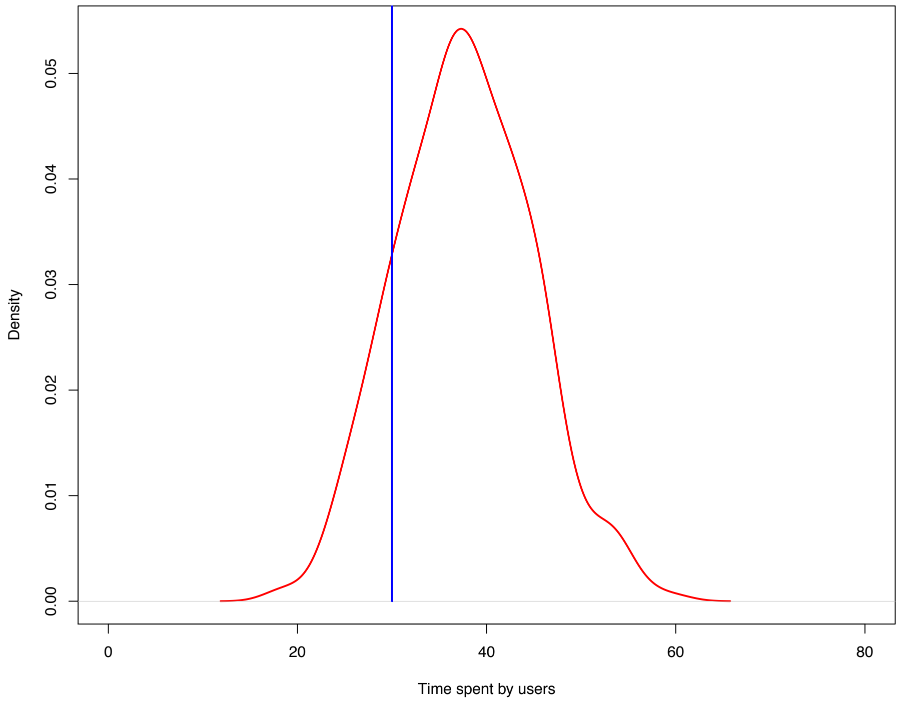
- One sample t-test: tests the mean of a single group against a known mean

Student’s t-test

- The t-test was introduced in 1908 by William Gosset, a chemist working for the Guiness Brewery in Dublin, Ireland — to monitor the quality of stout beer
- Because his employer did not want to reveal the fact that he was using statistics for quality control, Gosset published the test using a pen name “Student” since he was a student of Sir Ronald Fisher
- The test involved calculating the value of “t”
- For larger sample sizes (i.e., $m>30$) the t-distribution converges to normal

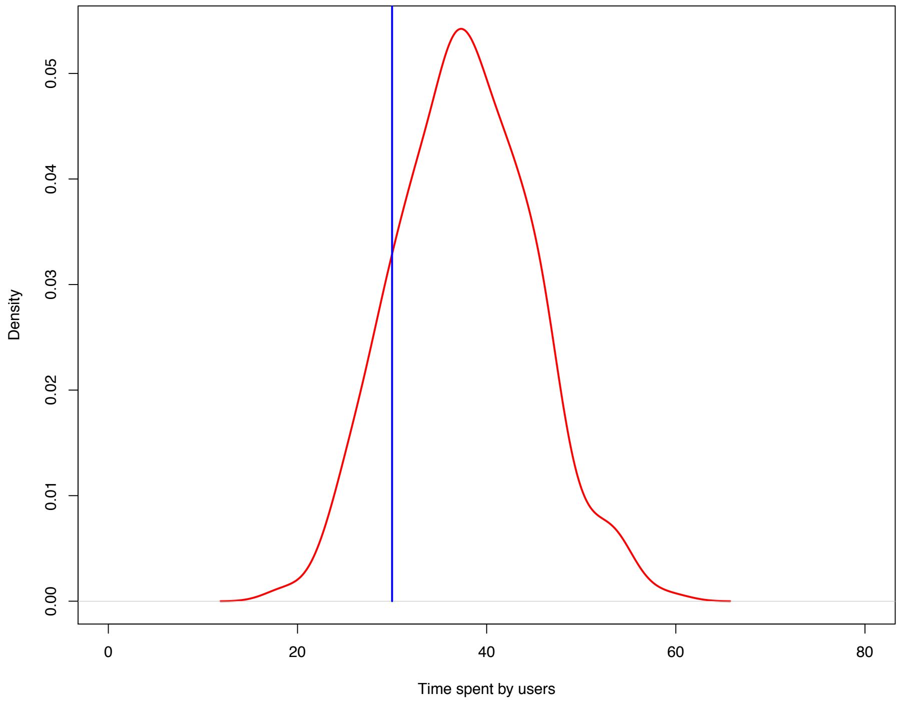
Example

- H_0 : Average user engagement (time spent on site) is 30s.
- H_1 : With new design (B), average user time is >30s.
- Measure user times under design B (red)
  - Observed average is 37.7s
  - Observed variance 51.4



Example

- H_0 : Average user engagement (time spent on site) is 30s.
- H_1 : With new design (B), average user time is >30s.
- Measure user times under design B (red)
  - Observed average is 37.7s
  - Observed variance 51.4



Example (cont)

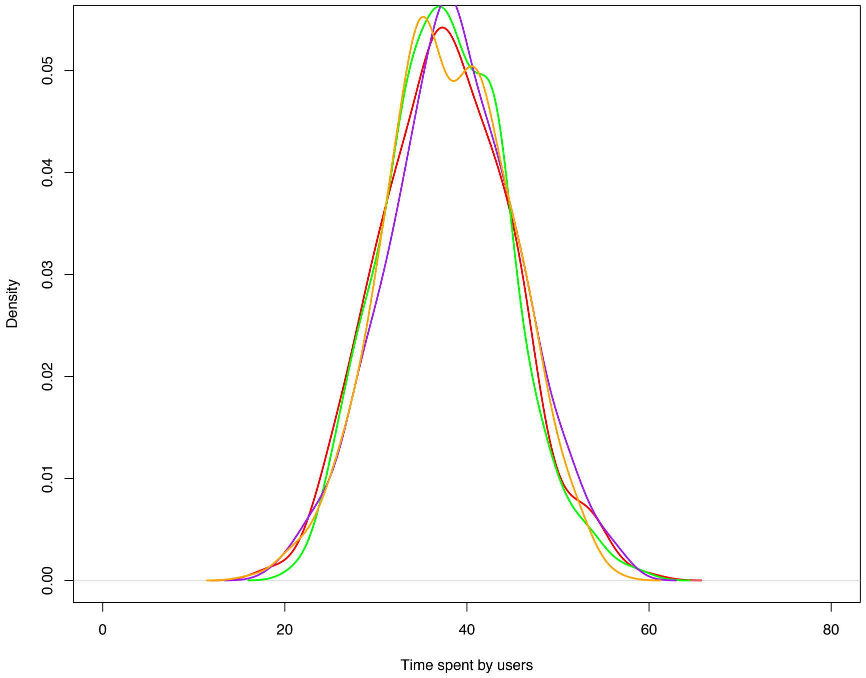
- The sample average is an empirical average over m independent samples from the distribution (in our example m=500)

- If we sampled a different 500 users the observed average may differ, e.g., (37.7, 37.8, 38.5, 37.9)

- Thus the sample average can be considered as a random variable, with mean and variance:

$$\mathbf{E}[\bar{x}] = \mathbf{E}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbf{E}[X] = \mathbf{E}[X] = \mu$$

$$\mathbf{Var}[\bar{x}] = \mathbf{Var}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] = \frac{1}{m^2}\sum_{i=1}^{m}\mathbf{Var}[X] = \frac{\sigma^2}{m}$$

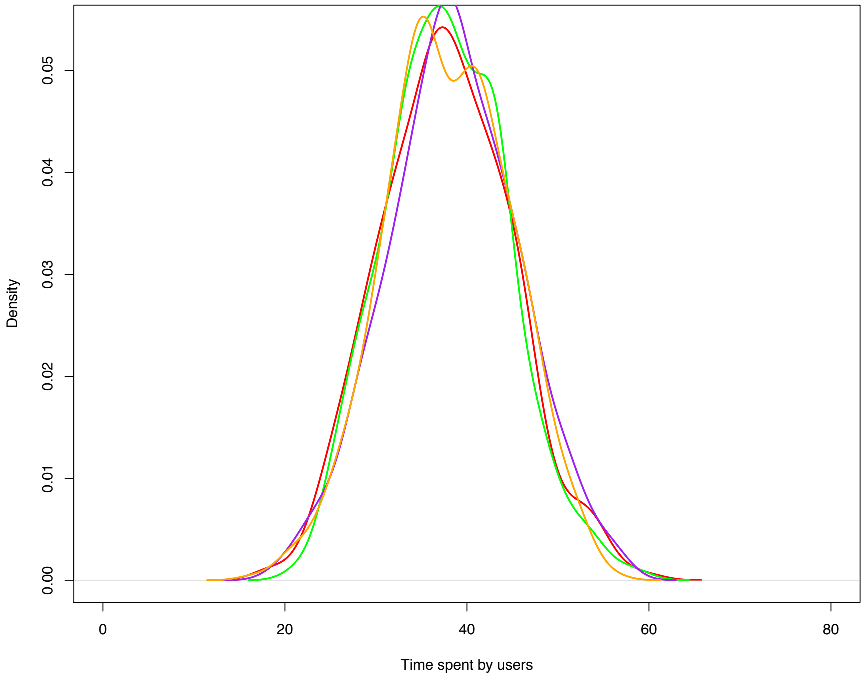


Example (cont)

- The sample average is an empirical average over m independent samples from the distribution (in our example m=500)
- If we sampled a different 500 users the observed average may differ, e.g., (37.7, 37.8, 38.5, 37.9)
- Thus the sample average can be considered as a random variable, with mean and variance:

$$\mathbf{E}[\bar{x}] = \mathbf{E}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbf{E}[X] = \mathbf{E}[X] = \mu$$

$$\mathbf{Var}[\bar{x}] = \mathbf{Var}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] = \frac{1}{m^2}\sum_{i=1}^{m}\mathbf{Var}[X] = \frac{\sigma^2}{m}$$



Central limit theorem

- The central limit theorem states that the sample mean $\bar{x}$ (for "reasonably sized" samples, e.g., m>30) has a Gaussian distribution regardless of the distribution of X

$$\bar{x} \to \mathcal{N}\left(\mu, \frac{\sigma^2}{m}\right)$$

- In practice, for m<30 and when we estimate σ^2 using the sample variance, we should use a Student's t-distribution with m-1 degrees of freedom

$$\bar{x} \to T\left(\mu, \frac{s^2}{m}\right)$$

Central limit theorem

- The central limit theorem states that the sample mean $\bar{x}$ (for “reasonably sized” samples, e.g., m>30) has a Gaussian distribution regardless of the distribution of X

$$\bar{x} \to \mathcal{N}\left(\mu, \frac{\sigma^2}{m}\right)$$

This converges to the population mean as the number of samples (m) → ∞

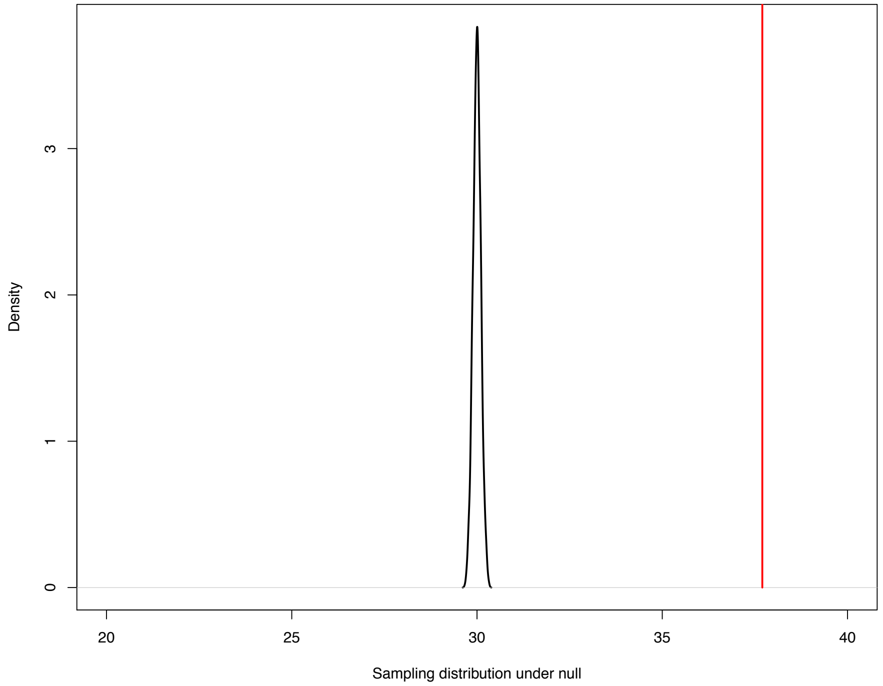
- In practice, for m<30 and when we estimate σ^2 using the sample variance, we should use a Student’s t-distribution with m-1 degrees of freedom

$$\bar{x} \to T\left(\mu, \frac{s^2}{m}\right)$$

Probability of observed data given null

- T-test computes distribution of sample mean, given H_0
- Mean = 30
- Variance = sample var / m
  = 51.4 / 500
- What is probability of observing mean of 37.7 (or higher) under the null hypothesis?

```
import math, scipy.stats as st
xobs = 37.7
xvar = 51.4/500
xnull = 30
1 - st.norm.cdf(xobs,loc=xnull,scale=math.sqrt(xvar))
0.0
```



Hypothesis testing errors

	H_0 true	H_1 true
Accept H_0	Correct	Type II error (false negative)
Reject H_0	Type I error (false positive)	Correct

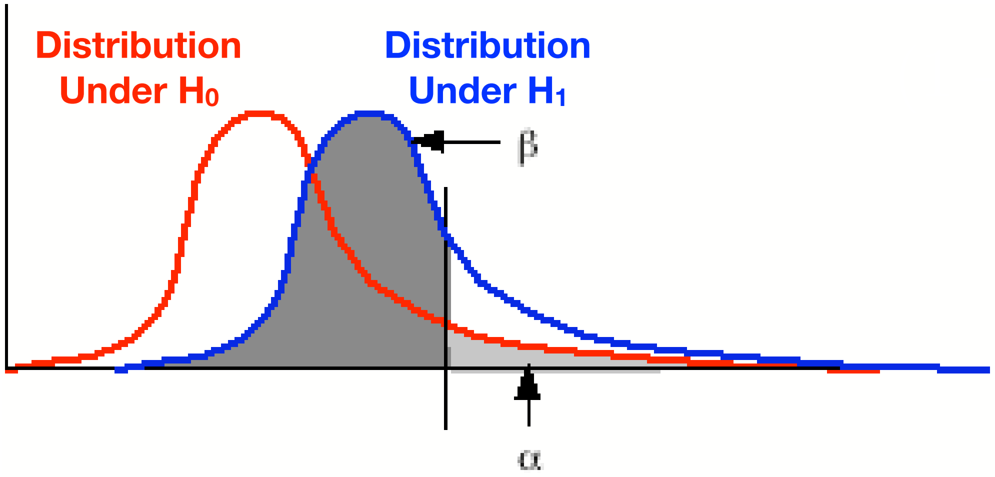
$p(\text{reject } H_0 | H_0 \text{ true})$ = "significance of test"

$p(\text{reject } H_0 | H_1 \text{ true})$ = "power of test"

- Type I error: null is rejected when it is true (generally considered to be most serious type of error)
  - E.g., conclude cancer drug increases life expectancy when in fact it doesn't
- Type II error: null is accepted when it is false
  - E.g., conclude that cancer drug does not increase life expectancy when in fact it does

Statistical power

- Lack of statistical significance does not necessarily imply that H_0 is true
- Test could have low statistical power: $(1 - \beta)$ portion of H_1 that is above threshold



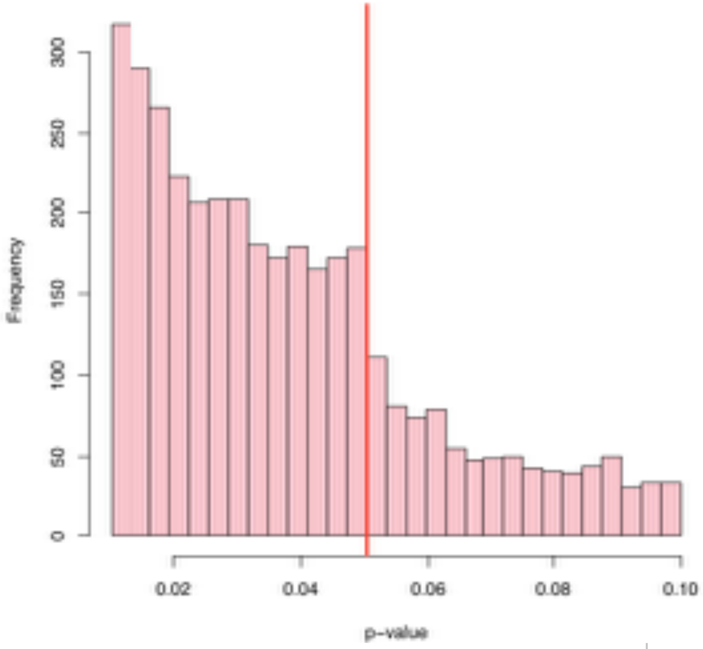
$$\beta = p(accept\ H_0 | H_0\ false) = p(type\ 2\ error)$$

How to increase statistical power

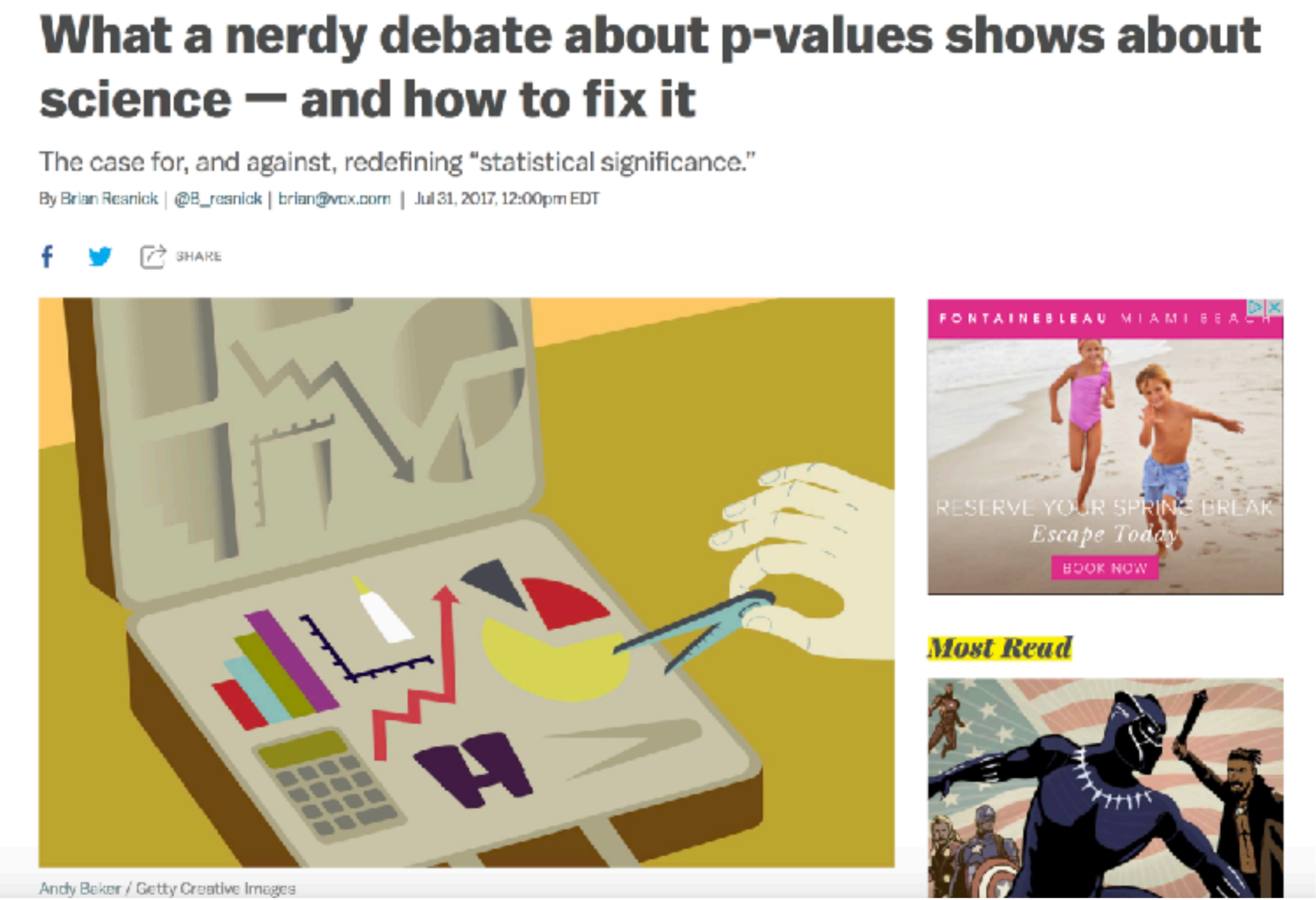
- Increase sample size
- Decrease sample variability
  - Pair observations, control for confounding variables, increase precision of measurements, …
- Increase effect size
  - More extreme experimental conditions, avoid ceiling/floor effects
- Increase α (e.g., from 0.05 to 0.10, but this increases Type I errors)

P values considered harmful

- P values correspond to the probability of your data given the null hypothesis, but frequently they are interpreted as the probability of the null hypothesis given the data
- The p value cannot tell you your hypothesis is correct
- The p value does not tell you the probability that your result occurred just by random chance
- P values cannot tell you the size of the effect, or the importance of a result
- $p < 0.05$ is not a line that separates real results from false ones — they are simply one piece evidence to consider in the investigation of significance of patterns



Histogram of p values from ~3,500 published journal papers
(from E. J. Masicampo and Daniel Lalande, *A peculiar prevalence of p values just below .05*, 2012)



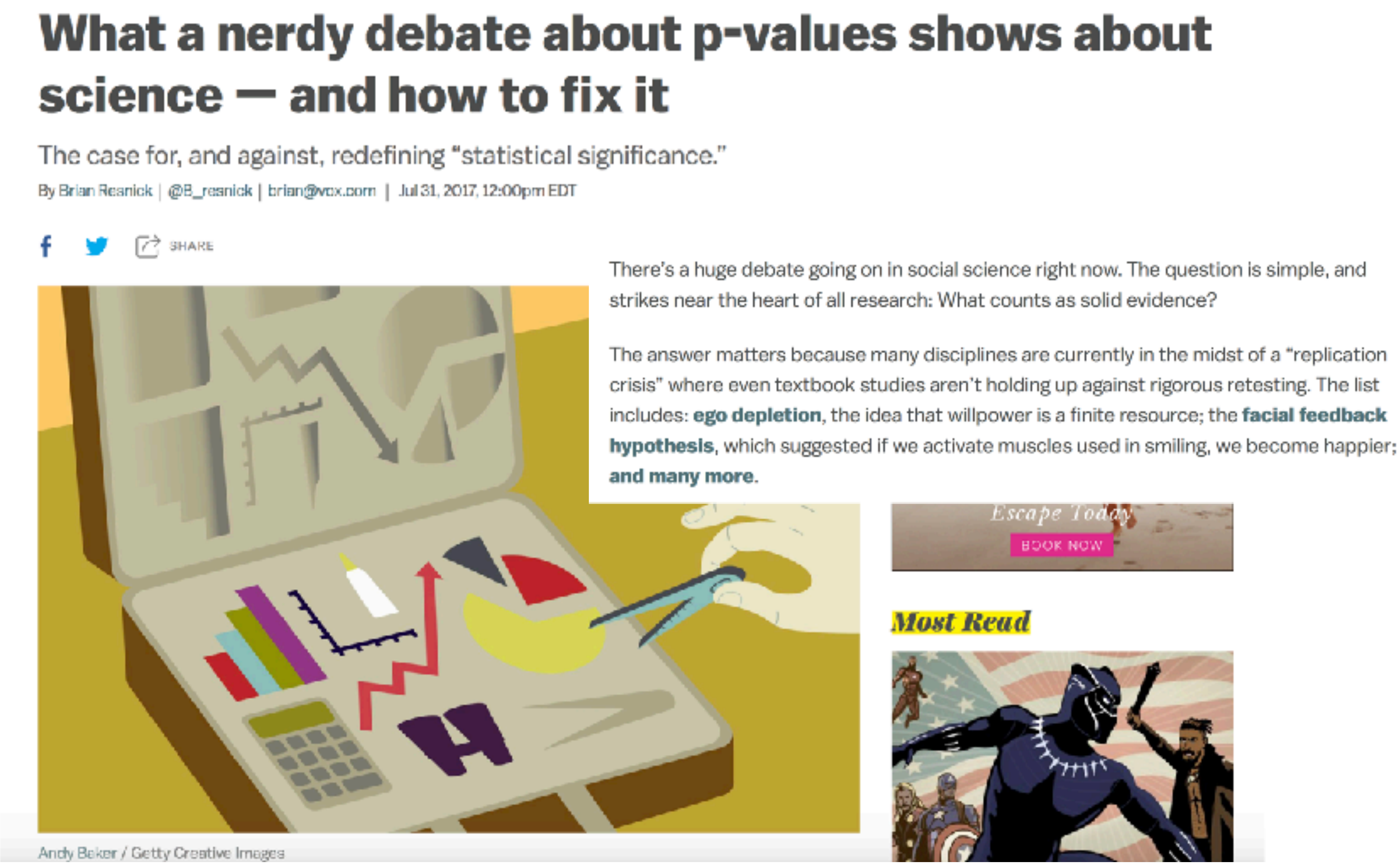
What a nerdy debate about p-values shows about science — and how to fix it

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

By Brian Resnick | @B_resnick | brian@vox.com | Jul 31, 2017, 12:00pm EDT

Andy Baker / Getty Creative Images

Most Read



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

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

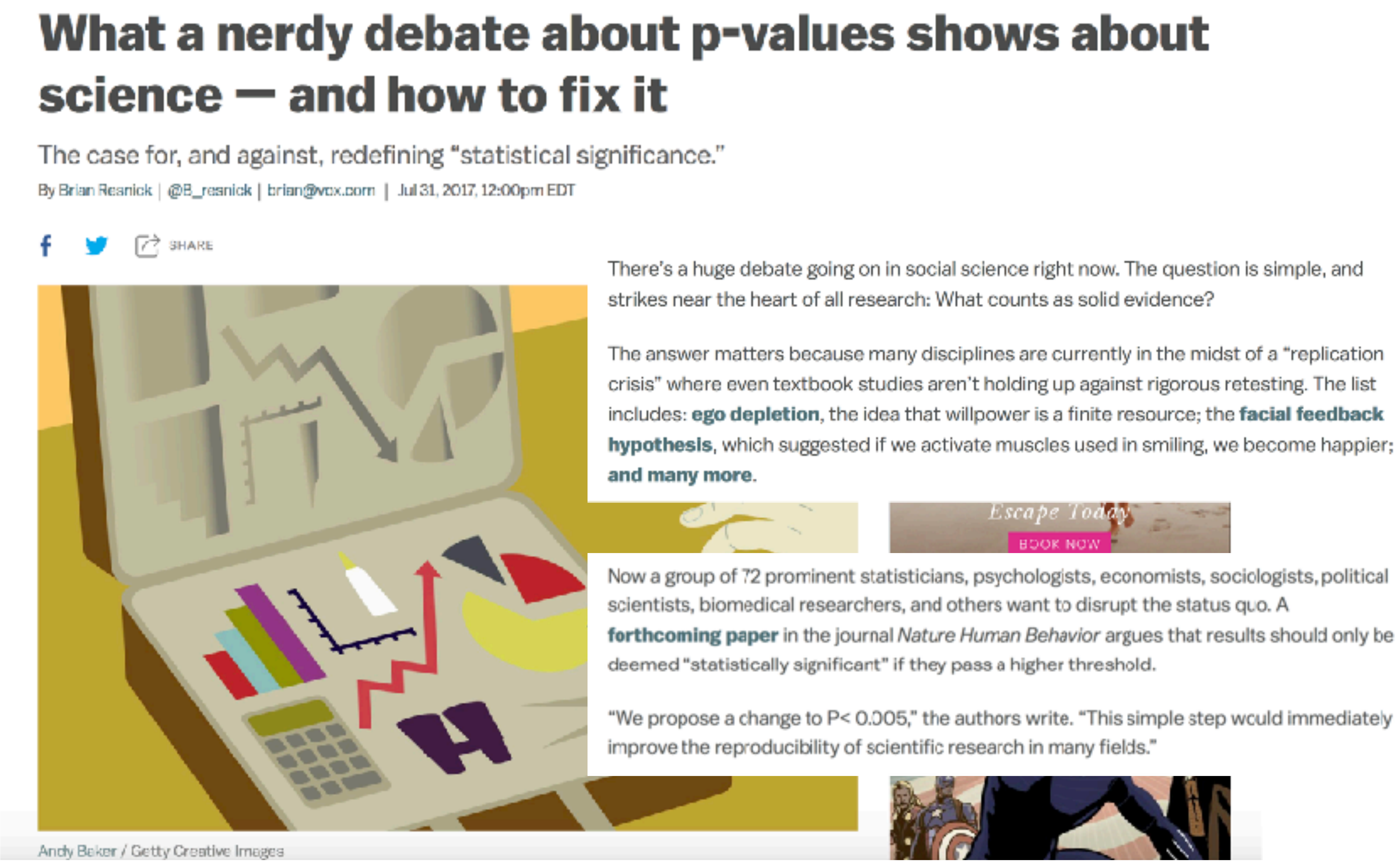
By Brian Resnick | @B_resnick | brian@vox.com | Jul 31, 2017, 12:00pm EDT

There's a huge debate going on in social science right now. The question is simple, and strikes near the heart of all research: What counts as solid evidence?

The answer matters because many disciplines are currently in the midst of a "replication crisis" where even textbook studies aren't holding up against rigorous retesting. The list includes: **ego depletion**, the idea that willpower is a finite resource; the **facial feedback hypothesis**, which suggested if we activate muscles used in smiling, we become happier; **and many more**.

Andy Baker / Getty Creative Images

Most Read



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

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

By Brian Resnick | @B_resnick | brian@vox.com | Jul 31, 2017, 12:00pm EDT

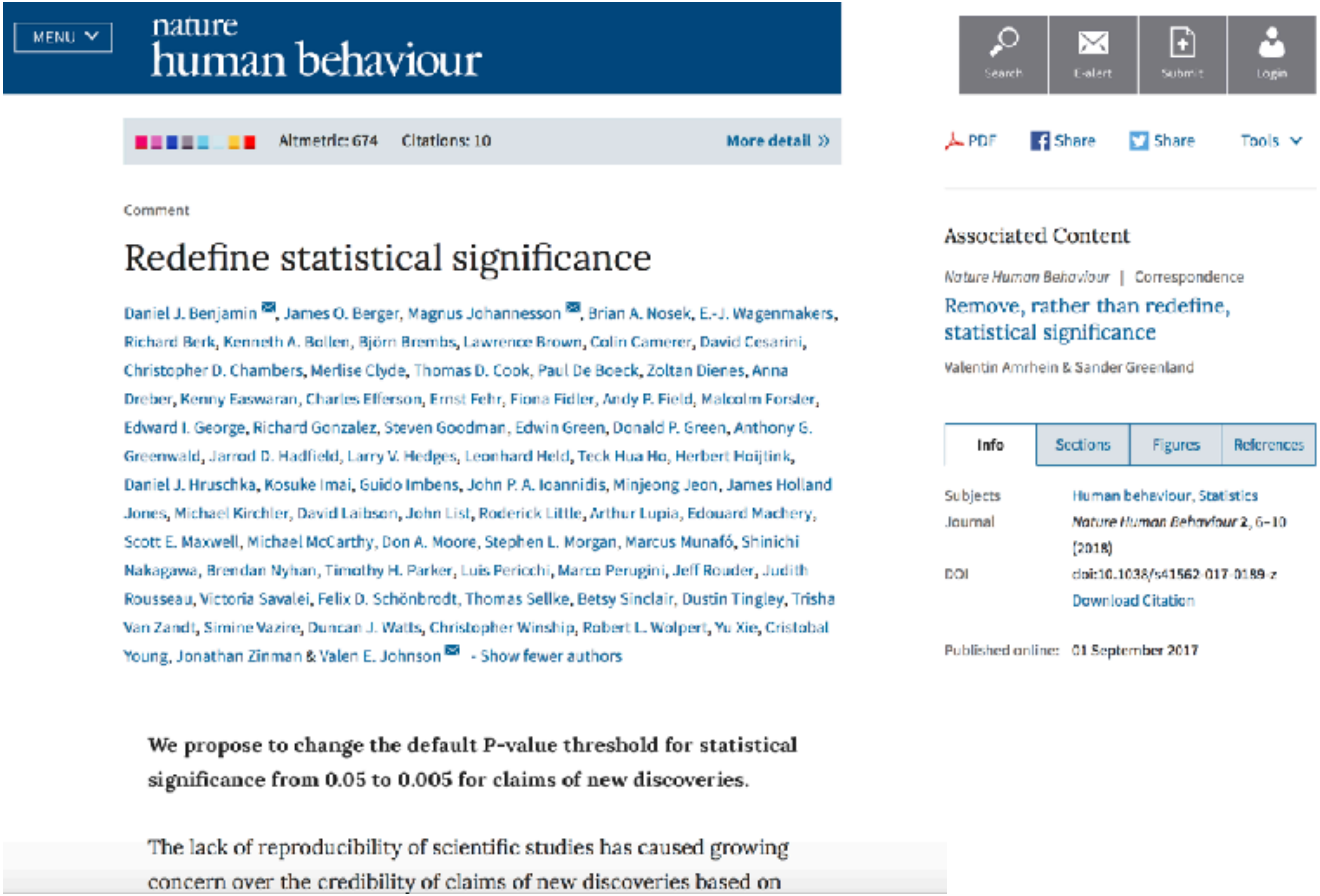
There's a huge debate going on in social science right now. The question is simple, and strikes near the heart of all research: What counts as solid evidence?

The answer matters because many disciplines are currently in the midst of a "replication crisis" where even textbook studies aren't holding up against rigorous retesting. The list includes: **ego depletion**, the idea that willpower is a finite resource; the **facial feedback hypothesis**, which suggested if we activate muscles used in smiling, we become happier; **and many more**.

Now a group of 72 prominent statisticians, psychologists, economists, sociologists, political scientists, biomedical researchers, and others want to disrupt the status quo. A **forthcoming paper** in the journal *Nature Human Behavior* argues that results should only be deemed "statistically significant" if they pass a higher threshold.

"We propose a change to $P < 0.005$," the authors write. "This simple step would immediately improve the reproducibility of scientific research in many fields."

Andy Baker / Getty Creative Images



Comment

Redefine statistical significance

Daniel J. Benjamin ✉, James O. Berger, Magnus Johannesson ✉, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony G. Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don A. Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman & Valen E. Johnson ✉ - Show fewer authors

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

Associated Content

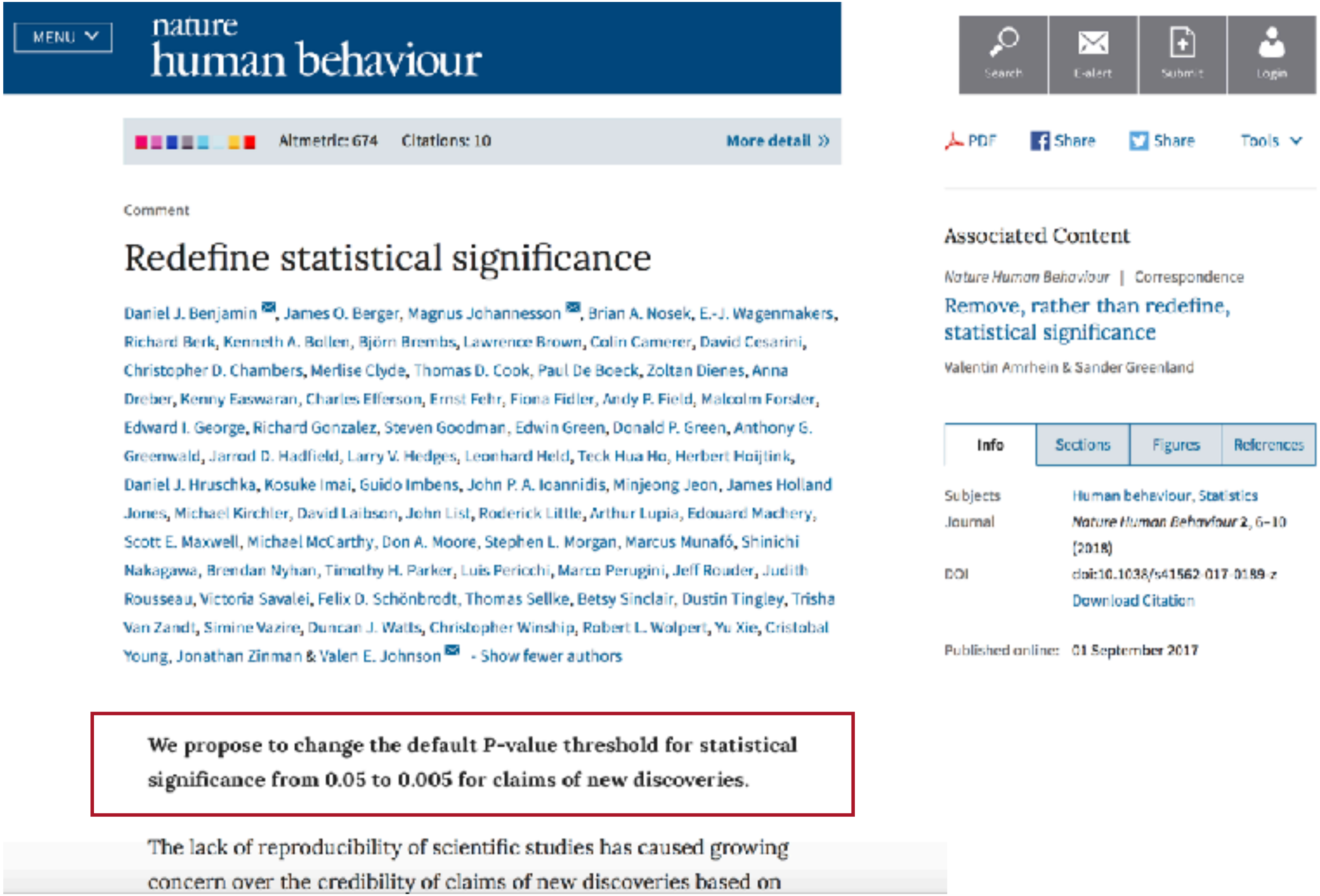
Nature Human Behaviour | Correspondence

Remove, rather than redefine, statistical significance

Valentin Amrhein & Sander Greenland

Subjects Human behaviour, Statistics
Journal *Nature Human Behaviour* **2**, 6–10 (2018)
DOI doi:10.1038/s41562-017-0189-z
Download Citation

Published online: 01 September 2017



nature
human behaviour

Altmetric: 674 Citations: 10

Comment

Redefine statistical significance

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony G. Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don A. Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman & Valen E. Johnson - Show fewer authors

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

Associated Content

Nature Human Behaviour | Correspondence

Remove, rather than redefine, statistical significance

Valentin Amrhein & Sander Greenland

Subjects Human behaviour, Statistics

Journal *Nature Human Behaviour* **2**, 6–10 (2018)

DOI doi:10.1038/s41562-017-0189-z

Download Citation

Published online: 01 September 2017