Response to the ASA's statement on p-values: context, process, and purpose

Edward L. Ionides
Department of Statistics, University of Michigan
and
Alexander Giessing
Department of Statistics, University of Michigan
and
Yaacov Ritov
Department of Statistics, University of Michigan
and
Scott E. Page
Departments of Complex Systems, Political Science
and Economics, University of Michigan

Version of a letter to appear in *The American Statistician*.
Submitted 16 March 2016, accepted 22 August 2016.

The ASA's statement on p-values: context, process, and purpose (Wasserstein & Lazar 2016) makes several reasonable practical points on the use of p-values in empirical scientific inquiry. The statement then goes beyond this mandate, and in opposition to mainstream views on the foundations of scientific reasoning, to advocate that researchers should move away from the practice of frequentist statistical inference and deductive science. Mixed with the sensible advice on how to use p-values comes a message that is being interpreted across academia, the business world, and policy communities, as, "Avoid p-values. They don't tell you what you want to know." We support the idea of an activist ASA that reminds the statistical community of the proper use of statistical tools. However, any tool that is as widely used as the p-value will also often be misused and misinterpreted. The ASA's statement, while warning statistical practitioners against these abuses, simultaneously warns practitioners away from legitimate use of the frequentist approach to statistical inference.

In particular, the ASA's statement ends by suggesting that other approaches, such as Bayesian inference and Bayes factors, should be used to solve the problems of using and interpreting p-values. Many committed advocates of the Bayesian paradigm were involved in writing the ASA's statement, so perhaps this conclusion should not surprise the alert reader. Other applied statisticians feel that adding priors to the model often does more to obfuscate the challenges of data analysis than to solve them. It is formally true that difficulties in carrying out frequentist inference can be avoided by following the Bayesian paradigm, since the challenges of properly assessing and interpreting the size and power for a statistical procedure disappear if one does not attempt to calculate them. However, avoiding frequentist inference is not a constructive approach to carrying out better frequentist inference.

On closer inspection, the key issue is a fundamental position of the ASA's statement on the scientific method, related to but formally distinct from the differences between Bayesian and frequentist inference. Let's focus on a critical paragraph from the ASA's statement: "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. These include methods that emphasize estimation over testing, such as confidence, credibility, or pre-

diction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discovery rates. All these measures and approaches rely on further assumptions, but they may more directly address the size of an effect (and its associated uncertainty) or whether the hypothesis is correct."

Some people may want to think about whether it makes scientific sense to "directly address whether the hypothesis is correct." Some people may have already concluded that usually it does not, and be surprised that a statement on hypothesis testing that is at odds with mainstream scientific thought is apparently being advocated by the ASA leadership. Albert Einstein's views on the scientific method are paraphrased by the assertion that, "No amount of experimentation can ever prove me right; a single experiment can prove me wrong" (Calaprice 2005). This approach to the logic of scientific progress, that data can serve to falsify scientific hypotheses but not to demonstrate their truth, was developed by Popper (1959) and has broad acceptance within the scientific community. In the words of Popper (1963), "It is easy to obtain confirmations, or verifications, for nearly every theory," while, "Every genuine test of a theory is an attempt to falsify it, or to refute it. Testability is falsifiability." The ASA's statement appears to be contradicting the scientific method described by Einstein and Popper. In case the interpretation of this paragraph is unclear, the position of the ASA's statement is clarified in their Principle 2: "P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone. Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither." Here, the ASA's statement misleads through omission: a more accurate end of the paragraph would read, "The p-value is neither. Nor is any other statistical test used as part of a deductive argument." It is implicit in the way the authors have stated this principle that they believe alternative scientific methods may be appropriate to assess more directly the truth of the null hypothesis. Many readers will infer the ASA to imply the inferiority of deductive frequentist methods for scientific reasoning. The ASA statement, in its current form, will therefore make it harder for scientists to defend a choice of frequentist statistical methods

during peer review. Frequentist papers will become more difficult to publish, which will create a cascade of effects on data collection, research design, and even research agendas.

Goodman (2016) provided an example of how the scientific community is interpreting the ASA's statement. Goodman (2016) noted that the title of the ASA's statement is "deceptively innocuous," and then proceeded to paraphrase the ASA's statement in support of inductive over deductive scientific reasoning: "What scientists want is a measure of the credibility of their conclusions, based on observed data. The P value neither measures that nor is part of a formula that provides it."

Gelman & Shalizi (2013) wrote a relevant discussion of the distinction between deductive reasoning (based on deducing conclusions from a hypothesis and checking whether they can be falsified, permitting data to argue against a scientific hypothesis but not directly for it) and inductive reasoning (which permits generalization, and therefore allows data to provide direct evidence for the truth of a scientific hypothesis). It is held widely, though less than universally, that only deductive reasoning is appropriate for generating scientific knowledge. Usually, frequentist statistical analysis is associated with deductive reasoning and Bayesian analysis is associated with inductive reasoning. Gelman & Shalizi (2013) argued that it is possible to use Bayesian analysis to support deductive reasoning, though that is not currently the mainstream approach in the Bayesian community. Bayesian deductive reasoning may involve, for example, refusing to use Bayes factors to support scientific conclusions. The Bayesian deductive methodology proposed by Gelman & Shalizi (2013) is a close cousin to frequentist reasoning, and in particular emphasizes the use of Bayesian p-values.

The ASA probably did not intend to make a philosophical statement on the possibility of acquiring scientific knowledge by inductive reasoning. However, it ended up doing so, by making repeated assertions implying, directly and indirectly, the legitimacy and desirability of using data to directly assess the correctness of a hypothesis. This philosophical aspect of the ASA statement is far from irrelevant for statistical practice, since the ASA position encourages the use of statistical arguments that might be considered inappropriate.

A judgment against the validity of inductive reasoning for generating scientific knowledge does not rule out its utility for other purposes. For example, the demonstrated utility

of standard inductive Bayesian reasoning for some engineering applications is outside the scope of our current discussion. This amounts to the distinction Popper (1959) made between "common sense knowledge" and "scientific knowledge."

References

Calaprice, A. (2005), *The new quotable Einstein*, Princeton University Press, Princeton, NJ.

Gelman, A. & Shalizi, C. R. (2013), 'Philosophy and the practice of Bayesian statistics', *British Journal of Mathematical and Statistical Psychology* **66**(1), 8–38.

Goodman, S. N. (2016), 'Aligning statistical and scientific reasoning', *Science* **352**(6290), 1180–1181.

Popper, K. R. (1959), *The logic of scientific discovery*, Hutchinson, London.

Popper, K. R. (1963), *Conjectures and refutations: The growth of scientific knowledge*, Routledge and Kegan Paul, New York, NY.

Wasserstein, R. L. & Lazar, N. A. (2016), 'The ASA's statement on p-values: context, process, and purpose', *The American Statistician,* pre-published online.