MATHEMATICA
CENTER FOR
IMPROVING RESEARCH EVIDENCE

To p or Not to p?

Understanding Statistical Significance and Its Role in Developing Evidence for Policymaking

A Center for Improving Research Evidence (CIRE) Forum

Washington, DC

March 1, 2017

Ann Person • Allen Schirm • Ronald Wasserstein
Stuart Buck • Molly Irwin • Timothy Day

Welcome

Ann Person, CIRE director

About CIRE

- The Center for Improving Research Evidence (CIRE):
  - Draws upon Mathematica's 40+ years of experience using rigorous evaluation designs to assess the impact of social policy and programs
  - Uses qualitative and quantitative analysis to build a better understanding of what programs work best, where, and for whom
  - Works to bridge the gap between research and practice

Moderator

Allen Schirm, Mathematica

Polling Question #1

Have you read the ASA statement on p-values and statistical significance?

- **Yes**
- **No**
- **I skimmed it!**

Today's Speakers

Molly Irwin, U.S. Department of Labor

Timothy Day, CMMI

Ron Wasserstein, American Statistical Association

Stuart Buck, Laura and John Arnold Foundation

MATHEMATICA
CENTER FOR
IMPROVING RESEARCH EVIDENCE

Doctor, It Hurts When I *p*

Ronald L. Wasserstein, Executive Director, ASA

Mathematica Forum

March 1, 2017

The Talk

- They think they know all about it already, because they learned about it from others like them.
- It is not nearly as interesting as they thought it would be.
- They've stopped listening before you've stopped talking.
- Chances are, they now understand it even less.

Why did the ASA issue a “statement on p-values and statistical significance?”

- "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

“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)

Why did the ASA issue a "statement on p-values and statistical significance?"

FEATURE HUMANS & SOCIETY, NUMBERS

Odds Are, It's Wrong

Science fails to face the shortcomings of statistics

BY TOM SIEGFRIED 2:40PM, MARCH 12, 2010

Magazine issue: Vol. 177 #7, March 27, 2010, p. 26

ScienceNews

MAGAZINE OF THE SOCIETY FOR SCIENCE & THE PUBLIC

Why did the ASA issue a "statement on p-values and statistical significance?"

Science fails to face the shortcomings of statistics

A journal went so far as to ban p-values

CONTEXT NUMBERS

P value ban: small step for a journal, giant leap for science

Editors reject flawed system of null hypothesis testing

BY TOM SIEGFRIED 3:18PM, MARCH 17, 2015

P-value “clarified” (in the 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*

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)

What goes into the p-value? Many things!

- Assumption that the null hypothesis is true is typically the only thing considered
- However, much more than that goes into the p-value. Many choices by the researcher can affect it.

ASA statement articulates 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.
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.

Does the ASA statement go far enough?

- The ASA statement does not go as far as it should go.
- However, it goes as far as it could go.

Biggest takeaway message from the 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)

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
- "... surely, God loves the .06 nearly as much as the .05." (Rosnell and Rosenthal 1989)

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

And, God forbid, 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

We want P(H|D) but p-values give P(D|H)

The problem illustrated (Carver 1978)

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

Obviously, it will be very high, perhaps .97 or higher.

The problem illustrated (Carver 1978)

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)?

This time the probability will undoubtedly be very low, perhaps .01 or lower.

The problem illustrated (Carver 1978)

No one would be likely to make the mistake of substituting the first estimate (.97) for the second (.01); that is, to accept .97 as the probability that a person has been hanged given that the person is dead.

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

Inference is hard work.

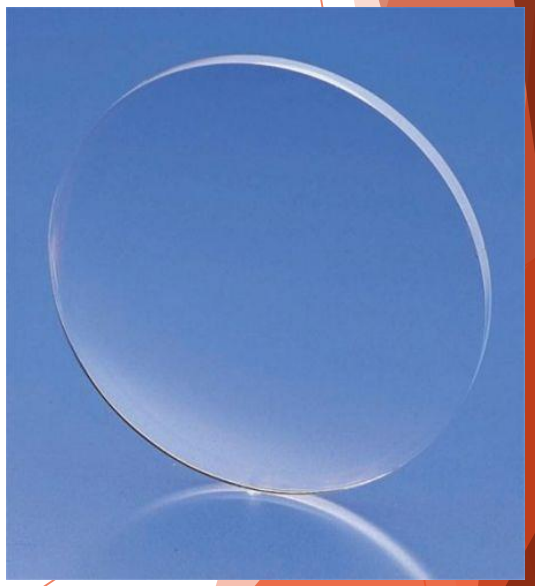
- Simplistic (“cookbook”) rules and procedures are not a substitute for this hard work.
- Cookbook + artificial threshold for significance = appearance of objectivity

In a world where $p<0.05$ carried no meaning…

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

You’d have to be convincing!

You will also have to be transparent



Wrapping up:

P-values themselves are not the problem, but…

- They are hard to explain
- They are easy to misunderstand
- They don't directly address the question of interest
- When mixed with bright line thinking, they lead to bad science.
- So, maybe if you have only been dating p-values, it's time to start seeing some other statistics.

Scientific Method for the 21st Century: A World Beyond $p < 0.05$

Haiku

Little p-value
what are you trying to say
of significance?

-Steve Ziliak

Questions?

ron@amstat.org

@RonWasserstein

Polling Question #2

Have you personally witnessed a misinterpretation of a p-value or significance test?

- **Yes**
- **No**
- **Can't remember**

Panel Discussion

Molly Irwin,
U.S. Department
of Labor

Timothy Day,
CMMI

Ron Wasserstein,
American
Statistical
Association

Stuart Buck,
Laura and
John Arnold
Foundation

Audience Questions?

- **Webinar audience: Submit questions with your name and organization through the Q&A widget**

- **In-person audience: State your name and organization before asking your question**

For More Information

- Mathematica's Center for Improving Research Evidence
  - CIRE@mathematica-mpr.com
  - Ann Person: aperson@mathematica-mpr.com

Networking Reception Starts
Now
Mathematica Lobby, 12th Floor
4:30–5:30 p.m.