The New England Journal of Statistics in Data Science Volume 2, 153–154 (2024)
DOI: https://doi.org/10.51387/24-NEJSDS22EDI

Editorial
Game-Theoretic Statistics and Safe Anytime-Valid Inference

Aaditya RAMDAS* and Peter GRÜNWALD

1. INTRODUCTION

It is an exciting time for the subfield that has recently named itself as "game-theoretic statistics and safe anytime-valid inference". Over the last 5 years, there have been many papers on this topic published at all the top journals in statistics, and all the top machine learning conferences, amongst other venues. Further, practitioners are paying attention, and several IT companies have already started implementing anytime-valid inferential tools like power-one sequential tests and confidence sequences as part of their internal pipelines. Meanwhile, we are also aware that a large part of the broader statistical community may be unaware of these advances, or even the terminology that these areas employ, what makes them "game-theoretic", and how the e-values underlying the progress fit into the broader landscape of statistics.

It is in this broader context that we had decided to launch and edit this special issue. Having several papers under a single theme appear together in such an issue allows for the topic to be discussed from philosophical, methodological, theoretical and practical viewpoints, and allows the name of the area to have a recognizable footprint. Luckily for us, several of the leading figures in the broader literature on the topic decided to submit an interesting paper to provide this special issue with an informal stamp of support, with the side benefit that it may now be easier for readers to find a starting point to access the broader literature on the topic. While none of these papers is by itself a survey of the literature, such a recent survey exists if readers need more background [3]. We now briefly describe what the final six accepted papers accomplished, and why they may be interesting for readers to examine further.

2. RESEARCH ARTICLES

[1] **Cost-Aware Generalized α-Investing for Multiple Hypothesis Testing.** The authors address the *online* multiple hypothesis testing problem, where a stream of hypotheses are tested one at a time, and decisions to reject must be irrevocably made in a manner that controls the false discovery rate. They consider a twist to the standard setting, where there are nontrivial data collection costs. Their cost-aware ERO algorithm posits a repeated two-step game between the investigator and nature, and designs α-investing rules in such a way that the α-wealth ends up being a martingale (while in general it could be a supermartingale or submartingale). The authors point to several possible extensions, such as relaxing the current assumption of risk-neutral investigator so as to hedge the risk of α-wealth loss.

[2] **A Safe Hosmer-Lemeshow Test.** The authors prose an e-value based alternative to the Hosmer-Lemeshow (HL) test for evaluating the calibration of probability forecasts for binary events. The test employs online isotonic regression to estimate the calibration curve as a betting strategy against the null hypothesis. The test has power against essentially all alternatives, making it superior to the HL test and resolving an instability problem of the latter. The authors provide both a simulation study and a study based on real-world data. These indicate that, while practically feasible versions of the proposed e-HL test may have reduced empirical power, they can still deliver results comparable to the classical HL test in practice.

[6] **The Anytime-Valid Logrank Test: Error Control under Continuous Monitoring with Unlimited Horizon.** The authors introduce the anytime-valid (AV) logrank test, a version of the logrank test that provides type-I error guarantees under optional stopping and continuation, based on e-values and test martingales. The test allows for cumulative meta-analysis with type-I error control, and the underlying method can be extended to define anytime-valid confidence intervals. The authors demonstrate type-I error guarantees for the test in a semiparametric setting of proportional hazards, show explicitly how to extend it to ties and to handle confidence sequences and indicate further extensions to the full Cox regression model. Using a Gaussian approximation on the logrank statistic, it is shown that the AV logrank test (which itself is always exact) has a similar rejection region to O'Brien-Fleming α-spending but with the potential to achieve 100% power by optional continuation. Although when used for study design, the anytime-valid approach requires a larger sample size, it turns out that the *expected* sample size is competitive by optional stopping.

[4] **Improving Data Analysis by Testing by Betting: Optional Continuation and Descriptive Statis-**

*Corresponding author.

tics. E-value based approaches can invariantly be understood in terms of *testing by betting*. This paper explores two consequences of this interpretation. When testing a statistical hypothesis, is it legitimate to deliberate on the basis of initial data about whether and how to collect further data? Game-theoretic probability's fundamental principle for testing by betting says yes, provided that you do not risk more capital than initially committed. Standard statistical theory uses Cournot's principle, which does not allow such optional continuation. Testing by betting can also help us with descriptive data analysis. To obtain a purely and honestly descriptive analysis using competing probability distributions, we have them bet against each other using the principle. The place of confidence intervals is then taken by sets of distributions that do relatively well in the competition. The author shows that in the simplest implementation, these sets coincide with R. A. Fisher's likelihood ranges.

[5] **E-detectors: A Nonparametric Framework for Sequential Change Detection.** The authors present a general reduction from sequential change detection to that of sequential testing using e-processes. While Lorden had already presented a reduction to sequential testing over 50 years ago, the new reduction is more efficient in theory and practice. The e-detector can perform sequential change detection in many settings in which this was previously not possible, employing the fact that new e-processes have been recently constructed for nonparametric and composite pre-change distributions (for example, using universal inference). The authors present a general detection delay analysis, and a computationally efficient discrete mixture method with components growing logarithmically with sample size. Several open problems are stated, in particular instantiating the general framework in new special cases, and tightening the analysis of detection delay and proving its first or second order optimality.

[7] **Nonparametric E-tests of Symmetry.** The authors design and study several e-values for testing the composite null hypothesis that the distribution of a univariate random variable is symmetric around the origin. While past work fully characterized all possible admissible e-values for testing symmetry, this work goes further by considering three particular nonparametric e-tests and studying their asymptotic (Pitman-type) relative efficiency for Gaussian data. They show that Fisher-type e-tests are efficient, sign e-tests have efficiency $2/\pi$ (like the standard case) and Wilcoxon e-tests have efficiency $3/\pi$ (like the standard case). The authors mention that future work could go beyond the Gaussian model, consider other notions of asymptotic efficiency, or generalize the notion of e-power that is considered (which, as is typical in this literature, is the expected log-wealth under the alternative).

ACKNOWLEDGEMENTS

We would like to thank the Editors-in-Chief of the New England Journal of Statistics in Data Science, Ming-Hui Chen and Min-ge Xie, for enabling this special issue and supporting us throughout the process. We would also like to thank the Managing Editor, Hai-Ying Wang, for always responding extremely promptly to our queries.

REFERENCES

[1] Cook, T., Dubey, H. V., Lee, J. A., Zhu, G., Zhao, T. and Flaherty, P. (2024). Cost-Aware Generalized α-Investing for Multiple Hypothesis Testing. *The New England Journal of Statistics in Data Science* 1–20. https://doi.org/10.51387/24-NEJSDS64.

[2] Henzi, A., Puke, M., Dimitriadis, T. and Ziegel, J. (2023). A Safe Hosmer-Lemeshow Test. *The New England Journal of Statistics in Data Science* 1–15. https://doi.org/10.51387/23-NEJSDS56.

[3] Ramdas, A., Grünwald, P., Vovk, V. and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science* **38**(4) 576–601. https://doi.org/10.1214/23-sts894. MR4665027

[4] Shafer, G. (2023). Improving Data Analysis by Testing by Betting: Optional Continuation and Descriptive Statistics. *The New England Journal of Statistics in Data Science* 1–14. https://doi.org/10.51387/23-NEJSDS55.

[5] Shin, J., Ramdas, A. and Rinaldo, A. (2023). E-detectors: A Nonparametric Framework for Sequential Change Detection. *The New England Journal of Statistics in Data Science* 1–32. https://doi.org/10.51387/23-NEJSDS51.

[6] ter Schure, J., Pérez-Ortiz, M., Ly, A. and Grünwald, P. (2024). The Anytime-Valid Logrank Test: Error Control Under Continuous Monitoring with Unlimited Horizon. *The New England Journal of Statistics in Data Science* 1–26. https://doi.org/10.51387/24-NEJSDS65.

[7] Vovk, V. and Wang, R. (2024). Nonparametric E-tests of Symmetry. *The New England Journal of Statistics in Data Science* 1–10. https://doi.org/10.51387/24-NEJSDS60.

Aaditya Ramdas. Department of Statistics and Data Science, Carnegie Mellon University, USA. E-mail address: aramdas@cmu.edu

Peter Grünwald. Machine Learning Group, CWI Amsterdam, The Netherlands (also affiliated with Leiden University, Leiden, the Netherlands). E-mail address: peter.grunwald@cwi.nl