The Significance of the ASA Statement on Statistical Significance and P-Values

Stephen T. ZILIAK

Little p-value
What are you trying to say
Of significance?

Far more in science and society than it positively should. That is one way to express the point of the American Statistical Association Statement on Statistical Significance and P-Values (ASA 2016). The cost of incorrect interpretation of Student's t-statistics, Fisher's p-values, and other tests of statistical significance has been shown to be unsustainably large (for a large scale survey see Ziliak and McCloskey (2008)). The ASA Statement on Statistical Significance and P-values is going to help.

Content aside, it is worth noting that the Statement emerged from a humane, Socratic, and Tocquevillean model of democracy and dialectic. (I'm an economist so I know the difference.) The Statement has seen more discerning eyes than a model on a catwalk. Statistician eyes, scientist eyes, journalist and business and Bayesian eyes, over and over. Still it stands, I think most will agree. Hats off to Ron Wasserstein and the ASA Board, who made openness and transparency, widespread democracy and dialectic the chief virtues of the drafting process. Over the course of the past year the Statement on Statistical Significance has evolved with the benefit of constant counsel from leading statisticians and scientists worldwide. Thus the Statement can be treated as a repeated and unbiased sample of best practice thinking about statistical significance (plus or minus a small error).

Despite occasional disagreements—some of them fundamental to the philosophy of science—the drafting Committee did what pundits and skeptics alike thought impossible. Together we agreed that the current culture of statistical significance testing, interpretation, and reporting has to go, and that adherence to a minimum of six principles can help to pave the way forward for science and society. Adherence to principles (2) through (6) will be productive, most of us believe, of a steady and rising stream of large net benefits to more than science. In economic policy. In health and drugs and medicine. And in every realm of life, from agronomy to zoology, including law, that is touched by the test of statistical significance.

Some, for example financial investors and publishing scientists, could begin immediately to reap the benefits of change implied by this Statement. There is a hunger for change among journal editors and referees; among grantors and journalists, lawyers, and decision-makers. Virtually no one is happy with mushy p's though they, as I and others have shown, are treated like the main dish of science. In abbreviated form I believe the most important principles for the much-anticipated paradigm shift are:

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.*"

The null hypothesis stipulating no numerical or practical difference between object A and object B may be true even when the p-value is low, below 0.05, for example. And an alternative hypothesis and effect size $A > B$ may be for legal or medical or commercial purposes important even when the p-value takes on higher values. The p-value is not an error-probability. And it doesn't measure the probability of a hypothesis given the evidence. Not even when searching for the Higgs boson or Einstein's ripples, through a lot of evidence. The ripples and boson probably exist. But the p-value or other test of statistical significance does not prove the role of random chance.

3. "*Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*"

Thus bright-line rules of acceptance and rejection, such as the false equations of $p > 0.05 =$ "insignificant, accept the null" and $p < 0.05 =$ "significant, reject the null" should be in most cases banished. (Many Presidents of the American Statistical Association have long argued such, from Kruskal and Zellner to Morganstein and Utts.) Recently the Supreme Court of the United States unanimously agreed in Matrixx v. Siracusano, 9-0, that statistical significance is neither necessary nor sufficient (Bialik 2011). Statistical significance does not mean scientific or business or policy "importance." And lack of statistical significance according to an arbitrary line and custom does not mean lack of importance. As Savage (1954, p. 116) noted long ago, "The cost of an observation in utility may be negative as well as zero or positive; witness the cook that tastes the broth"

4. "*Proper inference requires full reporting and transparency*."

For example, if the published regression models are the result of dropping and adding variables until the t, p, R-squared and other values reach a certain level of significance, the published report should be transparent and say so. Drawing on the work of Committee members and others, Ziliak (2016) describes a number of easy-to-adopt changes to improve inferences with the style of the scientific research paper.

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

Student's t-statistic is a signal-to-noise ratio, and the p-value

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Stephen T. Ziliak, Roosevelt University College of Arts and Sciences–Economics, 430 S. Michigan Ave, Chicago, IL 60605-1313 (Email: sziliak@roosevelt.edu).

is the probability of observing a t-statistic equal to or larger than the one you see in the data, assuming to be true the null hypothesis and other data and modeling assumptions. The point here is that in p, t, or other form, the signal-to-noise ratio is not telling us what we want mainly to know: the answer to the expected size-matters/how much question. William Sealy Gosset aka "Student" (1876–1937) himself would agree, and would feel that his test of significance has been much abused. In a letter of 1905 to Karl Pearson the Guinness brewer and pioneer of small sample theory and applications said:

> When I first reported on the subject [of "The Application of the 'Law of Error' to the Work of the Brewery" (1904)], I thought that perhaps there might be some degree of probability which is conventionally treated as sufficient in such work as ours and I advised that some outside authority [in mathematics, such as Karl Pearson] should be consulted as to what certainty is required to aim at in large scale work. However it would appear that in such work as ours the degree of certainty to be aimed at must depend on the *pecuniary advantage to be gained by following the result of the experiment, compared with the increased cost of the new method, if any, and the cost of each experiment* (quoted in Ziliak [2008]).

Suppose diet pills Oomph and Precision are priced the same and bring the same side-effects. Oomph promises to remove 20 pounds on average but it is uncertain in actual effect, at plus or minus 10 pounds on either side of 20. Pill Precision promises a 5-pound weight-loss but its variance is found in well-designed studies to be much lower, at plus or minus 0.5 pounds. What pill is best? The signal-to-noise ratio of pill Oomph is $2(20/\pm 10)$ while for pill Precision the signal rises five times higher, to an impressive $10(5/\pm 0.5)$. Pill Precision, though more "significant," has, so to speak, no oomph. Precision works at best less effectively than Oomph at its worst. When choosing between two diet pills—between two tax policies, two blood-thinning medicines, or two paths for climate change—the signal-to-noise ratio—Student's t—is not the point. The point, what matters most times, is the expected size and meaning of uncertain effects across the whole distribution. What pill or path you favor should depend on the expected size and net value—the expected loss function—of acting as if the favored pill or hypothesis is true. Begin by not favoring the test of statistical significance.

References

American Statistical Association (2016), "Statement on Statistical Significance and P-Values," *The American Statistician*, 70.

Bialik, C. (2011), "Making a Stat Less Significant," *The Wall Street Journal*, April 11th, The Numbers Guy column. *http://www.wsj.com/articles/SB10001424052748703712504576235683249040812*.

Savage, L. (1954), *The Foundations of Statistics*, New York: Dover.

Ziliak, S. (2008), "Guinnessometrics: The Economic Foundation of 'Student's' t," *Journal of Economic Perspectives*, 22, 199–216.

——— (2016), "Statistical Significance and Scientific Misconduct: Improving the Style of the Published Research Paper," *Review of Social Economy*, 74 (1), 83–97. *http://dx.doi.org/10.1080/00346764.2016.1150730*.

Ziliak, S., and McCloskey, D. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor: University of Michigan Press.