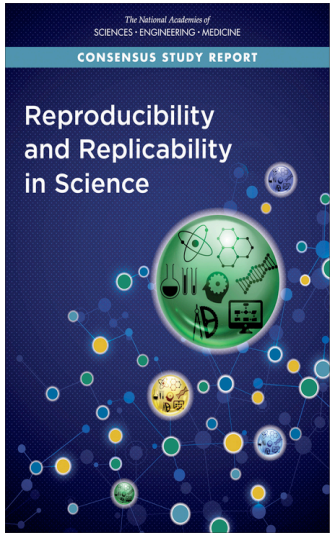
Replicability and Safe, Anytime Valid Inference (SAVI)

Alessandra Salvan and Luigi Pace

Department of Statistical Sciences, University of Padova
Department of Economics and Statistics, University of Udine

joint ongoing work with
Nicola Sartori, University of Padova
Claudia Di Caterina, University of Verona

Replicability and Statistics



Replicability is considered the hallmark (or gold standard, or myth,. . .) of science

- When findings are statistical in nature, i.e., subject to considerable haphazard variability, what counts as a replication is, however, subtle
- Replication of a random experiment is more elusive than replication of a deterministic experiment
- Even more problematic is to assess the uncertainty of conclusions based on limited data in order to provide hints at the generalizability of findings and comparison of results
- Think of estimation of a quantity and an associated estimation interval

Need of statistical reforms?

Romero (2019), Philosophy of science and the replicability crisis, *Philosophy Compass*:

5 | WHAT TO DO?

The big remaining question is normative: What should we do? Because the crisis is likely the result of multiple contributing factors, there is a big market of proposals. I classify them in three camps: *statistical reforms*, *methodological reforms*, and *social reforms*. I use this classification primarily to facilitate discussion. Indeed, there are few strict reformists of each camp. Most authors agree that science needs more than one kind of reform. Nonetheless, authors also tend to emphasize the benefits of particular interventions (in particular, the statistical reformists). I discuss some of the most salient proposals from each camp.

5.1 | Statistical reforms

Statistical reforms

- Attention to replicability requires rethinking statistical procedures
- These are interesting times for researchers working on foundations of statistical inference
- Many proposals of reform in the literature revolve around p-values and associated questionable research practices, QRP, such as p-hacking
- Indeed, replicability of p-values is almost by definition elusive
- To lower the threshold of significance for p-values from, say 0.05 to 0.005, as in Benjamin et al (2018), is not a solution
- Reject p-values altogether and embrace Bayesian inference?

Safe, Anytime Valid Inference (SAVI)

Ramdas, Grünwald, Vovk & Schafer (2023), Game-theoretic statistics and safe anytime-valid inference, *Statistical Science*, forthcoming:

> As emphasized by Johari et al. (2022); Howard et al. (2021); Grünwald, De Heide and Koolen (2023); Shafer (2021); Pace and Salvan (2020), amongst others, we need to go beyond disapproval of peeking, and we instead should give researchers tools to fully accommodate it. The branch of mathematical statistics that enables this, sequential analysis, was brilliantly launched in the 1940s and 1950s by Wald, Anscombe, Robbins, and others. The innovations introduced by Robbins, Darling, Siegmund and Lai included *confidence sequences* that are valid at any and all times and *tests of power one*. But these ideas occupied only a small niche in sequential analysis research until around 2017. Since then, interest has exploded and much conceptual progress has been made in parallel threads, which we attempt to summarize.

Inference with accumulating data

- SAVI is concerned with coherent results of inference with accumulating data (streaming data sets, A/B testing in on-line randomized experiments, drugs surveillance, ...)
- In a sequential environment, if a reliable representation of the true state of nature becomes eventually available, there might be a large reputational penalty from hasty announcements
- Think for instance of estimating the result of an election from early reporting counting areas, where the estimate is made only hours before a winner is declared
- Here, we focus on estimation intervals:

$$\textit{estimate} \pm \textit{uncertainty measure}$$

- Different forms of uncertainty measures are used with different context and aims

Reference example: normal mean with known variance

- Data $y_1, y_2, \ldots$ independent observations of a $N(\mu, \sigma_0^2)$
- With $y^{(n)} = (y_1, \ldots, y_n)$ having sample mean $\overline{y}_n = \sum_{i=1}^n y_i/n$
  - standard frequentist confidence interval

$$I^F_{1-\alpha}(y^{(n)}) = \overline{y}_n \pm z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}$$

  - Bayesian credible interval with prior $N(\eta_0, \tau_0^2)$

$$I^B_{1-\alpha}(y^{(n)}) = \frac{\eta_0/\tau_0^2 + n\overline{y}_n/\sigma_0^2}{1/\tau_0^2 + n/\sigma_0^2} \pm z_{1-\alpha/2}\sqrt{\frac{1}{1/\tau_0^2 + n/\sigma_0^2}}$$

    z_α denotes the α quantile of the $N(0,1)$ distribution

  - The two intervals substantially overlap when τ_0^2 or n is large

Properties of estimation intervals

- Frequentist coverage

$$P_\mu\left(\mu \in I^F_{1-\alpha}(Y^{(n)})\right) = 1-\alpha\,, \qquad 1-\alpha \text{ confidence level}$$

- Bayesian posterior probability

$$P_{\eta_0,\tau_0^2}\left(\mu \in I^B_{1-\alpha}(y^{(n)})|y^{(n)}\right) = 1-\alpha\,, \quad 1-\alpha \text{ credibility level}$$

Deficiencies of (fixed n) estimation intervals

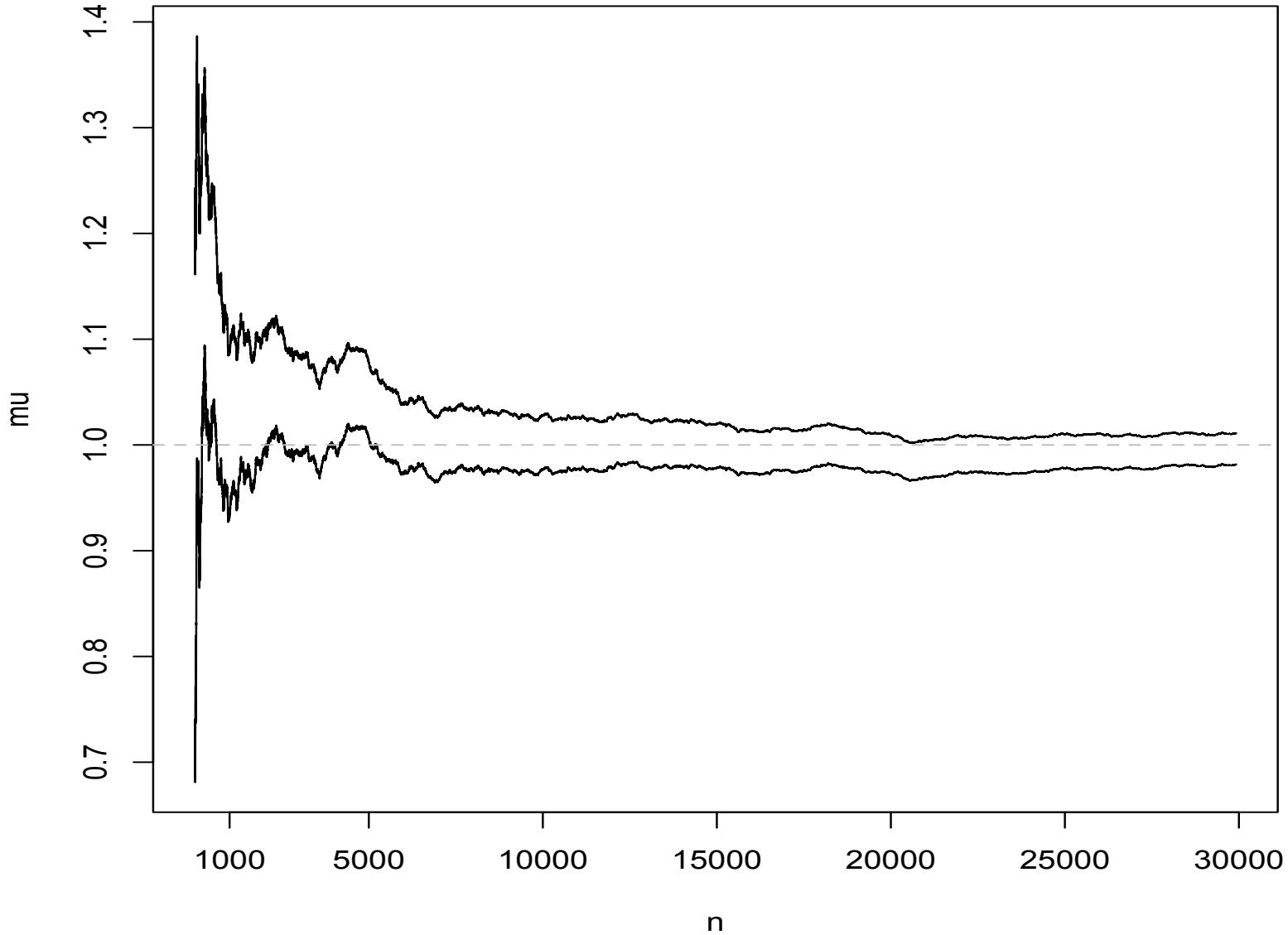
In a sequential setting with optional stopping time n

1. frequentist coverage properties are not satisfied, i.e., confidence level is no more $1 - \alpha$

2. contradictory conclusions (non-overlapping intervals) may be obtained when the sample is enlarged, i.e. when information increases

Non-coverages and contradictions in a sequence $I^F_{1-\alpha}(y^{(n)})$

$\mu = 1$, $\sigma_0^2 = 4$, $1 - \alpha = 0.8$, n from 100 to 30000



Simulation check of non-coverages and contradictions

Empirical percentages of contradictions and non-coverages of $I^F_{1-\alpha}(y^{(n)})$ in 10^4 sequences of samples with size from 10 to 4,000 from $N(0,1)$

$100(1-\alpha)$	90	95	99	99.5
contradictions	51.32	27.35	5.20	2.29
non-coverages	77.79	54.21	18.52	10.86

Fixed n confidence intervals and contradictions

- Even with revised higher levels, the probability that standard confidence intervals at sample sizes n and $n+m$ do not overlap is

$$P_\mu\left(I^F_{1-\alpha}(Y^{(n)}) \bigcap I^F_{1-\alpha}(Y^{(n+m)}) = \emptyset\right) > 0$$

- Therefore the probability is 1 that we can find in the sequence $I^F_{1-\alpha}(Y^{(n)})$ a pair of disjoint intervals, and consequently it is almost sure that we observe a sequence of samples giving rise to contradictory $(1-\alpha)$-level confidence intervals

- With large but finite sample sizes, though $I^F_{1-\alpha}(y^{(n)})$ shrink towards μ as n increases, conflicting conclusions may be reported at various stages of the data acquisition process, with large probability

Confidence sequences

- Confidence sequences introduced by Robbins (1970) remedy both deficiencies

- In the reference normal example often they still have the form

$$CS_{1-\varepsilon}(y^{(n)}) = \textit{estimate}_n \pm \textit{uncertainty measure}_n$$

  where $\textit{estimate}_n = \overline{y}_n$, while $\textit{uncertainty measure}_n$ is such that

$$P_\mu \left(\mu \in CS_{1-\varepsilon}(Y^{(n)}) \,, \quad \text{for every } n \right) \geq 1 - \varepsilon$$

- We call $1 - \varepsilon$ the persistence level; it is guaranteed independently of the stopping rule (anytime validity)

- Persistence level $1 - \varepsilon$ in turn implies that the probability of contradiction is also controlled (Pace and Salvan, 2020):

$$P_\mu \left(\bigcap_n CS_{1-\varepsilon}(Y^{(n)}) = \emptyset \right) \leq \varepsilon$$

Mixture confidence sequences

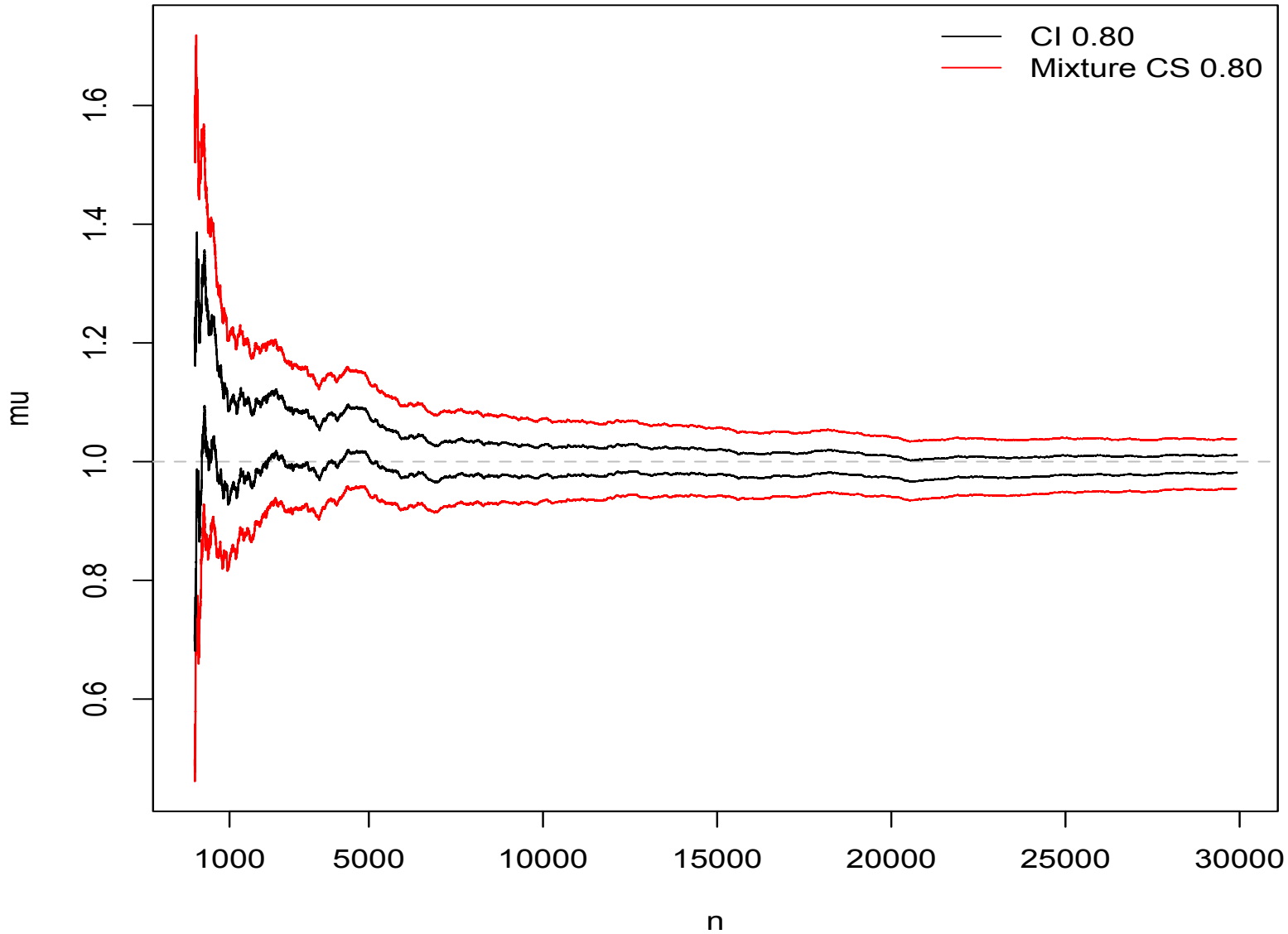
- Ville's inequality (Ramdas et al, 2023, Section 2.5) is the starting point to obtain various types of $CS_{1-\varepsilon}(y^{(n)})$
- For a statistical model with parameter $\theta \in \Theta \subseteq \mathbb{R}^p$ and densities $p_n(y^{(n)}; \theta)$, mixture confidence sequences (Robbins, 1970) have the form

$$CS_{1-\varepsilon}(y^{(n)}) = \left\{ \theta \in \Theta \,:\, p_n(y^{(n)}; \theta) \geq \varepsilon \int_\Theta p_n(y^{(n)}; \theta) \pi(\theta)\, d\theta \right\},$$

  where the weight function $\pi(\theta)$ is a preset probability density over Θ with $\pi(\theta) > 0$ for every $\theta \in \Theta$
- In the reference normal example, a conjugate $N(\mu_0, \tau_0^2)$ weight function gives a mixture confidence sequence with a simple closed form expression

Normal example: mixture $CS_{1-\varepsilon}(y^{(n)})$ vs $I_{1-\alpha}(y^{(n)})$

- $\mu = 1$, $\sigma_0^2 = 4$; $\mu_0 = 0$, $\tau_0^2 = 1$



Mixture confidence sequences and principles of inference

- Mixture CS's are among the rare examples of procedures with a frequentist interpretation that obey the strong likelihood principle, as well as the sufficiency and conditionality principles (see, e.g., Cox & Hinkley, 1974, Section 2.3, for a discussion of the principles of inference)
- Mixture CS's are able to incorporate prior information and have a Bayesian interpretation
- Indeed, when the mixing density $\pi(\theta)$ represents a prior distribution, the posterior with data $y^{(n)}$ is

$$\pi(\theta|y^{(n)}) = \frac{p_n(y^{(n)};\theta)\pi(\theta)}{\int_\Theta p_n(y^{(n)};\theta)\pi(\theta)\,d\theta}$$

and the mixture confidence sequence may be recast as

$$CS_{1-\varepsilon}(y^{(n)}) = \left\{\theta \in \Theta \,:\, \pi(\theta|y^{(n)}) \geq \varepsilon\pi(\theta)\right\},$$

based on the relative belief ratio (cf. Evans, 2016)

Bayesian and evidential properties of mixture CS's

- For every n the credibility level of $CS_{1-\varepsilon}(y^{(n)})$ is at least $1-\varepsilon$ (Pace & Salvan, 2020, Section 3)
- Unlike what happens for the usual $1-\alpha$ credible intervals, the Bayesian probability of non-contradiction of $CS_{1-\varepsilon}(y^{(n)})$ is at least $1-\epsilon$
- Mixture confidence sequences have also an evidential interpretation (Pawel, Ly & Wagenmakers, 2023). Indeed, the ratio

$$\frac{\int_\Theta p_n(y^{(n)};\theta)\pi(\theta)\,d\theta}{p_n(y^{(n)};\theta)}$$

  is a Bayes factor quantifying the strength of evidence against θ relative to the mixture density $\int_\Theta p_n(y^{(n)};\theta)\pi(\theta)\,d\theta$

Summary and generalizations

- CS's and replicability as controlled probability of non-contradiction
- Mixture CS's and inferential principles
- Generalizations:
  - Ramdas et al (2023): test martingales (Ville's inequality); nonparametric settings; robustness to dependence; Howard et al (2021) nonparametric CS's, estimands that vary with n
  - Approximate CS's, also based on pseudo-likelihoods (Waudby-Smith et al, 2021; Pace & Salvan 2020; Pace, Salvan & Sartori, 2023)
  - Computational aspects: efficient updating and applications in generalized linear models (ongoing work with N.Sartori and C.Di Caterina)

Some references

- Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J., Berk, R., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2, 6–10.
- Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman and Hall, London.
- Committee On Reproducibility And Replicability In Science (2019). *Reproducibility and Replicability in Science*. National Academies Press, Washington, D.C. https://doi.org/10.17226/25303
- Evans, M. (2016). Measuring statistical evidence using relative belief. *Comput. and Structural Biotechnology J.*, 14, 91–96.
- Howard, S.R., Ramdas, A., McAuliffe, J. & Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49, 1055–1080.
- Pace, L. & Salvan, A. (2020). Likelihood, replicability and Robbins' confidence sequences. *International Statistical Review*, 88, 599–615.
- Pace, L., Salvan, A. & Sartori, N. (2023). Confidence sequences with composite likelihoods. *Canadian J. Statist.*, 51, 877–896.
- Pawel, S., Ly, A. & Wagenmakers, E.-J. (2023). Evidential calibration of confidence intervals. *The American Statistician*, DOI: 10.1080/00031305.2023.2216239.
- Ramdas, A., Grünwald, P., Vovk V. & Schafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, forthcoming.
- Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41, 1397–1409.
- Romero, F. (2019). Philosophy of science and the replicability crisis, *Philosophy Compass*, 14:e12633.
- Waudby-Smith, I., Arbour, D., Sinha, R., Kennedy, E. H. & Ramdas, A. (2021). Time-uniform central limit theory and asymptotic confidence sequences. *arXiv:2103.06476*.