From JAMA March 22, 2018: "*Moving the P value threshold from .05 to .005 will shift about one-third of the statistically significant results of past biomedical literature to the category of just 'suggestive.' This shift is essential for those who believe (perhaps crudely) in black and white, significant or nonsignificant categorizations. For the vast majority of past observational research, this recategorization would be welcome*." – John Ioannidis

- That the dichotomization of an information measure as "significant" or "not significant" remains an accepted norm for research reporting should be a point of shame for statistics education (and is increasingly viewed as scandalous):
Imagine if sample sizes were only reported as "small" or "large", without giving the numbers observed.

For the growing number of scientists who want to stop the information loss and distortion from habitual dichotomization of evidence, the push for a lower cutpoint is akin to prescribing more bloodletting to cure weakness caused by bloodletting:
The literature is already gravely weakened by outcome-driven reporting (selection bias, informative censoring), and we are given a straightfaced proposal to fix this by making the the censoring heavier and even more biased!

Aggravating a bias is not a temporary fix, it's a categorical mistake. If we are to stop statistical distortion of scientific information, we must stop allowing study outcomes to determine reporting (not just whether an outcome is reported, but how much it calls attention to itself in the broader literature and press, including claims of "significance").

As for the form of report, the challenge is to replace crude categorizations with continuous measures of information that end users can appreciate intuitively better than significance tests (for which just about all common intuitions are wrong).

It seems meeting this challenge is mechanically simple.
Many renown writers have recommended such continuous reporting of *P* for decades (e.g., Lehmann's "Testing Statistical Hypotheses", D.R. Cox's writings).

And, it takes little effort and less space to simply report "*P*=..." instead of "significant" or "not significant".

So I'm led to ask:

1) Why has the sagely and easily implemented advice of so many authorities failed to staunch the routine and pointless degradation of test information into an arbitrary dichotomy?

Then there's the problem of null-bias being promoted as if it were an objective state, when it's just reflective of a loss function that is often not be shared by all stakeholders.

Which leads me to ask:

2) Why has statistics failed to advise and implement routine testing of alternatives to the null? The post-data information analog of pre-data power is the P-value for the same alternative at which power was calculated (using the same sidedness as the null test), not a post-data "power" calculation (confidence intervals provide only a dichotomized measure of the information against alternatives).

There are very plausible answers to (1) and (2) that come down to human demands for certainty; these demands are not well-matched to the extensive uncertainty that is ignored by formal statistical procedures. Thus the barriers to reform are not mechanical or theoretical; they are psychological and thus much harder to recognize and overcome (and can't be overcome without being recognized).

Finally, given that only a small minority of users can come close to defining let alone using a P-value correctly, I have also come to ask:

3) Why has statistics insisted on measuring evidence on such an extremely nonlinear scale as that of P-values, instead of a scale in which independent bits of information would simply add and would not invite naive inversions into posterior probabilities?

I suggest that the focus on P instead of information reflects a tradition of treating probability theory as more basic and intuitive than information theory. I suspect this tradition had seen the best of its days by 1950 and is now long overdue for retirement.

By 1957, Good had suggested rescaling P to $S = -\log(P)$, which is the *surprisal*, *self-information*, *logworth*, or *S*-value for the tail-event prediction called the *P*-value.

With *S* taken in base 2 logs and explained as bits of information against the tested hypothesis or model, might be better understood than the *P*-value; for example, *S* can be equated to a simple, additive event count.

For further lamentations (and advice, sagely or not) see Greenland, S. (2017), "The need for cognitive science in methodology," American Journal of Epidemiology, 186, 639–645. doi: 10.1093/aje/kwx259
open access at https://academic.oup.com/aje/article/186/6/639/3886035
see also
Amrhein, V., Korner-Nievergelt, F., and Roth, T. (2017), "The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research," PeerJ, 5, e3544. doi: 10.7717/peerj.3544.
Amrhein, V., and Greenland, S. (2018), "Remove, rather than redefine, statistical significance," Nature Human Behaviour, 2, 4.
McShane et al. (2017). Abandon Statistical Significance. https://arxiv.org/abs/1709.07588

-Sander Greenland, Department of Epidemiology and Department of Statistics, UCLA