Scientometrics
https://doi.org/10.1007/s11192-018-2817-4

Is NHST logically flawed? Commentary on: "NHST is still logically flawed"

Alexandre Galvão Patriota[1]

Received: 19 March 2018

Schneider (2015) presents an interesting review on the main differences of Fisher's 'significance tests' and Neyman–Pearson's 'hypothesis tests'. The author says that "[i]n scientific reasoning, the most definitive test of a hypothesis is the syllogism of *modus tollens*or 'proof by contradiction"'and that "[t]his is also the logical form used in NHST, however, the crucial predicament is that modus tollens becomes formally incorrect with probabilistic statements which may lead to seriously incorrect conclusions."

In this short note, I point out an important missing ingredient in the application of the *modus tollens* syllogism to the Null Hypothesis Significance Test (NHST) used by Schneider's original paper (Schneider 2015) and Schneider's response (Schneider 2018) to Wu's commentary (Wu 2018) on Schneider's original paper.

The syllogism of *modus tollens* follows:

$$\text{If } (A \rightarrow B) \wedge (\neg B), \text{ then } \neg A. \tag{1}$$

That is, if "A implies B" and "B is not true", then conclude that "A is not true". *Modus tollens* is a valid inference procedure and it is often employed in statistical inference, specially in NHST, in the following sense: if A is an assumptiom that implies an observable event B (attained from an experiment) and, after conducting the experiment, we observe the negation of B, then we must conclude that our assumption A is not true.

Schneider (2018) provides the following scheme to apply the *modus tollens* in NHST:

Premise 1: If H_0 (i.e., A) is true, then Q (i.e., B) is highly likely...
Premise 2: Not-Q (i.e., $\neg B$)... (2)
Conclusion: H_0 is highly unlikely

The statement Q is about the p-value being greater than a certain threshold value α (the significance level). By using scheme (2) and interpreting p-values as conditional probabilities, Schneider (2018) concludes that: ".. since NHST is based on one conditional

✉ Alexandre Galvão Patriota
patriota@ime.usp.br

[1] Departamento de Estatística, IME, Universidade de São Paulo, Rua do Matão, 1010, São Paulo, SP 05508-090, Brazil

Published online: 26 June 2018

probability alone and framed in a probabilistic *modus tollens* framework of reasoning, it is by definition logically invalid."

In order to illustrate the scheme (2) in statistical notation, let us consider the null hypothesis $H_0 : \theta \in \Theta_0$, the observed data $x \in \mathcal{X} \subseteq \mathbb{R}^n$ and a positive test statistic $T_{H_0}(x)$ that orders the sample space in the following sense: the more discrepant H_0 is from the observed data x, the larger is the observed value $T_{H_0}(x)$; for instance, $T_{H_0}(x) = -2\log(\lambda(H_0, x))$, where $\lambda(H_0, x) = \frac{\sup_{\theta\in\Theta_0} L(\theta_0;x)}{\sup_{\theta\in\Theta} L(\theta;x)}$ is the likelihood ratio statistics and $L(\theta; x)$ is the likelihood function. The following p-value's definition satisfies Fisher's requirements for significance tests:

$$p(H_0; x) = \sup_{\theta\in\Theta_0} P_\theta\big(T_{H_0}(X) \geq T_{H_0}(x)\big).$$

It is well known that, under H_0 and some regular conditions on the statistical model and on the geometry of Θ_0, asymptotically $p(H_0, X)$ has an uniform distribution. The p-value is the probability of observing an extreme event in the best scenario of H_0. Moreover, it is noteworthy that the p-value is **not** a conditional probability given H_0 as it is typically stated. It is not correct to operate the above p-value as it were a genuine conditional probability, since the events H_0 and $\{T_{H_0}(X) \geq T_{H_0}(x)\}$ are not measurable in the same space and therefore the conditional probability should not be employed for them. Conclusions on the validity of p-values based on conditional probability arguments are most incorrect in the domain of the classical or frequentist frameworks[1].

Let us return to the scheme (2). Notice that it can be rewritten in terms of statistical statements as

$$\text{If } \big[\underbrace{H_0}_{A} \rightarrow \underbrace{p(H_0; x) \geq \alpha}_{B}\big] \wedge (\underbrace{p(H_0; x) < \alpha}_{\neg B}), \text{ then } \underbrace{(\neg H_0)}_{\neg A}. \tag{3}$$

The problem with (3) is that the null hypothesis H_0 cannot alone guarantee that $p(H_0; x) \geq \alpha$, since we could have observed a "rare" event under H_0 such that $p(H_0; x) < \alpha$. That is, it is not the case that H_0 implies $p(H_0; x) \geq \alpha$. Therefore, the syllogism of *modus tollens* should not be applied in form of (3). Furthermore, this line of reasoning does not represent the Fisher's disjunction upon a significant result: "either a **rare event occurred** or H_0 **is not true**" (Fisher 1959). The missing ingredient in (3) is discussed in what follows.

Let R_{H_0} be a subset of the sample space that indicates the relevant rare event under the null hypothesis H_0 such that $H_0 \wedge (x \notin R_{H_0})$ implies $p(H_0; x) \geq \alpha$. We take A to be the statement "$H_0 \wedge (x \notin R_{H_0})$" and B the statement "$p(H_0; x) \geq \alpha$", then the syllogism presented in Eq. (1) for a significance test should read as follows

$$\text{If } \big[\underbrace{[H_0 \wedge (x \notin R_{H_0})]}_{A} \rightarrow \underbrace{p(H_0; x) \geq \alpha}_{B}\big] \wedge (\underbrace{p(H_0; x) < \alpha}_{\neg B}), \text{ then } \underbrace{[(\neg H_0) \vee (x \in R_{H_0})]}_{\neg A}. \tag{4}$$

The whole statement (4) is interpreted in plain English as follows: provided x is not a rare event and the null hypothesis H_0 is true, then $p(H_0; x) \geq \alpha$. If, however, we observe $p(H_0; x) < \alpha$, then we must conclude that either a **rare event occurred** or H_0 **is not true**. This seems to be very reasonable to me.

[1] I use the term "classical framework" when the classical statistical model is employed as a mathematical tool without necessarily adopting the frequentist paradigm

The statement (4) suggests that the responsability is all on the analyst to decide whether a significant result is relevant. We should not blame the statistical tool when in fact the problem lies in another domain. The human factor should be considered more seriously, since it seems to be common in modern science that some experiments are not reproducible (see Open Science Collaboration 2015) and also tend to overestimate effect sizes (Fanelli et al. 2017). Futhermore, although the usual *p*-value has some technical issues (see, Schervish 1996, for instance), they can be avoided by redefining it by means of confidence sets (Patriota 2013; Bickel and Patriota 2018).

Acknowledgements The author gratefully acknowledges Grants from CNPq (Brazil).

References

Bickel, D. R., & Patriota, A. G. (2018). Self-consistent confidence sets and tests of composite hypotheses applicable to restricted parameters, Bernoulli. http://www.bernoulli-society.org/index.php/publications/bernoulli-journal/bernoulli-journal-papers.

Fanelli, D., Costas, R., & Ioannidis, J. P. A. (2017). Meta-assessment of bias in science. *PNAS*, *114*(14), 3714–3719.

Fisher, R. A. (1959). *Statistical methods and scientific inference* (2nd ed.). Edinburgh: Oliver and Boyd.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716.

Patriota, A. G. (2013). A classical measure of evidence for general null hypotheses. *Fuzzy Sets and Systems*, *233*, 74–88.

Schervish, M. J. (1996). *P* values: What they are and what they are not. *The American Statistician*, *50*, 203–206.

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*, 411–432.

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

Wu, J. (2018). Is there an intrinsic logical error in null hypothesis significance tests? Commentary on: "Null hypothesis significance tests. A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations". *Scientometrics*, *115*, 621–625.