Journal of Statistical Research *ISSN 0256 - 422 X*
2016, Vol. 48-50, No. 1, pp. 11-12

COMMENTARY ON "CRISIS IN SCIENCE? OR CRISIS IN STATISTICS! MIXED MESSAGES IN STATISTICS WITH IMPACT ON SCIENCE"

A. GELMAN
Department of Statistics and Department of Political Science
Columbia University, New York, NY 10027, USA
Email: gelman@stat.columbia.edu

I agree with Fraser and Reid that the many abuses of p-values in the real world arise not so much because applied researchers have ignored the lessons of statistics, but because in many ways they have learned the lessons of statistics all too well. Thus, to the extent that education from statisticians to practitioners is part of the solution, what is needed is not merely to shout existing messages even louder, or to require scientists to take more of the usual sort of statistics courses, but rather for we, the statistics profession, to think carefully about what messages we are sending and how these messages can be encouraging statistics abuse.

Before going on, let me clarify that I think the real problem is not with p-values but with what is called "null hypothesis significance testing" (NHST): the practice by which a researcher seeks to reject a straw-man null hypothesis as evidence in favor of some favored alternative. This is rife in the literature, indeed is the standard use of statistical analysis in psychology, medicine, and other fields. Heres an example, one of many: Carney, Cuddy, and Yap (2010) presented evidence that when subjects held their body in a certain posture called the power pose, they gained a feeling of confidence and certain hormone levels increased, compared to a control position in which the subjects held an alternative posture. Key comparisons in this paper had p-values of less than 0.05. A few years later, Ranehill et al. (2014) did a larger-scale preregistered replication of the study and failed to find an effect. What went wrong? The problem was with the logic of the significance test: In their original paper, Carney, Cuddy, and Yap had the choice of many possible data analyses; as a result, the probability of attaining statistical significance in some way would be much higher than 5%, even in the absence of any effect. Indeed, effects are small enough and variation is high enough that it would be essentially impossible to untangle signal from noise in a study of that size. The problem with the p-value here is that it is contingent on the choice of what analyses might have been performed, had the data been different.

In the Cuddy, Carney, and Yap study, a similar problem would have arisen with the NHST logic even had some other method than p-values been used. For example if likelihood ratios or Bayes factors were used to determine statistical significance and were used to reject the null, one would again have to be concerned about the many forking paths in this analysis.

Fraser and Reid discuss a Bayesian interpretation of the p-value that is similar to that presented by Greenland and Poole (2013). In my discussion to that Greenland and Poole paper (Gelman, 2013), I wrote that I see the mathematical connection between the (one-tailed) p-value and the posterior probability Pr$(\theta > 0 \mid y)$ under a uniform prior distribution on θ — but I question that uniform prior. The problems where NHST is causing the most problems are where effects are small. For example, Gertler et al. presented a study of early childhood intervention in Jamaica, reporting an effect of 42% on adult income. (It was a longitudinal study in which the children, first observed before school age, were followed up into their twenties.) The estimate was statistically significant; thus the 95% confidence interval on the treatment effect was something like [2%, 82%]. But I don't believe this interval. I certainly don't believe the 82% on the high end and, thinking Bayesianly, my prior based on the literature of such interventions is that any effect will be small. Perhaps a normal prior with mean 0 and standard deviation 10% would be reasonable, in which case the resulting posterior inference would not nearly be so optimistic as implied by the uniform prior. From a frequentist perspective, I do not think this interval has good coverage because of selection—"researcher degrees of freedom", in the words of Simmons, Nelson, and Simonsohn (2011)—in the data processing and analysis.

The problem that I see in statistical education is that we present statistical methods as alchemy, a way to convert randomness into a sort of certainty, as associated with words such as "confidence" and "significance". Look at statistics textbooks — including my own! — and you'll see example after example in which data are collected, analysis is done, and then inference is conveniently summarized with statistically significant p-values and confidence intervals that comfortably exclude zero. It's no wonder that practitioners, trained from such books, go out into the world expecting to find such clean summaries. The message we (implicitly) teach is that if you're studying a real effect and you have a good design and reasonable sample size, you'll succeed in the sense of getting a low p-value or a high posterior probability or a confidence interval that excludes zero.

Now consider this from the point of view of a researcher, Dr. X, analyzing some data. Dr. X presumably thinks hes studying a real effect (otherwise why work on the problem at all) and that he did a good design, and he might have even performed a power analysis to check that his sample size is large enough. Such power analyses are typically wildly optimistic because published effect size estimates tend to be way too large, biased as they are by the statistical significance filter: big estimates are statistically significant and get published, while estimates near zero, being non-significant, never appear. But this is a subtle point not mentioned in textbooks and not, we suspect, recognized by most researchers. So here is Dr. X, sure he's doing everything right and expecting to see a positive result: it's no wonder that he might jiggle his data a bit to get everything to line up.

So, to get researchers to stop chasing their tails with NHST, I think we need to revise our education, to take away the message that statistical significance (or the Bayesian or confidence interval equivalent) will come as a matter of course. Rather, researchers need to learn to live with uncertainty.

References

[1] Gelman A. (2012). P-values and statistical practice. *Epidemiology* **24**(1), 69–72

[2] Greenland S. and Poole C. (2012). Living with p values: resurrecting a Bayesian perspective on frequentist statistic. *Epidemiology* **24**(1), 62–68.