Response to commentary on "Is NHST logically flawed"

Jesper W. Schneider

Danish Centre for Studies in Research and Research Policy, Department of Political Science, Aarhus University, Bartholins Allé 7, 8000-DK, Aarhus, Denmark

jws@ps.au.dk

I thank the Editor for giving me the opportunity to respond to Dr. Patriota's (this issue, XXXX) brief comment in relation to two of my pervious publications on null hypothesis significance testing (NHST) in this journal (Schneider, 2015; 2018). I shall be brief myself in my response.

If I understand Patriota correctly, he insists that I am mistaken when I describe p values as conditional probabilities. According to Patriota, from a frequentist point of view this is incorrect because p values supposedly do not "respect the basic properties of conditional probabilities" (Patriota, this issue, XXXX, p. 2).

I am familiar with Patriota's claim, one he shares with other frequentists, see for example Larry Wasserman's blog entry but also the critical comments accompanying it[1], as well as the parallel discussion on Andrew Gelman's blog[2]; Patriota was an avid debater in both discussions. But like Gelman and several other commenters in the debate, I do not think I have misunderstood anything. Whether the p value is stated as frequentists prefer Pr(d(X) ≥ d(x0); H0), or with Bayesian notation Pr(d(X) ≥ d(x0)| H0), for all practical purposes in my view, the p value, is indeed a probability conditional or conditioned on an assumption, the null hypothesis. I see no point in re-opening this debate here, clearly Patriot and I disagree.

Patriota further argues that the syllogism of probabilistic *modus tollens*, which I and others (e.g., Pollard & Richardson, 1987) have presented to demonstrate the logically fallacy inherent in the NHST procedure, should be altered as it does not align with Fisher's famous disjunction (Fisher, 1956). A new syllogism is presented by Patriota and he claims that this one leads to the "reasonable conclusion" that when "$p(H_0; x) < \alpha$ … either a rare event occurred or H_0 is true", Fisher's disjunction (Patriota, this issue, XXXX, p. 2).

From a logically point of view I do not think this is a "reasonable conclusion". First, I still consider $p(H_0; x)$, or in my notation $p(H_0 | D)$ to be an inverse probability fallacy in this context. But, more important, Fisher took the process of logical refutation as the model for his significance tests. He stated that "the force of a test of significance is logically that of the simple disjunction: Either an exceptionally rare chance has occurred, or the theory or random distribution [i.e., the null hypothesis] is not true" (Fisher, 1956, p. 39). But as Howson and Urbach (2006, p. 150) argue "… avoiding an unreasonably strong interpretation, Fisher fell back on one that is unhelpfully weak, for the significant or critical results in a test of significance are by definition improbable, relative to the null hypothesis. Inevitably, therefore, a significant result is either a 'rare chance' (an improbably

[1] https://normaldeviate.wordpress.com/2013/03/14/double-misunderstandings-about-p-values/
[2] http://andrewgelman.com/2013/03/12/misunderstanding-the-p-value/

event) or the null hypothesis is false, or both. And Fisher's claim amounts to no more than this empty truism." The same point has been made earlier by Hacking (1965).

Recently Briggs (2017) has forcefully argued that, "Fisher's "logical disjunction" is evidently not one, since the either-or describes different propositions. A real disjunction can however be found: Either the null is false and we see a small p-value, or the null is true and we see as mall p-value. Or just: Either the null is true or it is false and we see a small p-value. Since 'Either the null is true or it is false' is a tautology, and is therefore necessarily true, we are left with, 'We see a small p-value.' The p-value casts no light on the truth or falsity of the null" (2017, p. 897).

In my view, Fisher's so-called disjunction is not a "reasonable conclusion" from a logical point of view and brings trouble to the inference procedure. I agree with Neyman and Pearson (1933) that is a misconception of probability. With this kind of "inductive inference", we cannot be sure that our conclusion is a logical result of the premises, hence "inferences" becomes decisions made using p values which Neyman criticised as an "act of will" (Neyman, 1937). That is why Neyman and Pearson was more restrictive when they spoke of "inductive behaviour" and decisions. When it comes to inference, probabilistic *modus tollens* is logically flawed (Sober, 2008).

Finally, Patriota argues "that the responsibility is all on the analyst to decide whether a significant result is relevant". As he states, "[w]e should not blame the statistical tool when in fact the problem lies in another domain …". A well-known argument, sure we should "blame" the users for misusing the statistical tools, but to educate better uses an understanding of their very limited relevance for very specific situations would certainly help, something Ronald Fisher was fully aware of.

References

Briggs, W. M. (2017). The Substitute for p-Values. *Journal of the American Statistical Association*, 112(519), 897-898.

Fisher, R. A. (1956). *Statistical Methods and Statistical Inference*. Edinburgh: Oliver and Boyed.

Hacking, I. (1965). *The Logic of Statistical Inference*, Cambridge: Cambridge University Press.

Howson, C. & Urbach, P. (2006). *Scientific Reasoning. The Bayesian Approach*. 3rd Edition, Peru, Ill.: Open Court.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society London*. Series A 236, 333–380.

Neyman, J. & Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses, *Philosophical Transactions of the Royal Society London*, Series A 231, 289–333.

Patriota, A. G. (XXXX). Is NHST logically flawed? Commentary on: "NHST is still logically flawed". *Scientometrics*, XXXX

Pollard, P., & Richardson, J. T. (1987). On the probability of making type I errors. *Psychological Bulletin*, 102(1), 159–163.

Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics*, 102(1), 411-432.

Schneider, J. W. (2018). NHST is still logically flawed. *Scientometrics*, 115(1), 627-635.

Sober, E. (2008). *Evidence and Evoluation. The Logic Behind Science*. Cambridge, UK: Cambridge University Press.