8 February, 2016

Discussion

Roderick J. Little, University of Michigan

The P-value statement is good, but perhaps more prominence should be given to the problems arising from the use of p-values as an *isolated* statistical measure. With very large data sets and high precision, p-values are all but useless, and the main focus is on estimates and potential sources of bias. With smaller studies where precision matters, no single measure can simultaneously answer two questions – "is the effect large?" and "is the estimated effect signal or noise?" An estimate and p-value address these questions, but I find a confidence or credibility interval much more cohesive and intuitive, giving a range of parameter values consistent with the data, and avoiding the need for a null hypothesis.

I teach a basic course in biostatistics to public health students. Confidence intervals are no problem – ideas like margin of error have even entered the vernacular. The difficulties begin with hypothesis testing. Aside from the elaborate terminology, what's the null and what's the alternative? When should the test be one-sided and when two-sided? Who cares about a point null that is never true (for example Nester, 1996)? If a deviation in one direction is of interest, the appropriate test seems to be one-sided; but the p-value calculation is still based on the known distribution of the test statistic under the point null, ignoring all the other null values. Isn't this a sleight of hand? Then we get to power calculations, which elude most quantitatively challenged students...

I have to teach hypothesis testing since it is so prevalent in biomedical research, but life would be much easier if we could just focus on estimates with their associated uncertainty. The basic artifice of hypothesis testing as a concept is perhaps the root cause of the problem, and I doubt that it will be solved by judicious and subtle statements like this one from the ASA Board.

Johnson's (2013) work points to an excessively high level of significance (5%) as a factor contributing to the failure to replicate science. On a lighter note, this inspired the following limerick, which I offer as my lame contribution to the debate:

In statistics, one rule did we cherish:
P point oh five we publish, else perish!
Said Val Johnson, "that's out of date,
Our studies don't replicate
P point oh oh five, <u>then</u> null is rubbish!"

References

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proc. Nat. Acad. Sci.*, 110, 48,

19313–19317.

Nester, M. R. (1996). An applied statistician’s creed. *Applied Statistics*, 45, 4, 401-410.