Disengaging from Statistical Significance

Kenneth J. ROTHMAN

In the marketplace of scientific results, the preferred currency by which results have been valued has been statistical significance, expressed either as a dichotomous label or by the underlying p-value, which may be given as a number or an inequality. Like other modern currencies, the value of this one is not inherent but derived from widely held assumptions and expectations. Indeed, reliance on statistical significance is as misplaced as faith in some dubious paper monies. At the risk of stretching the analogy, I suggest that a version of Gresham's Law has been operating, allowing statistical significance to force out of circulation better ways to analyze data, and leaving us with results that are, all too often, astonishingly misleading.

As the ASA statement (ASA 2016) indicates, a fundamental flaw of relying on statistical significance for inference is the need to dichotomize all results into those that are significant or not significant. This practice degrades vast efforts to collect and analyze quantitative data into a mere label. Furthermore, if ever there were a false dichotomy, it is the dichotomy between significant and not. The label is assigned by an arbitrary rule and inevitably has less information than the p-value from which it derives. Moreover, the p-value itself is a handicapped approach to interpretation because it doesn't measure effect size. Instead, it blends together information on estimated effect size and the precision of that estimate (Lang, Rothman, and Cann 1998). Although p-values and confidence intervals are closely related, a confidence interval, in contrast to a p-value, expresses separately both effect size and precision (Poole 2001). This advantage of confidence intervals illustrates that it usually takes two numbers to measure two distinct characteristics. Unfortunately, all too often we have seen the reported confidence interval used merely to determine if the null value lies within it or not, debasing the confidence interval into a label, a surrogate significance test (Cumming 2012).

The correspondence between results that are statistically significant and those that are truly important is far too low to be useful. Consequently, scientists 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. These are pernicious problems, and not just in the metaphorical sense. 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 (Hauer 2004; Schmidt and Rothman 2014).

How do we fix this problem? The reliance on statistical significance testing is ingrained in the social system of many sciences, and therefore reflexive on the part of many members of those social systems, making it difficult to counter. Nonetheless, we can and should advise today's students of statistics that they should avoid statistical significance testing, and embrace estimation instead. Those who have tried offering this advice know it can be challenging. Students all too often fear that their success will be measured by publications and grants that are evaluated by reviewers who esteem statistical significance. Despite such inertia, in epidemiology there has been an encouraging trend toward reporting confidence intervals to supplement or even supplant statistical significance and p-values, toward using confidence intervals to measure effect size and to gauge precision rather than to test null hypotheses, and toward avoiding the fallacy of considering every statistically non-significant result as if it were evidence for a null relation.

Real change will take the concerted effort of experts to enlighten working scientists, journalists, editors and the public at large that statistical significance has been a harmful concept, and that the estimates of meaningful effect measures is a much more fruitful research aim than the testing of null hypotheses. This statement of the ASA does not go nearly far enough toward that end, but it is a welcome start and a hopeful sign.

References

American Statistical Association (ASA) (2016), "ASA Statement on Statistical Significance and P-values," *The American Statistician*, 70, DOI: 10.1080/00031305.2016.1154108.

Cumming, G. (2012), *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis,* New York: Routledge.

Hauer, E. (2004), "The Harm Done by Tests of Significance," *Accident Analysis and Prevention*, 36, 495–500.

Lang, J., Rothman, K. J., and Cann, C. I. (1998), "That Confounded P-Value," *Epidemiology*, 9, 7–8.

Poole, C. (2001), "Low P-values or Narrow Confidence Intervals: Which are More Durable?" *Epidemiology*, 12, 291–294.

Schmidt, M., and Rothman, K. J. (2014), "Mistaken Inference Caused by Reliance on and Misinterpretation of a Significance Test," *International Journal of Cardiology*, 177, 1089–1090.

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Kenneth J. Rothman, RTI International–Health Solutions, 3040 East Cornwallis Road, Research Triangle Park, NC 27709-2194 (Email: krothman@rti.org).