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

Did she just say “Deaflandia”??
Typical “Hero’s Journey” Narrative Arc

Intro

The Problem

Happily Ever After

Call to Action

The Solution

Today’s “P-Values’ Journey” Narrative Arc

Intro

The Problem

Call to Action

Happily Ever After

Eventually

One part of the Solution

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

1The 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 \( \{R_n\}_{n \in [0,1]} \) be a collection of \( \mathcal{P} \)-measurable subsets of \( \mathcal{X} \) such that (1) \( \mathbb{P}(R_n) = \alpha \) and (2) if \( \alpha' < \alpha \) then \( R_{\alpha'} \subseteq R_{\alpha} \). Then the p-value of \( H_0 \) for data \( X = x \) is \( \inf_{\alpha \in [0,1]} \{ \alpha : x \in R_{\alpha} \} \).

(Stark, 2016)

---

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?
P-Values, the xkcd version

![xkcd P-Value comic](https://xkcd.com/628/)

EXCREMENT

THAT'S THE P-VALUE!

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?


Wearing red: $58

Wearing blue: $32

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.
  • Can it be replicated under other conditions? Is it generalizable? How does it relate to other studies?
  • Or any combination of the above.

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.

R. A. Fisher called such results “significant.” To Fisher, this meant that the result was worth further scrutiny.
“You keep using that word. I don't think that it means what you think it means.”
– Inigo Montoya

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 spend the rest of my life with. Let’s publish and get grants together. I love you!

- 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/

- 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/
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/

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).

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, what is \( P(D|H) \)?

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

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)

\[
P(D|H) = \text{probability of Dying given that you were Hanged} = 98\% \\
P(H|D) = \text{probability of being Hanged given that you Died} = 0.01\%
\]

"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."

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."
- 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/quotsaghn.html](http://www.indiana.edu/~stigtsts/quotsaghn.html)

“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)
'Scientists 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)

Biggest takeaway from ASA Statement: **Bright line thinking is bad for science.**
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?

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.

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 . . .
There might have been a fluke, and we’re OK with that. We accept statistical uncertainty as a given.

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

The relationship was big enough for us to care about.

- 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.


- Wearing red: $58
- Wearing blue: $32
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.”
