The logic of p-values[1]

Jose D. Perezgonzalez (Massey University, New Zealand)

Wagenmakers et al. (2017) addressed the illogic use of *p*-values in inferential statistics in *Psychological Science under Scrutiny*. While historical criticisms (e.g., Harshbarger, 1977, onwards) mostly deal with 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. generalize such argumentation 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)—is a logically consistent *modus tollens* (also Sober, 2008): If P, then Q; not Q; consequently not P, which the authors parsed as: If H_0, then not y; y; consequently not H_0.

The authors defined 'y' 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), which seems incongruent. Indeed, 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. I guess 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. Thus, following Pollard and Richardson (1987), a congruent *modus tollens* would be:

$$\text{If } H_0\text{, then not } p < \text{sig; } p < \text{sig (observed); consequently not } H_0 \qquad (0)$$

[1] Preprint of, Perezgonzalez, J. D. (2017). Commentary: The need for Bayesian hypothesis testing in psychological science. *Front. Psychol*. 8:1434. doi: 10.3389/fpsyg.2017.01434

Wagenmakers et al.'s main argument is that a correct *modus tollens* is rendered inconsistent when made probabilistic, as: If H_0, then $p <$ sig very unlikely; $p <$ sig; consequently H_0 very unlikely (also Pollard and Richardson, 1987; Cohen, 1994; Falk, 1998). There are, however, three problems with the argument.

The first problem is stylistic. The first premise states that a significant result—which already implies an unlikely or improbable event under H_0—is unlikely: a redundant probability statement. Their probabilistic syllogism can thus be simplified as:

If H_0, then $p <$ sig (i.e., very unlikely p's); $p <$ sig; consequently H_0 very unlikely (1)

Correction (1) makes now quite evident the second problem. The second premise simply affirms that an unlikely result just happened (also Cortina and Dunlap, 1997), 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). Such realization of an unlikely event is obvious in the examples given: Tracy is a US congresswoman, Francis is the Pope, and John made money at the casino, each despite the odds against them, none denying the consequent (also Cohen, 1994; Beck-Bornholdt and Dubben, 1996; Cortina and Dunlap, 1997; Krämer and Gigerenzer, 2005; Rouder et al., 2016). A plausible correction, following Harshbarger (1977) and Falk (1998), would state:

If H_0, then not $p <$ sig; $p <$ sig; consequently probably not H_0 (2)

Correction (2) brings to light the third and most important problem. *Modus tollens* is in the form: If P, then Q; not Q; consequently not P. Therefore, whenever the consequent (Q) gets denied in the second premise, it leads to denying the antecedent (P) in the conclusion. The same operation ought to prevail with probabilistic premises (e.g., Oaksford and Chater, 2001, 2009; Evans, Thompson and Over, 2015), whereby a probable Q_p gets denied in the second premise without its probability warranting transposition onto a non-probabilistic antecedent P.

For example, if red cars (P) get stolen 95% of the time ($Q_{\geq.95}$) and we learn of a Lamborghini with little chance of so disappearing (not $Q_{\geq.95}$), it is logical to conclude that the Lamborghini is not red (not P). Equally, if John submits to Nature (Q) whenever his subjective probability of getting published soars above 20% ($P_{>.2}$), yet he will not submit his latest article (not Q), it is logical to conclude that he probably expects no publication (not $P_{>.2}$). Furthermore, if the probability of people playing lotto increases (Q_y) whenever winning is more probable (e.g., a 'Must Be Won' jackpot; P_x), yet ticket sales are rather flat (not Q_y), it is logical to conclude that there is probably no 'Must Be Won' jackpot on the cards (not P_x).

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* (0), contrary to what Wagenmakers et al.'s (and historical) arguments claim. Said otherwise, while Bayesian statistics allow for the antecedent to be probable (P_p), Fisher's and Neyman-Pearson's tests assume true 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 is not dependent on, such *p*-value, the *reductio* being determined exclusively by the chosen level of significance, whether conventional or not, and whether established a priori (α) or not. 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).

Therefore, the technology of frequentist testing holds their *modus tollens* logically. While historical critiques are unclear on whether they are (wrongly) criticizing frequentist tests or (correctly) criticizing the NHST mishmash, Wagenmakers et al.'s criticism of the *p*-value is faulty in that they allow for a probability transposition warranted neither by *modus tollens* nor by the technical apparatus of Fisher's and Neyman-Pearson's tests.

References

Adams, E. W. (1988). Modus tollens revisited. *Analysis* 48, 122-128. doi:10.2307/3328213

Beck-Bornholdt, H. P., and Dubben, H. H. (1996). Is the Pope an alien? *Nature* 381, 730. doi:10.1038/381730d0

Cohen, J. (1994). The Earth is round (p < .05). *Am. Psychol.* 49, 997-1003. doi:10.1037/0003-066X.49.12.997

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

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

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

Falk, R. (1998). In criticism of the null hypothesis statistical test. *Am. Psychol.* 53, 798-799.

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

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

Gigerenzer, G. (2004). Mindless statistics. *J. Soc. Econ.* 33, 587–606. doi:10.1016/j.socec.2004.09.033

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

Mayo, D. G. (2017, April 15). If you're seeing limb-sawing in p-value logic, you're sawing off the limbs of reductio arguments [Web log post]. Retrieved from 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.

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

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

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

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

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: 520-47. doi:10.1111/tops.12214.

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

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