Core idea: Testing by betting (Ville+Vovk/Shafer)

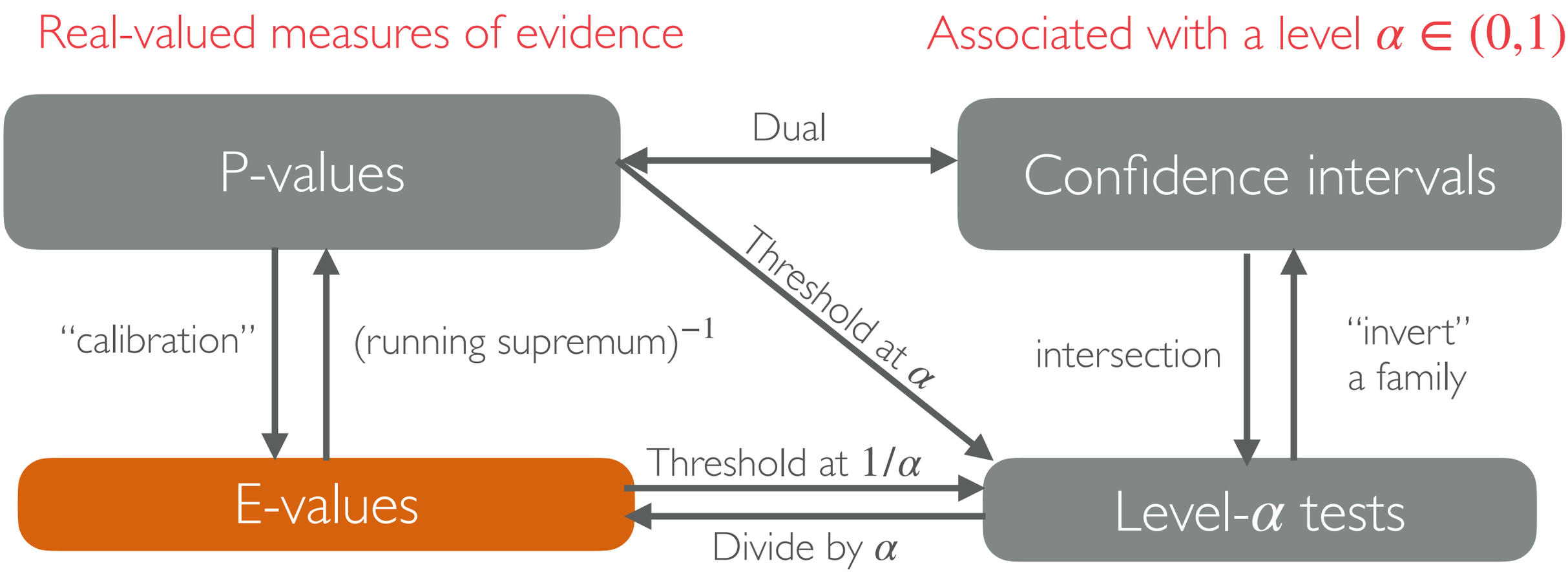
In order to test a hypothesis, one sets up a game such that:
if the null is true, no strategy can systematically make (toy) money,
but if the null is false, then a good betting strategy can make money.

Wealth in the game is directly a measure of evidence against the null.

Each strategy of the statistician = a different estimator or test statistic.
So there are "good" and "bad" strategies for betting,
just as there are good and bad estimators or test statistics.

Testing and estimation == game and strategy design.

Non-sequential testing/estimation



Every level-α hypothesis test can be recovered by thresholding an e-value at $1/\alpha$, and so

(a) nothing is lost when "restricting" attention to e-values

(b) e-values are equally fundamental to p-values and tests

A p-process (or anytime-valid p-value) for a null $H_0 : P \in \mathscr{P}$ is a sequence $(p_t)_{t\geq 1}$ that satisfies

For any stopping time τ, $P \in \mathscr{P} : P(p_\tau \leq \alpha) \leq \alpha$.

Johari et al. (2015, 2021),
Howard, Ramdas, et al. (2018, 2021)

An e-value for H_0 is a $[0,\infty]$-valued r.v. e s.t.

$\forall P \in \mathscr{P}$, $\mathbb{E}_P(e) \leq 1$. (e for evidence or expectation)

An e-process for H_0 is a sequence of e-values $(e_t)_{t\geq 1}$

$$\sup_{\tau} \sup_{P\in\mathscr{P}} \mathbb{E}_P(e_\tau) \leq 1.$$

As we will later see, $(1/\sup_{s\leq t} e_s)_{t\geq 1}$ yields a p-process.

Howard, Ramdas, et al. (2018-2021)
Grunwald et al. (2019-2021)
Shafer (2020), Vovk & Wang (2021)

A "level-α sequential test" for a null hypothesis $\mathscr{P}$ is an adapted binary sequence (ϕ_n) such that for every $P \in \mathscr{P}, P(\exists t \geq 1 : \phi_t = 1) \leq \alpha$.
For any stopping time $\tau, P \in \mathscr{P}, \mathbb{E}_P[\phi_\tau] \leq 1$.

A "confidence sequence (CS)" for a functional ψ is a sequence of confidence intervals (L_n, U_n) that are constructed from the first n samples, and have a uniform (simultaneous) coverage guarantee:

$$\mathbb{P}(\forall t \geq 1 : \psi(\mathbb{P}) \in (L_t, U_t)) \geq 1 - \alpha\,.$$

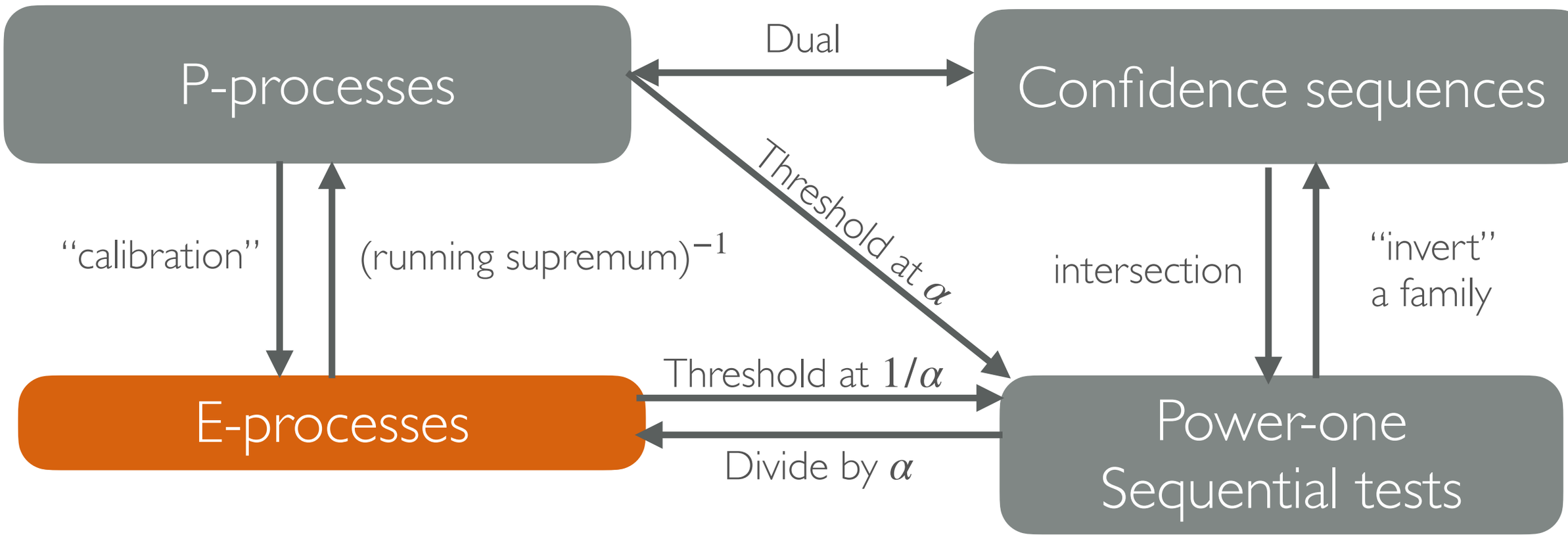
For any stopping time τ : $\mathbb{P}(\psi(\mathbb{P}) \notin (L_\tau, U_\tau)) \leq \alpha$

Darling, Robbins '67, '70s
Lai '76, '84
Robbins, Siegmund '70s

Sequential anytime-valid inference (SAVI)

Real-valued measures of evidence

Associated with a level $\alpha \in (0,1)$



Every level-α sequential test can be recovered by thresholding an e-process at $1/\alpha$, and so

(a) nothing is lost when "restricting" attention to e-processes

(b) nothing more general than an e-process is required

One way to test a composite null vs. composite alternative

$$\frac{\text{Likelihood of alternative } P_\nu}{\text{Likelihood of null } P_\theta}$$

"Frequentist"

$$\frac{\text{Maximum (alt) likelihood}}{\text{Maximum (null) Likelihood}}$$

Bayes

$$\frac{\text{Mixture (alt) likelihood}}{\text{Mixture (null) Likelihood}}$$

Game-theoretic
(wealth in a game)

$$\frac{\text{Mixture or plug-in (alt) likelihood}}{\text{Maximum (null) Likelihood}}$$

Can use "prior" information to bet
on the alternative.
But evidence is compared to best null.

Only the last option is an e-process (the "universal inference" e-process).
It has the asymptotically optimal growth rate (Dixit+Martin'23).

* : no dominating reference measure
for the set of distributions

Nontrivial (e-power) test martingales exist

Testing symmetry*

Two-sample testing*

Bounded means*

T-test (in shrunk filtration)

Exchangeability* (in shrunk filtration)

Independence testing* (in shrunk filtration)

Nontrivial test supermartingales exist

SubGaussian distributions* (or any bounded MGF)

Robust, heavy-tailed mean estimation*

Nontrivial test martingales exist

Testing symmetry*

Two-sample testing*

Bounded means*

T-test (in shrunk filtration)

Exchangeability* (in shrunk filtration)

Independence testing* (in shrunk filtration)

Nontrivial e-processes exist

Any composite $\mathscr{P}$: "universal inference"

Exchangeability* (in original filtration)

T-test (in original filtration)

Nontrivial test supermartingales exist

SubGaussian distributions* (or any bounded MGF)

Robust, heavy-tailed mean estimation*

Nontrivial test martingales exist

Testing symmetry*

Two-sample testing*

Bounded means*

T-test (in shrunk filtration)

Exchangeability* (in shrunk filtration)

Independence testing* (in shrunk filtration)

Nontrivial e-processes exist

If M is an e-process for $\mathscr{P}$,
then it is an e-process for the convex hull of $\mathscr{P}$.

Nontrivial test supermartingales exist

If M is a test supermartingale for $\mathscr{P}$,
then it is also one for the **fork-convex hull** of $\mathscr{P}$.

Informally, a fork-convex combination of P_1 and P_2 would be a distribution that followed P_1 until some stopping time, and then with some probability, the rest of the data is drawn from P_2 (conditional on what's been drawn so far).

Being careful about reference measures, taking limits and closures, yields the fork-convex hull.

Ramdas et al. "Testing Exchangeability…" (IJAR, 2022)

E versus P

In the P-world, we judge tests by probabilities:

$$P_{H_0}(\phi(X) = 1)$$

and

$$P_{H_1}(\phi(X) = 1)$$

This is a theory of decision making

In the E-world, we judge e-values by expectations:

$$\mathbb{E}_{H_0}[W]$$

and

$$\mathbb{E}_{H_1}[\log W]$$

("e-power" or "growth rate")

This is a theory of evidence

We are designing a complementary theory to (say) Neyman-Pearson.
When you see an e-value or e-process, ask about its e-power or growth rate,
not its power (a p-concept) — there is some loss in transforming one to other.

E-values offer a middle ground in the Bayesian-Frequentist debate

Bayesians criticize frequentists on two fronts:

- Cannot update evidence on seeing more data
- Too pessimistic/worst-case (not using prior knowledge)

Frequentists criticize Bayesians mainly on the subjectivity of their evidence.
For objectivity, prior must be revealed and agreeable.

Game-theoretic statistics balances the strengths:

- It decouples the actor (the statistician who bets) from the interpreter of evidence (anyone else). The evidence can be updated on the fly.
- The log-optimal betting strategy is Bayesian: if the statistician has a prior over the *alternative (how they believe the world works)*, they *should* use it to gamble.
- But the measure of evidence is objectively interpretable without knowledge or care for the priors employed by the gambler. E-values have frequentist validity.

My null hypothesis is that roulette tables tend to be fair.
Suppose you tell me: roulette Table 5 is skewed (and you have a hunch how).
Then, you go to roulette table 5 and multiply your initial wealth 10-fold.
No matter what your priors were or how you gambled, that's very strong evidence:
I am now inclined to believe the roulette table was indeed skewed.