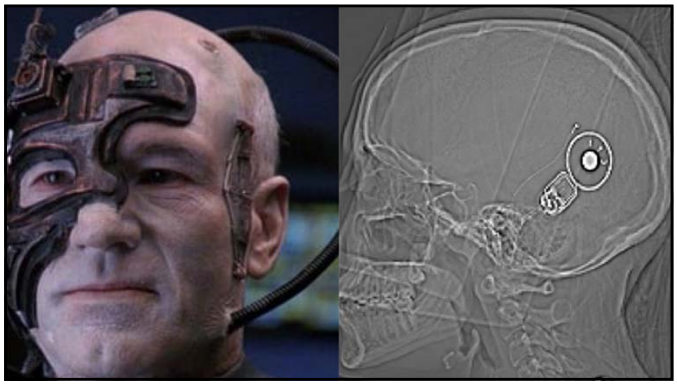
Statistical Significance, P-Values & Communicating Quantitative Research

Regina Nuzzo, Ph.D.
Senior Advisor for Statistics Communication, American Statistical Association

National Birth Defects Prevention Annual Meeting
March 10, 2020

1



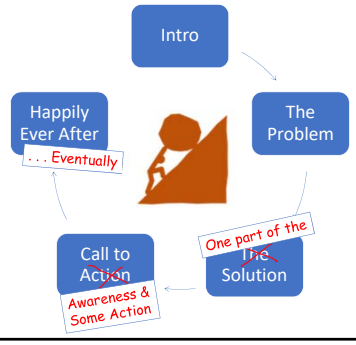
2

3

Typical "Hero's Journey" Narrative Arc

4

Today's "P-Values' Journey" Narrative Arc



5

P-value "clarified" in the 2016 ASA Statement

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*

6

6

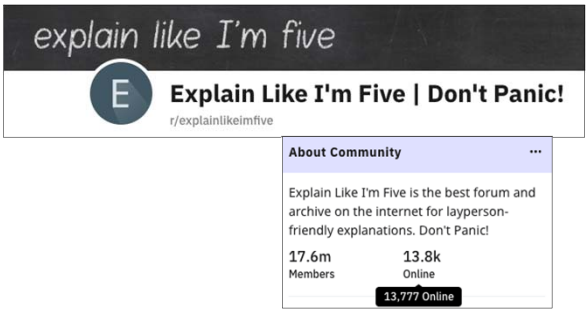
Perhaps this is clearer

[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)

7

7



8

8

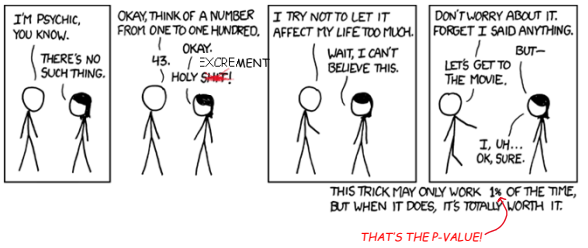
P-Values: The ELI5 ("Explain it like I'm 5") version

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

9

9

P-Values, the xkcd version



https://xkcd.com/628/

10

Romantic Red: Red Enhances Men's Attraction to Women

Andrew J. Elliot and Daniela Niesta
University of Rochester

In many nonhuman primates, the color red enhances males' attraction to females. In 5 experiments, the authors demonstrate a parallel effect in humans: Red, relative to other achromatic and chromatic colors,

"Imagine that you are going on a date with this person and have $100 in your wallet.

How much money would you be likely to spend on your date?"

Elliot, A. J., & Niesta, D. (2008). Journal of personality and social psychology, 95(5), 1150.

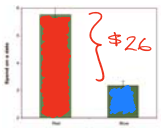
11

Romantic Red: Red Enhances Men's Attraction to Women

Andrew J. Elliot and Daniela Niesta
University of Rochester

In many nonhuman primates, the color red enhances males' attraction to females. In 5 experiments, the authors demonstrate a parallel effect in humans: Red, relative to other achromatic and chromatic colors,

The analysis on spend on a date revealed a significant effect of color, $t(21) = 3.19$, $p < .01$, $d = 1.35$. As displayed in Figure 5e,



Wearing red: $58

Wearing blue: $32

Elliot, A. J., & Niesta, D. (2008). Journal of personality and social psychology, 95(5), 1150.

12

P < 0.01 = "Holy 🙏 🙏"
What's behind this small p-value?

- **There was a fluke**.
  - Something unusual happened in the data just by chance.
  - *The smaller the sample size, the greater the chances of a fluke.*
- **Something was violated**.
  - There was a mismatch between *what was actually done* in the data analysis and *what needed to be done* for the p-value to be a valid indicator.
  - For example, was the data analysis planned before looking at the data? Were all analyses and results presented, no matter the outcome? And all strange, nit-picky rules that are part of the deal when using p-values?
  - *A small p-value might simply be a sign that data analysis rules were broken.*
- **There was a real but tiny relationship**, so tiny that we shouldn't care about it.
  - *The larger the sample size, the more clinically meaningless effects will be picked up.*
- **There was a relationship that is worth more study**.
  - Can it be replicated under other conditions? Is it generalizable? How does it relate to other studies?
- ***Or any combination of the above.***

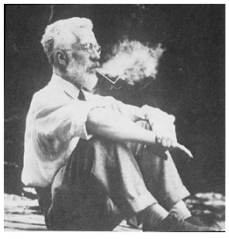
13

Common misinterpretations of p = 0.01

- There is only a 1% chance the two groups were different
- There is only a 1% chance of getting the result we did by chance alone
- The probability the null hypothesis is false is 99%
- If we were to repeat this study, there is a 99% chance of it replicating

14

14



R. A. Fisher called such results "significant"

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

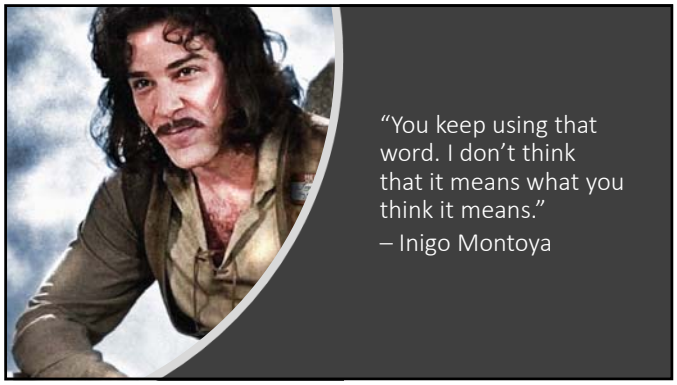
sig·nif·i·cant
/sig nifikənt/

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"

15

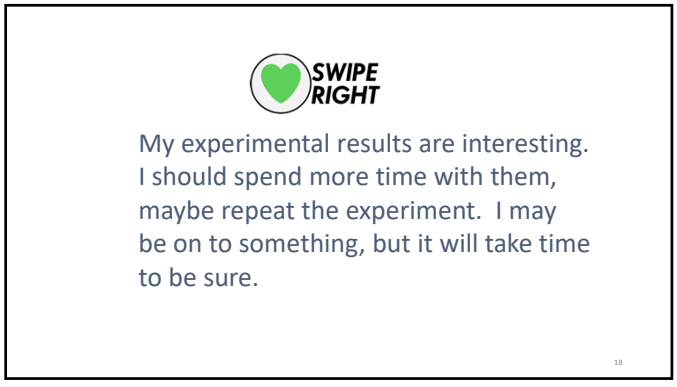
15


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

16


tinder
0.01
P-value
25
8

17


SWIPE RIGHT
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.

18

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!

19

19

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

Thanks to Matthew Hankins for these quotes
https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/

20

20

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

Thanks to Matthew Hankins for these quotes
https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/

21

21

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/

22

22

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).

23

23

A fundamental problem

Unfortunately, a p-value is a probability statement about our data assuming the hypothesis!

That is, P(D|H).

24

24

No equivalence here

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

25

25

The problem, illustrated

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

Lacking data, let's make up a number: P(D|H)=.98

(only 2% hanging survival rate)

26

26

The problem, illustrated

Now reverse:

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)

27

27

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

P(D|H) = probability of Dying given that you were Hanged = 98%

P(H|D) = probability of being Hanged given that you Died = 0.01%

"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."

28

28

Why the 2016 ASA statement?

- "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

29

29

"Let's be clear. Nothing in the ASA statement is new."

Statisticians and others have been sounding the alarm about these matters for decades, to little avail.

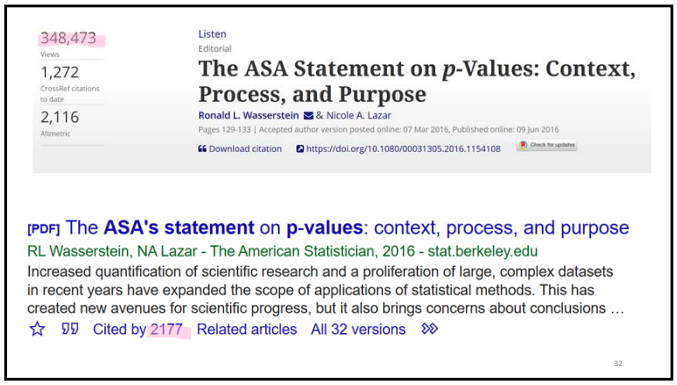
(Wasserstein and Lazar, 2016)

30

30


Odds Are, It's Wrong
Science fails to face the shortcomings of statistics
Magazine issue: Vol. 177 #7, March 27, 2010, p. 26
P value ban: small step for a journal, giant leap for science
nature
Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video
Archive | Volume 506 | Issue 7487 | News Feature | Article
NATURE | NEWS FEATURE
Scientific method: Statistical errors
P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.
Regina Nuzzo
12 February 2014
31

31


348,473
Views
1,272
CrossRef citations to date
2,116
Altmetric
Listen
Editorial
The ASA Statement on p-Values: Context, Process, and Purpose
Ronald L. Wasserstein & Nicole A. Lazar
Pages 129-133 | Accepted author version posted online: 07 Mar 2016, Published online: 09 Jun 2016
Download citation
https://doi.org/10.1080/00031305.2016.1154108
[PDF] The ASA's statement on p-values: context, process, and purpose
RL Wasserstein, NA Lazar - The American Statistician, 2016 - stat.berkeley.edu
Increased quantification of scientific research and a proliferation of large, complex datasets in recent years have expanded the scope of applications of statistical methods. This has created new avenues for scientific progress, but it also brings concerns about conclusions …
Cited by 2177 Related articles All 32 versions
32

32


vevo
0:34 / 4:01
Taylor Swift - Shake It Off
2,852,167,472 views • Aug 18, 2014
33

33

Scientific Studies: Last Week Tonight with John Oliver (HBO)

15M views · 3 years ago

34

34

"(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

(Rothman, supplement to the 2016 ASA statement)

35

35

Biggest takeaway from ASA Statement:

Bright line thinking is bad for science.

36

36

If we eliminate 'p < 0.05' bright-line thinking . . .

. . . What could you do to get your paper published, your research grant funded, your drug approved, your policy or business recommendation accepted?

37

37

What might be behind our p-value

- There was a fluke.
- Something was violated.
- There was a real but tiny relationship, so tiny that we shouldn't care about it.
- There was a relationship that is worth more study.

38

The case that we can make for our p-value and our findings

- There might have been a fluke, and we're OK with that.
- Something was NOT violated, and here's why we say that . . .
- The relationship was big enough for us to care about.
- Here is the other evidence behind our finding . . .

39

There might have been a fluke, and we're OK with that.

We accept statistical uncertainty as a given.

40

40

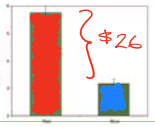
Something was *NOT* violated, because we followed best practices:

- Preregistration and prespecified analyses
- Separating exploratory and confirmatory data analyses
- Open data and open code
- Completeness in reporting

41

The relationship was big enough for us to care about.

The analysis on spend on a date revealed a significant effect of color, $t(21) = 3.19$, $p < .01$, $d = 1.35$. As displayed in Figure 5e,



Wearing red: $58

Wearing blue: $32

- Men viewing women in the Red condition were willing to spend an average of $26 more on dinner than men viewing women in the Blue condition.
- Other studies have shown that $10 is the minimum noticeable difference in dinner spending.
- The 95% CI here was [$9, $43], suggesting that the results are fairly consistent with a meaningful effect.

Elliot, A. J., & Niesta, D. (2008). Journal of personality and social psychology, 95(5), 1150.

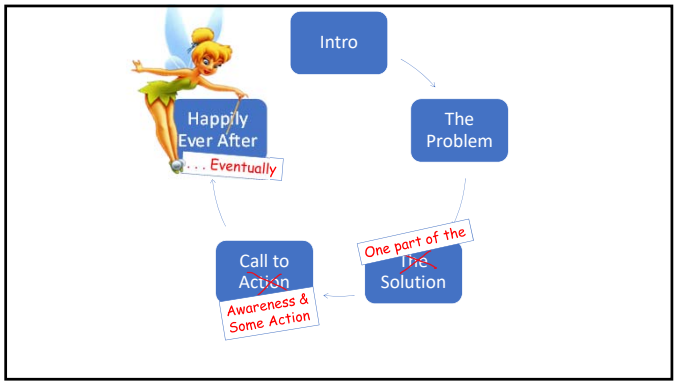
42

And here's the other evidence behind our findings, and why we think the relationship is worth more study . . .

Consider "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, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. "Abandon statistical significance." *The American Statistician* 73, no. sup1 (2019): 235-245.

43



44

45

Photo by Pierre Bamin - Creative Commons No known copyright restrictions

Photo by Muhammad Haikal Sjukri on Unsplash

Photo by Riley McCullough - Creative Commons No known copyright restrictions

46