Dollar-Correct, Time-Wrong: Twelve Backtest Bugs That Survive Reconciliation

Charlie Yan*

August 2026

Abstract

A backtest whose profit-and-loss reconciles to the cent can still be wrong by a factor of two to twelve in risk-adjusted terms—or can be measuring an exposure the strategy never claimed. The defect in such cases lies not in the dollars but in their *timing*, their *coverage*, the *identity* of the instrument marked, or the *attribution* of the P&L to the hypothesis under test. The standard defenses—reconciliation audits, multiple- testing corrections, deflated Sharpe ratios, walk-forward protocols, leakage screens—do not detect these failures, because every one of them takes the daily P&L series as given. We document twelve bug classes encountered in a single systematic options program, each with its mechanism, the distortion it produced on live research, a runtime-checkable invariant that prevents it, and a fully synthetic reproduction shipped as a test. Two classes warrant emphasis. First, exit-day lumping—booking a multi-day trade's P&L on its exit date—inflates the Sharpe ratio *only* when concurrent positions share shocks: in simulation the exit-day-to-MTM Sharpe ratio climbs from 0.64 (purely idiosyncratic P&L, where lumping *deflates*) to a plateau of 2.1–2.2 once shocks are substantially shared, bracketing the 2.2–2.4× inflation observed in production. This explains why the bug survives toy testing and why it is worst precisely for the correlated, stress-exposed books where accuracy matters most. Second, *attribution failure*—dollars, timing and coverage all correct while the P&L originates in an exposure the strategy was not testing—survives every check in the other eleven classes; in our worked example a tape passing all of them earned its money from a risk premium and a hedge path, and a side-randomized control scored the constant-side benchmark at five times the signal's Sharpe. We propose per-leg attribution plus a side-randomized negative control as the minimum standard of evidence for any signal tape, and we release the invariants as running code so that adoption costs an import rather than a culture change.

1 Introduction

The quantitative research literature has developed an impressive apparatus for asking whether a backtested result is *statistically* credible. We can price the multiplicity of a search (Harvey and Liu, 2015; Harvey et al., 2016), deflate a Sharpe ratio for the number of trials that produced it (Bailey et al., 2014; Bailey and López de Prado, 2014), estimate the probability that an in-sample optimum is

*Draft v0.3 (journal-length), August 2026. Companion code: `opentape`, an MIT-licensed library shipping every invariant below as a callable and every bug class as a synthetic failing-then-passing test; no market data is required to reproduce any exhibit in this paper. Every performance figure is labeled with its accounting basis. Figures marked INVALID are quoted *as* invalidated examples, with the corrected number alongside; none is a result of this paper. Code, pre-registration documents and a data-availability statement: https://github.com/charlieyanhx/backtest-bug-taxonomy.

overfit (Bailey et al., 2017), embed the whole exercise in a protocol (Arnott et al., 2019), and screen for the train-test contamination that has produced a reproducibility crisis in machine-learning-based science (Kapoor and Narayanan, 2023; Kapoor et al., 2024). This apparatus shares a working premise: that the per-strategy performance number being scrutinized was computed correctly in the first place. The premise is not universal— Arnott et al. (2019)'s protocol explicitly covers data and samples alongside economic motivation, multiple testing, cross-validation, model dynamics, complexity and research culture—but that category addresses data *quality* and sample construction (survivorship, outliers, selection) rather than the accounting mechanics by which a position file becomes a daily P&L series. It is those mechanics that concern us.

This paper is about the prior question. In building and auditing tapes for a systematic options program—a documented research record spanning several months of intensive work over a five-and-a-half-year data sample—the most damaging errors we encountered were not statistical. They were accounting errors: the dollars were right—reconciliations passed, totals tied to the cent—but the dollars were placed on the wrong days, attributed to the wrong instrument, computed over an incomplete calendar, or generated by an exposure the strategy never claimed to test. Each produced a headline number that survived every statistical defense we could apply, because those defenses operate on the daily series that the accounting error had already corrupted.

We document twelve such classes (Table 1), each with its mechanism, its measured distortion on real research, a build-time invariant, and a synthetic reproduction that lets a reader watch the bug appear and then be caught without access to our data or anyone else's.

We are explicit about what is and is not new here. Most of the twelve are engineering hygiene: known in principle to careful practitioners, absent from the published record chiefly because nobody writes down what they fixed last quarter. Their value in this paper is documentation and enforcement—measured magnitudes, and code that fails a build—not discovery. Two of the twelve are different, and they are the paper's actual claims.

Exit-day lumping is conditional, not universal (§4). Booking each trade's P&L on its exit date, rather than marking it daily, leaves totals unchanged and is widely known to distort risk statistics. What appears not to be known is *when*: under i.i.d. daily P&L, lumping does not inflate the Sharpe ratio at all—it deflates it. The inflation requires that concurrent positions share shocks, because mark-to-market booking concentrates a common shock onto one catastrophic day while exit-day booking disperses it across the staggered exit dates of every trade that lived through it. Figure 2 traces the inflation ratio from 0.64 to its 2.1–2.2 plateau as commonality rises. The practical implication is uncomfortable: the bug is invisible in the low-correlation test cases a developer is likely to construct, and maximal in the correlated, stress-exposed books where the risk statistic actually matters.

Attribution failure survives everything else (§6). A tape can pass coverage assertions, reconciliation, sign checks, force-close discipline and calendar-time exit location—and still be measuring a premium rather than a signal. Any structure with a nonzero risk-premium or beta exposure harvests it whatever the overlay signal says; a roughly side-symmetric signal then *dilutes* the harvest instead of adding to it, and the resulting Sharpe ratio is a diluted premium wearing the signal's name. The detection is cheap and, in our experience, decisive: decompose the P&L by leg, and re-run the identical tape with randomized and constant sides. In the worked example of §6 the constant-side control scored 1.75 against the signal's 0.34.

Our contribution is deliberately unglamorous. It is a taxonomy, a set of assertions, and a library. We argue that this is the right shape for the problem: statistical defenses are applied by researchers who choose to apply them, whereas invariants run on every build. The twelve classes below were

each expensive to learn once; none should be expensive to learn again.

2 Related literature

Selection and multiplicity. Lo (2002) showed how sensitive the Sharpe ratio is to the statistical properties of the return series it summarizes, and in particular to serial correlation—an early instance of the general point that the *shape* of the daily series, not merely its mean and variance, drives the statistic. Sullivan et al. (1999) formalized data-snooping bias with the reality check; Harvey and Liu (2015); Harvey et al. (2016) brought multiple-testing discipline to the cross-section of factors; Bailey and López de Prado (2014); Bailey et al. (2014) introduced the deflated Sharpe ratio and, with Bailey et al. (2017), the probability of backtest overfitting. Arnott et al. (2019) synthesize this line into a practical protocol whose scope is broader than multiplicity alone, covering data and samples, cross-validation, model dynamics and research culture. The centre of gravity of this line, though, is *how many* strategies were tried and how the winner's statistic should be discounted; where it addresses data, it addresses sample integrity—survivorship, outliers, selection—rather than whether the daily series under analysis was assembled correctly from the positions that generated it.

Return misdating and smoothed risk statistics. Getmansky et al. (2004) show that reported returns on illiquid portfolios are a moving average of true returns, that this smoothing understates volatility and inflates the Sharpe ratio, and that positive serial correlation in reported returns is its signature; they supply a smoothing-adjusted Sharpe ratio as the correction. Their source is stale pricing on assets that do not trade. Class 1 below is the same bias arriving from a different door: the assets are perfectly liquid and continuously priced, and the smoothing is introduced by an *accounting convention*. We show in §4 that exit-day marking induces exactly the serial-correlation signature GLM identify, and induces it precisely where it inflates — which makes their diagnostic and their correction directly applicable to a setting where nobody thinks to look for illiquidity, because there is none.

Leakage and reproducibility. Kapoor and Narayanan (2023) document leakage as the dominant reproducibility failure in machine-learning-based science and propose model-info sheets; Kapoor et al. (2024) extends this to reporting standards. In finance specifically, recent work catalogues spurious predictability arising from data leakage and look-ahead in ML pipelines. Leakage is adjacent to our subject—both produce wrong numbers from clean intentions—but distinct: leakage is information crossing a boundary it should not; our classes are correct information placed at the wrong time, on the wrong instrument, or credited to the wrong cause.

Costs and implementability. A parallel literature shows that anomalies survive statistically and die economically: Novy-Marx and Velikov (2016) on transaction costs in the anomaly zoo, Frazzini et al. (2018) on trading costs at scale, Chordia et al. (2014) on capacity. McLean and Pontiff (2016) document how much published anomaly performance decays post-publication. Our class 5 (mid-fill fantasy) is the accounting counterpart of this literature's economic point: the three-line convention we propose (mid, friction-aware, full cross) makes the friction sensitivity of a result a reported statistic rather than a discovered disappointment.

Replication. Hou et al. (2020) re-examine a large body of published results and find substantial non-replication; Chen (2021) bounds how much of the anomaly literature p-hacking alone could explain. Our classes suggest one under-explored contributor to non-replication: not that the original authors searched too hard, but that the original tape mis-timed, mis-covered, or mis-attributed its own P&L in ways no reader could detect from a published table.

#	Class	Measured distortion	Invariant
1	Exit-day lumping	Sharpe ×2.2–2.4 (conditional; §4)	one mark per held day
2	Trajectory truncation	a "12.6" Sharpe, 100% artifact	calendar-time exit location
3	Boundary drops	24% of trades, non-randomly	tape count = entry-log count
4	Missing-leg deferral	1.99 → 1.37; tail-day sign inverts	re-mark quote-less held days
5	Mid-fill fantasy	"8.58" → −8.20 at real quotes	three lines; decide on full cross
6	Active-day annualization	up to ×8 on sparse books	full-calendar padding
7	Sign bug	net exceeds gross	assert line 3 ≤ Line 1
8	Dropped/retried candidates	revoked a live deployment	force-close, never drop
9	Wrong-instrument lookup	P&L sign flips (§5)	exact-identity joins
10	Same-snapshot execution	manufactures edge from pure noise	next-quote fills
11	Calendar-indexed differencing	deletes every Monday	trading-day grid throughout
12	**Attribution failure**	**signal 0.34 vs constant-side 1.75**	**leg decomposition + side control**

Table 1: The twelve classes. Distortions are from live research in one program and are illustrative magnitudes, not population estimates. All twelve have synthetic reproductions in the companion library.

Marking and valuation. The accounting literature on fair-value measurement and the practitioner literature on P&L attribution both treat the timing of recognition as first-order; the Basel-era P&L attribution tests for internal models are, in effect, an institutional version of the leg-decomposition we advocate in class 12. What is missing, and what this paper supplies, is the translation of that discipline into research-stage backtests, where no auditor is watching and the researcher's own reconciliation is the only gate.

3 Setting and evidence standard

All examples arise from one systematic SPY options program: hourly chain data spanning 2020–2026, short-premium and relative-value strategies, retail-to-mid-scale friction, holding periods of days to weeks, and heavy position overlap during volatility episodes. The program's standing evidence rules—each adopted after a specific failure—are: three fill lines (mid; friction-aware; full cross, meaning buy at ask and sell at bid on every leg both ways, which is the only go/no-go number); mark-to-market-daily marking padded to the full business calendar with zeros on inactive days; a coverage line reported with every result (rows in, rows out, per filter); pre-registered decision bars; and a test-count ledger.

We stress that none of the twelve classes is exotic. Each was found in code written by competent researchers who were reconciling their numbers. That is the point: reconciliation is a necessary check that is comprehensively insufficient.

4 Class 1: exit-day lumping, and when it actually bites

Mechanism. The whole trade's P&L is booked on its exit day; days on which the position was held and moving show zero.

What it cost us. A skew strategy quoted at 1.91 (exit-day basis, INVALID) re-marked to 0.81; a premium sleeve 4.16 → 1.72; an aggregate book 4.78 → ≈2.2. Inflation of 2.36×, 2.42× and 2.17× respectively—a 2.2–2.4× band—with totals unchanged in every case.

The conditional result. We simulate 120 overlapping ten-day trades on a 252-day calendar, with

each trade's daily P&L a convex combination $\rho \cdot (\text{common factor}) + (1-\rho) \cdot (\text{idiosyncratic})$, the common factor carrying five planted crash days. Figure 1 shows the canonical case ($\rho = 1$): identical cumulative totals, visibly different daily distributions—the exit-day series simply does not contain the crash days, because each crash was dispersed across the later exit dates of the trades that were open through it. Figure 2 varies ρ:

ρ (weight on common shock)	0.0	0.2	0.4	0.5	0.6	0.8	1.0
exit-day / MTM Sharpe ratio	0.64	0.97	1.65	1.93	2.12	2.19	2.14

Under purely idiosyncratic P&L, exit-day marking *deflates* the Sharpe ratio; the inflation appears only as positions come to share shocks, saturating near 2.1–2.2×— which brackets what we measured in production, on a book whose positions share a volatility factor almost completely.

The mechanism is portfolio-level smoothing, and it leaves GLM's fingerprint. At the level of a single trade, exit-day marking *concentrates*: eleven days of P&L collapse onto one. At the level of the portfolio it *smooths*, because a shock common to many open positions is dispersed across their staggered exit dates. That is the same operation Getmansky et al. (2004) model for illiquid holdings, and it leaves the same trace. In the simulation, the first-order autocorrelation of the exit-day series rises monotonically with commonality — +0.02 at $\rho = 0$, +0.21 at $\rho = 0.5$, +0.29 at $\rho = 1$ — tracking the inflation almost exactly. The MTM-daily series moves the other way, from +0.16 at $\rho = 0$ to −0.03 at $\rho = 1$.

That second series is the important one, and it disciplines the diagnostic. A *correctly* marked daily series is not serially uncorrelated: with overlapping multi-day holds, consecutive days share open positions, and the resulting variation in inventory induces positive autocorrelation — here +0.16 — with no marking error whatsoever. So the naive reading, that positive serial correlation in a liquid-instrument P&L series indicates a marking convention, is wrong: it would flag every correctly marked book with overlapping holds. What separates the two is the *sign of the relationship with commonality*, not the level. Correct marking's autocorrelation is an overlap artifact that common shocks wash out, so it *falls* as positions correlate; lumping's is manufactured by the convention itself, so it *rises*. The usable test is therefore conditional: estimate the autocorrelation the position overlap alone implies, and treat only the excess — and specifically excess that grows with cross-position correlation — as evidence of a marking convention. Where that test fires, GLM's smoothing-adjusted Sharpe is the right correction to apply. We flag this as a diagnostic worth developing rather than a validated instrument: we have calibrated it on a simulation and one book, and its false-positive rate on a wider sample of correctly marked, overlapping-hold strategies is unmeasured.

Why this matters beyond bookkeeping. A developer testing a marking convention on synthetic i.i.d. P&L will conclude the convention is harmless. The convention is harmless *there*. It is maximally harmful for short-volatility, correlated, episode-clustered books —the books whose risk statistics are most consequential and most scrutinized. We know of no prior statement of this conditionality.

Invariant. One mark row per trade per held trading day, asserted at build time. The assertion is cheap because marks telescope: interior re-marks cannot change a trade's total, so enforcing coverage cannot alter reconciled dollars—only their timing.

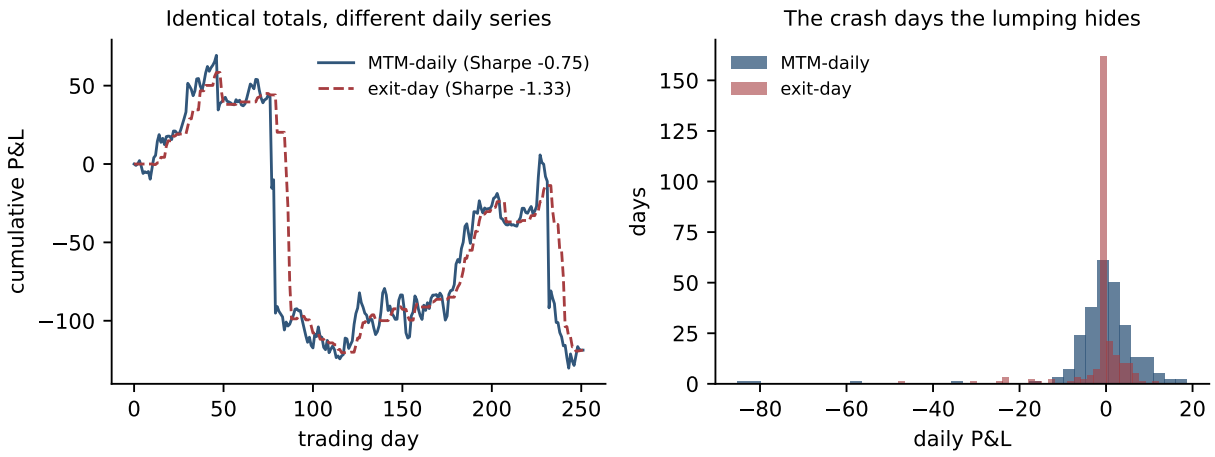

Figure 1: Exhibit 1. Synthetic 120-trade book, fully common shocks. Left: cumulative P&L is identical by construction; the daily paths are not. Right: the exit-day series is missing the crash days entirely—they have been smeared across later exit dates.

5 Classes 2–11: the catalogue

Class 2, trajectory truncation. The exit is located by *index* into a per-trade trajectory array. When a leg leaves the quoted panel mid-hold—delisting, a moneyness filter, or quote loss, all of which concentrate on stress days—the index silently reads a stale row. Cost: a "12.64" Sharpe (INVALID) that was 100% artifact, and a second strategy's 4.14 that died on the same re-audit. Invariant: locate exits on the calendar and force-close a leg that leaves the panel at its last real quote.

Class 3, boundary drops. Splitting a tape into periods drops trades straddling a boundary. The drop is not random—it selects long and escalating holds, i.e. the risk-carrying trades. Cost: a walk-forward missing 24% of its trades. Invariant: per-period tape count equals entry-log count.

Class 4, missing-leg deferral. Per-leg marks are written only on days the leg has a quote; deep in-the-money legs lose quotes on crash days, so the crash-day move lumps onto the next quoted day. Totals unchanged, tail-day *timing inverts*: the book can print a gain on the crash day. Cost: a core sleeve 1.99 → 1.37. Caught only when someone asked why a short-volatility book's *best* day was a crash day—which we now run as a standing diagnostic. Invariant: re-mark quote-less held days (carry plus intrinsic shift, floored at intrinsic: model-free and total-preserving) and assert coverage.

Class 5, mid-fill fantasy. One accounting line, at mid or at modeled fills. Cost: an "8.58" (modeled, INVALID) that was −8.20 at real crossed quotes and reached a live deployment before the three-line rule existed; separately, a +9.67 at mid that was −14.6 at 1% friction. Invariant: three lines always, decide on full cross, and report the line-1-minus-line 3 gap as the result's friction sensitivity.

Class 6, active-day annualization. Computing the Sharpe ratio over active days only and annualizing by $\sqrt{252}$. Cost: up to 8× on sparse books; one +15.97 (INVALID). Invariant: pad to the full calendar with zeros.

Class 7, sign bug. Net exceeding gross—the spread credited rather than paid—via sign conventions in multi-leg friction. Invariant: assert line 3 ≤ Line 1 every run.

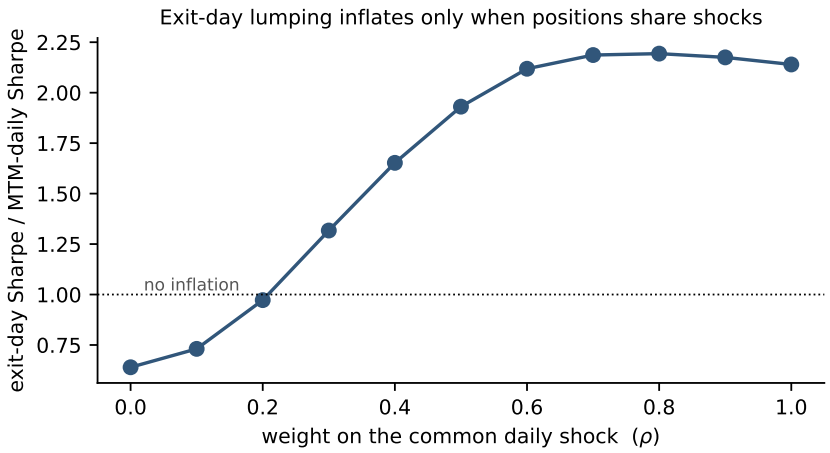

Figure 2: Exhibit 2. The inflation is conditional on shock commonality. At $\rho = 0$ exit-day marking deflates the Sharpe ratio (0.64×); at $\rho = 1$ it inflates it (2.14×). Median over 40 seeds per point.

Class 8, dropped or retried candidates. Entries that never resolved in-window quietly discarded, or retried until they resolve favorably: look-ahead selection on the exit. Cost: revoked a live deployment (honest Sharpe ≈ 0 from "8.04", INVALID). Invariant: force-close, never drop; class 3's count check catches the residue.

Class 9, wrong-instrument lookup. Forward P&L looked up by bucket (delta band, moneyness bin) rather than exact contract; as the underlying moves, the bucket re-maps to a different instrument and the "position" silently mutates. Our synthetic reproduction is deliberately stark: with two strikes and one day of spot movement, the bucket join reports -0.40 and the exact-identity join $+1.50$ on what is nominally the same trade. Invariant: join on exact (expiration, strike, type).

Class 10, same-snapshot execution. Filling at the quote that generated the signal. For any quote-derived signal this embeds the signal in the fill. In the reproduction, a "buy the dip" rule applied to pure i.i.d. quote noise earns a large apparent edge when filled at the signal quote and exactly zero when filled at the next one. Invariant: next-quote fills, strictly after signal time.

Class 11, calendar-indexed differencing. Differencing on a calendar-date index deletes every Monday and post-holiday observation, or converts a "48-hour" hold into a two-trading-day hold spanning a weekend. Invariant: all differencing and holding periods on the trading-day grid.

6 Class 12: attribution failure

Mechanism. Dollars right, timing right, coverage right, fills honest—and the P&L still does not come from the hypothesis. Any structure carrying a risk premium or a beta harvests it regardless of the overlay signal. If the signal is roughly side-symmetric, it splits the book between premium-collecting and premium-paying trades, and the net result is a *diluted premium* that the researcher reads as a signal result.

Worked example (all figures full cross fills, marked daily, full calendar, 2020-09 to 2026-05). A residual-fade tape passed every check in classes 1–11: coverage asserted, counts tied, sign check held, force-close applied, calendar-time exits. Headline: Sharpe 0.34, 95% CI $[-0.18, 0.90]$. Three diagnostics then located the money:

- **Concentration.** 98% of the P&L came from a single calendar year; another year lost roughly a third of the total.
- **Leg decomposition.** In the profitable year the *option leg lost* money while the delta hedge made it—i.e. the tape's P&L was a hedge path, not an option-selection result.
- **Side-randomized control.** Re-running the identical entries with the side replaced: signal 0.34, random sides -1.62, **always-short 1.75** $[0.69, 3.32]$, always-long -2.28. The strategy's own structure, stripped of its signal, scored five times better than the signal.

The signal's genuine discrimination was approximately \$4.9 per trade against a short-side premium of roughly \$24 per trade: real, and roughly five-fold too small to matter. No accounting audit could have reached this conclusion, because no accounting was wrong.

Invariant (two-part). (i) Decompose P&L by leg and require the thesis leg to carry the result. (ii) Re-run the identical tape with randomized and constant sides; if a constant side beats the signal, there is no signal. Both are inexpensive—the second is a loop over the existing tape with one array replaced—and in our experience neither is standard practice.

7 Two audits that passed and were still wrong

Case A. A competent pre-deployment leakage audit produced eight severity-ranked findings, correctly identified a real exit-fill leak worth -1.41 Sharpe, corrected the headline from 8.04 to 6.81 in-sample and 6.10 out-of-sample (that strategy's own basis, later INVALID in full), and certified the strategy operational. It was subsequently invalidated entirely by classes 1 and 8. The audit was not incompetent; it was scoped to check the tape against itself, and both defects lived upstream of the tape it was given.

Case B. A five-year projection turning \$250k into \$18.5M (Sharpe "6.7", INVALID) rested on a slippage constant of \$4.80 per contract that had never been measured. The measured value was \$9.68. At the measured value the strategy loses money at full-spread fills. This is class 5 with an unvalidated parameter in the friction-aware line: a single number, never checked, propagating through an otherwise careful analysis.

The lesson. Audits validate computations; invariants constrain constructions. An audit runs when someone schedules it, examines what it is pointed at, and cannot see the assumptions baked into its inputs. An invariant runs on every build and fails loudly. The twelve classes argue for shifting effort from the former to the latter.

8 What a screen can and cannot tell you

The class-12 example carries a final lesson worth separating out. The screen that motivated that tape projected approximately \$2.75 per contract net; the tape realized \$2.56—the screen's *mean* was accurate to within 7%. What the screen could not see was everything that determined whether the strategy was worth running: the concentration of P&L in one year, the 56% of gross consumed by the spread, the commissions taking half of the remainder, and the daily volatility that set the Sharpe ratio at 0.34.

Screens are accurate about means and uninformative about risk. The dangerous ones are precisely those validated on average P&L per trade, where concentration and volatility are structurally

invisible. We suggest that any screen used to justify building a tape should be reported with the explicit caveat that it constrains the mean only—and that the tape, not the screen, is what decides.

9 The library

`opentape` ships each invariant as a callable and each class as a synthetic failing-then-passing test. Marking: `mtm_daily_marks` with `assert_full_coverage`. Accounting: `three_lines`, `full_calendar_sharpe`, `sign_check`. Structure: `assert_calendar_exit`, `assert_trade_count_ties`. Attribution: `attribution_decompose`, `negative_control_sides`, `best_day_check`. No market data is included or required; every exhibit in this paper is reproducible from the test suite. A tape that passes all twelve invariants is not thereby correct; a tape that fails any one is wrong or unproven.

Suggested adoption order, by cost-to-benefit in our experience: (1) full-calendar padding and the coverage assertion, which together retire classes 1, 4 and 6 for a few lines; (2) the sign check and count check, which are one line each and retire classes 3, 7 and 8; (3) exact-identity joins and calendar-time exits, which are refactors rather than additions (classes 2, 9, 11); (4) the three-line convention (class 5), which requires quote data most programs already have; (5) next-quote fills (class 10); and (6) the attribution pair (class 12), which is last only because it is the one that requires interpreting a result rather than asserting a property.

10 Limitations

This is a single-program catalogue: one asset class, one friction regime, one team's failure modes. The magnitudes—2.2–2.4×, 8×, the simulated 0.64-to-2.2 curve—are illustrations of mechanism, not population estimates, and the simulation's saturation point depends on parameters we chose to resemble our book. The invariant set is necessary, not sufficient: we make no claim that twelve is the complete list, and we would expect a cash-equity or futures program to contribute classes we have never seen. Class-12 detection by constant-side controls applies to side-symmetric signals; other exposures require other controls (volatility-bucket-matched nulls, sector-neutral randomizations, and so forth), and constructing the right null for a given structure remains a matter of judgment rather than a recipe.

References

Arnott, R., Harvey, C., Markowitz, H. (2019). A backtesting protocol in the era of machine learning. *Journal of Financial Data Science* 1(1).

Bailey, D., López de Prado, M. (2014). The deflated Sharpe ratio: correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management* 40(5).

Bailey, D., Borwein, J., López de Prado, M., Zhu, Q. (2014). Pseudo-mathematics and financial charlatanism. *Notices of the AMS* 61(5).

Bailey, D., Borwein, J., López de Prado, M., Zhu, Q. (2017). The probability of backtest overfitting. *Journal of Computational Finance* 20(4).

Chen, A. (2021). The limits of p-hacking: some thought experiments. *Journal of Finance* 76(5).

Chordia, T., Subrahmanyam, A., Tong, Q. (2014). Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? *Journal of Accounting and Economics* 58(1).

Frazzini, A., Israel, R., Moskowitz, T. (2018). Trading costs. Working paper, AQR Capital Management.

Getmansky, M., Lo, A., Makarov, I. (2004). An econometric model of serial correlation and illiquidity in hedge fund returns. *Journal of Financial Economics* 74(3).

Harvey, C., Liu, Y. (2015). Backtesting. *Journal of Portfolio Management* 42(1).

Harvey, C., Liu, Y., Zhu, H. (2016). ... and the cross-section of expected returns. *Review of Financial Studies* 29(1).

Hou, K., Xue, C., Zhang, L. (2020). Replicating anomalies. *Review of Financial Studies* 33(5).

Kapoor, S., Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4(9).

Kapoor, S., et al. (2024). REFORMS: consensus-based recommendations for machine-learning-based science. *Science Advances* 10(18).

Lo, A. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal* 58(4).

McLean, R., Pontiff, J. (2016). Does academic research destroy stock return predictability? *Journal of Finance* 71(1).

Novy-Marx, R., Velikov, M. (2016). A taxonomy of anomalies and their trading costs. *Review of Financial Studies* 29(1).

Sullivan, R., Timmermann, A., White, H. (1999). Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Finance* 54(5).