Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1

Anytime-valid confidence sequences as a resolution to the Bayesian/frequentist interval debate

Alexander Ly[*], Udo Boehm[*] and Peter Grünwald[*,†]

Within the simple context of $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1)$, Campbell and Gustafson (2023) eloquently raised two issues concerning model-averaged credible intervals for θ based on

$$\pi(\theta \,|\, y^n) = \pi(\theta \,|\, y^n, \mathcal{M}_1)[1 - P(\mathcal{M}_0 \,|\, y^n)] + \delta_{\theta_0}(\theta) P(\mathcal{M}_0 \,|\, y^n), \tag{1}$$

which mixes $\pi(\theta \,|\, y^n, \mathcal{M}_1)$ with a point mass at θ_0 weighted by

$$P(\mathcal{M}_0 \,|\, y^n) = \frac{P(\mathcal{M}_0)}{P(\mathcal{M}_0) + \text{BF}_{10}(y^n)[1 - P(\mathcal{M}_0)]}. \tag{2}$$

Here, $P(\mathcal{M}_0) \in (0, 1)$ is a chosen prior model probability for $\mathcal{M}_0$, and $\text{BF}_{10}(y^{(n)})$ the Bayes factor based on $\theta \sim \mathcal{N}(0, g)$ with tuning prior variance $g > 0$, say, $g = 1$, where

$$\text{BF}_{10}(y^n) = \text{BF}_{10}(n, z\,;\, g) = (1 + ng)^{-\frac{1}{2}} \exp\left(\tfrac{ngz^2}{2(1+ng)}\right), \quad z := \sqrt{n}(\bar{y} - \theta_0). \tag{3}$$

The two issues deal with $n \to \infty$ while keeping z fixed, resulting in $P(\mathcal{M}_0 \,|\, z, n) \not\to 0$.

Issue 1: Since $P(\mathcal{M}_0 \,|\, z, n) \not\to 0$, the credible interval derived from Eq. (1) will not converge to the classical confidence interval.

Issue 2: The point mass prohibits the specification of an exact $1 - \gamma$ credible interval for certain $\gamma \in (0, 1)$. The latter can be roughly resolved by defining the $1 - \gamma$ credible interval as the smallest interval that has *at least*, rather than exactly, $1 - \gamma$ posterior probability. The resulting interval will then be well-defined and in potential only a wee bit wider.

Both concerns are bypassed with an anytime-valid confidence sequence (Grünwald, 2023; Howard et al., 2020; Ly et al., 2024b; Pawel et al., 2024; Wagenmakers et al., 2020) that collects all the null values θ_0 for which $\text{BF}_{10}(n, z) \leq 1/\alpha$; for Eq. (3) this confidence sequence is given by

$$\texttt{CS}(1 - \alpha) := \left[\bar{y} - \tfrac{1}{\sqrt{n}} \sqrt{\tfrac{1+ng}{ng} \log\left(\tfrac{1+ng}{\alpha^2}\right)}, \bar{y} + \tfrac{1}{\sqrt{n}} \sqrt{\tfrac{1+ng}{ng} \log\left(\tfrac{1+ng}{\alpha^2}\right)}\right]. \tag{4}$$

Irrespective of the choice of the prior on θ, $\texttt{CS}(1-\alpha)$ will be well-defined for all $\alpha \in (0, 1)$ and $n \in \mathbb{N}$, thus solving Issue 2.

arXiv: TBA
[*]Machine Learning, Centrum Wiskunde & Informatica, Alexander.Ly@cwi.nl; Udo.Boehm@cwi.nl
[†]Mathematical Institute, Leiden University, Peter.Grunwald@cwi.nl

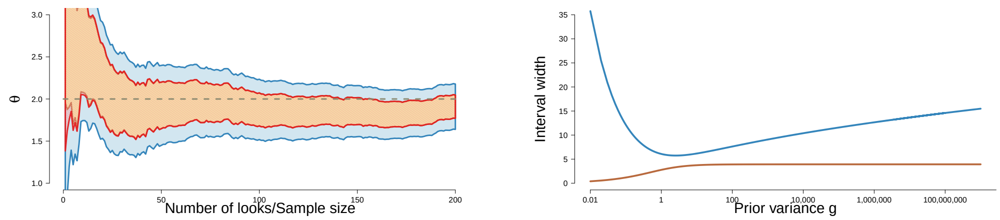

Figure 1: Left panel: Due to the Bernstein-von Mises theorem, the standard (not model-averaged) 95% credible interval (yellow, appearing orange due to the overlap) and the 95% classical confidence interval (red) converge to each other. Both do not cover the true data generating $\theta = 2$ at all times with 95% chance, whereas a 95% anytime-valid confidence sequence (blue) does. Right panel: The 95% $\texttt{CS}(1-\alpha)$ interval width expands as $g \downarrow 0$ and $g \uparrow \infty$, whereas the standard 95% credible interval width asymptotes to the 95% classical confidence interval width as $g \uparrow \infty$. Figures plotted with the `safestats` package (Ly et al., 2024b) in `R`.

$\texttt{CS}(1-\alpha)$ also circumvents Issue 1, as it will cover the data-governing θ with at least $1-\alpha$ chance regardless of when, or whether data collection is stopped. In fact, *without over-inflating the type I error* α, we can reject $\mathcal{M}_0$ and halt data collection at the first data-driven time τ at which θ_0 falls outside $\texttt{CS}(1-\alpha)$. In contrast, the same procedure with a classical confidence interval $\texttt{Conf}(1-\gamma)$ will lead to a type I error converging to one because $\texttt{Conf}(1-\gamma)$ has a width of $\mathcal{O}(\frac{1}{\sqrt{n}})$. The law of iterated logarithm requires a sequentially safe interval width of at least $\mathcal{O}(\frac{1}{\sqrt{n \log \log(n)}})$. Hence, from a sequential analysis point of view it is undesirable for interval estimates to align with classical confidence intervals. In other words, Issue 1 becomes irrelevant; Fig. 1 illustrates the advantage of $\texttt{CS}(1-\alpha)$ being different from $\texttt{Conf}(1-\gamma)$.

The time-uniform coverage guarantee is due to $\text{BF}_{10}(n, z)$ being a so-called E-process (Grünwald et al., 2024; Ramdas et al., 2023), that is, a non-negative stochastic process S for which the following holds:

$$\text{For all } \mathbb{P} \in \mathcal{M}_0 \text{ and any stopping time } \tau : \mathbb{E}_{\mathbb{P}}[S^{\tau}] \leq 1. \tag{5}$$

This defining property suffices for Ville's inequality to hold, which states that

$$\text{For any } \mathbb{P} \in \mathcal{M}_0, \mathbb{P}(\text{For all } n \, : \, S^n \leq 1/\alpha) \geq 1-\alpha. \tag{6}$$

Lemma 3 in Howard et al. (2020) shows that Eq. (6) still holds if n is replaced by a stopping time τ. It is worth noting that not all E-processes are Bayes factors and that not all Bayes factors are E-processes (e.g. Ly et al., 2024a). In case a Bayes factor is an E-process, Eq. (6) implies that for interval estimates it might be better to invert such a Bayes factor instead of using it to update prior model probability as in Eq. (1) (e.g. Grünwald, 2022).

Acknowledgments

The authors would like to extend their gratitude to EJ Wagenmakers for the numerous lively discussions on the Bayesian/frequentist debate over the years.

Funding

This research was supported by two Dutch scientific organisation VENI fellowship grants: "Increasing Scientific Efficiency with Sequential Methods" (VI.Veni.211G.040) to AL, and "Efficient Models of Decision-Making for Assessing Cognitive Processing States" (VI.Veni.201G.045) to UB.

References

Campbell, H. and Gustafson, P. (2023). "Defining a Credible Interval Is Not Always Possible with "Point-Null" Priors: A Lesser-Known Correlate of the Jeffreys-Lindley Paradox." *Bayesian Analysis*, 1(1): 1–16.

Grünwald, P. D. (2022). "Beyond Neyman-Pearson." *arXiv preprint arXiv:2205.00901*.

— (2023). "The e-posterior." *Philosophical Transactions of the Royal Society A*, 381(2247): 20220146.

Grünwald, P. D., de Heide, R., and Koolen, W. (2024). "Safe testing." *Journal of the Royal Statistical Society. Series B (Methodological)*. To appear and with discussion.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2020). "Time-uniform Chernoff bounds via nonnegative supermartingales." *Probability Surveys*, 17: 257–317.

Ly, A., Boehm, U., ter Schure, J., Grünwald, P. D., Ramdas, A., and van Ravenzwaaij, D. (2024a). "Safe Anytime-Valid Inference: Practical Maximally Flexible Sampling Designs for Experiments based on *e*-Values." *In Progress*.

Ly, A., Turner, R. J., Pérez-Ortiz, M. F., Boehm, U., ter Schure, J., and Grünwald, P. (2024b). ***safestats**: Safe Anytime-Valid Inference*. R package version 0.8.8, Maintainer: Alexander Ly (a.ly@jasp-stats.org), https://cran.r-project.org/package=safestats.

Pawel, S., Ly, A., and Wagenmakers, E.-J. (2024). "Evidential Calibration of Confidence Intervals." *The American Statistician*, 78(1): 47–57.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2023). "Game-theoretic statistics and safe anytime-valid inference." *Statistical Science*, 38(4): 576–601.

Wagenmakers, E.-J., Gronau, Q. F., Dablander, F., and Etz, A. (2020). "The support interval." *Erkenntnis*, 1–13.