Numeracy
Advancing Education in Quantitative Literacy

Volume 15 | Issue 2 Article 8

2022

Talking about Statistical Significance in *Numeracy*

Nathan D. Grawe
Carleton College, ngrawe@carleton.edu
Gizem Karaali
Pomona College, gizem.karaali@pomona.edu

Recommended Citation

Grawe, Nathan D., and Gizem Karaali. "Talking about Statistical Significance in *Numeracy*." *Numeracy* 15, Iss. 2 (2022): Article 8. DOI: https://doi.org/10.5038/1936-4660.15.2.1424

Talking about Statistical Significance in *Numeracy*

Abstract

In recent years, much debate has surrounded the potential for audiences to be mislead by several common practices when reporting statistical significance tests. Two editors of *Numeracy* share the journals perspectives on these questions. As an interdisciplinary journal, we recognize and honor the genre differences represented by our authors and audience members. As a consequence, the journal is open to many practices. Still, we acknowledge the concerns raised by the American Statistical Association and others and encourage authors to write with care and clarity, however results may be represented.

Keywords

quantitative literacy, statistical significance, p values

Cover Page Footnote

Nathan D. Grawe is Professor of Economics at Carleton College and Executive Editor of *Numeracy*.

Gizem Karaali completed her undergraduate studies at Boğaziçi University, Istanbul, Turkey. After receiving her Ph.D. in Mathematics from the University of California Berkeley, she taught at the University of California Santa Barbara for two years. She is currently Professor of Mathematics at Pomona College where she enjoys teaching a wide variety of courses and working with many interesting people. Her scholarly interests include humanistic mathematics, pedagogy, and quantitative literacy, as well as social justice implications of mathematics and mathematics education. She is a Senior Editor of *Numeracy*.

We recently saw an interesting Twitter thread begun by Ethan Mollick that discussed the origins of the widespread use of 0.05 as a *p*-value threshold for statistical significance.[1] In the resulting Twitter discussion, Daniël Lakens pointed to Leahey's (2005) study of papers published in the *American Sociological Review* and the *American Journal of Sociology*, two leading quantitative outlets in the field of sociology.[2] Lackens emphasized Leahey's Figure 1 which plotted, at five-year intervals between 1935 and 2000, the propensity of authors to employ statistical significance tests in general, to apply the 0.05 rule specifically, and to utilize "stars" as a means of calling attention to *p*-values below 0.05, 0.01, and 0.001.

Much of the debate concerned whether the rise of the "rule of 0.05" was best attributable to copyright law or software development. Supporters of the former hypothesis believe that the popularity of the 0.05 alpha value can be attributed to Fisher's (1935) *The Design of Experiments*. Hurlbert and Lombardi (2009) report that Fisher wished to reproduce the chi-squared table found in Pearson's (1914) *Tables for Statisticians and Biometricians*. Having found it difficult to get funding for the publication of statistical tables, Pearson was unwilling to allow a reproduction that would likely reduce royalties from *Tables*—royalties that funded further table publications. According to Hurlbert and Lombardi, because Fisher was unable to reproduce the tables, he was forced to take a more dichotomous approach to statistical inference, focusing on just two alpha values, 0.01 and 0.05. Supporters of this hypothesis note that the share of papers in Leahey's study employing the 0.05 rule rose from 0% in 1935 to almost 60% by 1950; the share employing statistical significance testing in general rose from about 35% to over 80% of papers in the same period.

But not all of the Twitter discussants were persuaded of Fisher's influence on modern practice. They noted that the propensity to use the 0.05 rule (and to engage in statistical significance testing more broadly) declined by almost 20 percentage points between 1950 and 1970. Perhaps Fisher's methodological influence would have waned had technology not intervened in the mid-1970s with the release of SPSS and SAS statistical software. From 1970 to 1985, the share of papers in Leahey's study employing the 0.05 rule rose to 70% while the share reporting statistical significance tests rose to almost 100%. During the same period, the three-star convention, which had been absent from papers in 1970, accounted for nearly half of all papers by 1995. The power of technology! We leave it to the reader to settle this particular debate.

Of course, the methodological role of statistical testing and *p*-values are recent matters of significant scholarly debate. In fact, in 2016 the Board of the American Statistical Association (ASA) released an official statement about *p*-values and their use (Wasserstein 2016). (Wasserstein and Lazar [2016] provide a useful

[1] https://twitter.com/emollick/status/1503061447835799566
[2] https://twitter.com/emollick/status/1503456643190824969

history of the context and process behind the statement.) Mollick's interesting Twitter thread reminded us that other journals have provided guidance to authors about their perspectives on these questions. (See, for example, Editors [2001], Ranstam [2012], and McBee and Matthews [2014].) Because such matters raise additional questions in a community gathered out of many disciplines, we provide here the perspective of *Numeracy* editors.

The principles articulated by the ASA are unassailable as matters of mathematical logic. For example, no one can doubt that the interpretation of any statistical test rests on "a specified statistical model" (Wasserstein 2016, 131). Similarly, sound decision-making cannot follow a dichotomous "rule of 0.05" (or of any other value); the weight of evidence is, after all, a continuous variable.[3] And maybe most importantly, a *p*-value, by construction, says nothing of the practical importance of a measured effect size. These claims are definitionally true and unequivocal.

Difference of opinions arise when thinking about how to implement these principles. For instance, if the weight of evidence for a model is a continuous variable such that *p*-values of 0.049 and 0.051 represent very similar statistical ideas, should we eschew the use of "stars" that call out results which rise to one or another level of statistical significance? Such practice does seem at odds with the incontrovertible principle, and yet one might argue that the reader is perfectly capable of recognizing the relative weight of various results, starred or not.

As *Numeracy* has grappled with these kinds of questions, we routinely note our interdisciplinary subject matter and author/audience base. In recent years we have published authors from mathematics, the natural sciences, social sciences, and the humanities. Each of our disciplines adopt different genre expectations of authors. If we are serious about the multi-disciplinary nature of quantitative literacy (QL), we must be careful not to adopt rules that undermine support for QL across the disciplines—particularly when the work comes from fields less regularly represented in our pages. So, the editors have neither prescribed nor proscribed specific manners of reporting statistical significance testing. We are willing to publish work that reports *p*-values, standard errors, critical values, confidence intervals, and even stars—so long as the text describing the results is appropriate to the content. Whatever the author's choice, the methods of analysis must be clear.

While the journal is open to many different practices concerning statistical significance reporting, we are more directive when it comes to the topic of effect size. Effect size, after all, is nearly always the subject of the studies we review. It seems off the mark to complete a study of the effect of x on y (or at least their correlation) and then fail to unpack the size of the relationship revealed by the data. Given that the weight of the evidence is non-dichotomous, we also welcome

[3] This comic from xkcd is useful in making this point with students: https://xkcd.com/1478/.

discussion of effect sizes when "standard thresholds" for statistical significance are unmet (with appropriate transparency that makes statistical evidence clear).

The multi-disciplinary nature of the QL community can create challenges as we pursue a common goal despite different training and convention. We welcome reviewers and authors to reach out to editors with questions or ideas for how we might better work together despite sometimes meaningful difference.

References

Editors. 2001. "The Value of *P*." *Epidemiology*, 12(3): 286. https://doi.org/10.1097/00001648-200105000-00002

Fisher, Ronald A. 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Hurlbert, Stuart H., and Celia M. Lombardi. 2009. "Final Collapse of the Neyman-Pearson Decision Theoretic Framework and Rise of the NeoFisherian." *Annales Zoologici Fennici*, 46(5): 311–349. https://doi.org/10.5735/086.046.0501

Leahey, Erin. 2005. "Alphas and Asterisks: The Development of Statistical Significance Testing Standards in Sociology." *Social Forces*, 84(1): 1–24. https://doi.org/10.1353/sof.2005.0108

McBee, Matthew T., and Michael S. Matthews. 2014. "Welcoming Quality in Non-Significance and Replication Work, but Moving Beyond the *P*-Value: Announcing New Editorial Policies for Quantitative Research in *JOAA*." *Journal of Advanced Academics*, 25(2): 73–87. https://doi.org/10.1177/1932202X14532177

Pearson, Karl, ed. 1914. *Tables for Statisticians and Biometricians*. Cambridge: The University Press. https://doi.org/10.5962/bhl.title.19414

Ranstam, Jonas. 2012. "Why the *P*-Value Culture Is Bad and Confidence Intervals a Better Alternative." *Oesteoarthritis and Cartilage*, 20(8): 805–808. https://doi.org/10.1016/j.joca.2012.04.001

Wasserstein, Ronald L. 2016. "ASA Statement on Statistical Significance and *P*-Values." *The American Statistician*, 70(2): 129–131. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. "ASA Statement on *P*-Values: Context, Process, and Purpose." *The American Statistician*, 70(2): 131–133. https://doi.org/10.1080/00031305.2016.1154108