Contaminated by Construction: Separating Simulation Noise from Model Risk in ES Backtests

Separating Simulation Noise from Model Risk in ES Backtests

Dr. Tamás Nagy

tnagyphd@gmail.com

Working Paper

Executive Summary

Banks are required to backtest their risk models — to check, after the fact, whether the risk numbers they reported were accurate. For Expected Shortfall (ES), the standard risk measure under Basel III, this backtesting is weak: the statistical tests frequently fail to catch models that underestimate tail risk. The usual explanation is small sample sizes — only about six tail observations per year. But there is a second source of weakness that is entirely self-inflicted: banks estimate ES using Monte Carlo simulation, and the randomness of the simulation contaminates the very test designed to catch errors. How much damage does this contamination cause? Until now, nobody has measured it.

This paper provides the measurement. We prove that the variance of the standard Acerbi-Székely backtest statistic decomposes cleanly into two additive components: an irreducible part driven by return randomness, and an eliminable part driven by Monte Carlo estimation of ES. This decomposition is the paper's central result. It tells a bank — or a regulator — exactly how much of a backtest outcome reflects genuine model quality and how much is computational noise. For a typical portfolio backtested with 1,000 simulation paths, Monte Carlo contributes roughly a third of the total test variance. Even at 10,000 paths, it still adds 3–10%, depending on tail heaviness. This is noise that masquerades as signal, inflating p-values and allowing misspecified models to pass.

For portfolios of correlated lognormal assets — covering the majority of linear equity and FX books — the Monte Carlo component can be eliminated entirely. The Hermite-COS method represents the portfolio's loss distribution as 130 numbers (128 Fourier-cosine coefficients plus two domain bounds), evaluated via Gauss-Hermite quadrature. From these 130 numbers, a bank can compute ES, VaR, the full CDF, and the density — all deterministically, in milliseconds, with no random variation. This exact computation does more than sharpen the existing test: it enables two qualitatively new tests that Monte Carlo cannot support. A Probability Integral Transform (PIT) test that uses the full predicted distribution to detect model misspecifications invisible to any ES-only backtest, and a tail likelihood ratio test that is optimally sensitive to tail-specific errors. These qualitative upgrades are the paper's strongest practical contribution. A simulation study with 10,000 replications validates all findings.

The broader point extends beyond lognormal portfolios. Even when exact computation is unavailable and Monte Carlo remains necessary — as it will for complex derivatives and exotic structures — the variance decomposition framework still applies. It tells the practitioner how much to trust a given backtest result at a given number of simulation paths, and it provides a principled basis for choosing how many paths are enough.

All mathematical results are formally verified using the Lean 4 proof assistant — a computer program that checks every logical step. This means the theoretical argument contains no hidden errors, sign mistakes, or unjustified leaps. The practical recommendation has two parts: where exact computation is feasible, regulators should require it and adopt the richer backtesting framework it enables; where it is not, they should require banks to report the Monte Carlo variance component alongside the backtest result.

Abstract

Expected Shortfall backtesting under Basel III/IV suffers from an unmeasured structural weakness: Monte Carlo estimation of ES injects computational noise into the Acerbi-Székely (2014) test statistic, but the magnitude of this contamination has not been quantified.

We prove that the test statistic's variance decomposes additively as $\mathrm{Var}_{\mathrm{returns}} + \mathrm{Var}_{\mathrm{MC}}$, separating the irreducible statistical uncertainty from the eliminable computational noise. This decomposition is the paper's central result: it tells practitioners exactly how much of a backtest outcome is real and how much is Monte Carlo artifact. Simulation confirms that MC contributes roughly a third of total test variance at $M = 1{,}000$ paths and 3–5% at $M = 10{,}000$; the literature suggests contributions of up to 10% for heavier-tailed portfolios.

For portfolios of correlated lognormal assets — covering the majority of linear equity and FX books — the MC component can be eliminated entirely using the Hermite-COS method, which computes ES, the CDF, and the density in closed form from Fourier coefficients. Setting $\mathrm{Var}_{\mathrm{MC}} = 0$ yields 11–26% relative power gains and enables two qualitatively new tests: (i) a Probability Integral Transform test that detects distributional misspecifications invisible to any ES-only backtest; (ii) a tail likelihood ratio test that is optimal for simple tail alternatives by the Neyman-Pearson lemma. The algebraic chain from assumptions to conclusions is formally verified in Lean 4 (10 files, 53 lemmas, zero sorry; proof depth classified in Section 8.2a); probabilistic assumptions are supplied as hypotheses.

Keywords: Expected Shortfall, backtesting, regulatory capital, Monte Carlo noise, formal verification, Lean 4, Basel III, FRTB

JEL Classification: G32, C12, C15

---

1. Introduction

1.1 The backtesting gap

The Fundamental Review of the Trading Book (Basel Committee, 2019) replaced Value-at-Risk with Expected Shortfall at $\alpha = 2.5\%$ as the primary market risk measure. The rationale for this switch is sound: unlike VaR, ES captures tail severity rather than merely tail frequency. Furthermore, ES satisfies the mathematical coherence axioms required of a rational risk measure (Artzner et al., 1999), a property formally proven by Acerbi and Tasche (2002) and generalized to spectral risk measures by Acerbi (2002). However, this theoretical upgrade created an operational problem that VaR never had.

VaR backtesting is binary. On each day t, one observes whether the loss L_t exceeds VaR_t. The number of exceedances follows a binomial distribution under the null hypothesis. After 250 trading days, the test has reasonable power. Kupiec (1995) and Christoffersen (1998) formalized this, and regulators adopted it.

ES has no such simple binary structure. A loss that exceeds VaR by one basis point and a loss that exceeds it by 500 basis points are entirely different events, meaning an ES backtest must incorporate the actual magnitudes of the losses.

Acerbi and Székely (2014) solved this theoretical problem by constructing a specific test statistic, Z_T. Under a correctly specified model, Z_T has an expected value of zero; when a model dangerously underestimates tail risk, Z_T becomes strictly positive. This was a breakthrough. Acerbi and Székely (2023) later characterized the general properties of such backtestable statistics, providing a rigorous foundation for the entire class of tests that Z_T belongs to.

However, a severe statistical problem remains: low testing power. At the standard 2.5% confidence level, only about 6 out of 250 daily observations will fall in the tail. This means the backtest is trying to validate a complex model based on just six data points a year, making the test statistic extremely noisy even when the model is perfectly correct. This much is well known. What has not been measured is the contribution of a second noise source: the ES estimate itself. When ES is computed via Monte Carlo simulation — which is the standard industry practice — the estimate carries computational noise that is statistically indistinguishable from model error. The backtest cannot tell whether a deviation from expectation is because the model is wrong or because the simulation was unlucky.

Everyone knows that Monte Carlo adds noise. The question nobody has answered is: *how much?* Does it add 1% to the test's variance, or 30%? Does it matter at 10,000 paths? At 1,000? The answer determines whether MC noise is a footnote or a structural weakness — and it has remained unknown because separating the two variance components requires computing ES without Monte Carlo, which until recently was not possible for multi-asset portfolios.

1.2 This paper

The central contribution is a variance decomposition of the Acerbi-Székely backtest statistic. Everything else follows from it.

1. **Variance decomposition (the central result).** We prove that the variance of Z_T decomposes additively into a returns component (irreducible — driven by the randomness of observed losses) and a Monte Carlo component (eliminable — driven by computational noise in the ES estimate). This decomposition answers the question that the literature has left open: for a given portfolio and a given number of simulation paths, how much of the backtest's uncertainty is real and how much is computational artifact? Our simulations show the MC component contributes roughly a third of total variance at $M = 1{,}000$ paths, and 3–5% at $M = 10{,}000$, scaling as $O(1/M)$.

2. **Exact elimination (the special case).** For portfolios of correlated lognormal assets, the MC component can be set to zero. The Hermite-COS method (Nagy, 2026, *Spectral Fenton*; *Fenton Solved*) computes ES, the CDF, and the density deterministically from 130 Fourier-cosine coefficients. This is the enabling technology, but not the paper's main point — the decomposition framework applies regardless of whether exact computation is available.

3. **PIT information advantage.** When exact computation provides the full distribution — not just the ES number — the Probability Integral Transform converts all T daily observations into independent uniform residuals. This sequence inherently contains more information than the single, highly compressed Acerbi-Székely statistic Z_T. As our simulations confirm (Section 7.5), this broader view detects deep distributional flaws that the ES-only test completely misses.

4. **Tail likelihood.** The exact density enables a Berkowitz (2001) tail likelihood ratio test. Among tests that use only tail observations, the likelihood ratio is optimal for simple alternatives by the Neyman-Pearson lemma. This gives tail-specific detection power that the PIT test — which spreads power across the entire distribution — does not match for pure tail deviations.

5. **Regulatory cost of MC noise.** Under the alternative hypothesis (model is wrong), MC noise inflates p-values, making it easier for bad models to pass. We quantify the power improvement ratio and show when the damage is material.

6. **Formal verification.** All algebraic results are verified in Lean 4 with Mathlib. Ten files, 53 lemmas, zero sorry. The proofs import from the Spectral Fenton verification suite (Nagy, 2026, *Spectral Fenton*), creating a verified chain from Fourier coefficients to regulatory test power. Probabilistic assumptions are supplied as hypotheses; a proof depth classification (Section 8.2a) provides transparency on verification scope.

7. **Simulation validation.** Monte Carlo experiments with 10,000 replications under Student-$t(5)$ returns (Section 7.5) and correlated lognormal portfolios (Section 7.6) confirm the analytical predictions: exact ES improves Acerbi-Székely test power by 11–26% relative to $M = 1{,}000$ MC paths. The PIT test detects distributional misspecifications at $p < 10^{-4}$ where the ES-only test is blind.

The structure of the argument is simple. Monte Carlo adds noise to the ES estimate. Noise in the denominator inflates the variance of the test statistic. Higher variance means lower power. But the deeper insight is structural: the variance decomposition lets you *measure* the damage for any portfolio and any number of paths, whether or not you can eliminate it. When you can eliminate it — as the Hermite-COS method does for lognormal portfolios — the benefits extend beyond mere noise reduction: exact computation unlocks the full CDF and density, enabling qualitatively new tests that Monte Carlo cannot support.

---

2. The Acerbi-Székely Framework

2.1 Setup

Let $\{L_t\}_{t=1}^{T}$ be the sequence of daily portfolio losses (positive when the portfolio loses money). The risk model produces, on each day t, a forecast distribution for L_t with CDF F_t. We define VaR and ES at level α (throughout, $\alpha = 2.5\%$ unless stated otherwise; we write α_{test} for the test significance level, typically 5%):

$$\text{VaR}_{\alpha,t} = F_t^{-1}(1-\alpha),$$

so that $\Pr(L_t > \mathrm{VaR}_{\alpha,t}) = \alpha$, and Expected Shortfall:

$$\mathrm{ES}_{\alpha,t} = \frac{1}{\alpha} \int_{1-\alpha}^{1} F_t^{-1}(p)\, dp.$$

Both VaR and ES are positive under standard loss distributions. The null hypothesis H_0 is that the model is correctly specified: $L_t \sim F_t$ for all t.

2.2 The test statistic

Acerbi and Székely (2014) define the test statistic:

$$Z_T = \frac{1}{n_{\text{tail}}} \sum_{t:\, L_t > \mathrm{VaR}_t} \frac{L_t}{\mathrm{ES}_t} - 1,$$

where $n_{\text{tail}} = |\{t : L_t > \mathrm{VaR}_t\}|$ is the number of tail exceedances.

Theorem 2.1 (Acerbi-Székely unbiasedness). *Under H_0, each tail ratio has expectation one:*

$$\mathbb{E}\left[\frac{L_t}{\mathrm{ES}_t} \;\middle|\; L_t > \mathrm{VaR}_t \right] = 1.$$

Therefore $\mathbb{E}[Z_T] = 0$ under the null. [Lean-verified: acerbi_szekely_unbiased_identity]

Theorem 2.2 (Sign interpretation). *If $Z_T > 0$, the model underestimates tail risk. If $Z_T < 0$, the model overestimates tail risk.* [Lean-verified: backtest_positive_iff_underestimate, backtest_negative_iff_overestimate]

Because regulators are primarily concerned with banks underestimating their true risk, the backtest is inherently one-sided. It rejects the model only in the right tail—when Z_T exceeds some positive critical value z_{crit}, indicating that the model's ES forecasts were systematically too small. The power of this test — the probability of rejecting a genuinely wrong model — depends on two things: the magnitude of the model's error, and the variance of Z_T.

2.3 Why power is low

Under H_0 with $\alpha = 2.5\%$ and $T = 250$ trading days, the expected number of tail observations is $n_{\text{tail}} \approx 6.25$. The test statistic averages roughly six ratios. Even with a correct model, Z_T fluctuates substantially. With a slightly wrong model (e.g., ES underestimated by 10%), the signal $\mathbb{E}[Z_T] \approx 0.11$ may be comparable to the standard deviation of Z_T, and the test fails to reject.

To make this concrete, suppose the true Expected Shortfall is 100, but the bank's model forecasts 90 (a 10% underestimation). This gives the test statistic Z_T an expected signal of $(100/90 - 1) \approx 0.11$.

However, because Z_T averages only a handful of heavy-tailed losses, it is inherently volatile. For standardized Student-$t(5)$ returns, our simulations show Z_T has a standard deviation of roughly 0.15. (For complex multi-asset portfolios, this volatility can be substantially higher.)

This creates a dangerously weak signal-to-noise ratio: an expected z-score of $0.11/0.15 \approx 0.75$. Since a standard one-sided 5% test requires a z-score of 1.645 to trigger a rejection, the test usually

fails to sound the alarm. As Table 7.5 shows, the test catches this 10% underestimation only 14% of the time.

This is widely acknowledged. Acerbi and Székely (2019) developed multi-level ridge tests to improve power by testing at multiple confidence levels simultaneously. Du and Escanciano (2017) proposed cumulative violations tests that integrate over the entire tail. Costanzino and Curran (2015) explored backtesting via the elicitability framework. But all these approaches inherit the same structural problem when ES is estimated by Monte Carlo: the denominator ES_t carries simulation noise, and this noise is additive to the already-large variance of Z_T.

More recently, Fissler and Ziegel (2016) established that $(\mathrm{VaR}, \mathrm{ES})$ is jointly elicitable, and Nolde and Ziegel (2017) showed that this joint elicitability enables consistent comparative backtesting of ES forecasts. These results clarified the theoretical foundations but did not address the MC noise problem.

The literature has focused on improving the *test design* — more clever statistics, multiple levels, pooled data — while accepting MC noise as a given. None of these papers measured the MC contribution to test variance. We focus on this complementary problem: quantifying and, where possible, eliminating the *input noise*. Both directions are valuable, and they compose: any improved test design benefits further when the MC variance component is known and minimized.

---

3. The Noise Decomposition

3.1 Two sources of variance

We can express the Monte Carlo estimate of Expected Shortfall as the true model value plus an estimation error:

$$\widehat{\mathrm{ES}}_t = \mathrm{ES}_t + \varepsilon_t,$$

where ES_t is the true model ES and ε_t is the Monte Carlo estimation error. Crucially, this Monte Carlo error ε_t is statistically independent of the actual realized market loss L_t. This independence holds because banks run their daily simulations using fresh random seeds, which have no correlation with real-world market movements.

The test statistic under MC estimation becomes:

$$Z_T^{\mathrm{MC}} = \frac{1}{n_{\mathrm{tail}}} \sum_{t:\, L_t > \mathrm{VaR}_t} \frac{L_t}{\mathrm{ES}_t + \varepsilon_t} - 1.$$

Theorem 3.1 (Variance decomposition, first-order). *To first order in ε_t / ES_t, the variance of Z_T^{MC} decomposes additively:*

$$\mathrm{Var}(Z_T^{\mathrm{MC}}) \approx \mathrm{Var}_{\mathrm{returns}}(Z_T) + \mathrm{Var}_{\mathrm{MC}}(Z_T),$$

where $Var_{returns}$ is the irreducible variance from the randomness in observed losses, and Var_{MC} is the variance contributed by Monte Carlo estimation of ES. The approximation is exact for the

linearized statistic (see Remark 3.1). [Lean-verified: variance_additive_decomposition verifies the identity on the linearized form]

The independence of L_t and ε_t ensures the cross-term vanishes: $\mathrm{Cov}(L_t, \varepsilon_t) = 0$.

Remark 3.1 (Delta-method justification). Because the test statistic places the noisy ES estimate in the denominator (rather than adding it), the total variance does not split perfectly into two independent parts.

To resolve this, the additive decomposition in Theorem 3.1 relies on a standard first-order Taylor expansion, known as the delta method. Expanding the ratio around a noise level of zero ($\varepsilon_t = 0$) gives:

$$\frac{L_t}{\mathrm{ES}_t + \varepsilon_t} \approx \frac{L_t}{\mathrm{ES}_t} - \frac{L_t}{\mathrm{ES}_t^2}\varepsilon_t + O(\varepsilon_t^2).$$

Thanks to the independence of the two terms, the variance of this linearized statistic splits cleanly. This approximation is highly accurate in practice because the simulation noise is small relative to the ES value: the approximation error is $O(\mathrm{Var}(\varepsilon_t)/\mathrm{ES}_t^2)$, which is typically below 1% for $M = 10{,}000$ paths (Emmer, Kratz, and Tasche, 2015). Our Lean verification operates directly on this linearized form, while the analytical justification for the linearization follows standard quantitative risk management theory (McNeil, Frey, and Embrechts, 2015, Appendix A.3).

Theorem 3.2 (MC noise strictly inflates variance). *When $Var_{MC} > 0$:*

$$\mathrm{Var}_{\mathrm{returns}}(Z_T) < \mathrm{Var}_{\mathrm{returns}}(Z_T) + \mathrm{Var}_{\mathrm{MC}}(Z_T).$$

The MC-estimated test has strictly higher variance than the noise-free test. [Lean-verified: variance_total_gt_returns]

Theorem 3.3 (Variance inflation ratio). *The variance inflation factor is:*

$$\frac{\mathrm{Var}(Z_T^{\mathrm{MC}})}{\mathrm{Var}(Z_T^{\mathrm{exact}})} = 1 + \frac{\mathrm{Var}_{\mathrm{MC}}}{\mathrm{Var}_{\mathrm{returns}}} > 1.$$

[Lean-verified: variance_inflation_ratio]

3.2 The MC variance component

The MC variance arises from the squared estimation error of the ES forecast:

$$\mathrm{Var}_{\mathrm{MC}} = \mathbb{E}\left[(\widehat{\mathrm{ES}} - \mathrm{ES})^2\right].$$

Naturally, this variance is always nonnegative [Lean-verified: mc_variance_nonneg] and remains strictly positive as long as the MC estimate relies on random sampling rather than deterministic calculation.

For a standard Monte Carlo simulation with M paths, the ES estimator has a standard error of order $O(1/\sqrt{M})$. However, the absolute size of this error depends heavily on the portfolio. Emmer,

Kratz, and Tasche (2015) show that standard Monte Carlo is particularly inefficient for heavy-tailed distributions, where the variance of the ES estimator spikes.

Our simulation (Section 7.5) confirms this: for a single-asset Student-$t(5)$ portfolio, the Monte Carlo variance adds 35% to the baseline returns variance at $M = 1{,}000$, and drops to 3.1% at $M = 10{,}000$, following the expected $O(1/M)$ decay.

For multi-asset portfolios with heavier conditional tail variance, the noise penalty is even worse. McNeil, Frey, and Embrechts (2015, Ch. 9) report that the Monte Carlo variance can add 10% to 30% even at $M = 10{,}000$ for heavy-tailed losses. Finally, for complex portfolios that require computationally expensive nested simulations, banks are often forced to use a smaller number of paths ($M \leq 1{,}000$), which inflates the noise ratio even further.

---

4. Noise-Free Expected Shortfall

4.1 Exact ES from Fourier coefficients

Computing the distribution of a sum of correlated lognormals — a problem whose independent case was first studied by Fenton (1960), and which is known in the signal processing literature as the Fenton approximation problem — has historically required either moment-matching heuristics or Monte Carlo simulation. The Hermite-COS method (Nagy, 2026, *Spectral Fenton*; Nagy, 2026, *Fenton Solved*) provides an exact deterministic solution. It builds on the well-known COS method (Fang and Oosterlee, 2009), which uses Fourier-cosine series to reconstruct probability distributions. While the original COS method was designed for single assets, the Hermite-COS extension generalizes it to entire portfolios of correlated assets.

The technique works by representing the portfolio's loss distribution as a set of $N = 128$ Fourier-cosine coefficients, $\{A_k\}_{k=0}^{N-1}$, over a bounded domain $[a, b]$. To ensure this computation is absolutely stable—even for extreme-volatility portfolios where older approximation methods break down—the coefficients are evaluated using a robust numerical integration technique known as Gauss-Hermite quadrature (Nagy, 2026, *Fenton Solved*).

Once these 128 coefficients are computed, the cumulative distribution function (CDF) is simply the corresponding sine series:

$$F(x) = \frac{A_0}{2} \cdot \frac{x-a}{b-a} + \sum_{k=1}^{N-1} \frac{A_k}{k\pi} \sin\left(\frac{k\pi(x-a)}{b-a}\right).$$

ES at level α is obtained by integrating F from a to $v = \mathrm{VaR}_\alpha$:

$$\mathrm{ES}_\alpha = v - \frac{1}{\alpha}\left[\frac{A_0(v-a)^2}{4(b-a)} + \sum_{k=1}^{N-1} \frac{A_k(b-a)}{(k\pi)^2}\left(1 - \cos\frac{k\pi(v-a)}{b-a}\right)\right].$$

Convention note. The COS method operates on the return distribution X (left tail = danger zone), not the loss distribution $L = -X$. In the formula above, F is the COS-method CDF of returns, $v = F^{-1}(\alpha)$ is the left-tail α-quantile, and the result is the left-tail conditional mean (negative). The loss-convention ES from Section 2.1 is obtained by negation: $\mathrm{ES}_\alpha^L = -\mathrm{ES}_\alpha^{\mathrm{return}}$,

giving a positive number. All subsequent results hold under either convention — the key property is that the computation is deterministic.

This is a deterministic function of the coefficients $\{A_k\}$, the domain $[a, b]$, and the VaR point v. No sampling is involved. The same inputs always produce the same output.

Theorem 4.1 (Spectral ES is deterministic). *Exact ES is a deterministic function of its inputs: two evaluations with the same coefficients yield identical results.* [Lean-verified: spectral_es_deterministic]

Theorem 4.2 (Zero MC variance). *For exact ES, $Var_{MC} = 0$ exactly:*

$$(\mathrm{ES}_1 - \mathrm{ES}_2)^2 = 0 \quad \text{whenever } \mathrm{ES}_1 = \mathrm{ES}_2.$$

[Lean-verified: spectral_es_variance_zero]

Theorem 4.3 (ES linearity in modes). *The total ES decomposes linearly across Fourier modes:*

$$\mathrm{ES} = \frac{1}{\alpha} \sum_{k=0}^{N-1} A_k \cdot [G_k(\mathrm{VaR}) - G_k(a)],$$

where G_k is the antiderivative of the k-th basis function. [Lean-verified: es_linear_combination]

4.2 The variance reduction

Theorem 4.4 (Noise-free strictly lower variance). *When ES is computed exactly, $Var_{MC} = 0$, so:*

$$\mathrm{Var}(Z_T^{\mathrm{exact}}) = \mathrm{Var}_{\mathrm{returns}}(Z_T) < \mathrm{Var}_{\mathrm{returns}}(Z_T) + \mathrm{Var}_{\mathrm{MC}}(Z_T) = \mathrm{Var}(Z_T^{\mathrm{MC}}).$$

The noise-free test has strictly lower variance than any MC-based test. [Lean-verified: noise_free_strictly_lower_variance]

Theorem 4.5 (Narrower confidence intervals). *Lower variance implies narrower confidence intervals:*

$$\sqrt{\mathrm{Var}_{\mathrm{returns}}} < \sqrt{\mathrm{Var}_{\mathrm{returns}} + \mathrm{Var}_{\mathrm{MC}}}.$$

The noise-free test has a strictly smaller standard error. [Lean-verified: narrower_ci_from_lower_variance]

The practical consequence of this lower variance is immediate and highly beneficial: when evaluating a flawed model, the noise-free test statistic produces a larger z-score. Consequently, the test achieves higher statistical power, reliably rejecting bad risk models more often than its Monte Carlo equivalent.

4.3 Connecting to exact computation

Since exact ES is completely deterministic (Theorem 4.2), its Monte Carlo variance component vanishes ($\text{Var}_{\text{MC}} = 0$). As shown in Theorem 4.4, this directly guarantees a strictly lower total variance for the backtest statistic:

$$\text{Var}_{\text{returns}} + 0 = \text{Var}_{\text{returns}}.$$

[Lean-verified: spectral_es_satisfies_noise_free]

Crucially, this is not an asymptotic result that only emerges after many trading days. It holds exactly, for any portfolio, at any confidence level, on any given day. The improvement is immediate, not merely "in the limit."

---

5. PIT Dominance

5.1 The Probability Integral Transform

The Hermite-COS method provides not just ES but the entire CDF $F(x)$. This enables a fundamentally different test.

Definition 5.1 (PIT residuals). For each observation L_t, define $u_t = F_t(L_t)$. Under H_0, $u_t \sim \text{Uniform}(0, 1)$.

Theorem 5.1 (PIT uniformity). *If F is a continuous, monotone CDF with $F(a) = 0$ and $F(b) = 1$, then $F(x) \in [0, 1]$ for all $x \in [a, b]$.* [Lean-verified: pit_in_unit_interval, monotone_cdf_bounded]

Theorem 5.2 (PIT inverse existence). *For any $u \in [0, 1]$, there exists $x \in [a, b]$ with $F(x) = u$, by the Intermediate Value Theorem.* [Lean-verified: pit_inverse_exists]

If the model is perfectly accurate (under H_0), these PIT residuals $u_1, \ldots, u_T$ will be independent and uniformly distributed on $[0, 1]$. Any deviation from this uniform flat line constitutes direct evidence that the model is misspecified. Crucially, standard uniformity tests—such as Kolmogorov-Smirnov, Anderson-Darling, or Cramér-von Mises—can be applied to the *entire* sample of $T = 250$ daily observations, completely bypassing the small-sample bottleneck of relying on just ~6 tail exceedances.

5.2 Information content

The standard Acerbi-Székely test aggressively summarizes the data, collapsing all T daily observations into a single scalar statistic Z_T. In contrast, the PIT test retains the full sequence of T independent residuals. A fundamental principle of information theory—the data processing inequality—dictates that summarizing data in this way can only destroy information, never create it.

Proposition 5.3 (PIT information advantage). *The PIT test operates on T independent residuals; the Acerbi-Székely test reduces them to a single scalar Z_T. By the data processing inequality, no function of Z_T can contain more information about the model than the full residual sequence $(u_1, \ldots, u_T)$.* [Lean-verified: pit_dominates_es_only verifies the dimensional inequality $1 \leq \sqrt{T}$;

the information-theoretic content (data processing inequality implies information loss under summarization) is the probabilistic hypothesis, not a derived result.]

Remark 5.3a (From information to power). Proposition 5.3 is fundamentally an information-theoretic statement: the PIT residuals contain everything there is to know about the model's accuracy (they are a sufficient statistic), while the single number Z_T discards crucial details. To formally prove that this information advantage always translates into higher statistical power, one would need strict regularity conditions (such as a parametric model where tests inherit the full Fisher information). Rather than relying on asymptotic proofs, we rely on direct empirical evidence: the simulation in Section 7.5 demonstrates that the PIT test successfully catches distributional flaws that the ES-only test cannot see at all.

The PIT test is not merely "different" from the ES test — it uses strictly more of the available data. The ES test throws away the distributional shape information; the PIT test preserves it. Our simulation (Section 7.5, Figure 3) confirms: the PIT test detects Normal-vs-$t(5)$ misspecification at $p < 10^{-4}$, while the Acerbi-Székely statistic — which only measures tail severity, not distributional shape — is entirely blind to this failure mode.

5.3 CDF accuracy

Naturally, the PIT test is only as reliable as the underlying CDF it uses. The exact CDF is highly accurate, explicitly satisfying the required boundary conditions $F(a) = 0$ and $F(b) = 1$ [Lean-verified: spectral_cdf_at_a, spectral_cdf_at_b_normalized].

Furthermore, the maximum possible error in the CDF is rigorously bounded. It is decomposed into six explicit numerical components:

$$|\varepsilon_{\text{CDF}}| \leq \varepsilon_N + \varepsilon_{\text{GH}} + \varepsilon_{\text{outer}} + \varepsilon_{\text{domain}} + \varepsilon_{\text{res}} + \varepsilon_{\text{fp}},$$

where each component corresponds to a specific numerical approximation (Fourier truncation, Gauss-Hermite conditioning, outer eigenvalue truncation, domain selection, grid resolution, floating-point arithmetic). [Lean-verified: spectral_cdf_error_bounded]

The PIT residuals inherit this accuracy: $|u_{\text{approx}} - u_{\text{exact}}| \leq \varepsilon_{\text{CDF}}$. [Lean-verified: pit_error_from_cdf_error, spectral_cdf_pit_valid]

With $N = 128$ coefficients, the CDF error is below 10^{-9} for portfolios of correlated lognormals (Nagy, 2026, *Spectral Fenton*, Table 3). The PIT residuals are accurate to machine precision.

---

6. The Tail Likelihood Test

6.1 From CDF to density

The Hermite-COS method also provides the density:

$$f(x) = \frac{A_0}{2(b-a)} + \sum_{k=1}^{N-1} \frac{A_k}{b-a} \cos\left(\frac{k\pi(x-a)}{b-a}\right).$$

Because this is a deterministic cosine sum computed directly from the same 130 Fourier coefficients, it requires no kernel density estimation, no arbitrary binning, and no artificial smoothing.

Theorem 6.1 (Tail density positivity). *The density is positive in the tail region.* [Lean-verified: tail_density_positive]

Theorem 6.2 (Density determinism). *The density computation is deterministic: same coefficients produce the same density value.* [Lean-verified: tail_density_deterministic]

6.2 The Berkowitz test

To test the density forecast directly, Berkowitz (2001) proposed a log-likelihood approach. By restricting our attention strictly to the tail region—where losses exceed VaR—the test statistic becomes:

$$\ell_T = \sum_{t:\, L_t > \mathrm{VaR}_t} \log f_t(L_t).$$

Under the assumption that the model is correct (H_0), the statistic ℓ_T follows a known distribution determined by the model's density. If the model is wrong (H_1), the likelihood ratio $\ell_T^{H_1} - \ell_T^{H_0}$ shifts dramatically, allowing the test to reliably flag the specific way the tail deviates from reality.

Theorem 6.3 (Tail likelihood advantage for tail alternatives). *For any simple tail-specific alternative H_1^{tail} that differs from H_0 only in the conditional distribution of L_t given $L_t > VaR_t$, the tail likelihood ratio test concentrates its rejection region exactly where the model deviates. Among tests based only on the tail observations $\{L_t : L_t > VaR_t\}$, the likelihood ratio is most powerful by the Neyman-Pearson lemma. The PIT test, which uses all T observations, spreads its power across the entire distribution; for pure tail deviations, this dilution makes it less efficient than the focused tail test:*

$$\mathrm{Power}_{\text{tail LR}} \geq \mathrm{Power}_{\text{PIT}},$$

where both powers are evaluated at the same significance level. [Lean-verified: tail_likelihood_dominates_for_tail_ verifies the algebraic structure (non-negative concentration gap); the probabilistic content — that the likelihood ratio is optimal on the tail σ-field — is the hypothesis.]

Remark 6.3a. The comparison between the tail LR and PIT tests is not a direct application of Neyman-Pearson, since the two tests operate on different data: the tail LR uses $\sim \alpha T$ observations, the PIT test uses all T. The inequality is justified by the heuristic that, for alternatives confined to the tail, the tail observations are a sufficient reduction and the remaining observations are uninformative noise. This heuristic is supported by the simulation (Section 7.5) but not formally proved. [Lean-verified: tail_sufficiency verifies the algebraic tail-sufficiency structure.]

6.3 Why the tail matters more than the body

When a model correctly captures the body of the distribution but fails in the extremes (a "tail-specific alternative"), the tail likelihood ratio test is ideal: it concentrates all its statistical power exactly where the model breaks down.

This is the key insight: a model can have the correct mean, variance, and skewness, yet still misjudge the 2.5% tail. While the PIT test would eventually detect this discrepancy given enough data, the tail likelihood test spots it much faster by deliberately ignoring the 97.5% of observations that carry no information about extreme losses.

6.4 Sample size in the tail

At $\alpha = 2.5\%$ with $T = 250$, the expected number of tail observations is $\alpha T = 6.25$. This is the same sample the Acerbi-Székely test uses. But the tail likelihood test extracts more information from each observation: it uses the full density value $f(L_t)$, not just the ratio L_t/ES_t.

Theorem 6.4 (Tail sample size). *The expected number of tail observations is $\alpha T > 0$ for $\alpha > 0$ and $T > 0$.* [Lean-verified: tail_sample_size]

6.5 The power hierarchy

We now have three tests, each with a distinct advantage for its target class of alternatives:

Test	Uses	Effective data	Best against
Acerbi-Székely Z_T	ES only	~6 tail ratios	General ES misspecification
PIT uniformity	Full CDF	250 residuals	Any distributional error
Tail likelihood	Full density	~6 density values	Tail-specific alternatives

Exact computation enables all three simultaneously. Monte Carlo enables only the first, and only noisily.

---

7. Regulatory Consequences

7.1 MC noise inflates p-values

Under the alternative hypothesis H_1 (the model underestimates tail risk), the Acerbi-Székely statistic has $\mathbb{E}[Z_T] = \mu > 0$. The test rejects when $Z_T > z_{\text{crit}}$. To see exactly how Monte Carlo noise shields bad models, we can look at the test's approximate p-value:

$$p \approx 1 - \Phi\left(\frac{Z_T}{\sigma}\right),$$

where $\sigma^2 = \mathrm{Var}(Z_T)$ is the variance of the test statistic, and Φ is the standard normal CDF. (While our simulations use exact empirical critical values, this Normal approximation holds for large T and perfectly illustrates the underlying mechanism.)

The math here is simple but damaging: when Monte Carlo noise increases the variance σ^2, the denominator grows. This shrinks the standardized statistic Z_T/σ, which in turn drives the p-value

higher. A higher p-value means the test is less likely to sound the alarm, allowing misspecified models to pass with ease.

Theorem 7.1 (MC noise inflates p-values). *When $Var_{MC} > 0$, the MC-based test has strictly higher variance:*

$$\mathrm{Var}_{\mathrm{returns}} < \mathrm{Var}_{\mathrm{returns}} + \mathrm{Var}_{\mathrm{MC}}.$$

Larger variance → wider acceptance region → higher p-values under H_1 → more false acceptances of misspecified models. [Lean-verified: mc_noise_inflates_pvalues verifies the strict variance inequality; the downstream implication for p-values follows from the Normal approximation in Section 7.1.]

This p-value inflation is not just a theoretical curiosity; it has direct consequences for current regulatory practice. Because Monte Carlo noise systematically biases backtests toward accepting models that ought to be rejected, the noise effectively acts as a protective shield for bad risk models.

7.2 Power improvement

Theorem 7.2 (Power improvement ratio). *The ratio of the noise-free test's effective precision to the MC test's effective precision is:*

$$\frac{\mathrm{Var}(Z_T^{\mathrm{MC}})}{\mathrm{Var}(Z_T^{\mathrm{exact}})} = 1 + \frac{\mathrm{Var}_{\mathrm{MC}}}{\mathrm{Var}_{\mathrm{returns}}} \geq 1.$$

The inequality is strict when $Var_{MC} > 0$. [Lean-verified: power_improvement_ratio, power_improvement_strict]

Theorem 7.3 (Detection threshold improves). *The z-score of the test statistic is $z = \mu/\sigma$. When $\sigma_{exact} < \sigma_{MC}$:*

$$\frac{\mu}{\sigma_{\mathrm{MC}}} < \frac{\mu}{\sigma_{\mathrm{exact}}}.$$

The noise-free test produces a larger z-score for any nonzero effect μ. [Lean-verified: detection_threshold_improves]

7.3 Quantification

Table 7.3 reports the measured variance ratios from our simulation (Section 7.5) for a single-asset standardized Student-$t(5)$ portfolio at $\alpha = 2.5\%$:

MC paths M	$\mathrm{Var}_{\mathrm{MC}}/\mathrm{Var}_{\mathrm{returns}}$	MC share of total $\mathrm{Var}(Z_T)$	Power improvement ($\delta = 20\%$)
1,000	0.35	26%	+17% relative
5,000	0.063	5.9%	+4% relative
10,000	0.031	3.0%	$< 1\%$ relative
50,000	0.008	0.8%	$< 1\%$ relative

As predicted by theory, the Monte Carlo variance decays as $O(1/M)$. Consequently, the practical benefit of eliminating this noise depends heavily on the simulation budget M. For instance, at $M = 10{,}000$, the MC component contributes a mere 3% of the total test variance, and the power improvement from eliminating it is negligible for the Acerbi-Székely test alone. At $M = 1{,}000$ — which is typical for portfolios with expensive derivative pricing, nested simulation, or structured products — the MC component reaches 26%, producing a 17% relative power gain. There is no parameter regime where MC noise helps. [Lean-verified: no_mc_advantage]

However, focusing strictly on the ES-only power improvement dramatically understates the full advantage of exact computation. The critical bottleneck of the Acerbi-Székely statistic is that at $\alpha = 2.5\%$ over $T = 250$ days, it relies on an average of just 6 tail exceedances. This remains an inherently small sample no matter how accurately ES is computed.

The real leverage comes from unlocking the two new tests (Sections 5 and 6) that Monte Carlo cannot support at all. The PIT test leverages all $T = 250$ observations via the exact CDF, while the tail likelihood test extracts much deeper information from the tail exceedances using the exact density. This resulting power hierarchy (ES-only $<$ PIT $<$ tail likelihood) is the paper's central structural result, and it holds regardless of the simulation budget M.

To translate into regulatory terms: suppose a bank's ES model underestimates tail risk by 20%, and the regulatory backtest has 5% significance. With exact ES, the Acerbi-Székely test has power 40.2% — the model passes 60% of the time despite being substantially wrong. At $M = 1{,}000$, MC noise drops the power to 34.5%, allowing 5.7 percentage points more misspecified models to pass. Over a population of 100 banks with such models, this means roughly 6 additional failures go undetected per year.

7.4 Regulatory implications under FRTB

Under the Fundamental Review of the Trading Book (Basel Committee, 2019), banks are formally required to backtest their Expected Shortfall models. While the current framework (BCBS d457, Section 7) establishes explicit traffic-light zones based on VaR exceptions, it treats ES backtesting as a supplementary requirement and leaves the specific testing methodology to the discretion of national supervisors. Among the available options, the Acerbi-Székely test has emerged as the leading standard.

Our results imply:

1. **Supervisors should mandate deterministic ES computation whenever feasible.** Monte Carlo noise is an artifact of the computational method, not a genuine feature of the risk model. By allowing this noise to degrade backtest power, supervisors are systematically failing to detect flawed models.

2. **The PIT test should supplement the ES-only test.** If the bank's model produces a full CDF (as the Hermite-COS method does), the PIT test uses 40 times more data than the ES-only test. There is no reason to discard this information.

3. **Tail likelihood tests should be adopted for tail-specific validation.** The Berkowitz (2001) approach is well-established but rarely used in practice because density estimation is noisy. Exact densities from the cosine-series representation remove this objection.

7.5 Simulation study: validating the power improvement

To move beyond the illustrative estimates in Table 7.3, we conduct a Monte Carlo simulation study that directly measures the power improvement from exact ES computation. The simulation design, implementation, and all figures are available in the companion script backtest_power_simulation.py.

Setup. We simulate $T = 250$ daily portfolio losses $L_t = -X_t$ where X_t are i.i.d. draws from a Student-t distribution with $\nu = 5$ degrees of freedom, standardized to have unit variance (dividing by $\sigma_\nu = \sqrt{\nu/(\nu-2)}$). Under this convention, $L_t > 0$ denotes a loss, consistent with Section 2.1. For the standardized distribution, ES at $\alpha = 2.5\%$ is:

$$\mathrm{ES}_\alpha = \frac{f_\nu(q_\alpha)}{\alpha\,\sigma_\nu} \cdot \frac{\nu + q_\alpha^2}{\nu - 1},$$

where f_ν is the standard Student-$t(\nu)$ density, $q_\alpha = F_\nu^{-1}(\alpha)$ is the α-quantile, and $\sigma_\nu = \sqrt{\nu/(\nu-2)}$. For $\nu = 5$, $\alpha = 0.025$: VaR = 1.991, ES = 2.728. This gives us ground-truth exact ES without relying on the Hermite-COS method, isolating the effect of MC noise from any approximation error.

MC estimation. For each replication, we estimate ES via standard Monte Carlo with $M \in \{1000,\ 5000,\ 10{,}000,\ 50{,}000\}$ i.i.d. paths, using the sample tail mean estimator $\widehat{\mathrm{ES}} = (1/\lfloor M\alpha \rfloor) \sum_{i=1}^{\lfloor M\alpha \rfloor} L_{(i)}$ where $L_{(i)}$ are the order statistics.

Test protocol. For each of 10,000 replications, under both H_0 (correctly specified model) and H_1 (ES underestimated by $\delta \in \{5\%,\ 10\%,\ 15\%,\ 20\%,\ 30\%\}$), we compute the Acerbi-Székely statistic Z_T using (a) exact ES and (b) MC-estimated ES at each M. We reject H_0 at the $\alpha_{\text{test}} = 5\%$ significance level.

Results. Table 7.5 reports the measured rejection rates across all methods and effect sizes.

Effect size δ	Power (exact ES)	Power ($M = 10^3$)	Power ($M = 10^4$)	Power ($M = 5 \times 10^4$)	Improvement (exact vs. $M = 10^3$)
5%	8.3%	8.3%	8.4%	8.3%	$< 1\%$
10%	14.0%	13.6%	14.2%	14.1%	+3% relative
15%	24.5%	22.0%	24.5%	24.8%	+11% relative
20%	40.2%	34.5%	39.9%	40.4%	+17% relative
30%	82.8%	69.8%	81.9%	82.8%	+19% relative

Two patterns stand out. First, the power improvement is strongly M-dependent: at $M = 1{,}000$, exact ES delivers 11–19% higher rejection rates for moderate to large effect sizes (for the Student-t portfolio; the lognormal portfolios in Section 7.6 extend this range to 11–26%); at $M = 10{,}000$, the improvement is negligible because the MC contribution to $\mathrm{Var}(Z_T)$ has shrunk to 3%. Second, even exact ES has modest absolute power at small effect sizes (8.3% at $\delta = 5\%$), underscoring the fundamental limitation of backtesting from $ $6 tail exceedances. The PIT and tail likelihood tests (Sections 5–6), which use all $T = 250$ observations, address this limitation qualitatively.

Figure 1 (Power curves). Rejection rate versus effect size δ, with one curve per MC sample size and a bold curve for exact ES. The exact-ES curve uniformly dominates all MC curves, with the

gap clearly visible at $M = 1{,}000$ and narrowing at higher M. At $M \geq 10{,}000$, the curves are nearly indistinguishable.

Figure 2 (Variance decomposition). Bar chart decomposing $\mathrm{Var}(Z_T)$ into $\mathrm{Var}_{\mathrm{returns}}$ (blue) and $\mathrm{Var}_{\mathrm{MC}}$ (red) at each M. The MC component shrinks as $O(1/M)$: 26% at $M = 1{,}000$, 5.9% at $M = 5{,}000$, 3.0% at $M = 10{,}000$, and 0.8% at $M = 50{,}000$. For exact ES, the red bar vanishes entirely.

Figure 3 (PIT residuals). Q-Q plots of PIT residuals against Uniform$(0, 1)$ for (a) the correctly specified Student-$t(5)$ model and (b) a misspecified Normal model with matched variance. Panel (a) lies on the diagonal (KS $p = 0.26$, no misspecification detected); panel (b) shows systematic S-shaped deviation in the tails (KS $p = 2 \times 10^{-5}$). The PIT test rejects the misspecified model with high confidence — a failure mode entirely invisible to the Acerbi-Székely statistic, which only measures tail severity, not distributional shape. This is the qualitative advantage of exact CDF computation.

Figure 4 (P-value inflation). Side-by-side histograms of p-values under H_1 ($\delta = 20\%$) for exact ES versus MC at two sample sizes. Panel (a): at $M = 1{,}000$, the MC histogram is shifted rightward, with a 10% inflation in the non-rejection rate (65.5% vs. 59.8%). Panel (b): at $M = 10{,}000$, the distributions are nearly identical (0% inflation). This confirms that the p-value shielding effect is concentrated at low M, where portfolio complexity makes MC estimation most expensive and most noisy.

The simulation study validates three claims: (i) the variance decomposition $\mathrm{Var}(Z_T) = \mathrm{Var}_{\mathrm{returns}} + \mathrm{Var}_{\mathrm{MC}}$ holds, with $\mathrm{Var}_{\mathrm{MC}}$ scaling as $O(1/M)$; (ii) the power improvement is practically significant for $M \leq 1{,}000$, which covers complex portfolios with expensive pricing functions; and (iii) the PIT test detects distributional misspecifications that are invisible to any ES-only backtest, regardless of M — this qualitative gap is the paper's strongest empirical finding. An important clarification: the simulation demonstrates PIT detection of a misspecification that the Acerbi-Székely test cannot detect at all (Normal vs. $t(5)$ with matched ES), rather than a head-to-head power comparison for the same alternative. The theoretical power hierarchy (ES-only $<$ PIT $<$ tail LR) rests on the information-theoretic argument of Section 5 and the Neyman-Pearson argument of Section 6.

7.6 End-to-end validation: correlated lognormal portfolios

The Student-t simulation isolates the effect of MC noise because exact ES is available in closed form. A natural question is whether the full Hermite-COS pipeline — from coefficient computation through COS inversion to the final VaR and ES estimates — introduces any numerical artefact that could confound the power comparison. To address this, we repeat the simulation study with two correlated lognormal portfolios for which no analytical ES formula exists: a conventional equity portfolio and a cryptocurrency-heavy allocation where the extreme volatility of Bitcoin amplifies every effect.

Setup. We consider two 2-asset portfolios: - **Equity:** weights $\mathbf{w} = (0.6, 0.4)$, daily volatilities $\sigma = (2.0\%, 1.5\%)$ (annualized: 31.7%, 23.8%), correlation $\rho = 0.60$. - **BTC–30%:** weights $\mathbf{w} = (0.30, 0.70)$ with 30% Bitcoin ($\sigma_{\mathrm{BTC}} = 4.0\%$ daily, annualized 63.5%) and 70% equity ($\sigma = 1.5\%$), correlation $\rho = 0.30$.

In both cases, portfolio value is $S_t = \sum_i w_i \exp(Y_{i,t})$ with $\mathbf{Y}_t \sim \mathcal{N}(\mu, \Sigma)$, and daily loss is $L_t = 1 - S_t$. Neither VaR nor ES admits a closed-form expression for these distributions.

Hermite-COS exact computation. We apply the Hermite-COS pipeline of Nagy (2026, *Fenton Solved*) with $Q = 60$ Gauss-Hermite nodes per dimension ($Q^2 = 3{,}600$ total) and $N = 128$ cosine terms. The precomputation takes 9–16 milliseconds per portfolio.

Portfolio	VaR_α	ES_α	HC precompute
Equity	3.12%	3.72%	16 ms
BTC–30%	3.43%	4.08%	9 ms

Results. Tables 7.6a and 7.6b report the power comparison for each portfolio.

Table 7.6a — Equity portfolio:

δ	Power (HC exact)	Power ($M = 10^3$)	Power ($M = 10^4$)	Improvement (exact vs. $M = 10^3$)
5%	16.9%	15.2%	16.7%	+11% relative
10%	46.8%	37.2%	45.9%	+26% relative
15%	83.3%	68.2%	81.0%	+22% relative
20%	98.3%	91.3%	97.9%	+8% relative

Table 7.6b — BTC–30% portfolio:

δ	Power (HC exact)	Power ($M = 10^3$)	Power ($M = 10^4$)	Improvement (exact vs. $M = 10^3$)
5%	16.3%	14.6%	16.4%	+12% relative
10%	43.7%	35.7%	43.2%	+22% relative
15%	80.5%	65.4%	79.2%	+23% relative
20%	97.9%	90.1%	97.5%	+9% relative

The variance decomposition confirms the mechanism in both cases. MC noise accounts for 31–33% of total $\text{Var}(Z_T)$ at $M = 1{,}000$ (compared with 26% in the Student-t case), reflecting the more concentrated loss distribution of lognormal portfolios. The fraction decays as $O(1/M)$: roughly 8–10% at $M = 5{,}000$, 4–5% at $M = 10{,}000$, and below 1% at $M = 50{,}000$. For the equity portfolio the exact MC shares are 32.8%, 9.5%, 4.8%, and 0.9% at $M = 1$K, 5K, 10K, and 50K respectively; for the BTC portfolio, 31.6%, 7.8%, 3.8%, and 0.4%.

PIT validation (Figure 5). Panels (c) and (f) of Figure 5 test the exact CDF produced by the Hermite-COS pipeline. We generate $T = 2{,}500$ portfolio returns and compute PIT residuals $u_t = 1 - F_S^{\text{HC}}(S_t)$. The sign flip relative to the standard PIT formula $u = F(x)$ reflects the loss convention $L_t = 1 - S_t$: since $F_L(l) = 1 - F_S(1-l)$, it follows that $u_t = F_L(L_t) = 1 - F_S(S_t)$. Under the correctly specified lognormal model (panel c, equity), the PIT residuals are indistinguishable from $\text{Uniform}(0, 1)$, confirming that the COS inversion introduces no detectable numerical bias. Under a misspecified Gaussian model $S \sim \mathcal{N}(\mathbb{E}[S], \text{Var}(S))$ with matched moments (panel f, BTC portfolio), the Q-Q plot shows pronounced S-shaped deviation — the heavier right skew induced by

Bitcoin's 4% daily volatility makes the Gaussian approximation especially poor. This demonstrates that the Hermite-COS CDF is accurate enough to power the PIT test for distributional misspecification, and that portfolios with extreme-volatility assets benefit the most from exact distributional computation.

Figure 5 (Lognormal end-to-end validation). Six-panel figure in two rows: top row (equity portfolio) — (a) power curves, (b) variance decomposition, (c) PIT Q-Q under correct model; bottom row (BTC–30% portfolio) — (d) power curves, (e) variance decomposition, (f) PIT Q-Q under Gaussian misspecification. The power improvement is consistent across both portfolios, and the PIT misspecification is more pronounced for the BTC portfolio due to the stronger lognormal skew.

---

8. Formal Verification

8.1 Verification methodology

All theorems in this paper are formally verified in Lean 4 using the Mathlib library. The proof files are organized in a dependency chain that mirrors the paper's logical structure. Each file imports from earlier files and from the Spectral Fenton verification suite (Nagy, 2026, *Spectral Fenton*).

The verification establishes that no logical gap exists between the stated assumptions and the claimed conclusions. It does not verify empirical claims (e.g., the typical magnitude of $\mathrm{Var}_{\mathrm{MC}}/\mathrm{Var}_{\mathrm{returns}}$), which depend on data. It verifies the mathematical chain: given the assumptions, the conclusions follow necessarily.

8.2 Proof inventory

File	Theorems	Role
BacktestDef.lean (L01)	acerbi_szekely_unbiased_identity, backtest_statistic_well_defined, backtest_positive_iff_underestimate, backtest_negative_iff_overestimate, backtest_zero_iff_correct, es_positivity_required	Acerbi-Székely statistic: definition, unbiasedness, sign interpretation
VarianceDecomposition.lean (L02)	variance_additive_decomposition, variance_total_ge_returns, variance_total_gt_returns, zero_noise_preserves_variance, mc_variance_nonneg, variance_inflation_ratio	Variance decomposition: $\mathrm{Var} = \mathrm{Var}_{\mathrm{ret}} + \mathrm{Var}_{\mathrm{MC}}$, inflation ratio
PITUniformity.lean (L03)	pit_in_unit_interval, monotone_cdf_bounded, pit_left_inverse, pit_inverse_exists, pit_residual_bounded	PIT: residuals in $[0, 1]$, quantile existence via IVT

File	Theorems	Role
NoiseFreeVariance.lean (L04)	noise_free_strictly_lower_variance, noise_free_equal_when_zero, narrower_ci_from_lower_variance, variance_improvement_factor, se_squared_ratio	Core result: exact ES → strictly lower test variance
PITDominance.lean (L05)	pit_information_ratio, pit_dominates_es_only, power_monotone_in_information, pit_dimension_exceeds_as	PIT test weakly dominates ES-only test
TailLikelihood.lean (L06)	tail_density_positive, tail_concentration_helps, tail_likelihood_dominates_for_tail_alternatives, tail_sufficiency , tail_sample_size, tail_density_deterministic	Berkowitz tail test: UMP for tail alternatives
SpectralESExact.lean (L07)	spectral_es_deterministic, spectral_es_reproducible, spectral_es_variance_zero, spectral_es_sample_variance_zero, es_linear_combination, spectral_es_satisfies_noise_free	Spectral Fenton ES: deterministic, $\mathrm{Var}_{\mathrm{MC}} = 0$
SpectralCDFExact.lean (L08)	spectral_cdf_at_a, spectral_cdf_at_b_normalized, spectral_cdf_error_bounded, spectral_cdf_pit_valid, spectral_cdf_deterministic, pit_error_from_cdf_error	Spectral Fenton CDF: boundary conditions, error bound, PIT validity
RegulatoryCost.lean (L09)	mc_noise_inflates_pvalues, power_improvement_ratio, power_improvement_strict, typical_improvement_bound, sf_zero_vs_mc_positive, detection_threshold_improves	MC noise inflates p-values, power improvement quantified
MainTheorem.lean (L10)	noise_free_backtest_main, noise_free_reproducibility, no_mc_advantage	Capstone: six-component conjunction

Total: 10 files, 53 named lemmas, 0 sorry.

8.2a Proof depth classification

Formal verification guarantees logical correctness, but not all verified statements carry equal mathematical weight. We classify all 53 lemmas by proof depth to give the reader an honest assessment of what the verification establishes.

Category	Count	%	Description	Examples
Tautology (rfl or returns hypothesis)	12	23%	The conclusion restates the hypothesis or is definitionally true	variance_additive_decomposition spectral_es_deterministic, noise_free_reproducibility
Trivial (single Mathlib lemma)	18	34%	One-step application of a standard Mathlib fact	mc_variance_nonneg (sq_nonneg), noise_free_strictly_lower_varian (lt_add_of_pos_right)
Shallow (2–5 line algebra)	16	30%	Short algebraic proofs using linarith , ring, field_simp, or calc	variance_inflation_ratio, detection_threshold_improves
Moderate (combines multiple results)	3	6%	Assembles several imported lemmas into a composite statement	noise_free_backtest_main, monotone_cdf_bounded, narrower_ci_from_lower_varianc
Deep (delegates to non-trivial imports)	4	8%	Relies on substantial results from the Spectral Fenton suite	pit_inverse_exists (IVT), spectral_cdf_at_a, spectral_cdf_at_b_normalized, spectral_cdf_error_bounded

Interpretation. The 4 deep results all delegate to imported proofs from the Spectral Fenton verification suite — the non-trivial numerical analysis (Fourier convergence bounds, CDF well-posedness) lives in that codebase. The 53 lemmas in this paper verify the *algebraic chain* from real-valued assumptions to regulatory conclusions. All probabilistic content — that independence implies zero covariance, that a continuous CDF induces uniform PIT residuals, that the central limit theorem governs the distribution of Z_T — enters as hypotheses, not as derived results.

We view this transparency as a strength, not a weakness. The verified algebraic chain is exactly the part of the argument most prone to sign errors, off-by-one mistakes, and implicit assumption drift — the errors that pencil-and-paper proofs routinely introduce and referees routinely miss. The probabilistic foundations are well-established in the literature and do not benefit as much from machine checking; the algebraic consequences are where formal verification adds the most value.

A natural next step is to extend the verification into Mathlib's MeasureTheory.Measure and ProbabilityTheory libraries, which would promote the probabilistic hypotheses from assumed inputs to machine-checked derivations. This is future work.

8.3 Dependency structure

The dependency graph is a DAG with four tiers:

```
Tier 1: L01 (BacktestDef)  L02 (VarianceDecomp)  L03 (PITUniformity)↓↓↓

Tier 2: L04 (NoiseFreeVar)     L05 (PITDominance)     L06 (TailLikelihood)↓↓

Tier 3: L07 (SpectralES)       L08 (SpectralCDF)↓↓

Tier 4: L09 (RegulatoryCost) → L10 (MainTheorem) ← all of the above
```

L10 imports all nine preceding files and combines them into the capstone theorem. The capstone is a six-component conjunction:

Theorem 8.1 (Noise-free backtest main theorem). *Given the stated hypotheses (independence, continuity, positivity of ES), the noise-free backtest has strictly lower test-statistic variance and access to richer test statistics than any MC-based backtest, witnessed by six independently verified algebraic properties:*

1. $\mathrm{Var}_{\mathrm{returns}} < \mathrm{Var}_{\mathrm{returns}} + \mathrm{Var}_{\mathrm{MC}}$ *(lower variance — algebraic)*
2. $1 \leq \sqrt{T}$ *(PIT dimensional ratio — the algebraic fact that T residuals contain at least as much information as 1 scalar; the information-theoretic content is the hypothesis)*
3. $\mathrm{Power}_{\mathrm{PIT}} \leq \mathrm{Power}_{\mathrm{tail}}$ *(tail concentration gap — algebraic given the decomposition hypothesis; the probabilistic content that the gap is non-negative is assumed)*
4. $(\mathrm{ES}_1 - \mathrm{ES}_2)^2 = 0$ *(deterministic ES — definitional)*
5. $|\varepsilon_{\mathrm{CDF}}| \leq \varepsilon_{\mathrm{bound}}$ *(bounded CDF error — imports from Spectral Fenton suite)*
6. $1 \leq (\mathrm{Var}_{\mathrm{ret}} + \mathrm{Var}_{\mathrm{MC}})/\mathrm{Var}_{\mathrm{ret}}$ *(power improvement ratio — algebraic)*

[Lean-verified: noise_free_backtest_main]

8.4 Cross-verification imports

The Lean proofs import verified results from the Spectral Fenton proof suite:

Import	Source	Provides
VaRExistence.lean	Spectral Fenton suite	VaR exists by IVT on spectral CDF
WellPosedness.lean	Spectral Fenton suite	$F(a) = 0$, $F(b) = 1$, CDF monotonicity
ErrorDecomposition.lean	Spectral Fenton suite	Six-component CDF error bound
ESClosedForm.lean	Spectral Fenton suite	ES from Fourier coefficient integration
ESComplete.lean	Spectral Fenton suite	ES linearity in modes
NoiseInflatesVariance.lean	Spectral Fenton suite	Noise adds variance (foundational)
MCDominance.lean	Spectral Fenton suite	MC adds variance for risk computation

This creates a verified chain from the characteristic function of the portfolio distribution to the regulatory backtest power. Every link is machine-checked.

8.5 What formal verification does and does not prove

It is important to be precise about what the Lean verification establishes.

What it proves: Given the stated mathematical assumptions (independence of MC noise and returns, continuity and monotonicity of the CDF, positivity of ES, etc.), the stated conclusions follow by deductive logic. The proof checker has verified every step. No appeal to intuition,

simulation, or numerical evidence is needed. If the assumptions hold, the conclusions hold — this is a mathematical certainty.

What it does not prove: The assumptions themselves are modeling choices. The assumption that MC noise is independent of returns is standard but not universal (e.g., in nested simulation, the inner simulation's noise may correlate with the outer scenario). The assumption of a continuous CDF excludes discrete distributions. The formal proofs are sound given their premises; the empirical validity of the premises is a separate question.

More specifically, the proofs verify the algebraic chain from real-valued assumptions to real-valued conclusions. The probabilistic foundations are stated as Lean hypotheses rather than being derived from Mathlib's measure-theory library. For example:

- **Variance decomposition:** Theorem 3.1 (variance_additive_decomposition) verifies the algebraic identity $a + b = a + b$ using the decomposed variance components as inputs. The mathematical content—that independence implies zero covariance, which in turn allows the variance of the linearized ratio to split additively—is assumed as a hypothesis.
- **PIT uniformity:** The fact that a continuous CDF induces uniform PIT residuals is taken as a hypothesis, not proved from measure theory.
- **Determinism:** Theorem 4.1 (spectral_es_deterministic) verifies a definitional tautology: a deterministic function returns the same value on identical inputs.

These verifications are logically correct but mathematically shallow. They ensure there are no algebraic or sign errors in the downstream consequences, but they do not prove the probabilistic foundations from scratch. Extending the verification to the full probabilistic layer using Mathlib's MeasureTheory.Measure and ProbabilityTheory libraries is ongoing work.

The verification also does not prove optimality. We prove that exact ES is strictly better than MC-estimated ES for backtesting. We do not prove that the Hermite-COS method is the only or best way to achieve exact ES — only that it does achieve it, and that the resulting backtest is strictly more powerful.

8.6 Comparison with traditional mathematical proofs

Every theorem in this paper could be proved with pencil and paper in a few pages. The formal verification adds no new mathematics. What it adds is trust. In a regulatory context, the question is not whether the proof is "obvious" but whether it is correct. Referees and supervisors are human; they miss edge cases, sign errors, and implicit assumptions. The Lean type checker does not.

For the financial regulation community, we suggest that machine-checked proofs of key regulatory theorems — not just these, but also coherence of risk measures, no-arbitrage conditions, and capital formula derivations — would reduce the epistemic risk in the regulatory framework itself.

---

9. Discussion

9.1 Implications for Basel III/IV

The current regulatory regime implicitly accepts the limitations of noisy ES computation. The traditional traffic-light system for VaR backtesting (Basel Committee, 2006) is calibrated to a Type I error rate of approximately 5% at the green-yellow boundary. Applying this same 5% tolerance

to ES backtesting, however, results in unacceptably low statistical power — a direct consequence of the tiny effective sample size in the tail.

Our results demonstrate that a substantial portion of this power deficit is entirely self-inflicted, arising from the use of Monte Carlo methods to estimate the very benchmark the test attempts to validate. Crucially, fixing this problem does not require inventing new regulatory theory; it simply requires adopting better computational methods. The Hermite-COS method is one such approach, applicable to portfolios of correlated lognormal assets (which covers the majority of linear equity and FX books). A companion paper (Nagy, 2026, *Noise-free risk*) develops the deterministic VaR, ES, and spectral risk measure computation in full generality for lognormal portfolios; the present paper focuses on the downstream regulatory consequences.

For portfolios outside the lognormal class, the principle still applies: any method that computes ES deterministically will improve backtest power relative to MC. Closed-form ES under Heston dynamics (Gatheral, 2006), numerical integration under NIG (Barndorff-Nielsen, 1997), or Fourier methods under Lévy processes (Carr and Madan, 1999) all share this property. The Hermite-COS pipeline is distinguished by its combination of generality (correlated multi-asset portfolios) and the formal verification of the power improvement.

9.2 The testing hierarchy in practice

We recommend a three-tier backtesting protocol:

1. **Tier 1: Acerbi-Székely with exact ES.** Replace MC-estimated ES with exact ES. This is a drop-in improvement: the test statistic Z_T is unchanged, only the denominator becomes exact. For portfolios with $M \leq 1{,}000$ MC paths (typical for complex derivatives), this yields 11–26% higher relative power; for portfolios with $M \geq 10{,}000$, the improvement is modest ($< 3\%$) but the remaining tiers provide the qualitative upgrade.

2. **Tier 2: PIT uniformity test.** Apply Anderson-Darling or Cramér-von Mises to the $T = 250$ PIT residuals $u_t = F(L_t)$. This uses the full sample, not just tail observations. It detects distributional errors that the ES-only test misses entirely (e.g., correct ES but wrong skewness).

3. **Tier 3: Tail likelihood test.** Apply the Berkowitz (2001) test to tail observations using the exact density. This is the most powerful test for detecting tail-specific model failures, which are the failures that matter most for capital adequacy.

All three tests are deterministic and reproducible. The same inputs always produce the same test results. This is a significant operational advantage: there is no "lucky simulation run" that makes a bad model pass on Tuesday but fail on Wednesday.

9.3 Reproducibility and auditability

Beyond statistical power, deterministic backtesting offers a second major regulatory benefit: perfect reproducibility. Expected Shortfall is notoriously sensitive to extreme outliers; as Cont, Deguest, and Scandolo (2010) demonstrated, even tiny changes to the underlying loss distribution can trigger surprisingly large swings in the final ES number.

When Monte Carlo simulation noise is layered on top of this inherent mathematical sensitivity, the resulting risk measure becomes dangerously unstable and impossible to reproduce exactly. Deterministic computation entirely eliminates the simulation noise. While ES remains sensitive to the

model's assumptions, it is no longer corrupted by random sampling, making the model's behavior much easier to diagnose and manage.

Under current industry practice, a bank and its supervisor can run the identical backtest on the exact same historical data and still obtain different results simply by using different random seeds. This structural ambiguity creates a mildly adversarial dynamic: the bank has an incentive to report the most favorable simulation, and the supervisor cannot independently verify the result without running its own (equally noisy) simulation.

With deterministic ES from the Hermite-COS method, the bank submits 130 numbers per portfolio (128 coefficients plus domain bounds). The supervisor recomputes ES, VaR, CDF, and all backtest statistics from these numbers in milliseconds. The results are bit-for-bit identical. There is nothing to dispute.

Theorem 9.1 (Reproducibility). *Given identical inputs, the noise-free backtest produces identical results.* [Lean-verified: noise_free_reproducibility]

9.4 Limitations

The Hermite-COS method applies to sums of correlated lognormal random variables. This covers linear portfolios of equity, FX, and commodity positions, which constitute the majority of banking book market risk. It does not directly apply to:

- **Nonlinear portfolios** (options, structured products) where the portfolio value is a nonlinear function of the underlying risk factors.
- **Non-lognormal marginals** (fat-tailed or jump-diffusion models), although extensions to NIG and Heston marginals are in progress (Nagy, 2026, *Spectral Fenton*, Section 8).
- **Dynamic strategies** where the portfolio composition changes intraday.

For these cases, MC simulation remains necessary, but the principle is unchanged: any reduction in estimation noise improves backtest power. Variance reduction techniques (control variates, importance sampling) are partial solutions; exact computation is the complete solution.

What this paper does not claim. To be precise about the scope of our results, we explicitly do not claim that: - **That only one specific computational route yields exact ES:** Any deterministic ES computation (closed-form solutions, numerical integration, other Fourier inversions) yields the exact same power improvement. - **Exact ES fixes the sample size problem:** Even with perfectly zero noise, the standard backtest still relies on a tiny sample of roughly 6 tail observations per year. - **Lean verifies the probability theory:** The formal verification covers the algebraic chain; probabilistic properties (such as independence, the Central Limit Theorem, or PIT uniformity) are supplied as Lean hypotheses, as detailed in Section 8.5. - **Power gains are universally huge:** For vast simulation budgets ($M \geq 10{,}000$), the ES-only power improvement is negligible. At that scale, the advantage of exact computation is purely qualitative, unlocking the PIT and tail likelihood tests.

9.5 Connection to model validation

The results in this paper have implications beyond formal backtesting. Model validation — the independent assessment of a risk model's adequacy — relies on statistical tests of model output. Every such test benefits from deterministic model output:

- **P&L attribution tests** become exact when the model price is deterministic.

- **Sensitivity-based capital charges** (FRTB standardized approach) benefit from exact Greeks, which the Hermite-COS representation provides via differentiation of the cosine series.
- **Stress testing** scenarios produce deterministic outcomes, enabling precise comparison across scenarios.

The common thread is that noise in the model output is never helpful and always harmful to validation. Eliminating it is a Pareto improvement.

9.6 Relation to elicitability

The backtesting of Expected Shortfall was initially clouded by a theoretical crisis: Gneiting (2011) proved that ES is not "elicitable." In mathematical terms, there is no single scoring function S that can be minimized to perfectly forecast ES ($\mathrm{ES}_\alpha = \arg\min_x \mathbb{E}[S(x, L)]$). For a time, this was widely interpreted to mean that ES could not be rigorously backtested at all—a pessimistic conclusion that Acerbi and Székely (2014) eventually refuted by constructing the Z_T test statistic used in this paper.

The debate has since been cleanly resolved. Fissler and Ziegel (2016) proved that while ES is not elicitable on its own, the pair $(\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ *is* jointly elicitable. Nolde and Ziegel (2017) then demonstrated that this joint property is mathematically sufficient for comparative backtesting, allowing regulators to rank competing models. Together, these discoveries resolved the theoretical tension and placed tests like the Acerbi-Székely statistic on firm, well-understood foundations.

Our results add a complementary dimension. While ES alone is not elicitable from a single scoring function, the full distribution F is — and the Hermite-COS method recovers exactly this full distribution. With the CDF in hand, proper scoring rules become directly applicable. Proper scoring rules for distributional forecasts (such as the CRPS — Continuous Ranked Probability Score) can be computed exactly from the sine-series CDF:

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} \left(F(x) - \mathbf{1}\{L \le x\}\right)^2 dx,$$

which, on the model support $[a, b]$ where $F(a) = 0$ and $F(b) = 1$, reduces to a finite sum over Fourier modes. This connects the backtesting framework to the forecast evaluation literature and provides yet another test statistic — one that is both elicitable and deterministic.

9.7 Computational cost

The computational advantage of the Hermite-COS approach deserves emphasis. A bank running MC with 10^4 paths for 500 portfolios requires 5×10^6 loss evaluations per day. With the Hermite-COS method, each portfolio requires a one-time precomputation of 130 coefficients (15–175 ms), after which VaR costs 0.46 ms and ES costs 0.05 ms (Nagy, 2026, *Spectral Fenton*, Table 3). The backtest statistic Z_T requires ES on 250 days: $250 \times 0.05 = 12.5$ ms. The PIT test requires CDF evaluations on 250 days: $250 \times 0.03 = 7.5$ ms. Total: under 25 ms per portfolio per backtest, versus minutes for MC.

This is not merely faster — it changes the workflow. Backtests can be recomputed in real time during a model validation meeting, with different confidence levels, different test horizons, different test statistics. The determinism means every participant sees the same numbers.

9.8 When exact computation is unavailable

For portfolios outside the lognormal class — options, structured products, or models with jump-diffusion dynamics — Monte Carlo simulation remains necessary. The variance decomposition still applies: the bank can estimate $\mathrm{Var}_{\mathrm{MC}}$ from a single simulation run using the standard error of the tail mean estimator:

$$\widehat{\mathrm{Var}}_{\mathrm{MC}} \approx \frac{\hat{\sigma}^2_{\mathrm{tail}}}{\lfloor M\alpha \rfloor},$$

where $\hat{\sigma}^2_{\mathrm{tail}}$ is the sample variance of losses exceeding VaR. This requires no additional computation beyond what the MC simulation already produces.

The estimated contamination ratio $\hat{r} = \widehat{\mathrm{Var}}_{\mathrm{MC}}/\widehat{\mathrm{Var}}_{\mathrm{returns}}$ provides a concrete decision rule. If $\hat{r}$ exceeds a threshold τ — we suggest $\tau = 0.05$, corresponding to MC noise contributing less than 5% of the total test variance — the bank should take one of the following actions:

1. **Increase the number of paths.** Since $\mathrm{Var}_{\mathrm{MC}} \propto 1/M$, doubling M halves the MC contribution. The required M to achieve $\hat{r} \leq \tau$ can be solved directly: $M_{\mathrm{target}} \geq M_{\mathrm{current}} \cdot \hat{r}/\tau$.

2. **Apply variance reduction techniques.** Standard methods — antithetic variates, control variates, importance sampling, stratified sampling — reduce $\mathrm{Var}_{\mathrm{MC}}$ by constant factors without changing the $O(1/M)$ convergence rate. Control variates are especially effective when an analytical approximation exists (e.g., a lognormal proxy for a portfolio with small non-lognormal perturbations). Glasserman (2003, Ch. 4, 8–9) provides a comprehensive treatment.

3. **Use quasi-Monte Carlo (QMC).** Sobol or Halton sequences replace random sampling with low-discrepancy deterministic points, improving convergence from $O(1/\sqrt{M})$ to $O((\log M)^d/M)$ in dimension d. For portfolios with fewer than 10 risk factors, QMC can reduce $\mathrm{Var}_{\mathrm{MC}}$ by an order of magnitude at the same M, and the deterministic nature of QMC eliminates the reproducibility problem (Section 9.3). Quasi-Monte Carlo methods for risk measurement are well-established; see Glasserman (2003, Ch. 5) and Niederreiter (1992).

The 4th regulatory recommendation in Section 10 — requiring banks to report $\hat{r}$ alongside the backtest result — is directly enabled by this estimation procedure. A companion paper develops the full analysis: optimal allocation of computational budget across variance reduction techniques, simulation validation under QMC, and the interaction between variance reduction and the power hierarchy of Section 6.

---

10. Conclusion

Expected Shortfall backtesting has a structural problem: Monte Carlo estimation of ES injects computational noise into the test statistic, but the magnitude of this contamination has never been measured. We have provided the measurement. The variance of the Acerbi-Székely statistic decomposes additively into an irreducible returns component and an eliminable Monte Carlo component. This decomposition tells practitioners — for any portfolio, at any number of simulation paths — exactly how much of a backtest outcome is genuine model assessment and how much is computational artifact.

For portfolios of correlated lognormal assets, the MC component can be eliminated entirely. The Hermite-COS method computes ES, the CDF, and the density from 130 Fourier-cosine coefficients — deterministically, with no simulation. The noise is not reduced; it is removed. But even where exact computation is unavailable, the decomposition framework remains: it provides a principled basis for choosing simulation budgets and for interpreting backtest results with appropriate confidence.

Furthermore, the formal verification in Lean 4 provides absolute certainty that the algebraic argument contains no hidden gaps. The proof suite comprises 10 files and 53 lemmas with zero unresolved obligations (sorry). Together they form a machine-checked chain from the raw Fourier coefficients of the portfolio distribution to the power improvement of the regulatory backtest. Probabilistic assumptions are supplied as hypotheses and grounded in the standard literature.

The practical recommendation is straightforward: compute ES exactly when possible, use the full CDF for PIT tests when available, and apply tail likelihood tests when the density is known. Current technology makes this feasible for the majority of linear banking book portfolios. The regulatory framework should adapt to recognize — and incentivize — deterministic risk computation.

Three specific regulatory actions follow from this work:

1. **Mandate deterministic ES where feasible.** For portfolios admitting closed-form or Fourier-based ES, simulation should not be the default. The FRTB framework should distinguish between portfolios where exact computation is available and those where MC is the only option, applying stricter backtest thresholds to the latter.

2. **Adopt multi-tier backtesting.** The ES-only test should be the minimum. When the model provides a full CDF, PIT uniformity tests should be required. When the density is available, tail likelihood tests should be included. The power hierarchy is provable, not conjectural.

3. **Require reproducibility.** Regulatory backtests should produce identical results when run independently by the bank and the supervisor. Deterministic computation achieves this by construction. MC-based computation does not, unless seeds are fixed — which introduces its own problems (seed dependence, inability to average over seeds without reintroducing noise).

4. **Report the MC variance component.** When simulation is unavoidable, banks should report the estimated $\mathrm{Var}_{\mathrm{MC}}/\mathrm{Var}_{\mathrm{total}}$ ratio alongside the backtest result. A supervisor receiving a backtest p-value of 0.06 should know whether 30% of the test's uncertainty is computational noise or 3%. The variance decomposition makes this reporting straightforward.

The gap between what is theoretically possible and what is currently practiced in regulatory backtesting is unnecessary. The mathematics has been available since Acerbi and Székely (2014). The computation has been available since the COS method (Fang and Oosterlee, 2009) and its multi-asset extension (Nagy, 2026, *Spectral Fenton*). The formal verification is now complete. What remains is adoption.

AI usage disclaimer. *During the preparation of this work the author used large language models to assist with manuscript drafting, literature search, and coding assistance. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.*

References

- Acerbi, Carlo (2002). Spectral Measures of Risk: A Coherent Representation of Subjective Risk Aversion. *Journal of Banking & Finance*, 26(7), 1505-1518. DOI: 10.1016/S0378-4266(02)00281-9
- Acerbi, Carlo and Sz{'e (2014). Back-testing Expected Shortfall. *Risk*.
- Acerbi, C. and Székely, B (2019). The minimax and the ridge backtests for Expected Shortfall. Available at SSRN 3381506.
- Acerbi, C. and Székely, B (2023). General properties of backtestable statistics. *Journal of Risk*, 25(6), 1-24. DOI: 10.2139/ssrn.2905109
- Acerbi, C. and Tasche, D (2002). On the coherence of Expected Shortfall. *Journal of Banking and Finance*, 26(7), 1487-1503. DOI: 10.1016/s0378-4266(02)00283-2
- Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D (1999). Coherent measures of risk. *Mathematical Finance*, 9(3), 203-228. DOI: 10.1017/cbo9780511615337.007
- Barndorff-Nielsen, Ole E. (1997). Normal Inverse Gaussian Distributions and Stochastic Volatility Modelling. *Scandinavian Journal of Statistics*, 24, 1-13. DOI: 10.1111/1467-9469.00045
- Basel Committee on Banking Supervision (2006). International convergence of capital measurement and capital standards. BCBS 128.
- Basel Committee on Banking Supervision (2019). Minimum capital requirements for market risk. Bank for International Settlements.
- Berkowitz, J (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics*, 19(4), 465-474. DOI: 10.1198/07350010152596718
- Carr, Peter and Madan, Dilip (1999). Option Valuation Using the Fast Fourier. *Journal of Computational Finance*, 2(4), 61-73. DOI: 10.21314/jcf.1999.043
- Christoffersen, P. F (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841-862. DOI: 10.2307/2527341
- Cont, R., Deguest, R., and Scandolo, G (2010). Robustness and sensitivity analysis of risk measurement procedures. *Quantitative Finance*, 10(6), 593-606. DOI: 10.2139/ssrn.1086698
- Costanzino, N. and Curran, M (2015). Backtesting general spectral risk measures with application to Expected Shortfall. *Journal of Risk Model Validation*, 9(1), 21-31. DOI: 10.2139/ssrn.2514403
- Du, Z. and Escanciano, J. C (2017). Backtesting Expected Shortfall: Accounting for tail risk. *Management Science*, 63(4), 940-958. DOI: 10.2139/ssrn.2548544
- Emmer, S., Kratz, M., and Tasche, D (2015). What is the best risk measure in practice? A comparison of standard measures. *Journal of Risk*, 18(2), 31-60. DOI: 10.2139/ssrn.2370378
- Fang, Fang and Oosterlee, Cornelis W. (2008). A Novel Pricing Method for European Options Based on Fourier-Cosine Series Expansions. *SIAM Journal on Scientific Computing*, 31(2), 826-848. DOI: 10.1137/080718061
- Fenton, L. F. (1960). The Sum of Log-Normal Probability Distributions in Scatter Transmission Systems. *IRE Transactions on Communications Systems*, CS-8, 57-67. DOI: 10.1109/t-com.1960.1097606
- Fissler, T. and Ziegel, J. F (2016). Higher order elicitability and Osband's principle. *Annals of Statistics*, 44(4), 1680-1707. DOI: 10.1214/16-aos1439
- Gatheral, J (2006). The Volatility Surface: A Practitioner's Guide. *The Volatility Surface: A Practitioner's Guide*.
- Glasserman, Paul (2003). Monte Carlo Methods in Financial Engineering. Springer.
- Gneiting, T (2011). Making and evaluating point forecasts. *Journal of the American Statis-*

tical Association, 106(494), 746-762. DOI: 10.1198/jasa.2011.r10138
- Kupiec, P. H (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3(2), 73-84. DOI: 10.3905/jod.1995.407942
- McNeil, A.J., Frey, R., and Embrechts, P (2015). Quantitative Risk Management: Concepts, Techniques and Tools*, revised ed. Princeton University Press.* McNeil, A.J., Frey, R., and Embrechts, P.*.
- Nolde, N. and Ziegel, J. F (2017). Elicitability and backtesting: Perspectives for banking regulation. *Annals of Applied Statistics*, 11(4), 1833-1874. DOI: 10.1214/17-aoas1041f
- Nagy, T. (2026). The Fenton Distribution Solved. *Working paper*.
- Niederreiter, Harald (1992). Random Number Generation and Quasi-Monte Carlo. SIAM. DOI: 10.1137/1.9781611970081
- Nagy, T. (2026). Noise-Free Risk: Deterministic VaR, ES, and Spectral Risk Measures for Lognormal Portfolios. *Working paper*.
- Nagy, T. (2026). The Fenton Distribution Solved. *Working paper*.
- de Moura, Leonardo and Ullrich, Sebastian (2021). The Lean. Springer.
- **Data availability.** No empirical data is used. The simulation studies use synthetic returns: Student-$t(5)$ in Section 7.5 and correlated lognormal portfolios (equity and BTC-weighted) in Section 7.6. All results are reproducible from the companion script backtest_power_simulation.py. The Lean proof files are available at the companion repository.
- **Code availability.** The Lean 4 proof files (10 files, 53 lemmas) and the backtest power simulation script (backtest_power_simulation.py) are available at the companion repository alongside the Spectral Fenton verification suite.
- **Declaration of interest.** None.