The New England Journal of Statistics in Data Science Volume 1, 138–141 (2023)
DOI: https://doi.org/10.51387/22-NEJSDS4A

Frequentism☆

Aad VAN DER VAART

Abstract

Discussion of "Four types of frequentism and their interplay with Bayesianism" by Jim Berger.

While the title of Berger's paper speaks of "types of frequentism", the introduction quickly makes it personal and promises a discussion of "types of frequentists". It made me immediately curious to learn my type, if any. Although I am quite pragmatic in these matters, and easily happy to play the 46.657th type of Bayesian, being a "frequentist" part of the time, I would consider a possibility.

Type III is out. Nobody wants to be this type. This is the type who misuses p-values. Berger is done with them in half a page.

Type IV, *conditional frequentism*, does not quite live up to Berger's promise of "ordering them from best to worst". He actually seems sympathetic to this type, but notes that finding good conditioning variables is difficult and the type is therefore limited to restrict their activities to a small spectrum of situations. For a moment Berger even seems to portray objective Bayesianism, which also gives conditioning, as a poor man's substitute to being Type IV frequentist. As an additional connection, I would refer to recent work (e.g. [11]) on *post-selection inference*, which tries to make confidence statements following model selection conditional on data. It is agreed that such ideas are interesting, but usually complicated.

Type I, *empirical frequentism*, is in for me. But it is not a discerning type. Berger argues that the principle set forth here is so natural that any Bayesian should belong to this type. In vague terms, persons of this type accept that if they repeatedly make a numerical claim on the accuracy of their statistical procedure, then they would lose their trustworthiness if the actual accuracy of their claim turned out to be wrong "on the average in the long run". It seems to me that the vague wording is what makes the principle so easy to accept. Berger's example of a statistical procedure is a confidence set $\{\theta \in C(X)\}$, with a percentage of correctness as accuracy claim, and the would-be Type I frequentist is then judged by the average frequency of coverage given a large number of confidence sets computed "in practice". This multitude of computed confidence sets need not be repetitions of a similar situation, making the principle stronger than a simplistic textbook description, but the principle is deliberately vague in not specifying how the θ and the X are obtained in the sequence of uses. The textbook frequentist may think of experiments when only the X is random and θ is fixed as a "true value", while the Bayesian may think of both θ and X being random, produced on a case-by-case basis, or with the θ repeatedly sampled from a prior, in the empirical Bayes formulation. In fact, no model or assumption on the origin of the multiple (θ, X) underlies the Type I frequentist principle, by its definition. The Type I frequentist is engaged in a sequence of "practical uses" of a statistical procedure and then does not want to be caught having made claims about frequencies that turn out to be untrue "in the long-run".

While everybody is a Type I frequentist, I must probably confess to be a Type II frequentist too, a *procedural frequentist*. Berger explains that I am in good company, because even though "less compelling", the principle includes "consistency" and even to a Bayesian this should be reasonable.

Actually I dislike the frequentist label, although I am occasionally categorised as "one of the consistency people" and when presenting theoretical findings on posterior distributions, often feel forced to use the term "frequentist Bayes", by lack of a better term. Berger defines "procedural frequentism" as the "evaluation of statistical procedures according to their frequentist properties, defined as properties that would arise from repeated imaginary application of the procedure to a specific problem". So while a Type I frequentist is a practising statistician, a Type II frequentist is a theoretician. There is no obstacle to being both, because these are not "types", but different activities. Certainly every Type I statistician will be interested in the "evaluation of statistical procedures". A measure of quality that is very relevant for the Type I frequentist, is the frequentist behaviour of their claimed accuracy levels. Given a model that allows to produce a sequence of pairs (θ, X), the Type I frequentist's desire to be correct on-the-average-in-the-long-run, can be proved to be fulfilled or not on this sequence. While the Type I frequentist is working in practice, the Type II frequentist can prove the accuracy without actually generating this sequence, or even imagining generating this sequence, by evaluating probabilities and expectations.

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

What I dislike about the frequentist label, is the unnecessary demand to think of a probability as a long-run frequency. One can perfectly think of probability without introducing imaginary sequences of averages converging to some number. Is this not what we teach (or should teach) students in their probability class? It is a level more abstract than the practicality of type I frequentists, but if our aim is to investigate theoretical properties, such abstraction is desirable. In this sense I don't want to be a Type II *frequentist*.

I would also comment that the theoretical study of "consistency" is not necessarily linked to repetitions. The most common case, where the data is a sample of size n and $n \to \infty$, does include repetitions, but they are actually not of Type II nature, because they happen within the procedure under investigation, and are not repetitions of the procedure. In the usual consistency theory our aim would be to investigate a procedure at a given n, which we could do by computing probabilities and expectations (in my view preferably not viewed as limits of averages) at that n or possibly simulations (and then as average of many repeated simulations), also at that n. In theoretical study $n \to \infty$ only to make this investigation tractable. In consistency studies n need not be sample size, but is a measure of the informativeness of the data. By considering the case that the data becomes more informative, we hope to gain insight in the finite n situation. For instance, the role of $n \to \infty$ could easily be played by the noise level tending to zero in an inverse problem, where the output of a system is observed with additive noise or the system is stochastic with intrinsic noise. Asymptotics are relevant if the noise is small, and do not involve frequencies.

Berger's description of the Type II frequentist also contains the reference to a "specific problem". This is an important aspect of his discussion, and it is a source for controversy, but it has nothing to do with frequentism in the sense of repetitions. Decision theory [2] takes as input a set of data distributions indexed by a parameter, and then for a given loss (or utility) function produces a risk function as a function of the parameter. This is the basis for the evaluation of statistical procedures as in the Type II principle, although it uses an expectation and not an average. Bayesians may add a prior over the parameter space. Non-Bayesians must deal with a full risk function. Asymptotic study of risk functions may yield insight [7, 13] without involving frequencies or repetitions.

It is viewed by some as a weakness that a risk function does not satisfy the likelihood principle, as it takes an expected value under a model and does involve potential observations that have not realised. There is plenty of discussion of this in the literature, too much to add to it here. My feeling is that the controversy often arises from having different aims. If one aims to *evaluate* procedures, then first one has to agree on a set of rules.

Decision theory is the basis of one set of rules, but even if this is accepted as a proper framework, one must still agree on the description of the "specific problem". The typical Bayesian framework consists of a generative model for the observed data, with a hierarchy of steps. For frequentist-Bayesian analysis (for lack of better terminology) within this framework, we have to draw a line in the hierarchy between the steps that are considered the prior and the steps that are considered to have really occurred in the real world. We may put the line above the top of the hierarchy, leaving no priors steps. Then the parameter was generated in the real world (from a known prior) and the Type II analysis will conclude that the Bayesian method (for instance a credible set, quoted with its credible level) works fine. We are in a situation that a statistician of any Type will agree that Bayes's rule is all that is needed. Berger's Type I frequentist principle, with its insistence on performance in practice, will be satisfied, because the prior describes the real world.

However, usually there will be a prior part to the hierarchy. Then frequentist-Bayesian-consistency analysis will ask the question whether, for any possible parameter value produced at the prior level, if that parameter gave the true state of the world, the Bayesian procedure satisfies the Type I frequentist principle. This Type II style question may not be so easily answered. Finite sample analysis is notably difficult, and even asymptotical insights are incomplete. In particular, we still have only a starting knowledge of Bayesian credible sets in high-dimensional parameter spaces or with high-dimensional data. We know credible sets may be misleading (see for instance Figure 1 in [6]), but there are also encouraging results (e.g. [10, 9, 12]).

In the empirical Bayes setting, mentioned by Berger in his discussion of Type I frequentism, the Bayesian hierarchy contains a layer in which for every individual observation a parameter (or latent variable) is generated: the likelihood becomes a mixture distribution. The line can be drawn above the distribution of these latent variables (the mixing distribution), or just below it, giving the so-called structural and incidental versions of the model. Which of the two is chosen makes a difference for the Type II analysis, which seems more convincing if carried out for the incidental version of the model, as this makes fewer assumptions. This was indeed the point of view of [8], which uses the mixing distribution as a working hypothesis. The Type I criterion, set by Berger, averages over the latent variables and seems in the spirit of the incidental version too.

The empirical Bayes setting is intriguing, and perhaps deserves its own Type. A textbook frequentist may adopt as a working hypothesis that the parameters were generated from a distribution, and could call on (e.g.) maximum likelihood for estimation of this distribution, as well as other parameters of the generative process. A Bayesian will see this distribution as just one part of the generative hierarchy, and perform a full Bayesian analysis or apply maximum likelihood as a computationally more efficient method to estimate hyper parameters. It seems that the resulting procedures work well in the Type I and Type II frameworks

in both the structural and incidental versions of the model (e.g. [5]). Empirical Bayes type of thinking can also bring multiple practical uses of statistical procedures, for instance in big data settings, together in a single analysis. Frequentist and Bayes may come together in procedures that estimate a prior, possibly even an nonparametric one. Which Type does this belong to?

If Type II analysis can start from decision theory, then considering type 1 test error should be fine. Type 1 test error is just one arm of the risk function, and we can use it for evaluation according to whichever rules we can agree on. In Section 2.2.5 Berger also calls it "useful" and together with type 2 test error "a key quantity to consider". However, the main message in his discussion is that it does "not satisfy the empirical frequentist principle". This seems to be in disaccord with the fact that we previously decided that everybody is a Type I frequentist. Can we both acknowledge that type 1 test error is a "key quantity" and that it does not satisfy a desirable first principle?

Half of Berger's paper is about testing. My understanding of his arguments is that the main trouble with the type 1 test error is not that it "does not satisfy" the Type I frequentist principle, but that it "cannot or should not be used" as a number to report the accuracy of a test. That seems obvious. It is just as true that the risk function at a specific parameter value cannot serve as a report of accuracy in a specific *practical* situation. However, it does not help me, that part of Berger's criticism is that "it is not unusual for people to interpret a level α as a surrogate", for what Berger feels should be a measure of accuracy. I am unconvinced that my medical colleagues do not understand the true meaning of α. They are smart, and it is not so difficult in the first place. However, it is an ongoing debate that already has raised too much emotion and I shall not add to it here.

Berger is a Bayesian. That is no news. I highlight it, because in his discussion of testing, Berger slips without much warning in the Bayesian mode. The warning comes in a clause in parentheses at the end of the first paragraph of Section 3: "(when we are incorporating Bayesian concepts)". Actually, Berger will be incorporating Bayesian concepts from thereon, all the time.

The question is what to report as "accuracy" when performing a statistical test, or multiple such tests. Berger's first answer is that it should be on-the-average-in-the-long-run the (average) probability $\Pr(H_0|\, H_0 \text{ is rejected})$. For textbook frequentists this quantity does not make sense: they will chastise their students for even thinking that there is something random about the null hypothesis, and so the expression $\Pr(H_0|\, H_0 \text{ is rejected})$ is illegal. In the Bayesian world, the probability arises naturally through Bayes's rule from the (Bayesian) prior probability $\Pr(H_0)$ of H_0 being true, and the ("frequentist") probabilities of the first and second kind, $\alpha = \Pr(H_0 \text{ is rejected}|\, H_0)$ and $1-\beta = \Pr(H_0 \text{ is rejected}|\, H_1)$. Formula (5) in Berger shows that $\Pr(H_0|\, H_0 \text{ is rejected})$ does arise as the limit of the (empirical) false discovery rate (FDR), when the tests would be repeated indefinitely. The empirical FDR is the fraction of rejected correct null hypotheses out of all rejected null hypotheses, and is not an observed quantity and so cannot be quoted as a measure of accuracy. Is that why Berger writes that it has a Type II frequentist but not a Type I frequentist justification?

Berger seems to have made the choice that $\Pr(H_0|\, H_0 \text{ is rejected})$ is the target for accuracy, and then quoting just α is miserable, and Type I invalid. But it seems that Berger's Type I principle leaves free the choice of *what* is to be correct on-the-average-in-the-long-run. If I evaluate the Type I validity of reporting α by one of the fractions N_0^r/N_0 or N_0^r/N, for N_0^r the number of correct null hypotheses rejected and N_0 the number of correct null hypotheses out of N test, then in both cases the limit is smaller than α. Can I then not conclude that (conservative) Type I validity is satisfied? Berger wants to quote $\Pr(H_0|\, H_0 \text{ is rejected})$, but it is impossible to do so, even on-the-average-in-the-long-run, without being a Bayesian and having a prior.

In the case of multiple testing, FDR seems precisely the relevant target, but in his discussion of multiple testing Berger changes the desired average measure of accuracy into the fraction of times that the sequence of tests rejects at least one correct null hypothesis out of the number of times that the sequence rejects at least one time. Berger's analysis then concludes that the Bonferroni procedure, which performs the individual tests at level α/m if m tests are performed, does only satisfy the Type I frequentist principle in the situation that $\Pr(H_0)$ is small and "is as bad as it can be" if $\Pr(H_0)^m \to 1$, as $m \to \infty$. But why is Bonferroni evaluated on a criterion that it was never meant to adhere to? The correct frequentist calculation is that the probability that the Bonferroni procedure rejects one or more correct null hypotheses is $\leq 1-(1-\alpha/m)^m \approx \alpha$. In his calculation (7), Berger has smuggled in the Bayesian quantity $\Pr(H_0)$, and changed the rules of the game by taking the probability relative to another one.

I don't want to imply that reporting α when performing a Bonferroni procedure is very informative about the data, but this is not different from reporting α when performing a single test. The Bonferroni method is simple and has the advantage of *strong control*: it is valid no matter the configuration of true and false hypotheses. As a sanity check, it makes perfect sense. It is definitely more useful than the claims by some Bayesians that their approaches "automatically" correct for multiple testing, put forward in the early days of genomics.

For many applications (like GWAS) controlling the family-wise error rate, as does the Bonferroni procedure, is considered unnatural and FDR is the standard. "Rejections" become "discoveries", which will be subjected to further scrutiny, and we wish to minimise the fraction of false

discoveries. The Benjamini-Hochberg (BH) procedure controls the FDR at level α in the strong sense of any configuration of true and false hypotheses. Therefore, it seems to me that this procedure with quoted accuracy α would be Type I and Type II frequentist valid, if interpreted in the way as was meant.

Interestingly, the BH procedure was developed in a frequentist setting, but it can be viewed as empirical Bayesian. We assume that $\Pr(H_0)$ makes sense, then obtain a data-dependent error probability $\Pr(H_{0,i} | P_i \leq x)_{|x=p_i}$ using Bayes's rule and finally replace the unknown $\Pr(H_0)$ and distribution of the p-values P_i by estimates ([4]). This is intriguing, as are histograms of observed p-values in the multiple testing situation, and what we can conclude from them. Not much of this intriguing interplay of frequentist and Bayesian reasoning seems to be captured in the Type I–IV discussion, but in my view it indicates that there is something valid about the procedure and its associated number α. Recent empirical Bayes and full Bayes testing procedures controlling FDR were considered in [1, 3].

In his discussion of data-dependent error probabilities, Berger returns to the Bayesian quantity $\Pr(H_0 |\, H_0$ is rejected) as the target for accuracy. He considers statistical procedures that reject the null hypothesis if a p-value is smaller than a pre-given number α, and then investigates the validity of quoting some transformation $\alpha(p)$ as the accuracy. Type I frequentist validity then holds if $\mathrm{E}\big(\alpha(p) | \, 0 \leq p \leq \alpha\big) = \Pr(H_0 |\, H_0$ is rejected). The "obvious" accuracy measure, which works, is $\alpha(p) = \Pr(H_0 | \, p)$, but this Bayesian quantity has the problem that it depends on the unknown $\Pr(H_0)$ and the unknown density f_1 of p given H_1. Reporting the p-value $\alpha(p) = p$ as measure of accuracy comes out as "terrible" from the Type I point of view. Berger discusses $\Pr(H_0) = 1/2$ and known f_1, a certain upper bound on f_1 or using odds ratios, but in my feeling the conclusion is that there is no real solution here.

And this concerns just the case of simple hypotheses. Berger refers to his earlier papers for composite hypotheses.

Perhaps the conclusion is that testing is just a way of behaving, in case we are forced to choose between alternatives? We can shield ourselves from making a wrong, harmful decision, but we cannot say how accurate we are. If there really are things like $\Pr(H_0)$, then it is a different story, and we go the Bayesian way, but setting $\Pr(H_0) = 1/2$ as a default has its own problems. If we do many tests simultaneously, it is a different story too. Then we may learn $\Pr(H_0)$ and other unknowns from the data, and can make statements on the "average H_0". In my feeling this empirical Bayes approach, also for estimation, still remains to be fully explored.

FUNDING

The research leading to these results is partly financed by a Spinoza prize awarded by the Netherlands Organisation for Scientific Research (NWO).

Accepted 16 August 2022

REFERENCES

[1] Abraham, K., Castillo, I. and Roquain, E. (2022). Empirical Bayes cumulative ℓ-value multiple testing procedure for sparse sequences. *Electron. J. Stat.* **16**(1) 2033–2081. https://doi.org/10.1214/22-ejs1979. MR4415394

[2] Berger, J. O. (1985) *Statistical decision theory and Bayesian analysis* second ed. *Springer Series in Statistics*. Springer-Verlag, New York. https://doi.org/10.1007/978-1-4757-4286-2. MR804611. https://doi-org.tudelft.idm.oclc.org/10.1007/978-1-4757-4286-2.

[3] Castillo, I. and Roquain, E. (2020). On spike and slab empirical Bayes multiple testing. *Ann. Statist.* **48**(5) 2548–2574. https://doi.org/10.1214/19-AOS1897. MR4152112.

[4] Efron, B. (2010) *Large-scale inference. Institute of Mathematical Statistics (IMS) Monographs* **1**. Cambridge University Press, Cambridge. Empirical Bayes methods for estimation, testing, and prediction. https://doi.org/10.1017/CBO9780511761362. MR2724758. https://doi-org.tudelft.idm.oclc.org/10.1017/CBO9780511761362.

[5] Jiang, W. and Zhang, C. -H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37**(4) 1647–1684. https://doi.org/10.1214/08-AOS638. MR2533467.

[6] Knapik, B. T., van der Vaart, A. W. and van Zanten, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.* **39**(5) 2626–2657. https://doi.org/10.1214/11-AOS920. MR2906881.

[7] Le Cam, L. (1986) *Asymptotic methods in statistical decision theory. Springer Series in Statistics*. Springer-Verlag, New York. https://doi.org/10.1007/978-1-4612-4946-7. MR856411. https://doi-org.tudelft.idm.oclc.org/10.1007/978-1-4612-4946-7.

[8] Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35** 1–20. https://doi.org/10.1214/aoms/1177703729. MR163407.

[9] Sniekers, S. and van der Vaart, A. (2020). Adaptive Bayesian credible bands in regression with a Gaussian process prior. *Sankhya A* **82**(2) 386–425. https://doi.org/10.1007/s13171-019-00185-0. MR4136240.

[10] Szabó, B., van der Vaart, A. and van Zanten, H. (2015). Honest Bayesian confidence sets for the L^2-norm. *J. Statist. Plann. Inference* **166** 36–51. https://doi.org/10.1016/j.jspi.2014.06.005. MR3390132.

[11] Taylor, J. E. (2018). A selective survey of selective inference. In *Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. IV. Invited lectures* 3019–3038. World Sci. Publ., Hackensack, NJ. MR3966521.

[12] van der Pas, S., Szabó, B. and van der Vaart, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.* **12**(4) 1221–1274. With a rejoinder by the authors. https://doi.org/10.1214/17-BA1065. MR3724985.

[13] van der Vaart, A. (2002). The statistical work of Lucien Le Cam. *Ann. Statist.* **30** 631–682. Dedicated to the memory of Lucien Le Cam. https://doi.org/10.1214/aos/1028674836. MR1922537. https://doi-org.tudelft.idm.oclc.org/10.1214/aos/1028674836.

Aad van der Vaart. Delft Institute of Applied Mathematics, Technical University Delft, Netherlands.
E-mail address: a.w.vandervaart@tudelft.nl