The New England Journal of Statistics in Data Science Volume 1, 147–148 (2023)
DOI: https://doi.org/10.51387/22-NEJSDS4REJ

Rejoinder of "Four Types of Frequentism and Their Interplay with Bayesianism"☆

James BERGER

Our thanks to all the discussants for their enlightening comments and valuable perspectives.

RESPONSE TO LUIS PERICCHI

Pericchi's Table 1 and Figure 1 are interesting, in that they indicate that the empirical error in testing (which is called fdr, therein) is more sensitive to π_0 than to β. This is also clear from the odds expression in equation 18 in the paper; the odds change more rapidly with changes in π_0 (the derivative of the log odds with respect to π_0 is $1/[\pi_0(1-\pi_0)]$) than with changes in β (the derivative of the log odds with respect to β is $1/\beta$). So refusal to even consider π_0 is questionable.

Indeed, when π_0 is unknown, it is natural for a Bayesian to treat it as just another unknown to be given a prior distribution. To frequentists, however, it is more common to estimate π_0 via, say, empirical Bayes.

Improving the $-e\,p\,\log p$ bound on a Bayes factor is certainly a worthy goal, and we wish Pericchi success in this endeavor.

RESPONSE TO JUDITH ROUSSEAU

Rousseau's concern about the lack of precision in the notion of empirical frequentism is understandable, since we purposely avoided trying to be precise, to allow for flexibility. She does make the helpful and clarifying distinction that, however it is defined, empirical frequentism should be based on a sequence of observable events, rather than a sequence of unobservable events. In hypothesis testing for instance, 'rejections' are observable events, so studying what happens under 'rejections' is compatible with empirical frequentism. But basing the evaluation on a series of unobservable events, such as the set of all true null hypotheses (the series of events used to define Type I error), would not qualify as empirical frequentism.

This is also complicated by the fact that empirical frequentism imagines that one learns the truth for the considered events, e.g., learns which of the null hypotheses are true in the sequence of rejected events. Sometimes this is somewhat realistic, in that rejections are ideally followed by efforts at replication. But we could never learn which nulls were true in the set of acceptances, so Type I error could not be determined from the series of real experiments.

Rousseau shows that one can make some rather strange and unhelpful empirical frequentist statements, reinforcing that care in the definition is needed.

Rousseau mentions E-values and states that $\log E$ might well have an empirical frequentist justification. That would be nice because E-values lack a procedural frequentist justification. An E-value, $E(x)$, satisfies the condition $P(1/E(x) \leq \alpha \mid H_0) \leq \alpha$, so the procedure "reject if $1/E(x) \leq \alpha$" does have the procedural frequentist property of having Type I error controlled at level α. But reporting $E(x)$ itself has no obvious procedural frequentist justification. (The situation is exactly the same as with a p-value: $P(p(x) \leq \alpha \mid H_0) = \alpha$, but directly reporting p does not have any procedural frequentist justification.)

RESPONSE TO AAD VAN DER VAART

Van Der Vaart starts out with a fun journey detailing his personal impressions of the various frequentist types. Everything he says here is sensible. I particularly liked the statement that an empirical frequentist is a practicing statistician, while a procedural frequentist is a theoretician, and that both have value. The comments on consistency are also nice; indeed, consistency does not exactly fit the definition of procedural frequentism. The suggestion that one probably needs more refinement in the 'types' of frequentism, such as making empirical Bayes its own 'type' certainly has merit; this is reiterated later in the discussion, when referring to multiple testing.

In regards to ordinary testing, Van Der Vaart notes that there are possible empirical frequentist targets other than the empirical false discovery rate. He mentions two, the fraction of incorrect rejections amongst all true nulls, and the fraction of incorrect rejections amongst all tests, and notes that both are bounded by α. The first is just the Type I error and we would argue that this is not a valid empirical frequentist target, as it is not based on what happens with observables. The second is a valid empirical frequentist target, but not a reasonable one; why normalize the incorrect rejections by all tests N, rather than just the rejections?

☆Main article: https://doi.org/10.51387/22-NEJSDS4.

In the discussion of multiple testing in the paper, note that each E_i in the sequence E_1, E_2, $\ldots$ is itself a multiple testing scenario, so we are looking at a sequence of different multiple tests; it is this sequence of different multiple tests that is evaluated according to empirical frequentism. Bonferroni and regular FDR are both procedural frequentist properties, computed under the null hypothesis, so the goal was to study the empirical frequentist performance of such reports in repeated use. As with ordinary testing, it does not seem to be possible find a sensible empirical frequentist measure that avoids involvement of the prior probabilities of the hypotheses. As Van Der Vaart notes, such involvement is clearly feasible in situations where π_0 can be estimated.

Accepted 18 July 2022

James Berger. Department of Statistical Science, Duke University, USA.
E-mail address: berger@duke.edu