What Should We Teach About p-values in Introductory Statistics?

Nicole A. Lazar

Department of Statistics
Pennsylvania State University

A Crisis of Reproducibility? Is the p-value to Blame?

Ioannidis (2005): Why Most Published Research Findings are False. Other papers follow, that point out misuses and abuses of statistical methodology.

Trafimow (2014): *Basic and Applied Social Psychology* will no longer require null hypothesis significance testing procedures (NHSTP) or inferential statistics more broadly.

Trafimow and Marks (2015): *Basic and Applied Social Psychology* bans the use of p-values, as well as "all vestiges of the NHSTP (p-values, t-values, F-values, statements about 'significant' differences or lack thereof, and so on)."

A Replicability/Reproducibility Crisis?

Concurrently, problems with replicability and reproducibility in various areas of science (neuroscience, social psychology, cancer trials, ...) gain more attention, including in popular press (e.g. *The New Yorker* (2010): The Truth Wears Off).

Open Science Framework Reproducibility Project (2011, psychology) finds low levels of reproducibility in published research. Ongoing project in cancer biology (2019).

Crisis in science? What is the role of statistics – and specifically the use of thresholds such as $p < 0.05$ for publication? P-hacking, researcher degrees of freedom, many signs of abuse of statistical methods in the literature.

What To Do?

Major proposals include:

- Ban p-values altogether (e.g. Trafimow and colleagues)
- Keep p-values but lower the default threshold for declaring "statistical significance" from 0.05 to 0.005 (Benjamin *et al.* 2017; 72 signed authors)
- Keep p-values but pick the significance threshold according to circumstances and context – "justify α" (Lakens *et al.* 2018; 88 signed authors)
- Keep p-values but avoid the use of thresholds and declarations of statistical significance (e.g. Wasserstein *et al.* 2019)
- Keep the status quo but do a better job of educating students/scientists/the public

A Confusing State of Affairs

Ideas around inference are in flux! Students ask: "What is next? What should we do?"

Instructors also have questions (e.g. a recent email with subject line: "Help for an instructor who hopes to move to a world beyond $p < 0.05$; webinars aimed at community college statistics instructors; . . .)

Not enough to tell students, instructors, and researchers "Don't do this . . . "

As a community, we also need to provide alternatives.

One Framework: The ATOM(IC) Principles

Accept uncertainty

Be **T**ransparent

Be **O**pen

Be **M**odest

Institutional

Change

More Generally: Emphasize Interpretation

A lot of focus in Introductory Statistics classes is *formulaic*: compute a test statistic using a formula; compute a p-value (hard to understand what it is) using a table or software; declare "statistically significant" or not; call it a day.

Leaves students with the impression that the main focus should be on whether or not $p < 0.05$ (typical α).

From there, easy to conclude that $p < 0.05$ means "the finding is true" and $p > 0.05$ means "nothing at all of interest here".

What about *interpretation*?

1. What does a p-value really mean?
2. What is the **scientific context** of the problem (questions of interest)?
3. What is the **statistical context** of the problem (study design, sample size, quality of measurements, model assumptions)?
4. What does the observed effect size mean *in the context of the problem*?
5. What does the p-value mean *in the context of the problem*?

One Framework: Simulation-Based Inference (SBI)

Rather than formulae ("plug and chug") use simulation to make difficult concepts like the p-value more concrete.

Some obvious advantages include:

1. Can use effects of interest (e.g. difference in proportions) directly, rather than a more complicated test statistic.
2. Can see what "typical" (under the null; under some set of assumptions) samples and outcomes look like; easier to identify oddball behavior.
3. Results in exact – continuous scale – empirical p-values, which are observably linked to sample effect size value (or other quantity of interest, such as a regression coefficient).
4. Can be easily implemented in standard software, or using specialized packages.

Guiding Principles

Simulation is a powerful tool for helping students to understand natural variability of effects (difference in means, difference in proportions, regression coefficients, etc.) under the null. SBI can help students understand where the p-value comes from, what it represents.

Some empirical evidence to show that this approach does improve understanding over the traditional (formula-based) method.

Reduce emphasis on bright-line thinking by reporting the actual p-values, and not just $p < 0.05$ (or any other chosen α).

Guiding Principles

Tie course elements together: the p-value does not (should not) stand in isolation. Interpreting in context means considering

1. study design (including sample size)
2. quality of measurements
3. other studies or other external evidence (theory)
4. what are the model assumptions and are they met?
5. practical significance of effects

These concepts can be taught on their own, and also integrated into the bigger picture of an analysis.

Guiding Principles

Develop intuition by *considering extremes* – what does a p-value of 0.003 really mean? what does a p-value of 0.3 really mean? $p = 0.049$ versus $p = 0.051$.

Develop intuition by *considering instructive cases* – small p-value with scientifically meaningless effect (sample size?); large p-value with scientifically important effect (standalone study? previous studies to support?)

Warn of the dangers of $p < 0.05$.

1. p-hacking
2. multiple tests performed; which results get published?
3. false sense of certainty

The End of p-values in the Intro Course?

NO!

Likewise, traditional formula-based inference needs to continue for now, as students will be expected to have some familiarity with this, published research typically won't be simulation-based and will invoke statistical significance and thresholding.

Students should be aware of significance thresholds and their dangers.

Change is needed, but it will be gradual.

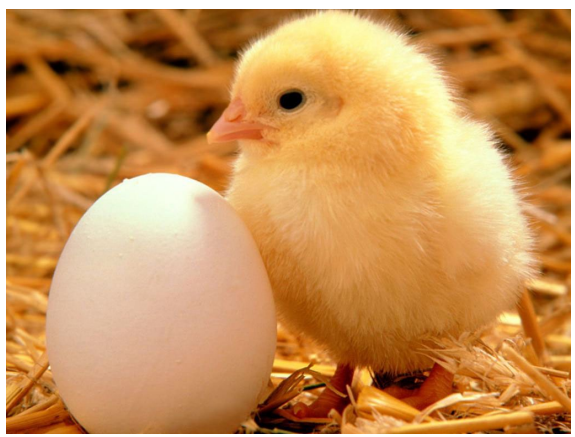
Or, As George Cobb Put It . . .

Q: Why do we teach $p < 0.05$ in our classes?

A: Because that's what the scientific community uses.

Q: Why do so many people still use $p < 0.05$?

A: Because that's what they were taught in their classes.



Acknowledgments

- The "teaching beyond $p < 0.05$" discussion group, UGA Fall 2019 (Catherine Case, Maduranga Dassanayake, Megan Lutz, Lynne Seymour) for helpful tips, examples, suggestions, discussions.
- Students in STAT 4/6230 (Applied Regression), UGA Fall 2019 for gamely attempting to move away from statistical significance and for bravely tackling simulation projects.
- Kari Lock Morgan for helpful discussions on teaching introductory statistics students to think past statistical significance.