It's Not the *P*-Values' Fault

Yoav BENJAMINI

I argue that ASA board statement about the *p*-values may be read as discouraging the use of *p*-values because they can be misused, while the other approaches offered there might be misused in much the same way. In particular, ignoring the effect of selection on statistical inferences is common yet potentially very harmful to the replicability of research results.

KEY WORDS: ASA board; Industrialized science; Selective inference.

When I was invited to participate in ASA committee, my initial response was that it would be better for the committee to draft a statement about the appropriate use of statistical tools for addressing the crisis of reproducibility and replicability (R&R) in science. Unfortunately, in response to outcries about the role of Statistics, which focused on the perceived role of the widely used *p*-values, the ASA board fell into the trap of formulating a statement about the *p*-values. The well-phrased statement demonstrates our mistake in singling out the *p*-value: posing the *p*-value as a culprit, rather than the way most statistical tools are used in the new world of industrialized science.

Admittedly, most statisticians reading this statement will agree with most of its principles (Bayesians may not agree to principle 1, frequentists will have difficulties understanding principle 6), but all principles stated are only about *p*-values and statistical significance. The result is a statement that will be read by our target audience as expressing very negative ASA attitude towards the *p*-value. As stated, the *p*-value "can be useful" providing "one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data" (Principle 1).

On the other hand:

Principle 2: "*p-values do not measure the probability*. . . ;

Principle 3: "scientific decisions *should not be based only on whether a p-value passes a specific threshold*" as this "*. . . leads to considerable distortion of the scientific process*";

Principle 4: "*P-values and related analyses should not be reported selectively*";

Principle 5: "*A p-value, or statistical significance, does not measure the size of an effect or the importance of a result*"

Principle 6: " . . . a *p*-value near 0.05 taken by itself offers only weak evidence against the null hypothesis"

Nonstatistical scientists, editors, policy makers or a judges, who read these principles will conclude that the *p*-value is indeed a very risky statistical tool, as advertised by its opponents. Avoiding its use and discouraging its use by others is just a matter of common sense. This will be the case especially since the ASA statement offers *Other Approaches: "In view of the prevalent misuses of and misconceptions concerning p-values, some statisticians prefer to supplement or even replace p-values with other approaches.*"

Yet all of these other approaches, as well as most statistical tools, may suffer from many of the same problems as the *p*-values do. What level of likelihood ratio in favor of the research hypothesis will be acceptable to the journal? Should scientific discoveries be based on whether posterior odds pass a specific threshold (P3)? Does either measure the size of an effect (P5)? Isn't our best effect size estimator useless as a single measure if not supported by a statement about its uncertainty? How can we decide about the sample size needed for a clinical trial—however analyzed—if we do not set a specific bright-line decision rule? Finally, 95% confidence intervals or credence intervals (both sharing the limitations in P2) offer no protection against selection when only those that do not cover 0, are selected into the abstract (P4).

What made the *p*-value so useful and successful in Science throughout the 20th century, despite of the misconceptions so well described in the statement? In some sense it offers a first line of defense against being fooled by randomness, separating signal from noise, because the models it requires are simpler than any other statistical tool needs. Likelihood ratios, effect size estimates, confidence intervals, and Bayesian methods all rely on assumed models over a wider range of situations, not merely under the tested null; Bayesian tools need further modeling, in the form of priors and hierarchical structures. Most important, the model needed to calculate of the *p*-value can be guaranteed to hold under appropriately designed and executed randomized experiments.

The *p*-value is a very valuable tool, but when possible it should be complemented—not replaced—by confidence intervals and effect size estimates. The end of a 95% confidence interval that extends towards 0 indicates by how much the difference can be separated from 0 (in a statistically significant way at level 5%. . .). The mean difference, when supported by an assessment of uncertainty is again useful. Disappointingly, in some areas of science these methods are grossly underutilized.

Sometimes, especially when using emerging new scientific technologies, the *p*-value is the only way to quantify uncertainty, and can be mapped and compared across conditions (e.g.

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Yoav Benjamini, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel (Email: ybenja@post.tau.ac.il). The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° [294519] PSARPS. The analysis of the papers of the Reproducibility Project was led by Yoav Zeevi, with Sofi Estashenko, Tal Galili, Tal Kozlovski, and Dan Warshavski (see replicability.tau.ac.il). Thanks to Henry Braun, Abba Krieger, and Philip Stark for comments.

functional MRI, Gene Expression, Genome Wise Association Studies). It is recognized that merely "full reporting and transparency" (Principle 4) is not enough, as selection is unavoidable in these large problems. Selection takes many forms: selection by a table, selection into the abstract, selection by highlighting in the discussion, selection into a model, or selection by a figure. Further statistical methods must be used to address the impact of selective inference, otherwise the properties each method has on the average for a single parameter (level, coverage or unbiasedness) will not hold even on the average over the selected parameters. Therefore, in those same areas, the p-value bright-line is not set at the traditional 5% level. Methods for adaptively setting it to directly control a variety of false discovery rates or other error rates are commonly used. More generally, addressing the effect of selection on inference has been a very active research area, resulting in new strategies and sophisticated tools for testing, confidence intervals, and effect-size estimates, in different setups. It deserves a separate review.

The transition in large complex problems illustrates the process occurring throughout science: the industrialization of the scientific process at the turn of the century. Experimentation is done by high throughput industrial processes and their outcomes are analyzed automatically, resulting in a large number of inferences to select from. With the availability of ever-larger databases and the ease of computations, other areas of science are undergoing similar industrialization processes, yet are slow to realize these changes. For example, the estimated number of *reported* inferences in the 100 studies included in the "reproducibility project" in Experimental Psychology (Open Science Collaboration, 2015) range from 5 to 730, with an average of 77 ($\pm$ 10) per study. We currently study the actual selection process in these complex studies (rather than merely counting) but it is enough to note that only 11 studies included any partial effort to address selection. Facing such ignorance I prefer to eyeball a set of p-values to assess the effect of selection rather than view a set of confidence intervals.

In summary, when discussing the impact of statistical practices on R&R, the p-value should not be singled out nor its use discouraged: its more likely the fault of selection, not the p-values'.

Reference

Open Science Collaboration (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, aac4716. DOI: 10.1126/science.aac4716.