The New England Journal of Statistics in Data Science Volume 1, 145–146 (2023)
DOI: https://doi.org/10.51387/23-NEJSDS4C

Discussion of: Four Types of Frequentism and Their Interplay with Bayesianism, by J. Berger[☆]

Judith ROUSSEAU

Jim Berger proposes an interesting review of different ways of addressing the problem of error reporting from a frequentist point of view and their connections to Bayesian ways of thinking. In a way this paper echoes Neyman (1977) [2] – at least as far as testing is concerned. Jim Berger has repeatedly made major contributions on the questioning of what makes a relevant measure of uncertainty or reported error and again this article is thought provoking.

Interestingly Neyman in Neyman (1977) [2] justifies (or advocate) the empirical frequentist criteria or error measures, although the Neyman–Pearson is defined as a procedural frequentist approach (in Jim Berger's terminology). If I agree with Jim Berger's point that the justification of the Neyman-Pearson procedure from an empirical frequentist point of view is not fully convincing, I don't quite agree with his arguments. This might be due to the interpretation of the definition of empirical frequentism and a difficulty for me is making sense to this definition which is quite vague.

Let us consider the type I error in a test of a simple null hypothesis versus a simple alternative hypothesis (or not simple, it does not really matter although in the latter the definition of type I and type II errors can be debatable). As in the paper, consider a series of tests with nominal type I error α and power β. One problem with reporting the type I error only is that it provides a very partial picture of the error (it provides no information if the true distribution is not in the null). Following the example of Section 3.2.1 of the paper and recalling that $\alpha = P_{H_{0i}(\mathcal{R}_i)}$ is the type I error for each experiment. Hence it only makes sense to report it when H_{0i} holds (i.e. $\theta_i = \theta_{0i}$) in which case we have

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\theta_i=\theta_{0i}}(\alpha - \mathbb{1}_{x_i\in\mathcal{R}_i}) \to 0 \qquad (1)$$

in probability as soon as the experiments are independent (or more generally as soon as a weak law of large numbers is valid). This is what is suggested in Neyman (1977) [2], pages 108–109. What is not satisfying in (1) is that the reported error only makes sense when non observable events ($\theta_i = \theta_{0i}$) occur. However since

$$\liminf_N \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\theta_i=\theta_{0i}}(\alpha - \mathbb{1}_{x_i\in\mathcal{R}_i}) \geq 0,$$

the type I error α can still be viewed as valid from an empirical frequentist view point, but obviously is much less interesting in the latter inequality. It appears very limited as an accuracy measure. In the theory of minimax estimation the typical risk function for a test is the sum of the type I and type II errors: $\alpha + 1 - \beta$. Interestingly this quantity suffers from the same drawback as the type I error:

$$\frac{1}{N}\sum_{i=1}^{N}[\mathbb{1}_{\theta_i=\theta_{0i}}(\alpha - \mathbb{1}_{x_i\in\mathcal{R}_i}) + \mathbb{1}_{\theta_i=\theta_{1i}}(1-\beta - \mathbb{1}_{x_i\notin\mathcal{R}_i})] = 0,$$

and needs a reporting strategy which depends on non observables for a long run justification, although the following inequality holds true:

$$\begin{aligned}\alpha + 1 - \beta &\geq \frac{1}{N}\sum_{i=1}^{N}[\mathbb{1}_{\theta_i=\theta_{0i}}\alpha + \mathbb{1}_{\theta_i=\theta_{1i}}(1-\beta)] \\ &\geq \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\theta_i=\theta_{0i}}\mathbb{1}_{x_i\in\mathcal{R}_i} + \mathbb{1}_{\theta_i=\theta_{1i}}\mathbb{1}_{x_i\notin\mathcal{R}_i} + o(1).\end{aligned}$$

The above inequality clearly shows the limit of reporting $\alpha + 1 - \beta$: in the long run if the proportion of null and of alternatives are of the same order then

$$\frac{1}{N}\sum_{i=1}^{N}[\mathbb{1}_{\theta_i=\theta_{0i}}\alpha + \mathbb{1}_{\theta_i=\theta_{1i}}(1-\beta)] \approx \frac{\alpha+1-\beta}{2},$$

which is significantly smaller than $\alpha+1-\beta$. The same holds if $\alpha \approx 1-\beta$ and actually reporting $\max(\alpha, 1-\beta)$ is closer to the lower bound. The same reasoning holds for the multiple testing problem under the Bonferroni correction. Again it is not clear to me that the problem with reporting α comes from the reported error not being empirical, but rather that it only makes sense when the null is true, which is often interpreted wrongly as: *it only makes sense when the null is rejected.*

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

One of the issues I have with the notion of empirical frequentism is, as I said, that it is quite vague: to which extent need the experiments be related or close to replicates? should the θ_i in the various experiments be considered as deterministic or random? Empirical frequentist justification of a reported error depends on how we answer these questions. In particular the posterior risk can be validated under the assumption that the θ_i are random and come from the (same) prior distribution. The targets suggested in Eqs. (3.1), (3.3) or (3.7) of the present paper clearly aim at giving an error measure when the null is rejected (i.e. given that the null is rejected) but this inevitably requires to model at least the probability that the null is verified and typically also the distribution under the alternative, which somewhat involves a prior distribution, say π_0 and/or f_1. Reporting error in a statistical test has long been a subject of debate, and much more than in other inference problems (estimation, confidence/credible regions etc) is still largely unresolved. There has been recent growing interests on E-values, as measures of accuracy in a testing procedure, as in Shafer (2021) [3] or Grunwald et al. (2020) for instance, which are strongly related to Bayes factors (BF) (see Grunwald et al. (2020) [1]). A difficulty with Bayes factors or E values – although some new results and propositions are made towards answering it for the latter – is that their scale is not known. Jim Berger shows that they do not enjoy an empirical frequentist justification either. But what about the log Bayes factors? In the safe test approach, Grunwald et al. (2020) [1] relates $\log E$ (which can be viewed as a log-BF) to the Kullback–Leibler divergence between *the alternative* and the null (roughly speaking), when studied under the alternative, which gives an empirical frequentist justification to $\log E$. The fact that E (or BF) does not have an empirical frequentist justification while $\log E$ could have, makes me wonder about the usefulness of considering empirical frequentist justifications.

FUNDING

The author has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 834175).

Accepted 16 August 2022

REFERENCES

[1] Grünwald, P., de Heide, R. and Koolen, W. M. (2020). Safe testing. In *2020 Information Theory and Applications Workshop (ITA)* 1–54. IEEE.

[2] Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese* **36** 97–131. https://doi.org/10.1007/BF00485695. MR0652325

[3] Shafer, G. et al. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **184**(2) 407–431. https://doi.org/10.1111/rssa.12647. MR4255905

Judith Rousseau. Department of Statistics, University of Oxford, UK. E-mail address: rousseau@ceremade.dauphine.fr