Fit-for-Purpose Inferential Methods: Abandoning/Changing P-values Versus Abandoning/Changing Research

John P.A. IOANNIDIS

P-values continue to be widely used and misused, but until now there has been a lack of consensus in the scientific community about how grave the misuse has been, how serious the consequences are, and how exactly we should proceed to remedy the situation. Many competing options exist to change the paradigm. While the very best statisticians and methodologists do not agree on the optimal future agenda, the current status quo is perpetuated, often with prominent misconceptions that we all recognize as highly problematic. At a minimum, I hope that the positions of the ASA Statement, which reflect a high (even if not perfect) level of consensus, may offer a solid launching ground for further remedial efforts.

Currently some P-values are reported in the majority of papers that perform any statistical analysis in any empirical data (Chavalarias, Wallach, Li, and Ioannidis 2016). Conversely reporting of effect sizes is less frequent and reporting of measures of uncertainty, such as confidence intervals, is far less frequent. The use/reporting of Bayesian statistics or false-discovery rate approaches remains overall exceedingly uncommon across the published literature (Chavalarias, Wallach, Li, and Ioannidis 2016). There are certainly exceptions to this overall average pattern. Bayesian statistics (Goodman1999) and false-discovery rate (Benjamini and Hochberg 1995) may be common in specific research discipline islands. Different fields of scientific inquiry are accustomed to using different inferential tools, but unfortunately this is driven mostly by tradition and by mimicking behavior, rather than careful thinking about fit-for-purpose. Different fields also differ a lot in their accepted rules of claiming success and in their silently agreed expectations, whenever they opt to use P-values. Most fields in biomedicine and social sciences are accustomed to spurious thresholds of $P < 0.05$ for making automated claims in their inferential machinery (Gigerenzer 2004). This leads to inferential blunders, especially in an era of low prior odds for a nonnull effect, highly exploratory analyses and hidden multiplicity coupled with selective reporting (Ioannidis 2005). Other fields, e.g. genomics or experimental particle physics, typically use far more stringent P-value thresholds.

The ASA Statement may offer a sound basis on contemplating how to improve the use of statistical inferences in each field and how to forgo long-established practices in favor of others that are better suited to what each scientific field aims to achieve. The best recipe is unlikely to be the same in all scientific disciplines and it is unlikely that there will be only one optimal recipe in each discipline. But some current practices immediately seem to be grossly misleading. It is fine to correct those misleading practices, regardless of what exactly they are replaced with, among several reasonable alternatives.

For example, many observational research fields such as large segments of nutritional epidemiology, electronic health record-based investigations using routinely collected data or other big data compilations, can be described as high-output machines producing copious P-value trash. With a combination of large datasets, confounding, flexibility in analytical choices (Patel, Burford, and Ioannidis 2015), and superimposed selective reporting bias, using a $P < 0.05$ value threshold to declare "success," any result can become statistically significant, but this means next to nothing. As Jeffreys put it over half a century ago: "A null hypothesis is set up and 'tested' against data: It is merely something set up like a coconut to stand until it is hit" (Jeffreys 1961) Many scientific fields are accustomed to taking an endless number of shots until they (unfortunately) hit the coconut.

Raising the bar to more stringent P-value thresholds may reduce some of that trash, but does not attack the root of the problem, and of course it may generate also some false-negatives (Ioannidis, Tarone, and McLaughlin 2011) One has to look calmly at the main principles. Does a null-hypothesis even make sense to test? Perhaps in several of the current P-value-chasing investigations nobody should really have cared about rejecting the null-hypothesis. In fact, is there any possibility that a null-hypothesis can avoid being rejected, if one can assemble a larger and larger sample size, in a setting where confounding and bias are impossible to eradicate to the point that whatever signals can be separated from the noise? Why test against the null when the null is impossible and/or meaningless? The principles of the Statement should lead to some thought before running any statistical analysis.

Sometimes, P-values should be avoided and other methods should be used instead, or simply descriptive metrics might suffice. Other times, it is doing the research that should be avoided, if the results are likely to be misleading, regardless of the inferential methods used. Some of the most prolific fields of current research (in terms of publication volume) are practically not contributing knowledge, but just expressing repeatedly how big bias can be in their domain. Then it is not an issue of abandoning P-values, it is an issue of abandoning poor research. Misleading use of P-values is so easy and automated that, especially when rewarded with publication and funding, it can become addictive. Investigators generating these torrents of P-values should be seen with sympathy as drug addicts in need of rehabilitation that will help them live a better, more meaningful scientific life in the future.

In many other fields, inferences using P-values will continue

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. John P.A. Ioannidis, C.F. Rehnborg Professor in Disease Prevention, Professor of Medicine, of Health Research and Policy, and of Statistics, Director of the Stanford Prevention Research Center, and Co-Director of the Meta-Research Innovation Center at Stanford (METRICS), Stanford University, 1265 Welch Rd, MSOB X306, Stanford, CA 94305 (Email: jioannid@stanford.edu).

to offer helpful insights, if properly used and interpreted. Other inferential methods may need to be used more frequently. In some fields, their use is overdue. For example in clinical trials and their meta-analyses, presenting effect sizes and their uncertainty should be the default and P-values can be nicely complemented, if not largely replaced, by Bayesian inferences. Using alternative inferential tools will still not solve, nevertheless, some of the problems that cause many misleading claims in this literature, in particular those related to hidden multiplicity and selective reporting biases (Dwan et al. 2013). If success is defined based on passing some magic threshold, biases may continue to exert their influence regardless of whether the threshold is defined by a P-value, Bayes factor, false-discovery rate, or anything else. Efforts to promote transparency in study design, conduct and reporting may have more to offer in this setting than blaming P-values. Studying how these efforts can be most successful is an entire field of research on its own (Ioannidis, Fanelli, Dunne, and Goodman 2015).

References

Chavalarias, D., Wallach, J., Li, A., and Ioannidis J.P. (2016), "Evolution of Reporting of P-values in the Biomedical Literature, 1990–2015," *Journal of the Amewrican Medical Association*, in press.

Goodman, S.N. (1999), "Toward Evidence-Based Medical Statistics. 2: the Bayes Factor," *Annals of Internal Medicine*, 130, 1005–1013.

Benjamini, Y., and Hochberg. Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300.

Gigerenzer, G. (2004), "Mindless Statistics," *Journal of Socioeconomics*, 33, 567–606.

Ioannidis, J.P. (2005), "Why Most Published Research Findings are False," *PLoS Medicine*, 2:e124.

Patel, C.J., Burford, B., and Ioannidis, J.P. (2015), "Assessment of Vibration of Effects Due to Model Specification can Demonstrate the Instability of Observational Associations," *Journal of Clinical Epidemiology*, 68, 1046–1058.

Jeffreys, H. (1961), *Theory of Probability*, Oxford, Oxford University Press.

Ioannidis, J.P., Tarone, R., and McLaughlin, J.K. (2011), "The False-Positive to False-Negative Ratio in Epidemiologic Studies," *Epidemiology*, 22, 450–456.

Dwan, K., Gamble, C., Williamson, P.R., Kirkham, J.J.; Reporting Bias Group (2013), "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias—An Updated Review," *PLoS One*, 8:e66844.

Ioannidis, J.P., Fanelli, D., Dunne, D.D., and Goodman, S.N. (2015), "Meta-research: Evaluation and Improvement of Research Methods and Practices," *PLoS Biology*, 13:e1002264.