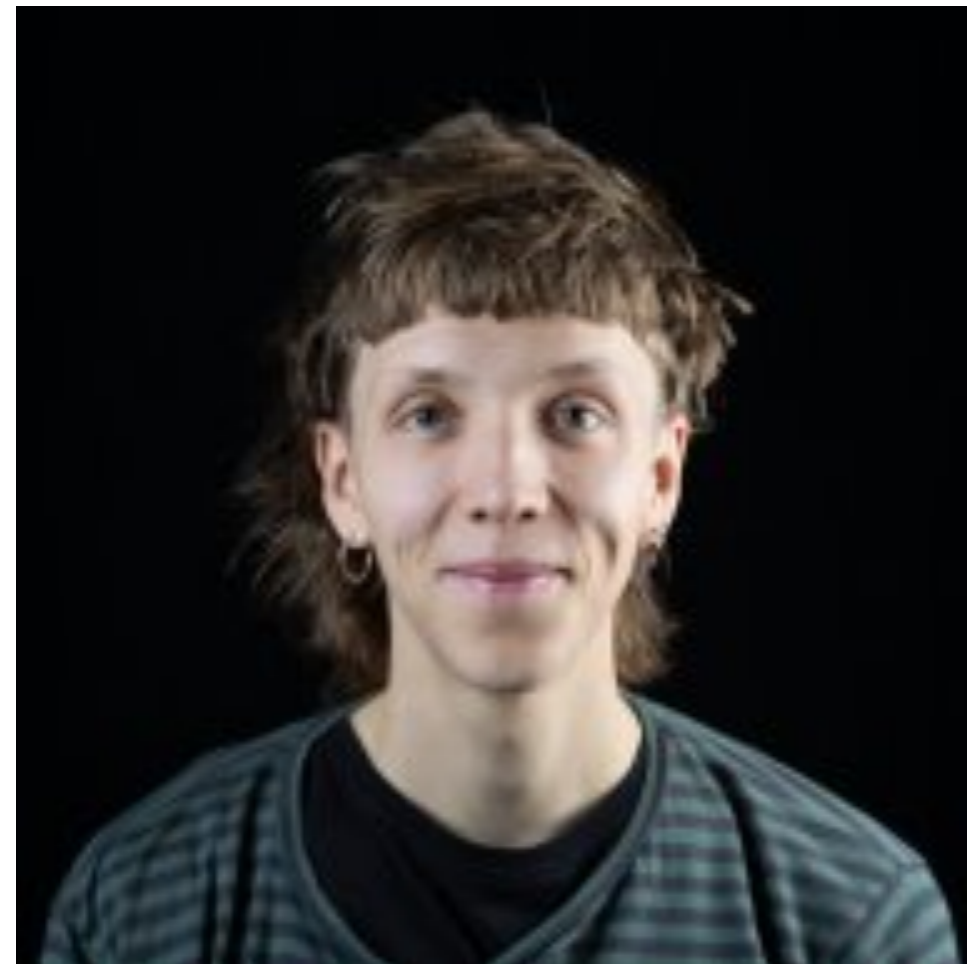


Betting on Bets: **Anytime-Valid Tests for Stochastic Dominance**

Sebastian Arnold, CWI Amsterdam
Yo Joong “YJ” Choe, INSEAD



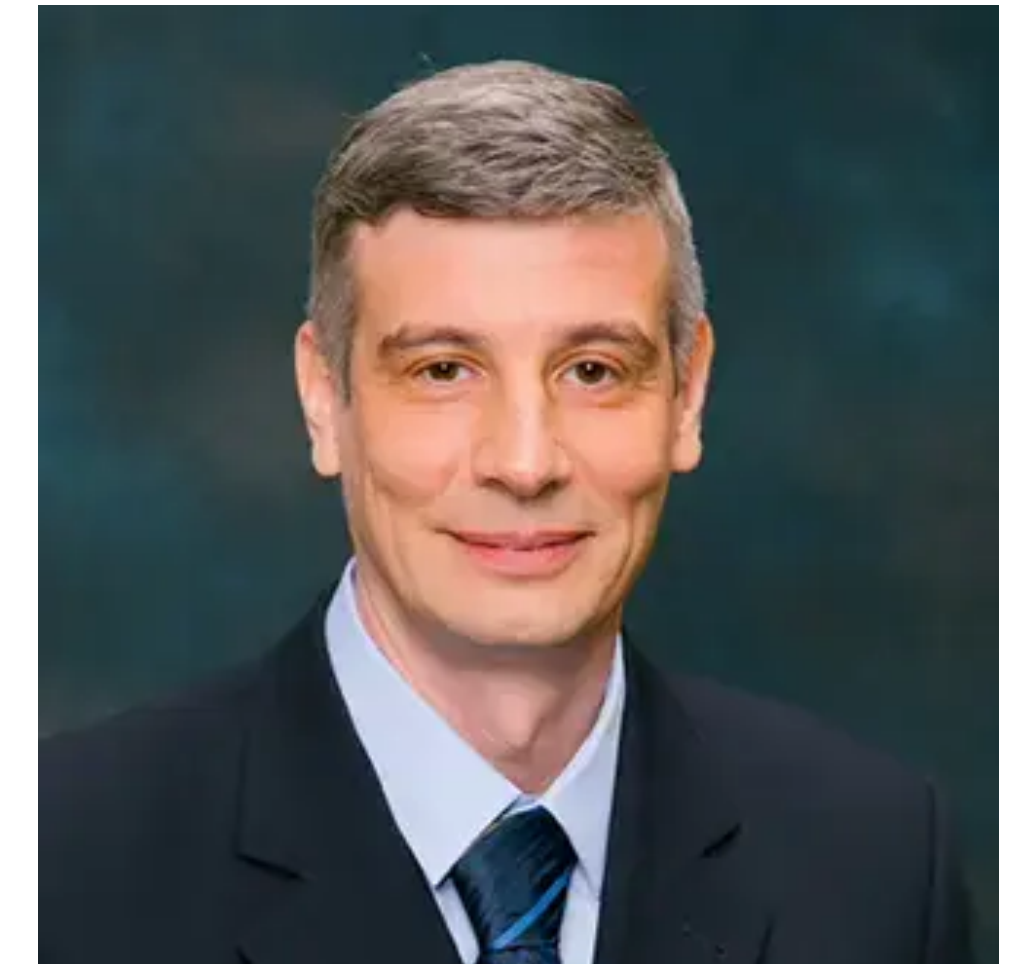
Sebastian Arnold*
CWI Amsterdam



YJ Choe*
INSEAD



Marco Scarsini
LUISS



Ilia Tsetlin
INSEAD

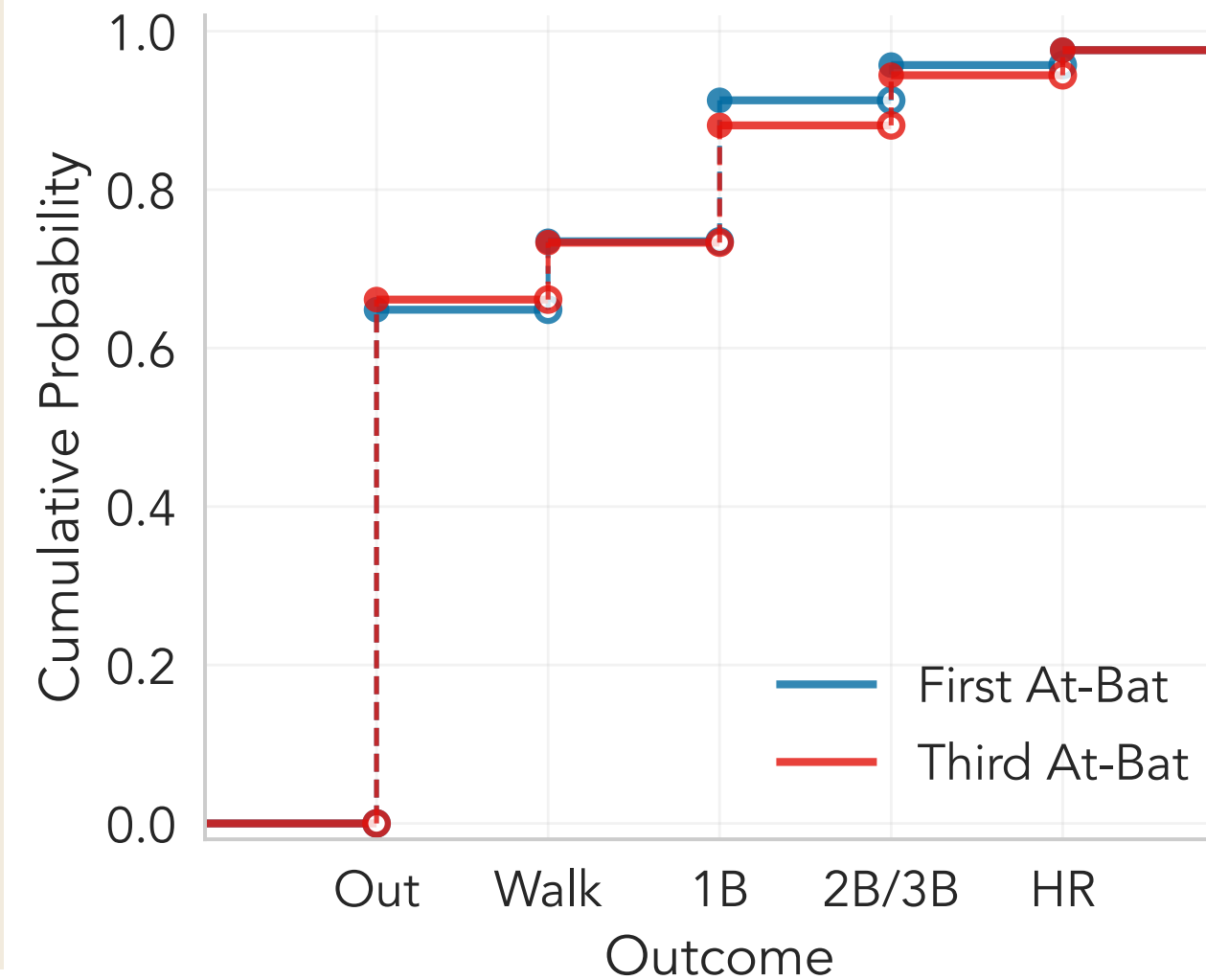
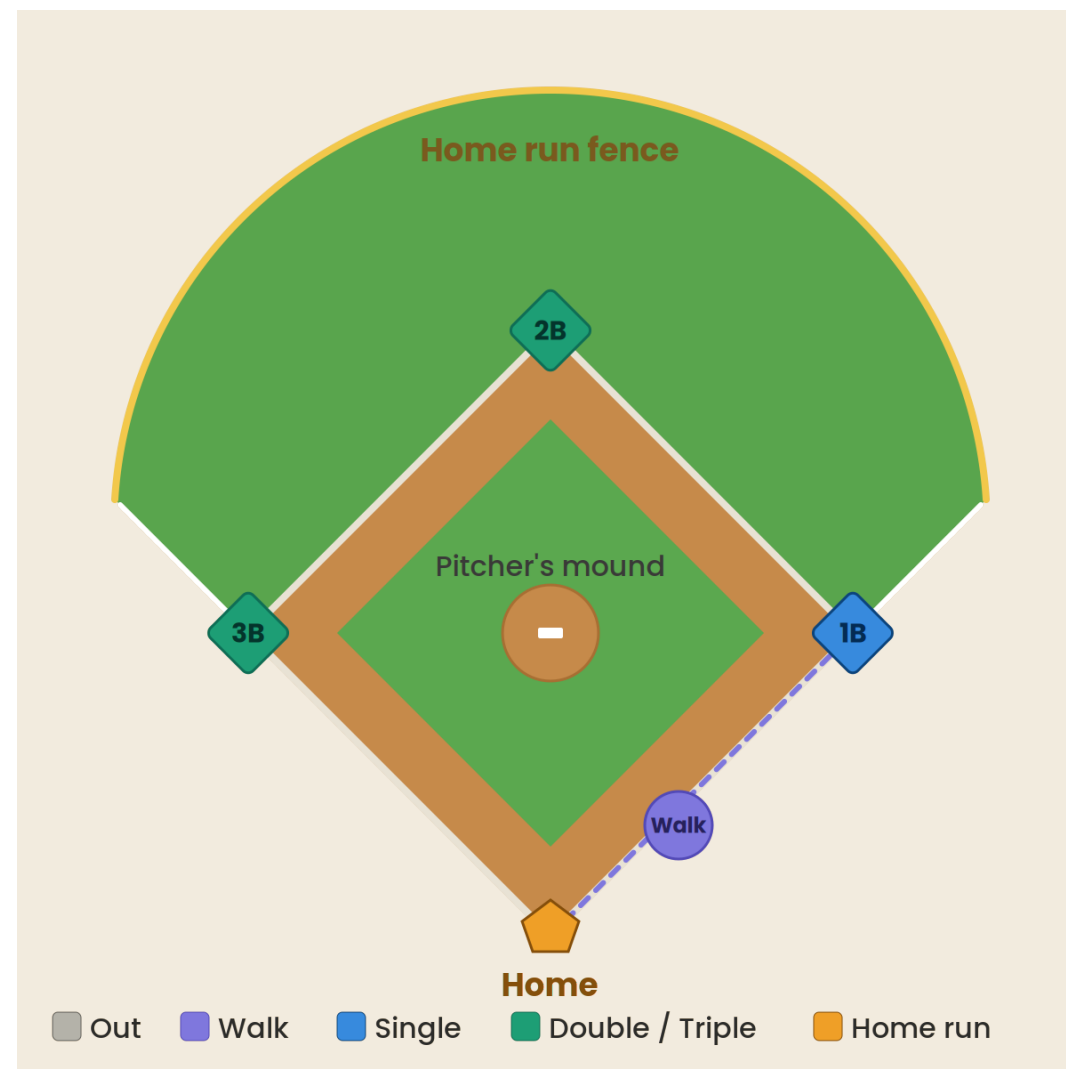
**Equal contribution.*

Motivation

Motivation #1: Does one prospect have an **upside** over another?

🏆 Comparing player performances:

Do batters have an upside when facing the same pitcher for the third time?



💰 Testing distributional effects:

Is there any upside in future incomes to veteran status? (Plot: Abadie, 2002)

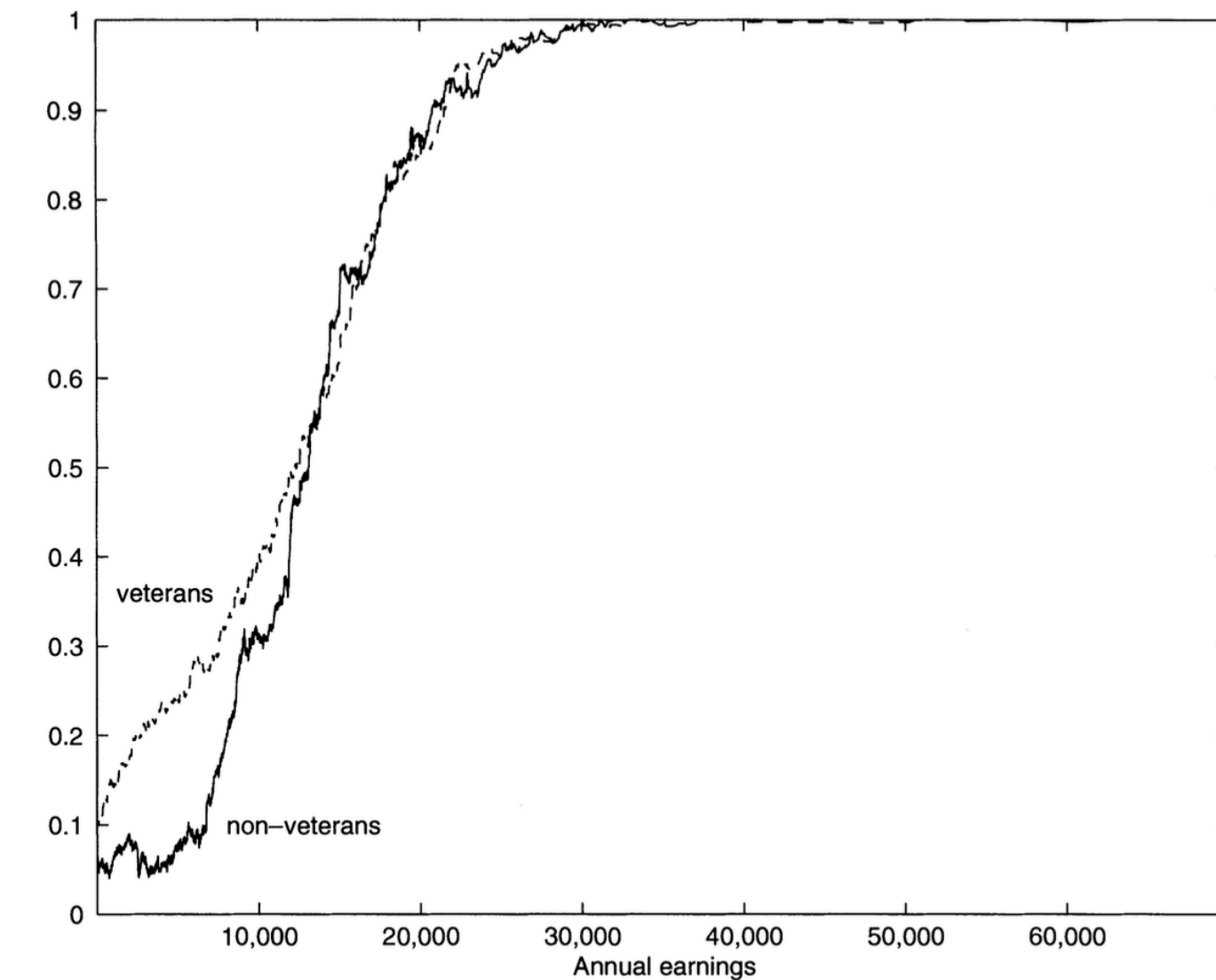


Figure 2. Estimated Distributions of Potential Earnings for Compliers.

These examples motivate **distributional** comparisons of random variables, beyond just their means, in different regions of the support.

Motivation #2: Can I monitor this test in real time?

In reality, these are *sequential* inference problems (literally, game time decisions!)

The Manager, the Ace and a Decision That Will Haunt the Rays



Blake Snell was dominating the Dodgers in a must-win World Series game. Kevin Cash still pulled him, raising painful questions of what might have been.



Blake Snell was taken out of Game 6 in the sixth inning despite having allowed just two hits. The Rays went on to lose, 3-1, ending the World Series. Kevin Jairaj/USA Today Sports, via Reuters

We are motivated to **sequentially** test/monitor for upsides, with validity at data-dependent stopping times.

The Goal

Develop **anytime-valid** methods for testing **stochastic dominance** that can be monitored for upsides at adaptive sample sizes.

...without sacrificing statistical power.

Stochastic Dominance

(a.k.a. Stochastic Ordering)

First-order stochastic dominance (1-SD)

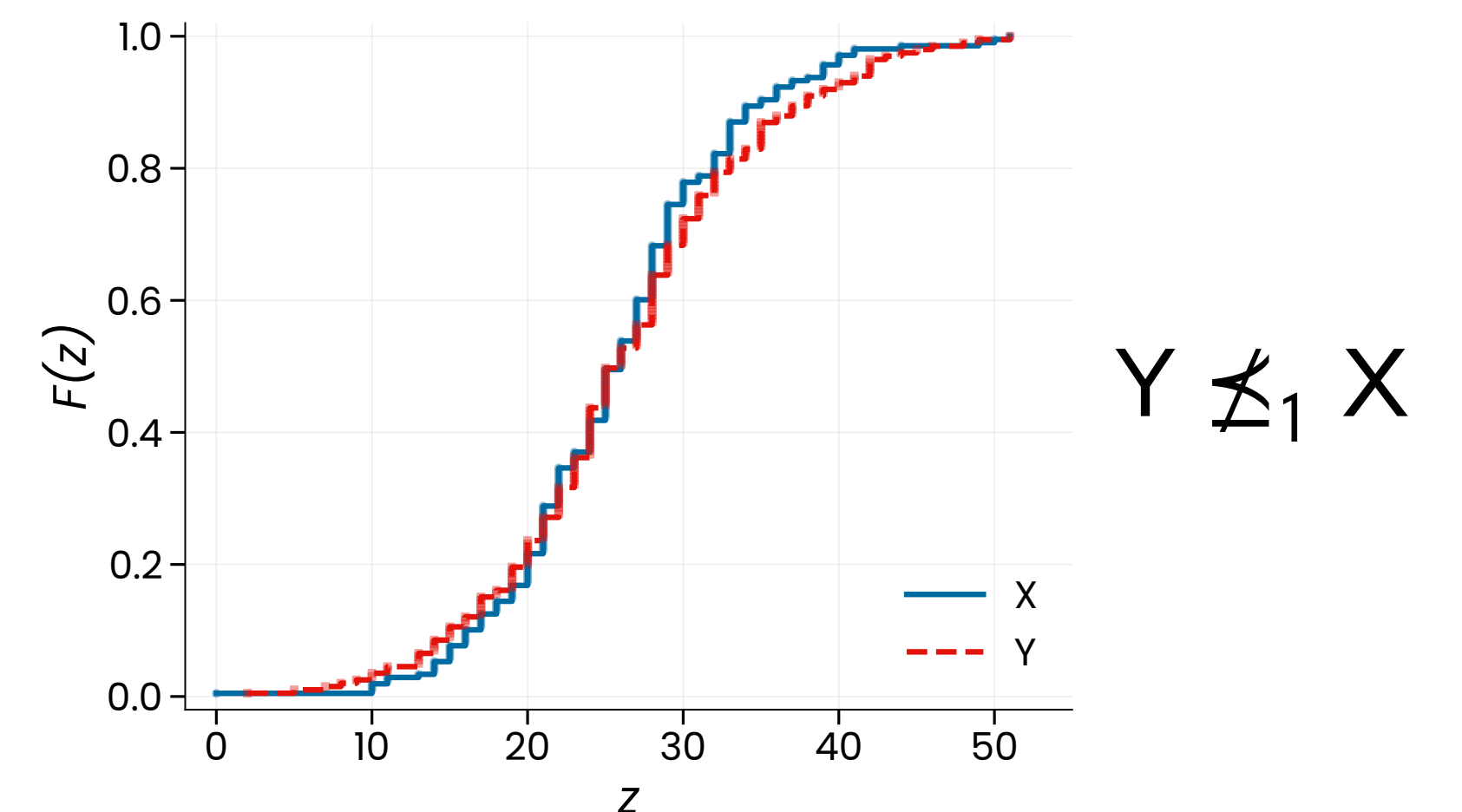
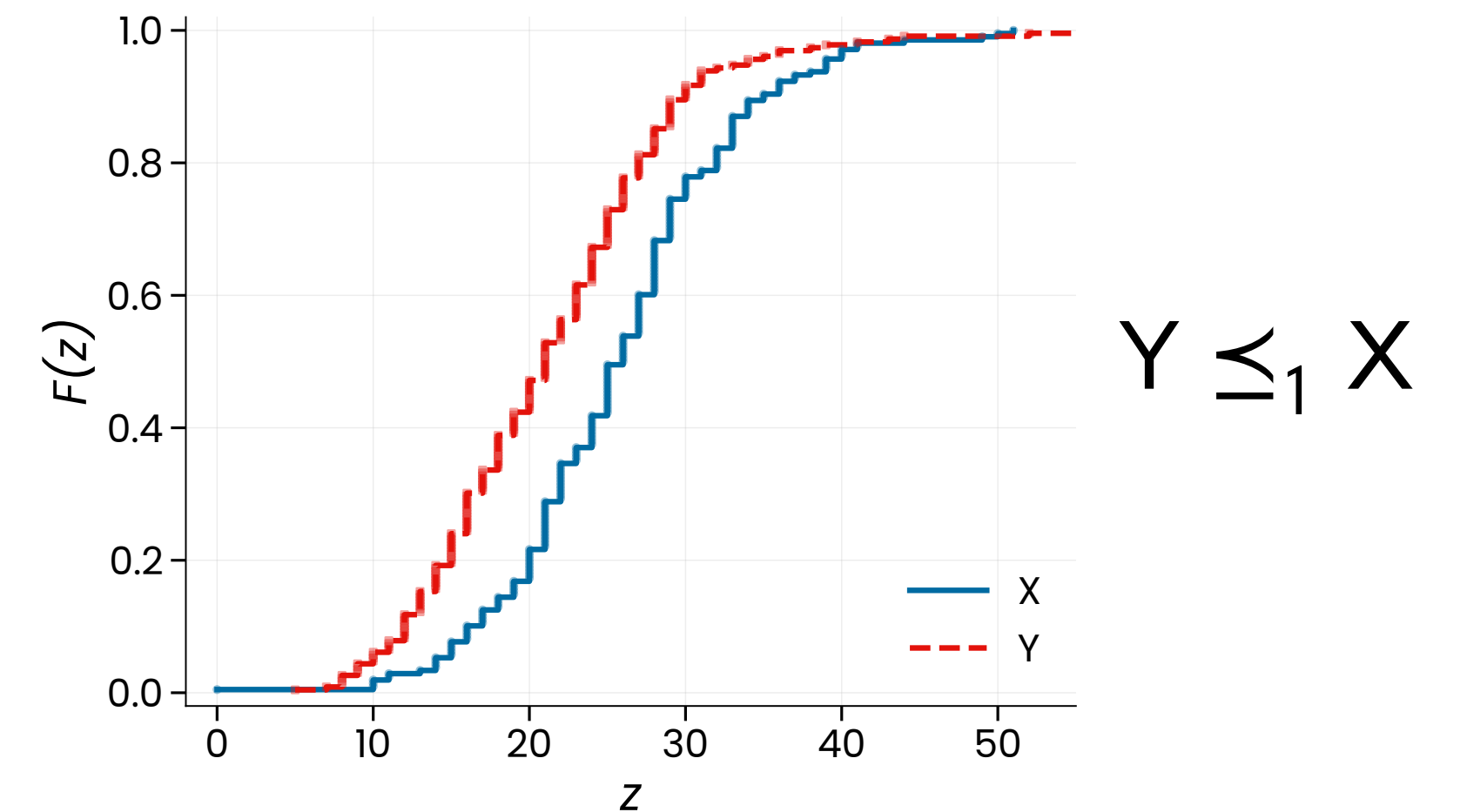
- Let X and Y be random variables with CDFs F_X and F_Y .
- Defn:** X stochastically dominates Y in the first order if the CDF of X is entirely below the CDF of Y :

$$Y \preceq_1 X \iff F_X(z) \leq F_Y(z), \forall z$$

- Expected utility view:** $Y \preceq_1 X$ if and only if

$$\mathbb{E}[u(Y)] \leq \mathbb{E}[u(X)]$$

for every increasing function u .



Testing for **upside** = Testing the 1-SD null

- Suppose we observe pairs of observations $(X_1, Y_1), (X_2, Y_2), \dots \sim \mathbb{P}_{XY}$.
- Define the (nonparametric and composite) null hypothesis that X dominates Y :

$$H_0 = \{ \mathbb{P} : Y \preceq_1 X \text{ under } \mathbb{P} \}$$

vs.

$$H_1 = \{ \mathbb{P} : Y \not\preceq_1 X \text{ under } \mathbb{P} \}$$

- The null hypothesis H_0 says that **Y has no upside over X** over the entire support ($z \in \mathbb{R}$).
- **Rejecting $H_0 \equiv$ there is an upside with which Y has an advantage over X .**

E-values & intersection nulls

- Given n data points $\mathbf{X}_1, \dots, \mathbf{X}_n$ and a null hypothesis \mathbf{H}_0 (possibly composite), an **e-value** $\mathbf{E} = \mathbf{E}_n(\mathbf{X}_1, \dots, \mathbf{X}_n)$ is any non-negative random variable satisfying:

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

- For intersection nulls, e-values can be combined easily under arbitrary dependence:**

Given e-values $\mathbf{E}^{(z)}$ for $\mathbf{H}_0^{(z)}$, their **(weighted) average** is an e-value for $\mathbf{H}_0 = \bigcap_{z \in Z} \mathbf{H}_0^{(z)}$:

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

Betting on bets:

An optimal e-variable for testing SD nulls

First step: Decompose the null hypothesis

$$H_0 = \{\mathbb{P} : Y \preceq_1 X \text{ under } \mathbb{P}\} = \{\mathbb{P} : \mathbb{P}(X \leq z) \leq \mathbb{P}(Y \leq z), \forall z\}.$$

- Highly composite family of distributions. However, it is an intersection null:

$$H_0 = \bigcap_z H_0(z), \quad \text{where} \quad H_0(z) = \{\mathbb{P} : \mathbb{P}(X \leq z) \leq \mathbb{P}(Y \leq z)\}.$$

- If we have an e-value $\mathbf{E}(z)$ for each $\mathcal{H}_0(z)$, then **any mixture** over \mathbf{z} is an e-value:

$$\mathbf{E} = \int \mathbf{E}(z) d\psi(z) \text{ is an e-value for } H_0.$$

The building-block GRO e-value

- For each “test threshold” $z \in \mathbb{R}$, let $\mathbf{D}(z) = \mathbf{1}(X \leq z) - \mathbf{1}(Y \leq z)$, and consider

$$\mathcal{H}_0(z) = \{\mathbb{P} : \mathbb{P}(X \leq z) \leq \mathbb{P}(Y \leq z)\} = \{\mathbb{P} : \mathbb{E}_{\mathbb{P}}[\mathbf{D}(z)] \leq 0\}.$$

Lemma.

(a) For any bet $\lambda \in [0, 1]$, $S(\lambda, z) = 1 + \lambda \mathbf{D}(z)$ is an **e-value** for $\mathcal{H}_0(z)$.

(b) For each alternative $\mathbb{Q} \notin \mathcal{H}_0(z)$, there is a **growth-rate optimal (GRO)** bet:

$$\lambda^*(z) = \frac{\mathbb{Q}(X \leq z < Y) - \mathbb{Q}(Y \leq z < X)}{\mathbb{Q}(X \leq z < Y) + \mathbb{Q}(Y \leq z < X)} = \frac{\mathbb{Q}(X \leq z) - \mathbb{Q}(Y \leq z)}{\mathbb{Q}(X \leq z < Y) + \mathbb{Q}(Y \leq z < X)} > 0.$$

- Growth rate:** $g_z(\lambda) = \mathbb{E}_{\mathbb{Q}}[\log S(\lambda, z)] = \mathbb{Q}(\mathbf{D}(z) = 1) \cdot \log(1 + \lambda) + \mathbb{Q}(\mathbf{D}(z) = -1) \cdot \log(1 - \lambda).$

Handling the intersection hypothesis

- Now that we have an optimal e-value for each threshold $\mathbf{z} \in \mathbf{Z}$, we can take any weighted average to obtain a valid e-value for the SD null \mathcal{H}_0 (intersection over z's).

$$E = \int_{\mathbf{Z}} E(\mathbf{z}) d\psi(\mathbf{z}), \text{ for any mixture distribution } \psi \text{ on } \mathbf{Z}.$$

- Then, across multiple rounds of data (sequentially observed), we can simply **multiply** these mixtures of GRO e-values!

Main result: A powerful e-process & a test of power one for SD

- At $\mathbf{t} \in \mathbb{N}$, let $\hat{\lambda}_{\mathbf{t}}^{\star}(\mathbf{z})$ the (empirical) predictable plug-in of $\lambda_{\mathbf{t}}^{\star}(\mathbf{z})$ using the first $(\mathbf{t} - 1)$ obs.
- Define the mixture process with a predictable *mixture* $(\psi_{\mathbf{t}})_{\mathbf{t} \in \mathbb{N}}$ over \mathbf{z} :

$$\mathbf{E}_{\mathbf{t}} = \prod_{\ell=1}^{\mathbf{t}} \mathbf{S}_{\ell}, \text{ where } \mathbf{S}_{\ell} = \int_{\mathbf{z}} \mathbf{S}(\hat{\lambda}_{\ell}^{\star}(\mathbf{z}), \mathbf{z}) d\psi_{\ell}(\mathbf{z}).$$

Theorem (Anytime-validity and power of the GRO e-process).

(a) $(\mathbf{E}_{\mathbf{t}})_{\mathbf{t} \in \mathbb{N}}$ is an e-process for the 1-SD null \mathbf{H}_0 .

(b) For reasonable* choices of $(\psi_{\mathbf{t}})_{\mathbf{t} \in \mathbb{N}}$, the e-process is powerful under any alternative \mathbb{Q} :

$$\mathbb{Q} \left(\liminf_{\mathbf{t} \rightarrow \infty} \frac{1}{\mathbf{t}} \log(\mathbf{E}_{\mathbf{t}}) > 0 \right) = 1, \text{ and the resulting sequential test has asymptotic power one.}$$

*Eventually assign enough mass on non-SD region: $\psi_{\mathbf{t}}(\{\mathbf{z} : \mathbb{Q}_X(\mathbf{z}) - \mathbb{Q}_Y(\mathbf{z}) > \varepsilon\}) > \delta$ for large \mathbf{t} .

Choosing the predictable mixture weights

- For each threshold, we already have a GRO e-value. So, we want to make sure that the mixture weights is adaptive and do not “miss” any important (non-dominance) regions.
- The simplest choice,

$$\psi_t = \psi = \frac{1}{m} \sum_{i=1}^m \delta_{z_i'}$$

can work well if $\{\mathbf{z}_1, \dots, \mathbf{z}_m\}$ “sufficiently cover” the non-dominance regions of interest.

- When the support is unbounded, we recommend adjusting the m thresholds to the empirical quantiles of observed data, and use *exponential weights* adapted to those quantiles:

$$\psi_t = \sum_{i=1}^{n_t} w_t(\mathbf{z}_i^t) \delta_{z_i^t}, \text{ where } w_t(\mathbf{z}_i^t) \propto \exp \left(\eta \cdot \frac{\hat{F}_X^{t-1}(\mathbf{z}_i^t) - \hat{F}_Y^{t-1}(\mathbf{z}_i^t)}{\hat{\sigma}_{t-1}} \right)$$

*Why “betting on bets”?

- Consider two uncertain outcomes X and Y , say the daily returns of two financial assets.
- Fix a “test threshold” (say, $z = 0\%$). Bookmaker proposes a **double-or-nothing-or-push bet** with the following rules.
 - *If Y nets a positive return ($Y > z$) but X does not, then you **double** your bet.*
 - *If X nets a positive return ($X > z$) but Y does not, then you **lose** your bet.*
 - *Otherwise, nothing happens (“push”).*
- (Bookmaker hypothesis = $H_0(z)$ = “Probability of Y greater than z is no better than of X .”)
- Skeptic places a **fraction** of her money on this bet.

Over repeated rounds, Skeptic’s wealth is an **e-process** for “upside” (SD) testing.
(even when Skeptic can split money across different test thresholds!)

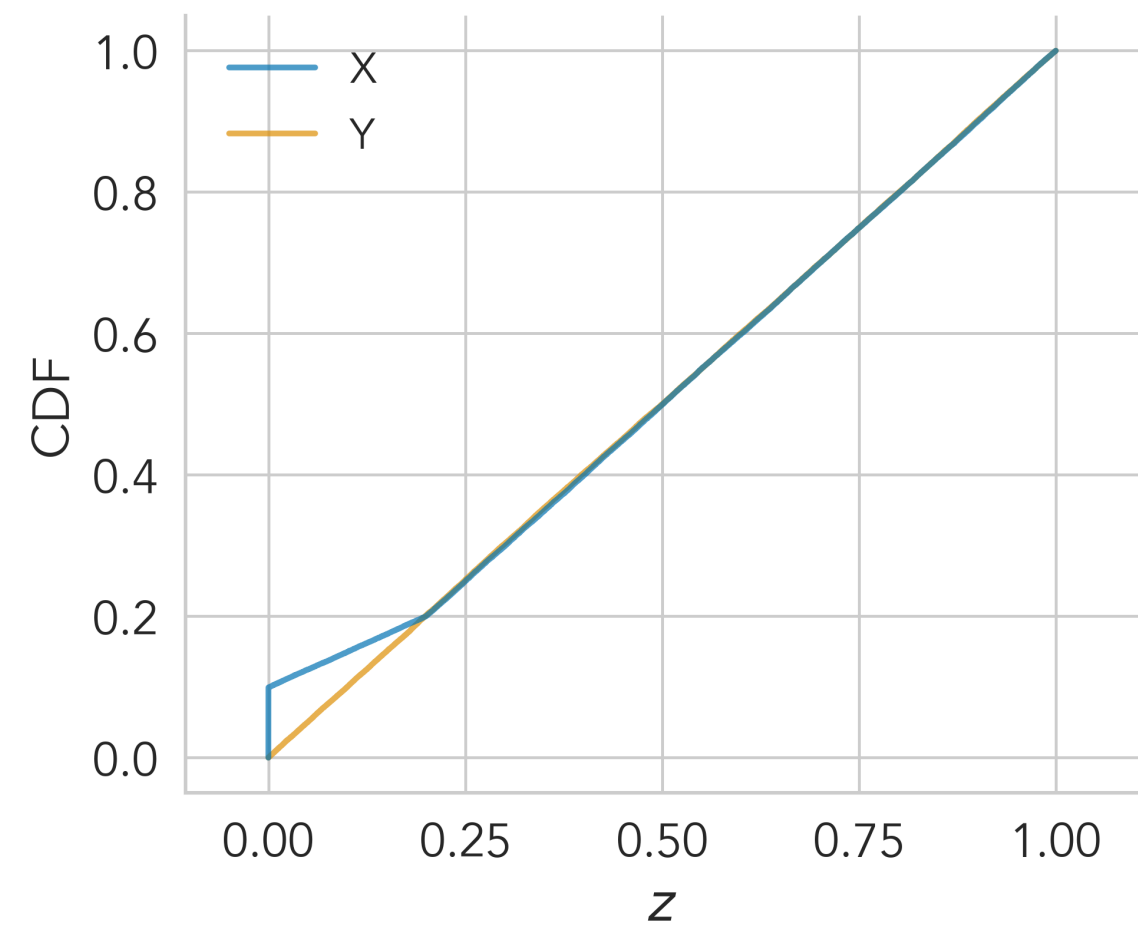
Simulations

Baselines

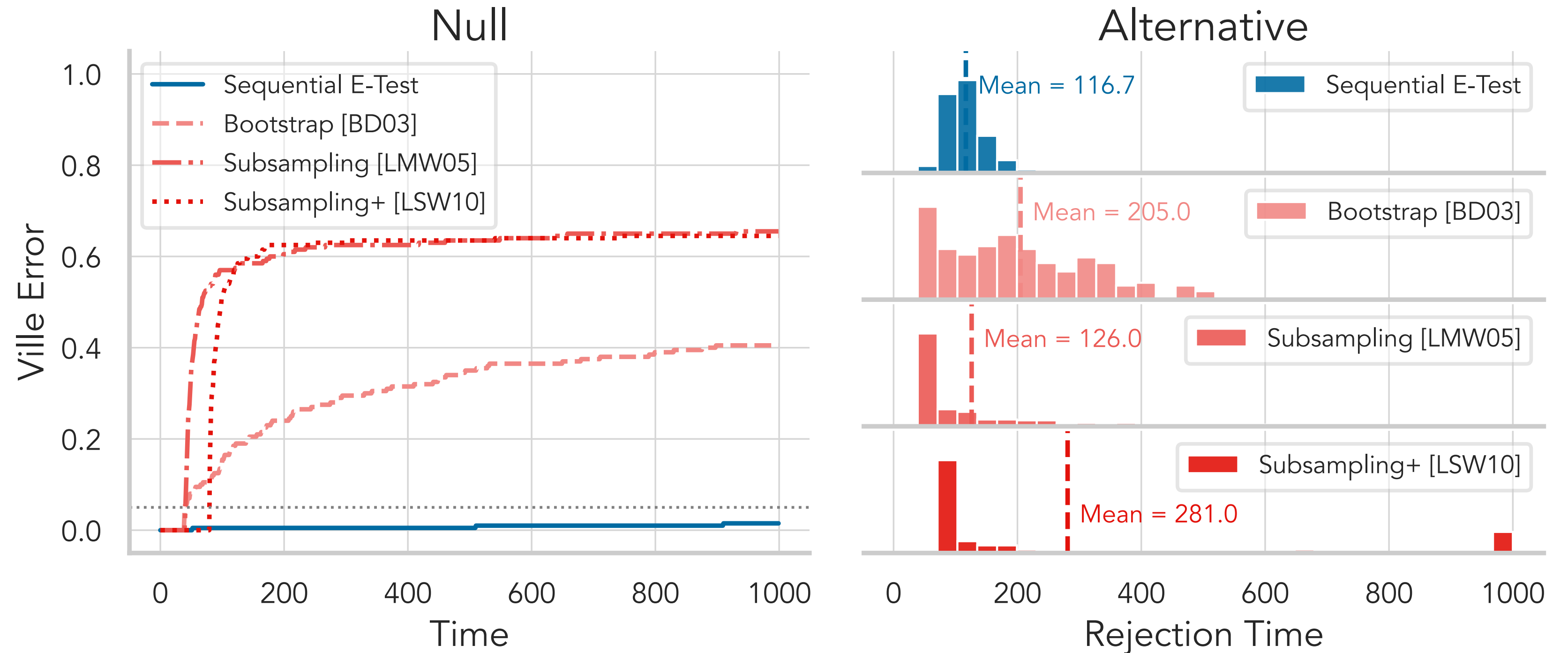
Denote the difference $\Delta_{t-1}(\mathbf{z}) = \sum_{\ell=1}^{t-1} D_{\ell}(\mathbf{z}) = t \left[\hat{F}_X^{t-1}(\mathbf{z}) - \hat{F}_Y^{t-1}(\mathbf{z}) \right]$, and $V_{t-1}(\mathbf{z}) = \sum_{\ell=1}^{t-1} D_{\ell}^2(\mathbf{z})$.

- **AdaGRO-Exp**: exponential weights with self-normalization; $\mathbf{w}_t(\mathbf{z}) \propto \exp \left\{ \eta \cdot \Delta_{t-1}(\mathbf{z}) / \sqrt{V_{t-1}(\mathbf{z})} \right\}$
- **AdaGRO-Linear**: linear weights using GRO bets themselves; $\mathbf{w}_t(\mathbf{z}) \propto \hat{\lambda}_t^{\star}(\mathbf{z})$
- **AdaGRO-Greedy** (Shekhar & Ramdas, 2023): all-in on the max; $\mathbf{w}_t(\mathbf{z}) = 1 \left\{ \mathbf{z} = \operatorname{argmax}_{\mathbf{z}} \Delta_{t-1}(\mathbf{z}) \right\}$
- **GRO**: non-adaptive GRO baseline with simple averaging; $\psi_t = \psi = \frac{1}{m} \sum_{i=1}^m \delta_{z_i}$
- **Constant**: non-adaptive, constant-bet (non-GRO) baseline; $\lambda_t(\mathbf{z}) = 0.1$

Simulation #1: Comparison with classical, non-SAVI tests



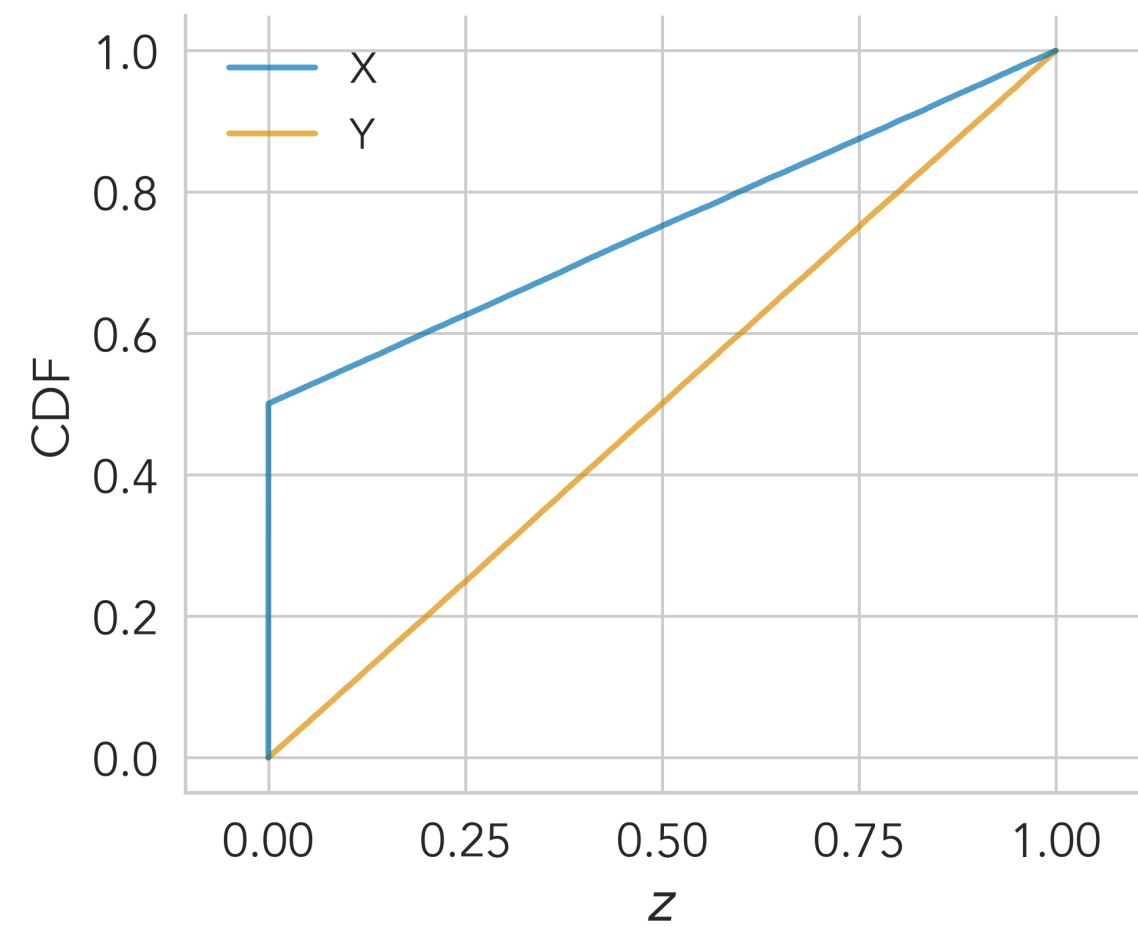
(the “**contact**” case)



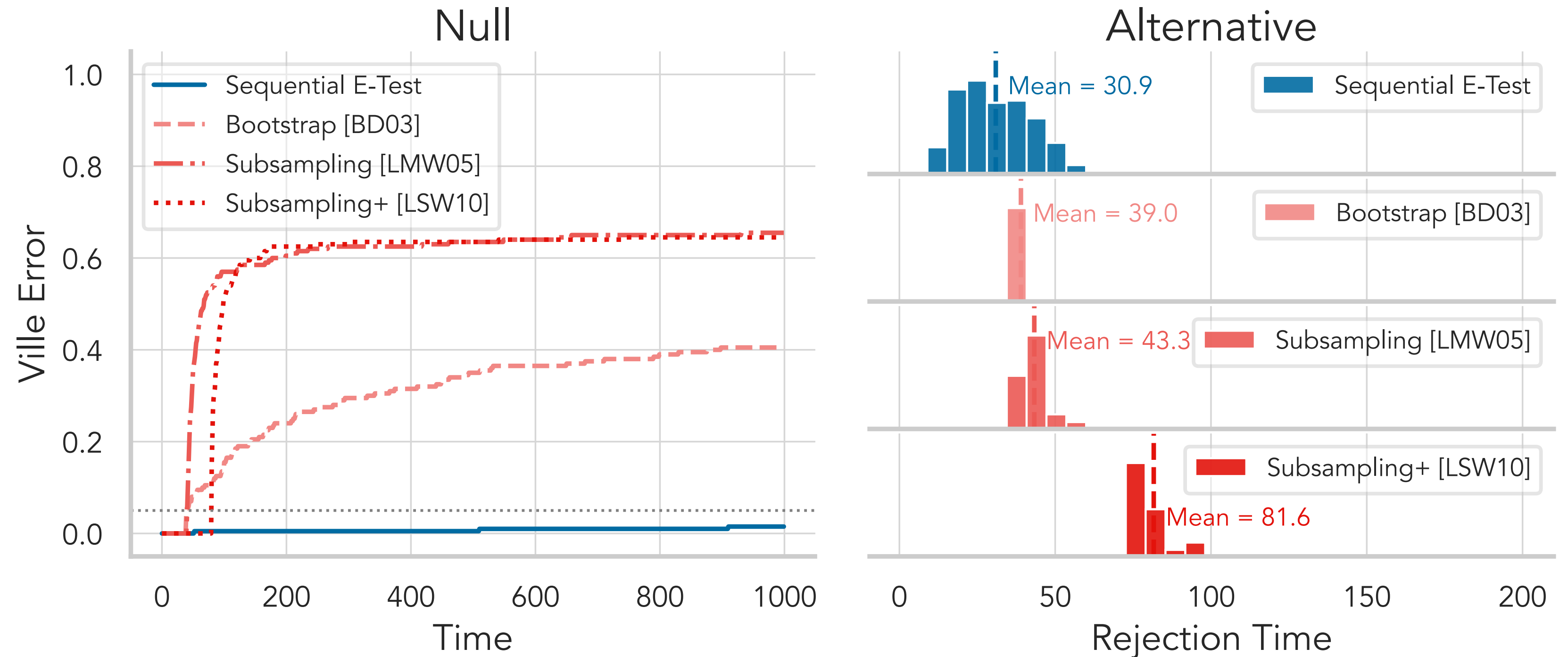
$$\text{Ville Error: } \hat{\mathbb{P}}(\exists t : E_t \geq 1/\alpha)$$

$$\text{Rejection Time: } \tau_\alpha = \inf \{t : E_t \geq 1/\alpha\}$$

Simulation #1: Comparison with classical, non-SAVI tests



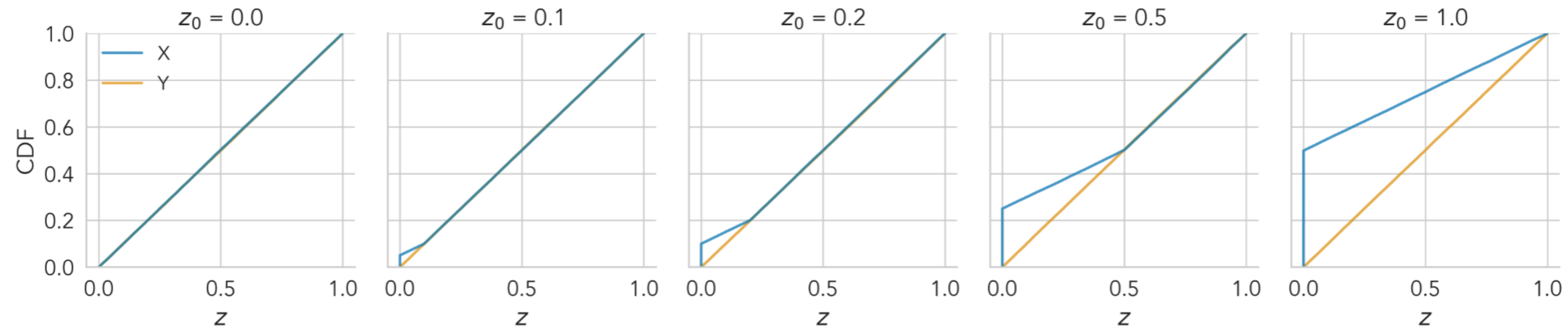
(the “no-contact” case)



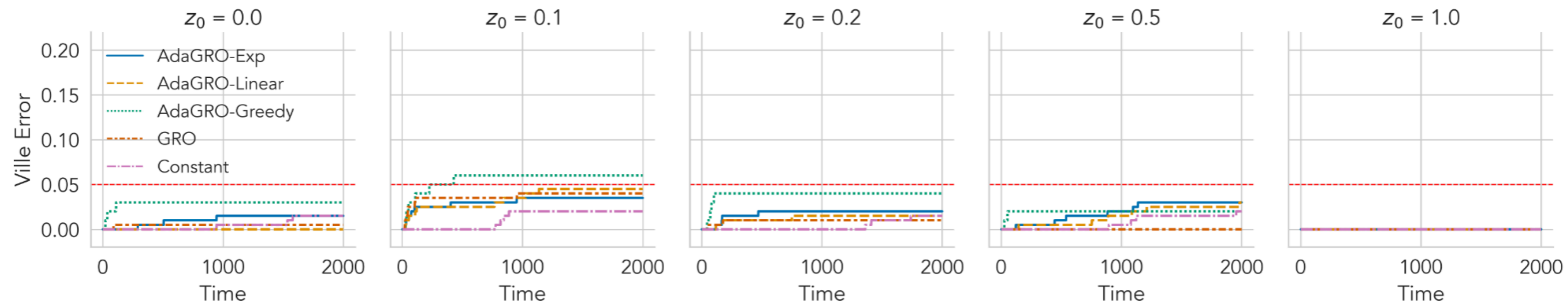
$$\text{Ville Error: } \hat{\mathbb{P}}(\exists t : E_t \geq 1/\alpha)$$

$$\text{Rejection Time: } \tau_\alpha = \inf \{t : E_t \geq 1/\alpha\}$$

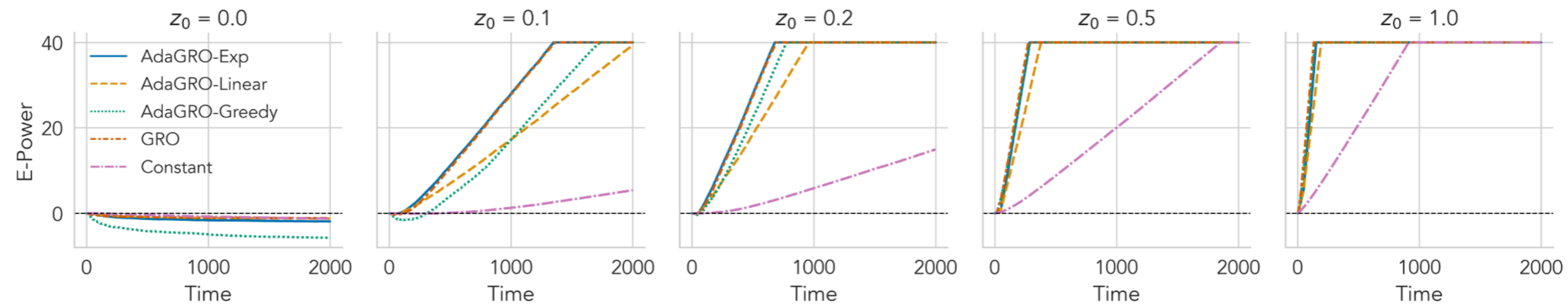
Simulation #1: Robustness to “contact sets” between CDFs



(a) CDFs of X and Y for $z_0 \in \{0.0, 0.1, 0.2, 0.5, 1.0\}$.

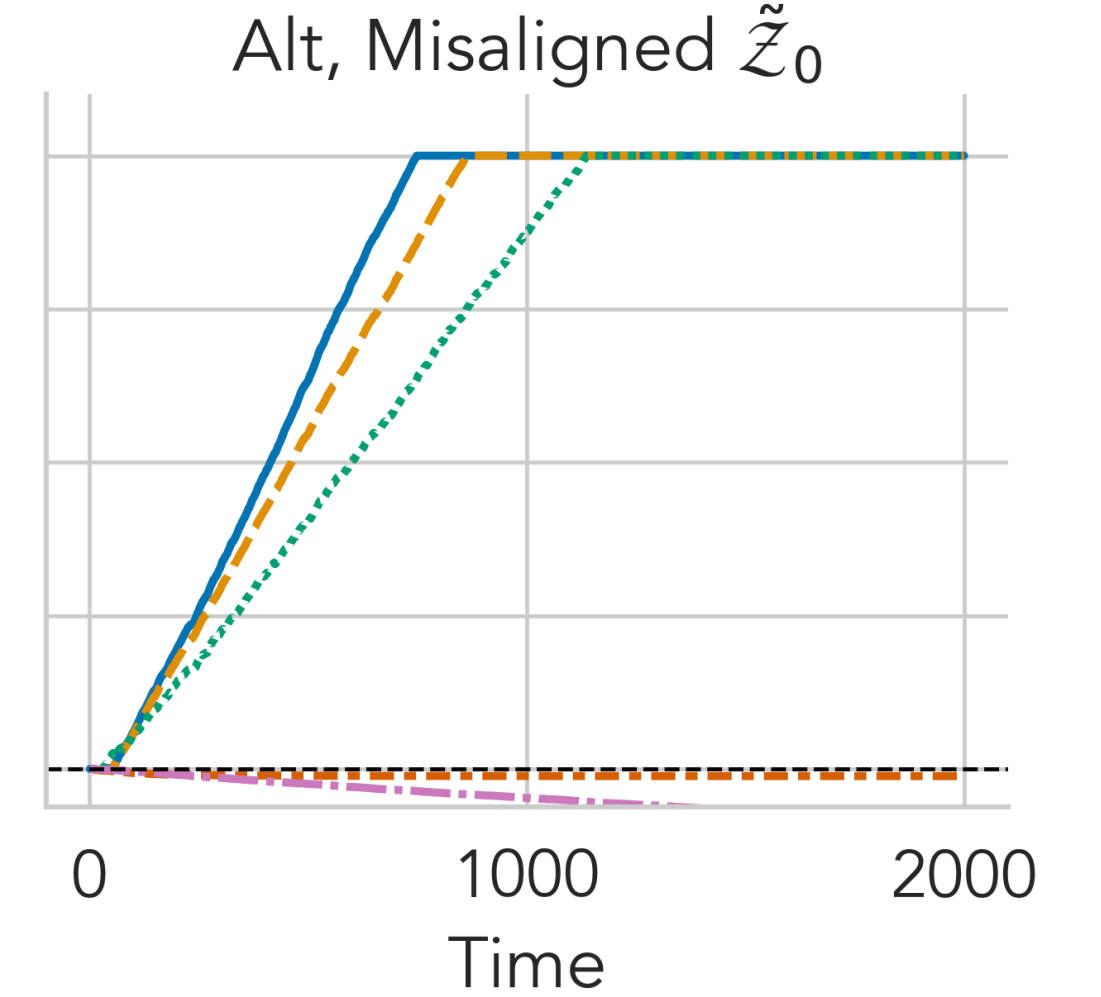
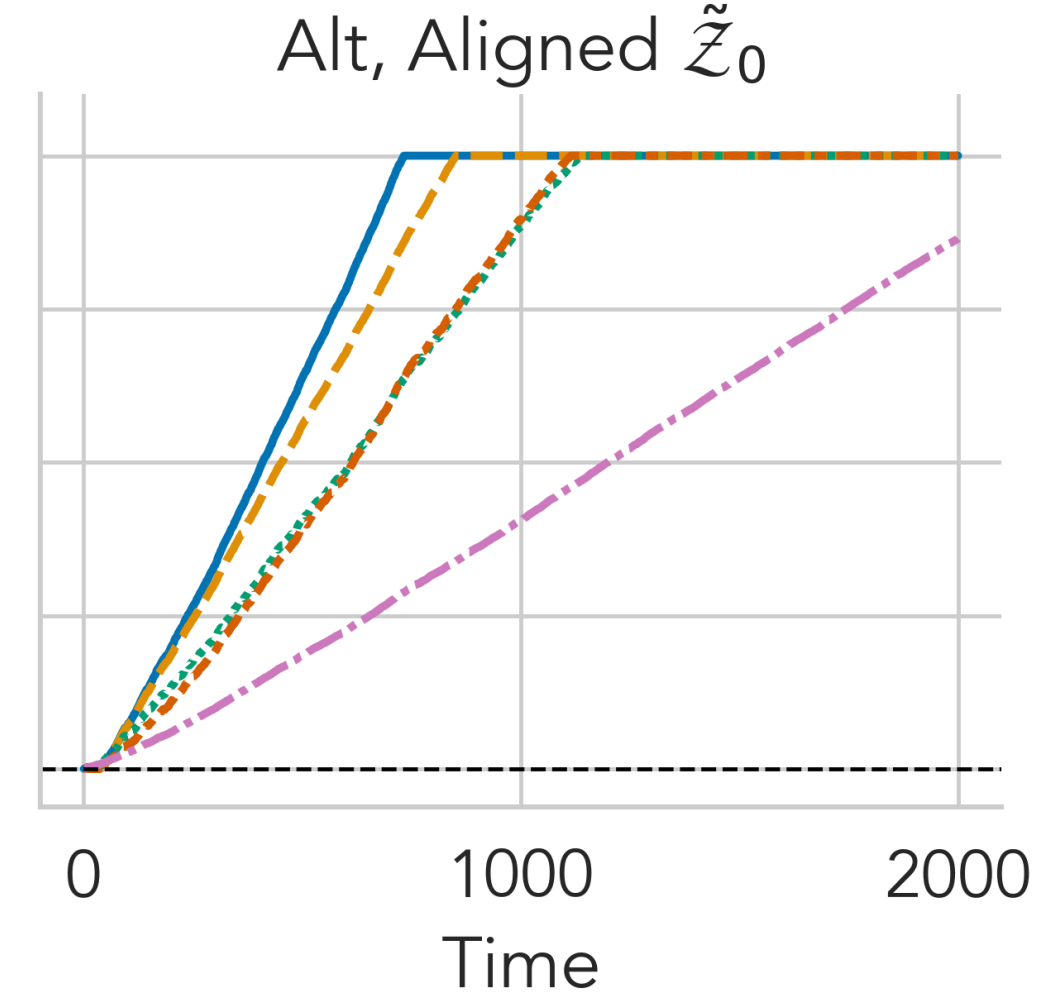
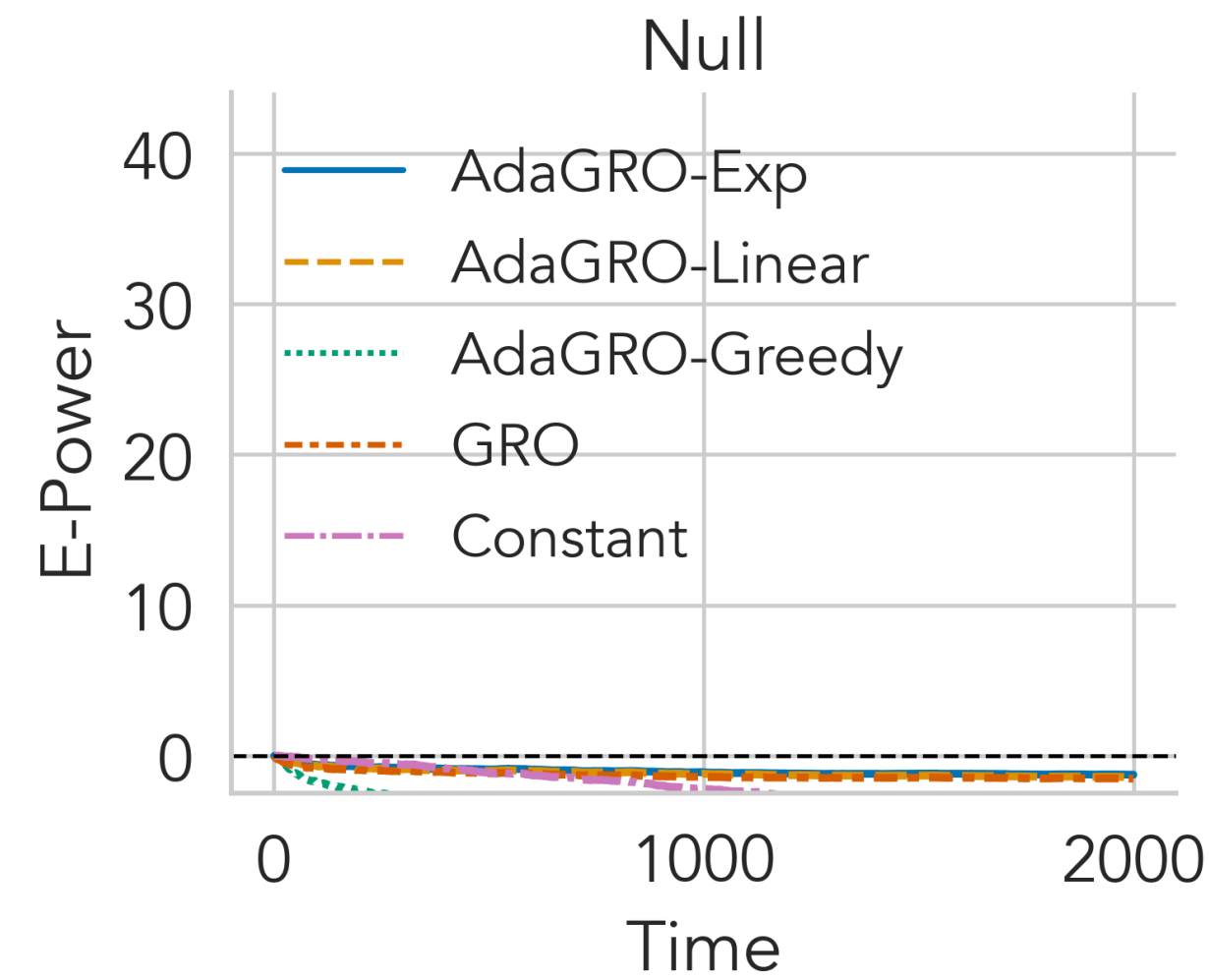
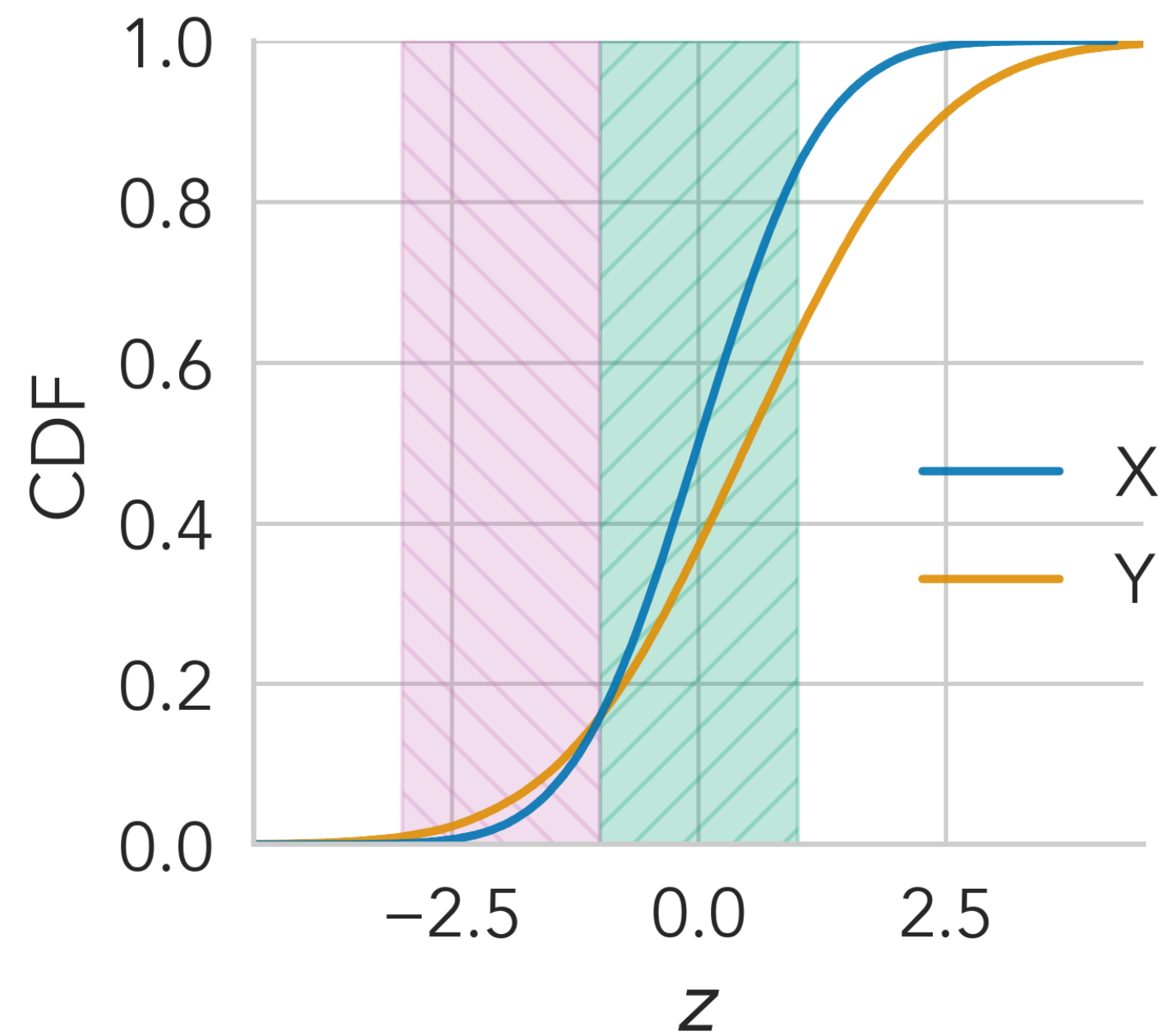


(b) Ville error for testing $\mathcal{H}_0 : X \leq_1 Y$.



(c) E-power against $\mathcal{H}_0 : Y \leq_1 X$.

Simulation #2: Adaptivity to non-dominance regions



Application: The third time-through-the-order penalty in baseball

Examining the 3rd time-through-the-order (3TTO) penalty

The Manager, the Ace and a Decision That Will Haunt the Rays

Blake Snell was dominating the Dodgers in a must-win World Series game. Kevin Cash still pulled him, raising painful questions of what might have been.

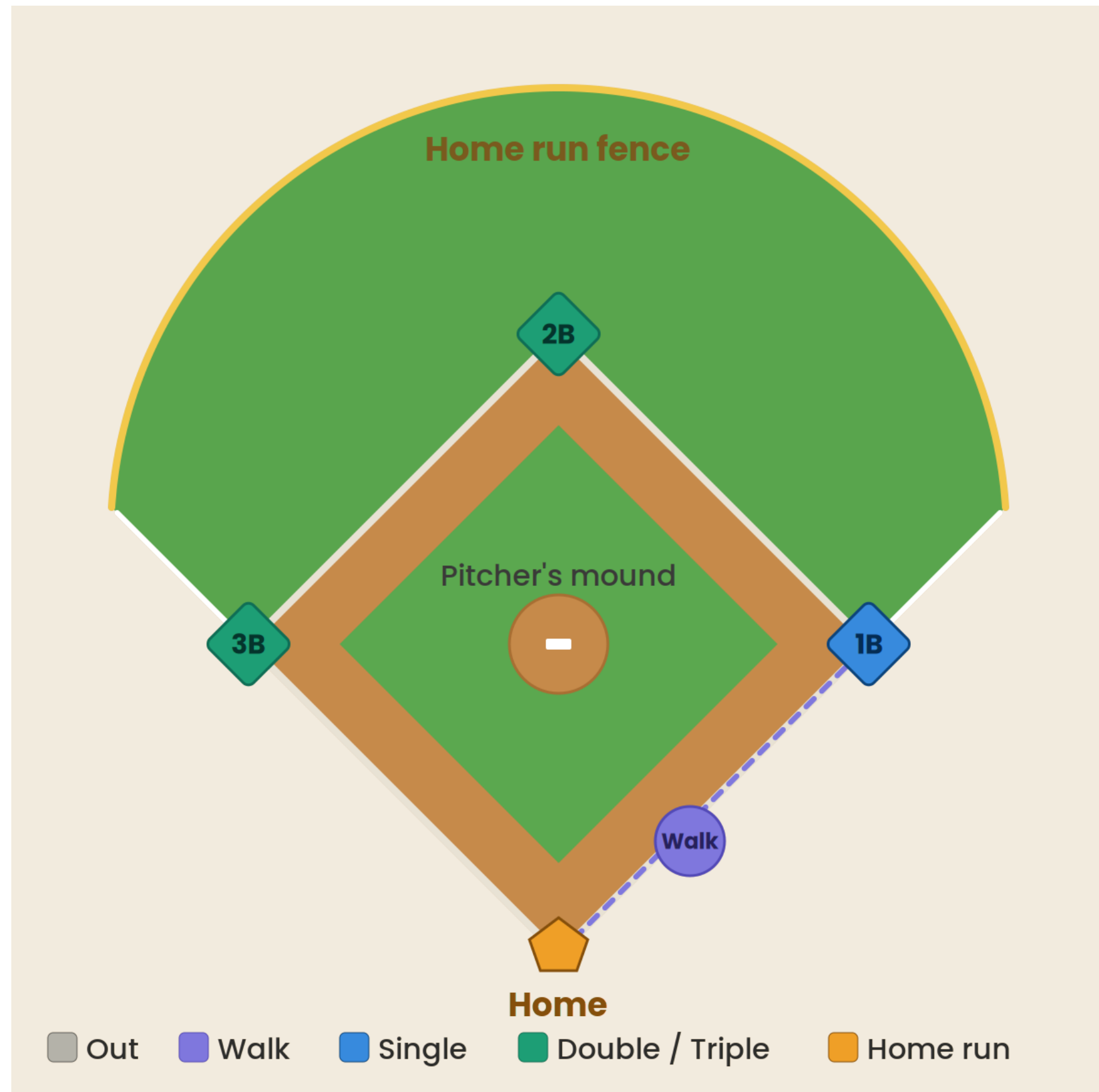


Blake Snell was taken out of Game 6 in the sixth inning despite having allowed just two hits. The Rays went on to lose, 3-1, ending the World Series. Kevin Jairaj/USA Today Sports, via Reuters

A starting pitcher typically faces the same lineup of nine batters multiple times in a game.

Should the manager replace the pitcher before batters face him for the third time?

Examining the 3rd time-through-the-order (3TTO) penalty



PC: Claude 4.8

Data: paired at-bats for MLB pitcher Blake Snell (2016–2025).

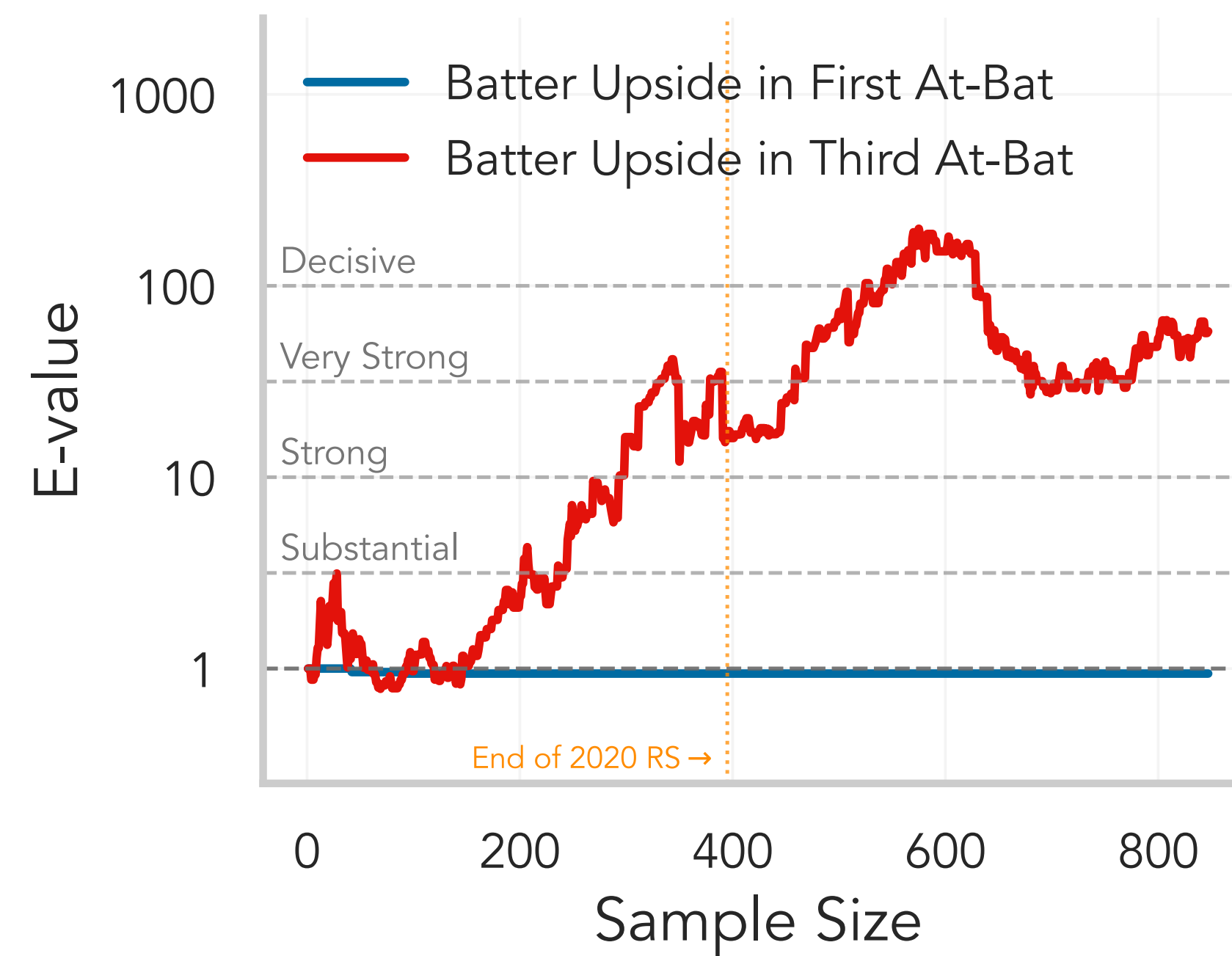
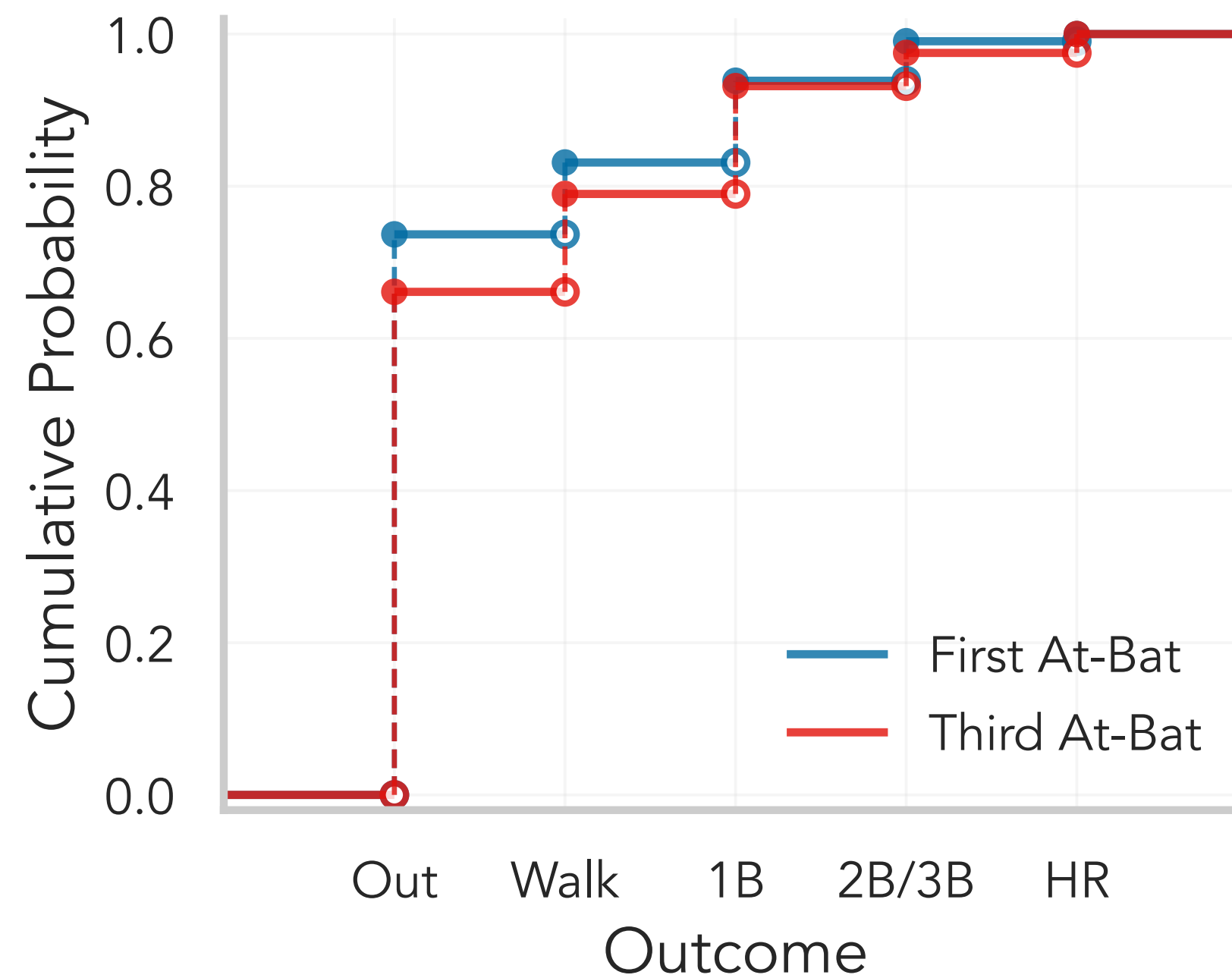
X: outcome in 1TTO. **Y:** outcome in 3TTO.

Order: Out < Walk < 1B < 2/3B < HR.

H0: No upside for batters in 3TTO.

H1: Batters have an upside in 3TTO.

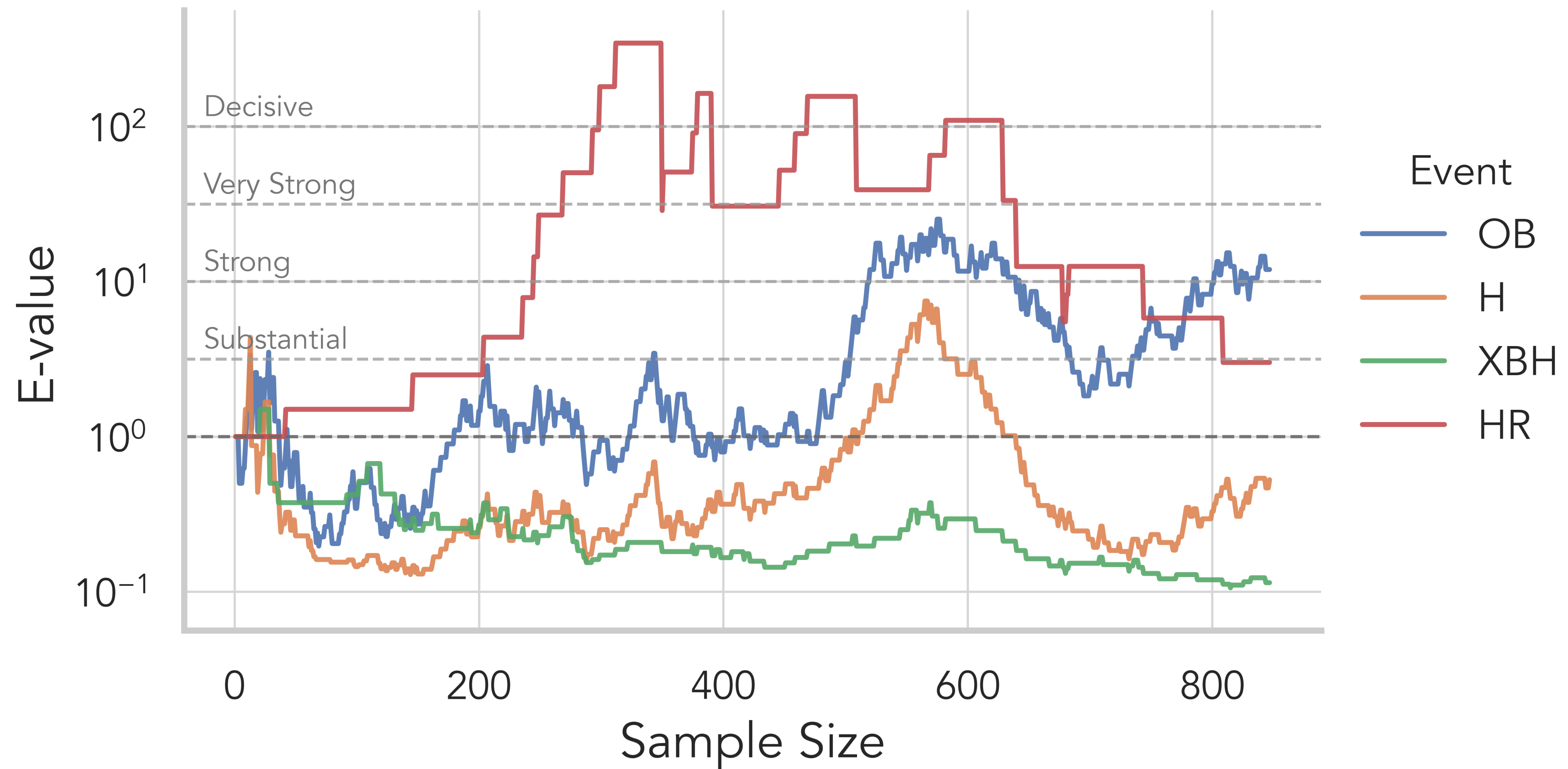
Application: Monitoring the third-time effect in baseball



Data: Every paired at-bat outcome against MLB pitcher Blake Snell, from 2016 to 2025 (regular seasons).

Fine-grained testing for each $H_0^{(z)}$

OB (on-base): Walk, 1B, 2B, 3B, HR
H (hit): 1B, 2B, 3B, HR
XBH (extra-base hit): 2B, 3B, HR



Affirming SD (i.e., *definite* upside):
Testing the non-dominance null

An impossibility result for testing non-SD

- Suppose we want to show the *stronger* claim that Y has a **definite upside** over X .
- In this case, we want to reject the the *non-SD null* (in first order)

$$\mathcal{H}'_0 = \{\mathbb{P} : \mathbb{P}(X \leq \mathbf{z}) < \mathbb{P}(Y \leq \mathbf{z}) \text{ for some } \mathbf{z}\} = \{\mathbb{P} : \mathbb{P}_X \not\preceq \mathbb{P}_Y\},$$

a **non-convex** union-hypothesis. The null & alternative are switched!

Corollary (No nontrivial e-variable exists for the non-SD null). There exists no e-variable E for \mathcal{H}'_0 which is nontrivial for all $\mathbb{Q} \in (\mathcal{H}'_0)^c$ (that is, $\mathbb{E}_{\mathbb{Q}} \log E > 0$) simultaneously.

follows from Theorem 6.2 by Zhang et al. (2024)

The minimum GRO e-value (finite support)

- Assume **finite** (and known) support and add separation from the null:

$$\mathcal{Q}(\varepsilon) = \{Q : Q(X \leq z_i) > Q(Y \leq z_i) + \varepsilon, \text{ for all } \mathbf{z} = \mathbf{z}_1, \dots, \mathbf{z}_{m-1}\}.$$

Proposition (Validity and asymptotic optimality of the min-approach)

(a) $\mathbf{S} = \min_{i=1, \dots, m-1} \mathbf{S}(\lambda_i, \mathbf{z}_i)$ is an e-value for \mathcal{H}'_0 , for any bet $\lambda_i \in [0, 1]$,

(b) For the predictable GRO plug-in, $\mathbf{e}_t = \min_{i=1, \dots, m-1} \left(\prod_{\ell=1}^t \mathbf{S}_{i,\ell} \right) \rightarrow \infty$, for any $Q \in \mathcal{Q}(\varepsilon)$,

and we have the asymptotically optimal growth-rate: $\liminf_{t \rightarrow \infty} \frac{1}{t} (\log \mathbf{e}_t - \log E_t) \geq 0$.

Affirming FSD for continuous outcomes

- Continuous CDFs F_X and F_Y on \mathbf{Z} will *always* cross at the boundaries/tails.
- Thus, we have to **restrict** the domain of interest to $\tilde{\mathbf{Z}} \subseteq \mathbf{Z}$ (Davidson and Duclos, 2013).
- Again, consider *separated* (and *restricted*) alternative

$$\mathcal{Q}(\varepsilon, \tilde{\mathbf{Z}}) = \{Q : Q(X \leq z_i) > Q(Y \leq z_i) + \varepsilon, \text{ for all } z \in \tilde{\mathbf{Z}}\}.$$

Proposition. We can construct sequential tests for the non-SD null that achieve *asymptotic power one* against alternatives in $Q \in \mathcal{Q}(\varepsilon, \tilde{\mathbf{Z}})$.

- These tests are based on *time-uniform CDF bands*.
Howard and Ramdas (2022), Mineiro and Howard (2023), and Clerico et al (2026).

Conclusion

Conclusion

- **We developed a novel family of e-processes & sequential tests for various SD notions.**
 - *These methods are fully nonparametric and robust to dependence between X and Y .*
 - *The e-based approach performs competitively in power with classical, non-SAVI methods.*
- **We discuss potential extensions to non-i.i.d., unpaired, and multivariate data.**
- **We also discuss extensions to general, integral stochastic orders.**
 - Notable examples: increasing convex orders (i.e., risk-seeking DMs) & infinite-order SD.
- **Non-SD null testing remains challenging, and there are various extensions for future work.**
 - In particular, affirming weaker notions SD (higher-order SD; almost SD; ...) remains hard.



Thank you!

Any questions?

<https://arxiv.org/abs/2604.21851>

Appendix

Testing higher-order SD

Higher-order SD

- More generally, \mathbf{X} stochastically dominates \mathbf{Y} in the \mathbf{k} -th order if

$$\mathbb{E}[u(\mathbf{Y})] \leq \mathbb{E}[u(\mathbf{X})] \text{ for every "utility" function } u \in \mathcal{U}^{[k]}.$$

- E.g., 2nd order utility class $\mathcal{U}^{[2]}$ consists of all **increasing & concave** functions (risk-averse DMs).
- Generally, the \mathbf{k} -th order utility class consists of functions that alternate signs in their first \mathbf{k} derivatives.
 - 3rd order models “prudent” DMs (marginal utility is convex): implies positive skew
 - 4th order captures “temperate” DMs (reluctance to accept additional risk): implies thinner tails
- **Testing for higher-order SD = Testing for upside by a DM with a partially specified utility function.**

Characterizing k-th order SD with generators

- Formally, X stochastically dominates Y in the k -th order (**k-SD**), or $Y \preceq_k X$, if

$$F_X^{[k]}(z) \leq F_Y^{[k]}(z), \quad \forall z \in \mathbb{R}, \quad \text{where} \quad F^{[k]}(z) = \int_{-\infty}^z F^{[k-1]}(u) du.$$

(we set $F^{[1]} \equiv F$).

Lemma (Characterizations of k-SD). The following statements are equivalent:

(a) $Y \preceq_k X$, or $F_X^{[k]}(z) \leq F_Y^{[k]}(z)$ for all $z \in \mathbb{R}$.

(b) [Utility View] $\mathbb{E}[u(Y)] \leq \mathbb{E}[u(X)]$ for any k -th order utility function $u \in \mathcal{U}^{[k]}$.

(c) [Generator View] $\mathbb{E}[(z - X)_+^{k-1}] \leq \mathbb{E}[(z - Y)_+^{k-1}]$ for all $z \in \mathbb{R}$.

Extension via the integral identity

- Proof of the generator characterization is based on the following identity (via Fubini):

$$F^{[k]}(\mathbf{z}) = \frac{1}{(k-1)!} \mathbb{E} \left[(z - X)_+^{k-1} \right].$$

- Thus, we can rewrite **the k-SD null hypothesis** as another intersection null over \mathbf{z} !

$$\mathcal{H}_0^{[k]} = \bigcap_{z \in \mathbb{R}} \mathcal{H}_0^{[k]}(z), \quad \text{where} \quad \mathcal{H}_0^{[k]}(z) = \{ \mathbb{P} : \mathbb{E}_{\mathbb{P}}[(z - X)_+^{k-1}] \leq \mathbb{E}_{\mathbb{P}}[(z - Y)_+^{k-1}] \}.$$

Main result: Asymptotically powerful e-processes for testing k-SD

Assumption. The support of X and Y is bounded from below: $Z = [a, \infty)$.

- Define bet $S^{[k]}(\lambda, z) = 1 + \lambda[u_z(Y) - u_z(X)]$, for the k-th order, *normalized* utility generator

$$u_z(x) = - \left[\frac{(z - x)_+}{z - a} \right]^{k-1} \in [-1, 0], x \in Z$$

Theorem (Anytime-validity and power for k-th order SD null).

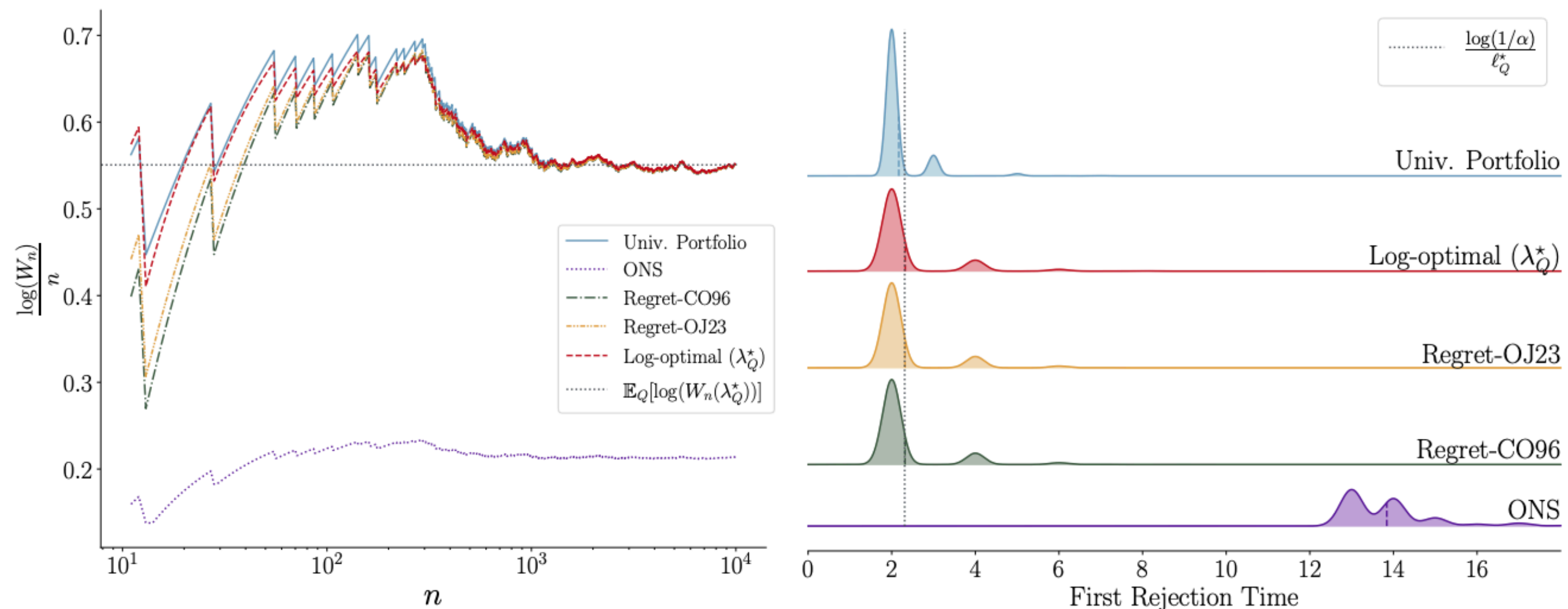
- (a) For any predictable mixture $(\psi_t)_{t \in \mathbb{N}}$ and bets $(\lambda_t(\mathbf{z}))_{t \in \mathbb{N}}$, we have an e-process for $\mathcal{H}_0^{[k]}$:

$$E_t^{[k]} = \prod_{\ell=1}^t S_\ell^{[k]}, \text{ where } S_\ell^{[k]} = \int_Z S^{[k]}(\lambda_\ell(\mathbf{z}), \mathbf{z}) d\psi_\ell(\mathbf{z}).$$

- (b) If we learn $(\lambda_t(\mathbf{z}))_{t \in \mathbb{N}}$ by UP, we have again exp. growth under any alternative \mathbb{Q}^* .

*How do we choose the bets?

- For k-SD, we don't have a growth-rate optimal (GRO) bet; bet outcome is no longer ternary.
- But we have a *bounded* outcome, and many betting strategies that we know would work.
- For each threshold \mathbf{z} , we default to **universal portfolio bets** $\lambda_t^{\text{UP}}(\mathbf{z})$, which are asymptotically log-optimal and achieves sublinear portfolio regret (Waudby-Smith et al., 2025, figure below).



Additional Materials

E-processes quantify evidence in sequential experiments

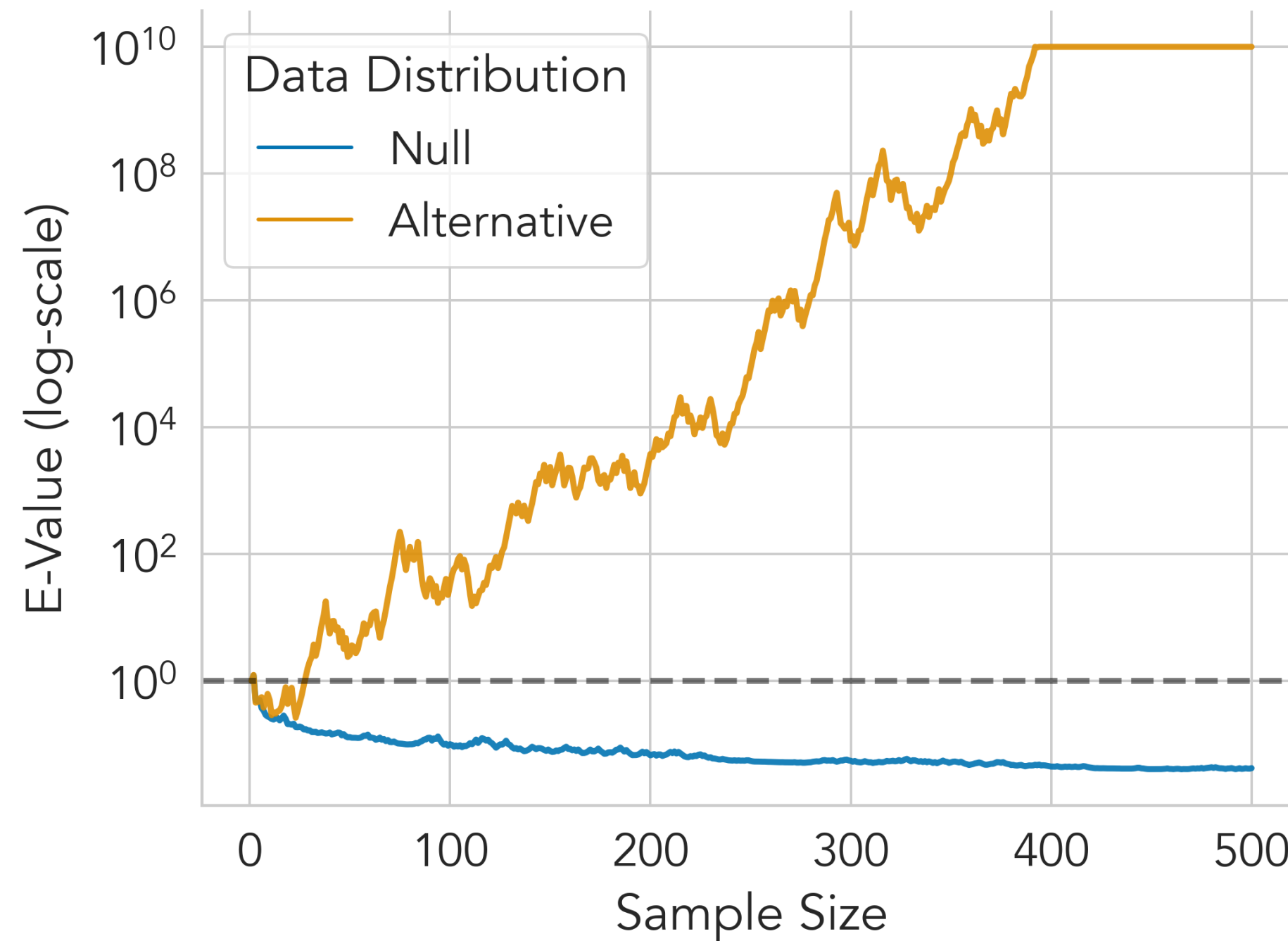
E-process $(E_t)_{t \geq 0}$

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:

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

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

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



Jean Ville

Alternatives based on time-uniform confidence bands

- Instead of taking mixtures or minima over z , we can also consider testing the KS parameter:

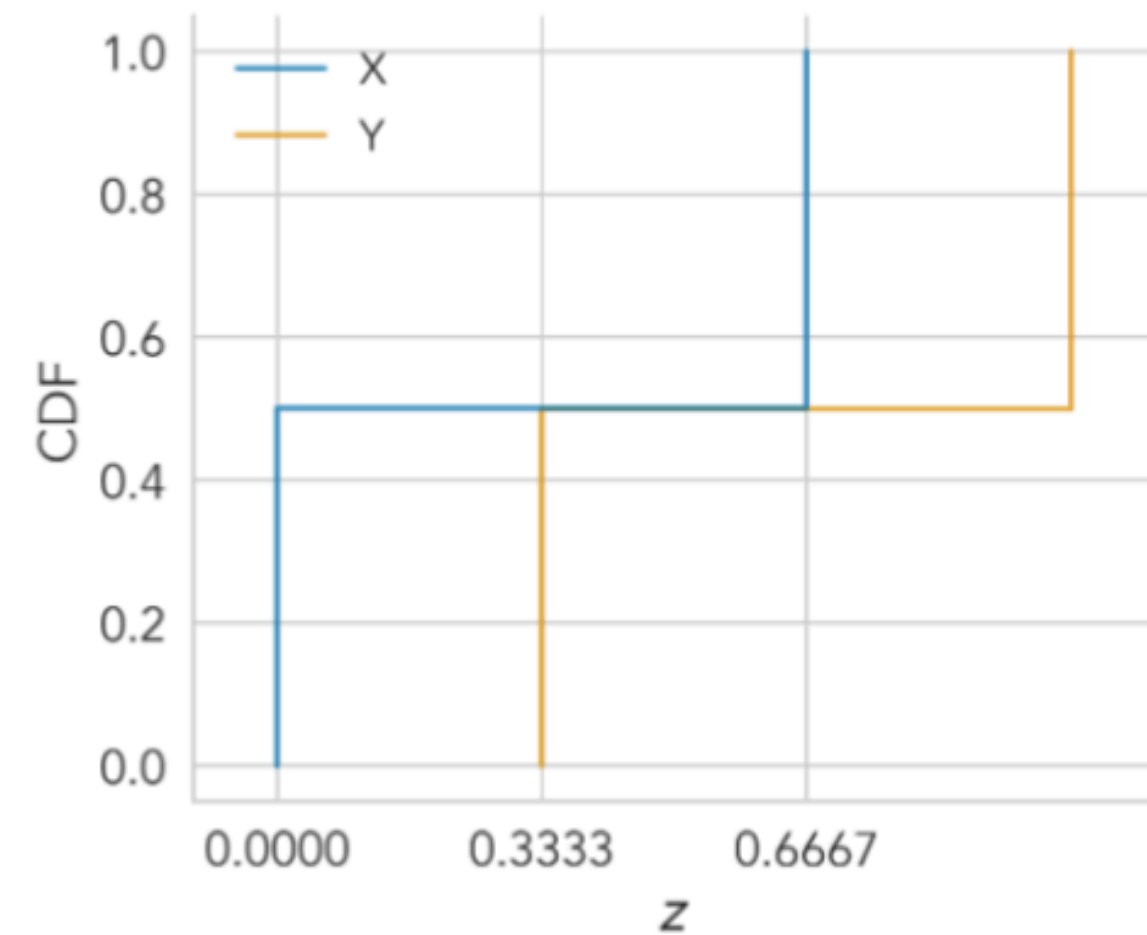
$$\theta = \sup_z [F_Y(z) - F_X(z)]_+ \leq \|F_X - F_Y\|_\infty.$$

- The theory of reverse martingales yields a **time-uniform concentration* around θ** , with which we can either construct a reverse martingale or a confidence sequence around θ :

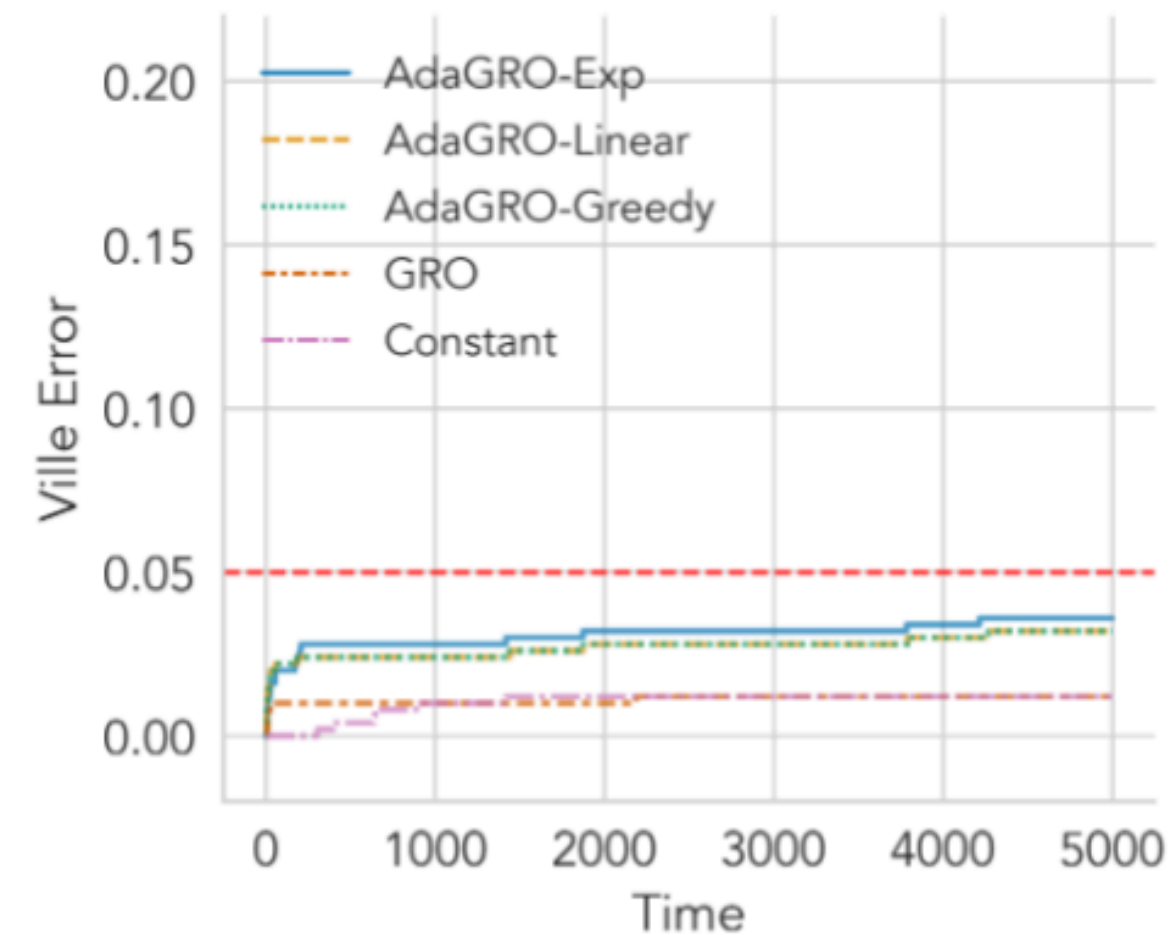
$$P(\exists t \geq 1 : \theta \in C_t^{(1-\alpha)}) \geq 1 - \alpha, \text{ where}$$
$$C_t^{(1-\alpha)} = \left(\|\hat{F}_X(t) - \hat{F}_Y(t)\|_\infty - \kappa_t, \|\hat{F}_X(t) - \hat{F}_Y(t)\|_\infty + \gamma_t \right).$$

- More conservative in practice for testing SD.

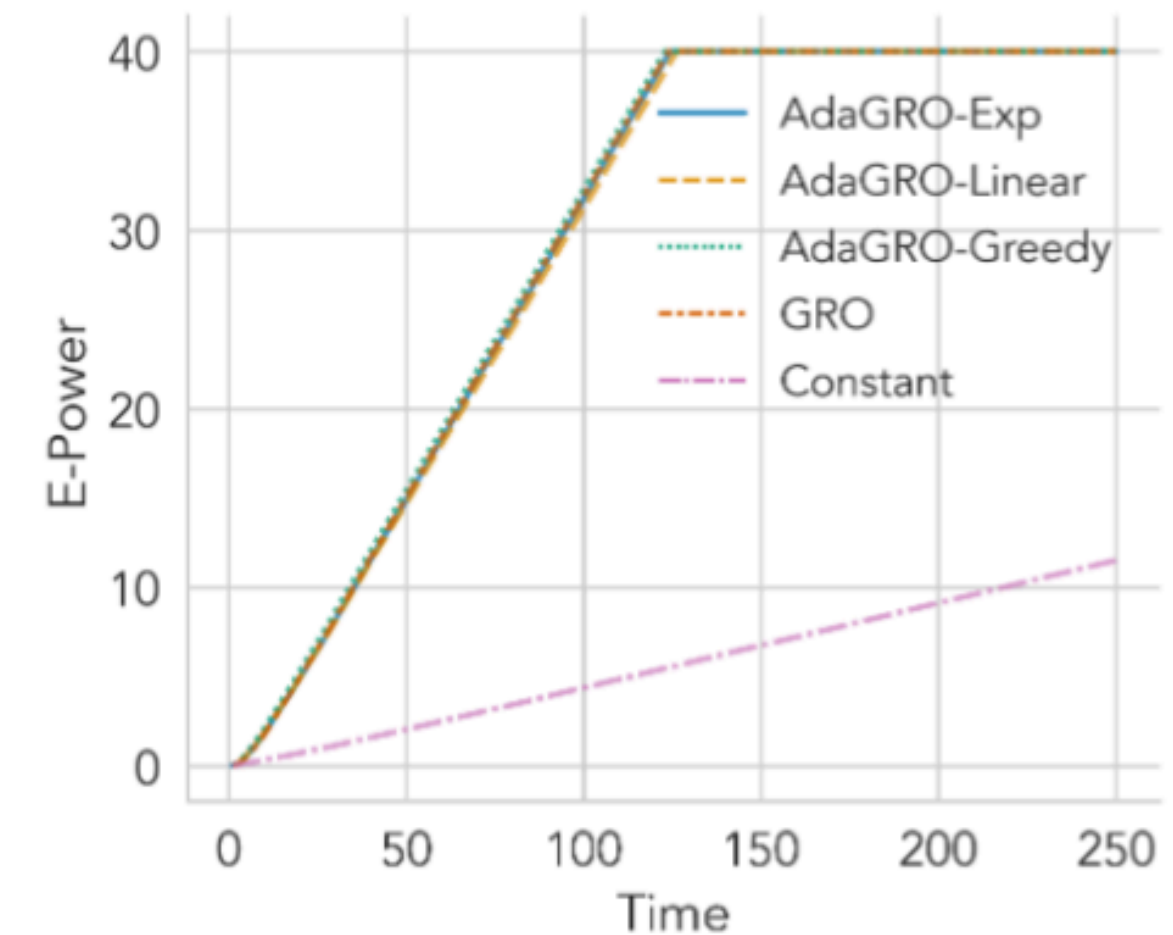
Simulation #A: Finite support with anti-monotonicity



(a) F_X and F_Y .



(b) Ville error for $\mathcal{H}_0 : X \leq_1 Y$.



(c) E-power against $\mathcal{H}_0 : Y \leq_1 X$.

$$\mathbb{P}(X = 0, Y = 1) = 1/2$$

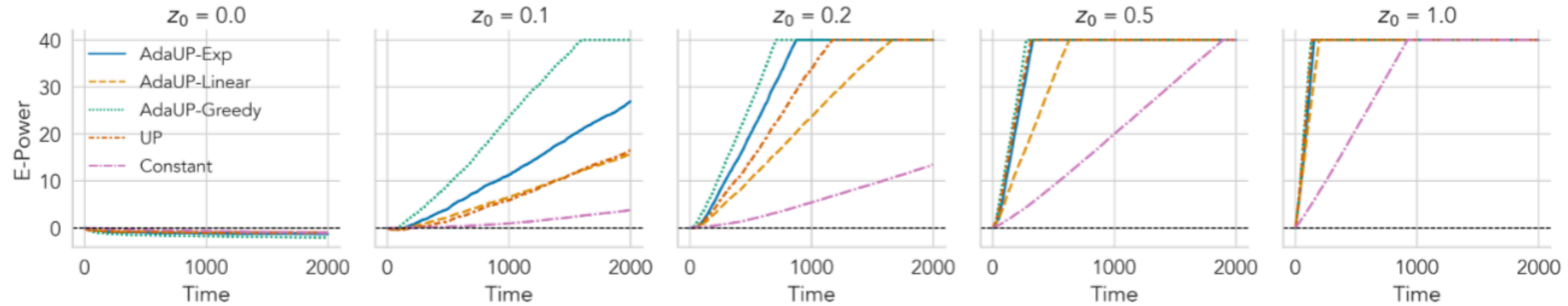
$$\mathbb{P}(X = 2/3, Y = 1/3) = 1/2$$

$$(\rho(X, Y) = -1)$$

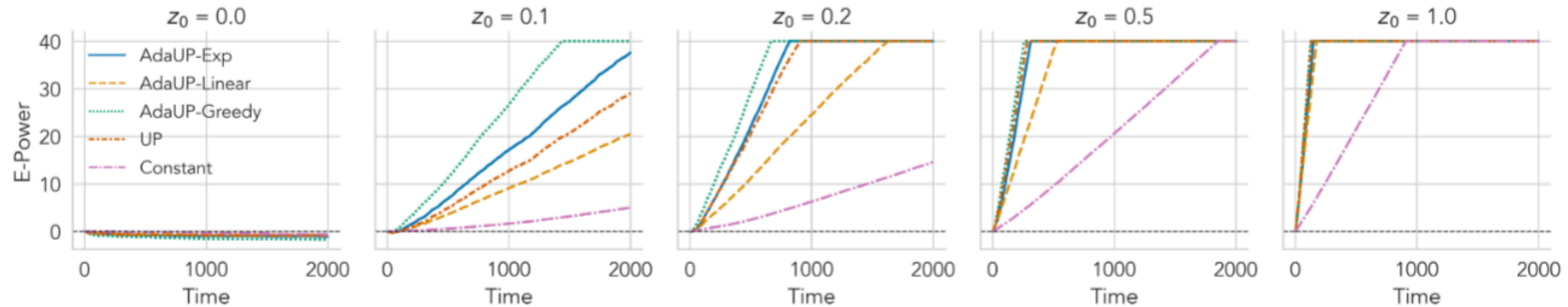
$$\text{Ville Error: } \hat{\mathbb{P}}(\exists t : E_t \geq 1/\alpha)$$

$$\text{Rejection Time: } \tau_\alpha = \inf \{t : E_t \geq 1/\alpha\}$$

Simulation #B: Testing 2-SD and 3-SD nulls

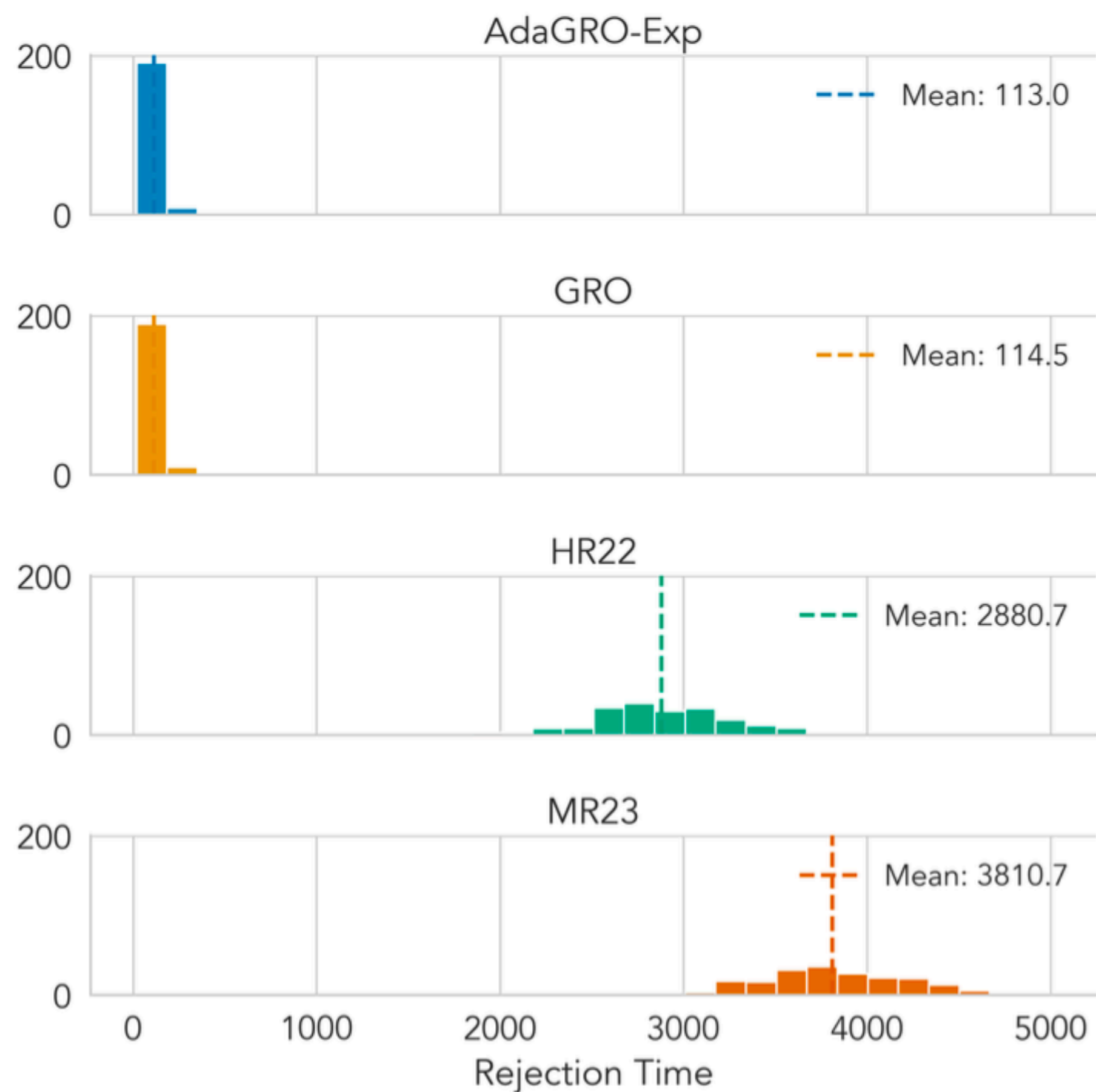


(a) E-power against $\mathcal{H}_0 : Y \preceq_2 X$ (2-SD).

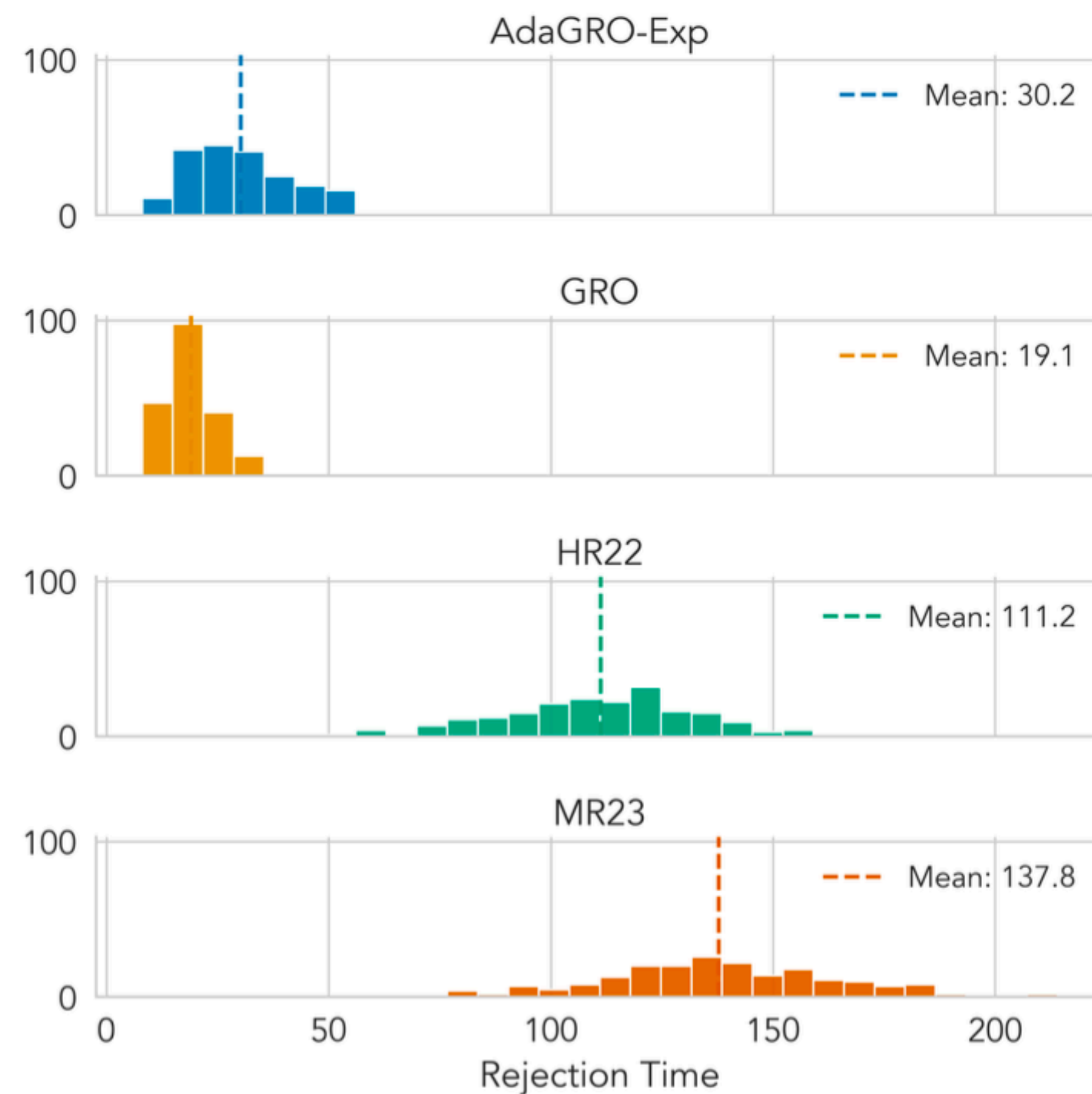


(b) E-power against $\mathcal{H}_0 : Y \preceq_3 X$ (3-SD).

Simulation #C: Comparison with time-uniform CDF bands



(a) Substantial contact between CDFs ($z_0 = 0.2$).



(b) No contact between CDFs ($z_0 = 1.0$).

The End