Received: 7 May 2021 | Revised: 20 September 2021 | Accepted: 29 September 2021

DOI: 10.1111/cobi.13861

Conservation Biology

EDITORIAL

The statistics wars and intellectual conflicts of interest

How should journal editors react to heated disagreements about statistical significance tests in applied fields, such as conservation science, where statistical inferences often are the basis for controversial policy decisions? They should avoid taking sides. They should also avoid obeisance to calls for author guidelines to reflect a particular statistical philosophy or standpoint. The question is how to prevent the misuse of statistical methods without selectively favoring one side.

The statistical-significance-test controversies are well known in conservation science. In a forum revolving around Murtaugh's (2014) "In Defense of P values," Murtaugh argues, correctly, that most criticisms of statistical significance tests "stem from misunderstandings or incorrect interpretations, rather than from intrinsic shortcomings of the P value" (p. 611). However, underlying those criticisms, and especially proposed reforms, are often controversial philosophical presuppositions about the proper uses of probability in uncertain inference. Should probability be used to assess a method's probability of avoiding erroneous interpretations of data (i.e., error probabilities) or to measure comparative degrees of belief or support? Wars between frequentists and Bayesians continue to simmer in calls for reform.

Consider how, in commenting on Murtaugh (2014), Burnham and Anderson (2014 : 627) aver that "P-values are not proper evidence as they violate the likelihood principle (Royall, 1997)." This presupposes that statistical methods ought to obey the likelihood principle (LP), a long-standing point of controversy in the statistics wars. The LP says that all the evidence is contained in a ratio of likelihoods (Berger & Wolpert, 1988). Because this is to condition on the particular sample data, there is no consideration of outcomes other than those observed and thus no consideration of error probabilities. One should not write this off because it seems technical: methods that obey the LP fail to directly register gambits that alter their capability to probe error. Whatever one's view, a criticism based on presupposing the irrelevance of error probabilities is radically different from one that points to misuses of tests for their intended purpose—to assess and control error probabilities.

Error control is nullified by biasing selection effects: cherry-picking, multiple testing, data dredging, and flexible stopping rules. The resulting (nominal) p values are not legitimate p values. In conservation science and elsewhere, such misuses can result from a publish-or-perish mentality and experimenter's flexibility (Fidler et al., 2017). These led to calls for preregistration of hypotheses and stopping rules–one of the most effective ways to promote replication (Simmons et al., 2012). However, data dredging can also occur with likelihood ratios, Bayes factors, and Bayesian updating, but the direct grounds to criticize inferences as flouting error probability control is lost. This conflicts with a central motivation for using p values as a "first line of defense against being fooled by randomness" (Benjamini, 2016). The introduction of prior probabilities (subjective, default, or empirical)–which may also be data dependent–offers further flexibility.

Signs that one is going beyond merely enforcing proper use of statistical significance tests are that the proposed reform is either the subject of heated controversy or is based on presupposing a philosophy at odds with that of statistical significance testing. It is easy to miss or downplay philosophical presuppositions, especially if one has a strong interest in endorsing the policy upshot: to abandon statistical significance. Having the power to enforce such a policy, however, can create a conflict of interest (COI). Unlike a typical COI, this one is intellectual and could threaten the intended goals of integrity, reproducibility, and transparency in science.

If the reward structure is seducing even researchers who are aware of the pitfalls of capitalizing on selection biases, then one is dealing with a highly susceptible group. For a journal or organization to take sides in these long-standing controversies—or even to appear to do so—encourages groupthink and discourages practitioners from arriving at their own reflective conclusions about methods.

The American Statistical Association (ASA) Board appointed a President's Task Force on Statistical Significance and Replicability in 2019 that was put in the odd position of needing to "address concerns that a 2019 editorial [by the ASA's executive director (Wasserstein et al., 2019)] might be mistakenly interpreted as official ASA policy" (Benjamini et al., 2021)—as if the editorial continues the 2016 ASA Statement on p-values (Wasserstein & Lazar, 2016). That policy statement merely warns against well-known fallacies in using p values. But Wasserstein et al. (2019) claim it "stopped just short of recommending that declarations of 'statistical significance' be abandoned" and announce taking that step. They call on practitioners not to use the phrase *statistical significance* and to avoid p value thresholds. Call this the no-threshold view. The 2016 statement was largely uncontroversial; the 2019 editorial was anything but. The President's Task Force should be commended for working to resolve the confusion (Kafadar, 2019). Their report concludes: "P-values are valid statistical measures that provide convenient

conventions for communicating the uncertainty inherent in quantitative results" (Benjamini et al., 2021). A disclaimer that Wasserstein et al., 2019 was not ASA policy would have avoided both the confusion and the slight to opposing views within the Association.

The no-threshold view has consequences (likely unintended). Statistical significance tests arise "to test the conformity of the particular data under analysis with [a statistical hypothesis] H_0 in some respect to be specified" (Mayo & Cox, 2006: 81). There is a function D of the data, the test statistic, such that the larger its value (d), the more inconsistent are the data with H_0. The p value is the probability the test would have given rise to a result more discordant from H_0 than d is were the results due to background or chance variability (as described in H_0). In computing p, hypothesis H_0 is assumed merely for drawing out its probabilistic implications. If even larger differences than d are frequently brought about by chance alone (p is not small), the data are not evidence of inconsistency with H_0. Requiring a low p *value* before inferring inconsistency with H_0 controls the probability of a type I error (i.e., erroneously finding evidence against H_0).

If p is low, then there is a high probability, $1 - p$, that the test would have produced a result that accords better with H_0, if one were dealing with chance variability alone. So, a low p value indicates inconsistency with H_0. The H_0 may be seen as statistically falsified (at the indicated level) if the low p value is not merely an "isolated result" but is brought about reliably (Fisher, 1947).

Such an indication is not automatically evidence of a hypothesis that explains the effect. Neyman–Pearson (N-P) tests are explicit that rejecting H_0 only indicates the alternative statistical hypothesis H_1, where H_0 and H_1 together exhaust the possibilities for the test. The simple (Fisherian) statistical significance test, with a single null hypothesis, has important uses in testing model assumptions; and both Bayesians and frequentists use them to this end (Gelman and Shalizi, 2013). However, a fair comparison of tests and confidence intervals must look to N-P tests, there being a duality between the two. Tests can be specified to control the probability of both type I and type II errors (i.e., erroneously failing to find evidence against the null hypothesis). Setting a low type II error probability against alternatives of interest ensures high power to detect them. Power turns on there being a threshold value for D beyond which data are taken as evidence against H_0.

Whether interpreting a simple Fisherian or an N-P test, avoiding fallacies calls for considering one or more discrepancies from the null hypothesis under test. Consider testing a normal mean $H_0: \mu \leq \mu_0$ versus $H_1: \mu > \mu_0$. If the test would fairly probably have resulted in a smaller p value than observed, if $\mu = \mu_1$ were true (where $\mu_1 = \mu_0 + \gamma$, for $\gamma > 0$), then the data provide poor evidence that μ exceeds μ_1. It would be unwarranted to infer evidence of $\mu > \mu_1$. Tests do not need to be abandoned when the fallacy is easily avoided by computing p values for one or two additional benchmarks (Burgman, 2005; Hand, 2021; Mayo, 2018; Mayo & Spanos, 2006).

The same is true for avoiding fallacious interpretations of nonsignificant results. These are often of concern in conservation, especially when interpreted as no risks exist. In fact, the test may have had a low probability to detect risks. But nonsignificant results are not uninformative. If the test very probably would have resulted in a more statistically significant result were there a meaningful effect, say $\mu > \mu_1$ (where $\mu_1 = \mu_0 + \gamma$, for $\gamma > 0$), then the data are evidence that $\mu < \mu_1$. (This is not to infer $\mu \leq \mu_0$.) "Such an assessment is more relevant to specific data than is the notion of power" (Mayo & Cox, 2006: 89). This also matches inferring that μ is less than the upper bound of the corresponding confidence interval (at the associated confidence level) or a severity assessment (Mayo, 2018). Others advance equivalence tests (Lakens, 2017; Wellek, 2017). An N-P test tells one to specify H_0 so that the type I error is the more serious (considering costs); that alone can alleviate problems in the examples critics adduce (H_0 would be that the risk exists).

Many think the no-threshold view merely insists that the attained p value be reported. But leading N-P theorists already recommend reporting p, which "gives an idea of how strongly the data contradict the hypothesis…[and] enables others to reach a verdict based on the significance level of their choice" (Lehmann & Romano, 2005: 63−64). What the no-threshold view does, if taken strictly, is preclude testing. If one cannot say ahead of time about any result that it will not be allowed to count in favor of a claim, then one does not test that claim. There is no test or falsification, even of the statistical variety. What is the point of insisting on replication if at no stage can one say the effect failed to replicate? One may argue for approaches other than tests, but it is unwarranted to claim by fiat that tests do not provide evidence. (For a discussion of rival views of evidence in ecology, see Taper & Lele, 2004.)

Many sign on to the no-threshold view thinking it blocks perverse incentives to data dredge, multiple test, and p hack when confronted with a large, statistically nonsignificant p value. Carefully considered, the reverse seems true. Even without the word *significance*, researchers could not present a large (nonsignificant) p value as indicating a genuine effect. It would be nonsensical to say that even though more extreme results would frequently occur by random variability alone that their data are evidence of a genuine effect. The researcher would still need a small p value, which is to operate with a threshold. However, it would be harder to hold data dredgers culpable for reporting a nominally small p value obtained through data dredging. What distinguishes nominal p values from actual ones is that they fail to meet a prespecified error probability threshold.

The no-threshold view is in tension with the U.S. Food and Drug Association's "long established drug review procedures that involve comparing p-values to significance thresholds for Phase III drug trials" (Wasserstein et al., 2019: 10). In response to a request by ASA officials to revise their author guidelines, *The New England Journal of Medicine* (2019) refuses to relinquish their requirement that the use of statistics in "claiming an effect or association should be limited to analyses for which the analysis plan outlined a method for controlling Type I error…" (Harrington et al., 2019: 286).

While it is well known that stopping when the data look good inflates the type I error probability, a strict Bayesian is not required to adjust for interim checking because the posterior

probability is unaltered. Advocates of Bayesian clinical trials are in a quandary because "The [regulatory] requirement of Type I error control for Bayesian [trials] causes them to lose many of their philosophical advantages, such as compliance with the likelihood principle" (Ryan et al., 2020: 7).

It may be retorted that implausible inferences will indirectly be blocked by appropriate prior degrees of belief (informative priors), but this misses the crucial point. The key function of statistical tests is to constrain the human tendency to selectively favor views they believe in. There are ample forums for debating statistical methodologies. There is no call for executive directors or journal editors to place a thumb on the scale. Whether in dealing with environmental policy advocates, drug lobbyists, or avid calls to expel statistical significance tests, a strong belief in the efficacy of an intervention is distinct from its having been well tested. Applied science will be well served by editorial policies that uphold that distinction.

ACKNOWLEDGMENTS

I thank M. Burgman, J. Miller, N. Schachtman, and anonymous reviewers for important corrections and constructive suggestions on earlier drafts.

Deborah G. Mayo

Virginia Tech, Blacksburg, Virginia, USA

Correspondence
Deborah G. Mayo, 235 Major Williams Hall, Philosophy Dept., Virginia Tech, Blacksburg, VA 24061, USA.
Email: mayod@vt.edu

Article impact statement: Editorial policies in conservation science should not selectively favor one side on controversies about statistical significance tests.

ORCID

Deborah G. Mayo https://orcid.org/0000-0001-8252-9968

LITERATURE CITED

Benjamini, Y. (2016). It's not the P-values' fault. Comment on Wasserstein and Lazar (2016), supplemental material (online). Available from: https://tandf.figshare.com/articles/dataset/The_ASA_s_statement_on_p_values_context_process_and_purpose/3085162/7?file=5368448.

Benjamini, Y., De Veaux, R., Efron, B., Evans, S., Glickman, M., Graubard, B., He, X., Meng, X.-L., Reid, N., Stigler, S., Vardeman, S., Wikle, C., Wright, T., Young, L., & Kafadar, K.. (2021). The ASA President's Task Force statement on statistical significance and replicability. *Annals of Applied Statistics*, *15*(3), 1084–1085.

Berger, J., & Wolpert, R. (1988). *The likelihood principle* (2nd edition). Lecture Notes-Monograph Series 6. Hayward, CA: Institute of Mathematical Statistics.

Burgman, M. (2005). *Risks and decisions for conservation and environmental management*. Cambridge: Cambridge University Press.

Burnham, K., & Anderson, D. (2014). *P* values are only an index to evidence: 20th-vs. 21st-century statistical science. *Ecology*, *95*(3), 627–630.

Fidler, F., Chee, Y., Wintle, B., Burgman, M., McCarthy, M., & Gordon, A. (2017). Metaresearch for evaluating reproducibility in ecology and evolution. *Bioscience*, *67*, 282–289.

Fisher, R. A. (1947). *The design of experiments* (4th edition). Edinburgh: Oliver and Boyd.

Gelman, A., Shalizi, C. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*, (1), 8. –38. http://doi.org/10.1111/j.2044-8317.2011.02037.x

Hand, D. J. (2021). Trustworthiness of statistical inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, http://doi.org/10.1111/rssa.12752

Harrington, D., D'Agostino, R., Gatsonis, C., Hogan, J., Hunter, M., Normand, S.-L., Drazen, J., & Hamel, M. (2019). New guidelines for statistical reporting in the journal. *New England Journal of Medicine*, *381*(3), 285–286.

Kafadar, K. (2019). The year in review…And more to come. President's corner. *Amstatnews*, *510*, 3–4.

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362.

Lehmann, E., & Romano, J. (2005). *Testing statistical hypotheses* (3rd edition). New York: Springer.

Mayo, D. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge: Cambridge University Press.

Mayo, D., & Cox, D. (2006). Frequentist statistics as a theory of inductive inference. In J. Rojo (Ed.), *Optimality: The Second Erich L. Lehmann Symposium*. Lecture Notes-Monograph series 49. Beachwood, OH: Institute of Mathematical Statistics. pp. 77–97.

Mayo, D., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *British Journal for the Philosophy of Science*, *57*(2), 323–357.

Murtaugh, P. A. (2014). In defense of *p* values. *Ecology*, *95*(3), 611–617.

New England Journal of Medicine (2019). Author guidelines. Available from: https://www.nejm.org/authorcenter/new-manuscripts.

Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. Boca Raton, FL: Chapman and Hall, CRC Press.

Ryan, E., Brock, K., Gates, S., & Slade, D. (2020). Do we need to adjust for interim analyses in a Bayesian adaptive trial design? *BMC Medical Research Methodology*, *20*(1), 1–9.

Simmons, J., Nelson, L., & Simonsohn, U. (2012). A 21 word solution. *Dialogue*, *26*(2), 4–7.

Taper, M., & Lele, S., (2004). *The nature of scientific evidence: Statistical, philosophical, and empirical considerations*. Chicago, IL: University of Chicago Press.

Wasserstein, R., & Lazar, N. (2016). The ASA's statement on *p*-values: Context, process and purpose. *American Statistician*, *70*(2), 129–133.

Wasserstein, R., Schirm, A., & Lazar, N. (2019). Moving to a world beyond "$p < 0.05$". *American Statistician*, *73*(S1), 1–19.

Wellek, S. (2017). A critical evaluation of the current "*p*-value controversy". *Biometrical Journal*, *59*(5), 854–872.