Betting on Bets:

Anytime-Valid Tests for Stochastic Dominance


Sebastian Arnold* *CWI Amsterdam*


YJ Choe* *INSEAD*


Marco Scarsini *LUISS*


Ilia Tsetlin *INSEAD*

arXiv: 2604.21851

Motivation

📊 **Does Y have an upside over X? Particularly when...**

1. *their means are not meaningfully different, or*
2. *"means" aren't even well-defined (e.g., ordinal data).*

📈 **Can we test for upsides in real time? We want...**

1. *validity under continuous monitoring, and*
2. *comparable power to classical (e.g., bootstrap) tests.*

Effects of veteran status on future incomes (Abadie, 2002)

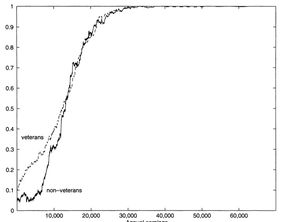
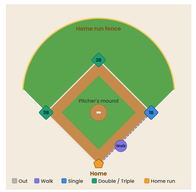


Figure 2. Estimated Distributions of Potential Earnings for Compliers.

"3rd time effect" in baseball



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.

Building block: Growth-rate optimal e-values

Def. An ***e-process*** $(E_t)_{t\in\mathbb{N}}$ for a composite hypothesis H is a sequence of nonnegative variables s.t.

$$\mathbb{E}_{\mathbb{P}}[E_\tau] \le 1 \text{ for every } \mathbb{P} \in H \text{ and stopping time } \tau.$$

At any given time t, we refer to E_t as an ***e-value*** for H.

Example: *likelihood ratio process* of any alternative $\mathbb{Q}$ over $\mathbb{P}$. An e-process is a nonparametric & composite generalization.

🎲 Interpretation as *statistical evidence:* how much wealth we multiply by betting against the odds put forth by $\mathbb{P}$.

Lemma. Fix any test threshold $z \in \mathbb{R}$. For any $\lambda \in [0,1]$,

$$S(\lambda, z) = 1 + \lambda[\mathbf{1}(X \le z) - \mathbf{1}(Y \le z)]$$

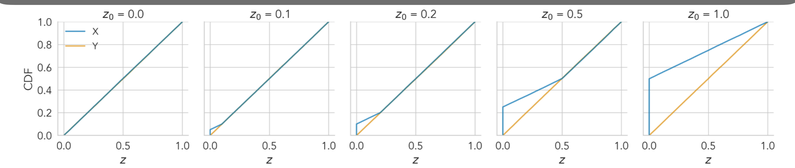
is an **e-value** for $H_0(z)$. (λ is the "betting parameter".)

Proposition. Given any alternative $\mathbb{Q} \notin H_0(z)$, there is a **growth-rate optimal (GRO)** bet:

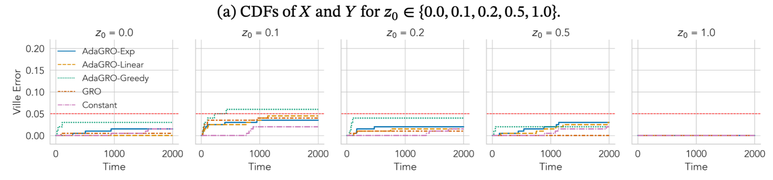
$$\lambda^{\mathrm{GRO}}(z) = \frac{\mathbb{Q}(X \le z < Y) - \mathbb{Q}(Y \le z < X)}{\mathbb{Q}(X \le z < Y) + \mathbb{Q}(Y \le z < X)}.$$

Multiplying these e-values yields an e-process: $E_t = \Pi_{\ell=1}^{t} S_\ell$.

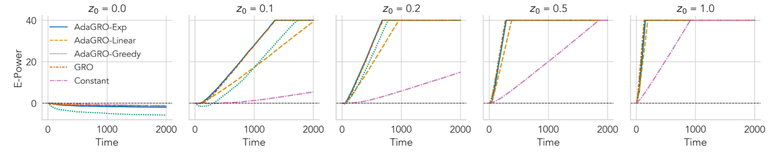
Simulation Experiments



(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 $\mathcal{H}_0 : X \preceq_1 Y$.



(c) E-power against $\mathcal{H}_0 : Y \preceq_1 X$.

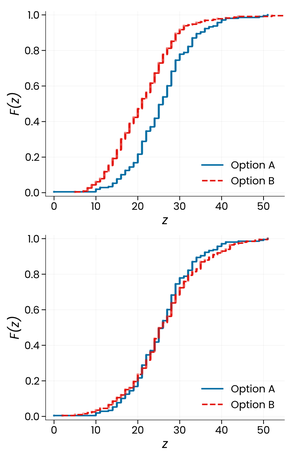
- *Ville Error (cumulative type I error):* $\hat{\mathbb{P}}(\exists \ell \le t : E_\ell \ge 1/\alpha)$
- *E-Power:* $\hat{\mathbb{E}}[\log E_t]$
- *Rejection Time:* $\tau_\alpha = \inf\{t : E_t \ge 1/\alpha\}$

The sequential SD testing problem

X, Y: RVs with marginal CDFs F_X, F_Y.

Def. X ***stochastically dominates*** Y if F_X falls entirely below F_Y:

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



"Y has no upside": *At every threshold* z, *the prob. that* Y *exceeds* z *is no more than the prob. that* X *exceeds* z.

The sequential SD testing problem:

With a stream of data $(X_1, Y_1), (X_2, Y_2), \ldots \sim \mathbb{P}_{XY}$, test:

$$H_0 : Y \preceq_1 X \text{ vs. } H_1 : Y \not\preceq_1 X.$$

Fact. The null H_0 is an **intersection:** $H_0 = \cap_z H_0(z)$, where $H_0(z) = \{\mathbb{P} : \mathbb{P}(X \le z) \le \mathbb{P}(Y \le z)\}$.

Main result: *Mixture* GRO e-process is powerful

To test the *intersection* null H_0, we can take mixtures over multiple test thresholds $Z_t \subset \mathbb{R}$ at each time t.

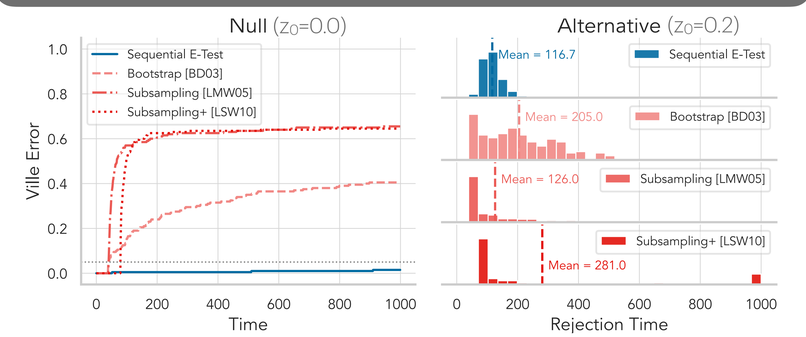
Theorem (GRO e-process is anytime-valid & powerful).

(a) $(E_t)_{t\in\mathbb{N}}$ is an e-process for the 1-SD null H_0.

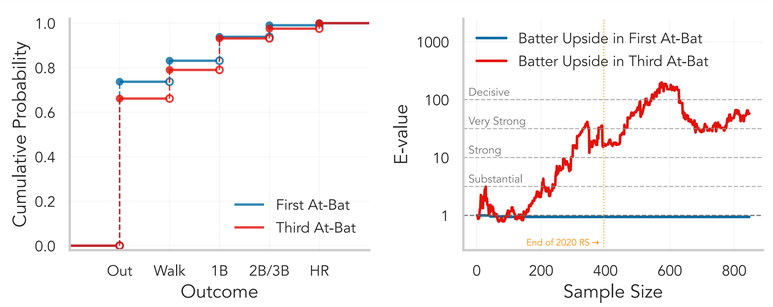
(b) For "reasonable" predictable mixtures $(\psi_t)_{t\in\mathbb{N}}$, $(E_t)_{t\in\mathbb{N}}$ is **powerful** against any non-1-SD alternative $\mathbb{Q}$:

$$\mathbb{Q}\left(\liminf_{t\to\infty} E_t = \infty\right) = 1\,. \textit{ (yields a test of power one)}$$

Comparison with non-SAVI methods



⚾ Application: Monitoring the 3TTO effect



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