Commentary: the need for bayesian hypothesis testing in psychological science

Health & Medicine

A commentary on

The Need for Bayesian Hypothesis Testing in Psychological Science

by Wagenmakers, E. J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., et al. (2017). Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions, eds S. O. Lilienfeld and I. D. Waldman (Chichester: JohnWiley & Sons), 123–138. .

Wagenmakers et al. (2017) argued the need for a Bayesian approach to inferential statistics in *Psychological Science under Scrutiny* . Their primary goal was to demonstrate the illogical nature of *p* -values, while, secondarily, they would also defend the philosophical consistency of the Bayesian alternative. In my opinion, they achieved their secondary goal but failed their primary one, thereby this contribution. I will, thus, comment on their interpretation of the logic underlying *p* -values without necessarily invalidating their Bayesian arguments.

Historical criticisms (e. g., Harshbarger, 1977 , onwards) have already delved in the illogical nature of null hypothesis significance testing (NHST)—a mishmash of Fisher's, Neyman-Pearson's, and Bayes's ideas (e. g., Gigerenzer, 2004 ; Perezgonzalez, 2015a). Wagenmakers et al.'s original contribution is to generalize similar criticisms to the *p* -value itself, the statistic used by frequentists when testing research data.

Wagenmakers et al. assert that Fisher's disjunction upon obtaining a significant result—i. e., either a rare event occurred or H_0 is not true (Fisher, 1959)—follows from a logically consistent *modus tollens* (also Sober, 2008):

If P, then Q; not Q; therefore not P, which the authors parsed as, If H_0, then not y; y; therefore not H_0.

" Y" is defined as " the observed data...[summarized by] the *p* -value" (p. 126). Therefore, their first premise proposes that, if H_0 is true, the observed *p* -values cannot occur (also Cohen, 1994 ; Beck-Bornholdt and Dubben, 1996). This seems incongruent, as the first premise of a correct *modus tollens* states a general rule—H_0 implies " not y"—while the second premise states a specific test to such rule—" this y" has been observed. If the authors meant for " y" to represent " significant data" as a general category in the first premise and as a specific realization in the second, a congruent *modus tollens* would ensue, as follows (also Pollard and Richardson, 1987):

If H_0, then not $p < sig$; $p < sig$ (observed); therefore not H_0 (1)

Wagenmakers et al.'s (also Pollard and Richardson, 1987 ; Cohen, 1994 ; Falk, 1998) main argument is that a correct *modus tollens* is rendered inconsistent when made probabilistic, as follows:

If H_0, then $p < sig$ very unlikely; $p < sig$; therefore probably not H_0 (2)

There are, however, three problems with (2), problems which I would like to comment upon. One problem is stylistic: The first premise states a redundant probability; that is, that a significant result—which already implies an unlikely or improbable event under H_0—is unlikely. Therefore, the syllogism could be simplified as follows:

If H_0, then $p < sig$; $p < sig$; therefore probably not H_0 (3)

Correction (3) now highlights another of the problems: The second premise simply affirms that an unlikely result just happened (also Cortina and Dunlap, 1997), something which is neither precluded by the first premise (no contrapositive ensues; Adams, 1988) nor formally conducive to a logical conclusion under *modus tollens* (Evans, 1982). Indeed, in the examples given (also by Cohen, 1994 ; Beck-Bornholdt and Dubben, 1996 ; Cortina and Dunlap, 1997 ; Krämer and Gigerenzer, 2005 ; Rouder et al., 2016), Tracy is a US congresswoman, Francis is the Pope, and John made money at the casino, each despite their odds against. Yet, none of those realizations deny the consequents. A correction, following Harshbarger (1977) and Falk (1998) , would state:

If H_0 , then not $p <$ sig; $p <$ sig; therefore probably not H_0 (4)

Correction (4) brings to light the most important problem: *Modus tollens* is in the form, If P, then Q; not Q; therefore not P. Thus, whenever the consequent (Q) gets denied in the second premise, it leads to denying the antecedent (P) in the conclusion. Such operation ought to prevail with probabilistic premises, as well (e. g., Oaksford and Chater, 2001 , 2009 ; Evans et al., 2015), whereby a probable consequent (Q_p) may be denied without its probability warranting transposition onto a non-probabilistic antecedent (P). For example, if all red cars (P) have a 95% chance of getting stolen ($Q \geq 0.95$) and we learn of a Lamborghini with a lesser probability of so disappearing (not $Q \geq 0.95$), it is logical to conclude that the Lamborghini is not red (not P).

In comparison, Bayesian logic allows for the antecedent to be probable. For example, if John always submits to Nature (Q) whenever his subjective probability of getting published soars above 20% ($P > 0.2$), yet he is not submitting his latest article (not Q), it is logical to conclude that he probably expects no publication (not $P > 0.2$).

We can, thus, envisage P or Q, or both, as probable without either warranting inter-transposition of their probabilities, which brings us back to a valid *modus tollens* (1). Said otherwise, while Bayesian statistics allow for the antecedent to be probable (P_p), Fisher's and Neyman-Pearson's tests assume exact antecedents (P); therefore, a probabilistic conclusion does not hold with frequentist tests (Mayo, 2017).

It ought to be noted that the *p* -value is a statistic descriptive of the probability of the data under H_0 [$p(D|H_0)$] (Perezgonzalez, 2015b). The *reductio ad absurdum* argument may be informed by, but it is not dependent on, such *p* -value, the *reductio* being determined exclusively by the chosen level of significance. For " it is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result. It is obvious that an experiment would be useless of which no possible result would satisfy him" (Fisher, 1960 , p. 13).

In conclusion, the technology of frequentist testing holds their *modus tollens* logically. Wagenmakers et al.'s criticism of the *p* -value is faulty in that they allow for a probability transposition not warranted either by *modus tollens* or by the technical apparatus of Fisher's and of Neyman-Pearson's tests. This

critique, however, does not extend to their Bayesian argumentation, an approach much needed for testing hypotheses—rather than just testing data—in contemporary science.

Author Contributions

The author confirms being the sole contributor of this work and approved it for publication.

Conflict of Interest Statement

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Adams, E. W. (1988). Modus tollens revisited. *Analysis* 48, 122–128. doi: 10. 1093/analys/48. 3. 122

CrossRef Full Text | Google Scholar

Beck-Bornholdt, H. P., and Dubben, H. H. (1996). Is the Pope an alien? *Nature* 381: 730. doi: 10. 1038/381730d0

CrossRef Full Text

Cohen, J. (1994). The earth is round (p < 0. 05). *Am. Psychol.* 49, 997–1003. doi: 10. 1037/0003-066X. 49. 12. 997

CrossRef Full Text

Cortina, J. M., and Dunlap, W. P. (1997). On the logic and purpose of significance testing. Psychol. *Methods* 2, 161–172. doi: 10. 1037/1082-989X. 2. 2. 161

CrossRef Full Text | Google Scholar

Evans, J. St. B. T. (1982). *The Psychology of Deductive Reasoning* . London: Routledge & Kegan Paul.

Google Scholar

Evans, J. St. B. T., Thompson, V. A., and Over, D. E. (2015). Uncertain deduction and conditional reasoning. Front. *Psychol.* 6: 398. doi: 10. 3389/fpsyg. 2015. 00398

CrossRef Full Text | Google Scholar

Falk, R. (1998). In criticism of the null hypothesis statistical test. Am. *Psychol.* 53, 798–799. doi: 10. 1037/0003-066X. 53. 7. 798

CrossRef Full Text

Fisher, R. A. (1959). *Statistical Methods and Scientific Inference, 2nd Edn* . Edinburgh: Oliver and Boyd.

Fisher, R. A. (1960). *The Design of Experiments, 7th Edn* . Edinburgh: Oliver and Boyd.

Gigerenzer, G. (2004). Mindless statistics. J. Soc. *Econ.* 33, 587–606. doi: 10. 1016/j. socec. 2004. 09. 033

CrossRef Full Text | Google Scholar

Harshbarger, T. R. (1977). *Introductory Statistics: A Decision Map, 2nd Edn* . New York, NY: Macmillan.

Krämer, W., and Gigerenzer, G. (2005). How to confuse with statistics or: the use of misuse of conditional probabilities. Stat. *Sci.* 20, 223–230. doi: 10. 1214/088342305000000296

PubMed Abstract | CrossRef Full Text

Mayo, D. G. (2017). *If You're Seeing Limb-Sawing in p-value Logic, You're Sawing off the Limbs of Reductio Arguments [Web Log Post]* . Available online at: https://errorstatistics. com/2017/04/15/if-youre-seeing-limb-sawing-in-p-value-logic-youre-sawing-off-the-limbs-of-reductio-arguments/

Oaksford, M., and Chater, N. (2001). The probabilistic approach to human reasoning. Trends Cogn. *Sci.* 5, 349–357. doi: 10. 1016/S1364-6613(00)01699-5

CrossRef Full Text | Google Scholar

Oaksford, M., and Chater, N. (2009). Précis of bayesian rationality: the probabilistic approach to human reasoning. Behav. *Brain Sci.* 32, 69–84. doi: 10. 1017/S0140525X09000284

CrossRef Full Text | Google Scholar

Perezgonzalez, J. D. (2015a). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Front. *Psychol.* 6: 223. doi: 10. 3389/fpsyg. 2015. 00223

PubMed Abstract | CrossRef Full Text | Google Scholar

Perezgonzalez, J. D. (2015b). P-values as percentiles. Commentary on: " Null hypothesis significance tests. A mix–up of two different theories: the basis for widespread confusion and numerous misinterpretations." Front. *Psychol.* 6: 341. doi: 10. 3389/fpsyg. 2015. 00341

CrossRef Full Text | Google Scholar

Pollard, P., and Richardson, J. T. E. (1987). On the probability of making type I errors. Psychol. *Bull.* 102, 159–163. doi: 10. 1037/0033-2909. 102. 1. 159

CrossRef Full Text | Google Scholar

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., and Wagenmakers, E. J. (2016). Is there a free lunch in inference? Top. Cogn. *Sci.* 8, 520–547. doi: 10. 1111/tops. 12214

CrossRef Full Text | Google Scholar

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

Google Scholar

Wagenmakers, E. J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., et al. (2017). “ The need for Bayesian hypothesis testing in psychological science,” in *Psychological Science under Scrutiny: Recent Challenges and Proposed Solutions* , eds S. O. Lilienfeld and I. D. Waldman (Chichester: John Wiley & Sons), 123–138.

Google Scholar