Safe inference in multi-agent systems

Ben Chugg

December 2025

1 The terrible, horrible, no good, very bad truth about statistics

Traditional statistical methodology is remarkably brittle. For instance, the social scientist's favorite tool, the p-value, only behaves as advertised under a long list of fragile conditions. In particular, it requires that both the sample size and the significance level be fixed in advance of gathering any data. This prohibits any form of optional stopping (deciding to stop the experiment early), optional continuation (deciding to gather more data), or updating evidential standards in response to surprising data.

These problems are not mere mathematical curiosities. In modern applications, these restrictions are almost impossible to respect. Online platforms are running thousands of experiments weekly; clinical trials require monitoring the subjects for unintended adverse side-effects; researchers want to gather more data based on budgetary constraints and the strength of the evidence so far. Such use cases demand continuous, adaptive, and flexible experimentation methods.

There are also deeper conceptual issues. P-values and related tools are defined in terms of the probability of outcomes that *could have happened* but did not. To justify their use, we have to imagine that in every counterfactual scenario the analyst followed the prespecified plan and never engaged in optional stopping, optional continuation, or model tweaking [25, 31]. This is an unrealistically high bar for human analysts, and the growing literature on the "replication crisis" suggests that it is routinely violated in practice [4]. For example, surveys in psychology report that more than half of researchers admit to practices such as selectively reporting significant results or deciding to stop data collection based on interim significance [20].

In recent years there has been a movement to develop novel statistical techniques which do not fall prey to these issues. One influential line of work is *game-theoretic statistics* [29], which builds a foundation for statistical inference on top of betting strategies between two players (the statistician vs nature), formalized in terms of supermartingales, or equivalently, sequences of "betting scores" that behave like wealth in a fair game. The application of these ideas to modern statistics and machine learning has come to be known as "safe, anytime-valid inference" (SAVI) [27].

The past five years have witnessed an explosion of work in this area—including some of my own—developing the theory into practical methodology for adaptive experimentation, online and sequential estimation, and model evaluation. The power and flexibility of these tools is evidenced by their uptake in industry. They have been recently adopted by many companies: Spotify [28], Netflix [22], Eppo [23], Growthbook [15], Abridge AI [24], Vowpal Wabbit (Microsoft) [30], and Adobe [2], to name a few.

2 Resuscitation via e-value

Within SAVI, the central objects are *e-values* and *e-processes* (together, *e-statistics*). E-values are nonnegative random variable whose expectation is at most 1 under a set of distributions $\mathcal{P}$, and e-processes are their sequential analogues. E-statistics have the fruitful interpretation as the wealth of a skeptic betting against $\mathcal{P}$; large values are thus evidence against $\mathcal{P}$ and, crucially, the properties of e-statistics are preserved under arbitrary stopping and many forms of adaptivity and data dependence.

Despite their seeming simplicity, e-statistics have proven remarkably useful across a wide variety of statistical tasks, from FDR control to mean estimation [26, 16]. Among other benefits, they have helped us solve the issues mentioned above by enabling (i) *time-uniform* inference and (ii) *post-hoc* decision-making in situations where they were previously impossible.

Time-uniform guarantees allow an analyst to continuously monitor the data as it arrives, deciding whether to stop or continue the experiment based on the observed results so far. Meanwhile, post-hoc guarantees retain their validity when the analyst changes the significance level or selection rule after seeing the data. That is, one can decide *after the fact* how strict to be, and make truly data-driven decisions.[1]

My work has focused on both time-uniform inference and post-hoc decision-making. I've worked on the underlying theory [11, 12, 13, 7, 32], and on applications to practical problems such as auditing, hypothesis testing, and sequential decision-making [9, 10, 6, 19, 18].

To elaborate briefly on these contributions: I showed how e-statistics can generalize PAC-Bayesian learning theory, unifying the vast majority of previous literature into a singular framework [11]. The techniques of this work proved general enough to be applicable outside of learning theory. They can be leveraged to provide time-uniform concentration of e-processes, resulting in state-of-the-art bounds for the means of random vectors [13] and dimension-free self-normalized inequalities [7]. Such inequalities are the foundation for most regret analysis in contextual bandits, and are thus immediately applicable [1]. We recently extended some of these results to infinite-dimensional Banach spaces [32].

Regarding post-hoc inference, I introduced a framework for studying "admissible" hypothesis tests in this setting [12]—tests which are in some sense unimprovable. We show that all admissible post-hoc procedures are characterized by e-values, once again highlighting the centrality of e-statistics in modern inferential methods. This has spawned several followup works, one of which develops a framework for "asymptotic post-hoc inference." Asymptotic methods are so widely used in practice because they require very few assumptions on the data. Our work, currently a working paper, develops post-hoc inferential methods which are just as widely applicable.

3 E-values on networks

Thus far, the theory of e-statistics has been developed primarily in single-statistician environments. I would like to extend this theory to networked environments, where many agents—whether human or AI—collect data, interact, and respond strategically. Drawing inspiration from Michael Jordan's recent whitepaper [21], this perspective naturally leads to

[1] This may not sound important (we forgive the reader if they are not immediately agog), but has significant results in practice. Consider the following: In the traditional framework, if the significance level is set to 0.05, a p-value of 10^{-6} and 0.04 have precisely the same implications. This is bizarre, seeing as the p-value is ostensibly supposed to be a measure of evidence against the null.

several research directions at the intersection of statistical inference, mechanism design, and multi-agent systems.

The first direction begins with the familiar view of e-values as measures of evidence. Imagine a network in which each node has a local e-statistic, constructed from that agent's data as well as information received from its neighbors. Under what conditions can these local objects be aggregated into a valid, global e-value that certifies claims about the network as a whole?

One can, for instance, view collaborative science as such a network: each lab runs experiments, reports e-values, and (sometimes) conditions its next study on others' reports. Likewise in public health and epidemiology, where different hospitals and labs contribute partial data to an overall prevalence estimate. Or in federated and distributed learning, where many devices monitor performance and/or safety locally and periodically transmit summaries.

It's worth noting that related questions sometimes arise in the study of emergent behavior in distributed systems. For example, in the theory of population protocols and CRNs (e.g., [3, 8, 17]), in distributed consensus and detection [33], and in industrial control networks [14]. One hopes that a new e-value based perspective on networks can either borrow insights from, or add insights to, these existing research areas.

A second, complementary direction views e-values not just as flows of evidence, but as flows of *payoffs* between agents bidding on contracts. Bates et al. [5] recently studied principal–agent hypothesis testing and showed that incentive-compatible contracts correspond to menus of e-values. In other words, they connect *statistical contract theory* directly to e-statistics: if a regulator wants to design a test-based payment or licensing scheme that is incentive compatible, they must be based on e-values.

In reality, there are typically many principals and many agents. Different principals (e.g., regulators, or online platforms, or research labs) offer different contracts, and multiple agents choose (and possibly compete for) which contracts to accept. This suggests extending principal–agent hypothesis testing to multi-agent systems and studying statistical contracts on networks. The goal is to design mechanisms in which an agent's payoff is tied to their reported e-statistic, truthful reporting (and perhaps honest experimental design) forms an equilibrium, and global error rates remain controlled.

This perspective on e-values suggests many specific questions, including: How should we design menus of contracts on a graph consisting of principals and agents? How do competing principals interact: do e-based statistical contracts be designed to avoid arbitrage opportunities or free-riding? How should we account for externalities, where one agent's experiment changes others' opportunities?

To modify an example of Bates et al. [5], consider drug approval as a networked principal–agent problem. Pharmaceutical companies act as agents who run clinical trials, while multiple regulators (e.g., in different jurisdictions) and payers (insurers, healthcare systems) are principals offering contracts. Each regulator specifies a menu of e-value based contracts: if an agent delivers an e-value above some threshold they obtain approval to market their drug. Principals decide which regulators to submit to, how much data to collect, and how much money to spend on their trials. How should the contracts be designed to ensure there's no profitable misreporting or cherry-picking, and error rates are maintained across the system as a whole?

Overall, these directions point toward a broader research program on e-statistics in networked environments. While e-statistics play a fundamental role in single-analyst inference and in basic incentive-compatible games, we hope to extend them as the go-to statistical primitives in multi-agent systems.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[2] Adobe. Experimentation panel. `https://experienceleague.adobe.com/en/docs/analytics-platform/using/cja-workspace/panels/experimentation`, September 2025. Adobe Customer Journey Analytics documentation, last updated 17 Sep 2025, accessed 2025-12-08.

[3] James Aspnes and Eric Ruppert. An introduction to population protocols. *Middleware for Network Eccentric and Mobile Applications*, pages 97–120, 2009.

[4] George C Banks, Steven G Rogelberg, Haley M Woznyj, Ronald S Landis, and Deborah E Rupp. Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology*, 31(3):323–338, 2016.

[5] Stephen Bates, Michael I Jordan, Michael Sklar, and Jake A Soloff. Principal-agent hypothesis testing. *arXiv preprint arXiv:2205.06812*, 2022.

[6] Ben Chugg and Daniel E Ho. Reconciling risk allocation and prevalence estimation in public health using batched bandits. *NeurIPS ML for Public Health*, 2021.

[7] Ben Chugg and Aaditya Ramdas. A variational approach to dimension-free self-normalized concentration. *arXiv preprint arXiv:2508.06483*, 2025.

[8] Ben Chugg, Hooman Hashemi, and Anne Condon. Output-oblivious stochastic chemical reaction networks. In *22nd International Conference on Principles of Distributed Systems*, 2019.

[9] Ben Chugg, Santiago Cortes-Gomez, Bryan Wilder, and Aaditya Ramdas. Auditing fairness by betting. *Advances in Neural Information Processing Systems*, 36:6070–6091, 2023.

[10] Ben Chugg, Peter Henderson, Jacob Goldin, and Daniel E Ho. Entropy regularization for population estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12198–12204, 2023.

[11] Ben Chugg, Hongjian Wang, and Aaditya Ramdas. A unified recipe for deriving (time-uniform) PAC-Bayes bounds. *Journal of Machine Learning Research*, 24(372):1–61, 2023.

[12] Ben Chugg, Tyron Lardy, Aaditya Ramdas, and Peter Grünwald. On admissibility in post-hoc hypothesis testing. *arXiv preprint arXiv:2508.00770*, 2025.

[13] Ben Chugg, Hongjian Wang, and Aaditya Ramdas. Time-uniform confidence spheres for means of random vectors. *Transactions on Machine Learning Research*, 2025.

[14] Brendan Galloway and Gerhard P Hancke. Introduction to industrial control networks. *IEEE Communications surveys & tutorials*, 15(2):860–880, 2012.

[15] GrowthBook. Sequential testing. `https://docs.growthbook.io/statistics/sequential`, 2025. GrowthBook documentation, accessed 2025-12-08.

[16] Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024.

[17] Hooman Hashemi, Ben Chugg, and Anne Condon. Composable computation in leaderless, discrete chemical reaction networks. In *26th International Conference on DNA Computing and Molecular Programming (DNA 26)(2020)*, pages 3–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2020.

[18] Peter Henderson, Ben Chugg, Brandon Anderson, and Daniel E Ho. Beyond ads: Sequential decision-making algorithms in law and public policy. In *Proceedings of the 2022 Symposium on Computer Science and Law*, pages 87–100, 2022.

[19] Peter Henderson, Ben Chugg, Brandon Anderson, Kristen Altenburger, Alex Turk, John Guyton, Jacob Goldin, and Daniel E Ho. Integrating reward maximization and population estimation: Sequential decision-making for internal revenue service audit selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5087–5095, 2023.

[20] Leslie K John, George Loewenstein, and Drazen Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5):524–532, 2012.

[21] Michael I Jordan. A collectivist, economic perspective on AI. *arXiv preprint arXiv:2507.06268*, 2025.

[22] Michael Lindon, Chris Sanden, Vache Shirikian, Yanjun Liu, Minal Mishra, and Martin Tingley. Sequential A/B testing keeps the world streaming netflix part 1: Continuous data. `https://netflixtechblog.com/sequential-a-b-testing-keeps-the-world-streaming-netflix-part-1-continuous-data-cba6c7ed49df`, February 2024. Netflix Technology Blog, accessed 2025-12-08.

[23] Ryan Lucht. What is sequential testing? (with examples). `https://www.geteppo.com/blog/sequential-testing`, July 2024. Eppo blog, accessed 2025-12-08.

[24] Michael Oberst, Davis Liang, and Zachary C. Lipton. Pioneering the science of ai evaluation. `https://www.abridge.com/ai/science-ai-evaluation`, September 2024. Abridge whitepaper, published 2024-09-19, last updated 2025-08-07, accessed 2025-12-08.

[25] John W Pratt. The foundations of statistical inference. *Quarterly of Applied Mathematics*, pages 170–172, 1964.

[26] Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 2025.

[27] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

[28] Mårten Schultzberg and Sebastian Ankargren. Choosing a sequential testing framework — comparisons and discussions. `https://engineering.atspotify.com/2023/03/choosing-sequential-testing-framework-comparisons-and-discussions`, March 2023. Spotify Engineering Blog, accessed 2025-12-08.

[29] Glenn Shafer and Vladimir Vovk. *Game-theoretic foundations for probability and finance*. John Wiley & Sons, 2019.

[30] VowpalWabbit contributors. Vowpal wabbit (version 9.5.0). `https://github.com/VowpalWabbit/vowpal_wabbit/releases/tag/9.5.0`, October 2022. GitHub release, published 14 Oct 2022, accessed 2025-12-08.

[31] Eric-Jan Wagenmakers. A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, 14(5):779–804, 2007.

[32] Justin Whitehouse, Ben Chugg, Diego Martinez-Taboada, and Aaditya Ramdas. Mean estimation in banach spaces under infinite variance and martingale dependence. *arXiv preprint arXiv:2411.11271*, 2024.

[33] Yang Xiao, Ning Zhang, Jin Li, Wenjing Lou, and Y Thomas Hou. Distributed consensus protocols and algorithms. *Blockchain for Distributed Systems Security*, 25:40, 2019.