Statistics and Unintended Consequences

Many years ago, Allan Wilks spoke about the experiences he and Richard Becker and John Chambers (co-developers of *S*, the progenitor of *S-Plus* and R) encountered among users. One of his remarks has remained with me all these years. He was surprised at the ways *S* was being used, ways they never imagined. "For example, one person called to say that *S* was incredibly slow. All he wanted was an identity matrix and it took a half hour. I was puzzled; the command `diag(1000)` takes a fraction of a second. It turns out that he was creating the matrix with for loops: `for (i in 1:1000) {for (j in 1:1000) {if (i == j) then A[i,j]=1 else A[i,j]=0}}`. It never occurred to us that people would use our package in this way."

Recently, at chapter meetings, conferences, and other events, I've had the good fortune to meet many of our members, many of whom feel queasy about the effects of differing views on *p*-values expressed in the March 2019 supplement of *The American Statistician (TAS)*. The guest editors—Ronald Wasserstein, Allen Schirm, and Nicole Lazar—introduced the ASA Statement on *P*-Values (2016) by stating the obvious: "Let us be clear. Nothing in the ASA statement is new." Indeed, the six principles are well-known to statisticians. The guest editors continued, "We hoped that a statement from the world's largest professional association of statisticians would open a fresh discussion and draw renewed and vigorous attention to changing the practice of science with regards to the use of statistical inference."

The authors of the March 2019 supplement of *TAS* offered change. Yet, as the editors noted, "The voices in the 43 papers in this issue do not sing as one. … To us, these are all the sounds of statistical inference in the 21st century, the sounds of a world learning to venture beyond $p < 0.05$."

A healthy debate about statistical approaches can lead to better methods. But, just as Wilks and his colleagues discovered, unintended consequences may have arisen: Nonstatisticians (the target of the issue) may be confused about what to do. Worse, "by breaking free from the bonds of statistical significance" as the editors suggest and several authors urge, researchers may read the call to "abandon statistical significance" as "abandon statistical methods altogether."

We agree with the editors' hope that "statistics in science and policy will become more significant than ever." Since this recent *TAS* supplement appeared, its guest editors have been busy traveling around the country and fielding phone calls to discuss and clarify the issues with *p*-values, the term "statistical significance," and "alternatives to *p*-values."

But we may need more. How exactly are researchers supposed to implement this "new concept" of statistical thinking? Without specifics, questions such as "Why is getting rid of *p*-values so hard?" may lead some of our scientific colleagues to hear the message as, "Abandon *p*-values"—despite the guest editors' statement: "We are not recommending that the calculation and use of continuous *p*-values be discontinued."

Karen Kafadar

Brad Efron once said, "Those who ignore statistics are condemned to re-invent it." In his commentary ("It's not the *p*-value's fault") following the 2016 ASA Statement on *P*-Values, Yoav Benjamini wrote, "The 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." Indeed, *p*-values (and all statistical methods in general) can be misused. (So may cars and computers and cell phones and alcohol. Even words in the English language get misused!) But banishing them will not prevent misuse; analysts will simply find other ways to document a point—perhaps better ways, but perhaps less reliable ones. And, as Benjamini further writes, *p*-values have stood the test of time in part because they offer "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"—especially now that Efron's bootstrap has become a familiar tool in all branches of science for characterizing uncertainty in statistical estimates.

Conceptually, likelihood ratios (LRs) and hierarchical Bayes models and probability distributions (on which LRs and Bayesian models are based) are useful additions to *p*-values. But they have uncertainty, too. Moreover, try explaining those statistical concepts to nonstatisticians. (I've tried. And so have we all when we work with nonquantitative scientists. The bootstrap is a lot easier to explain.) Our challenge continues to be to effectively explain these concepts to nonstatisticians.

In the March 2019 *TAS* supplement, Ronald Fricker and his colleagues looked at 31 articles published in a 2016 issue of *Basic & Applied Social*

Psychology (BASP) one year after its editors banned the use of inferential statistics. "We found multiple instances of authors overstating conclusions beyond what the data would support if statistical significance had been considered. Readers would be largely unable to recognize this because the necessary information to do so was not readily available." They conclude, "In our opinion, the practices we have observed in the papers published in *BASP* post-ban will not help to solve this problem [proper inference]; in fact, we believe they will make it worse." Fricker et al. also recall the recommendations of the American Psychological Association's Task Force on Statistical Inference (1999), which included Donald Rubin, Frederick Mosteller, and John Tukey: "Some had hoped that this task force would vote to recommend an outright ban on the use of significance tests in psychology journals. Although this might eliminate some abuses, the committee thought that there were enough counterexamples … to justify forbearance."

Where will moving to a world beyond $p < 0.05$ take us? Will "statistics in science and policy become more significant than ever" as the *TAS* authors propose? Or will it lead to more confusion, less interpretable studies, and more associations claimed to be important, but maybe no more than one would expect from having calculated thousands of Pearson correlation coefficients? If other journals cite peer-reviewed publications in ASA journals as justification for revising their editorial policies to banish *p*-values, the core of our profession will be threatened, and we may not see "statistics in science and policy become more significant than ever."

It is reassuring that "*Nature* is not seeking to change how it considers statistical evaluation of papers at this time," but this line is buried in its March 20 editorial, titled "It's Time to Talk About Ditching Statistical Significance." Which sentence do you think will be more memorable? We can wait to see if other journals follow *BASP*'s lead and then respond. But then we're back to "reactive" versus "proactive" mode (see February's column), which is how we got here in the first place.

Indeed, the ASA has a professional responsibility to ensure good science is conducted—and statistical inference is an essential part of good science. Given the confusion in the scientific community (to which the ASA's peer-reviewed 2019 *TAS* supplement may have unintentionally contributed), we cannot afford to sit back. After all, that's what started us down the "abuse of *p*-values" path. (See the April column.)

In an unpublished manuscript that he kindly shared with me while I was preparing this column, Stephen Stigler suggests "A Novel Solution to the 'Crisis' in Significance Testing: Read Fisher!" Quoting from Fisher's classic, *The Design of Experiments*:

> In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result."

Stigler concludes, "It is clear that Fisher would not have considered a different threshold, even one as small as 0.005, as a solution to a problem. It is also clear that Fisher was an ardent advocate of reproducible science." And that—reproducibility—is the real heart of the problem. (See the recently released report from the National Academy of Science, *Reproducibility and Replication in Science*.) As Benjamini said, "It's not the *p*-value's fault."

Tukey wrote years ago about Bayesian methods: "It is relatively clear that discarding Bayesian techniques would be a real mistake; trying to use them everywhere, however, would in my judgment, be a considerably greater mistake." In the present context, perhaps he might have said: "It is relatively clear that trusting or dismissing results based on a single *p*-value would be a real mistake; discarding *p*-values entirely, however, would in my judgment, be a considerably greater mistake."

We should take responsibility for the situation in which we find ourselves today (and during the past decades) to ensure that our well-researched and theoretically sound statistical methodology is neither abused nor dismissed categorically. I welcome your suggestions for how we can communicate the importance of statistical inference and the proper interpretation of *p*-values to our scientific partners and science journal editors in a way they will understand and appreciate and can use with confidence and comfort—before they change their policies and abandon statistics altogether. Please send me your ideas!