

HOW TO USE THE SHARPE RATIO

Marcos López de Prado, Alexander Lipton, and Vincent Zoonekynd

ADIA Lab Research Paper Series, No. 19

March 7, 2026

Marcos López de Prado is Global Head, Quantitative Research & Development, Abu Dhabi Investment Authority (ADIA); Advisory Board Member, ADIA Lab; Professor of Practice, College of Engineering, Cornell University; Research Fellow, Applied Mathematics & Computational Research Department, Lawrence Berkeley National Laboratory.

marcos.lopezdeprado@adia.ae

Alexander Lipton is Global Head, Quantitative Research & Development, Abu Dhabi Investment Authority (ADIA); Advisory Board Member, ADIA Lab.

alexander.lipton@adia.ae

Vincent Zoonekynd is Quantitative Research & Development Lead, Abu Dhabi Investment Authority (ADIA); Research Affiliate, ADIA Lab.

vincent.zoonekynd@adia.ae

HOW TO USE THE SHARPE RATIO

ABSTRACT

The Sharpe ratio is the dominant metric for evaluating investment skill, yet inference based on it is routinely flawed—often leading to false confidence, incorrect conclusions, and costly decisions. This paper proposes a new standard for Sharpe ratio inference and reporting by diagnosing common sources of error and providing practical corrections grounded in modern statistical theory. We identify five recurring pitfalls: (i) reporting point estimates without statistical significance; (ii) biased inference caused by wrongly assuming independent and identically distributed Normal returns; (iii) ignoring test power and minimum sample length requirements; (iv) misinterpreting p -values as probabilities that the null is true; and (v) failing to correct for multiple testing and selection effects. To address these issues, we solve a long-standing open problem in financial econometrics: the derivation of a closed-form approximation to the sampling distribution of the Sharpe ratio estimator when returns are jointly non-Normal and serially correlated. Monte Carlo experiments confirm that the proposed framework yields more reliable inference than classical t -statistics and standard multiple-testing adjustments. The key message is straightforward: the Sharpe ratio remains useful for manager ranking, strategy selection, portfolio construction, and asset allocation, but only when paired with a comprehensive inference framework and disciplined reporting—otherwise it becomes a powerful generator of false discoveries. Results can be replicated using the code available at <https://github.com/zoonek/2025-sharpe-ratio>.

KEY TAKEAWAYS

- Standard Sharpe ratio inference is biased in finite samples and further distorted by serial correlation, non-Normality, and multiple testing. Without correction, their reported values provide unwarranted confidence and lead to suboptimal investment decisions.
- Corrected measures improve reliability: Tools such as the Probabilistic Sharpe Ratio (PSR), Minimum Track Record Length (MinTRL), Observed Bayesian False Discovery Rate (oFDR), and Deflated Sharpe Ratio (DSR) provide statistically sound inference, accounting for non-Normal returns, serial correlation, short samples, and selection bias.
- Different corrections suit different contexts: FWER-based methods (like DSR) are more appropriate for contexts where decisions have system-wide repercussions, while FDR-based methods (like SFDR) better fit contexts where decisions have local repercussions.

Keywords: Sharpe ratio, non-Normality, serial correlation, multiple testing, inference.

JEL Classification: G0, G1, G2, G15, G24, E44.

AMS Classification: 91G10, 91G60, 91G70, 62C, 60E.

A central principle of modern portfolio theory is that investors are willing to bear risk only to the extent that they expect to be compensated for it. Investment efficiency is commonly defined as the amount of return achieved per unit of risk. The most widely accepted measure of investment efficiency is the Sharpe ratio, which expresses excess return relative to volatility, and was introduced in a series of seminal papers by Sharpe [1966, 1975, 1994].¹ For example, Sharpe ratios are used to report actual and backtested performance;² identify new investment factors; rank, hire and fire portfolio managers; filter investment strategies (e.g., in due diligence questionnaires); define targets and constraints in portfolio optimization programs, etc.

While the Sharpe ratio is reported ubiquitously in academic and practitioner publications, the inference done on it is often wrong, for at least five reasons. First, a point estimate of the Sharpe ratio does not convey information about its statistical significance. A more meaningful way to measure and compare investment efficiency is to report the Sharpe ratio in the probability space. Second, Sharpe ratio inference is biased by several factors, including sample length, skewness, kurtosis, serial correlation, and multiple testing. It is critical to correct the estimated Sharpe ratio for those variables before making an investment decision. Third, when practitioners and academics use the Sharpe ratio to test an investment's efficiency, they almost never report the power of the test. Without this information, they may be accepting an unreasonably high type-II error (the proportion of false negatives or misses), thus concealing issues such as using a sample length that is too short for the effect being measured. Fourth, practitioners and academics often compute the p -value of the Sharpe ratio (i.e., the upper-tail probability at the observed Sharpe ratio, conditional on the null hypothesis of zero efficiency being true), but use it as if it represented its Bayesian posterior (i.e., the probability that the null hypothesis of zero efficiency holds, conditional on observing a Sharpe ratio at least as large as the realized value). Computing the probability of the null hypothesis given the data is arguably more relevant for Sharpe ratio inference, and yet this is rarely done in practice. Fifth, the False Strategy Theorem (Bailey and López de Prado [2014]) proved that it is trivial to achieve any arbitrary value of the Sharpe ratio through multiple testing. There is no fixed rejection threshold that controls for a given familywise false positive rate (type-I error), and the rejection threshold increases with the number of independent trials and the variance of the Sharpe ratios across trials.³ Similarly, controlling for a given false discovery rate (i.e., the proportion of false strategies among all positives) requires making assumptions regarding the proportion of efficient investments, and their expected Sharpe ratio. Without adjusting for multiple tests, Sharpe ratios are essentially useless.

Unless these five pitfalls are addressed, the Sharpe ratios computed on backtested or real returns will lead to incorrect inference, inefficient portfolios and, what is worse, investment losses. The

¹ The 1966 definition of the Sharpe Ratio computed the standard deviation on absolute returns. Sharpe revised his definition in 1994 to compute the standard deviation on excess returns. Throughout this paper we refer to the 1994 revision, as published in *The Journal of Portfolio Management*.

² For a discussion of the three types of backtests, see Joubert et al. [2024].

³ Intuitively, a trial takes place every time a Sharpe ratio is computed to choose an investment among several independent candidates. For example, a researcher may conduct 1,000 backtests before settling for a particular design of an investment strategy, or a hedge fund may interview 100 portfolio managers before hiring one. The more independent trials take place, the higher is the probability that at least one false positive will appear. Also, the larger is the variance of the Sharpe ratios across all trials, the greater is the expected value of the maximum Sharpe ratio for a given number of independent trials. For additional details, see Appendix 3.

goals of this paper are to develop better tools for doing inference on the Sharpe ratio, to provide a clear manual for their use, and to improve current standards for decision-making and reporting.

LITERATURE REVIEW

There is a large literature on statistical inference for the Sharpe ratio. In this section we focus on the most influential contributions. In a landmark paper, Lo [2002] derived an asymptotic framework for Sharpe ratio inference and provided explicit results under the standard assumption of independent and identically distributed (i.i.d.) Normal returns. For the empirically more realistic case of serial dependence and non-Normality, he formulated the problem in HAC/GMM (heteroskedasticity- and autocorrelation-consistent long-run covariance estimation within the Generalized Method of Moments) form but did not provide an explicit, implementable closed-form approximation to the sampling distribution of the Sharpe ratio estimator.⁴ Because the i.i.d. Normal assumption is empirically problematic (Lo and MacKinlay [1999], Kat and Lu [2002], Brooks and Kat [2002], Agarwal et al. [2004], Fung et al. [2008]), subsequent work extended Sharpe ratio inference to more realistic cases. Mertens [2002] produced a closed-form expression for non-Normal returns, but it still assumed serial independence. Ledoit and Wolf [2008] proposed robust Sharpe ratio inference based on HAC standard errors and a studentized time-series bootstrap. While their resampling approach avoided the i.i.d. Normal returns assumption, it did not yield a closed-form approximation to the Sharpe ratio’s sampling distribution.

A point estimate of the Sharpe ratio does not convey information about statistical significance. A more actionable piece of information for an investor is to estimate the probability of observing a Sharpe ratio below the realized one, conditional on the null hypothesis. To that purpose, Bailey and López de Prado [2012] introduced: (i) the probabilistic Sharpe ratio (PSR), which expresses an observed Sharpe ratio in the probabilistic space, while adjusting for sample length, skewness, and kurtosis; (ii) the minimum track record length (MinTRL), which computes the smallest number of observations needed to reject a null hypothesis with a given confidence level; and (iii) the concept of Sharpe ratio efficient frontier, which allows investors to optimize a portfolio in risk-adjusted terms while controlling for the uncertainty derived from track record length and non-Normal returns.

Applying Extreme Value Theory, Bailey and López de Prado [2014] and Bailey et al. [2014] introduced the deflated Sharpe ratio (DSR), a Sharpe-ratio-specific familywise error rate (FWER) correction that adjusts for sample length, non-Normal returns, and selection bias under multiple testing. Fabozzi and López de Prado [2018] proposed a template for disclosing multiple tests to investors. López de Prado [2020] derived closed-form equations for estimating the power of DSR. DSR incorporates information about the number of trials, the variance of Sharpe ratios across those trials, and the effective correlation structure of the backtests (López de Prado and Lewis [2019]). These investment-relevant features allow DSR to deliver more precise and powerful inference than generic multiple-testing methods applied to the Sharpe ratio.

⁴ Lo [2002] proposed using a GMM framework to obtain the asymptotic variance of the Sharpe ratio under serial dependence, but stopped short of deriving a closed-form expression, stating instead the problem in GMM form (see his equations (14) and (A15), and compare them to our equation (3)). Lo [2002, equation (20)] derived a closed-form expression for time aggregation (annualization) of Sharpe ratios under serial correlation, which addresses a distinct problem from the sampling distribution of the Sharpe ratio estimator. For a detailed explanation, see our Appendix 2.

Several general FWER procedures have been adapted for Sharpe ratio inference, without developing Sharpe-specific sampling theory. White [2000] and Hansen [2005] introduced bootstrap-based tests designed to correct for data-snooping bias, which can be applied to comparisons of Sharpe ratios across competing strategies. Romano and Wolf [2005, 2016] proposed stepdown and resampling-based FWER controls, formulated for arbitrary test statistics, that have been used in empirical finance to adjust inference on Sharpe ratios. Under the assumption of i.i.d. Normal returns, Harvey and Liu [2015] treat the Sharpe ratio as a re-scaled t-statistic and apply classical Bonferroni and Holm corrections—generic multiple-testing tools rather than Sharpe-specific inference methods. López de Prado [2018] introduced Combinatorial Purged Cross-Validation (CPCV), a simulation-based method that uses resampling to deflate Sharpe ratios under backtest overfitting, complementing the DSR framework. With the exception of DSR, none of the above FWER methods operate on the Sharpe ratio’s generalized sampling distribution.

As an alternative to FWER controls, researchers have proposed methods that target the proportion of false strategies among all positives (false discovery proportion, FDP), or the expected value of that proportion (false discovery rate, FDR). Under a frequentist framework, Romano et al. [2008] develop multiple-testing procedures that control FDP. Also under a frequentist framework, Harvey and Liu [2015] use the Benjamini–Hochberg–Yekutieli adjustment to derive an FDR-controlled rejection threshold for t-statistics, which indirectly yields an implied Sharpe rejection threshold. Harvey and Liu [2020] employ Efron’s double-bootstrap methodology to obtain an FDR-calibrated hurdle for generic performance t-statistics. Under an empirical-Bayes framework, Harvey et al. [2025] adopt Efron’s local false discovery rate (local FDR) methodology, extending it to accommodate cross-sectional dependence among test statistics and to settings where the total number of tests may be unknown or only partially observed. These FDR/FDP methods were not developed specifically for the Sharpe ratio, hence they do not operate on the Sharpe ratio’s sampling distribution. Their rejection thresholds can be mapped to Sharpe ratio thresholds through the usual Sharpe-to-t conversion, however this conversion comes at the cost of assuming i.i.d. Normal returns.

INNOVATIONS

While the primary objective of this paper is educational, it contains several notable innovations. First, we derive an explicit, implementable closed-form approximation to the sampling distribution of the Sharpe ratio’s plug-in estimator by applying the functional delta method to an M-estimator and using a long-run (HAC) covariance to accommodate serial dependence and non-Normal returns. As explained earlier, Lo [2002] developed inference for the Sharpe ratio under the standard i.i.d. Normal assumption, and formulated the more realistic dependent case as a HAC/GMM problem, without carrying the derivation to an explicit closed-form expression.⁵ The importance of developing such closed-form results for non-i.i.d. and non-Normal returns was recognized shortly thereafter and acknowledged as an open challenge (Wolf [2003]; Lo [2003]).⁶ Notably, despite its clear practical importance, this problem remained unresolved for more than two decades

⁵ See Lo [2002, equation (14)]. For a detailed explanation, see our Appendices 1 and 2.

⁶ In a letter to the editor of the *Financial Analysts Journal*, Lo [2003] explained that “the IID case was meant primarily to be illustrative and is only of limited practical value because the IID assumption is often violated for financial data.” Despite that author’s disclaimer and warning, Sharpe ratio p -values continue to be misreported regularly in academic and practitioner publications, due to the incorrect assumption that returns are i.i.d. Normal.

after it was first articulated.⁷ Bailey and López de Prado [2012] derived closed-form formulas for the i.i.d. non-Normal case; in this paper we answer Wolf and Lo’s challenge by further dropping the i.i.d. assumption, thus solving the problem that Lo left stated in HAC/GMM form. To our knowledge, the framework presented here enables Sharpe ratio inference under substantially more general assumptions than prior closed-form treatments.

Second, we derive the Sharpe ratio’s Planned Bayesian False Discovery Rate (that is, the probability that the null is true conditional on its rejection), denoted as pFDR, without assuming i.i.d. Normal returns. Equivalently, pFDR is the probability complementary to precision: the probability that the observed Sharpe ratio was drawn from the null distribution, conditional on rejecting the null.

Third, we derive the Sharpe ratio’s Observed Bayesian False Discovery Rate (that is, the probability that the null is true conditional on observing a Sharpe ratio at least as large as the realized value), denoted as oFDR, also without assuming i.i.d. Normal returns. Researchers often misinterpret p -values as the probability of the null hypothesis given the evidence (Wasserstein et al. [2019]), rather than their true meaning as the probability of the evidence given the null hypothesis, and both (Planned and Observed) Bayesian False Discovery Rates help resolve this confusion.

Fourth, we show that the standard Extreme Value Theory (EVT) approximation for the dispersion of the maximum test statistic is systematically misspecified. When the location and scale normalizers are jointly estimated from the same finite sample, a non-vanishing covariance term arises. Ignoring this dependence leads to structural under- or over-control of the FWER at empirically relevant values of K . By explicitly accounting for this effect, we derive a corrected variance expression. The resulting rejection thresholds differ materially from classical EVT benchmarks and yield qualitatively different strategy-selection outcomes, demonstrating that this correction is not a refinement but a necessary condition for valid inference when normalizers are jointly estimated from the same sample.

Fifth, we introduce a multiple-testing control for investment strategies that differs fundamentally from existing FWER and FDR approaches, called *sequential* FDR (SFDR). Classical FDR methods—such as Benjamini and Hochberg [1995], Benjamini and Yekutieli [2001], Storey [2002, 2003], Efron et al. [2001], and Efron [2004, 2008]—are batch procedures designed to control the expected proportion of false discoveries within a contemporaneous cross-section of hypotheses, including Bayesian and empirical-Bayes variants that estimate posterior error rates from a cross-sectional mixture. Online multiple-testing procedures (e.g., Foster and Stine [2008], Javanmard and Montanari [2015, 2018], and Ramdas et al. [2018]) instead handle streams by allocating time-varying test levels and controlling long-run aggregate error rates defined through

⁷ A closed-form solution is particularly valuable in practice because it replaces heuristic bandwidth choices, resampling schemes, and matrix-based long-run covariance estimation with an explicit, auditable expression that can be evaluated instantly and implemented consistently across systems. This is not merely a matter of convenience: Sharpe ratios are widely embedded in portfolio construction routines, optimization engines, risk budgeting, and investment approval processes, where inference must be fast, reproducible, and explainable. In the absence of tractable closed-form inference under serially correlated, non-Normal returns, practitioners are often forced either to revert to the unrealistic i.i.d. Normal assumption for the sake of tractability or to rely on computationally intensive and difficult-to-govern resampling procedures.

cumulative false discoveries and rejections. In contrast, SFDR requires neither a batch nor an alpha-allocation mechanism: it is formulated as a sequential approval rule that sets a Sharpe ratio cutoff to ensure that, among approved strategies, the probability of approving a false strategy remains below a pre-specified level. This yields a per-approval posterior-error guarantee tailored to sequential investment committee decisions and, to our knowledge, is distinct from both the classical/empirical-Bayes FDR literature and the online-testing literature.

Sixth, we introduce an algorithm for the calculation of the rejection threshold that targets a user-defined pFDR level. While other algorithms may have been developed in the past for FDR controls, we believe that this is the first of its class that applies the Sharpe ratio’s sampling distribution and that it does not assume i.i.d. Normal returns.

Seventh, we synthesize these methodological advances into an improved standard for Sharpe ratio reporting and decision-making. Rather than treating the Sharpe ratio as a standalone point estimate, this standard formalizes the joint reporting of estimation uncertainty under realistic return processes, significance, power, posterior false discovery probabilities, and appropriate multiple-testing adjustments. This unified framework bridges classical inference, Bayesian decision theory, Extreme Value Theory, machine learning, and practical investment governance, providing actionable guidance for both researchers and practitioners. As far as we know, no prior work has articulated a comprehensive reporting standard for Sharpe ratio inference that is simultaneously statistically coherent, decision-theoretically grounded, and applicable to sequential investment approval processes.

THE SHARPE RATIO

Consider a sample of T excess returns of an investment strategy, $\{r_t\}_{t=1,\dots,T}$, from a stationary process with finite population mean μ and variance σ^2 . The true (unobserved) Sharpe ratio (SR) is defined as

$$SR = \frac{\mu}{\sigma} \tag{1}$$

Sharpe [1966] originally called this the “reward-to-variability” ratio. It assesses an investment’s efficiency in terms of the reward investors receive per unit of dispersion, and it can be interpreted as a measure of risk-adjusted performance, signal over noise, or skill over luck. The Sharpe ratio satisfies several important properties: (i) the tangency portfolio in Markowitz’s efficient frontier has maximum Sharpe ratio; (ii) its definition only requires that the first two moments exist and be finite; (iii) under linear and sufficiently low funding costs, it is approximately scale-free within a given sampling frequency, which facilitates the assessment of skill across investments with different levels of leverage.

In the context of property (i), ranking portfolios by mean and variance—or equivalently by their Sharpe ratios—yields the same ordering as expected-utility maximization under either of two sufficient conditions. First, when these returns follow the Normal distribution (more generally, when returns follow an elliptical distribution), all investors with increasing and concave utility functions rank portfolios solely by their means and variances, rendering higher-order moments irrelevant for ranking. In this case, efficient portfolios are ordered by their Sharpe ratios. Second,

when investor preferences are quadratic, expected utility depends only on mean and variance, regardless of the return distribution, and Sharpe ratio ordering again coincides with mean–variance ranking. Outside these special cases, mean–variance analysis remains valid whenever investor preferences admit a mean–variance representation, either exactly or as a second-order approximation, as originally emphasized by Markowitz [1952, 1959]. When returns are non-Normal or serially correlated, variance may no longer fully characterize downside or tail risk,⁸ but it continues to measure dispersion, and the Sharpe ratio retains its economic interpretation as expected excess return per unit of standard deviation.⁹

A common misunderstanding regarding property (ii) is that, because mean and variance fully determine the Normal distribution, and the Sharpe ratio is defined on these two moments, then users of the Sharpe ratio must implicitly assume that returns are i.i.d. Normal. This argument is incorrect. The Sharpe ratio is well defined for any process with finite mean and variance, regardless of higher-order moments or temporal dependence. Normality and independence are therefore not structural assumptions underlying the Sharpe ratio itself; they are modeling conveniences often introduced later to simplify statistical inference. The same situation occurs in regression analysis, where t-values of estimated betas can be computed without assuming that observations are i.i.d. Normal; those t-values may not follow a Student’s t distribution, but asymptotically valid inference can still be conducted using the appropriate generalized sampling distribution. Following the same rationale, next we derive the generalized sampling distribution of the Sharpe ratio.

GENERALIZED SAMPLING DISTRIBUTION OF THE SHARPE RATIO

Hedge fund strategy returns are characterized by short sample lengths, a positive Sharpe ratio, positive serial correlation, negative skewness, and positive excess kurtosis. These five features increase the sampling variance of the Sharpe ratio estimator.

Exhibit 1 confirms that the literature’s standard assumptions of Normality and serial independence are unwarranted and unrealistic. Statistics are computed on the monthly return series of Hedge Fund Research’s main style indices (Equity Hedge, Event Driven, Relative Value and Macro), as well as the weighted composite, from January 1990 (the start of the series) to November 2025 (the last available observation). For all cases, we must reject the hypothesis of Normality (see Jarque-Bera statistics and *p*-values) and serial independence (see 10-lag Ljung-Box statistics and *p*-values) at conventional significance levels. Throughout this paper, we compute Sharpe ratios in the frequency of the observations, without annualizing them, because annualization is unnecessary for inference.

⁸ Later in this paper we show that PSR incorporates downside and tail-risk information, providing yet another reason to prefer PSR over point estimates of the Sharpe ratio.

⁹ For utility functions that depend on higher-order moments, it is possible in principle to construct alternative reward-to-risk ratios that reflect investor preferences better than the Sharpe ratio when returns are non-Normal. In practice, however, investors rarely agree on a particular utility function, which makes it difficult to define a universally accepted alternative ratio. Moreover, even if consensus were reached on such an alternative preference-reflecting ratio, developing a coherent inferential framework for it—comparable to the one derived here for the Sharpe ratio—would pose substantial theoretical, mathematical, and practical challenges. These considerations help explain the Sharpe ratio’s historical success and suggest that, despite its limitations, it is likely to remain the dominant benchmark for assessing investment efficiency.

HFR Indices	Composite	Equity Hedge	Event-Driven	Relative Value	Macro
BBG Code	HFRIFWI Index	HFRIEHI Index	HFRIEDI Index	HFRIRVA Index	HFRIMI Index
Mean	0.007	0.009	0.008	0.007	0.007
StDev	0.019	0.026	0.020	0.012	0.020
Skew	-0.711	-0.319	-1.425	-2.703	0.694
Kurt	6.381	5.303	9.889	22.897	4.611
AR(1)	0.249	0.191	0.300	0.365	0.176
T	431	431	431	431	431
JB (stat)	234.920	99.130	974.210	7457.520	79.160
JB (p)	0.000	0.000	0.000	0.000	0.000
LB-10 (stat)	41.820	31.960	53.810	82.520	55.150
LB-10 (p)	0.000	0.000	0.000	0.000	0.000

Exhibit 1 – Non-Normality and serial-correlation in hedge fund returns (monthly frequency)

Under these stylized facts, the generalized sampling distribution of the Sharpe ratio's plug-in estimator (\widehat{SR}) is, for a sufficiently large T , approximately

$$\widehat{SR} = \frac{\hat{\mu}}{\hat{\sigma}} \underset{a}{\sim} \mathcal{N} \left[SR, \frac{1}{T} \left(\frac{1+\rho}{1-\rho} - \frac{1+\rho+\rho^2}{1-\rho^2} \gamma_3 SR + \frac{1+\rho^2}{1-\rho^2} \frac{\gamma_4-1}{4} SR^2 \right) \right] \quad (2)$$

where \mathcal{N} denotes the Normal distribution, γ_3 is the skewness of the excess returns, γ_4 is Pearson's kurtosis of the excess returns (with value 3 when returns are Normal), and ρ is the first-order autocorrelation of the excess returns, with $\rho \in (-1,1)$. Appendix 1 provides a proof of the above statement.

Applying the estimator \widehat{SR} on the sample $\{r_t\}_{t=1,\dots,T}$ we obtain a particular estimate \widehat{SR}^* . Replacing the above parameters with their estimates, the estimated variance of the Sharpe ratio's estimator evaluated at (or under the assumption that) $SR = \widehat{SR}^*$, denoted by $\sigma^2[\widehat{SR}^*]$, is¹⁰

$$\begin{aligned} \sigma^2[\widehat{SR}^*] &= V[\widehat{SR} | SR = \widehat{SR}^*] \\ &= \frac{1}{T} \left(\frac{1+\hat{\rho}}{1-\hat{\rho}} - \frac{1+\hat{\rho}+\hat{\rho}^2}{1-\hat{\rho}^2} \hat{\gamma}_3 \widehat{SR}^* + \frac{1+\hat{\rho}^2}{1-\hat{\rho}^2} \frac{\hat{\gamma}_4-1}{4} \widehat{SR}^{*2} \right) \end{aligned} \quad (3)$$

A numerical example may help clarify these concepts. Consider a portfolio manager with a two-year track record of monthly returns, where $(\hat{\mu}, \hat{\sigma}, \hat{\gamma}_3, \hat{\gamma}_4, \hat{\rho}, T) = (0.036\%, 0.079\%, -2.448, 10.164, 0.2, 24)$. The estimated Sharpe ratio is $\widehat{SR}^* = 0.456$, with an estimated standard deviation of $\sigma[\widehat{SR}^*] = 0.379$. However, assuming i.i.d. Normal returns, the estimated standard deviation would be approximately 43% smaller, $\sigma[\widehat{SR}^*] = 0.214$. This evidences that ignoring the non-Normality and serial correlation of returns can lead to a gross underestimation of the Sharpe ratio's variance, which in turn means a higher than expected rate of false positives.

¹⁰ Some readers may prefer to use the notation $\hat{\sigma}^2[\widehat{SR}^*]$ to refer to $V[\widehat{SR} | SR = \widehat{SR}^*]$.

PROBABILISTIC SHARPE RATIO

Following Bailey and López de Prado [2012], let SR_0 be the Sharpe ratio threshold that an investor sets to separate false strategies ($SR \leq SR_0$) from true strategies ($SR > SR_0$). We can assess whether a strategy with observed Sharpe ratio \widehat{SR}^* is a true strategy by testing the one-sided null hypothesis $H_0: SR \leq SR_0$ against the alternative $H_1: SR > SR_0$. Under H_0 , the test statistic ($z^*[SR_0]$) is

$$z^*[SR_0] = \frac{\widehat{SR}^* - SR_0}{\sigma[SR_0]} \xrightarrow{d} \mathcal{N}[0,1] \quad (4)$$

$$\begin{aligned} \sigma[SR_0] &= \sqrt{V[\widehat{SR}|SR = SR_0]} \\ &= \sqrt{\frac{1}{T} \left(\frac{1 + \hat{\rho}}{1 - \hat{\rho}} - \frac{1 + \hat{\rho} + \hat{\rho}^2}{1 - \hat{\rho}^2} \hat{\gamma}_3 SR_0 + \frac{1 + \hat{\rho}^2}{1 - \hat{\rho}^2} \frac{\hat{\gamma}_4 - 1}{4} SR_0^2 \right)} \end{aligned} \quad (5)$$

The test statistic compares the estimated Sharpe ratio (\widehat{SR}^*) to *the least favorable case* in the null hypothesis (SR_0), and adjusts the difference by the standard error under the null hypothesis ($\sigma[SR_0]$).¹¹ The equation for $\sigma[SR_0]$ results from evaluating the estimator under the least favorable case of the null hypothesis, $V[\widehat{SR}|SR = SR_0]$.

Practitioners often compare Sharpe ratios computed on different sampling frequencies (daily, weekly, monthly, ...) by scaling them into an “annualized” equivalent. When returns are i.i.d., the scaling factor is the square root of the number of observations per year. Lo [2002, “Time Aggregation”] derived the scaling factor under non-i.i.d. returns. However, we argue that re-scaling is not the correct way of comparing Sharpe ratios. Even if two Sharpe ratios are computed on the same sampling frequency, they are not directly comparable, because the sample length (T) may be different, among other variables that influence the estimator’s variance. The test statistic $z^*[SR_0]$, computed on the original sampling frequency, makes it possible to compare Sharpe ratios without the need for annualization.¹²

The one-sided test mirrors the standard decision problem faced by investors, where false strategies must be filtered out before deciding the asset allocation among the true ones. The significance level α (false positive rate, type-I error) is the probability of rejecting H_0 when it is true,

$$\alpha = P[\widehat{SR} \geq SR_c | H_0] = 1 - Z \left[\frac{SR_c - SR_0}{\sigma[SR_0]} \right] \quad (6)$$

¹¹ Least favorable in the sense of minimizing the chances of rejecting the null hypothesis.

¹² Recent work questions the common practice of time aggregation and annualization of Sharpe ratios across horizons. Welch [2025] shows that the familiar square-root-of-time scaling ignores compounding effects and can lead to misleading conclusions about long-horizon risk–reward tradeoffs, implying that Sharpe ratios need not increase with investment horizon and may ultimately lose economic interpretability. This critique is consistent with our paper’s emphasis that Sharpe ratios computed at different horizons are not directly comparable, both because heuristic time aggregation distorts their economic interpretation and because statistical uncertainty itself depends on the evaluation horizon. Accordingly, our analysis focuses on valid inference conditional on a given evaluation horizon rather than on extrapolating Sharpe ratios across horizons through heuristic scaling rules.

where $Z[\cdot]$ denotes the CDF of the standard Normal distribution. The critical value of the test (SR_c) can be computed as

$$z_{1-\alpha} = Z^{-1}[1 - \alpha] \quad (7)$$

$$SR_c = SR_0 + \sigma[SR_0]z_{1-\alpha} \quad (8)$$

We reject H_0 with confidence $(1 - \alpha)$ if $z^*[SR_0] \geq z_{1-\alpha} \Leftrightarrow \widehat{SR}^* \geq SR_c$. The Probabilistic Sharpe Ratio (PSR) is the probability of observing a Sharpe ratio below \widehat{SR}^* conditional on H_0 being true,

$$PSR = P[\widehat{SR} < \widehat{SR}^* | H_0] = Z[z^*[SR_0]] = 1 - P[\widehat{SR} \geq \widehat{SR}^* | H_0] = 1 - p \quad (9)$$

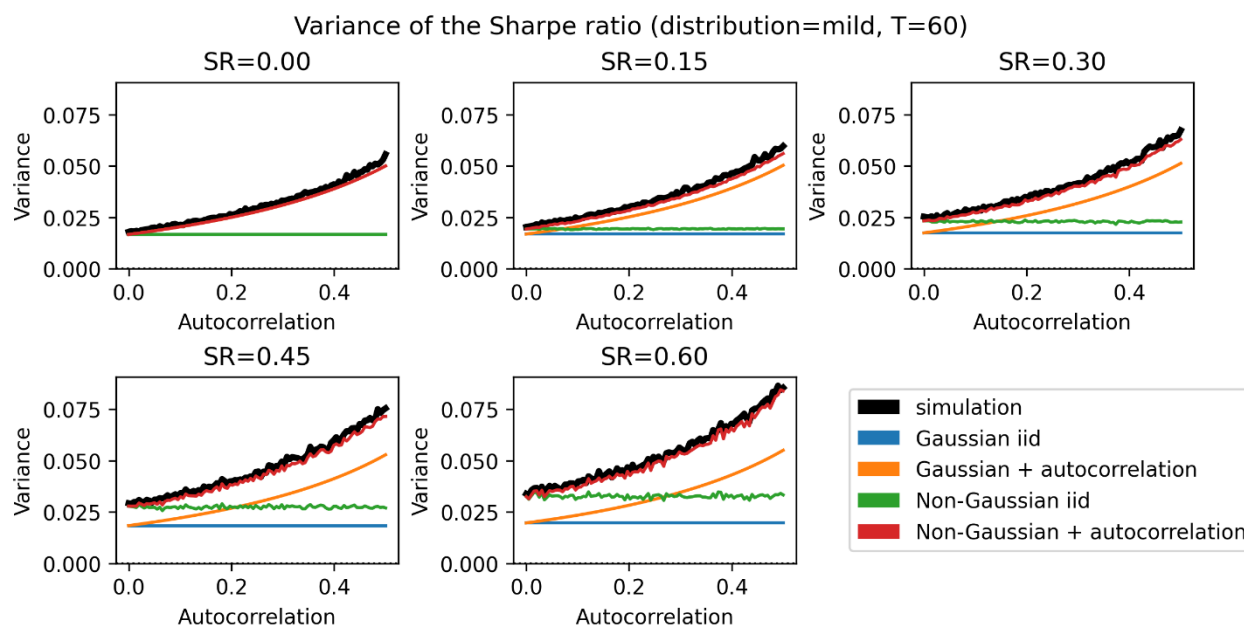
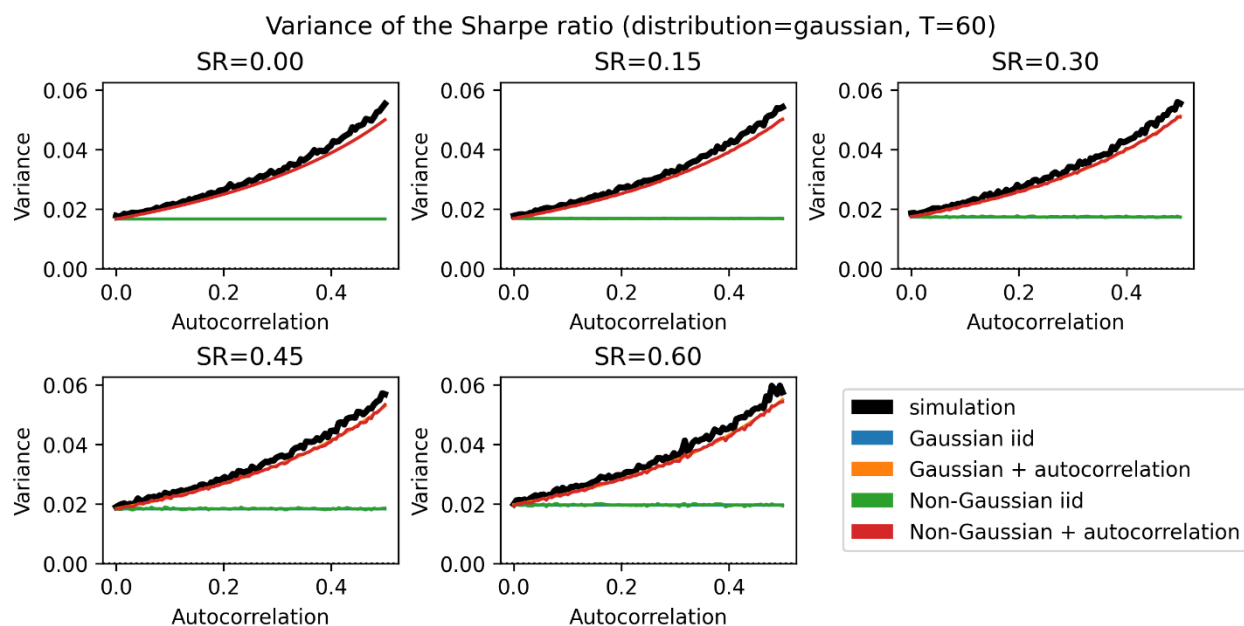
where $p = P[\widehat{SR} \geq \widehat{SR}^* | H_0]$ is known as the test's p -value. This may also be interpreted as the maximum confidence with which the null hypothesis can be rejected after observing \widehat{SR}^* . Following with our numerical example, under the null hypothesis where $SR_0 = 0$, then $PSR = Z[z^*[0]] = Z\left[\frac{\widehat{SR}^*}{\sigma[SR_0]}\right] = 0.966$, but under the null hypothesis where $SR_0 = 0.1$, then $PSR = 0.900$.

Note that under the null hypothesis where $SR_0 = 0$ and i.i.d. returns, the value of $z^*[0]$ reduces to $\widehat{SR}^* \sqrt{T}$, which coincides with the statistic of the non-central Student's t-distribution test. PSR and Student's t tests are also equivalent under i.i.d. Normal returns. PSR and Student's t tests differ under non-i.i.d. returns, and also under i.i.d. non-Normal returns when $SR_0 \neq 0$. PSR's generality is a strong reason for preferring it over other tests that assume i.i.d. Normal returns.¹³

An important property of PSR is that, unlike Sharpe ratio point estimates, it assigns lower values to investments exposed to downside, tail and drawdown-related risks. This follows directly from equation (5), which penalizes $SR_0 > 0$ strategies when they exhibit negative skewness (downside risk), positive excess kurtosis (tail risk), or positive serial correlation (a common driver of volatility clustering and drawdowns). To our knowledge, PSR is the only risk-adjusted performance measure that simultaneously satisfies the following properties: (i) explicitly penalizes uncertainty due to short samples, downside, tail and drawdown-related risks; (ii) is directly usable for ranking and decision thresholds; (iii) is grounded in modern portfolio theory via the Sharpe ratio; and (iv) admits a closed-form analytical inferential framework.¹⁴

¹³ Under i.i.d. Normal returns and very small sample ($T < 30$), a non-central Student's t-distribution test (with $T - 1$ degrees of freedom and non-centrality parameter $\delta = SR_0 \sqrt{T}$) may be more precise than PSR, however that advantage vanishes when returns are not i.i.d. Normal, making PSR a better choice in practice.

¹⁴ Numerous ratios have been proposed to emphasize downside or tail risk, such as the Sortino ratio (Sortino and van der Meer [1991]), which replaces total variance with downside semi-deviation; distribution-based measures such as the Omega ratio (Keating and Shadwick [2002]); tail-based measures such as return-to-CVaR ratios (Boudt et al. [2008]); and path-dependent measures such as the return-to-drawdown (Calmar) ratio (Young [1991]). While these alternatives may be useful as descriptive or ranking heuristics in specific contexts, they are not grounded in modern



portfolio theory—unlike the Sharpe ratio, which characterizes the tangency portfolio—and they lack a well-developed and widely accepted inferential framework. Their sampling distributions, estimation uncertainty, test power, and multiple-testing corrections are not computable analytically in closed-form manner. As a result, statistical inference for these ratios typically relies on simulation, resampling, or ad hoc approximations rather than on closed-form sampling theory (see, e.g., Zakamouline and Koekebakker [2009]). These shortcomings may explain why none of these alternatives has been widely accepted as a replacement for the Sharpe ratio.

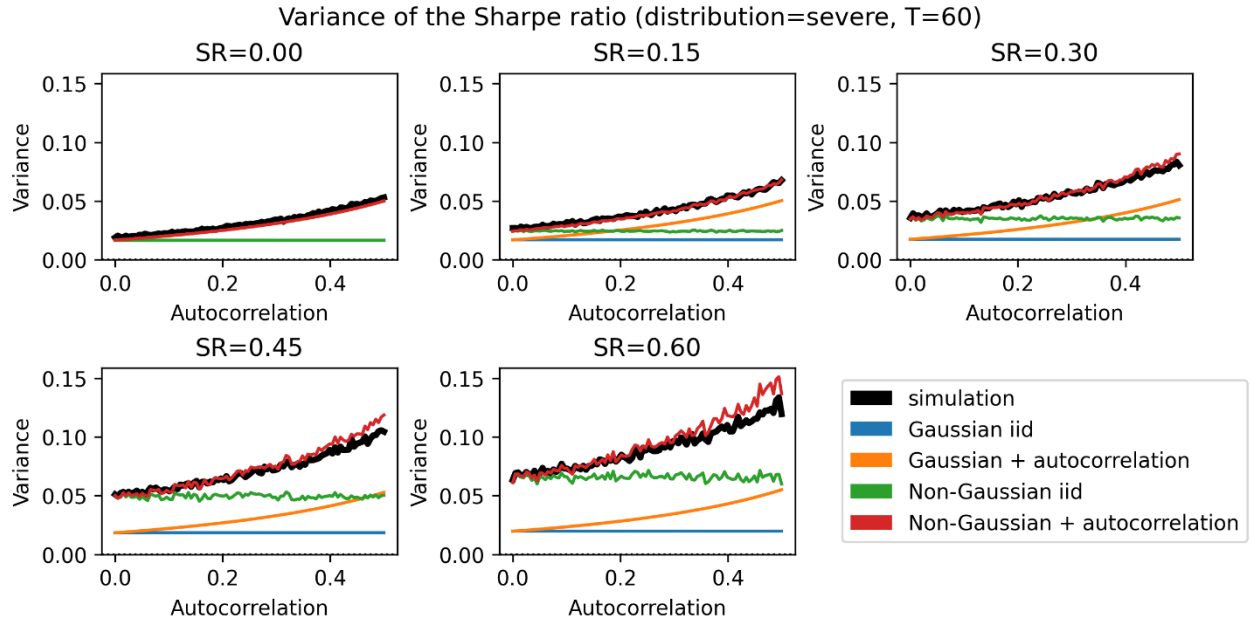
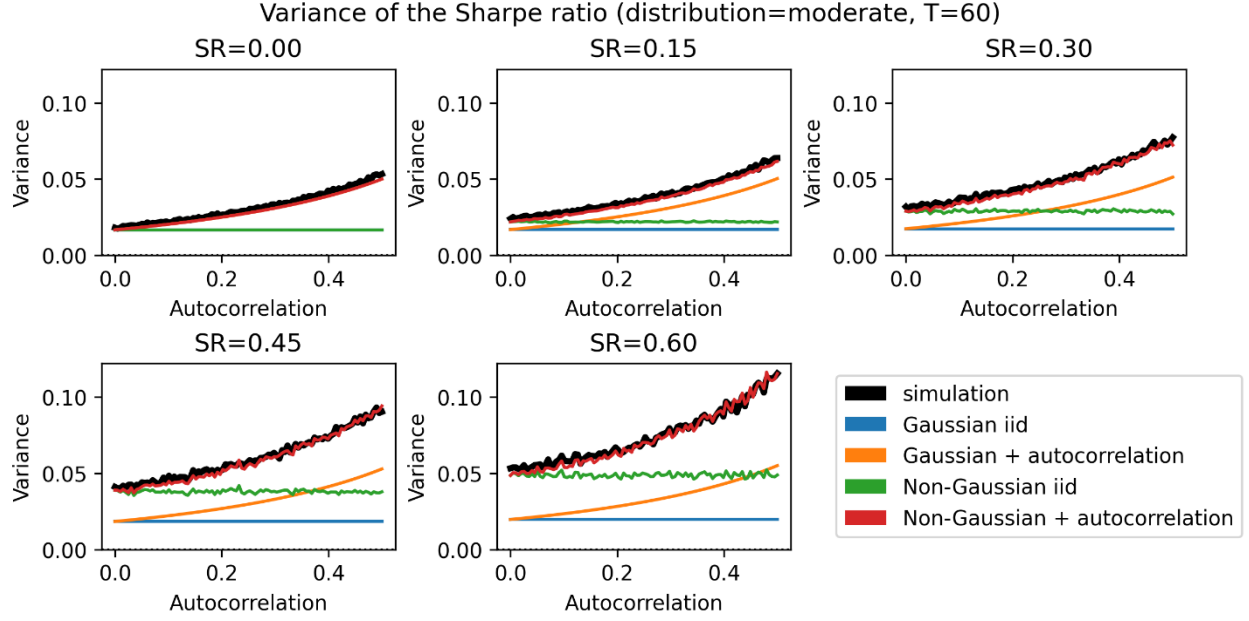


Exhibit 2 – Variance of the Sharpe ratio’s estimator as a function of autocorrelation, under different degrees of non-Normality (monthly frequency, $T = 60$)

To evaluate the accuracy of the proposed variance formula under realistic return dynamics, we conduct a Monte Carlo experiment based on controlled data-generating scenarios. Innovations are drawn i.i.d. from Mixtures of Gaussians calibrated to exhibit four degrees of non-Normality (Gaussian, mild, moderate, and severe), consistent with the stylized facts reported in Exhibit 1. Serial dependence is then introduced by filtering these innovations through a linear AR(1) structure with autocorrelation coefficients ranging from $\rho = 0$ to $\rho = 0.5$. For each of the four distributional scenarios, we consider five values of the true Sharpe ratio, $SR_0 \in \{0, 0.15, 0.3, 0.45, 0.6\}$, yielding 20 controlled data-generating scenarios in total. Each simulation consists of $T = 60$ monthly

observations (five years), and 10,000 independent replications are generated for each scenario. All parameters—including skewness, kurtosis, and serial correlation—are re-estimated on each simulated sample, so that inference explicitly reflects the estimation error inherent in short return histories.

Exhibit 2 reports results for the 20 aforementioned controlled data-generating scenarios. Each panel plots the variance of the Sharpe ratio estimator as a function of the autocorrelation coefficient ρ , with the sample length fixed at $T = 60$. The “Simulation” curve represents the benchmark variance obtained from Monte Carlo experiments under the specified data-generating scenario. The remaining curves apply equation (5) under progressively less restrictive modeling assumptions. “Gaussian i.i.d.” reproduces Lo [2002], hence assuming Normal and serially independent returns. “Gaussian + autocorrelation” relaxes independence while maintaining Normality. “Non-Gaussian i.i.d.” incorporates empirical skewness and kurtosis but ignores serial dependence, as in Bailey and López de Prado [2012]. Finally, “Non-Gaussian + autocorrelation” applies the full expression in equation (5), using estimates of skewness, kurtosis, and serial correlation. Some of the curves overlap under certain scenarios, as it can be deduced from equation (5).

The plots show that assuming i.i.d. Normal returns leads to a severe understatement of the sampling variance whenever returns exhibit either serial dependence or higher-order moments, and especially when both are present. Realistic scenarios show that the actual variance of the Sharpe ratio can be four or more times larger than its estimate under the i.i.d. Normal assumption. Partial corrections—accounting for serial correlation but not non-Normality, or vice versa—capture only a fraction of the true variance. Only the full expression in equation (5) closely tracks the Monte Carlo benchmark across all scenarios, without requiring a large T . Although skewness, kurtosis, and autocorrelation are estimated with noise in small samples, the experiment demonstrates that incorporating these noisy estimates yields substantially more accurate inference than imposing incorrect i.i.d. Normal assumptions. The improvement of PSR over classical Student-t-based inference increases with (i) the severity of non-Normality, (ii) the strength of serial correlation, and (iii) the magnitude of the true Sharpe ratio being tested (SR_0).¹⁵

MINIMUM TRACK RECORD LENGTH

Following Bailey and López de Prado [2012], the minimum track record length (MinTRL) is defined as the minimum sample size T such that the observed \widehat{SR}^* (together with $\hat{\rho}$, $\hat{\gamma}_3$, $\hat{\gamma}_4$) allows the rejection of H_0 at significance level α . Formally, the problem can be stated as

$$MinTRL = \min_T \{P[\widehat{SR} \geq \widehat{SR}^* | H_0] \leq \alpha\} \quad (10)$$

with solution when $\widehat{SR}^* > SR_0$

¹⁵ The code needed to reproduce the exhibits and numerical examples in this paper is available at <https://github.com/zoonek/2025-sharpe-ratio>

$$MinTRL = \left(\frac{1 + \hat{\rho}}{1 - \hat{\rho}} - \frac{1 + \hat{\rho} + \hat{\rho}^2}{1 - \hat{\rho}^2} \hat{\gamma}_3 SR_0 + \frac{1 + \hat{\rho}^2}{1 - \hat{\rho}^2} \frac{\hat{\gamma}_4 - 1}{4} SR_0^2 \right) \left(\frac{z_{1-\alpha}}{\widehat{SR}^* - SR_0} \right)^2 \quad (11)$$

Equivalently, MinTRL can be defined as the minimum sample size T such that PSR is not less than $(1 - \alpha)$. Following with our numerical example, for $\alpha = 0.05$ and under the null hypothesis where $SR_0 = 0$, then $MinTRL = 19.543$ months, however under the null hypothesis where $SR_0 = 0.1$, then the minimum track record length more than doubles, to $MinTRL = 39.369$ months. It takes a longer sample to reject a SR_0 that is closer to the observed \widehat{SR}^* . One way to validate these results is to replace in the PSR equation the value of T with MinTRL, thus obtaining $(1 - \alpha)$.

TRUE POSITIVE RATE (POWER, RECALL, SENSITIVITY)

Following López de Prado [2020], let SR_1 be the expected value of the alternative hypothesis, $H_1: SR > SR_0$. In practice, SR_1 can be set to the average Sharpe ratio observed among strategies that have yielded acceptable performance as defined by an investor. Then, the false negative rate (β , type-II error) is defined as the probability of not rejecting H_0 given that H_1 is true,

$$\beta = P[\widehat{SR} < SR_c | H_1] = Z \left[\frac{SR_c - SR_1}{\sigma[SR_1]} \right] \quad (12)$$

Power is defined as the probability of rejecting the null when it is false,¹⁶

$$P[\widehat{SR} \geq SR_c | H_1] = 1 - \beta \quad (13)$$

Power is determined ex-ante by test parameters, not the observed \widehat{SR}^* . When conditioning on the alternative hypothesis, the analogue of the p -value is the upper-tail probability evaluated at the observed statistic, $P[\widehat{SR} \geq \widehat{SR}^* | H_1]$. This quantity is sometimes referred to as *achieved power*, although it is rarely used in practice (not to be confused with *observed power*, often defined as $P[\widehat{SR} \geq SR_c | SR = \widehat{SR}^*]$).

The choice of α (false positive rate) determines β (false negative rate), hence also $1 - \beta$ (true positive rate). To see this, note that

$$SR_c = SR_0 + \sigma[SR_0] z_{1-\alpha} \quad (14)$$

$$\begin{aligned} 1 - \beta &= P[\widehat{SR} \geq SR_c | H_1] = 1 - Z \left[\frac{SR_c - SR_1}{\sigma[SR_1]} \right] \\ &= 1 - Z \left[\frac{SR_0 + \sigma[SR_0] z_{1-\alpha} - SR_1}{\sigma[SR_1]} \right] \end{aligned} \quad (15)$$

¹⁶ In different fields, power is sometimes also denoted recall, sensitivity or true positive rate.

$$\begin{aligned}\sigma[SR_1] &= \sqrt{V[\widehat{SR}|SR = SR_1]} \\ &= \sqrt{\frac{1}{T} \left(\frac{1 + \hat{\rho}}{1 - \hat{\rho}} - \frac{1 + \hat{\rho} + \hat{\rho}^2}{1 - \hat{\rho}^2} \hat{\gamma}_3 SR_1 + \frac{1 + \hat{\rho}^2}{1 - \hat{\rho}^2} \frac{\hat{\gamma}_4 - 1}{4} SR_1^2 \right)}\end{aligned}\quad (16)$$

In particular, for $SR_0 = 0$, the value of β can be simplified into

$$\beta = Z \left[\frac{z_{1-\alpha} \sqrt{\frac{1 + \hat{\rho}}{1 - \hat{\rho}}} - SR_1 \sqrt{T}}{\sqrt{\frac{1 + \hat{\rho}}{1 - \hat{\rho}} - \frac{1 + \hat{\rho} + \hat{\rho}^2}{1 - \hat{\rho}^2} \hat{\gamma}_3 SR_1 + \frac{1 + \hat{\rho}^2}{1 - \hat{\rho}^2} \frac{\hat{\gamma}_4 - 1}{4} SR_1^2}} \right] \quad (17)$$

These equations show that, for a given SR_0 and SR_1 , we can increase the power of the test either by increasing α (at the expense of more type-I errors) or by increasing the sample length T . Note that $\hat{\gamma}_3$, $\hat{\gamma}_4$ and $\hat{\rho}$ are strategy characteristics typically not under the direct control of the researcher. One possibility would be to use the above equation to derive the value of T needed to achieve a target power $(1 - \beta)$, as an alternative (or complement) to MinTRL.¹⁷

Non-Normality	Skew	Kurt	AR(1)	SR1	Precision	Recall	F1
gaussian	0.0	3.0	0	0.15	0.861	0.316	0.463
gaussian	0.0	3.0	0	0.3	0.930	0.751	0.831
gaussian	0.0	3.0	0	0.45	0.950	0.966	0.958
gaussian	0.0	3.0	0	0.6	0.947	0.999	0.972
gaussian	0.0	3.0	0.2	0.15	0.865	0.255	0.394
gaussian	0.0	3.0	0.2	0.3	0.921	0.596	0.724
gaussian	0.0	3.0	0.2	0.45	0.942	0.889	0.915
gaussian	0.0	3.0	0.2	0.6	0.949	0.980	0.964
Non-Normality	Skew	Kurt	AR(1)	SR1	Precision	Recall	F1
mild	-0.9	5.7	0	0.15	0.844	0.352	0.497
mild	-0.9	5.7	0	0.3	0.916	0.736	0.816
mild	-0.8	5.5	0	0.45	0.938	0.949	0.944
mild	-0.8	5.3	0	0.6	0.937	0.993	0.964
mild	-0.8	5.5	0.2	0.15	0.806	0.251	0.382
mild	-0.8	5.5	0.2	0.3	0.887	0.582	0.703
mild	-0.9	5.6	0.2	0.45	0.933	0.867	0.899
mild	-0.7	5.1	0.2	0.6	0.925	0.980	0.952

¹⁷ For example, $\min_T \{P[\widehat{SR} \geq SR_c | H_1] \geq 1 - \beta\}$ for a target power $(1 - \beta)$. One disadvantage of such approach is the need to assume a value for SR_1 , which MinTRL does not require.

Non-Normality	Skew	Kurt	AR(1)	SR1	Precision	Recall	F1
moderate	-1.7	10.6	0	0.15	0.836	0.374	0.517
moderate	-1.7	10.3	0	0.3	0.899	0.735	0.809
moderate	-1.6	9.9	0	0.45	0.925	0.926	0.925
moderate	-1.5	9.3	0	0.6	0.924	0.990	0.956
moderate	-1.8	10.4	0.2	0.15	0.795	0.283	0.417
moderate	-1.7	10.4	0.2	0.3	0.875	0.572	0.692
moderate	-1.6	9.9	0.2	0.45	0.913	0.842	0.876
moderate	-1.6	9.9	0.2	0.6	0.919	0.961	0.939
Non-Normality	Skew	Kurt	AR(1)	SR1	Precision	Recall	F1
severe	-2.5	17.1	0	0.15	0.812	0.403	0.539
severe	-2.4	16.6	0	0.3	0.889	0.736	0.805
severe	-2.3	15.9	0	0.45	0.909	0.913	0.911
severe	-2.2	14.9	0	0.6	0.911	0.981	0.945
severe	-2.4	16.2	0.2	0.15	0.800	0.372	0.508
severe	-2.5	17.2	0.2	0.3	0.881	0.586	0.704
severe	-2.3	15.7	0.2	0.45	0.919	0.842	0.879
severe	-2.2	15.1	0.2	0.6	0.920	0.953	0.937

Exhibit 3 – Precision and recall of PSR under different processes (monthly frequency, $T = 60$)

Exhibit 3 reports precision ($P[H_1|\widehat{SR} \geq SR_c]$) and recall ($P[\widehat{SR} \geq SR_c|H_1]$) rates for the Monte Carlo experiment described earlier, where $SR_0 = 0$, with various non-Normality scenarios and values of SR_1 and ρ . Column F1 shows the harmonic mean of precision and recall. The results demonstrate that PSR's power does not decrease with non-Normality or serial correlation across different levels of signal strength, evidencing that the method works as designed.

Following with our numerical example, for $\alpha = 0.05$ and under the alternative hypothesis where $SR_1 = 0.5$, then the false negative rate is $\beta = 0.411$. Incorrectly assuming that returns are i.i.d. Normal would yield a false negative rate of only $\beta = 0.224$, an underestimation of 45%.

PLANNED BAYESIAN FALSE DISCOVERY RATE

The Sharpe ratio's planned Bayesian false discovery rate, denoted as pFDR, is the probability that the null hypothesis is true given that it was rejected,

$$pFDR = P[H_0|\widehat{SR} \geq SR_c] \quad (18)$$

Since precision is defined as the probability that the alternative hypothesis is true given that the null has been rejected, $P[H_1|\widehat{SR} \geq SR_c]$, precision is equal to one minus pFDR. Like power, pFDR is determined ex-ante by test parameters, not the observed \widehat{SR}^* . We can compute pFDR as an application of Bayes' theorem,

$$P[H_0|\widehat{SR} \geq SR_c] = \frac{P[\widehat{SR} \geq SR_c|H_0]P[H_0]}{P[\widehat{SR} \geq SR_c]} \quad (19)$$

From the law of total probability, we know that

$$\begin{aligned} P[\widehat{SR} \geq SR_c] &= P[\widehat{SR} \geq SR_c | H_0]P[H_0] + P[\widehat{SR} \geq SR_c | H_1]P[H_1] \\ &= \alpha P[H_0] + (1 - \beta)(1 - P[H_0]) \end{aligned} \quad (20)$$

resulting in

$$\begin{aligned} P[H_0 | \widehat{SR} \geq SR_c] &= \frac{\alpha P[H_0]}{\alpha P[H_0] + (1 - \beta)(1 - P[H_0])} \\ &= \left(1 + \frac{(1 - \beta)P[H_1]}{\alpha P[H_0]}\right)^{-1} \end{aligned} \quad (21)$$

In a Bayesian interpretation, $P[H_0]$ represents the prior probability that a randomly evaluated strategy is false. In practice, this quantity can be elicited in a data-informed manner as follows: (i) to compute power, the investor defines the average Sharpe ratio of true strategies, denoted SR_1 ; (ii) all evaluated strategies, whether rejected or not, are sorted in descending order by their test statistic, $z^*[SR_0]$; (iii) the subset of top-performing strategies whose average Sharpe ratio is closest to SR_1 is identified; and (iv) $P[H_1]$ is elicited as the proportion of strategies included in that subset, with $P[H_0] = 1 - P[H_1]$. This procedure yields a data-informed prior consistent with SR_1 that reflects the implied prevalence of skill among the evaluated strategies.¹⁸

Following with our numerical example, suppose that $P[H_1] = 0.1$, $\alpha = 0.05$ and $\beta = 0.411$, then $pFDR = 0.433$. This illustrates how a test with relatively high power (at a 58.9% level) can still have a high planned false discovery rate (at a 43.3% level) compared to the targeted false positive rate (at 5% level) when true strategies are relatively rare (10% probability). Incorrectly assuming that returns are i.i.d. Normal would yield a $pFDR = 0.367$, an underestimation of 15%.

OBSERVED BAYESIAN FALSE DISCOVERY RATE

The previous equations show that $pFDR$ is a function of the test characteristics $(\alpha, \beta, P[H_0])$,¹⁹ not the observed \widehat{SR}^* . This invites the question: what is the probability that H_0 is true conditional on the estimated Sharpe ratio being greater or equal to the observed one? This probability, $P[H_0 | \widehat{SR} \geq \widehat{SR}^*]$, denoted as $oFDR$, can be understood as the Bayesian posterior probability associated with the prior, $P[H_0]$, after incorporating the evidence summarized by the p -value, $p = P[\widehat{SR} \geq \widehat{SR}^* | H_0]$. From Bayes' theorem,

$$oFDR = P[H_0 | \widehat{SR} \geq \widehat{SR}^*] = \frac{P[\widehat{SR} \geq \widehat{SR}^* | H_0]P[H_0]}{P[\widehat{SR} \geq \widehat{SR}^*]} \quad (22)$$

From the law of total probability, we know that

¹⁸ Alternatively, an investor could also derive a coherent pair $(P[H_1], SR_1)$, by fitting a mixture of two Gaussians on the test statistics $z^*[SR_0]$, through an algorithm like the one introduced in López de Prado and Foreman [2014]. The main difference between these two approaches is that the first (elicited) method combines data with economic judgment, injected via the investor's choice of SR_1 , whereas the second (mixture) method is purely statistical and may not align with economically meaningful definitions of skill.

¹⁹ In the i.i.d. Normal case, β is a function of α , T and SR_1 , which are determined before the observations take place. In the non-Normal serial-correlation case, β is also a function of $\hat{\rho}$, $\hat{\gamma}_3$ and $\hat{\gamma}_4$, which depend on the observed returns but not \widehat{SR}^* .

$$P[\widehat{SR} \geq \widehat{SR}^*] = P[\widehat{SR} \geq \widehat{SR}^* | H_0]P[H_0] + P[\widehat{SR} \geq \widehat{SR}^* | H_1]P[H_1] \quad (23)$$

$$= pP[H_0] + (1 - Z[z^*[SR_1]])(1 - P[H_0])$$

where $z^*[SR_1] = \frac{\widehat{SR}^* - SR_1}{\sigma[SR_1]}$, resulting in

$$P[H_0 | \widehat{SR} \geq \widehat{SR}^*] = \frac{pP[H_0]}{pP[H_0] + (1 - Z[z^*[SR_1]])(1 - P[H_0])} \quad (24)$$

Following with our numerical example, for $SR_0 = 0$, $SR_1 = 0.5$ and $P[H_1] = 0.1$, then the p -value is $P[\widehat{SR} \geq \widehat{SR}^* | H_0] = 1 - PSR = 0.034$, while the $oFDR$ is $P[H_0 | \widehat{SR} \geq \widehat{SR}^*] = 0.361$. This evidences that an investment may have a statistically significant Sharpe ratio at a 3.4% p -value, and yet the observed false discovery rate can be relatively high (at a 36.1% level), because true strategies are relatively rare. Incorrectly assuming that returns are i.i.d. Normal would yield an $oFDR = 0.165$, an underestimation of 54%.

MULTIPLE TESTING CORRECTIONS

Researchers rarely test a single Sharpe ratio, as they are constantly searching for new anomalies that expand the limits of knowledge, and diversify investment portfolios. Under those circumstances, performing inference as if a single trial had taken place grossly underestimates the probability of a false positive. To see why, consider a sample of K observed Sharpe ratios, $\{\widehat{SR}_k^*\}_{k=1, \dots, K}$, independently drawn from the same distribution.²⁰ We would like to test the null hypothesis $H_0: SR_k \leq SR_0$ for all $k = 1, \dots, K$. Suppose that we set at α the false positive probability in every test k . For $K > 1$ and $0 < \alpha < 1$, the probability that there is at least one false positive is not α , but a larger value α_K , called familywise error rate (FWER),

$$\alpha_K = 1 - (1 - \alpha)^K \quad (25)$$

Two questions arise naturally: (a) what is the new rejection threshold (SR_c) for the strategy with the highest observed Sharpe ratio among K candidates ($\max_k \{\widehat{SR}_k^*\}$), such that it controls for a given FWER level α_K ?; and (b) what is the new rejection threshold (SR_c) such that the proportion of false strategies among all selected strategies (i.e., those with $\widehat{SR}_k^* \geq SR_c$) is controlled at a given $pFDR$ level q ? Question (a) addresses the risk of deploying the single best strategy selected after a broad search, and question (b) addresses the risk associated with approving strategies sequentially, one decision at a time. These two questions control for two different probabilities, and therefore have two different answers.

CASE A: SEARCH-AWARE CONTROL OF THE FAMILYWISE ERROR RATE

Classical FWER methods (Bonferroni [1936], Šidák [1967], Holm [1979], Hochberg [1988]) contemplate the rejection of multiple nulls while controlling for α_K . That classical setting is often not directly applicable to financial research, where authors report only the best outcome after searching across K experiments. Under those circumstances, conducting inference requires

²⁰ We later relax this independence assumption, but it is useful at this point for expositional purposes.

modeling the distribution of the maximum Sharpe ratio. To that purpose, we assume that the K observed Sharpe ratios, $\{\widehat{SR}_k^*\}_{k=1,\dots,K}$, are independently (or approximately independently) drawn under H_0 from a Normal distribution with mean $E[\{\widehat{SR}_k^*\}] = SR_0$ and variance $V[\{\widehat{SR}_k^*\}]$. Note that this assumption applies to the Sharpe ratios observed across trials, and it is compatible with non-Normal serially correlated returns.

Remark 1: Effective Number of Trials

In practice, the trials are often dependent. When that is the case, K can be approximated as the effective number of independent trials, via clustering methods (López de Prado [2019]), or through the eigenvalues of the correlation matrix of trials' returns series (López de Prado [2018, section 18.7], López de Prado [2020]). See Appendix 3 for details and experimental validation.

Remark 2: Cross-Sectional Variance

Regardless of whether trials are dependent or independent, the cross-sectional variance $V[\{\widehat{SR}_k^*\}]$ comprises two sources of variation: (i) expected sampling variance $E[\{\sigma^2[SR_k]\}]$, where $\sigma^2[SR_k]$ is the sampling variance derived from k th-trial's parameters $(T_k, \rho_k, \gamma_{3,k}, \gamma_{4,k}, SR_k)$, following equation (2); and (ii) true heterogeneity across trials, $V[\{SR_k\}]$.²¹ A high $V[\{SR_k\}]$ is consistent with an extensive search outside a predefined theoretical framework, which is more likely to yield Sharpe ratios that are irreproducible out-of-sample (due to backtest overfitting or statistical flukes). Accounting for (ii) helps protect investors from these irreproducible outcomes, which is generally desirable even if that comes at the expense of lower power. Alternatively, investors may choose to impose the sharp-null configuration $SR_k = SR_0$ for all $k = 1, \dots, K$, by using the less conservative estimate $V[\{\widehat{SR}_k^*\} | SR_k = SR_0] = E[\{\sigma^2[SR_k]\}] = \frac{1}{K} \sum_{k=1}^K \sigma_k^2[SR_0]$, where $\sigma_k^2[SR_0]$ is the sampling variance derived from k th-trial's estimates $(T_k, \hat{\rho}_k, \hat{\gamma}_{3,k}, \hat{\gamma}_{4,k}, SR_0)$, following equation (5).

Exact Distribution of the Maximum

For a finite number of trials K (independent or effectively independent), the maximum of Normal variables follows a skewed order-statistic distribution with CDF (see David and Nagaraja [2003]),

$$P \left[\max_k \{\widehat{SR}_k\} < x \right] = \left(Z \left[\frac{x - SR_0}{\sqrt{V[\{\widehat{SR}_k^*\}]}} \right] \right)^K \quad (26)$$

After appropriate centering and scaling, this distribution converges asymptotically to a Gumbel law, see Leadbetter et al. [1983]. The rejection threshold that controls for a given FWER level α_K can be computed as

²¹ Let $X_k = \mu_k + \varepsilon_k$, with $V[\mu_k | k] = 0$, $E[\varepsilon_k | k] = 0$, and $V[\varepsilon_k | k] = \sigma_k^2$. By the law of total variance, for any random draw k^* from $\{1, \dots, K\}$, then $V[X_{k^*}] = V[E[X_{k^*} | k^*]] + E[V[X_{k^*} | k^*]] = V[\mu_{k^*}] + E[\sigma_{k^*}^2]$. In particular, if k^* is uniform, then $V[\mu_{k^*}] = V[\{\mu_k\}]$, and $E[\sigma_{k^*}^2] = E[\{\sigma_k^2\}] = \frac{1}{K} \sum_{k=1}^K \sigma_k^2$.

$$SR_c = SR_0 + Z^{-1}[(1 - \alpha_K)^{1/K}] \sqrt{V[\{\widehat{SR}_k^*\}]} \quad (27)$$

One practical limitation of working directly with the exact distribution of the maximum is that it yields only: (i) a PSR-type lower-tail probability for the selected strategy, $P\left[\max_k\{\widehat{SR}_k\} < \max_k\{\widehat{SR}_k^*\} | H_0\right]$; and (ii) the corresponding rejection threshold (SR_c), that controls for the FWER α_K . However, other quantities defined earlier in the paper require a search-adjusted null Sharpe ratio ($SR_0|K$) and a search-adjusted standard error ($\sigma[SR_0|K]$). Without these quantities, closed-form expressions for important estimands—such as MinTRL, power, pFDR, and oFDR—cannot be derived within a unified Sharpe-ratio-based framework. These limitations are overcome by estimating the expected value and standard deviation of the maximum Sharpe ratio.²²

Expected Value of the Maximum Sharpe Ratio

The False Strategy Theorem (Bailey and López de Prado [2014]) derives the expected value of the maximum as:

$$E\left[\max_k\{\widehat{SR}_k^*\}\right] \approx SR_0 + \sqrt{V[\{\widehat{SR}_k^*\}]} \left((1 - \gamma)Z^{-1}\left[1 - \frac{1}{K}\right] + \gamma Z^{-1}\left[1 - \frac{1}{Ke}\right] \right) \quad (28)$$

where $\gamma = 0.5772156649 \dots$ is the Euler-Mascheroni constant, and e is Euler's number. See López de Prado and Bailey [2021] for estimated error bounds of the above expression.

In a single-trial test of $H_0: SR \leq SR_0$, the least favorable case is the boundary $SR = SR_0$. Under a multiple testing search, however, the object of interest is the selected strategy, whose Sharpe ratio is an extreme order statistic. The one-trial boundary SR_0 is no longer the “least favorable case” in a multiple testing search. Instead, the search-adjusted “least favorable case” null, $SR_{0,K}$, is

$$SR_{0,K} = E\left[\max_k\{\widehat{SR}_k^*\}\right] \quad (29)$$

Standard Deviation of the Maximum Sharpe Ratio

Similarly, under multiple trials, we must estimate the standard deviation of the maximum Sharpe ratio across K strategies, $\sigma[SR_{0,K}]$. This requires re-scaling $\sqrt{V[\{\widehat{SR}_k^*\}]}$ by the standard deviation of the maximum of K standard Normal variables,

$$\sigma[SR_{0,K}] = \sqrt{V\left[\max_k\{\widehat{SR}_k^*\}\right]} = \sqrt{V[\{\widehat{SR}_k^*\}]} \sqrt{V\left[\max_k\{X_k\}\right]} \quad (30)$$

²² Using the mean and variance of the maximum corresponds to a second-order approximation of its distribution, which preserves location and scale while discarding higher-order cumulants (e.g., skewness). This is sufficient when inference depends on moderate tail probabilities and when the statistic of interest is used as a standardized pivot, as in PSR and MinTRL. Higher-order corrections would materially affect only extreme quantiles and are therefore unnecessary for the inferential objectives pursued here.

where $\{X_k\}_{k=1,\dots,K}$ are K i.i.d. standard Normal variables. The re-scaling factor $\sqrt{V\left[\max_k\{X_k\}\right]}$ can be computed as derived in Appendix 4.

	1	2	3	4	5	6	7	8	9	10
0	1.00000	0.82565	0.74798	0.70122	0.66898	0.64492	0.62603	0.61065	0.59779	0.58681
10	0.57728	0.56889	0.56143	0.55473	0.54867	0.54315	0.53808	0.53341	0.52909	0.52507
20	0.52131	0.51780	0.51449	0.51138	0.50844	0.50565	0.50301	0.50050	0.49811	0.49582
30	0.49364	0.49155	0.48954	0.48762	0.48577	0.48399	0.48228	0.48062	0.47903	0.47748
40	0.47599	0.47455	0.47315	0.47180	0.47048	0.46921	0.46797	0.46676	0.46559	0.46445
50	0.46334	0.46226	0.46120	0.46017	0.45917	0.45819	0.45723	0.45629	0.45538	0.45448
60	0.45361	0.45275	0.45192	0.45109	0.45029	0.44950	0.44873	0.44797	0.44723	0.44650
70	0.44579	0.44508	0.44439	0.44372	0.44305	0.44240	0.44175	0.44112	0.44050	0.43989
80	0.43929	0.43870	0.43811	0.43754	0.43698	0.43642	0.43587	0.43533	0.43480	0.43428
90	0.43376	0.43325	0.43275	0.43226	0.43177	0.43129	0.43081	0.43034	0.42988	0.42942

Exhibit 4 – Standard deviation re-scaling factors, from $K = 1$ to $K = 100$

For the reader’s convenience, Exhibit 4 reports the values of re-scaling factors, from $K = 1$ to $K = 100$. Appendix 5 derives an approximation to this expression as

$$\sqrt{V\left[\max_k\{X_k\}\right]} \approx \sqrt{\frac{\pi^2}{6} - \frac{\gamma^2}{1+\gamma}} \left(Z^{-1}\left[1 - \frac{1}{Ke}\right] - Z^{-1}\left[1 - \frac{1}{K}\right] \right) \quad (31)$$

Deflated Sharpe Ratio

Applying these two adjustments to PSR (i.e., replacing SR_0 with $SR_{0,K}$ and replacing $\sigma[SR_0]$ with $\sigma[SR_{0,K}]$) gives the deflated Sharpe ratio (DSR), and applying the same adjustments to MinTRL prevents an underestimation of the minimum sample length. One advantage of DSR over general-purpose correction methods (Bonferroni, Šidák, Holm, Hochberg, etc.) is that those methods do not account for both sources of cross-sectional variation in constructing the null distribution of the selected strategy: $E[\{\sigma^2[\widehat{SR}_k^*]\}]$ and $V[\{SR_k\}]$.²³ The variance of the Sharpe ratios across trials tends to be large in overly complex models and in models found through brute force searches over a large parameter space, i.e. where researchers did not constrain the search to a well-defined theoretical framework. When controlling for false positives and Sharpe ratio inflation, model complexity is not a virtue.²⁴ The False Strategy Theorem demonstrates the virtue of conducting research within the confines of a theoretical causal framework, see López de Prado [2023] and López de Prado and Zoonekynd [2025]. Since $V[\{\widehat{SR}_k^*\}]$ can be much larger than $\sigma^2[SR_0]$ of the

²³ To see this, consider the case $K = 10$ and $\alpha_K = 0.05$. The rejection threshold under Šidák’s correction is $SR_c = SR_0 + \sigma[SR_0]z_{0.995}$, because $\alpha = 1 - (1 - \alpha_K)^{1/K} \approx 0.005$. In contrast, the rejection threshold under DSR is $SR_c = SR_{0,K} + \sigma[SR_{0,K}]z_{0.95}$, where both $SR_{0,K}$ and $\sigma[SR_{0,K}]$ are a function of $V[\{\widehat{SR}_k^*\}]$. This illustrates how Šidák’s correction does not account for cross-sectional dispersion in Sharpe ratio estimates. This critique extends to all generic FWER procedures: by operating solely through adjusted significance levels applied to per-strategy sampling distributions, they do not adjust the null for the extreme-value selection effect driven by cross-sectional dispersion.

²⁴ Complexity is rarely a virtue in finance, see Berk [2023], Buncic [2025], Cartea et al. [2025], Fallahgoul [2025], Harvey et al. [2025], and Nagel [2025], among many others.

selected model, DSR should be preferred over generic FWER corrections in financial applications.²⁵

A Monte Carlo experiment can assess the effectiveness of this DSR control. We can generate time series of monthly returns with length $T = 60$ under different scenarios subject to the null hypothesis being true ($SR_0 = 0$), compute the Sharpe ratios, select the maximum Sharpe ratio out of $K = 10$, and apply the DSR adjustments to derive the rejection threshold SR_c that controls for FWER at level $\alpha_K = 0.05$. We can then compute $P[\widehat{SR} \geq SR_c | H_0]$, and compare that value with the target $\alpha_K = 0.05$. Exhibit 5(a) reports the error $P[\widehat{SR} \geq SR_c | H_0] - \alpha_K$ in column Diff, demonstrating that DSR adjustments work as designed even for small sample sizes. In particular, the control's effectiveness does not materially degrade with the presence of serial correlation. Performance degrades when returns are severely non-Normal, in which case it is recommended to increase the sample length (e.g., sampling with daily rather than monthly frequency), or to apply a different control (e.g., derive the correct SR_c experimentally, via Monte Carlo). Exhibit 5(b) repeats the experiment on 5-years' worth of daily returns ($T = 1300$ observations, the equivalent to 60 months).

Non-Normality	Skew	Kurt	AR(1)	SR_c	Diff
gaussian	0.0	3.0	0	0.337	0.007
gaussian	0.0	3.0	0.2	0.416	0.008
mild	-0.9	5.6	0	0.340	0.044
mild	-0.9	5.6	0.2	0.416	0.034
moderate	-1.7	10.2	0	0.347	0.073
moderate	-1.7	10.2	0.2	0.418	0.068
severe	-2.3	16.1	0	0.352	0.086
severe	-2.3	16.1	0.2	0.421	0.094

Exhibit 5(a) – DSR control under different processes (monthly frequency, $T = 60$, $\alpha_K = 0.05$)

Non-Normality	Skew	Kurt	AR(1)	SR_c	Diff
gaussian	0.0	3.0	0	0.075	0.006
gaussian	0.0	3.0	0.2	0.090	0.002
mild	-0.9	5.6	0	0.075	0.008
mild	-0.9	5.6	0.2	0.089	0.020
moderate	-1.7	10.3	0	0.075	0.024
moderate	-1.7	10.3	0.2	0.089	0.019
severe	-2.4	16.6	0	0.075	0.036
severe	-2.4	16.6	0.2	0.088	0.024

Exhibit 5(b) – DSR control under different processes (daily frequency, $T = 1300$, $\alpha_K = 0.05$)

These adjustments also enable the correct estimation of power, pFDR and oFDR. Following with our numerical example, for $K = 10$, and $V[\{\widehat{SR}_k^*\}] = 0.1$, the one-trial-equivalent to $SR_0 = 0$ is

²⁵ As an alternative to DSR, the CPCV method can also be used to deflate the Sharpe ratio. An advantage of CPCV is that it enables the simulation of trials when that information is missing, however it requires access to the strategy's algorithm. See López de Prado [2018, chapter 12] for details.

shifted to $SR_{0,K} = E \left[\max_k \{\widehat{SR}_k^*\} \right] = 0.498$, and the cross-sectional standard deviation $\sqrt{V[\{\widehat{SR}_k^*\}]} = 0.316$ is rescaled to $\sigma[SR_{0,K}] = 0.186$, which results in $DSR = 0.410$ (compared to the one-trial PSR of 0.966). In words, after accounting for the multiple tests, the high observed Sharpe ratio turns out to be below what would be expected from zero skill (a coin toss).

CASE B: CONTROLLING FOR SEQUENTIAL FALSE DISCOVERY RATE

Benjamini and Hochberg [1995] introduced methods for controlling the expected proportion of false discoveries among a batch of simultaneously rejected null hypotheses. This classical FDR setting is not aligned with typical investment management practice, where strategies are usually evaluated individually over successive meetings of an investment committee rather than selected as a batch. The classical FDR framework would be appropriate if one wished to control the average proportion of false strategies selected within each meeting. In this section, we introduce an alternative FDR formulation that controls the posterior probability of error in each individually approved strategy. When the same rule is applied sequentially across committee decisions, this posterior error probability converges to the expected long-run proportion of errors across all selected strategies. We call this framework *sequential FDR* (SFDR), to differentiate it from the classical (Benjamini-Hochberg) batch FDR that is more relevant in other empirical disciplines.²⁶

Suppose that a researcher wishes to select strategies while ensuring that the posterior probability that *each* selected strategy is false does not exceed q . This is equivalent to solving for the rejection threshold SR_c such that pFDR targets a level q ,

$$\begin{aligned} P[H_0 | \widehat{SR} \geq SR_c] &= \frac{\alpha P[H_0]}{\alpha P[H_0] + (1 - \beta)(1 - P[H_0])} \\ &= \left(1 + \frac{(1 - \beta)P[H_1]}{\alpha P[H_0]} \right)^{-1} = q \end{aligned} \quad (32)$$

Replacing α and β in the above equation, we obtain the equilibrium condition

$$q = \left(1 + \frac{\left(1 - Z \left[\frac{SR_c - SR_1}{\sigma[SR_1]} \right] \right) (1 - P[H_0])}{\left(1 - Z \left[\frac{SR_c - SR_0}{\sigma[SR_0]} \right] \right) P[H_0]} \right)^{-1} \quad (33)$$

²⁶ In the factor investing literature, population-level FDR studies estimate the expected fraction of false discoveries among statistically significant published factors. Studies reach differing conclusions regarding the magnitude of multiple-testing distortions (e.g., Harvey et al. [2016]; Chen and Zimmermann [2022]). Importantly, even a low population-level FDR would not necessarily imply low SFDR values for individual factors. The distinction arises because FDR averages error across discoveries at the population level, whereas SFDR evaluates decision-level error by conditioning explicitly on the selection rule applied to that factor. This perspective helps explain mathematically subsequent empirical evidence documenting limited factor replication and diminished post-selection performance (e.g., Chen et al. [2025]), two patterns associated with selection bias under multiple testing (Bailey et al. [2017]). SFDR therefore reflects the decision environment faced by investors, who allocate capital to individually selected factors rather than across the full population of published predictors.

See Appendix 6 for a step-by-step derivation of this expression. A root-finding algorithm applied to the above expression yields the threshold SR_c that satisfies the condition $P[H_0 | \widehat{SR} \geq SR_c] = q$. Note that under the SFDR framework the researcher chooses *all* strategies above this threshold SR_c , and not only the one with the highest Sharpe ratio among all trials (as is the case in FWER settings). For this reason, SR_0 does not need to be adjusted for the number of trials, like in the DSR correction, and variable K is not part of the equilibrium condition.

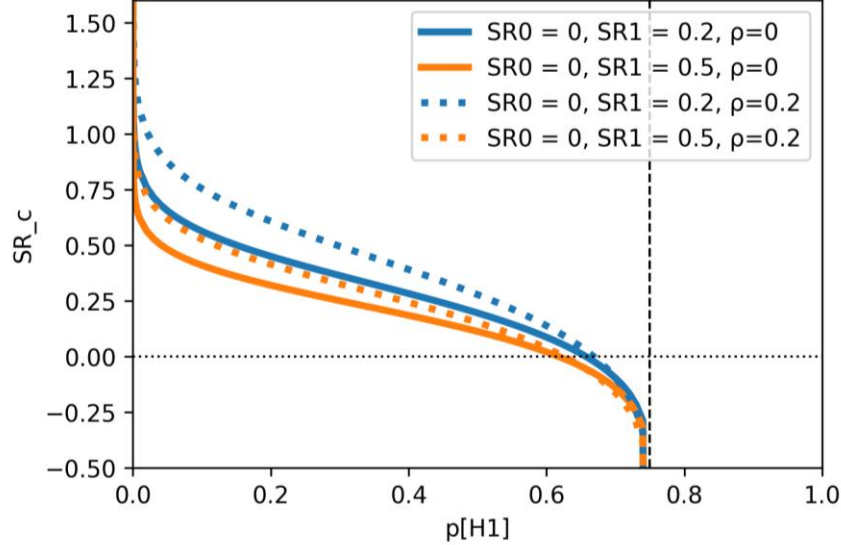


Exhibit 6 – Sharpe ratio rejection thresholds (SR_c) that control for a constant false discovery rate ($q = 0.25$) as a function of the unconditional probability of a true strategy ($P[H_1]$), under $SR_1 = 0.2$, $SR_1 = 0.5$, $\rho = 0$, and $\rho = 0.2$ (monthly frequency, $T = 24$)

Following with our numerical example, Exhibit 6 plots the SR_c thresholds that control for a constant pFDR at a level $q = 0.25$ as a function of $P[H_1]$. The orange line shows results under $SR_0 = 0$, $SR_1 = 0.5$, $\sigma[SR_0] = 0.250$, and $\sigma[SR_1] = 0.393$. For instance, under $P[H_1] = 0.1$, then $SR_c = 0.53$ controls for a constant pFDR at a level $q = 0.25$. Accordingly, the strategy with an observed Sharpe ratio $\widehat{SR}^* = 0.456$ would be discarded. It may seem at first paradoxical that SR_c can be negative. The reason is, as $P[H_1]$ approaches the value $1 - q$, no strategy is discarded regardless of how negative its \widehat{SR}^* is, because the probability that the strategy is false is below the tolerance for false discoveries. Exhibit 6 also illustrates that relaxing the alternative hypothesis, from $SR_1 = 0.5$ (orange line) to $SR_1 = 0.2$ (blue line), has the effect of increasing the rejection thresholds. The reason is, when the Sharpe ratio of true strategies is lower, it is harder to separate true from false strategies, and the threshold must adjust for the increased probability of a false discovery. A similar effect takes place when serial correlation increases from $\rho = 0$ (solid lines) to $\rho = 0.2$ (dashed lines), because that change increases the variance of the Sharpe ratio's estimator.

A Monte Carlo experiment can assess the effectiveness of the SFDR control. We can generate time series of monthly returns with length $T = 60$ under different scenarios subject to the null hypothesis being true ($SR_0 = 0$) with probability $P[H_0] = 0.9$, and the alternative hypothesis $SR_1 = 0.5$ being true with probability $P[H_1] = 0.1$. Using the procedure described earlier, we can

derive the rejection threshold SR_c that controls for $q = 0.25$. Exhibit 7(a) reports the error $P[H_0 | \widehat{SR} \geq SR_c] - q$ in column Diff, as well as the precision, recall and F1, demonstrating that SFDR adjustments work as designed even for small sample sizes. In particular, the control's effectiveness does not materially degrade with the presence of serial correlation. Performance degrades when returns are severely non-Normal, in which case it is recommended to increase the sample length (e.g., sampling with daily rather than monthly frequency), or to apply a different control (e.g., derive the correct SR_c experimentally, via Monte Carlo). Exhibit 7(b) repeats the experiment on 5-years' worth of daily returns ($T = 1300$ observations, the equivalent to 60 months).

Non-Normality	Skew	Kurt	AR(1)	SR1	P[H1]	SR_c	Precision	Recall	F1	Diff
gaussian	0.0	3.0	0	0.15	0.1	0.397	0.709	0.037	0.071	0.041
gaussian	0.0	3.0	0	0.3	0.1	0.257	0.732	0.628	0.676	0.018
gaussian	0.0	3.0	0	0.45	0.1	0.234	0.731	0.962	0.831	0.019
gaussian	0.0	3.0	0	0.6	0.1	0.231	0.728	0.997	0.841	0.022
gaussian	0.0	3.0	0.2	0.15	0.1	0.576	0.667	0.004	0.008	0.083
gaussian	0.0	3.0	0.2	0.3	0.1	0.346	0.705	0.407	0.516	0.045
gaussian	0.0	3.0	0.2	0.45	0.1	0.296	0.723	0.817	0.767	0.027
gaussian	0.0	3.0	0.2	0.6	0.1	0.285	0.734	0.972	0.836	0.016
Non-Normality	Skew	Kurt	AR(1)	SR1	P[H1]	SR_c	Precision	Recall	F1	Diff
mild	-0.9	5.6	0	0.15	0.1	0.370	0.600	0.082	0.144	0.150
mild	-0.9	5.6	0	0.3	0.1	0.259	0.679	0.633	0.656	0.071
mild	-0.8	5.5	0	0.45	0.1	0.236	0.654	0.938	0.771	0.096
mild	-0.8	5.4	0	0.6	0.1	0.232	0.688	0.996	0.814	0.062
mild	-0.9	5.6	0.2	0.15	0.1	0.518	0.610	0.035	0.067	0.140
mild	-0.8	5.5	0.2	0.3	0.1	0.343	0.652	0.469	0.545	0.098
mild	-0.9	5.5	0.2	0.45	0.1	0.300	0.676	0.816	0.740	0.074
mild	-0.8	5.5	0.2	0.6	0.1	0.287	0.698	0.957	0.808	0.052
Non-Normality	Skew	Kurt	AR(1)	SR1	P[H1]	SR_c	Precision	Recall	F1	Diff
moderate	-1.7	10.2	0	0.15	0.1	0.353	0.565	0.127	0.208	0.185
moderate	-1.7	10.1	0	0.3	0.1	0.260	0.572	0.605	0.588	0.178
moderate	-1.6	10.0	0	0.45	0.1	0.239	0.623	0.899	0.736	0.127
moderate	-1.6	9.9	0	0.6	0.1	0.233	0.626	0.981	0.765	0.124
moderate	-1.7	10.3	0.2	0.15	0.1	0.485	0.550	0.047	0.087	0.200
moderate	-1.7	10.1	0.2	0.3	0.1	0.341	0.645	0.479	0.550	0.105
moderate	-1.6	10.0	0.2	0.45	0.1	0.303	0.616	0.788	0.691	0.134
moderate	-1.6	9.8	0.2	0.6	0.1	0.290	0.639	0.936	0.759	0.111
Non-Normality	Skew	Kurt	AR(1)	SR1	P[H1]	SR_c	Precision	Recall	F1	Diff
severe	-2.4	16.2	0	0.15	0.1	0.344	0.493	0.154	0.234	0.257
severe	-2.3	16.1	0	0.3	0.1	0.261	0.573	0.621	0.596	0.177
severe	-2.3	15.7	0	0.45	0.1	0.241	0.612	0.904	0.730	0.138
severe	-2.2	15.5	0	0.6	0.1	0.235	0.605	0.974	0.747	0.145
severe	-2.3	16.0	0.2	0.15	0.1	0.467	0.491	0.084	0.143	0.259
severe	-2.3	16.1	0.2	0.3	0.1	0.340	0.561	0.487	0.522	0.189
severe	-2.3	15.8	0.2	0.45	0.1	0.305	0.601	0.775	0.677	0.149
severe	-2.2	15.3	0.2	0.6	0.1	0.293	0.625	0.920	0.744	0.125

Exhibit 7(a) – SFDR control under different processes (monthly frequency, $T = 60$, $q = 0.25$)

Non-Normality	Skew	Kurt	AR(1)	SR1	P[H1]	SR_c	Precision	Recall	F1	Diff
gaussian	0.0	3.0	0	0.15	0.1	0.050	0.701	1.000	0.824	0.049
gaussian	0.0	3.0	0	0.3	0.1	0.050	0.693	1.000	0.818	0.057
gaussian	0.0	3.0	0	0.45	0.1	0.050	0.688	1.000	0.815	0.062
gaussian	0.0	3.0	0	0.6	0.1	0.050	0.703	1.000	0.826	0.047
gaussian	0.0	3.0	0.2	0.15	0.1	0.061	0.702	0.998	0.824	0.048
gaussian	0.0	3.0	0.2	0.3	0.1	0.061	0.714	1.000	0.833	0.036
gaussian	0.0	3.0	0.2	0.45	0.1	0.061	0.711	1.000	0.831	0.039
gaussian	0.0	3.0	0.2	0.6	0.1	0.061	0.723	1.000	0.839	0.027
Non-Normality	Skew	Kurt	AR(1)	SR1	P[H1]	SR_c	Precision	Recall	F1	Diff
mild	-0.9	5.7	0	0.15	0.1	0.050	0.702	1.000	0.825	0.048
mild	-0.9	5.7	0	0.3	0.1	0.050	0.697	1.000	0.821	0.053
mild	-0.9	5.7	0	0.45	0.1	0.050	0.682	1.000	0.811	0.068
mild	-0.9	5.6	0	0.6	0.1	0.050	0.680	1.000	0.810	0.070
mild	-0.9	5.7	0.2	0.15	0.1	0.061	0.718	0.991	0.833	0.032
mild	-0.9	5.7	0.2	0.3	0.1	0.061	0.723	1.000	0.839	0.027
mild	-0.9	5.7	0.2	0.45	0.1	0.061	0.702	1.000	0.825	0.048
mild	-0.9	5.6	0.2	0.6	0.1	0.061	0.696	1.000	0.821	0.054
Non-Normality	Skew	Kurt	AR(1)	SR1	P[H1]	SR_c	Precision	Recall	F1	Diff
moderate	-1.7	10.6	0	0.15	0.1	0.050	0.676	1.000	0.806	0.074
moderate	-1.7	10.5	0	0.3	0.1	0.050	0.669	1.000	0.801	0.081
moderate	-1.7	10.4	0	0.45	0.1	0.050	0.679	1.000	0.809	0.071
moderate	-1.7	10.2	0	0.6	0.1	0.050	0.676	1.000	0.807	0.074
moderate	-1.8	10.6	0.2	0.15	0.1	0.061	0.716	0.990	0.831	0.034
moderate	-1.7	10.5	0.2	0.3	0.1	0.061	0.712	1.000	0.831	0.038
moderate	-1.7	10.4	0.2	0.45	0.1	0.061	0.716	1.000	0.834	0.034
moderate	-1.7	10.2	0.2	0.6	0.1	0.061	0.724	1.000	0.840	0.026
Non-Normality	Skew	Kurt	AR(1)	SR1	P[H1]	SR_c	Precision	Recall	F1	Diff
severe	-2.5	17.1	0	0.15	0.1	0.050	0.674	0.999	0.805	0.076
severe	-2.4	16.9	0	0.3	0.1	0.050	0.664	1.000	0.798	0.086
severe	-2.4	16.7	0	0.45	0.1	0.050	0.663	1.000	0.798	0.087
severe	-2.4	16.4	0	0.6	0.1	0.050	0.675	1.000	0.806	0.075
severe	-2.5	17.0	0.2	0.15	0.1	0.061	0.693	0.987	0.814	0.057
severe	-2.4	16.9	0.2	0.3	0.1	0.061	0.691	1.000	0.817	0.059
severe	-2.4	16.6	0.2	0.45	0.1	0.061	0.720	1.000	0.837	0.030
severe	-2.4	16.3	0.2	0.6	0.1	0.061	0.706	1.000	0.828	0.044

Exhibit 7(b) – SFDR control under different processes (daily frequency, $T = 1,300$, $q = 0.25$)

WHICH MULTIPLE TESTING CONTROL SHOULD BE APPLIED?

It is important to understand that FWER and FDR measure different probabilities, and their respective corrections control for different goals. One correction is not superior to the other, but depending on the context, one is more appropriate than the other.

Controlling for FWER is important in instances where a selected model overrides the rest, making it critical to control the probability that the one selected model is a false positive. The stringency of FWER corrections is warranted by the fact that, should the discovery be false, repercussions may propagate system-wide, impacting a large portion of the investment community. This is the standard situation in academic publishing, scientific discovery, and policy design, where the community relies on a particular finding to the detriment of competing explanations. In the context of finance, FWER corrections are more appropriate in foundational discoveries, like factor models

for risk and investing, valuation models for collateral requirements, monetary and fiscal policy, or microstructural models used for executing orders in central risk books.

Controlling for FDR is important in instances where all models that satisfy a minimum threshold are applied concurrently, therefore the focus is on controlling the percentage of errors (a quality control). The leniency of FDR corrections is warranted by the fact that, should a particular product be faulty, repercussions will be localized, impacting only a small fraction of users or assets. In the context of finance, FDR corrections are more appropriate in industrial applications, like the recruitment of portfolio managers, or the selection and decommissioning of strategies by an investment committee.

A NEW STANDARD FOR SHARPE RATIO INFERENCE

The results of this paper motivate an improved standard for the reporting and application of the Sharpe ratio in both academic research and investment practice. Many of the persistent misuses of the Sharpe ratio do not arise from the metric itself, but from treating it as a standalone point estimate, detached from sampling uncertainty, decision context, and multiplicity of tests that typically accompany its use. The methodological contributions reviewed and developed in this paper naturally assemble into a coherent framework that addresses these shortcomings and clarifies how the Sharpe ratio should be communicated and acted upon.

In view of our findings, we recommend that academics and practitioners follow an improved standard for Sharpe ratio reporting and decision-making. In particular:

1) Avoid ranking / selection based on annualized Sharpe ratio point estimates.

Annualization obscures sampling uncertainty and amplifies inflation arising from short samples, serial dependence, and non-Normal returns. Sharpe ratios should not be compared or ranked solely on point estimates, whether annualized or not.

2) Express Sharpe ratios in probability space.

Inference should be based on statistics such as the Probabilistic Sharpe Ratio (PSR), which assesses false positive rates (type-I error) relative to a set target while explicitly accounting for sample length, higher-order moments, and serial dependence through a generalized sampling variance.

3) Report minimum sample requirements and test power.

Any assessment of statistical significance should be accompanied by the Minimum Track Record Length (MinTRL) and the power of the test. They quantify the risk of making decisions on insufficient evidence and the risk of using underpowered tests that produce too many false negatives (type-II error).

4) Report posterior error probabilities.

Rather than relying solely on classical p -values, practitioners should report the probability that the null hypothesis is true given the evidence, using planned or observed Bayesian false discovery rates (pFDR and oFDR). These quantities directly address common misinterpretations of statistical significance and provide decision-relevant measures of uncertainty.

5) Correct for multiple testing.

When Sharpe ratios are used in procedures involving multiple trials—such as factor discovery, strategy search, or repeated evaluation—reported estimates should be adjusted for multiple testing using appropriate tools, such as the Deflated Sharpe Ratio (DSR) or the Sequential False Discovery Rate (SFDR), depending on whether decisions introduce system-wide or local risks.

Exhibit 8 summarizes these recommendations and contrasts the proposed standard with prevailing practice. Together, these requirements shift Sharpe ratio analysis from *ad hoc* point estimation toward a principled, decision-theoretic framework that makes uncertainty, error trade-offs, and multiplicity explicit.

Share Ratio Use	Current Standard	New Standard
Comparison & selection	Annualized Sharpe ratio	Probabilistic Sharpe ratio (PSR)
Estimation uncertainty	Often ignored	Explicitly quantified, reported
Sampling variance	Assumes i.i.d. Normal returns	Generalized variance, under non-Normal and AR(1) returns
Control for Type I Error	Confidence bands, p-value	Report PSR and MinTRL
Control for Type II Error / Recall	Often ignored	Report Power
Posterior Error / Precision	Often ignored	Report pFDR, oFDR
Control for Multiple Testing	Almost always ignored	Apply DSR, SFDR controls

Exhibit 8 - An improved standard for Sharpe ratio reporting

Different measurements and null hypotheses should be applied and reported, depending on which stage of the investment process the strategy is in. For example, after discovering a market anomaly, researchers may assess its risk premium through a backtest. Given a Sharpe ratio backtested on a discovery set \widehat{SR}_D^* , a PSR where $P[\widehat{SR} < \widehat{SR}_D^* | SR = 0] \gg 0.5$ is consistent with a statistically significant premium (i.e., distinguishable from noise). Before a strategy's approval, it is useful to assess whether the discovery backtest's performance is replicable in a sample unseen by researchers, called embargo.²⁷ Given a Sharpe ratio backtested on an embargo set \widehat{SR}_E^* , a PSR where $P[\widehat{SR} < \widehat{SR}_E^* | SR = \widehat{SR}_D^*] \ll 0.5$ is consistent with a structural break in the strategy's embargo performance, likely due to overfitting. During ramp-up, it is useful to corroborate that live performance is consistent with the simulated performance from the discovery and embargo periods. Given a live Sharpe ratio \widehat{SR}_L^* , and a Sharpe ratio \widehat{SR}_{D+E}^* covering discovery and live periods, a PSR where $P[\widehat{SR} < \widehat{SR}_L^* | SR = \widehat{SR}_{D+E}^*] \ll 0.5$ is consistent with a structural break in the strategy's live performance, likely due to data or execution issues incorrectly modeled in simulations. Finally, during full deployment, it is useful to monitor performance to identify alpha decay. Given an out-of-sample Sharpe ratio (covering embargo and live periods) \widehat{SR}_{L+E}^* , a PSR where $P[\widehat{SR} < \widehat{SR}_{L+E}^* | SR = \widehat{SR}_D^*] \ll 0.5$ is consistent with significant alpha decay relative to

²⁷ In this context, we denote by discovery backtest the performance simulated on the same sample used by a researcher to identify an investment opportunity and develop a strategy. This differs from the performance simulated on a sample that was withheld from researchers during strategy design, called embargo backtest. The minimum sample length for both backtests (discovery and embargo) is determined by MinTRL.

simulated performance. Should $P[\widehat{SR} < \widehat{SR}_{L+E}^* | SR = 0] \ll 0.5$, the strategy may be decommissioned.

Stage of Lifecycle	Main Decision	Inference Tool						
		PSR	MinTRL	Power	pFDR	oFDR	DSR	SFDR
Discovery	Is this pattern signal or noise?	Primary	Useful	Primary	Useful	Useful	Useful	Useful
Deflation	Is this discovery real after K trials?	Useful	Primary	Primary	Rare	Primary	Primary	Rare
Validation	Is Embargo performance consistent with discovery?	Primary	Primary	Useful	Rare	Primary	Rare	Rare
Allocation	Control long-run posterior error across approvals	Useful	Useful	Useful	Primary	Useful	Useful	Primary
Live Testing	Is execution/data degrading the signal?	Primary	Primary	Useful	Rare	Primary	Rare	Rare
Full Deployment	Is alpha decaying? decommission?	Primary	Rare	Rare	Rare	Primary	Rare	Rare

Exhibit 9 – Relative importance of inference tools per stage in the strategy’s lifecycle

Exhibit 9 summarizes how inference and multiple-testing controls should be aligned with the stages of a strategy’s lifecycle, in the same sequence in which decisions are made. First, during the discovery stage, the primary question is whether an observed pattern is signal or noise, which motivates the use of PSR-based inference together with ex-ante measures of evidentiary adequacy such as power and MinTRL. Second, during the deflation stage, the question shifts to whether the discovery remains real after accounting for selection bias under multiple testing. This motivates applying multiple-testing corrections such as DSR and interpreting the evidence through posterior error probabilities such as oFDR. Third, during the validation stage, the relevant question is whether embargo performance is consistent with the discovery performance, for which PSR, MinTRL, and oFDR provide decision-relevant diagnostics under realistic return dynamics. Fourth, during the allocation stage, the objective is to control long-run posterior error across approvals, which motivates SFDR as a process-level control, complemented by planned error measures such as pFDR. Fifth, during the live testing stage, the focus becomes whether execution or data issues are degrading the signal, again calling for PSR, MinTRL, and oFDR. Sixth, during the full deployment stage, inference shifts to detecting alpha decay and supporting decommissioning decisions, where PSR (e.g., against a zero-efficiency null) and oFDR remain the most informative tools.

Importantly, this standard does not call for abandoning the Sharpe ratio. On the contrary, it provides a coherent way to interpret and communicate what the Sharpe ratio does and does not imply under realistic return dynamics, finite samples, and decision settings. By making estimation uncertainty, power, posterior error probabilities, and multiple testing explicit, the proposed standard improves transparency, comparability, and accountability in both empirical finance and investment governance.

CONCLUSIONS

The Sharpe ratio remains the most widely used measure of investment efficiency, yet its naive application leads to misleading inference and financial losses. This paper has shown that valid inference requires addressing five key pitfalls: (a) neglect of statistical significance; (b) biased inference due to wrongly assuming that returns are i.i.d. Normal; (c) insufficient test power, and lack of assessment of minimum sample length requirements; (d) the confusion between classical p -values and the probability of the null hypothesis given the evidence; and (e) failure to correct for multiple testing.

To address these shortcomings, we have reviewed several statistical frameworks, outlined in Exhibit 10. Sharpe-specific methods derive inference directly from the Sharpe ratio's sampling distribution, accounting for finite-sample effects, non-Normality, and serial dependence. Non-Sharpe-specific methods apply generic test statistics (typically t -statistics that require a Sharpe-to- t conversion) that do not operate on the Sharpe ratio's sampling distribution.

The Probabilistic Sharpe Ratio (PSR) expresses observed Sharpe ratios in probability space, adjusting for skewness, kurtosis, serial correlation, and sample length. The Minimum Track Record Length (MinTRL) quantifies the data requirements to achieve meaningful inference. Bayesian false discovery rate methods estimate the posterior probability that a strategy is false, conditioning respectively on the test's rejection region (pFDR) and on the observed test statistic/ p -value (oFDR); these posterior error probabilities are important quantities that p -values do not reflect. The Deflated Sharpe Ratio (DSR) is a parametric, extreme-value-based, selection-aware test designed to control the false positive probability of deploying the single best strategy after search. It corrects for selection bias and backtest overfitting by accounting for the number and correlation of trials. The sequential false discovery rate (SFDR) procedure offers a decision-theoretic framework to control for the proportion of false strategies among those sequentially approved by investment committees. DSR and SFDR are Sharpe-specific multiple testing controls that are better aligned with investment practice than general-purpose methods.

Monte Carlo experiments confirm that the methods presented in this paper provide more reliable inference than classical t -tests and general-purpose multiple-testing corrections under serially-correlated non-Normal returns. Realistic scenarios show that the actual variance of the Sharpe ratio can be four or more times larger than its estimate under the i.i.d. Normal assumption. PSR yields high precision and recall, while DSR and SFDR achieve their targeted probabilities (respectively, FWER and the pFDR). The choice between FWER and FDR corrections depends on the context: FWER is more appropriate in settings where a single discovery supersedes the rest, and FDR is better suited for settings where competing discoveries are deployed simultaneously.

Beyond its methodological and governance contributions, this paper articulates an improved standard for Sharpe ratio reporting and decision-making. Rather than treating the Sharpe ratio as a standalone point estimate, the proposed standard requires explicit communication of estimation uncertainty, sampling variance under realistic return dynamics, statistical power, posterior error probabilities, and appropriate corrections for multiple testing. This framework integrates frequentist inference, Bayesian decision theory, Extreme Value Theory, machine learning, and practical investment governance into a coherent set of reporting principles. Adoption of this

standard would improve transparency, reduce false discoveries, and align Sharpe ratio–based decisions with the sequential and high-stakes nature of real-world investment processes.

Method	Authors	Correction Type	Sharpe Specific?	Notes
Lo's Significance Test	Lo [2002]	Single-test inference	Yes	Adjusts for sample length, under Normal returns
Bootstrap Test	Ledoit & Wolf [2008]	Single-test inference	Yes	HAC standard errors and a studentized time-series bootstrap
Probabilistic Sharpe Ratio (PSR)	Bailey & López de Prado [2012]	Single-test inference	Yes	Adjusts for skewness, kurtosis, sample length
Minimum Track Record Length (MinTRL)	Bailey & López de Prado [2012]	Sample size adequacy	Yes	Computes required minimum observations needed to reject the null hypothesis
Sharpe Ratio Efficient Frontier	Bailey & López de Prado [2012]	Portfolio optimization framework	Yes	Extends Sharpe ratio to efficient frontier under non-Normality
Generalized Variance of the Sharpe Ratio	López de Prado, Lipton & Zoonekynd [2025]	Single-test inference	Yes	Variance of the Sharpe ratio's estimator under non-Normal & AR(1) returns
Reality Check	White [2000]	FWER	Adapted	Bootstrap test against best-performing strategy
SPA Test	Hansen [2005]	FWER	Adapted	Improves on Reality Check; less conservative
Stepdown Resampling	Romano & Wolf [2005, 2016]	FWER	Adapted	Resampling-based multiple testing correction
Deflated Sharpe Ratio (DSR)	Bailey & López de Prado [2014]	FWER	Yes	Corrects for non-Normality, sample length and multiple testing
Bonferroni [1936] and Holm [1979] tests	Harvey & Liu [2015]	FWER	Adapted	Applied classical FWER corrections to the Sharpe ratio
Combinatorial Purged Cross-Validation (CPCV)	López de Prado [2018]	FWER	Adapted	Bootstrapping of Sharpe ratio's distribution under different scenarios
Power of the Sharpe Ratio	López de Prado [2020]	FWER	Yes	Computes the type-II error associated with a Sharpe ratio rejection threshold
False Discovery Probability Control	Romano, Shaikh and Wolf [2008]	FDP (Frequentist)	Adapted	General purpose batch multiple-testing FDP control
Benjamini-Yekutieli [2001] tests	Harvey & Liu [2020]	FDR (Frequentist)	Adapted	Benjamini-Hochberg-Yekutieli FDR control applied to the Sharpe ratio
Efron [2004] test	Harvey & Liu [2020]	FDR (Frequentist)	Adapted	Efron-style bootstrap Sharpe hurdle linked to false positives
Efron [2008] test	Harvey, Sancetta & Zhao [2025]	FDR (Bayesian)	Adapted	Efron-style local FDR test, with cross-sectional correlation and unknown number of tests
Bayesian oFDR / pFDR	López de Prado, Lipton & Zoonekynd [2025]	FDR (Bayesian)	Yes	Bayesian tail-area FDR, under serially-correlated non-Normal returns
Sequential FDR	López de Prado, Lipton & Zoonekynd [2025]	FDR (Bayesian)	Yes	Control for the posterior probability of error in each individually approved strategy

Exhibit 10 – Methods applied to conduct inference on the Sharpe ratio

For every parameter that determines the variance of the Sharpe ratio, hedge fund indices feature the trait that increases rather than decreases that variance: negative skewness, rather than positive; positive excess kurtosis, rather than negative; positive serial correlation, rather than negative; shorter sample length, rather than longer. Higher variance means greater chance of conflating luck with skill, and greater potential for losses. These characteristics are less prevalent in other investments where Sharpe ratios tend to be lower or are not used as the primary criterion for ranking and selection, such as factor strategies, mutual funds, and passive investments. We therefore conjecture that the widespread misuse of the Sharpe ratio for ranking and selection has instituted an adverse selection mechanism, favoring strategies whose return-generating processes inflate Sharpe ratios. The very traits that inflate the Sharpe ratio and therefore elevate those strategies in the rankings are also the traits that make those strategies dangerous. If investors ranked and selected investments using PSR instead of annualized Sharpe ratios, they would not favor hedge funds with inflated performance. We hope that adoption of the standards presented in this paper will contribute to the application of improved controls among investors, such that strategies with hazardous traits are properly penalized, and the aforementioned adverse selection is mitigated.

Financial applications of the mathematical findings presented in this paper include: (i) improved tools for filtering out false strategies prior to portfolio construction and allocation decisions; (ii) improved estimation of the Sharpe-efficient frontier (Bailey and López de Prado [2012]), i.e., the frontier comprising portfolios that maximize expected Sharpe ratio at each level of Sharpe ratio estimation uncertainty; (iii) the development of a new family of robust portfolio optimization frameworks in which inferential measures—such as PSR and oFDR, as well as DSR, MinTRL, and uncertainty-adjusted Sharpe thresholds—can operate as objectives, targets, or as soft/hard constraints within allocation programs (Jacquier et al. [2022]). We expect application (iii), in particular, to motivate further research and attract increased attention from investors in the coming years.

Beyond finance, the findings in this paper are relevant to a broad range of scientific and engineering settings where inference is conducted on a signal-to-noise measure—an estimated mean effect divided by an estimated dispersion—under conditions that violate the i.i.d. Normal paradigm. In economics, observations are typically serially and cross-sectionally dependent due to persistence, aggregation, and shared shocks; in engineering and industrial processes, measurements are autocorrelated because systems exhibit inertia, drift, and feedback control; in physics and chemistry, experimental signals are often contaminated by correlated noise, instrument response, and nonstationary backgrounds; and in biomedical and materials sciences, data commonly display heavy tails, outliers, and mixture behavior driven by heterogeneity across samples, devices, or environments. In all these domains, the i.i.d. Normal assumption is therefore not a harmless simplification but an empirically fragile modeling convenience, and it can lead to substantial underestimation of uncertainty when used for inference on efficiency metrics. Moreover, discovery is rarely based on a single test: researchers typically screen many candidate models, designs, molecules, or control policies, and proceed with the most promising ones, which introduces a universal selection bias (winner’s curse) analogous to multiple strategy backtesting. The generalized sampling theory and multiple-testing controls developed in this paper address

these ubiquitous departures from idealized assumptions, enabling closed-form, decision-relevant inference that remains valid under non-Normality, serial dependence, and selection effects.

Ultimately, the Sharpe ratio remains a valuable tool only if properly adjusted, reported and interpreted. Researchers and practitioners must move beyond raw point estimates, incorporating corrections for non-Normality, serial correlation, small samples, and multiple testing. Failure to do so turns the Sharpe ratio from a measure of efficiency into a systematic source of error. By applying the framework laid out in this paper, investors can control more effectively for false positives and false negatives, reduce backtest overfitting, improve the efficiency of their portfolios, and establish a more rigorous standard for performance assessment and reporting.

ACKNOWLEDGEMENTS

We are especially grateful to Andrew Lo (MIT) and Michael Wolf (University of Zürich), whose 2003 letters to the *Financial Analysts Journal* defined the challenge we addressed in this paper. We thank fellow ADIA Lab board members Robert F. Engle and Guido W. Imbens for their comments and collegial support. We also thank our ADIA colleagues, especially Koushik Balasubramanian, Illya Barziy, Walter Distaso, Jacques Joubert, Dmitri Maksarov, Dragan Sestovic, and Blaz Zlicar. We are grateful for useful comments provided by Patrick Cheridito (ETH Zürich), Frank Fabozzi (EDHEC), Campbell Harvey (Duke University), Alessia López de Prado Rehder (University of Zürich), Emilio Porcu (Khalifa University), Riccardo Rebonato (EDHEC), Alessio Sancetta (Royal Holloway, University of London), Luis Seco (University of Toronto), Horst Simon (ADIA Lab), Josef Teichmann (ETH Zürich), and Jorge Zubelli (Khalifa University). The views expressed in this paper are the authors', and do not necessarily represent the opinions of the organizations they are affiliated with.

APPENDIX

A.1. DISTRIBUTION OF THE SHARPE RATIO UNDER NON-NORMAL AND SERIALY CORRELATED RETURNS

Consider a stationary, ergodic and weakly dependent series of T excess returns, $\{r_t\}_{t=1,\dots,T}$, with a finite fourth moment. We define the following notation:

$$\begin{aligned}
 \mu &= E[r_t] \\
 x_t &= r_t - \mu \\
 \sigma^2 &= E[x_t^2] \\
 v_3 &= E[x_t^3] \\
 v_4 &= E[x_t^4] \\
 \gamma_3 &= v_3/\sigma^3 \\
 \gamma_4 &= v_4/\sigma^4 \\
 \rho &= \text{Cor}[x_t, x_{t+1}] \\
 SR &= \mu/\sigma
 \end{aligned} \tag{34}$$

For notational convenience, let us write $r_T \stackrel{a}{\sim} \mathcal{N}\left[\theta, \frac{1}{T}\Sigma\right]$ to mean

$$\sqrt{T}(r_T - \theta) \xrightarrow{d} \mathcal{N}[0, \Sigma] \tag{35}$$

The delta method states that, if $r_T \stackrel{a}{\sim} \mathcal{N}\left[\theta, \frac{1}{T}\Sigma\right]$, then $f[r_T] \stackrel{a}{\sim} \mathcal{N}\left[f[\theta], \frac{1}{T}\dot{f}[\theta]\Sigma\dot{f}[\theta]'\right]$, where $\dot{f}[\theta]$ is the Jacobian of f at θ , and $\dot{f}[\theta]'$ is its transpose matrix.²⁸ In particular, if $r_t \stackrel{iid}{\sim} \mathcal{N}[\mu, \sigma^2]$, we know that

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} \stackrel{a}{\sim} \mathcal{N}\left[\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \frac{1}{T}\begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right] \tag{36}$$

and we can apply the delta method to

$$\begin{aligned}
 f[a, b] &= a/\sqrt{b} \\
 \dot{f}[a, b] &= \left(\frac{1}{\sqrt{b}}, -\frac{1}{2}\frac{a}{b^{3/2}}\right)
 \end{aligned} \tag{37}$$

yields the asymptotic behavior of the ratio $\hat{\mu}/\hat{\sigma}$ under i.i.d. Normal returns,

$$V\left[\frac{\hat{\mu}}{\hat{\sigma}}\right] = \frac{1}{T}\begin{pmatrix} \frac{1}{\sigma} & -\frac{1}{2}\frac{\mu}{\sigma^3} \end{pmatrix} \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma} \\ -\frac{\mu}{2\sigma^3} \end{pmatrix} = \frac{1}{T}\left(1 + \frac{1}{2}\frac{\mu^2}{\sigma^2}\right) \tag{38}$$

²⁸ We assume that f is differentiable at θ and that $\dot{f}[\theta]$ is of full rank.

$$\frac{\hat{\mu}}{\hat{\sigma}} \stackrel{a}{\sim} \mathcal{N} \left[\frac{\mu}{\sigma}, \frac{1}{T} \left(1 + \frac{1}{2} \frac{\mu^2}{\sigma^2} \right) \right]$$

This was the key contribution in Lo [2002]. We wish to extend and generalize this result by dropping the joint assumptions of returns independence and Normality. In particular, the generalized closed-form solution should account for $\gamma_3 \neq 0$, $\gamma_4 \neq 3$ and $\rho \neq 0$. We can estimate $\theta = (\mu, \sigma^2)'$ using the generalized method of moments (GMM), from the following moments,

$$\phi_t = \begin{pmatrix} r_t - \mu \\ (r_t - \mu)^2 - \sigma^2 \end{pmatrix} \quad (39)$$

Under standard regularity conditions (stationarity, weak dependence, existence of moments), the GMM estimator $\hat{\theta}$ satisfies

$$\hat{\theta} \stackrel{a}{\sim} \mathcal{N} \left[\theta, \frac{1}{T} H^{-1} \Sigma (H^{-1})' \right] \quad (40)$$

where

$$\begin{aligned} H &= \lim_{T \rightarrow \infty} E \left[\frac{1}{T} \sum_{t=1}^T \frac{\partial \phi_t}{\partial \theta} \right] \\ \Sigma &= \lim_{T \rightarrow \infty} E \left[\frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T \phi_t \phi_s' \right] \end{aligned} \quad (41)$$

The Jacobian can be computed as

$$\begin{aligned} H &= \lim_{T \rightarrow \infty} E \left[\frac{1}{T} \sum_{t=1}^T \frac{\partial \phi_t}{\partial \theta} \right] = \lim_{T \rightarrow \infty} E \left[\frac{1}{T} \sum_{t=1}^T \begin{pmatrix} -1 & 0 \\ -2(r_t - \mu) & -1 \end{pmatrix} \right] \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \begin{pmatrix} -1 & 0 \\ -2E[r_t - \mu] & -1 \end{pmatrix} = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \\ &= -I \end{aligned} \quad (42)$$

The variance can be computed as follows (using $x_t = r_t - \mu$, to simplify formulas),

$$\begin{aligned} \Sigma &= \lim_{T \rightarrow \infty} E \left[\frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T \phi_t \phi_s' \right] \\ &= \lim_{T \rightarrow \infty} E \left[\frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T \begin{pmatrix} (r_t - \mu)(r_s - \mu) & (r_t - \mu)[(r_s - \mu)^2 - \sigma^2] \\ [(r_t - \mu)^2 - \sigma^2](r_s - \mu) & [(r_t - \mu)^2 - \sigma^2][(r_s - \mu)^2 - \sigma^2] \end{pmatrix} \right] \\ &= \lim_{T \rightarrow \infty} E \left[\frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T \begin{pmatrix} x_t x_s & x_t (x_s^2 - \sigma^2) \\ (x_t^2 - \sigma^2) x_s & (x_t^2 - \sigma^2)(x_s^2 - \sigma^2) \end{pmatrix} \right] \end{aligned} \quad (43)$$

$$\begin{aligned}
&= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T \begin{pmatrix} E[x_t x_s] & E[x_t(x_s^2 - \sigma^2)] \\ E[(x_t^2 - \sigma^2)x_s] & E[(x_t^2 - \sigma^2)(x_s^2 - \sigma^2)] \end{pmatrix} \\
&= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T \begin{pmatrix} E[x_t x_s] & E[x_t x_s^2] - E[x_t] \sigma^2 \\ E[x_t^2 x_s] - E[x_s] \sigma^2 & E[x_t^2 x_s^2] - \sigma^2 E[x_t^2] - \sigma^2 E[x_s^2] + \sigma^4 \end{pmatrix} \\
&= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T \begin{pmatrix} E[x_t x_s] & E[x_t x_s^2] \\ E[x_t^2 x_s] & E[x_t^2 x_s^2] - \sigma^4 \end{pmatrix}
\end{aligned}$$

Note that this involves not only the auto-covariances $E[x_t x_s]$, but also the co-skewness $E[x_t x_s^2]$ and the co-kurtosis $E[x_t^2 x_s^2]$. So far, our use of the GMM did not make any assumptions (except for the existence of the moments). At this point, we introduce the assumption that $\{x_t\}_{t=1, \dots, T}$ follows a stationary AR(1) process, i.e.

$$x_t = \rho x_{t-1} + \varepsilon_t \quad (44)$$

where $\{\varepsilon_t\}_{t=1, \dots, T}$ is i.i.d., with zero mean, and $\rho \in (-1, 1)$. We need to compute the relation between the variance of x_t and the variance of ε_t :

$$E[x_t^2] = E[(\rho x_{t-1} + \varepsilon_t)^2] = \rho^2 E[x_{t-1}^2] + E[\varepsilon_t^2] = \rho^2 E[x_t^2] + E[\varepsilon_t^2] \quad (45)$$

resulting in

$$E[\varepsilon_t^2] = (1 - \rho^2) E[x_t^2] \quad (46)$$

With this AR(1) assumption, we can now compute the required higher moments. For $h \geq 0$,

$$E[x_t x_{t+h}] = E[x_t (\rho^h x_t + \varepsilon_{t+h} + \rho \varepsilon_{t+h-1} + \dots + \rho^{h-1} \varepsilon_{t+1})] = \rho^h E[x_t^2] \quad (47)$$

$$E[x_t x_{t+h}^2] = E[x_t (\rho^h x_t + \varepsilon_{t+h} + \rho \varepsilon_{t+h-1} + \dots + \rho^{h-1} \varepsilon_{t+1})^2] = \rho^{2h} E[x_t^3] \quad (48)$$

$$E[x_t^2 x_{t+h}] = E[x_t^2 (\rho^h x_t + \varepsilon_{t+h} + \rho \varepsilon_{t+h-1} + \dots + \rho^{h-1} \varepsilon_{t+1})] = \rho^h E[x_t^3] \quad (49)$$

$$\begin{aligned}
E[x_t^2 x_{t+h}^2] &= E[x_t^2 (\rho^h x_t + \varepsilon_{t+h} + \rho \varepsilon_{t+h-1} + \dots + \rho^{h-1} \varepsilon_{t+1})^2] \\
&= \rho^{2h} E[x_t^4] + E[x_t^2] E[(1 + \rho^2 + \dots + \rho^{2(h-1)}) \varepsilon_t^2] \\
&= \rho^{2h} E[x_t^4] + \frac{1 - \rho^{2h}}{1 - \rho^2} E[x_t^2] E[\varepsilon_t^2] \\
&= \rho^{2h} E[x_t^4] + \frac{1 - \rho^{2h}}{1 - \rho^2} E[x_t^2] (1 - \rho^2) E[x_t^2] \\
&= \rho^{2h} E[x_t^4] + (1 - \rho^{2h}) (E[x_t^2])^2 = \rho^{2h} v_4 + (1 - \rho^{2h}) \sigma^4 \\
&= \sigma^4 + \rho^{2h} (v_4 - \sigma^4)
\end{aligned} \quad (50)$$

We now need to compute the limit of the average of those higher moments. In the course of the computations, we will use the following result

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T t \rho^t = \lim_{T \rightarrow \infty} \rho \left(\frac{1 - \rho^T}{T(1 - \rho)^2} - \frac{\rho^T}{1 - \rho} \right) = 0 \quad (51)$$

because, since $|\rho| < 1$, the series $\sum t \rho^t$ is absolutely convergent. We can now compute the limit of the average of those higher moments. For the auto-covariance,

$$\Sigma_{1,1} = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T \rho^{|t-s|} \sigma^2 \quad (52)$$

We can decompose the sum along the diagonals,

$$\begin{aligned} \sum_{t=1}^T \sum_{s=1}^T \rho^{|t-s|} &= T + 2(T-1)\rho + 2(T-2)\rho^2 + \dots + 2\rho^{T-1} \\ &= T + 2 \sum_{t=1}^{T-1} (T-t)\rho^t = T + 2T \sum_{t=1}^{T-1} \rho^t - 2 \sum_{t=1}^{T-1} t \rho^t \\ &= T + 2T\rho \frac{1 - \rho^{T-1}}{1 - \rho} - 2 \sum_{t=1}^{T-1} t \rho^t = T \frac{1 + \rho}{1 - \rho} - \frac{2\rho(1 - \rho^T)}{(1 - \rho)^2} \end{aligned} \quad (53)$$

This gives

$$\begin{aligned} \Sigma_{1,1} &= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T \rho^{|t-s|} \sigma^2 = \lim_{T \rightarrow \infty} \frac{1}{T} \left(T \frac{1 + \rho}{1 - \rho} - \frac{2\rho(1 - \rho^T)}{(1 - \rho)^2} \right) \sigma^2 \\ &= \frac{1 + \rho}{1 - \rho} \sigma^2 \end{aligned} \quad (54)$$

The computation for the co-kurtosis is similar, with ρ^2 instead of ρ :

$$\begin{aligned} \Sigma_{2,2} &= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T \rho^{2|t-s|} (v_4 - \sigma^4) \\ &= (v_4 - \sigma^4) \lim_{T \rightarrow \infty} \left(\frac{1 + \rho^2}{1 - \rho^2} - \frac{2\rho^2(1 - \rho^{2T})}{T(1 - \rho^2)^2} \right) \\ &= \frac{1 + \rho^2}{1 - \rho^2} (v_4 - \sigma^4) \end{aligned} \quad (55)$$

For the co-skewness, we can decompose the sum into three parts: the diagonal, the upper triangular part, and the lower triangular part.

$$\begin{aligned}
\Sigma_{1,2} &= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T E[x_t x_s^2] \\
&= \lim_{T \rightarrow \infty} \frac{1}{T} \left(\sum_{t=s} E[x_t x_s^2] + \sum_{t>s} E[x_t x_s^2] + \sum_{t<s} E[x_t x_s^2] \right) \\
&= \lim_{T \rightarrow \infty} \frac{1}{T} \left(T v_3 + \sum_{h=1}^{T-1} (T-h) \rho^h v_3 + \sum_{h=1}^{T-1} (T-h) \rho^{2h} v_3 \right) \\
&= v_3 \lim_{T \rightarrow \infty} \left[\left(1 + \frac{\rho}{1-\rho} + \frac{\rho^2}{1-\rho^2} \right) - \frac{1}{T} \left(\frac{\rho}{(1-\rho)^2} + \frac{\rho^2}{(1-\rho^2)^2} \right) \right. \\
&\quad \left. + \frac{1}{T} \left(\frac{\rho^{T+1}}{(1-\rho)^2} + \frac{\rho^{2(T+1)}}{(1-\rho^2)^2} \right) \right] = \frac{1+\rho+\rho^2}{1-\rho^2} v_3
\end{aligned} \tag{56}$$

For an AR(1) process, the GMM estimators of the mean and variance are asymptotically Normal,

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} \overset{a}{\sim} \mathcal{N} \left(\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \frac{1}{T} \begin{pmatrix} \frac{1+\rho}{1-\rho} \sigma^2 & \frac{1+\rho+\rho^2}{1-\rho^2} v_3 \\ \frac{1+\rho+\rho^2}{1-\rho^2} v_3 & \frac{1+\rho^2}{1-\rho^2} (v_4 - \sigma^4) \end{pmatrix} \right) \tag{57}$$

For compactness, we compute the scaled variance $V[\widehat{SR}]T$ under AR(1) returns and first four moments as

$$\begin{aligned}
&V[\widehat{SR}]T \\
&= \begin{pmatrix} \frac{1}{\sigma} \\ -\frac{\mu}{2\sigma^3} \end{pmatrix} \begin{pmatrix} \left(1 + \frac{2\rho}{1-\rho}\right) \sigma^2 & \left(1 + \frac{\rho}{1-\rho} + \frac{\rho^2}{1-\rho^2}\right) v_3 \\ \left(1 + \frac{\rho}{1-\rho} + \frac{\rho^2}{1-\rho^2}\right) v_3 & \left(1 + \frac{2\rho^2}{1-\rho^2}\right) (v_4 - \sigma^4) \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma} \\ -\frac{\mu}{2\sigma^3} \end{pmatrix} \\
&= \frac{1}{\sigma^2} \left(1 + \frac{2\rho}{1-\rho}\right) \sigma^2 + 2 \left(1 + \frac{\rho}{1-\rho} + \frac{\rho^2}{1-\rho^2}\right) v_3 \frac{1-\mu}{\sigma^2 \sigma^3} \\
&\quad + \left(1 + \frac{2\rho^2}{1-\rho^2}\right) (v_4 - \sigma^4) \left(\frac{-\mu}{2\sigma^3}\right)^2 \\
&= \left(1 + \frac{2\rho}{1-\rho}\right) - \left(1 + \frac{\rho}{1-\rho} + \frac{\rho^2}{1-\rho^2}\right) \frac{v_3 \mu}{\sigma^3 \sigma} + \left(1 + \frac{2\rho^2}{1-\rho^2}\right) \frac{v_4 - \sigma^4}{4\sigma^4} \left(\frac{\mu}{\sigma}\right)^2 \\
&= \left(1 + \frac{2\rho}{1-\rho}\right) - \left(1 + \frac{\rho}{1-\rho} + \frac{\rho^2}{1-\rho^2}\right) \gamma_3 SR + \left(1 + \frac{2\rho^2}{1-\rho^2}\right) \frac{\gamma_4 - 1}{4} SR^2 \\
&\quad = \frac{1+\rho}{1-\rho} - \frac{1+\rho+\rho^2}{1-\rho^2} \gamma_3 SR + \frac{1+\rho^2}{1-\rho^2} \frac{\gamma_4 - 1}{4} SR^2
\end{aligned} \tag{58}$$

This concludes the proof.

A.2. EXTENDING LO [2002] TO SERIALY CORRELATED RETURNS

Lo [2002] did not provide a closed-form solution for the sampling variance of the Sharpe ratio under serially correlated returns. Instead, that paper derived a “Time Aggregation” scaling factor when Normal returns are serially correlated. To see the difference more clearly, let us set $(v_3, v_4) = (0, 3\sigma^4)$ in equation (57) to reflect the Normal case, resulting in

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} \overset{a}{\sim} \mathcal{N} \left[\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \frac{1}{T} \begin{pmatrix} \frac{1+\rho}{1-\rho} \sigma^2 & 0 \\ 0 & 2 \frac{1+\rho^2}{1-\rho^2} \sigma^4 \end{pmatrix} \right] \quad (59)$$

This expression gives us the long-run covariance matrix of $(\hat{\mu}, \hat{\sigma}^2)'$ under Normal AR(1) dependence. Applying the functional delta method yields,

$$\begin{pmatrix} \hat{\mu} \\ \frac{\hat{\mu}}{\hat{\sigma}} \end{pmatrix} \overset{a}{\sim} \mathcal{N} \left[\begin{pmatrix} \mu \\ \frac{\mu}{\sigma} \end{pmatrix}, \frac{1}{T} \begin{pmatrix} \frac{1+\rho}{1-\rho} & \frac{1}{2} \frac{1+\rho^2}{1-\rho^2} \frac{\mu^2}{\sigma^2} \\ \frac{1}{2} \frac{1+\rho^2}{1-\rho^2} \frac{\mu^2}{\sigma^2} & \frac{1+\rho^2}{1-\rho^2} \frac{\mu^2}{\sigma^2} \end{pmatrix} \right] \quad (60)$$

Compare our equation (60) with Lo’s equation (22) for “Time Aggregation.” The solution to the sampling variance problem is not a simple sample size adjustment: the variance in the i.i.d. Normal case has two terms, 1 and SR^2 , and they are adjusted differently,

$$\begin{aligned} T_1^{eff} &= \frac{1-\rho}{1+\rho} T \\ T_{SR^2}^{eff} &= \frac{1-\rho^2}{1+\rho^2} T \end{aligned} \quad (61)$$

We hope that this derivation clarifies and settles the apparent confusion among practitioners and academics regarding the assumptions and derivations in Lo [2002].

A.3. EXPERIMENTAL VALIDATION OF EFFECTIVE NUMBER OF TRIALS

In practice, trials are rarely independent. The “effective number of trials” K is defined as the number of approximately independent latent sources of variation underlying a collection of correlated strategy backtests, where each source contributes independently to the dispersion and extreme values of the observed Sharpe ratio estimates.

We propose three practical approaches to estimate K . The first approach follows López de Prado [2019]: (i) compute the correlation matrix of the return time series across all trials; (ii) cluster this matrix and determine the optimal number of clusters by maximizing the t-statistic of the mean Silhouette score; (iii) form one representative return series per cluster as a minimum-variance weighted combination of its constituents (this prevents that the most volatile trials dominate); (iv) using those derived time series, compute one Sharpe ratio per cluster; (v) treat the number of clusters as an estimate of the effective number of independent trials K , since each cluster represents

a largely distinct independent source of variation contributing to the distribution of extremes. As a sanity check, compute the cross-sectional variance of Sharpe ratios derived in step (iv), and verify that it is of comparable order of magnitude to that observed across the full set of correlated trials.

The second approach is inspired by López de Prado [2020]: (i) compute the correlation matrix of the time series of returns from all trials; (ii) fit the Marchenko-Pastur distribution, and count the eigenvalues of the correlation matrix that exceed the distribution's upper bound; (iii) remove those non-trivial eigenvalues, and iterate the previous steps until there are no more eigenvalues beyond the limit; (iv) estimate K as the number of removed non-trivial eigenvalues.

The third approach follows López de Prado [2018, section 18.7], which derives K as the effective rank (also known as eigenvalue entropy) of the correlation matrix of the time series of returns from all trials. This approach is intended to produce an upper bound for K , as it will likely exaggerate its value for the reason explained next.

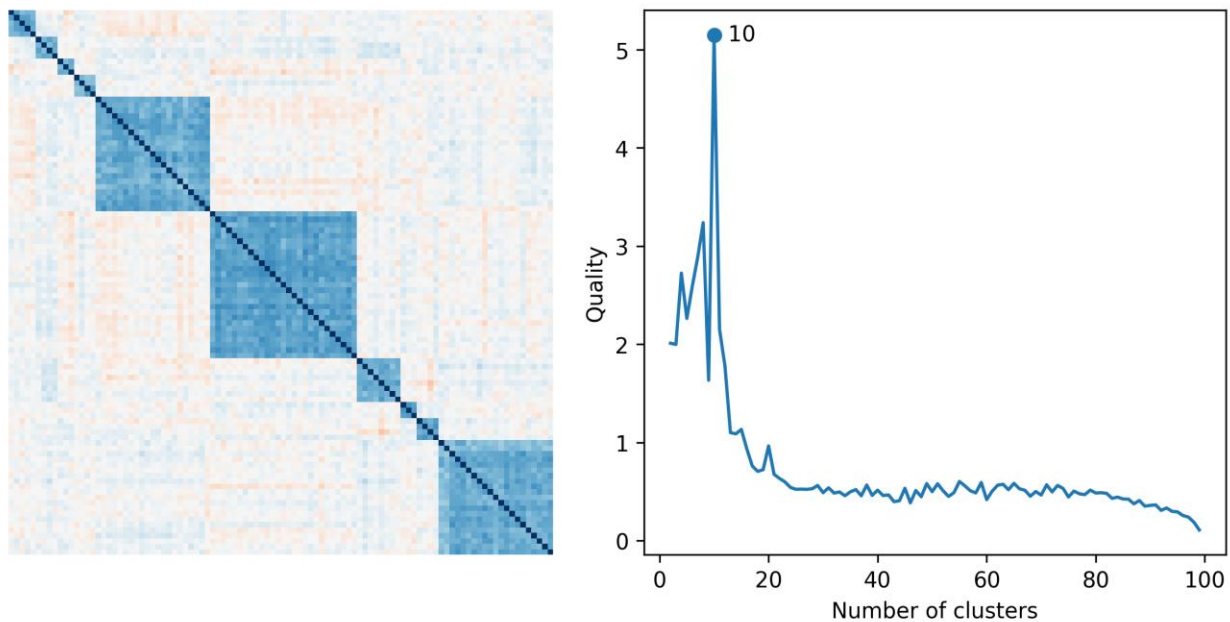


Exhibit 11 – Optimal number of clusters, using the Silhouette method

The first two approaches give an estimate of the number of independent ideas tested, without accounting for the number of variants of each idea; the third approach accounts for all those variants, resulting in a higher number. Exhibit 11 shows a correlation matrix with 10 clusters (left) and the quality (average Silhouette score divided by the standard deviation of the Silhouette score) of a k -means clustering as k varies. Exhibit 12 shows the distribution of the estimated effective number of clusters for the three aforementioned approaches on simulated non-Normal data, with the ground truth (10 clusters) marked as a dotted line. The first two approaches give very similar estimates, where both are very close to the ground truth.

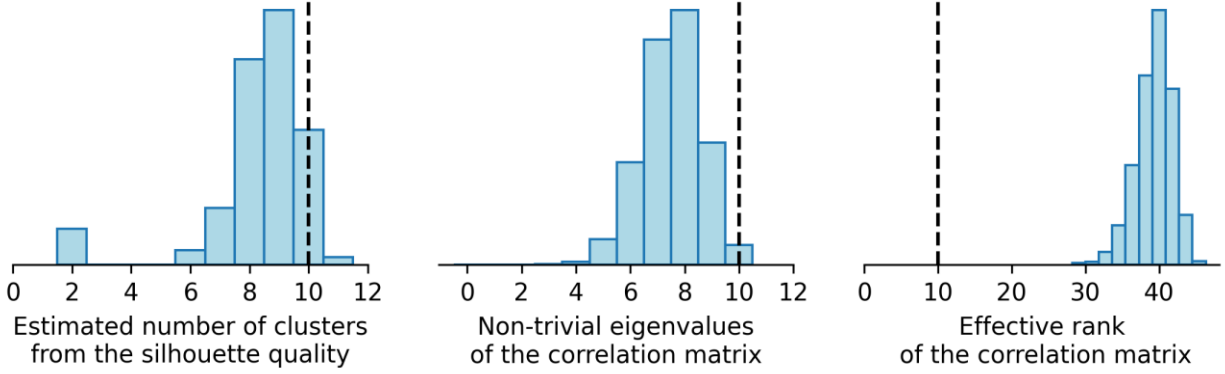


Exhibit 12 – Optimal number of clusters, using the Silhouette method

A.4. VARIANCE OF THE MAXIMUM

Consider K i.i.d. standard Normal variables, X_1, \dots, X_K . We would like to compute the variance of $M = \max\{X_1, \dots, X_K\}$. There is no closed-form expression for this variance, but the CDF of the maximum M is

$$F[m] = P[M \leq m] = P\left[\bigcap_{k=1}^K (X_k \leq m)\right] = \prod_{k=1}^K P[(X_k \leq m)] = Z[m]^K \quad (62)$$

where $Z[m]$ is the CDF of the standard Normal. Accordingly, the density is

$$f[m] = K\phi[m]Z[m]^{K-1} \quad (63)$$

where $\phi[m]$ is the PDF of the standard Normal. We can compute the moments of this distribution as

$$E[M^r] = \int m^r f[m] dm = K \int m^r \phi[m] Z[m]^{K-1} dm = KE[X^r Z[X]^{K-1}] \quad (64)$$

where $X \sim \mathcal{N}[0,1]$. Using the above expression, we can compute the variance of the maximum as

$$V[M] = E[M^2] - E[M]^2 \quad (65)$$

These expectations $E[M^r]$ can be computed numerically using the Gauss-Hermite quadrature. For an implementation, see Numpy's module `np.polynomial.hermite.hermgauss`.

A.5. APPROXIMATED VARIANCE OF THE MAXIMUM

Let X_1, \dots, X_K be i.i.d. Normal random variables with mean μ and variance σ^2 , and define

$$\begin{aligned} M_K &= \max_{1 \leq k \leq K} \{X_k\} \\ u_1 &= \mu + \sigma Z^{-1}\left[1 - \frac{1}{K}\right] \end{aligned} \quad (66)$$

$$u_2 = \mu + \sigma Z^{-1} \left[1 - \frac{1}{Ke} \right]$$

$$\Delta = u_2 - u_1$$

where $Z^{-1}[\cdot]$ is the quantile function of the standard Normal. The above definition applies the quantiles used in the False Strategy Theorem (Bailey and López de Prado [2014]), $1 - \frac{1}{K}$ and $1 - \frac{1}{Ke}$.

In Extreme Value Theory (EVT), the Normal distribution belongs to the Gumbel maximum domain of attraction (Leadbetter et al. [1983, Theorem 1.5.3]). In particular, there exist normalizing sequences a_K and $b_K > 0$ such that

$$\frac{M_K - a_K}{b_K} \xrightarrow{d} G \quad (67)$$

where G is a standard Gumbel random variable. A standard choice of normalization is to set the location and scale using high quantiles of the Normal distribution,

$$\begin{aligned} a_K &= u_1 \\ b_K &= u_2 - u_1 = \Delta \end{aligned} \quad (68)$$

This choice ensures that, as $K \rightarrow \infty$, the normalized maximum converges in distribution to a standard Gumbel random variable. Applying this normalization,

$$\frac{M_K - u_1}{\Delta} \xrightarrow{d} G \quad (69)$$

The standard Gumbel distribution satisfies

$$\begin{aligned} E[G] &= \gamma \\ V[G] &= \frac{\pi^2}{6} \end{aligned} \quad (70)$$

where γ is the Euler–Mascheroni constant. Writing the maximum as $M_K = u_1 + \Delta G$, the variance of the maximum is $V[M_K] = \Delta^2 V[G]$, which scales like Δ^2 .

Classical EVT treats the normalizers (u_1, Δ) as deterministic sequences. In empirical applications, however, (u_1, Δ) are typically unknown and must be inferred from finite samples (for example by estimating (μ, σ) from the data, or by fitting the normalization constants from paired tail quantiles/order statistics). We denote by $(\hat{u}_1, \hat{\Delta})$ the corresponding sample-based estimators of (u_1, Δ) , defined from the same finite sample used to compute the maximum. This induces a dependence between \hat{u}_1 and $\hat{\Delta}$, and contributes additional dispersion to the plug-in representation of the maximum. Under the above Gumbel normalization and to first order in the dispersion of $(\hat{u}_1, \hat{\Delta})$, we approximate the normalized maximum by

$$M_K \approx \hat{u}_1 + \hat{\Delta} G \quad (71)$$

treating the limiting Gumbel variable G as asymptotically independent of $(\hat{u}_1, \hat{\Delta})$ to leading order for the purposes of this variance approximation. Under this approximation, the variance of the plug-in representation $\hat{u}_1 + \hat{\Delta}G$ can be computed exactly by direct expansion, yielding

$$V[\hat{u}_1 + \hat{\Delta}G] = V[\hat{u}_1] + 2E[G]\text{Cov}(\hat{u}_1, \hat{\Delta}) + E[G^2]V[\hat{\Delta}] + V[G]E[\hat{\Delta}]^2 \quad (72)$$

Out of these four terms, only two scale like Δ^2 and can therefore be considered principal components: $V[G]E[\hat{\Delta}]^2$, and a cross term, $2E[G]\text{Cov}(\hat{u}_1, \hat{\Delta})$.²⁹ With regards to the first term, using $E[\hat{\Delta}] \approx \Delta$, we derive the classical EVT result

$$V[G]E[\hat{\Delta}]^2 \approx \frac{\pi^2}{6} \Delta^2 \quad (73)$$

Regarding the second term, when u_1 and Δ are jointly estimated from the paired quantiles $1 - \frac{1}{K}$ and $1 - \frac{1}{Ke}$, under the associated Gumbel normalization, we obtain our correction

$$\begin{aligned} \text{Cov}[\hat{u}_1, \hat{\Delta}] &\approx -\frac{\gamma}{2(1+\gamma)} \Delta^2 \\ 2E[G]\text{Cov}(\hat{u}_1, \hat{\Delta}) &\approx -\frac{\gamma^2}{(1+\gamma)} \Delta^2 \end{aligned} \quad (74)$$

The negative sign is consistent with $\hat{\Delta} = \hat{u}_2 - \hat{u}_1$. Putting all these pieces together, the refined variance approximation inspired by the False Strategy Theorem's quantiles is

$$V[M_K] \approx \Delta^2 \left(\frac{\pi^2}{6} - \frac{\gamma^2}{1+\gamma} \right) \quad (75)$$

Classical EVT characterizes the distribution of the maximum through deterministic normalizing sequences a_K and b_K , which in the present Normal setting correspond to $a_K = u_1$ and $b_K = \Delta$. This leads to the first-order variance approximation $V(M_K) \approx \left(\frac{\pi^2}{6}\right) \Delta^2$ in the Gumbel domain. This approximation treats the normalizers as fixed and therefore ignores the fact that, in empirical applications, both location and scale are jointly estimated from the same finite sample. Our result departs from this convention by explicitly propagating the estimation uncertainty of the normalizing constants into the dispersion of the maximum. In particular, when u_1 and Δ are inferred from paired tail quantiles, their induced dependence contributes a correction at the same Δ^2 scale as the classical EVT variance. Accounting for this effect yields the above refined variance

²⁹ The other two terms, $V[\hat{u}_1]$ and $E[G^2]V[\hat{\Delta}]$, are estimation-variance terms, and their contribution to the variance of the maximum is marginal relative to the Δ^2 -scale contributions.

expression that is absent from standard EVT treatments and is directly relevant for inference on extrema under multiple testing.

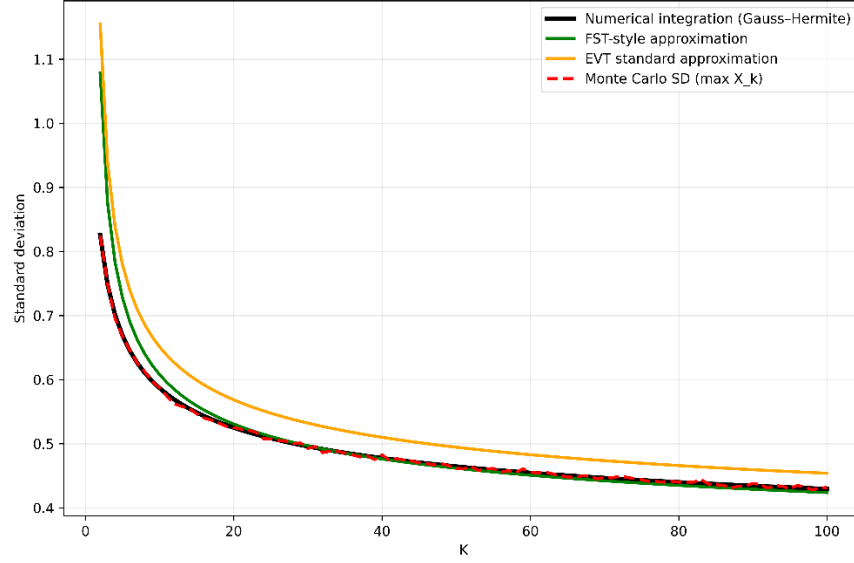


Exhibit 13 – Comparison of estimates of the Standard Deviation of the Maximum

Exhibit 13 confirms the higher accuracy of our approximation, relative to EVT’s standard approximation. For the range $K \in [2, 100]$, it compares the values obtained: (i) applying the numerical integration introduced in Appendix 4 (black bold line); (ii) applying the above False-Strategy-Theorem-style approximation (green line); (iii) applying EVT’s standard approximation (orange line); and (iv) results from a Monte Carlo experiment (dashed red line). For $K > 7$, the False-Strategy-Theorem-style approximation has a relative error below 5%.

A.6. REJECTION THRESHOLD THAT CONTROLS FOR SFDR

Consider a random variable X , where X is drawn from a distribution $H_0: \mathcal{N}[\mu_0, \sigma_0^2]$ with probability $P[H_0]$, and X is drawn from a distribution $H_1: \mathcal{N}[\mu_1, \sigma_1^2]$ with probability $P[H_1] = 1 - P[H_0]$. Given an observed value of X that exceeds a H_0 -rejection threshold c , we denote as false discovery rate the probability that this observation was indeed drawn from H_0 , namely $P[H_0|X \geq c]$. For a target false discovery rate q , we would like to compute the threshold c that achieves

$$q = P[H_0|X \geq c] \quad (76)$$

Let us denote the probabilities

$$\begin{aligned} \alpha &= P[X \geq c|H_0] \\ \beta &= P[X < c|H_1] \end{aligned} \quad (77)$$

We can compute these probabilities as

$$\alpha = P[X \geq c|H_0] = P\left[\frac{X - \mu_0}{\sigma_0} \geq \frac{c - \mu_0}{\sigma_0}\right] = 1 - Z\left[\frac{c - \mu_0}{\sigma_0}\right] \quad (78)$$

$$\beta = P[X < c|H_1] = P\left[\frac{X - \mu_1}{\sigma_1} < \frac{c - \mu_1}{\sigma_1}\right] = Z\left[\frac{c - \mu_1}{\sigma_1}\right] \quad (79)$$

Then, our target probability can be computed as

$$\begin{aligned} q = P[H_0|X \geq c] &= \frac{P[H_0 \cap (X \geq c)]}{P[X \geq c]} \\ &= \frac{P[X \geq c|H_0]P[H_0]}{P[X \geq c|H_0]P[H_0] + P[X \geq c|H_1]P[H_1]} \end{aligned} \quad (80)$$

We can introduce the probabilities defined earlier,

$$q = \frac{\alpha P[H_0]}{\alpha P[H_0] + (1 - \beta)(1 - P[H_0])} = \left(1 + \frac{(1 - \beta)(1 - P[H_0])}{\alpha P[H_0]}\right)^{-1} \quad (81)$$

For fixed parameters $(\mu_0, \sigma_0, \mu_1, \sigma_1, P[H_0])$, $q[c] = P[H_0|X \geq c]$ is monotone decreasing in c , hence the threshold c that achieves a target q (when it exists) is unique. Finally, we can compute q as a function of c ,

$$q = \left(1 + \frac{\left(1 - Z\left[\frac{c - \mu_1}{\sigma_1}\right]\right)(1 - P[H_0])}{\left(1 - Z\left[\frac{c - \mu_0}{\sigma_0}\right]\right)P[H_0]}\right)^{-1} \quad (82)$$

and the value of c is the solution to a root-finding algorithm applied on the above expression.

REFERENCES

- Agarwal, V. and N. Naik (2004): “Risks and Portfolio Decisions Involving Hedge Funds.” *Review of Financial Studies*, Vol. 17, No. 1, pp. 63–98. Available at doi:10.1093/rfs/hhg044
- Bailey, D. and M. López de Prado (2012): “The Sharpe Ratio Efficient Frontier.” *Journal of Risk*, Vol. 15, No. 2, pp. 3–44.
- Bailey, D. and M. López de Prado (2014): “The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality.” *The Journal of Portfolio Management*, Vol. 40, No. 5, pp. 94–107.
- Bailey, D., J. Borwein, M. López de Prado, and J. Zhu (2014): “Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-Of-Sample Performance.” *Notices of the American Mathematical Society*, Vol. 61, No. 5, pp. 458–471.
- Bailey, D., J. Borwein, M. López de Prado, and J. Zhu (2017): “The Probability of Backtest Overfitting.” *Journal of Computational Finance*, Vol. 20, No. 4, pp. 39–70.
- Benjamini, Y. and Y. Hochberg (1995): “Controlling the false discovery rate: a practical and powerful approach to multiple testing.” *Journal of the Royal Statistical Society, Series B*, Vol. 57, pp. 289–300.
- Benjamini, Y. and D. Yekutieli (2001): “The control of the false discovery rate in multiple testing under dependency.” *Annals of Statistics*, Vol. 29, pp. 1165–1188.
- Berk, J. B. (2023): “Comment on ‘The Virtue of Complexity in Return Prediction.’” Working paper. Available at SSRN: <https://ssrn.com/abstract=4410125>
- Buncic, D. (2025): “Simplified: A Closer Look at the Virtue of Complexity in Return Prediction.” Working paper. Available at SSRN: <https://ssrn.com/abstract=5239006>
- Bonferroni, C. (1936): “Teoria statistica delle classi e calcolo delle probabilità.” *Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, Vol. 8, pp. 3–62.
- Boudt, K., P. Carl, and B. G. Peterson (2008): “To CVaR or to MAD, That’s the Question.” *The Journal of Portfolio Management*, Vol. 34, No. 4, pp. 50–57.
- Brooks, C. and H. Kat (2002): “The Statistical Properties of Hedge Fund Index Returns and Their Implications for Investors.” *Journal of Alternative Investments*, Vol. 5, No. 2, pp. 26–44.
- Cartea, Á., Q. Jin and Y. Shi (2025): “The Limited Virtue of Complexity in a Noisy World.” Working paper. Available at SSRN: <https://ssrn.com/abstract=5202064>
- Chen, A. and T. Zimmermann (2022): “Publication Bias and the Cross-Section of Stock Returns.” *Review of Asset Pricing Studies*, Vol. 10, No. 2, pp. 249–289.

Chen, A., A. Lopez-Lira, and T. Zimmermann (2025) “Does Peer-Reviewed Research Help Predict Stock Returns?” Working paper. Available at ArXiv: <https://arxiv.org/abs/2212.10317>

David, H. and Nagaraja, H. (2003): *Order Statistics*. Wiley, 1st edition.

Efron, B., R. Tibshirani, J. Storey, and V. Tusher (2001): “Empirical Bayes analysis of a microarray experiment.” *Journal of the American Statistical Association*, Vol. 96, pp. 1151–1160.

Efron, B. (2004): “Large-scale simultaneous hypothesis testing: the choice of a null hypothesis.” *Journal of the American Statistical Association*, Vol. 99, pp. 96–104.

Efron, B. (2008): “Microarrays, empirical Bayes and the two-groups model.” *Statistical Science*, Vol. 23, pp. 1–22.

Fabozzi, F. and M. López de Prado (2018): “Being Honest in Backtest Reporting: A Template For Disclosing Multiple Tests.” *The Journal of Portfolio Management*, Vol. 45, No. 1, pp. 141–147.

Fallahgoul, H. (2025): “High-Dimensional Learning in Finance.” Working paper. Available at SSRN: <https://ssrn.com/abstract=5281959>

Foster, D., and R. Stine (2008): “ α -Investing: A Procedure for Sequential Control of Expected False Discoveries.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* Vol. 70, No. 2, pp. 429–444.

Fung, W., D. Hsieh, N. Naik, and T. Ramadorai (2008): “Hedge Funds: Performance, Risk, and Capital Formation.” *Journal of Finance*, Vol. 63, No. 4, pp. 1777–1803. Available at: doi:10.1111/j.1540-6261.2008.01375.x

Hansen, P. R. (2005): “A Test for Superior Predictive Ability.” *Journal of Business & Economic Statistics*, Vol. 23, No. 4, pp. 365–380.

Harvey, C. and Y. Liu (2015): “Backtesting.” *The Journal of Portfolio Management*, Vol. 42, No. 1, pp. 13–28.

Harvey, C., Y. Liu, and H. Zhu (2016): “... and the Cross-Section of Expected Returns.” *Review of Financial Studies*, Vol. 29, No. 1, pp. 5–68.

Harvey, C. and Y. Liu (2020): “False (and Missed) Discoveries in Financial Economics.” *The Journal of Finance*, Vol. 75, No. 5, pp. 2503–2553.

Harvey, C., A. Sancetta, and Y. Zhao (2025): “What Threshold Should be Applied to Tests of Factor Models?” Working paper, available at SSRN: <https://ssrn.com/abstract=5925386>

Hochberg, Y. (1988): “A sharper Bonferroni procedure for multiple tests of significance.” *Biometrika*, Vol. 75, No. 4, pp. 800–802.

- Holm, S. (1979): “A simple sequentially rejective multiple test procedure.” *Scandinavian Journal of Statistics*, Vol. 6, pp. 65–70.
- Jacquier, A., O. Kondratyev, A. Lipton, and M. López de Prado (2022): *Quantum Machine Learning and Optimisation in Finance: On the Road to Quantum Advantage*. Packt Publishing, 1st edition.
- Javanmard, A., and A. Montanari (2015): “On Online Control of False Discovery Rate.” Working paper. Available at arXiv:1502.06197.
- Javanmard, A., and A. Montanari (2018): “Online Rules for Control of False Discovery Rate and False Discovery Exceedance.” *The Annals of Statistics*, Vol. 46, No. 2, pp. 526–554.
- Joubert, J., D. Sestovic, I. Barziy, W. Distaso, and M. López de Prado (2024): “Enhanced Backtesting for Practitioners.” *The Journal of Portfolio Management*, Vol. 51, No. 2, pp. 12-27.
- Kat, H. and S. Lu (2002): “An Excursion into the Statistical Properties of Hedge Fund Returns.” Working paper, available at SSRN: <https://ssrn.com/abstract=310227>
- Keating, C. and W. F. Shadwick (2002): “A Universal Performance Measure.” *Journal of Performance Measurement*, Vol. 6, No. 3, pp. 59–84.
- Leadbetter, M., G. Lindgren, H. Rootzén (1983): *Extremes and Related Properties of Random Sequences and Processes*. Springer Verlag, 1st edition.
- Ledoit, O. and M. Wolf (2008): “Robust performance hypothesis testing with the Sharpe ratio.” *Journal of Empirical Finance*, Vol. 15, No. 5, pp. 850-859.
- Lo, A. (2002): “The Statistics of Sharpe Ratios”. *Financial Analysts Journal*, Vol. 58, No. 4, pp. 36-52.
- Lo, A. (2003): “The Statistics of Sharpe Ratios: Author’s Response.” *Financial Analysts Journal*, Vol. 59, No. 5, p. 17.
- Lo, A. and C. MacKinlay (1999): *A Non-Random Walk Down Wall Street*. Princeton University Press, 1st edition.
- López de Prado, M. (2018): *Advances in Financial Machine Learning*. Wiley, 1st edition.
- López de Prado, M. (2019): “A Data Science Solution to the Multiple-Testing Crisis in Financial Research.” *Journal of Financial Data Science*, Vol. 1, No. 1, pp. 99–110
- López de Prado, M. (2020): *Machine Learning for Asset Managers*. Cambridge University Press, 1st ed.
- López de Prado, M. (2023): *Causal Factor Investing*. Cambridge University Press, 1st edition.

López de Prado, M. and D. Bailey (2021): “The False Strategy Theorem: A Financial Application of Experimental Mathematics.” *American Mathematical Monthly*, Vol. 128, No. 9, pp. 825-831.

López de Prado, M. and M. Foreman (2014): “A Mixture of Gaussians Approach to Mathematical Portfolio Oversight: The EF3M Algorithm.” *Quantitative Finance*, Vol. 14, No. 5, pp. 913-930.

López de Prado, M. and M. Lewis (2019): "Detection of False Investment Strategies Using Unsupervised Learning Methods." *Quantitative Finance*, Vol. 19, No. 9, pp. 1555-1565.

López de Prado, M. and V. Zoonekynd (2025): “Correcting the Factor Mirage: A Research Protocol for Causal Factor Investing.” *The Journal of Portfolio Management*, forthcoming.

Markowitz, H. (1952). “Portfolio Selection.” *The Journal of Finance*. Vol. 7, No. 1, pp. 77–91.

Markowitz, H. (1959): *Portfolio Selection: Efficient Diversification of Investments*. John Wiley & Sons, 1st edition.

Mertens, E. (2002): “Variance of the IID estimator in Lo (2002)”. Working paper, University of Basel.

Nagel, S. (2025): “Seemingly Virtuous Complexity in Return Prediction.” Working paper. Available at SSRN: <https://ssrn.com/abstract=5335012>

Ramdas, A, T. Zrnic, M. Wainwright, and M. Jordan (2018): “SAFFRON: An Adaptive Algorithm for Online Control of the False Discovery Rate.” In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Vol. 80, pp. 4286–4294.

Romano, J. P., A. Shaikh, M. Wolf (2008): “Formalized Data Snooping Based on Generalized Error Rates.” *Econometric Theory*, Vol. 24, pp. 404-447.

Romano, J. P. and M. Wolf (2005): “Stepwise Multiple Testing as Formalized Data Snooping.” *Econometrica*, Vol. 73, No. 4, pp. 1237–1282.

Romano, J. P. and M. Wolf (2016): “Efficient Computation of Adjusted P-values for Resampling-Based Stepdown Multiple Testing.” *Statistics & Probability Letters*, Vol. 113, pp. 38–40.

Sharpe, W. (1966): “Mutual Fund Performance”, *Journal of Business*, Vol. 39, No. 1, pp. 119–138.

Sharpe, W. (1975): “Adjusting for Risk in Portfolio Performance Measurement”, *The Journal of Portfolio Management*, Vol. 1, No. 2, Winter, pp. 29-34.

Sharpe, W. (1994): “The Sharpe ratio”, *The Journal of Portfolio Management*, Vol. 21, No. 1, Fall, pp. 49-58.

Šidák, Z. K. (1967). “Rectangular Confidence Regions for the Means of Multivariate Normal Distributions.” *Journal of the American Statistical Association*, Vol. 62, No. 318, pp. 626–633.

Sortino, F. A. and R. van der Meer (1991): “Downside Risk.” *The Journal of Portfolio Management*, Vol. 17, No. 4, pp. 27–31.

Storey, J. (2002): “A direct approach to false discovery rates.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 64, No. 3, pp. 479–498.

Storey, J. D. (2003): “The positive false discovery rate: a Bayesian interpretation and the q-value.” *Annals of Statistics*, Vol. 31, pp. 2013–2035.

van der Vaart, A. (1998): *Asymptotic Statistics*. Cambridge University Press, 1st edition.

Wasserstein, R., A. Schirm, and N. Lazar (2019): “Moving to a World Beyond ‘ $p < 0.05$ ’.” *The American Statistician*, Vol. 73, No. 1, pp. 1–19. Available at <https://doi.org/10.1080/00031305.2019.1583913>

Welch, I. (2025): “Long-Term Risk-Reward Tradeoffs and Sharpe Ratios.” Working paper. Available in SSRN: https://ssrn.com/abstract_id=5709087

White, H. (2000): “A Reality Check for Data Snooping.” *Econometrica*, Vol. 68, No. 5, pp. 1097–1126.

Wolf, M. (2003): “The Statistics of Sharpe Ratios: A Comment.” *Financial Analysts Journal*, Vol. 59, No. 5, p. 17.

Young, P. (1991): “Maximum Drawdown.” *Risk*, Vol. 4, No. 10, pp. 32–37.

Zakamouline, V. and S. Koekebakker (2009): “Portfolio Performance Evaluation with Generalized Sharpe Ratios: Beyond the Mean and Variance.” *Journal of Banking & Finance*, Vol. 33, No. 7, pp. 1242–1254.