SAVI 2025: 3rd Workshop on Game-theoretic Statistics and Sequential, Anytime-Valid Inference

Organized by
Ruodo Wang (Waterloo)
Aaditya Ramdas (CMU)
Peter Grünwald (Leiden)
Johanna Ziegel (ETH)

Written by
Ben Chugg (CMU)

June 29th - July 4th, 2025

Abstract

Sequential, anytime-valid inference (SAVI) seeks to widen the scope of traditional statistical tools. Grounded in the theory of supermartingales, its methods provide valid inference under adaptive sampling, optional stopping, and other forms of data-dependent continuation that invalidate classical guarantees. The resulting statistical tests are based on *e-values* rather than p-values, which are both mathematically more flexible and significantly more interpretable. The third SAVI workshop, held in 2025 at the Chennai Mathematical Institute, brought together theoreticians and practitioners from across statistics, machine learning, economics, and related fields, all interested in the theory and practice of e-values.

1 Introduction by the Organizers

The third workshop on *Game-theoretic Statistics and Sequential, Anytime-Valid Inference* (SAVI) was held from June 29 to July 4 2025, at the Chennai Mathematical Institute (CMI) as a BIRS 5-day workshop. The meeting was organized by Ruodu Wang, Aaditya Ramdas, Peter Grünwald, and Johanna Ziegel. It continued the SAVI series after the 2022 meeting at EURANDOM and the 2024 Oberwolfach workshop, and drew an engaged group of theoreticians and practitioners. Topics and interests spanned statistics, probability, machine learning, and their applications in experimental design, biomedicine, economics, law, and reinforcement learning.

SAVI sits at a lively moment in its development. In addition to three recent JRSSB discussion papers, 2025 marked the publication of the first textbook devoted entirely to e-values [2]. The 2025 BIRS workshop reflected the community's rapid evolution, with exciting talks being given on a wide variety of topics and several presenting groundbreaking results.

There were 45 in person participants and 33 talks. Talks were either 30 or 45 minutes long and most stimulated lively discussions. The workshop drew a genuinely mixed cohort, both in seniority and geography. Career stages ranged from professor to student: 24 professors, 14 PhD students, 6 postdocs, and 1 engineer. Institutionally, attendees came from across the world—roughly 40% from Europe (e.g., CWI/Leiden/Delft,

Oxford, Paris-Saclay/Dauphine, Padova, Bremen), 30% from North America (CMU, UNC, UMass Amherst, UChicago, Toronto, MIT, Stanford), 20% from India (IISc, CMI, several IITs, TIFR, Ashoka), 10% from the rest of Asia (INSEAD/NUS/KAUST), plus one attendee from Oceania (La Trobe).

As with prior SAVI meetings, this workshop emphasized both theoretical and conceptual progress in addition to practical reach. A central through line was of course designing concrete procedures for testing and estimation that remain valid under various forms of feedback and adaptivity, and how to deploy them in modern experimental pipelines (such as A/B testing, bandits, off-policy evaluation, and conformal prediction).

Several talks made explicit connections between SAVI methodology and standard tools, discussing how e-values and e-processes can improve or strengthen traditional guarantees. Others pushed on practical "pain points": robustifying SAVI methods, sharpening thresholds, and leveraging side information across streams while retaining anytime validity. While a great many subjects were discussed, we might cluster the talks into the following themes:

Foundations and methodology of e-values and e-processes. There was significant focus on the theoretical foundations of e-values. Nick Koning started things off by discussing the link between e-values and testing [1]. Several talks then discussed power and optimality: Zhenyuan Zhang gave an elegant talk discussing when powerful e-values exist for composite hypotheses using the theory of optimal transport [3], and Eugene Clarico discussed optimal e-variables for various constrained hypothesis classes [7]. Peter Grünwald discussed the e-posterior [6] and some links between e-values and Bayesian decision theory. Wouter Koolen showed that the t-likelihood ratio test is an e-process under the appropriate one-sided null [5], thus solving an open-problem posed by Wang and Ramdas [4]. Christopher Blier-Wong returned to fundamental question of how to threshold e-values, showing that the traditional choice of $1/\alpha$ should change if one makes stronger distributional assumptions [26].

Several talks addressed deep, measure-theoretic properties of e-values and e-processes: YJ Choe discussed the possibility of combining evidence across filtrations [25] and Hien Nguyen spoke about when and how universal inference can overcome certain measurability assumptions.

Aaditya Ramdas closed out the workshop with an energetic final talk in which he examined common criticisms of e-values, giving his perspective on to what extent such criticisms are valid and how they ought to be addressed.

Sequential testing and confidence sequences. Several talks investigated stopping-time behavior, efficiency, and robustness for sequential procedures. Shubhada Agrawal discussed tight upper and lower bounds for power-one sequential tests [8] and Aytijhya Saha discussed how one can extend likelihood-ratio testing to handle Huber noise [9]. Lasse Fisher discussed the surprising fact that one can improve the approximate sequential likelihood ratio test by avoiding overshoot [10], thereby strictly improving the power of a classical and popular method.

The dual side of a sequential test is a sequential confidence interval, i.e., a confidence *sequence*. Hamish Flynn and Claudia Di Caterina discussed confidence sequences for generalized linear models [11, 13], Diego Martinez-Taboada showed how one can obtain sharp empirical Bernstein confidence sequences for bounded observations in smooth Banach spaces [14], and Francois Caron explored how to use priors in conjunction with the extended Ville's inequality to obtain robust confidence sequences [12]. Francesco Orabona gave a very well-received talk in which he showed how, using an insight going back to Cover, betting strategies can be modified in the fixed-time setting to yield tighter confidence intervals [15]. The core insight is that if one has only one timestep remaining, then a betting strategy can be more aggressive at no loss.

Multiple testing and post-hoc inference. There was significant activity around FDR control and multiple testing. Nikos Ignatiadis discussed asymptotic and compound e-values and the relationship to empirical Bayes [16] and Thorsten Dickhaus discussed anytime-valid bounds for FDP [20]. Junu Lee discussed improvements to e-BH via boosting [27] and Sanjit Dandapanthula introduced the problem of multiple testing in multi-stream changepoint detection [22], proving several negative results and proposing a new metric called *error over patience* (EOP). Ben Chugg discussed post-hoc hypothesis testing [17]. Neil Xu gave a remarkable talk on uniform improvements to the e-Benjamini-Hochberg and the Benjamini-Yekutieli procedures, showing that all multiple testing methods are a special case of a general closed testing procedure based on

e-values [19]. That such prevalent methods can be uniformly improved is astonishing—this was one of most surprising results presented at the workshop.

Connections to learning, experimentation, and decision-making. Many talks sought to connect recent theoretical advances in SAVI to make progress on classical problems in online experimentation. Stephan Bongers discussed anytime-valid off-policy evaluation for reinforcement learning, and Vikas Deep and Sandeep Juneja discussed optimal A/B testing and best arm identification [21, 28]. Muriel Pérez-Ortiz showed how one can construct anytime-valid tests for sparse anomalies [33], and Jonathan Williams discussed asymptotic efficiency and a possibilist version of the Bernstein von-Mises theorem [32]. Georgios Gavrilopoulos discussed how SAVI can provide flexible methods for sequential model selection [24]; Patrick Flaherty how such methods can be used to hedge the risk of financial bankruptcy in the business of experimentation [29]. Inspired in part by the practice of DNA matching in law, David Bickel gave a fascinating talk on the "marginal e-value," in which a prior is placed on the null hypothesis [30]. Etienne Gauthier relayed his work on how e-values can be used to expand the guarantees offered by conformal prediction, including enabling post-hoc valid conformal sets [31].

Overall, the talks were of a very high caliber. Tea and lunch breaks were filled with discussions and conjectures (Aaditya Ramdas could be seen making ample use of the blackboard in the tea room). Indeed, several collaborations are already under way, building on ideas spawned at the workshop.

Logistically, participants stayed at Gokulam Park Sabari with group transport to CMI each morning. The mood was lighthearted and undoubtedly helped by the excellent food provided by BIRS for lunch and by the hotel for breakfasts and dinners. On Wednesday afternoon, a break was taken from the technical talks and spent on an enjoyable excursion to Mahabalipuram (Mamallapuram). Many participants expressed regret at having to leave Chennai after the workshop.

We are grateful to BIRS and CMI for hosting and supporting SAVI 2025, and we acknowledge the broader BIRS consortium and funders. The workshop benefited from the cross-disciplinary mix of statisticians, probabilists, and ML researchers, and we hope to see the scope widen ever further in the coming years. The momentum from EURANDOM (2022) and Oberwolfach (2024) was undoubtedly carried forward in Chennai. We are looking forward to SAVI 2026!

2 Summary of Talks

The marginal e-value: Testing by betting given a prior probability of the null hypothesis

David R. Bickel

Abstract. The infamous player called "Skeptic" bets \$1 against the simple null hypothesis that a random sample X will be drawn from a distribution of probability density function f_0, knowing that it will otherwise be drawn from a distribution of probability density function f_1. Skeptic also knows the prior probability of the null hypothesis to be π_0, a number in $[0, 1]$. In return for the \$1, Skeptic chooses to receive the payout that is log-optimal according to the prior predictive distribution of probability density function $f = \pi_0 f_0 + (1 - \pi_0) f_1$. That payout is the e-variable $E = f(X)/f_0(X)$. With x as the observed sample, the e-value that realizes E is $e = f(x)/f_0(x)$, where $f(x)$ is known as the marginal likelihood. Considering e as the degree to which the null hypothesis is disproven resolves certain pathologies in evidence theory while reflecting the prior probability of the null hypothesis. To generalize that, let $E_{(0)}$ denote any e-variable that tests a simple or composite null hypothesis, and let $e_{(0)}$ be its e-value for $X = x$. The corresponding marginal e-variable is $E_{\pi_0} = \pi_0 + (1 - \pi_0) E_{(0)}$, and its realization, the marginal e-value, is $e_{\pi_0} = \pi_0 + (1 - \pi_0) e_{(0)}$. For $\pi_0 > 0$, as is suitable for many genetics and genomics applications, e_{π_0} is regularized toward 1 to the extent that the null hypothesis has high prior probability.

Boosting e-BH via conditional calibration

June Lee

Abstract. The e-BH procedure is an e-value-based multiple testing procedure that provably controls the false discovery rate (FDR) under any dependence structure between the e-values. Despite this appealing theoretical FDR control guarantee, the e-BH procedure often suffers from low power in practice. In this paper, we propose a general framework that boosts the power of e-BH without sacrificing its FDR control under arbitrary dependence. This is achieved by the technique of conditional calibration, where we take as input the e-values and calibrate them to be a set of "boosted e-values" that are guaranteed to be no less – and are often more – powerful than the original ones. Our general framework is explicitly instantiated in three classes of multiple testing problems: (1) testing under parametric models, (2) conditional independence testing under the model-X setting, and (3) model-free conformalized selection. Extensive numerical experiments show that our proposed method significantly improves the power of e-BH while continuing to control the FDR. We also demonstrate the effectiveness of our method through an application to an observational study dataset for identifying individuals whose counterfactuals satisfy certain properties.

Fuzzy / Continuous Testing: Unifying (Optimal) Tests and E-values

Nick Koning

Abstract. While the e-value is swiftly rising in prominence in many applications of hypothesis testing and multiple testing, its formal relationship to the classical theory of testing is not yet fully settled. We unify e-values and classical testing, by describing how e-values naturally arise as the 'continuous' or 'fuzzy' generalization of a test. This cements the foundational role of the e-value in hypothesis testing. Such continuous tests may be viewed as directly interpreting the rejection probability of classical randomized tests as evidence, offering the benefits of randomized tests without the downsides of a randomized decision. By generalizing the traditional notion of power, we obtain a unified theory of optimal testing which nests both classical Neyman-Pearson-optimal tests and log-optimal e-values as special cases. This suggests the only difference between typical classical tests and typical e-values is a different choice of power target. Finally, we describe the relationship to the traditional p-value, and show that continuous tests offer a stronger and arguably more appropriate guarantee than p-values when used as a continuous measure of evidence.

Improving Wald's (approximate) sequential probability ratio test by avoiding overshoot

Lasse Fischer

Abstract. Wald's sequential probability ratio test (SPRT) is a cornerstone of sequential analysis. Based on desired type-I, II error levels α, β, it stops when the likelihood ratio crosses certain thresholds, guaranteeing optimality of the expected sample size. However, these thresholds are not closed form and the test is often applied with approximate thresholds $(1-\beta)/\alpha$ and $\beta/(1-\alpha)$ (approximate SPRT). When $\beta > 0$, this neither guarantees error control at α, β nor optimality. When $\beta = 0$ (power-one SPRT), this method is conservative and not optimal. The looseness in both cases is caused by *overshoot*: the test statistic overshoots the thresholds at the stopping time. Numerically calculating thresholds may be infeasible, and most software packages do not do this. We improve the approximate SPRT by modifying the test statistic to avoid overshoot. Our 'sequential boosting' technique *uniformly* improves power-one SPRTs ($\beta = 0$) for simple nulls and alternatives, or for one-sided nulls and alternatives in exponential families. When $\beta > 0$, our techniques provide guaranteed error control at α, β, while needing less samples than the approximate SPRT in our simulations. We also provide several nontrivial extensions: confidence sequences, sampling without replacement and conformal martingales.

On the existence of powerful p-values and e-values for composite hypotheses testing

Zhenyuan Zhang

Abstract. Given a composite null $\mathcal{P}$ and composite alternative $\mathcal{Q}$, when and how can we construct a p-value whose distribution is exactly (or stochastically larger than) uniform under the null, and stochastically smaller than uniform under the alternative? Similarly, when and how can we construct an e-value whose expectation exactly equals (or is smaller than) one under the null, but its expected logarithm under the alternative is positive? We give very neat answers to these basic questions when $\mathcal{P}$ and $\mathcal{Q}$ are convex polytopes (in the space of probability measures on a Polish space $\mathfrak{X}$). We prove that such constructions are possible if and only if $\mathcal{Q}$ does not intersect the span of $\mathcal{P}$. If the p-value is allowed to be stochastically larger than uniform under $P \in \mathcal{P}$, and the e-value can have expectation at most one under $P \in \mathcal{P}$, then it is achievable whenever $\mathcal{P}$ and $\mathcal{Q}$ are disjoint. More generally, even when $\mathcal{P}$ and $\mathcal{Q}$ are not polytopes, we characterize the existence of a bounded nontrivial e-variable whose expectation exactly equals one under any $P \in \mathcal{P}$. The proofs utilize convex geometry and recently developed techniques in simultaneous transport of measures. We also provide an iterative construction that explicitly constructs such p/e-values, and under certain conditions it finds the one that grows fastest under a specific alternative Q.

Sequential Model Confidence Sets

Georgios Gavrilopoulos

Abstract In most prediction and estimation situations, scientists consider various statistical models for the same problem, and naturally want to select amongst the best. Hansen et al. (2011) provide a powerful solution to this problem by the so-called model confidence set, a subset of the original set of available models that contains the best models with a given level of confidence. Importantly, model confidence sets respect the underlying selection uncertainty by being flexible in size. However, they presuppose a fixed sample size which stands in contrast to the fact that model selection and forecast evaluation are inherently sequential tasks where we successively collect new data and where the decision to continue or conclude a study may depend on the previous outcomes. In this article, we extend model confidence sets sequentially over time by relying on sequential testing methods. Recently, e-processes and confidence sequences have been introduced as new, safe methods for assessing statistical evidence. Sequential model confidence sets allow to continuously monitor the models' performances and come with time-uniform, nonasymptotic coverage guarantees.

STAR-Bets: Sequential TArget-Recalculating Bets for Tighter Confidence Intervals

Francesco Orabona

Abstract The construction of confidence intervals for the mean of a bounded random variable is a classical problem in statistics with numerous applications in machine learning and virtually all scientific fields. In particular, obtaining the tightest possible confidence intervals is vital every time the sampling of the random variables is expensive. The current state-of-the-art method to construct confidence intervals is by using betting algorithms. This is a very successful approach for deriving optimal confidence sequences, even matching the rate of law of iterated logarithms. However, in the fixed horizon setting, these approaches are either sub-optimal or based on heuristic solutions with strong empirical performance but without a finite-time guarantee. Hence, no betting-based algorithm guaranteeing the optimal $\mathcal{O}(\sqrt{\frac{\sigma^2 \log \frac{1}{\delta}}{n}})$ width of the confidence intervals is known. This work bridges this gap. We propose a betting-based algorithm to compute confidence intervals that empirically outperforms the competitors. Our betting strategy uses the optimal strategy in every step (in a certain sense), whereas the standard betting methods choose a constant strategy in advance. Leveraging this fact results in strict improvements even for classical concentration inequalities, such as the ones of Hoeffding or Bernstein. Moreover, we also prove that the width of our confidence intervals is optimal up to a constant factor.

Optimal Best-Arm Identification in Bandits with access to Offline Data

Sandeep Juneja

Abstract Learning paradigms based solely on offline data and those based solely on sequential online learning have been well studied in the literature. In this talk, we consider the combination of offline data with online learning, an area less studied but of obvious practical importance. We consider the stochastic multi-armed bandit problem, where our goal is to identify the arm with the highest mean in the presence of relevant offline data, with confidence $1 - \delta$. We perform a lower-bound analysis on policies that provide such probabilistic correctness guarantees. We also identify a boundary region such that zero online samples are needed when the number of offline samples is above that region. We further develop plug-and-play algorithms that rely on the popular generalized likelihood ratio (GLR) method that asymptotically matches the lower bound even up to the multiplicative constant when off-line samples are sufficiently lower than the boundary region. When the offline samples are sufficiently close to the boundary region, the situation is more delicate. We conduct a detailed analysis to bring out conditions under which online sample complexity continues to match the lower bound, and where it does not due to fundamental statistical noise-driven limitations of the GLR method. Algorithms based on plug-and-play into the lower bound, while optimal, can be computationally prohibitive. In online settings, faster top-2 algorithms that are also computationally efficient have been devised. We observe that such algorithms can be suboptimal in our online-offline setting. We shed light on performance degradation and, through nuanced analysis, develop clairvoyant top-2 methods that perform asymptotically optimally. Further, we outline the fluid behavior of the proposed algorithm that illuminates its comparative advantage.

Responding to e-rroneous criticisms of e-values

Aaditya Ramdas

Abstract This talk discusses some common misconceptions surrounding e-values, their power, and their applications. For example, to the criticism that e-based methods are less powerful than others, we note that any level-α test can be recovered by thresholding a suitably chosen e-value. Likewise, any valid FDR procedure can be realized as e-BH on appropriate compound e-values. A second criticism is that sequential methods based on Ville's inequality must be loose. However, every sequential test uses Ville's inequality implicitly. Moreover, sequential tests are often as tight or tighter than fixed-n tests. We end by discussing a universality result for e-processes.

Data Integration Via Analysis of Subspaces (DIVAS)

Jan Hannig

Abstract A major challenge in the age of Big Data is the integration of disparate data types into a data analysis. That is tackled here in the context of data blocks measured on a common set of experimental subjects. This data structure motivates the simultaneous exploration of the joint and individual variation within each data block. This is done here in a way that scales well to large data sets (with blocks of wildly disparate size), using principal angle analysis, careful formulation of the underlying linear algebra, and differing outputs depending on the analytical goals. Ideas are illustrated using mortality and neuroimaging data sets. Connections to e-values will be explored.

Asymptotically optimal adaptive A/B test for ATE

Vikas Deep

Abstract Motivated by practical applications in clinical trials and online platforms, we study A/B testing with the aim of estimating a confidence interval (CI) for the average treatment effect (ATE) using the minimum expected sample size. This CI should have a width at most ϵ while ensuring that the probability of the CI not containing the true ATE is at most δ. To answer this, we first establish a lower bound on the expected sample size needed for any adaptive policy which constructs a CI of ATE with desired properties. Specifically, we prove that the lower bound is based on the solution to a non-convex max-min optimization problem for small δ. Tailoring the "plug-in" approach for the ATE problem, we construct an adaptive policy that is asymptotically optimal, i.e., matches the lower bound on the expected sample size for small δ. Interestingly, we find that, for small ϵ and δ, the asymptotically optimal fraction of treatment assignment for A and B is proportional to the standard deviation of the outcome distributions of treatments A and B, respectively. However, as the proposed approach can be computationally intensive, we propose an alternative adaptive policy. This new policy, informed by insights from our lower bound analysis, is computationally efficient while remaining asymptotically optimal for small values of ϵ and δ. Numerical comparisons demonstrate that both policies perform similarly across practical values of ϵ and δ, offering efficient solutions for A/B testing.

The E-Posterior

Peter Grünwald

Abstract I review definition and meaning of the *E-Posterior*, a concept I introduced in a 2023 paper. Essentially, the e-posterior of θ given Y is just the reciprocal of the e-value of Y, where the e-value is based on taking θ as representing a null hypothesis. It is shown how the graph of the e-posterior can be revealingly compared to the graph of the *contour function* of a standard posterior and the *p-value function* of a system of confidence intervals. The latter also gives rise to an *IM (inferential model)* which is prominent in Ryan Martin's works. The e-posterior is a special type of IM, which is considerably looser (implies weaker probabilistic statements about θ) than a standard IM. This looseness pays off in that it allows you to make, to some extent, decisions relative to *data-dependent* loss functions. This is illustrated by comparing four versions of the same game, in which, after Y is observed, a bookie offers a certain gamble to a decision maker. It is shown how the game is favourable to decision maker in quite narrow circumstances if the decision-maker employs a confidence distribution, somewhat wider circumstances if she employs an IM, significantly wider circumstances if she employs an e-posterior, and the widest circumstances if she is willing to make strong Bayesian assumptions.

On Stopping Times of Power-one Sequential Tests: Tight Lower and Upper Bounds

Shubhada Agrawal

Abstract We prove two lower bounds for stopping times of sequential tests between general composite nulls and alternatives. The first lower bound is for the setting where the type-1 error level α approaches zero, and equals $\log(1/\alpha)$ divided by a certain infimum KL divergence, termed $\mathrm{KL}_{\mathrm{inf}}$. The second lower bound applies to the setting where α is fixed and $\mathrm{KL}_{\mathrm{inf}}$ approaches 0 (meaning that the null and alternative sets are not separated) and equals $c\mathrm{KL}_{\mathrm{inf}}^{-1}\log\log\mathrm{KL}_{\mathrm{inf}}^{-1}$ for a universal constant $c > 0$. We also provide a sufficient condition for matching the upper bounds and show that this condition is met in several special cases. Given past work, these upper and lower bounds are unsurprising in their form; our main contribution is the generality in which they hold, for example, not requiring reference measures or compactness of the classes.

Anytime-valid simultaneous lower confidence bounds for the true discovery proportion

Thorsten Dickhaus

Abstract We propose a method that combines the closed testing framework with the concept of safe anytime-valid inference (SAVI) to compute lower confidence bounds for the true discovery proportion in a multiple testing setting. The proposed procedure provides confidence bounds that are valid at every observation time point and that are simultaneous for all possible subsets of hypotheses. While the hypotheses are assumed to be fixed over time, the subsets of interest may vary. Anytime-valid simultaneous confidence bounds allow us to sequentially update the bounds over time and allow for optional stopping. This is a desirable property in practical applications such as neuroscience, where data acquisition is costly and time-consuming. We also present a computational shortcut which makes the application of the proposed procedure feasible when the number of hypotheses under consideration is large. We illustrate the performance of the proposed method in a simulation study and give some practical guidelines on the implementation of the proposed procedure.

Asymptotic and compound e-values: multiple testing and empirical Bayes.

Nikos Ignatiadis

Abstract We explicitly define the notions of (bona fide, approximate or asymptotic) compound p-values and e-values, which have been implicitly presented and extensively used in the recent multiple testing literature. While it is known that the e-BH procedure with compound e-values controls the FDR, we show the converse: every FDR controlling procedure can be recovered by instantiating the e-BH procedure with certain compound e-values. Since compound e-values are closed under averaging, this allows for combination and derandomization of FDR procedures. We then connect compound e-values to empirical Bayes. In particular, we use the fundamental theorem of compound decision theory to derive the log-optimal simple separable compound e-value for testing a set of point nulls against point alternatives: it is a ratio of mixture likelihoods. As one example, we construct asymptotic compound e-values for multiple t-tests, where the (nuisance) variances may be different across hypotheses. Our construction may be interpreted as a data-driven instantiation of the optimal discovery procedure (ODP), and our results provide the first type-I error guarantees for data-driven ODP.

Universal inference without measurability

Hien Nguyen

Abstract When a model is correctly specified, universal inference allows for the construction of finite sample correct confidence sets and hypothesis tests, regardless of the available estimator at hand. This enables the application of universal inference in exotic settings where few other competing methods are available. However, in such settings, the prevalence of measure-theoretic pathologies becomes more common, making the process of maximization and the available parameter estimators non-measurable. In empirical process asymptotics, a reasonable treatment of non-measurability is via the use of the outer expectation and outer probability. In this talk, we borrow this perspective and seek to demonstrate that many constructions of universal inference are valid when expectation and probability are replaced by their outer counterparts.

Confidence Sequences for Generalized Linear Models via Regret Analysis.

Hamish Flynn

Abstract. We develop a methodology for constructing confidence sets for parameters of statistical models via a reduction to sequential prediction. Our key observation is that for any generalized linear model (GLM), one can construct an associated game of sequential probability assignment such that achieving low regret in the game implies a high-probability upper bound on the excess likelihood of the true parameter of the GLM. This allows us to develop a scheme that we call online-to-confidence-set conversions, which effectively reduces the problem of proving the desired statistical claim to an algorithmic question. We study two varieties of this conversion scheme: 1) analytical conversions that only require proving the existence of algorithms with low regret and provide confidence sets centered at the maximum-likelihood estimator 2) algorithmic

conversions that actively leverage the output of the online algorithm to construct confidence sets (and may be centered at other, adaptively constructed point estimators). The resulting methodology recovers all state-of-the-art confidence set constructions within a single framework, and also provides several new types of confidence sets that were previously unknown in the literature..

Hedging in Sequential Experiments

Patrick Flaherty

Abstract. We build on the game-theoretic statistics framework to explore how an investigator can hedge their bets against the null hypothesis and thus avoid ruin. First, we describe a method by which the investigator's test martingale wealth process can be capitalized by solving for the risk-neutral price. Then, we show that a portfolio that comprises the risky test martingale and a risk-free process is still a test martingale which enables the investigator to select a particular risk-return position. Finally, we show that a function that is derivative of the test martingale process can be constructed and used as a hedging instrument by the investigator or as a speculative instrument by a risk-seeking investor who wants to participate in the potential returns of the uncertain experiment wealth process. This is ongoing work and we hope to discuss where and how concepts from mathematical finance can be used to allocate or share risk in biological experiments.

Asymptotic efficiency of inferential models and a possibilistic Bernstein–von Mises theorem

Jonathan Williams

Abstract. The inferential model (IM) framework offers an alternative to the classical probabilistic (e.g., Bayesian and fiducial) uncertainty quantification in statistical inference. A key distinction is that classical uncertainty quantification takes the form of precise probabilities and offers only limited large-sample validity guarantees, whereas the IM's uncertainty quantification is imprecise in such a way that exact, finite-sample valid inference is possible. But are the IM's imprecision and finitesample validity compatible with statistical efficiency? That is, can IMs be both finite-sample valid and asymptotically efficient? This paper gives an affirmative answer to this question via a new possibilistic Bernstein–von Mises theorem that parallels a fundamental Bayesian result. Among other things, our result shows that the IM solution is efficient in the sense that, asymptotically, its credal set is the smallest that contains the Gaussian distribution with variance equal to the Cramér–Rao lower bound. Moreover, a corresponding version of this new Bernstein–von Mises theorem is presented for problems that involve the elimination of nuisance parameters, which settles an open question concerning the relative efficiency of profiling-based versus extension-based marginalization strategies.

Optimal classes of e-variables.

Eugenio Clerico

Abstract. We discuss the existence and characterise in simple cases the smallest class of e-variables that dominates all other e-variables. Some open question are outlined.

Approximate mixture confidence sequences in generalized linear models

Claudia Di Caterina

Abstract. We illustrate a simple method for computing approximate mixture confidence sequences for scalar components of the coefficient vector in generalized linear models. The sequential calculation is based on the renewable estimation algorithm proposed by Luo and Song (2020, *JRSSB*). We evaluate the empirical performance of the procedure through simulations within the logistic regression framework, including a comparison with an approximate version of the asymptotic Gaussian mixture confidence sequences.

Combining evidence across filtrations

Yo Joong (YJ) Choe

Abstract. In sequential anytime-valid inference, any admissible procedure must be based on *e-processes*: generalizations of test martingales that quantify the accumulated evidence against a composite null hypothesis at any stopping time. We propose a method for combining e-processes constructed in different filtrations but for the same null. Although e-processes in the same filtration can be combined effortlessly (by averaging), e-processes in different filtrations cannot because their validity in a coarser filtration does not translate to a finer filtration. This issue arises in sequential tests of randomness and independence, as well as in the evaluation of sequential forecasters. We establish that a class of functions called *adjusters* can lift arbitrary e-processes across filtrations. The result yields a generally applicable "adjust-then-combine" procedure, which we demonstrate on the problem of testing randomness in real-world financial data. Furthermore, we prove a characterization theorem for adjusters that formalizes a sense in which using adjusters is necessary. There are two major implications. First, if we have a powerful e-process in a coarsened filtration, then we readily have a powerful e-process in the original filtration. Second, when we coarsen the filtration to construct an e-process, there is a logarithmic cost to recovering validity in the original filtration.

Bringing Closure to FDR Control With a Uniform Improvement of the e-Benjamini-Hochberg Procedure

Ziyu Xu (Neil)

Abstract. We present a novel necessary and sufficient principle for multiple testing methods. This principle asserts that every multiple testing method is a special case of a general closed testing procedure based on e-values. It generalizes the standard closure principle, known to underlie all methods controlling familywise error and tail probabilities of false discovery proportions, to a large class of error rates — in particular, this generalized closure principle applies to methods controlling the false discovery rate (FDR). By writing existing methods as special cases of this procedure, we can achieve uniform improvements of these methods, and we show this in particular for the eBH and the BY procedures, as well as the self-consistent method of Su (2018). We also show that methods derived using the closure principle have several valuable properties. They generally control their error rate not just for one rejected set, but simultaneously over many, allowing post hoc flexibility for the researcher. Moreover, we show that because all multiple testing methods for all error rates are special cases of the same procedure, researchers may even choose the target error rate post hoc. Under certain conditions, this flexibility even extends to post hoc choice of the nominal error rate. In addition, the closure principle allows methods to exploit logical relationships between hypotheses to gain power.

E-Values Expand the Scope of Conformal Prediction

Etienne Gauthier

Abstract. Conformal prediction is a powerful framework for distribution-free uncertainty quantification. The standard approach to conformal prediction relies on comparing the ranks of prediction scores: under exchangeability, the rank of a future test point cannot be too extreme relative to a calibration set. This rank-based method can be reformulated in terms of p-values. In this paper, we explore an alternative approach based on e-values, known as conformal e-prediction. E-values offer key advantages that cannot be achieved with p-values, enabling new theoretical and practical capabilities. In particular, we present three applications

that leverage the unique strengths of e-values: batch anytime-valid conformal prediction, fixed-size conformal sets with data-dependent coverage, and conformal prediction under ambiguous ground truth. Overall, these examples demonstrate that e-value-based constructions provide a flexible expansion of the toolbox of conformal prediction.

Empirical Bernstein in smooth Banach spaces

Diego Martinez-Taboada

Abstract. Existing concentration bounds for bounded vector-valued random variables include extensions of the scalar Hoeffding and Bernstein inequalities. While the latter is typically tighter, it requires knowing a bound on the variance of the random variables. We derive a new vector-valued empirical Bernstein inequality, which makes use of an empirical estimator of the variance instead of the true variance. The bound holds in 2-smooth separable Banach spaces, which include finite dimensional Euclidean spaces and separable Hilbert spaces. The resulting confidence sets are instantiated for both the batch setting (where the sample size is fixed) and the sequential setting (where the sample size is a stopping time). The confidence set width asymptotically exactly matches that achieved by Bernstein in the leading term.

Multiple testing in multi-stream sequential change detection

Sanjit Dandapanthula

Abstract. Multi-stream sequential change detection involves simultaneously monitoring many streams of data and trying to detect when their distributions change, if at all. Here, we theoretically study multiple testing issues that arise from detecting changes in many streams. We point out that any algorithm with finite average run length (ARL) must have a trivial worst-case false detection rate (FDR), family-wise error rate (FWER), per-family error rate (PFER), and global error rate (GER); thus, any attempt to control these Type I error metrics is fundamentally in conflict with the desire for a finite ARL (which is typically necessary in order to have a small detection delay). One of our contributions is to define a new class of metrics which can be controlled, called error over patience (EOP). We propose algorithms that combine the recent e-detector framework (which generalizes the Shiryaev-Roberts and CUSUM methods) with the recent e-Benjamini-Hochberg procedure and e-Bonferroni procedures. We prove that these algorithms control the EOP at any desired level under very general dependence structures on the data within and across the streams. In fact, we prove a more general error control that holds uniformly over all stopping times and provides a smooth trade-off between the conflicting metrics. Additionally, if finiteness of the ARL is forfeited, we show that our algorithms control the worst-case Type I error.

Anytime-Valid Tests for Sparse Anomalies

Muriel Pérez

Abstract. We study Anytime-Valid (AV) tests for the presence of anomalies in a large number of data streams. AV tests are built that monitor mixture likelihood ratios continuously and their performance is judged using log-optimality criteria. We propose a framework to assess AV tests for this problem and show the fundamental limits of detection in a Gaussian-location instance of the problem. These results are related to but not implied by existing results for fixed-sample tests. Additionally, in the Gaussian-location setting, we show tests that are optimal even when the parameters of the problem are unknown. The methods that we develop set the benchmarks for more general optimal AV tests for sparse anomalies.

Huber-robust likelihood ratio tests for composite nulls and alternatives

Aytijhya Saha

Abstract. We present an e-value based framework for testing composite nulls against composite alternatives when an ϵ fraction of the data can be arbitrarily corrupted. Our tests are inherently sequential, being valid at arbitrary data-dependent stopping times, but they are new even for fixed sample sizes, giving type-I error control without any regularity conditions. We achieve this by modifying and extending a proposal by Huber (1965) in the point null versus point alternative case. Our test statistic is a nonnegative supermartingale under the null, even with a sequentially adaptive contamination model where the conditional distribution of each observation given the past data lies within an ϵ (total variation) ball of the null. The test is powerful within an ϵ ball of the alternative. As a consequence, one obtains anytime-valid p-values that enable continuous monitoring of the data, and adaptive stopping. We analyze the growth rate of our test supermartingale and demonstrate that as $\epsilon \to 0$, it approaches a certain Kullback-Leibler divergence between the null and alternative, which is the optimal non-robust growth rate. A key step is the derivation of a robust Reverse Information Projection (RIPr). Simulations validate the theory and demonstrate excellent practical performance.

Improved thresholds for e-values

Christopher Blier-Wong

Abstract. The rejection threshold used for e-values and e-processes is by default set to $1/\alpha$ for a guaranteed type-I error control at α, based on Markov's and Ville's inequalities. This threshold can be wasteful in practical applications. We discuss how this threshold can be improved under additional distributional assumptions on the e-values; some of these assumptions are naturally plausible and empirically observable, without knowing explicitly the form or model of the e-values. For small values of α, the threshold can roughly be improved (divided) by a factor of 2 for decreasing or unimodal densities, and by a factor of e for decreasing or unimodal-symmetric densities of the log-transformed e-value. Moreover, we propose to use the supremum of comonotonic e-values, which is shown to preserve the type-I error guarantee. We also propose some preliminary methods to boost e-values in the e-BH procedure under some distributional assumptions while controlling the false discovery rate. Through a series of simulation studies, we demonstrate the effectiveness of our proposed methods in various testing scenarios, showing enhanced power.

This is joint work with Ruodu Wang. 20 minute talk.

Confidence sequences with informative, bounded-influence priors

François Caron

Abstract. Confidence sequences are collections of confidence sets that, at a prescribed level, simultaneously cover the true parameter for every sample size. Obtaining tight confidence sequences is of great interest, and this can be achieved by incorporating prior information via the method of mixtures. However, confidence sequences constructed with informative priors can be sensitive to misspecifications and may become arbitrarily wide - and therefore vacuous - when the prior conflicts with the data. Here, we focus on obtaining confidence sequences for the mean of Gaussians observations with known variance. We show that, when using a prior whose tails decay exponentially or polynomially, together with the extended Ville inequality, the width of the confidence sequence remains uniformly bounded. In the polynomial-tail case, the sequence reverts to the one obtained under an improper prior when the prior strongly conflicts with the data. Our analysis therefore yields confidence sequences that are tighter than standard ones when the prior is well-specified, while guaranteeing uniformly bounded width under misspecification.

Anytime-valid off-policy evaluation for reinforcement learning

Stephan Bongers

Abstract. Off-policy evaluation (OPE) in reinforcement learning (RL) concerns the evaluation of hypothetical policies without having to deploy them for exploration, which can be costly, risky, and/or unfeasible. A key limitation of the applicability of OPE in RL is that it is known to suffer from high-variance in long-horizon domains, which make reliable quantification of the uncertainty crucial. Often, the uncertainty is

quantified using confidence intervals (CIs) which are only valid for sample sizes that are fixed in advance. In this paper, we provide for the first time asymptotic confidence sequences (CS) for the evaluation of a policy in RL that are asymptotically valid at arbitrary (data-dependent) sample sizes. Asymptotic CSs are the anytime-valid analog of asymptotic CIs that allow for asymptotically valid inference at any point in time during the data collection process. Concretely, we construct asymptotic CSs for various existing estimators of interest in OPE. These include the doubly robust estimators that are known to be efficient in the sense that they attain the lowest possible asymptotic variance. We show that our CSs are valid for a wide range of settings, including those that have (1) a Markovian and/or time-invariant structure, and (2) a finite or infinite horizon. In particular, we show that Neyman orthogonality holds in these settings and provides the conditions that guarantee the validity of the asymptotic CSs. We empirically demonstrate the benefits of leveraging our asymptotic CSs.

On admissibility in post-hoc hypothesis testing

Ben Chugg

Abstract. The validity of classical hypothesis testing requires the significance level—typically denoted as α—to be fixed before any statistical analysis takes place. This is a stringent requirement. For instance, it prohibits updating α during (or after) an experiment due to changing concern about the cost of false positives, or to reflect unexpectedly strong evidence against the null. Perhaps most disturbingly, witnessing a p-value $p \ll \alpha$ vs $p \leq \alpha$ has no (statistical) relevance for any downstream decision-making. Building on observations of Abraham Wald in 1939 and following recent work of Grünwald (2024), we develop a theory of *post-hoc* hypothesis testing, enabling α to be chosen after seeing and analyzing the data. To study "good" post-hoc tests we introduce Π-admissibility, where Π is a set of adversaries which map the data to a significance level. A test is Π-admissible if, roughly speaking, there is no other test which performs at least as well and sometimes better across all adversaries in Π. For point nulls and alternatives, we prove general properties of any Π-admissible test for any Π and classify the set of admissible tests for various specific Π.

The t-test is a supermartingale after all

Wouter Koolen

Abstract. The t-statistic is a widely-used scale-invariant statistic for testing the null hypothesis that the mean is zero. Martingale methods enable sequential testing with the t-statistic at every sample size, while controlling the probability of falsely rejecting the null. For one-sided sequential tests, which reject when the t-statistic is too positive, a natural question is whether they also control false rejection when the true mean is negative. We prove that this is the case using monotone likelihood ratios and sufficient statistics. We develop applications to the scale-invariant t-test, the location-invariant -test and sequential linear regression with nuisance covariates.

References

[1] Nick W Koning, Continuous testing: Unifying tests and e-values, *arXiv preprint arXiv:2409.05654* (2024).

[2] Aaditya Ramdas and Ruodu Wang, Hypothesis testing with e-values, *Foundations and Trends in Statistics* **1** (2025).

[3] Zhenyuan Zhang, Aaditya Ramdas, and Ruodu Wang, On the existence of powerful p-values and e-values for composite hypotheses, *The Annals of Statistics* **52**(5) (2024) 2241—-2267.

[4] Hongjian Wang and Aaditya Ramdas, Anytime-valid t-tests and confidence sequences for Gaussian means with unknown variance, *Sequential Analysis* **44**(1) (2025) 56—-110.

[5] Peter D Grünwald and Wouter M Koolen, Supermartingales for one-sided tests: Sufficient monotone likelihood ratios are sufficient, *arXiv preprint arXiv:2502.04208* (2025).

[6] Peter D Grünwald, The e-posterior, *Philosophical Transactions of the Royal Society A* **381**(2247) (2023) 20220146.

[7] Eugenio Clerico, On the optimality of coin-betting for mean estimation, *International Journal of Approximate Reasoning* (2025) 109550.

[8] Shubhada Agrawal and Aaditya Ramdas, On Stopping Times of Power-one Sequential Tests: Tight Lower and Upper Bounds, *arXiv preprint arXiv:2504.19952* (2025).

[9] Aytijhya Saha and Aaditya Ramdas, Huber-robust likelihood ratio tests for composite nulls and alternatives, *arXiv preprint arXiv:2408.14015* (2024).

[10] Lasse Fischer and Aaditya Ramdas, Improving the (approximate) sequential probability ratio test by avoiding overshoot, *arXiv e-prints* (2024) arXiv—-2410.

[11] Eugenio Clerico, Hamish Flynn, Gergely Neu, and others, Confidence Sequences for Generalized Linear Models via Regret Analysis, *arXiv preprint arXiv:2504.16555* (2025).

[12] Stefano Cortinovis, Valentin Kilian, and François Caron, Confidence sequences with informative, bounded-influence priors, *arXiv preprint arXiv:2506.22925* (2025).

[13] Claudia Di Caterina, Alessandra Salvan, Nicola Sartori, and others, Mixture confidence sequences for regression coefficients in generalized linear models. In *Proceedings of the 37th International Workshop on Statistical Modelling*, **3**, 420—-423, 2023.

[14] Diego Martinez-Taboada and Aaditya Ramdas, Empirical Bernstein in smooth Banach spaces, *arXiv preprint arXiv:2409.06060* (2024).

[15] Václav Voráček and Francesco Orabona, STaR-Bets: Sequential Target-Recalculating Bets for Tighter Confidence Intervals, *arXiv preprint arXiv:2505.22422* (2025).

[16] Nikolaos Ignatiadis, Ruodu Wang, and Aaditya Ramdas, Asymptotic and compound e-values: multiple testing and empirical Bayes, *arXiv preprint arXiv:2409.19812* (2024).

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

[18] S. Bongers, M. T. J. Spaan, and F. A. Oliehoek., Anytime-valid off-policy evaluation for reinforcement learning, *In Progress* (2025).

[19] Ziyu Xu, Aldo Solari, Lasse Fischer, Rianne de Heide, Aaditya Ramdas, and Jelle Goeman, Bringing Closure to False Discovery Rate Control: A General Principle for Multiple Testing, *arXiv preprint arXiv:2509.02517* (2025).

[20] Friederike Preusse, Anytime-valid simultaneous lower confidence bounds for the true discovery proportion, *arXiv preprint arXiv:2505.17803* (2025).

[21] Vikas Deep, Achal Bassamboo, and Sandeep Kumar Juneja, Asymptotically optimal and computationally efficient average treatment effect estimation in A/B testing. In *Forty-first International Conference on Machine Learning*, 2024.

[22] Sanjit Dandapanthula and Aaditya Ramdas, Multiple testing in multi-stream sequential change detection, *arXiv preprint arXiv:2501.04130* (2025).

[23] Jack Prothero, Meilei Jiang, Jan Hannig, Quoc Tran-Dinh, Andrew Ackerman, and JS Marron, Data integration via analysis of subspaces (DIVAS), *TEST* **33**(3) (2024) 633—-674.

[24] Sebastian Arnold, Georgios Gavrilopoulos, Benedikt Schulz, and Johanna Ziegel, Sequential model confidence sets, *arXiv preprint arXiv:2404.18678* (2024).

[25] Yo Joong Choe and Aaditya Ramdas, Combining evidence across filtrations, *arXiv preprint arXiv:2402.09698* (2024).

[26] Christopher Blier-Wong and Ruodu Wang, Improved thresholds for e-values, *arXiv preprint arXiv:2408.11307* (2024).

[27] Junu Lee and Zhimei Ren, Boosting e-BH via conditional calibration, *arXiv preprint arXiv:2404.17562* (2024).

[28] Agniv Bandyopadhyay, Sandeep Juneja, and Shubhada Agrawal, Optimal top-two method for best arm identification and fluid analysis, *Advances in Neural Information Processing Systems* **37** (2024) 66568—-66646.

[29] Thomas Cook and Patrick Flaherty, Hedging in Sequential Experiments, *arXiv preprint arXiv:2406.15867* (2024).

[30] David R Bickel, The marginal e-value: Testing by betting given a prior probability of the null hypothesis, *Preprint* (2025).

[31] Etienne Gauthier, Francis Bach, and Michael I Jordan, E-values expand the scope of conformal prediction, *arXiv preprint arXiv:2503.13050* (2025).

[32] Ryan Martin and Jonathan P Williams, Asymptotic efficiency of inferential models and a possibilistic Bernstein–von Mises theorem, *International Journal of Approximate Reasoning* **180** (2025) 109389.

[33] Muriel F Pérez-Ortiz and Rui M Castro, Anytime-Valid Tests for Sparse Anomalies, *arXiv preprint arXiv:2506.22588* (2025).