Testing mean and variance by e-processes with applications in finance

Zhanyi Jiao

https://zhanyij.github.io/

Department of Mathematics
Illinois State University

Perspectives on Actuarial Risks in Talks of Young Researchers
May 19, 2025

Agenda

Testing mean and variance

A tester is interested in testing whether

$$\mathbb{E}[X_i|\mathcal{F}_{i-1}] \leqslant \mu_i \quad \text{and} \quad \text{Var}(X_i|\mathcal{F}_{i-1}) \leqslant \sigma_i^2 \quad \text{for each } i.$$

- X_i: data points arrive sequentially $(i = 1, \cdots, n)$
  - possibly dependent,
  - each from an unknown distribution (possibly different).
- $\mathcal{F}_i$: σ-field generated by $X_1, X_2, \cdots, X_{i-1}$.
- μ_i and σ_i are $\mathcal{F}_{i-1}$ measurable.

Testing mean and variance

1. Testing both the mean and the variance
   - Time-series switch away from a given regime (mean and variance bounds).
   - Quality control/financial risk assessment/clinical trials...
2. Testing the mean under the knowledge of an upper bound on the variance
   - Useful when comparing with the literature (Howard et al.'21 AOS, Wang/Ramdas'23 SPA, Waudby-Smith/Ramdas'24 JRSSB)
3. Testing the mean and the variance with shape information of the distribution
   - Symmetry, unimodality and their combination.

Challenges

- Data is neither independent nor identically distributed
  - Inference on the underlying distributions is problematic.
- Each observation may come from an unknown and different distribution
  - Non-stationarity invalidates many classical models.
- Irregular composite models
  - Incorporating shape constraints makes testing even more challenging.
- Sequential adaptation
  - Dynamically controlling false positives (Type I errors) is far more complex.

Progress

What is an e-value?

Definition (e-variables, e-values, and e-processes)

1. An e-variable for testing $\mathcal{H}$ is a non-negative random variable $E : \Omega \to [0, \infty]$ that satisfies $\int E \, \mathrm{d}Q \leqslant 1$ for all $Q \in \mathcal{H}$.
   - Realized values of e-variables are e-values.
2. Given a filtration, an e-process for testing $\mathcal{H}$ is a non-negative process $(E_t)_{t=0,1,\ldots,n}$ such that $\int E_\tau \, \mathrm{d}Q \leqslant 1$ for all stopping times τ and all $Q \in \mathcal{H}$.

Remark: For simple hypothesis $\{\mathbb{P}\}$

- e-variable: non-negative random variable with mean $\leqslant 1$
- e-process: (e.g.) non-negative supermartingale with initial value $\leqslant 1$

An approachable interpretation of e-value (Grünwald et al.'24 JRSSB)

- Imagine a gamble/lottery (contract, ticket, investment) that one can buy for \$1, and pays \$$E$ if he won (fractional amounts are allowed).
- Null hypothesis: one expects NOT to gain any money by buying such lotteries.
- For any $r \in \mathbb{R}^+$, upon buying r lotteries one expected to end up with $r\mathbb{E}[E] \leqslant r$.
- If the observed value of E is large, say $20 = 1/0.05$ - one would have gained a lot.

  $\Rightarrow$ something might be wrong about the null

A naive e-process

- Sequential gambles: \$1 for payoff $\$E_i$ for i-th gamble.
- Start by investing \$1 in 1-st gamble (E_1) and, after observing E_1, reinvest all our new capital $\$E_1$ into 2-nd gamble (E_2).
- After observing E_2, new capital becomes $\$E_1 \cdot E_2$.
- Keep playing the gamble games and reinvest all new capital to the next gamble.

  $\Rightarrow$ After t-th gamble, the capital becomes: $M_t = \prod_{i=1}^{t} E_i$ (e-process).
- If, under the null, we do not expect to gain any money for any of the gamble.

  $\Rightarrow$ we should not expect to gain any money under whichever strategy we employ for deciding whether or not to reinvest.

Basic settings

1. Construct a e-variable for the hypothesis based on one data point
   - Testing a global null (meta-analysis)
   - Sequential testing
   - Testing by betting (relevant to e-value)
2. Combine each e-value using different procedures (e-process)
3. Compare the e-value with its threshold level

Non-parametric composite hypotheses (one data point)

- $\mathcal{H}(\mu,\sigma) = \{Q : \mathbb{E}^Q[X] \leqslant \mu,\ \ \mathrm{Var}^Q(X) \leqslant \sigma^2\}$
- $\mathcal{H}_{\mathrm{S}}(\mu,\sigma) = \{Q \in \mathcal{H}(\mu,\sigma) :$ X is symmetrically distributed$\}$
- $\mathcal{H}_{\mathrm{U}}(\mu,\sigma) = \{Q \in \mathcal{H}(\mu,\sigma) :$ X is unimodally distributed$\}$
- $\mathcal{H}_{\mathrm{US}}(\mu,\sigma) = \mathcal{H}_{\mathrm{U}}(\mu,\sigma) \cap \mathcal{H}_{\mathrm{S}}(\mu,\sigma)$

E-variables for hypotheses (one data point)

- $\mathcal{H}(\mu,\sigma) = \{Q : \mathbb{E}^Q[X] \leqslant \mu,\ \mathrm{Var}^Q(X) \leqslant \sigma^2\}$.

  $\Rightarrow$ e-variable: $E_0 = (X-\mu)_+^2/\sigma^2$.

- $\mathcal{H}_{\mathrm{S}}(\mu,\sigma) = \{Q \in \mathcal{H}(\mu,\sigma) : X \text{ is symmetrically distributed}\}$.

  $\Rightarrow$ e-variable: $E = 2E_0$.

- $\mathcal{H}_{\mathrm{U}}(\mu,\sigma) = \{Q \in \mathcal{H}(\mu,\sigma) : X \text{ is unimodally distributed}\}$.

  $\Rightarrow$ e-variable: $E = E_0$.

- $\mathcal{H}_{\mathrm{US}}(\mu,\sigma) = \mathcal{H}_{\mathrm{U}}(\mu,\sigma) \cap \mathcal{H}_{\mathrm{S}}(\mu,\sigma)$.

  $\Rightarrow$ e-variable: $E = 2E_0$.

E-process and betting process

The e-process is constructed in the sense of wealth process

$$M_t = (1 - \lambda_t + \lambda_t E_t) M_{t-1} = \prod_{i=1}^{t} (1 - \lambda_i + \lambda_i E_i)$$

- Initial capital: $M_0 = 1$.
- Payoff for each bet (for \$1): $1 - \lambda_i + \lambda_i E_i$.
- Betting process/strategy: $\lambda_i \in [0, 1] \Rightarrow$ how much fraction to bet (Kelly criterion) $\Rightarrow$ maximize expected value of the log wealth.
- Type-I error control guaranteed by Ville's inequality

$$\mathbb{P}\left(\max_{t \in \{1,\ldots,n\}} M_t \geqslant 1/\alpha \right) \leqslant \alpha.$$

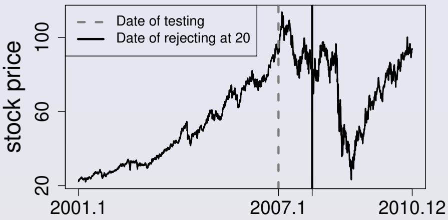
Progress

Empirical study with financial data

- Test $\mathcal{H}(\mu, \sigma)$ on the daily losses of stocks during 2007-2008 financial crisis.
  - Calculate daily losses by $L_t = -(S_{t+1} - S_t)/S_t$, where S_t is the close price at day t.
  - Estimate the $\hat{\mu}$ and variance $\hat{\sigma}$ for the null with data from 2001.1.1 - 2006.12.31.
  - Compute e-values for daily losses from 2007.1.1, fed into e-processes. (e-mixture and e-GREE)
- Calculate the number of trading days to reject the null.
  - Early warning: $e \in (2, 10^{1/2}) \Rightarrow$ threshold $e = 2$.
  - Substantial evidence: $e \in (10^{1/2}, 10) \Rightarrow$ threshold $e = 5$ and $e = 10$.
  - Very strong evidence: $e \in (10, 10^{3/2}) \Rightarrow$ threshold $e = 20$[1].
- Choose 22 stocks from 11 sectors of the S&P 500 with largest capitalization.

[1] In accordance with Jeffrey's rule of thumb about e-values.

An example



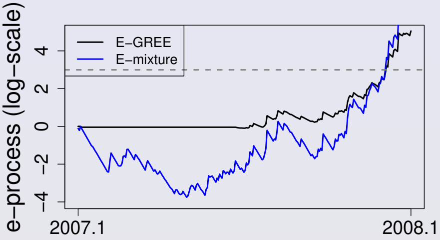


Figure: Sample mean and variance $\hat{\mu} = -0.001028$ and $\hat{\sigma} = 0.012123$ estimated from historical data for Simon Property (SPG) stock from 2001.1.1 to 2006.12.31.

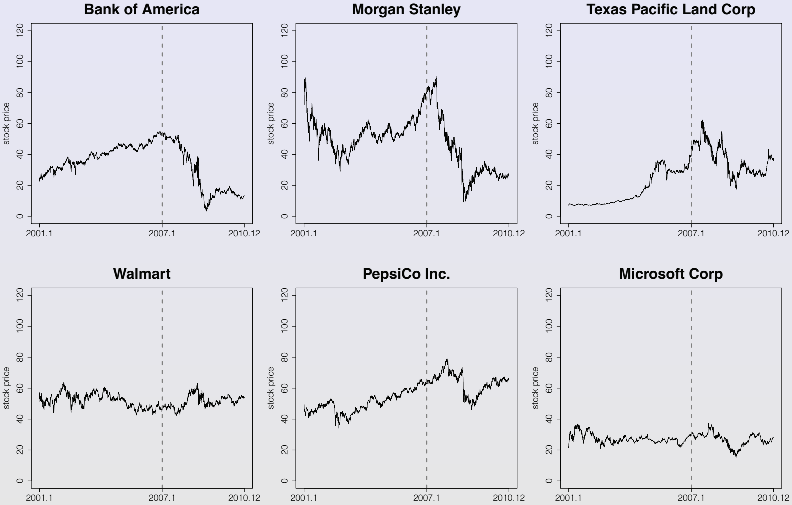
Other stocks

		E-GREE				E-mixture			
	Threshold	2	5	10	20	2	5	10	20
Financials	Bank of America	378	385	385	393	393	394	395	403
	Morgan Stanley	429	439	445	447	447	447	447	447
Utilities	The Southern	-	-	-	-	-	-	-	-
	Duke Energy	-	-	-	-	-	-	-	-
Communication Services	Verizon Comms.	-	-	-	-	-	-	-	-
	AT&T	-	-	-	-	-	-	-	-
Consumer Staples	Walmart	-	-	-	-	-	-	-	-
	PepsiCo	-	-	-	-	-	-	-	-
Consumer Discretionary	Ford Motor	476	491	498	565	546	594	594	594
	Las Vegas Sands	442	445	447	450	451	454	457	457
Energy	Texas Pacific Land	158	244	261	269	242	261	261	263
	Pioneer	496	622	-	-	-	-	-	-
Material	Southern Copper	476	496	537	-	539	-	-	-
	Air Products	476	516	537	-	-	-	-	-

Other stocks (cont.)

		E-GREE				E-mixture			
	Threshold	2	5	10	20	2	5	10	20
Health Care	Johnson & Johnson	-	-	-	-	-	-	-	-
	Pfizer	-	-	-	-	-	-	-	-
Technology	Int. Business Machines	-	-	-	-	-	-	-	-
	Microsoft	-	-	-	-	-	-	-	-
Industrials	General Electric	537	546	578	-	-	-	-	-
	United Parcel Service	476	524	542	632	542	604	-	-
Real Estate	Simon Property	165	224	242	254	223	239	250	253
	Prologis	264	271	271	296	270	271	271	275

Stock price sample paths



Progress

Bactesting risk measure

Daily observations

- ρ forecast r_t (risk measure to be tested)
- ϕ forecast z_t (auxiliary statistic)
- realized loss L_t

Hypothesis to test

$$\mathcal{H}_0 : \rho(L_t|\mathcal{F}_{t-1}) \leqslant r_t \text{ and } \phi(L_t|\mathcal{F}_{t-1}) = z_t \text{ for } t = 1, \ldots, T$$

In our case (mean and variance):

- Risk measure to be backtest: $\rho(L_t \mid \mathcal{F}_{t-1}) = \mathrm{Var}(L_t \mid \mathcal{F}_{t-1}) \leqslant \sigma_t^2$;
- Auxiliary statistics: $\phi(L_t \mid \mathcal{F}_{t-1}) = \mathbb{E}(L_t \mid \mathcal{F}_{t-1}) = \mu_t$.

Model-free e-statistics

Definition (Model-free e-statistics)

A model-free e-statistic for $(\rho, \phi) : \mathcal{M} \to \mathbb{R}^2$ is a measurable function $e : \mathbb{R}^3 \to [0, \infty]$ satisfying $\int e(x, \rho(F), \phi(F))\, \mathrm{d}F \leqslant 1$ for each $F \in \mathcal{M}$.

Example:

- $(\mathbb{E}, \mathrm{Var})$: $e(x, r, z) = (x - z)^2/r$. (Fan/J./Wang'24 Biometrika)
- (VaR_α): $e(x, r) = \mathbb{1}_{\{x>r\}}/(1-\alpha)$. (Wang/Wang/Ziegel'25 MS)
- $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$: $e(x, r, z) = (x - z)_+/((1-\alpha)(r - z))$. (Wang/Wang/Ziegel'25 MS)
- (ex_α): $e(x, r) = |1 - \alpha - \mathbb{1}_{\{x>r\}}| \left(\frac{x}{r} - 1\right) + 1$. (J./Wang/Zhao'25)
- $(\mathrm{var}_\alpha, \mathrm{ex}_\alpha)$: $e(x, r, z) = \alpha(x - z)_+^2 + (1-\alpha)(x - z)_-^2/r$. (J./Wang/Zhao'25)

E-backtesting

1. Risk measure ρ forecast r_t and ϕ forecast z_t.
2. Observe realized loss L_t
3. Calculate the e-value $E_t = e(L_t, r_t, z_t)$.
4. Decide betting process $\lambda_t \in [0, 1]$ (e.g., e-mixure, e-GREE).
5. Compute the e-process $(M_0 = 1)$

$$M_t = (1 - \lambda_t + \lambda_t E_t) M_{t-1} = \prod_{s=1}^{t} (1 - \lambda_s + \lambda_s E_s).$$

6. Compare with the thresholds (e.g., $E \geqslant 2$, $E \geqslant 10$, $E \geqslant 20$).

Reference

- Fan, Y., Jiao, Z. and Wang, R. (2025). Testing the mean and variance by e-processes. *Biometrika*, **112**(1).
- Grünwald, P., de Heide, R. and Koolen, W. M. (2024). Safe testing. *Journal of the Royal Statistical Society Series B*, **86**(5) 1091–1128.
- Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, **49**(2), 1055–1080.
- Vovk, V. and Wang, R. (2021). E-values: Calibration, combination, and applications. *The Annals of Statistics*, **49**(3), 1736–1754.
- Wang, Q., Wang, R. and Ziegel, J. F. (2025). E-backtesting. *Management Science*, forthcoming.
- Wang, H. and Ramdas, A. (2023). Catoni-style confidence sequences for heavy-tailed mean estimation. *Stochastic Processes and their Applications*, **163**, 168–202.
- Wang, R. and Ramdas, A. (2024). Hypothesis Testing with E-values. *arXiv*:2410.23614.
- Waudby-Smith, I. and Ramdas, A. (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B*, **86**(1): 1–27.

Thank you for your kind attention!

https://zhanyij.github.io/

Hypotheses to test

- $\mathcal{H}$: a collection of probability measures (hypothesis)

$$\mathcal{H} = \{Q : \mathbb{E}^Q[X_i|\mathcal{F}_{i-1}] \leqslant \mu_i,\ \mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leqslant \sigma_i^2 \text{ for } i = 1, \ldots, n\} \tag{1}$$

  - μ_i and σ_i are $\mathcal{F}_{i-1}$-measurable for each i
  - Can be absorbed into X_i by considering $(X_i - \mu_i)/\sigma_i$ instead of X_i

$$\Downarrow$$

$$\mathcal{H}(\mu, \sigma) = \{Q : \mathbb{E}^Q[X_i|\mathcal{F}_{i-1}] \leqslant \mu,\ \mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leqslant \sigma^2 \text{ for } i = 1, \ldots, n\} \tag{2}$$

Comparison with p-value

For a simple hypothesis $\{\mathbb{P}\}$:

E-values E	P-values P
$\mathbb{E}^{\mathbb{P}}[E] \leqslant 1$ and $E \geqslant 0$	$\mathbb{P}(P \leqslant \alpha) \leqslant \alpha$ for $\alpha \in (0,1)$
betting scores, stopped martingales	probability of a more extreme observation
$\mathbb{E}^{\mathbb{P}}[M_t \vert \mathbf{X}]$	$\mathbb{P}(T' \leqslant T(\mathbf{X}) \vert \mathbf{X})$
(conditional) expectation	(conditional) probability
reject when e(data) is large	reject when p(data) is small

Why e-values

E-values	P-values
Valid for arbitrary dependence	Usually need stronger dependence
Valid at any stopping time	Typically no optional validity
Valid for infinite samples	Usually resort to asymptotics
Model free	Often need full distributional information of test statistics and parametric model
Easy to merge	Complicated to merge

Some examples (Wang/Ramdas'22 JRSSB)

- **Robust to dependence**: Suppose we observe non-negative data $X_1, \cdots, X_n$, we wish to test $H_0 : \mathbb{E}[X_i] \leqslant \mu$ for all i. Then, $(X_1 + \cdots + X_n)/(n\mu)$ is an e-value for any dependence structure without knowing the distribution.
- **Robust to misspecfication**: Instead of assuming that the data X is Gaussian to build a p-value (utilize the entire distribution), we may instead assume that it is symmetric about the origin. Then $\exp(\lambda X - \lambda^2 X^2/2)$ is a valid e-variable.
- **Sequential inference**: Suitable for risk measure backtesting where data usually arrives sequentially.
- **Easy to merge**: Hypothesis will be tested with future evidence (by other scientists), meta-analysis...

Two-sided e-values testing the mean given variance

Two-sided null hypothesis $\mathcal{H}(\mu^L, \mu^U, \sigma)$:

$$\mathcal{H}(\mu^L, \mu^U, \sigma) = \left\{Q : \mathbb{E}^Q[X_i|\mathcal{F}_{i-1}] \in [\mu^L, \mu^U] \text{ and } \mathrm{Var}^Q(X_i|\mathcal{F}_{i-1}) \leqslant \sigma^2 \text{ for } i = 1, \dots n\right\}.$$

- If $\mu^L \leqslant \mu^U$, the e-value of one data point:

$$E = \frac{(X - \mu^U)_+^2 + (X - \mu^L)_-^2}{\sigma^2}.$$

- If $\mu^L = \mu^U = \mu$, the e-value of one data point:

$$E = \frac{(X - \mu)^2}{\sigma^2}.$$

Two batch methods

Use all data directly. Independence among $X_1, \ldots, X_n$ is required. A natural statistic is the sample mean $T = \sum_{i=1}^n X_i/n$ (mean at most μ, variance at most σ^2/n)

- E-batch method: An e-variable for $\mathcal{H}(\mu, \sigma)$ or $\mathcal{H}_{\mathrm{U}}(\mu, \sigma)$ is

$$E_0 = n(T - \mu)_+^2/\sigma^2,$$

an e-variable for $\mathcal{H}_{\mathrm{S}}(\mu, \sigma)$ or $\mathcal{H}_{\mathrm{US}}(\mu, \sigma)$ is

$$E_{\mathrm{S}} = 2n(T - \mu)_+^2/\sigma^2.$$

- P-batch method: A p-variable for $\mathcal{H}(\mu, \sigma)$ or $\mathcal{H}_{\mathrm{U}}(\mu, \sigma)$ is

$$P_0 = (1 + E_0)^{-1},$$

a p-variable for $\mathcal{H}_{\mathrm{S}}(\mu, \sigma)$ or $\mathcal{H}_{\mathrm{US}}(\mu, \sigma)$ is

$$P_{\mathrm{S}} = \min\{(2E_0)^{-1}, P_0\}.$$

Choose λ_t

Method 1: Heuristic choice of constant $\lambda_i = \lambda \in [0, 1]$

- E-mixture method: Choose $\lambda_i = \lambda = 0.01 \times \{1, \ldots, 20\}$, average the resulting e-processes over these choices.

Method 2: Adaptive choices of dependent on observed data

- E-GRO (growth-rate optimal) method:
  (Kelly'56, Grünwald/de Heide/Koolen'24 JRSSB)

$$\lambda_i = \arg\max_{\lambda \in [0,1)} \mathbb{E}^{Q_i}[\log(1 - \lambda + \lambda E) \mid \mathcal{F}_{i-1}].$$

- E-GREE (growth-rate for empirical e-statistics) method:
  (Waudby-Smith/Ramdas'24 JRSSB, Wang/Wang/Ziegel'25 MS)

$$\lambda_i = \arg\max_{\lambda \in [0,1)} \frac{1}{i-1} \sum_{j=1}^{i-1} \log(1 - \lambda + \lambda E_j).$$

Simulation settings

- Set $\mu = 0$ and $\sigma = 1$ without loss of generality.
- Concentrate on the null hypothesis $\mathcal{H}(0, 1)$.
- Data generating process $\mathrm{NL}(\nu, \eta^2)$ (independent but not identical)
  - $X_1, X_3, \ldots,$ follow a Normal distribution.
  - $X_2, X_4, \ldots,$ follow a Laplace distribution.
  - Two distribution has same mean ν and the same variance η^2.
- Consider two alternatives:
  1. Data generated from $\mathrm{NL}(0, \eta^2)$ where $\eta > 1$.
  2. Data generated from $\mathrm{NL}(\nu, 1)$ where $\nu > 0$.
- Compute the rejection rate over 1000 runs using the thresholds of $E \geqslant 1/\alpha$ and $P \leqslant \alpha$, with $\alpha = 0.05$. ($\mathbb{P}(E \geqslant 1/\alpha) \leqslant \alpha$)

A comparison of different methods

(a) **P-Fisher:** $P_F = 1 - \chi_{2n}(-2(\log P_1 + \cdots + \log P_n))$.

(b) **P-Simes:** $P_S = \min_{i \in [n]} \frac{n}{i} P_{(i)}$.

(c) **E-mixture:** $M_t^{mix} = \prod_{i=1}^{t}(1 - \lambda + \lambda E_i)$, where $\lambda = 0.01 \times \{1, \ldots, 20\}$.

(d) **E-GREE:** $M_t^{G} = \prod_{i=1}^{t}(1 - \lambda_i + \lambda_i E_i)$, where $\lambda_i = \arg\max_{\lambda \in [0,1)} \frac{1}{i-1} \sum_{j=1}^{i-1} \log(1 - \lambda + \lambda E_j)$.

(e) **E-batch:** $E_0 = n(T - \mu)_+^2 / \sigma^2$, where $T = \sum_{i=1}^{n} X_i / n$.

(f) **P-batch:** $P_0 = (1 + E_0)^{-1}$.

Rejection rates for all methods

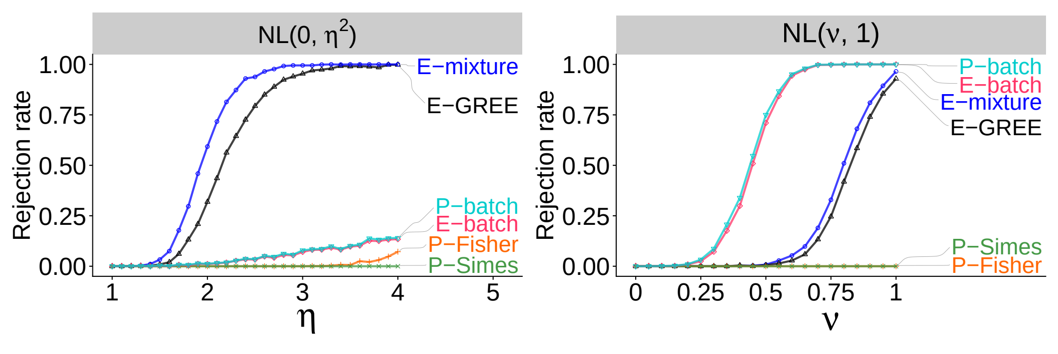


Figure: Rejection rates for all methods for testing $\mathcal{H}(0,1)$ with sample size $n = 100$ over 1000 runs using the threshold 20 for e-value methods, and threshold 0.05 for p-value methods.

Rejection rates for $\mathcal{H}_{\mathsf{S}}$, $\mathcal{H}_{\mathsf{U}}$, $\mathcal{H}_{\mathsf{US}}$

	E-mixture	E-GREE	P-Fisher	P-Simes	E-batch	P-batch
$\mathcal{H}$	0.419	0.315	0.000	0	0.639	0.664
$\mathcal{H}_{\mathsf{S}}$	0.998	0.882	0.000	0	0.900	0.900
$\mathcal{H}_{\mathsf{U}}$	0.419	0.315	0.006	0	0.639	0.664
$\mathcal{H}_{\mathsf{US}}$	0.998	0.882	0.063	0	0.900	0.900

Table: Rejection rates of testing $\mathcal{H}(0,1)$, $\mathcal{H}_{\mathsf{S}}(0,1)$, $\mathcal{H}_{\mathsf{U}}(0,1)$ and $\mathcal{H}_{\mathsf{US}}(0,1)$ with $n = 100$ data generated from the model $\mathrm{NL}(0.5, 2)$.

Elicitability of risk measure

Definition (Elicitablity)

A statistical functional Y is said to be elicitable if there exists a scoring function $S(x, y)$ such that, when forecasting Y, the expected score (expected forecast error)

$$\mathbb{E}[S(x, Y)] = \int S(x, y) dF_Y(y)$$

is uniquely minimized when the forecast x equals the true value Y of the distribution.

- Ensures that the optimal forecast is the true functional value.
  - When $S(x, y) = (x - y)^2$, the optimal forecasts are the mean functional

$$\rho^*(F_Y) = \arg\min_x \mathbb{E}[S(x, Y)] = \mathbb{E}[Y]$$

- Non-elicitable $\Rightarrow$ Cannot simply plug in your forecasts and observed losses into a single consistent metric (score function) to see if your forecasts match the true underlying risk.

Coelicitability (Joint elicitability)

Definition (Coelicitability)

Single-valued statistical functionals $\rho_1(\cdot), \ldots, \rho_k(\cdot)$, $k \geqslant 1$, are called *coelicitable* with respect to a class of distributions $\mathcal{P}$ if there exists a forecasting objective function $S : \mathbb{R}^{k+1} \to \mathbb{R}$ such that

$$(\rho_1(F), \ldots, \rho_k(F)) = \arg \min_{(x_1, \ldots, x_k) \in \mathbb{R}^k} \int S\big(x_1, \ldots, x_k, y\big)\, dF(y), \quad \forall F \in \mathcal{P}.$$

Backtesting risk measures

- Risk measure ρ to backtest
- Define $\mathcal{F}_{t-1}$: all available information up to $t-1$
- Daily observations
  - risk measure forecast r_t for $\rho(L_t)$ given $\mathcal{F}_{t-1}$
  - realized loss L_t

Hypothesis to test

$$\mathcal{H}_0 : \rho(L_t|\mathcal{F}_{t-1}) \leqslant r_t \text{ for } t = 1, \ldots, T.$$

- Only works for sole elicitable risk measure.
  (e.g., expectation, Value-at-Risk, expectile, median shortfall)
- Non-elicitable risk measures cannot be directly backtest.
  (e.g., variance, Expected Shortfall)

Bactesting risk measure with auxiliary statistic

- The model space $\mathcal{M}$ is a set of distributions on $\mathbb{R}$
- ρ is the risk measure to be tested
  - treated as a mapping on either $\mathcal{M}$ or $\mathcal{X}$
- $\phi : \mathcal{M} \to \mathbb{R}$ represents auxiliary statistic
- $\psi = (\rho, \phi)$ represents the collection of available statistical information

Remark:

- If ϕ is a constant (we can take $\phi = 0$), then only the predicted value of ρ is used
- ϕ may be d-dimensional in general, but we focus on $d = 0$ (constant ϕ) or $d = 1$