INFORMS Advances in Decision Analysis Conference
June 24, 2026

Comparing Sequential Forecasters by Betting

Based on joint work with Aaditya Ramdas (CMU/Stanford)

Yo Joong "YJ" Choe
INSEAD Decision Sciences

Part I: Testing Forecasters By Betting

Growing popularity of betting-based inference in the statistics community

Received: 15 March 2019 | Accepted: 11 May 2020

DOI: 10.1111/rssa.12647

ORIGINAL ARTICLE

Discussion Paper

Testing by betting: A strategy for statistical and scientific communication

Glenn Shafer

Rutgers University, Newark, NJ, USA

Correspondence
Glenn Shafer, Rutgers University, Newark, NJ, USA.
Email: gshafer@business.rutgers.edu

Abstract

The most widely used concept of statistical inference—the *p*-value—is too complicated for effective communication to a wide audience. This paper introduces a simpler way of reporting statistical evidence: report the outcome of a bet against the null hypothesis. This leads to a new role for likelihood, to alternatives to power and confidence, and to a framework for meta-analysis that accommodates both planned and opportunistic testing of statistical hypotheses and probabilistic forecasts. This framework builds on the foundation for mathematical probability developed in previous work by Vladimir Vovk and myself.

KEYWORDS

betting score, game-theoretic probability, likelihood ratio, *p*-value, statistical communication, warranty

Journal of the Royal Statistical Society Series B: Statistical Methodology, 2024, **86**, 1–27
https://doi.org/10.1093/jrsssb/qkad009
Advance access publication 16 February 2023

Discussion Paper

Estimating means of bounded random variables by betting

Ian Waudby-Smith[1] and Aaditya Ramdas[1,2]

[1]Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA

Address for correspondence: Aaditya Ramdas, Machine Learning, Carnegie Mellon University, 5000 Forbes Ave, 132 Baker Hall, Pittsburgh, PA 15213, USA. Email: aramdas@cmu.edu

Read before The Royal Statistical Society at the Discussion Meeting organized by the Research Section on Tuesday, 23 May 2023, Dr Robin Evans in the Chair.

Journal of the Royal Statistical Society Series B: Statistical Methodology, 2024, **86**, 1091–1128
https://doi.org/10.1093/jrsssb/qkae011
Advance access publication 7 March 2024

Discussion Paper

Safe testing

Peter Grünwald[1,2], Rianne de Heide[3] and Wouter Koolen[1,4]

[1]Machine Learning Group, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[2]Mathematical Institute, Leiden University, Leiden, The Netherlands
[3]Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[4]Statistics Group,University of Twente, Enschede, The Netherlands

Address for correspondence: Peter Grünwald, Centrum Wiskunde & Informatica, Science Park 123, Amsterdam 1098 XG, The Netherlands. Email: pdg@cwi.nl

Read before The Royal Statistical Society at a meeting organized by the Research Section on Wednesday, 24th January, 2024, Professor Robin Evans in the Chair.

Warmup: Testing whether a coin is fair

Protocol (Testing whether a coin is fair). **Skeptic** is endowed with a dollar ($M_0 = 1$).

For rounds $t = 1, 2, \ldots$:

1. **Skeptic** announces a signed "bet" $\beta_t \in [-1, 1]$.
2. **Nature** flips the coin and reveals the outcome $y_t \in \{-1, 1\}$.
3. **Skeptic** ends up with wealth $M_t = M_{t-1} \cdot [1 + \beta_t y_t]$.

1. *If the coin is fair, Skeptic is not expected to grow his wealth (M_t is a martingale).*
2. *If the coin is biased, Skeptic can get **rich** by strategically choosing the bets!*

The general framework: Testing-by-betting 🎲

Protocol (Testing a probability by betting). **Skeptic** is endowed with a dollar ($\mathsf{M}_0 = 1$).

Forecaster announces a probability distribution $\mathsf{P} \in \mathscr{P}(\mathscr{Y})$. Then, for rounds $\mathsf{t} = 1, 2, \ldots$:

1. **Skeptic** announces a *betting function* $\beta_t : \mathscr{Y} \to \mathbb{R}$ that satisfies $\mathbb{E}_{\mathsf{P}}[\beta_t(\mathsf{Y})] \leq 1$.
2. **Nature** reveals the outcome $\mathsf{y}_t \in \mathscr{Y}$.
3. **Skeptic** ends up with wealth $\mathsf{M}_t = \mathsf{M}_{t-1} \cdot \beta_t(\mathsf{y}_t)$.

In the previous example, we had
$\mathsf{P} \equiv \mathsf{Ber}(1/2),\ \beta_t(\mathsf{Y}) = 1 + \beta_t \mathsf{Y}.$

(M_t) *forms a* ***test (super)martingale*** *under the protocol's filtration.*

On the other hand, if $\mathsf{Y}_t \overset{\text{iid}}{\sim} \mathsf{Q} \neq \mathsf{P}$*, then Skeptic has a betting strategy* (β_t) *to get rich!*

Shafer's fundamental principle of testing-by-betting

🎲 **Fundamental Principle of Testing-By-Betting:**

In the protocol, the skeptic ***discredits*** P *to the extent that* M_t *is large.*

- *Evidential interpretation of wealth:* (M_t) *is a* ***"graduated appraisal of evidence"*** *against* P.
  - *"If this bet multiplies the money it risks by a large factor, we have evidence against the hypothesis, and the factor measures the strength of evidence." (Shafer, JRSSA'21)*
  - *"Multiplying our money by 5 might merit attention; multiplying it by 100 or 1000 might be considered conclusive." (Shafer, JRSSA'21)*
- *Key rules: Skeptic's bets must be* ***predictable*** *at each round* **t**, and they cannot bet more than what they have at each round **(no-bankruptcy)**.

*Connections to likelihood ratios

- Under the rules of predictability & no-bankruptcy, (M_t) is a test (super)martingale.
- Now, say the betting odds are $(1-p) : p$, corresponding to the claim that $Y_t \sim \text{Ber}(p)$.
- Skeptic instead believes that $Y_t \sim \text{Ber}(q)$, $q \neq p$. How should they bet?

Growth-rate optimal bet (Kelly, 1956): $\beta_t^{\text{GRO}} = \text{argmax}_{\beta_t \in [-1,1]} \mathbb{E}_q[\log \beta_t] = 2q - 1.$

The resulting bet is simply the **likelihood ratio:** $\beta_t^{\text{GRO}}(Y) = \dfrac{q^Y(1-q)^{1-Y}}{p^Y(1-p)^{1-Y}}.$

- In fact, any betting strategy induces an *implied alternative*.

E-value/e-process as statistical evidence

To highlight the role of test martingales as *evidence*, the literature shifted to the terms **e-value** (single round) and **e-process** (sequential rounds). Key characteristics (relatively):

Compared to p-values...

1. *Both achieve **frequentist validity** (type I error control in repeated trials).*
2. *Unlike usual p-values, e-values can **flexibly handle optional stopping.***
3. *Multiple e-values can be **combined easily** under arbitrary dependence.*

Compared to Bayes Factors...

1. *When built correctly, both can **flexibly handle optional stopping.***
2. *Unlike Bayes factors, e-values always achieve **frequentist validity.***
3. *E-values are often **"distribution-free"** (both handle composite hypotheses).*

Part II: Comparing Sequential Forecasters By Betting

Based on: Y. J. Choe & A. Ramdas. “Comparing Sequential Forecasters.” *Operations Research* (2023).

What does this have to do with forecasters?

- A forecaster is the **bookmaker** in the hypothetical betting game.
  - *In his book, Shafer explicitly uses the term "forecaster" as the player proposing* P*.*
  - *In testing terms, the forecaster proposes the null hypothesis.*
- In practice, forecasters (e.g., meteorologists) make **sequential** predictions.
  - *This is very natural in the betting game framework & makes the forecaster's role explicit.*
  - *This (eventually) leads to a flexible and interpretable framework for testing forecasters.*

Testing a sequential forecaster by betting: The binary case

Suppose a **Skeptic** tests a **Forecaster** making a sequence of predictions on binary outcomes.

Protocol (Testing a sequential forecaster). **Skeptic** is endowed with $M_0 = 1$. For $t = 1, 2, \ldots$:

1. **Forecaster** announces a probability distribution $p_t \in [0, 1]$.
2. **Skeptic** announces a bet $\beta_t(y) = 1 + \beta_t(y - p_t)$, $\beta_t \in [-c_0, c_1]$. Note that $\mathbb{E}_{p_t}[\beta_t(Y)] = 1$.
3. **Nature** reveals the outcome $y_t \in \{0, 1\}$.
4. **Skeptic** ends up with wealth $M_t = M_{t-1} \cdot \beta_t(y_t)$.

$c_0, c_1 > 0$ are chosen s.t. Skeptic cannot go bankrupt.

Under any law $\mathbb{P}$, the forecasts p_t have to be ***conditionally calibrated:***

$$H_0 : \mathbb{E}_{\mathbb{P}}[Y_t \mid \mathcal{F}_{t-1}, p_t] = p_t, \ \forall t\,.$$

Comparative *ex post* evaluation with scoring rules

- Essentially, for well-parametrized forecasts, we can test for **conditional calibration.** (Giacomini & White, 2006; Lai et al., 2011)

- Beyond this (e.g., distributional/nonparametric forecasts), we need **(proper) scoring rules.** (Winkler & Murphy, 1968; Gneiting & Raftery, 2007; …)

- In practice, we are usually interested in ***relative* proper scores** against a baseline.

1. *Testing **pointwise** score differences (HZ'22):* $\delta_t = \mathbb{E}[s(P_t, Y) - s(Q_t, Y) \mid \mathcal{F}_{t-1}, P_t, Q_t]$

2. *Testing **mean** score differences (CR'24):* $\Delta_t = \frac{1}{t} \sum_{i=1}^{t} \delta_i$

Null hypothesis = Forecaster P is no better than Q

- For such betting games to make sense, we implicitly bring in a bookmaker on the forecasters (a "meta-forecaster") claiming that ***P will perform no better than Q over time.***
- **The exact form of the betting function** determines the specific form of the null hypothesis.
  - ***Pointwise dominance:*** $\beta_t(d_t) = 1 + \beta_t d_t$, *where* $\beta_t \in [0, 1]$.
  - ***Average-sense (mean) dominance, assuming bounded scores:***

$$\beta_t(d_t) = \exp\left\{\beta d_t - \psi_E(\beta)(d_t - D_{t-1})^2\right\}, \text{ where } D_{t-1} = \frac{1}{t-1}\sum_{i=1}^{t-1} d_i \,.$$

*$\psi_E(\beta) = -\log(1-\beta) - \beta$

Comparing sequential forecasters by betting

Game (Comparing Sequential Forecasters). Let $M_0 = 1$. For rounds $t = 1, 2, \ldots$:

1. **Forecasters 1 & 2** each announce a forecast, $P_t \in \mathscr{P}$ and $Q_t \in \mathscr{P}$, respectively.
2. **Bookmaker** announces a *negative* mean score difference* as $\delta_t \leq 0$.
3. **Skeptic** announces a bet $\beta : \mathbb{R} \to \mathbb{R}_{\geq 0}$ satisfying $\mathbb{E}\left[\beta(D_t) \mid \mathscr{F}_{t-1}, \delta_t\right] \leq 1$.
4. **Nature** announces the outcome $y_t \in \{0, 1\}$, determining $d_t = s(P_t, y_t) - s(Q_t, y_t)$.
5. **Skeptic's** wealth is updated: $M_t = M_{t-1} \cdot \beta(d_t)$.

*We make our betting choice independent of δ_t:
Choose some $\beta \in [0, 1]$ s.t. $\beta(d_t) = \exp\{\beta d_t - \psi_E(\beta)(d_t - D_{t-1})^2\}$.

Main result #1: Sequential comparison of time-varying forecast scores

- Let Δ_t denote the difference in *mean* conditional scores over time:

$$\Delta_t = \frac{1}{t}\sum_{i=1}^{t} \delta_i = \frac{1}{t}\sum_{i=1}^{t} \mathbb{E}\left[s(p_i, Y_i) - s(q_i, Y_i) \mid \mathcal{F}_{i-1}, p_i, q_i\right].$$

- Let $\hat{\sigma}_t^2 = (1/t)\Sigma_{i=1}^{t}(d_i - D_{i-1})^2$ denote the running variance estimate.

- Let H_0 be the null hypothesis that *"P is no better than Q on average", or* $H_0 : \Delta_t \leq 0, \forall t$.

Theorem. Let s be a scoring rule bounded in $[0, 1]$ (e.g., Brier). Then:

(a) $M_t := \int_0^1 \exp\left\{\lambda t D_t - \psi_E(\lambda) t \hat{\sigma}_t^2\right\} dF(\lambda)$ is an **e-process** for $H_0 : \Delta_t \leq 0, \forall t$.

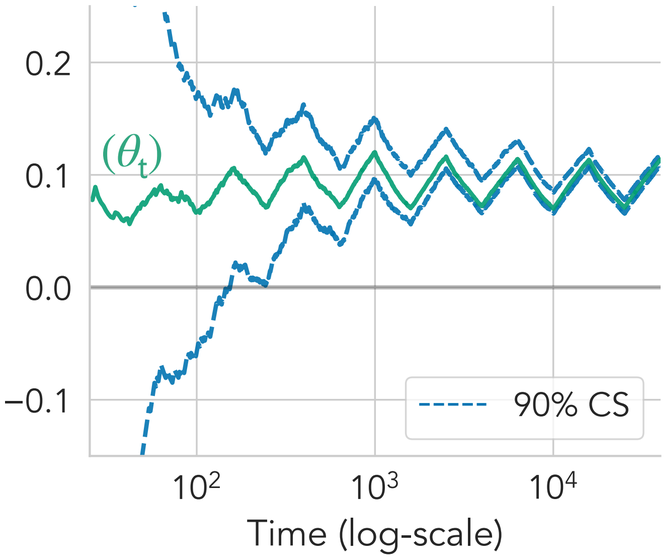
M_t has a closed form when F is a Gamma distribution.

(b) Under any alternative $\mathbb{Q}$ (that is, $\exists t : \Delta_t > 0$), (M_t) yields a **test of power one.**

Main result #2: Continuous estimation of time-varying forecast scores

- For parameter estimation, we may use **confidence sequences (CS):**

$$(1-\alpha)\text{-level CS } (C_t)_{t\geq 1} \text{ for } (\theta_t) \text{ iff } P\left(\forall t \geq 1 : \theta_t \in C_t\right) \geq 1-\alpha.$$



- In our case, the parameter of interest is a *time-varying mean* (Δ_t).

Theorem. Let s be a scoring rule bounded in $[0, 1]$ (e.g., Brier). For $\alpha \in (0, 1)$,

$$C_t := \left(D_t \pm c_\alpha \hat{\sigma}_t \cdot \sqrt{\frac{\log\log t}{t}} \right) \text{ forms a } (1-\alpha)\text{-level CS for } \Delta_t.$$

($c_\alpha \asymp \sqrt{\log(1/\alpha)}$ is a constant.)

*Handling the log score with Winkler's normalization

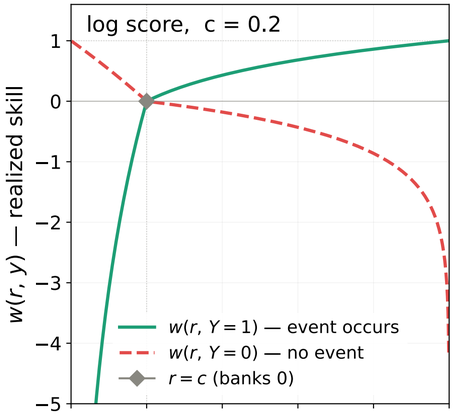
- Let **s** be any proper scoring rule on binary forecasts (e.g., the log score), and

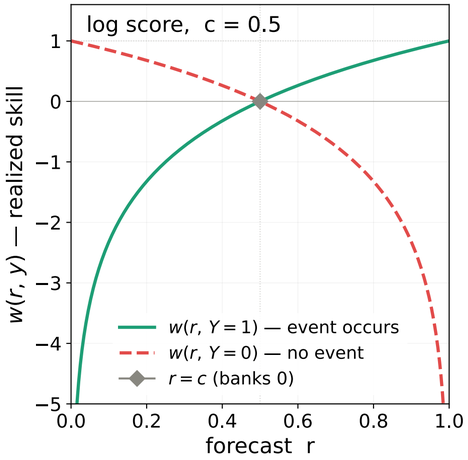
$$\text{“Winkler score”} = w(p, y; c) = \frac{s(p, y) - s(c, y)}{T(p, c)}.$$

where $T(p, c) = [s(1, 1) - s(c, 1)]1(p \geq c) + [s(0, 0) - s(c, 0)]1(p < c)$.

- For any fixed $c \in (0, 1)$, **w** is an *asymmetric* proper score (Winkler, 1994), measuring the "net demonstrated skill over the base rate" (perfect = +1).

- We can then test **whether P underperforms Q in mean Winkler-log score:**

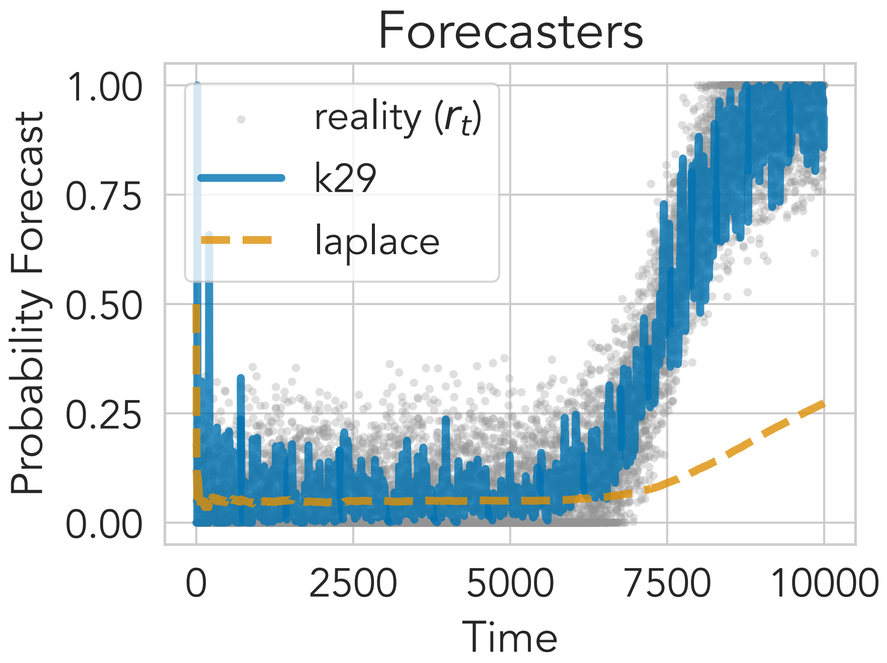


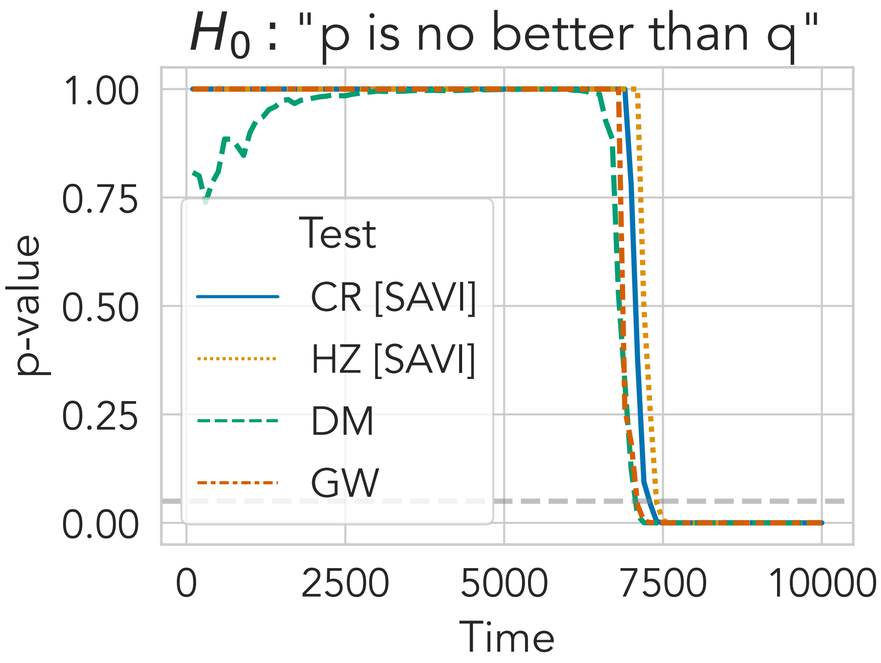


Proposition. Let $H_0 : W_t \leq 0, \forall t$, where $W_t = t^{-1}\Sigma_{i=1}^{t}\mathbb{E}[w(p_i, Y_i; c) \mid \mathcal{F}_{i-1}, p_i]$.
Then, we can derive an (analogous) **e-process** and a **test of power one for H_0.**

Statistical power relative to classical tests

Betting/SAVI (CR & HZ) vs. Diebold-Mariano (DM) & Giacomini-White (GW)





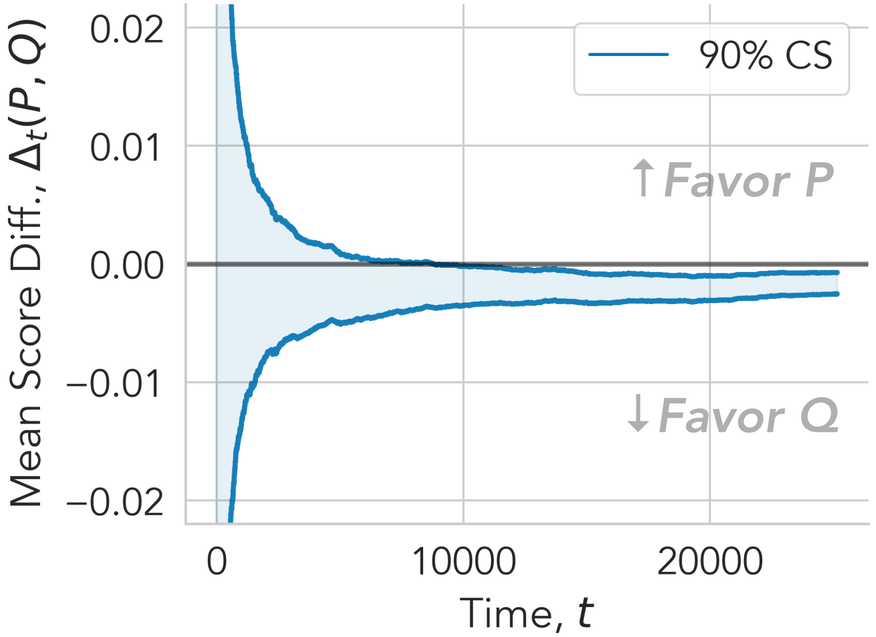
Among these, only the "SAVI" methods have validity under optional stopping.

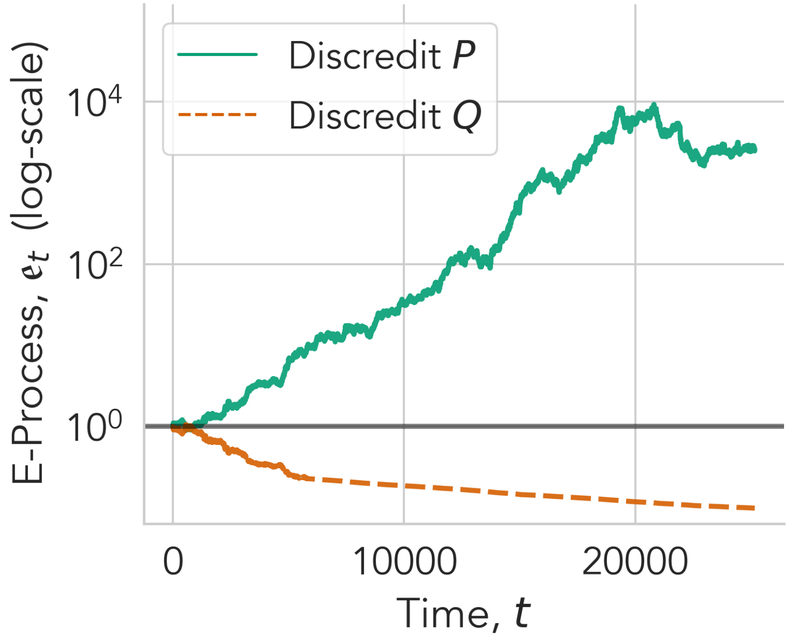
Choe & Ramdas (CR): this work
Henzi & Ziegel (HZ): concurrent work on testing a "strong null"

Data: Every MLB game's win/loss outcomes from 2010 to 2019.

Illustration on real-world baseball forecasters

FiveThirtyEight *(P)* vs. Vegas odds *(Q)* on MLB games. Scoring rule = Brier





Confidence Sequence (C_t)

E-Process (e_t)

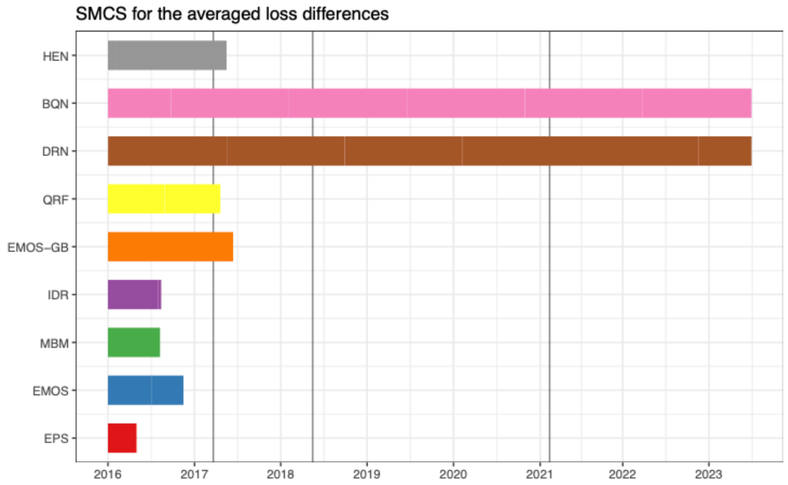
*$H_0 : \Delta_t(P, Q) \leq 0, \forall t$

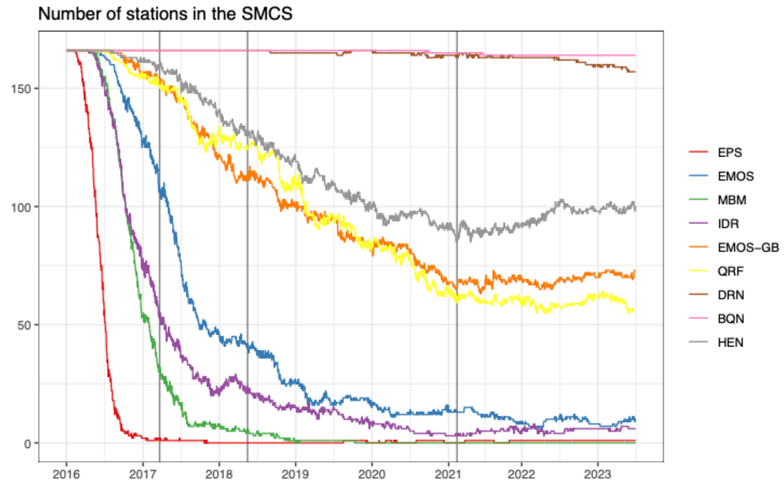
Part III: Concluding Remarks

Extensions

- **Comparing multiple forecasters**

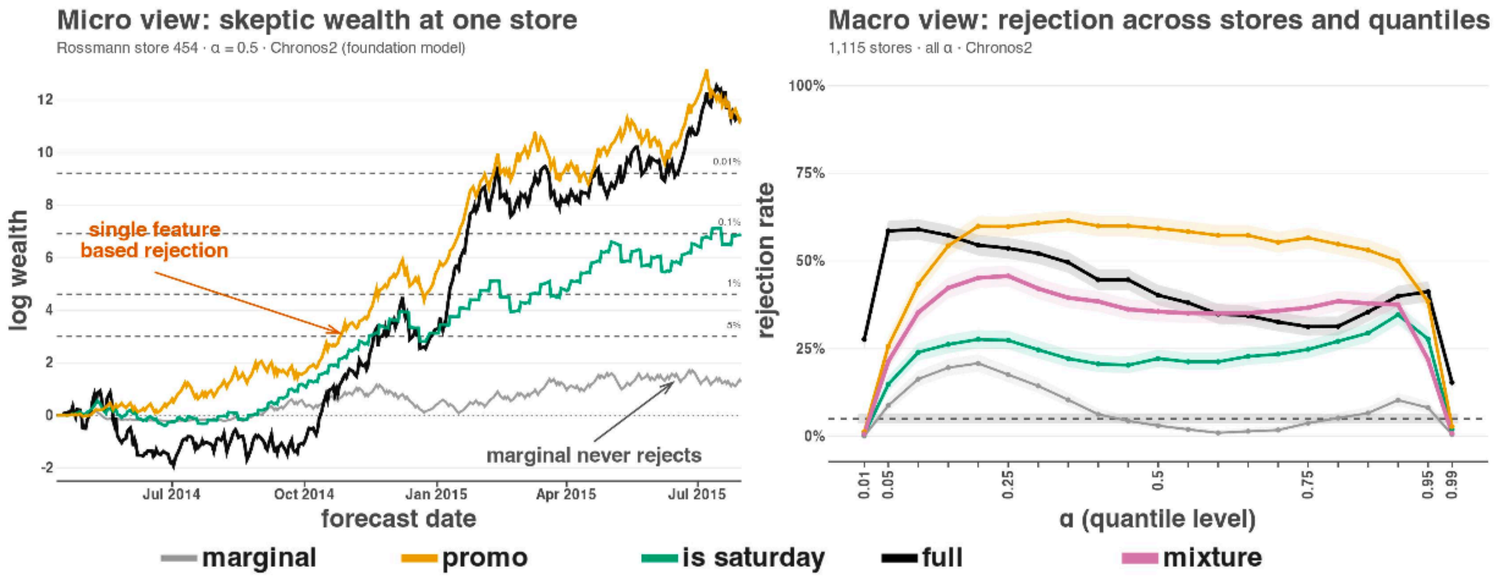
  Arnold*, Gavrilopoulos*, Schulz, & Ziegel. "Sequential model confidence sets." *JRSS-B* (2026).





- **Feature-aware audits for sequential quantile forecasters**

  Antonov*, Mukherjee*, Pibernik, & Choe. "Bet on features: Anytime-valid and feature-aware auditing of conditional quantile forecasters." *Working paper* (2026).



Summary

- A betting-based approach to sequential forecaster comparison under various scoring rules
- *Anytime-valid:* controls false positives under optional stopping & continuous monitoring
- *Tight estimation & test of power one:* CS has optimal rate up to log terms
- *"Worry-free":* No assumptions on forecasters or outcomes + flexible choice of sample sizes

Thank You

Any Questions?

For more, visit https://yjchoe.github.io/

Appendix

See, e.g., Ramdas & Wang (2025)

Key desiderata: Anytime-validity & positive growth rate

- **Anytime-validity:** type I error control under optional stopping (or continuous monitoring).

At any stopping time τ, $\mathbb{E}_P[M_\tau] \leq 1$.

(Ville's inequality) At any $\alpha \in (0,1)$, $P\left(\exists t \geq 1 : M_t \geq 1/\alpha\right) \leq \alpha$.

- *Note: By Ville's inequality,* $\mathfrak{p}_t = 1/M_t$ *is an "anytime-valid" p-value.*

- **Positive growth rate** under the alternative: wealth grows exponentially fast ("e-power").

Under any alternative Q, we want $\mathbb{E}_Q[\log(M_t/M_{t-1})] = \mathbb{E}_Q[\log \beta_t] > 0$ at each t.

(Robbins' test of power one) For $\tau_\alpha = \inf\{t : M_t > 1/\alpha\}$, $Q(\tau < \infty) = 1$.

Composite generalizations of likelihood ratios

Simple-vs-Simple LR

$$\frac{q(Y_1, \ldots, Y_t)}{p(Y_1, \ldots, Y_t)}$$

Frequentist LR

$$\frac{\max_{q \in \mathcal{Q}} q(Y_1, \ldots, Y_t)}{\max_{p \in \mathcal{P}} p(Y_1, \ldots, Y_t)}$$

Game-Theoretic Wealth ("E-Process")

$$\frac{\int_{\mathcal{Q}} q(Y_1, \ldots, Y_t)\pi_t(q \mid \mathcal{Q})dq}{\max_{p \in \mathcal{P}} p(Y_1, \ldots, Y_t)}$$

Bayes Factor

$$\frac{\int_{\mathcal{Q}} q(Y_1, \ldots, Y_t)\pi(q \mid \mathcal{Q})dq}{\int_{\mathcal{P}} p(Y_1, \ldots, Y_t)\pi(p \mid \mathcal{P})dp}$$

*π_t can be either a prior or a learned mixture over time

Bonus from Winkler (1977)

REWARDING EXPERTISE IN PROBABILITY ASSESSMENT[+]

Robert L. Winkler
Graduate School of Business
Bloomington, U.S.A.

1. Introduction

Experts are valuable sources of information for individuals or groups with decision-making problems. Moreover, the theory of personal, or subjective, probability, as developed by de Finetti (1937) and Savage (1954), provides a framework within which experts can represent their uncertainties in a quantitative fashion. Morris (1974, pp. 1233-1234) writes as follows:

> It is a rare decision that is not made in the context of significant uncertainty. In attempting to resolve this uncertainty a decision maker often must rely upon the judgment of one or more other persons. We shall refer to such a person who provides a judgment concerning uncertain matters as an expert ... The most detailed and most interesting representation of an expert's judgment pertaining to an uncertain quantity is the probability function he assigns to it.

the expert is penalized for any dishonesty. Given a fixed p, the expert has no control over sharpness. Thus, the only way to improve the expected score is to obtain further information about the situation of interest in an attempt to change p in the direction of increasing sharpness. In this sense, proper scoring rules encourage the acquisition of expertise. Since the acquisition of expertise seems desirable, this result indicates that the encouragement of honesty is not the only raison d'être for scoring rules in probability assessment.

In addition to their role in probability assessment, scoring rules also play an important role in probability evaluation. Probability evaluation involves ex post considerations (i.e., considerations after the event or variable of interest has been observed), while the concern in this paper has been with ex ante considerations. It is possible, of course, for an individual to have high expected scores that are not realized ex post, so it is important not to make the mistake of giving ex post interpretations to the results developed here. Given certain qualifications, some ex post implications can be drawn from the ex ante results, and a brief sketch of some implications might help to put matters in proper perspective. The link between ex ante and ex post results is reliability, or calibration: an expert is said to be reliable if in a large number of situations for which the expert's probability of an event is p, the relative frequency of occurrence of the event is also p. If an expert is reliable, then the expert's average actual score will equal the average expected score. Given reliability, then, sharpness is re-

Unifying developments across (very) different fields

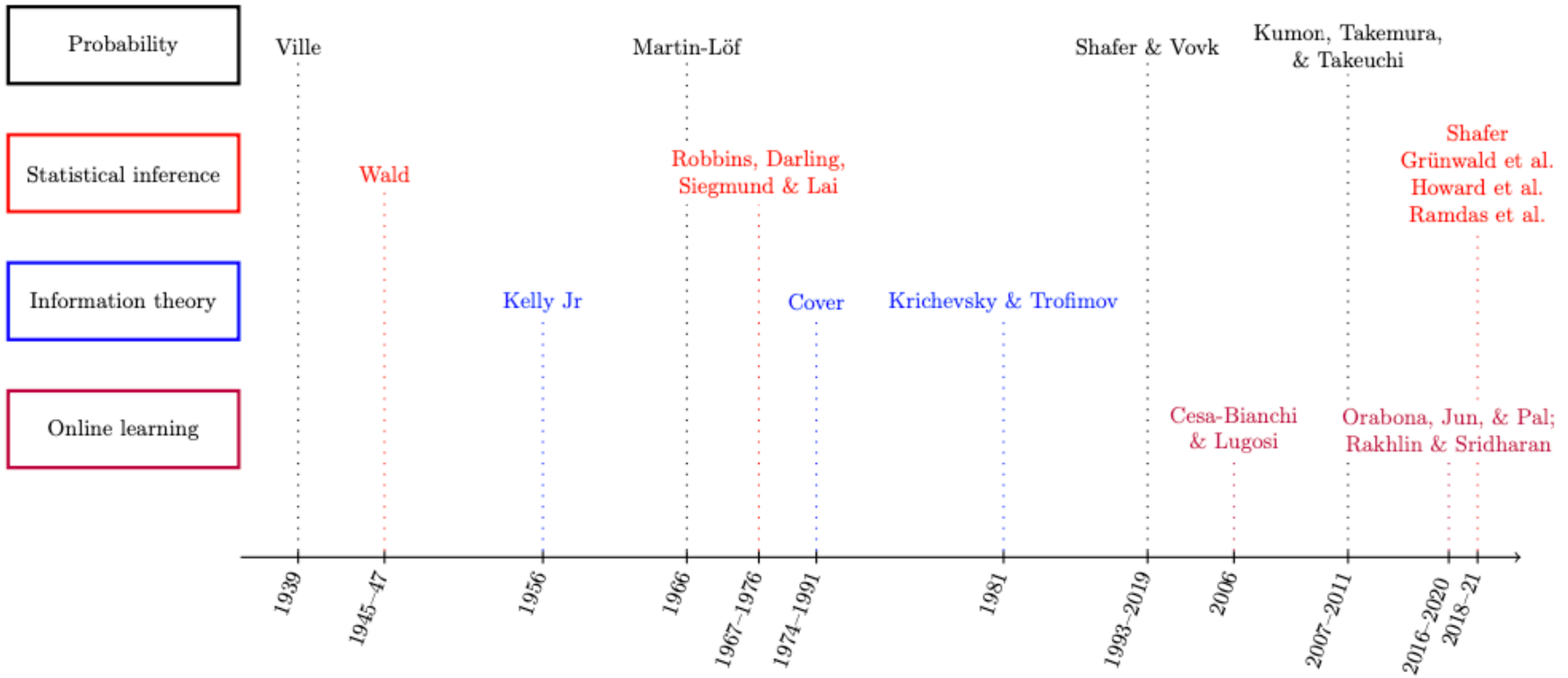
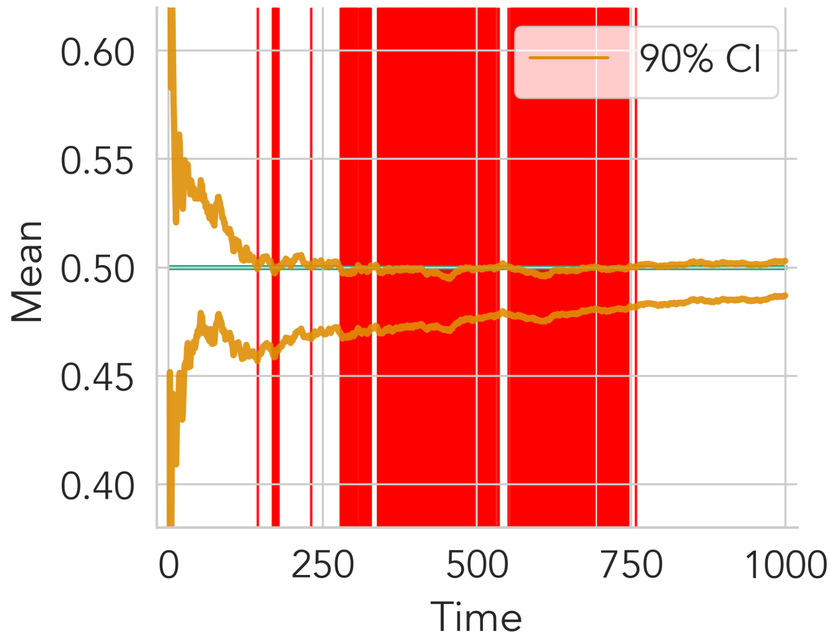


Figure 9: A brief selective history of betting ideas appearing (often implicitly) in various literatures. As discussed further in Section F, these subfields have rarely cited each other, but ideas are now beginning to permeate. Several authors did not explicitly use the language of betting, and their inclusion above is due to reinterpreting their work in hindsight.

Continuous Monitoring Invalidates Classical Statistical Guarantees

- Suppose we estimate a score difference *by constructing a new CI after every new outcome.*
- ***The longer you experiment, the more likely you'll run into a miscoverage (at least once)!***
- Analogous phenomenon for **optional stopping.**



Green: target. Red: miscoverage.

*Comparison with frequentist sequential methods

1. **Group-sequential / α-spending / naive correction methods**
   - *Requires a pre-specified total sample size / maximum number of interim checks…*
   - *…or is very conservative under continuous monitoring*
2. **Wald's sequential probability ratio test**
   - *Achieves validity at a particular stopping time (& requires a likelihood function)*

Existing methods can be viable in some cases; however,
they are NOT generally flexible (or powerful) for **continuous monitoring.**

E-value: "E is the new P"

- Given **n** data points $X_1, \ldots, X_n$ and a null hypothesis H_0 (possibly composite), an **e-value** $E = E_n(X_1, \ldots, X_n)$ is any non-negative random variable satisfying:

$$\mathbb{E}_{H_0}[E] \leq 1.$$

e.g., $E = \prod_{i=1}^{n} \frac{q(X_i)}{p(X_i)}$

- **E-values can be used for testing H_0:** for any $\alpha \in (0, 1)$, by Markov's inequality,

$$P(E \geq 1/\alpha) \leq \alpha, \quad \forall P \in H_0.$$

- **For intersection nulls, e-values can be combined easily under arbitrary dependence:**
  Given e-values $E^{(z)}$ for $H_0^{(z)}$, their **(weighted) average** is an e-value for $H_0 = \cap_{z \in Z} H_0^{(z)}$:

$$\mathbb{E}_{H_0}\left[\sum_{z \in Z} w^{(z)} E^{(z)}\right] = \sum_{z \in Z} w^{(z)} \mathbb{E}_{H_0}\left[E^{(z)}\right] \leq 1.$$

A key benefit of using e-values over p-values!

E-processes quantify evidence in sequential experiments

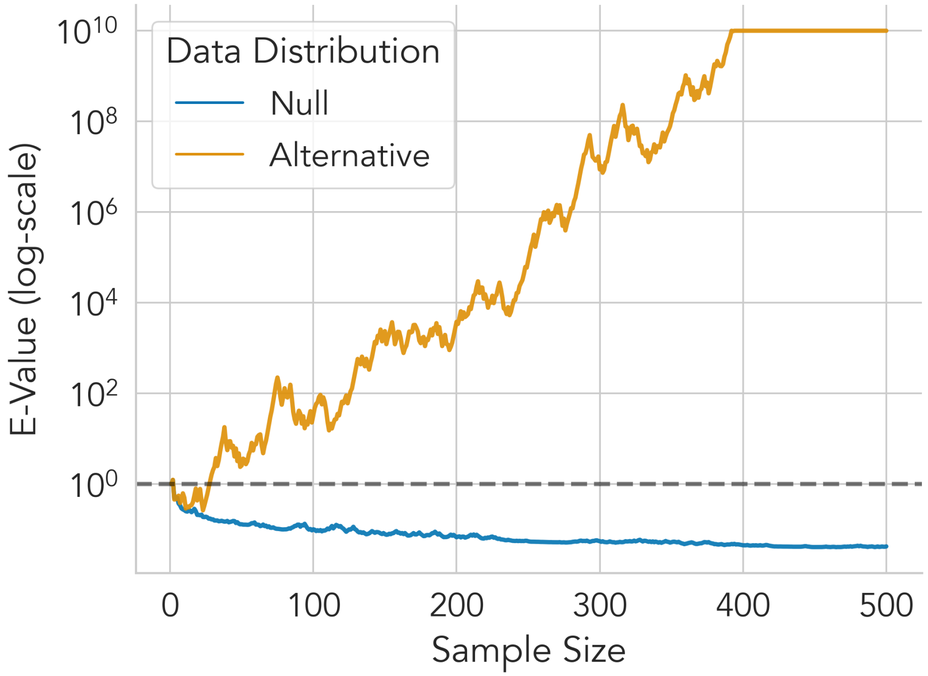
E-process $(E_t)_{t\geq 0}$

A non-negative process for H_0

For any stopping time τ,

$$\mathbb{E}_{H_0}[E_\tau] \leq 1.$$

"ANYTIME-VALIDITY"



An e-process is expected to be small under the null.
We want it to grow large under the alternative.

Ville's inequality: From e-processes to sequential tests

- Let $\alpha \in (0, 1)$ be any significance level.
- **Ville's inequality** for test martingales & e-processes:

$$P\left(\exists t \geq 1 : E_t \geq 1/\alpha\right) \leq \alpha, \ \forall \alpha \in (0, 1).$$

Jean Ville

- This is **equivalent** to a time-uniform guarantee for sequential testing:

$$P\left(\exists t \geq 1 : E_t \geq 1/\alpha\right) \leq \alpha, \ \forall \alpha \in (0, 1).$$

🎲 **Game-Theoretic View:** If the proposed odds (H_0) were accurate, then it is unlikely ($\leq$ 0.1 chance) to see my wealth grow by a lot ($\geq$10 fold), ***ever.***

Testing a binary forecaster in a betting game, single round

Suppose a **Skeptic** wants to test a **Forecaster** for a binary outcome $Y \in \{0, 1\}$ (e.g., rain).

Protocol (Testing a forecaster by betting). **Skeptic** is endowed with a dollar to make a bet.

1. **Forecaster** announces the *betting odds* $(1 - p) : p$, depending on the outcome.
2. **Skeptic** announces a fraction $\lambda \in [0, 1]$ of money to be placed on $Y = 1$; the rest goes to $Y = 0$.
3. **Nature** reveals the outcome $y \in \{0, 1\}$.
4. **Skeptic** ends up with wealth $S(y; \lambda) = (1 - y)\dfrac{1 - \lambda}{1 - p} + y\dfrac{\lambda}{p} = 1 + \gamma(y - p)$, where $\gamma = \dfrac{\lambda - p}{p(1 - p)}$.

Testing one sequential forecaster by betting on proper scores

Skeptic and **Forecaster** agree to play the game, given **a proper score** $s : \mathscr{P}(\mathscr{Y}) \times \mathscr{Y} \to \mathbb{R}$.

Protocol (Testing a sequential forecaster). **Skeptic** is endowed with $M_0 = 1$. For $t = 1, 2, \ldots$:

1. **Forecaster** announces a probability distribution $P_t \in \mathscr{P}(\mathscr{Y})$.
2. **Skeptic** announces a *betting function on scores* β_t s.t. $\mathbb{E}[\beta_t(s(P_t, Y), \text{past}) \,|\, P_t, \text{past}] \leq 1$.
3. **Nature** reveals the outcome $y_t \in \mathscr{Y}$.
4. **Skeptic** ends up with wealth $M_t = M_{t-1} \cdot \beta_t(s(P_t, y_t), \text{past})$.

Ideally, large wealth = evidence of deficient proper score relative to the oracle.

CS vs. CI: Cumulative Miscoverage & Interval Width

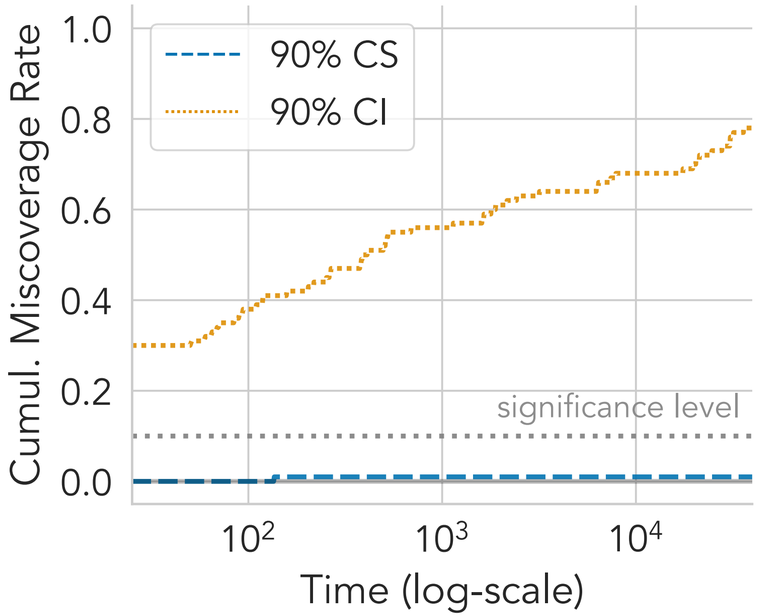
For any time-varying sequence of parameters $(\Delta_t)_{t \geq 1}$,

CS $\quad P\left(\forall t : \Delta_t \in C_t\right) \geq 1 - \alpha.$

CI $\quad \forall t, P\left(\Delta_t \in C_t\right) \geq 1 - \alpha.$

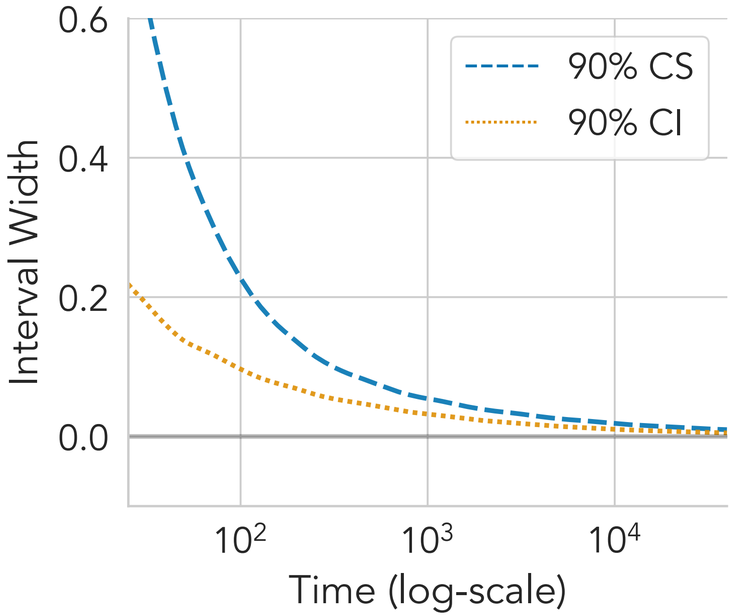
Cumulative Miscoverage Rate

$$\text{miscov}_t = \hat{\mathbb{P}}\left(\exists i \leq t : \Delta_i \notin C_i\right)$$



Interval Width

$$U_t - L_t, \text{ where } C_t = (L_t, U_t)$$



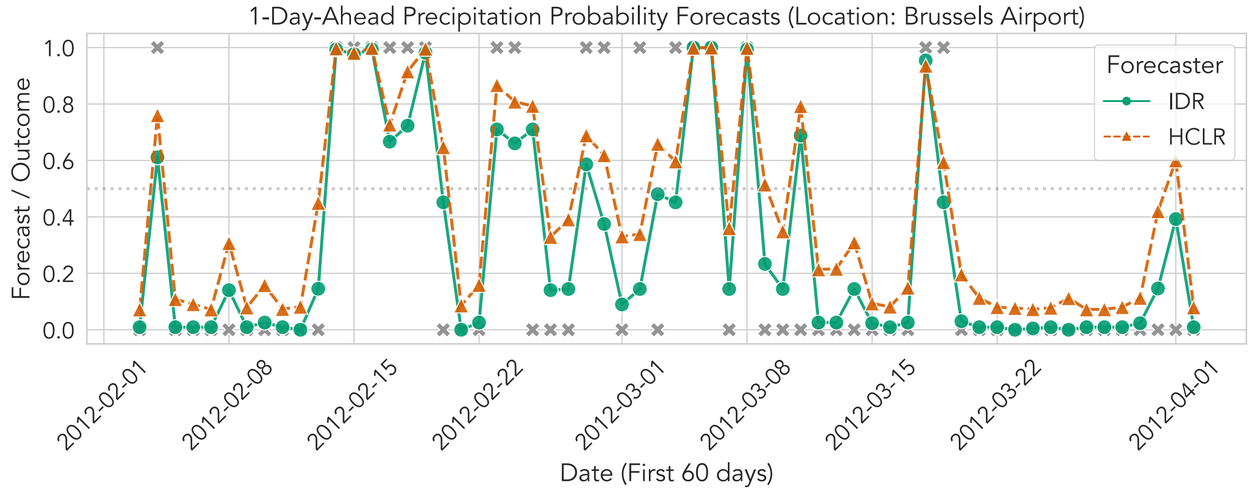
*Fixed-time CI due to Lai et al. (2011).

Data source: ECMWF; experiment adapted from Henzi & Ziegel (2022).

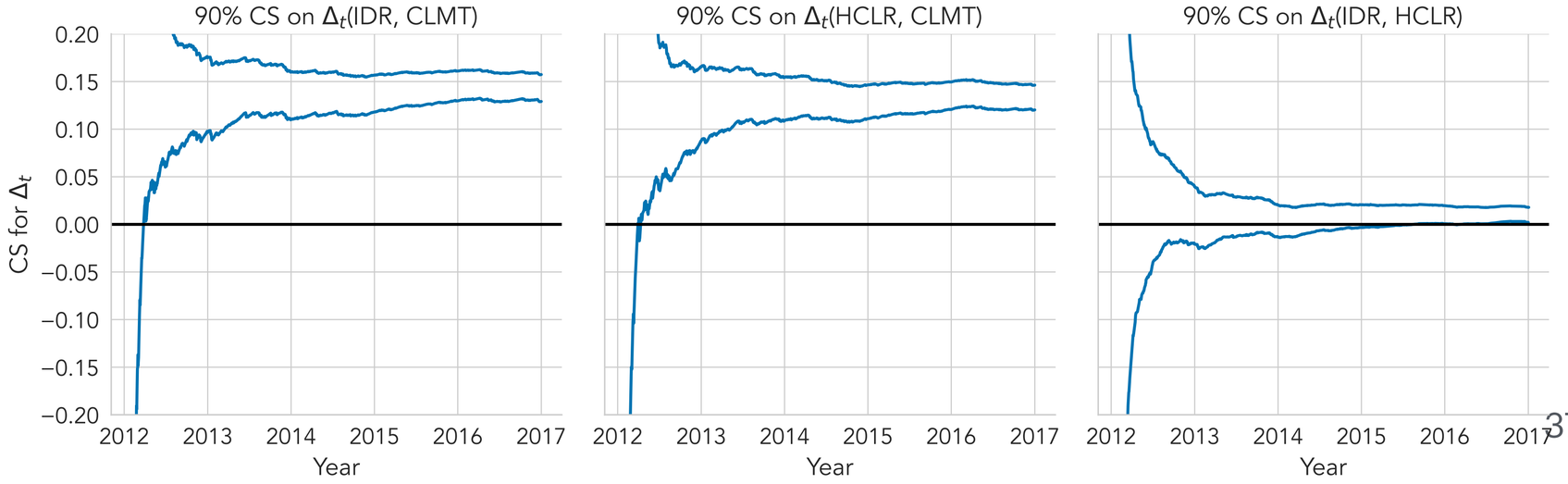
Illustration on ensemble weather forecasters ☔

Data: Daily precipitation forecasts & outcomes at major European airports from 2012 to 2017

Forecasts & Outcomes:
(first 60 days)

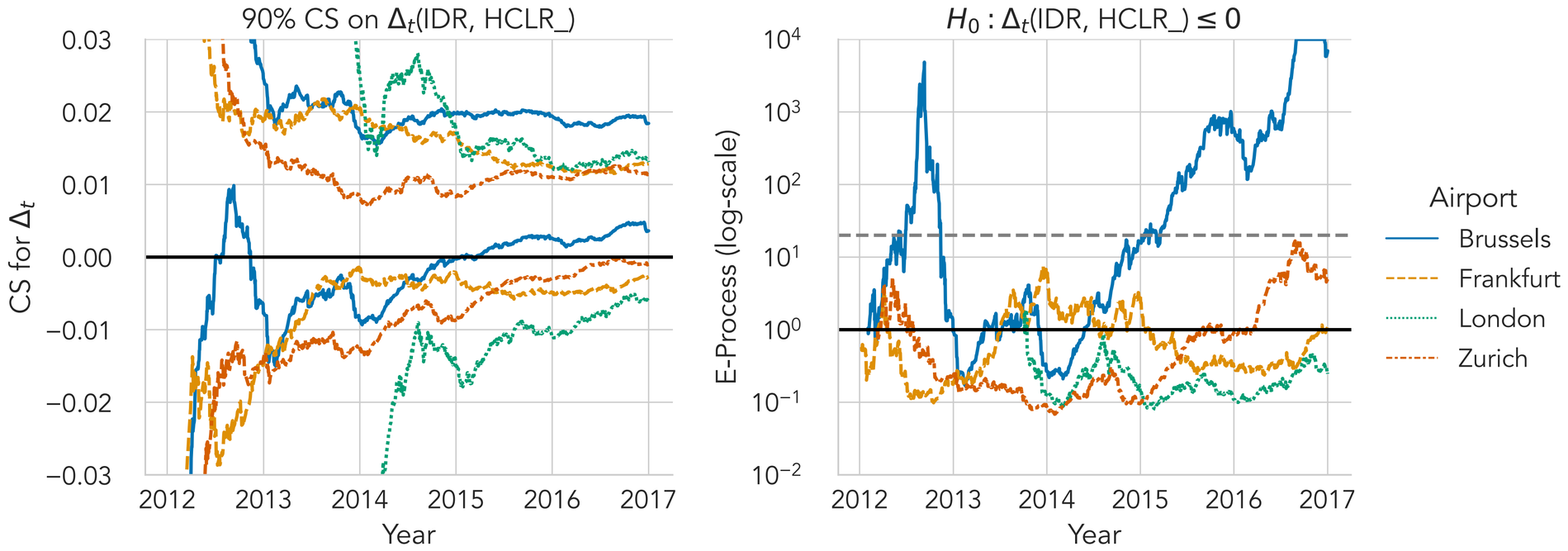


90% CS:
(incl. comparisons
w/ climatology baseline)



Comparing Ensemble Weather Forecasters ☔

Daily precipitation forecasts & outcomes at four major European airports (2012-2017)



Confidence Sequence (C_t)

E-Process (e_t)

*$H_0 : \Delta_t(P, Q) \leq 0, \forall t$

Data source: ECMWF; experiment adapted from Henzi & Ziegel (2022).

Game-theoretic vs. classical statistics/probability

Game-Theoretic View	Classical View
Forecaster's probability (or set of probabilities)	Null hypothesis (either simple or composite/imprecise)
Skeptic's choice of the betting function	Alternative hypothesis
Skeptic's wealth process	A composite likelihood ratio (a nonnegative martingale starting at 1)
Inverse of skeptic's wealth	(Conservative) P-value
Games for which skeptic's wealth can grow large (Huygens, 1600s; Ville, 1939; Ruf et al., 2022; ...)	Events of small probability

The End