HOW TO USE THE SHARPE RATIO

Marcos López de Prado, Alexander Lipton, and Vincent Zoonekynd

ADIA Lab Research Paper Series, No. 19

September 23, 2025

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 most widely used measure of investment efficiency, yet its statistical inference is often conducted incorrectly. This paper reviews the pitfalls of naive Sharpe ratio analysis and provides a comprehensive framework for its proper use. We identify five main pitfalls: (a) the Normality assumption; (b) neglect of statistical significance and minimum sample length assessment; (c) insufficient test power; (d) confusion between classical *p*-values and the probability of the null hypothesis given the data; and (e) failure to correct for multiple testing. To address these issues, we survey and extend a set of methods, including the Probabilistic Sharpe Ratio (PSR), Minimum Track Record Length (MinTRL), the Sharpe ratio's Observed Bayesian Tail-Area False Discovery Rate (oFDR), and the Deflated Sharpe Ratio (DSR). Monte Carlo experiments confirm that these corrections yield more reliable inference than traditional t-statistics and general-purpose multiple-testing adjustments. We further distinguish between familywise error rate (FWER), false discovery rate (FDR), and hybrid FWER-FDR frameworks, showing their respective suitability for academic versus industrial applications. The central conclusion is that the Sharpe ratio remains a valuable metric only when properly corrected; without these adjustments, it risks misleading researchers and practitioners alike.

KEY TAKEAWAYS

- The standard Sharpe ratio estimator is biased in finite samples and further distorted by skewness, kurtosis, 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), Deflated Sharpe Ratio (DSR) and Observed Bayesian False Discovery Rate (oFDR) provide statistically sound inference, accounting for non-Normal returns, limited samples, and selection bias.
- Different corrections suit different contexts: FWER-based methods (like DSR) are more appropriate for academic discovery of factors, while FDR and hybrid FWER-FDR approaches better fit industrial applications where many strategies or managers are selected simultaneously.

Keywords: Sharpe ratio, statistical inference, non-normality, power, *p*-value, Bayesian FDR, FWER.

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 understood as the amount of return achieved per unit of risk. One of the most widely accepted measures 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]. For example, Sharpe ratios are used to rank portfolio managers, assess investment strategies, discover new investment factors, and grade hedge funds.

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, the Sharpe ratio estimator is biased by several factors, including sample length, skewness, kurtosis, and multiple testing. It is critical to correct the estimated Sharpe ratio for those variables before making an investment decision. Second, the estimated value of the Sharpe ratio does not convey information about its statistical significance. A more useful way to measure investment efficiency is to report the Sharpe ratio in the probability space. 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 sample length that is too short for the effect being measured. Fourth, while some practitioners and academics may compute the *p*-value of the Sharpe ratio (i.e., the probability of observing a Sharpe ratio at least as large as the one obtained, conditional on the null hypothesis of zero efficiency being true), this probability is arguably less actionable than its posterior (i.e., the posterior probability that the null hypothesis of zero efficiency holds, conditional on having observed a Sharpe ratio at least as large as the realized value). Computing the probability of the null hypothesis given the data is typically more relevant for Sharpe ratio inference. 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 false positive rate (type-I error), and the rejection threshold increases with the number of trials and the variance of the Sharpe ratios across trials.[1] Similarly, controlling for a given false discovery rate (i.e., the proportion of negatives 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.

Without addressing these five caveats, the Sharpe ratio computed on backtest and historical performance will lead to incorrect inference and, what is worse, investment losses. The goal of this paper is to provide a clear manual for doing inference on the Sharpe ratio.[2]

[1] Intuitively, a trial takes place every time the Sharpe ratio is computed in the context of a decision. 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 trials take place, the higher is the probability that a 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 trials.

[2] In this paper we have not relaxed the assumption of serial independence, as there is no known analytical solution for the interaction between non-Normality and serial dependence. One possible approach is to reduce the sample frequency to a level where the assumption of serial independence is realistic.

LITERATURE REVIEW

There is a long and robust literature that develops the statistical framework for carrying out inference on the Sharpe ratio. To cite a few influential publications, Engle [1982] introduced the ARCH model, providing a framework for modelling time-varying volatility that enabled more accurate forecasts of conditional variance. Engle [2002] introduced the dynamic conditional correlation model, allowing for parsimonious estimation of time-varying correlations, and thus more robust Sharpe ratio analysis at the portfolio level. Lo [2002] derived the asymptotic distribution of the Sharpe ratio under the assumption of independent and identically distributed (i.i.d.) Normal returns. The assumption of Normal returns is problematic, particularly when considering hedge fund investment strategies, see Lo and MacKinlay [1999], Brooks and Kat [2002], Agarwal et al. [2004], Fung et al. [2008]. Accordingly, Mertens [2002] extended Lo's result to i.i.d. non-Normal returns, and Christie [2005] proved that Mertens' result holds under very general conditions, namely stationary and ergodic returns.

An observed Sharpe ratio does not convey information about its statistical significance. A more actionable piece of information for an investor is to estimate the probability that the true expected Sharpe ratio is non-negative, or above a target threshold. To that purpose, Bailey and López de Prado [2012] introduced the probabilistic Sharpe ratio (PSR), which expresses an observed Sharpe ratio in the probabilistic space, while adjusting for sample length, skewness and kurtosis. They also introduced the minimum track record length (MinTRL), which computes the smallest number of observations needed to reject a null hypothesis with a given confidence level. Furthermore, they introduced the concept of Sharpe ratio efficient frontier, which allows investors to optimize a portfolio under non-Normal, leveraged returns while incorporating the uncertainty derived from track record length.

Applying extreme value theory, Bailey and López de Prado [2014] and Bailey et al. [2014] introduced the deflated Sharpe ratio (DSR), a familywise error rate (FWER) adjustment of the Sharpe ratio estimator that corrects for selection bias under multiple testing (in addition to correcting for sample length and non-Normal returns). DSR incorporates information about the number of trials and the variance of the Sharpe ratios across those trials, and it accounts for the correlation structure of the backtests through the estimation of the effective number of trials, see López de Prado and Lewis [2019].[3] Accounting for these many features allows investors to test the Sharpe ratio with higher precision and power compared to using general-purpose multiple testing correction methods. Under the assumption of Normal returns, Harvey and Liu [2015] treat the Sharpe ratio as a t-statistic, and apply classical Bonferroni, Holm and Benjamini-Hochberg-Yekutieli corrections to it, while accounting for the average correlation across the backtests.

In addition to the above approaches, several other strands of research have been adapted to Sharpe ratio inference under multiple testing. White [2000] and Hansen [2005] introduced bootstrap-based procedures designed to address data-snooping bias, and have been explicitly applied to Sharpe ratio comparisons across competing strategies. Romano and Wolf [2005, 2016] developed stepdown and resampling-based multiple testing corrections that, while not Sharpe-specific, have been used to adjust Sharpe ratio inference in empirical finance. More recently, López de Prado [2018] introduced Combinatorial Purged Cross-Validation (CPCV), a simulation-based method to

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

deflate Sharpe ratios under backtest overfitting, complementing the Deflated Sharpe Ratio framework.

A standard result in statistics is that all tests of hypothesis suffer from a trade-off between type-I (false positive) errors and type-II (false negative) errors. In a nutshell, one cannot reduce the probability of selecting a low-Sharpe strategy without increasing the probability of missing high-Sharpe strategies. To control for this trade-off, López de Prado [2020] estimated the power of DSR, and showed that DSR effectively allows researchers to achieve higher levels of power while targeting a certain FWER level. Similarly, Harvey and Liu [2020] use a double-bootstrap method to establish a t-statistic hurdle that is associated with a proportion of false positives. Using a Local False Discovery Rate (FDR) approach, Harvey et al. [2025] compute a t-statistic hurdle that controls for a proportion of false discoveries.

INNOVATIONS

While the primary objective of this paper is educational, it contains several innovations. First, we derive the Sharpe ratio's Planned Bayesian Tail-Area False Discovery Rate (that is, the probability of making a false discovery at the rejection level), denoted as pFDR, without assuming the Normality of returns. Second, we derive the Sharpe ratio's Observed Bayesian Tail-Area False Discovery Rate (that is, the probability of making a false discovery at the observed statistic), denoted as oFDR, also without assuming the Normality of returns. This complements and reinforces the Local False Discovery Rate presented in Harvey et al. [2025] with an additional tool. *P*-values are often misinterpreted as the probability of the null hypothesis given the observed statistic (Wasserstein et al. [2019]), rather than their true meaning as the probability of the observed statistic given the null hypothesis, and both (Planned and Observed) Bayesian False Discovery Rates help resolve this confusion.

A third innovation is that we introduce a hybrid FWER-FDR method, namely a method for applying a FWER correction to oFDR. The need for this correction arises from the situation where capital is allocated to N strategies, and each strategy was chosen for having the highest Sharpe ratio among K (with the other $K-1$ strategies having been discarded). To our knowledge, this is the first time that a hybrid FWER-FDR correction is proposed and computed.

A fourth innovation is that we introduce a new algorithm for the calculation of the rejection threshold that controls for a user-targeted proportion of false discoveries (FDR). While other algorithms may have been developed in the past to this purpose, to our knowledge this is the first of its class that does not assume Normal returns.

SHARPE RATIO

Following Bailey and López de Prado [2012], consider a sample of T excess returns, $\{r_t\}_{t=1,\dots,T}$, with expected value μ and variance σ^2. These excess returns are assumed to be i.i.d., or at least stationary and ergodic. The true (unobserved) Sharpe ratio (SR) is defined as

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

An important property of the Sharpe ratio is its invariance to leverage, which facilitates comparisons across investments.[4] The maximum likelihood estimator for the Sharpe ratio ($\widehat{SR}$) is,

$$\widehat{SR} = \frac{\hat{\mu}}{\hat{\sigma}} \xrightarrow{a} \mathcal{N}\left[SR, \frac{1 - \gamma_3 SR + \frac{\gamma_4 - 1}{4} SR^2}{T}\right] \tag{2}$$

where γ_3 is the skewness of the excess returns, and γ_4 is the kurtosis of the excess returns (with value 3 when returns are Normal). Applying the estimator $\widehat{SR}$ on the sample $\{r_t\}_{t=1,\dots,T}$ we obtain a particular estimate $\widehat{SR}^*$. The estimated variance of the Sharpe ratio under the assumption that $SR = \widehat{SR}^*$ is

$$\hat{\sigma}_{SR}^2 = \frac{1 - \hat{\gamma}_3 \widehat{SR}^* + \frac{\hat{\gamma}_4 - 1}{4} \widehat{SR}^{*2}}{T} \tag{3}$$

Throughout this paper, we compute Sharpe ratios in the frequency of the observations, without annualizing, because annualization is unnecessary for inference. For example, consider a portfolio manager with a 2-year track record of monthly returns, where $(\hat{\mu}, \hat{\sigma}, \hat{\gamma}_3, \hat{\gamma}_4, T) = (0.036\%, 0.079\%, -2.448, 10.164, 24)$. The estimated Sharpe ratio is $\widehat{SR}^* = 0.456$, with a standard deviation of $\hat{\sigma}_{SR} = 0.329$.[5] However, assuming Normally distributed returns, the standard deviation would be approx. 35% smaller, $\hat{\sigma}_{SR} = 0.214$. This evidences that ignoring the non-Normality 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.[6]

PROBABILISTIC SHARPE RATIO

Following Bailey and López de Prado [2012], we can assess whether $\widehat{SR}^*$ is statistically significant by testing the null hypothesis $H_0: SR \leq SR_0$ against the alternative $H_1: SR > SR_0$. The test statistic ($z^*[SR_0]$) is

$$z^*[SR_0] = \frac{\widehat{SR}^* - SR_0}{\hat{\sigma}_{SR_0}} \xrightarrow{a} Z \tag{4}$$

$$\hat{\sigma}_{SR_0} = \sqrt{\frac{1 - \hat{\gamma}_3 SR_0 + \frac{\hat{\gamma}_4 - 1}{4} SR_0^2}{T}} \tag{5}$$

[4] In practice, this comparability is hampered by the fact that leverage is not free, hence a strategy with different leverage levels may have somewhat different Sharpe ratios.

[5] The variance of the Sharpe ratio does not scale linearly with the frequency of the observations, hence it should be computed on the original frequency of the data before applying the linear scaling (assuming that observations are i.i.d.).

[6] Alternative ratios have been proposed to address non-Normal returns, such as the Sortino ratio, which estimates the standard deviation only on negative returns. While it may be a useful heuristic in particular settings, the Sortino ratio does not have a well-developed theoretical foundation like the Sharpe ratio.

where Z is the standard Normal distribution, SR_0 reflects *the least favorable case* in the null hypothesis (e.g., $SR_0 = 0$ in the case of a single trial).[7] The equation for $\hat{\sigma}_{SR_0}$ results from applying the least favorable case on the estimator $\hat{\sigma}^2_{SR}$. The significance level α (false positive rate, type I error) is the probability of rejecting H_0 when it is true,

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

The critical value of the test (SR_c) can be computed as

$$z_{1-\alpha} = Z^{-1}[1-\alpha] \tag{7}$$

$$SR_c = SR_0 + \hat{\sigma}_{SR_0} z_{1-\alpha} \tag{8}$$

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 less extreme than $\widehat{SR}^*$ subject to H_0 being true,

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

where $p = P\left[\widehat{SR} \geq \widehat{SR}^* | H_0\right]$ 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 earlier example, under the null hypothesis where $SR_0 = 0$, then $PSR = Z\left[z^*[0]\right] = Z\left[\frac{\widehat{SR}^*}{\hat{\sigma}_{SR_0}}\right] = 0.987$, but under the null hypothesis where $SR_0 = 0.1$, then $PSR = 0.939$.

Note that under the null hypothesis where $SR_0 = 0$, 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. Therefore, non-Normality only makes a difference when doing inference on $SR_0 \neq 0$. As we will see later, this is particularly important when accounting for multiple testing, and a strong reason for preferring PSR over other tests that assume Normal returns.[8]

The effectiveness of PSR can be demonstrated through the following Monte Carlo experiment. We generate 10,000 returns time series, each representing 5 years' worth of daily observations, by drawing from Mixtures of Gaussians with negative skewness, positive excess kurtosis, and a target Sharpe ratio SR_0. For each of these series we compute the PSR and Student's t statistics, and estimate their respective right-tail probabilities. If the tests work as designed, those right-tail probabilities should follow a Uniform distribution under the null hypothesis. Exhibit 1 reports the value of the Kolmogorov-Smirnov test statistics. The conclusion is that PSR produces more reliable inference than Student's t-distribution test under non-Normal returns. The advantage of

[7] Least favorable in the sense of minimizing the chances of rejecting the null hypothesis.

[8] Under 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 $\delta = \widehat{SR}^*\sqrt{T}$) may be more precise than PSR, however that advantage vanishes under non-Normal returns, making PSR a better choice in practice.

PSR over Student's t-distribution test increases with: (a) the severity of the non-Normality of the returns; and (b) the magnitude of the threshold Sharpe ratio being tested (SR_0).[9]

Non-Normal	T	Annual SR0	Avg Skew	Avg Ex. Kurt	KS_PSR	KS_t	Diff
mild	1260	0	-0.9	2.7	0.010	0.010	0.000
mild	1260	0.5	-0.9	2.7	0.006	0.426	0.420
mild	1260	1	-0.9	2.7	0.009	0.728	0.719
mild	1260	1.5	-0.9	2.7	0.018	0.895	0.877
mild	1260	2	-0.9	2.7	0.007	0.968	0.961
moderate	1260	0	-1.7	7.5	0.014	0.014	0.000
moderate	1260	0.5	-1.7	7.5	0.010	0.427	0.417
moderate	1260	1	-1.7	7.5	0.012	0.718	0.706
moderate	1260	1.5	-1.7	7.5	0.014	0.884	0.870
moderate	1260	2	-1.7	7.5	0.018	0.962	0.945
severe	1260	0	-2.4	13.8	0.017	0.017	0.000
severe	1260	0.5	-2.4	13.8	0.008	0.417	0.409
severe	1260	1	-2.4	13.8	0.013	0.715	0.701
severe	1260	1.5	-2.4	13.8	0.016	0.877	0.861
severe	1260	2	-2.4	13.8	0.022	0.952	0.931

Exhibit 1 – Kolmogorov-Smirnov statistics of PSR and t-test vs Uniform distribution under non-Normal returns

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 we can reject H_0 with confidence $(1-\alpha)$. Formally, the problem can be stated as

$$MinTRL = \min_T \left\{ P\left[\widehat{SR} < \widehat{SR}^* | H_0\right] = 1-\alpha \right\} \tag{10}$$

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

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

Following with our earlier example, for $\alpha = 0.05$ and under the null hypothesis where $SR_0 = 0$, then $MinTRL = 13.029$ months, however under the null hypothesis where $SR_0 = 0.1$, then the minimum track record length more than doubles, to $MinTRL = 27.109$ months. One way to validate these results is to replace in the PSR equation the value of T with MinTRL, thus obtaining $(1-\alpha)$.

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

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 positive excess returns. 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,

$$\beta = P\left[\widehat{SR} < SR_c | H_1\right] = Z\left[\frac{SR_c - SR_1}{\hat{\sigma}_{SR_1}}\right] \tag{12}$$

Then, power is defined as the probability of rejecting the null when it is false,[10]

$$P\left[\widehat{SR} \geq SR_c | H_1\right] = 1 - \beta \tag{13}$$

Power is determined by test parameters, not the observed $\widehat{SR}^*$. The choice of α (false positive rate) determines β (false negative rate), hence also $1 - \beta$ (true positive rate). To see this, note that

$$SR_c = SR_0 + \hat{\sigma}_{SR_0} z_{1-\alpha} \tag{14}$$

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

$$\hat{\sigma}_{SR_1} = \sqrt{\frac{1 - \hat{\gamma}_3 SR_1 + \frac{\hat{\gamma}_4 - 1}{4} SR_1^2}{T}} \tag{16}$$

where $\hat{\sigma}_{SR_1} > \hat{\sigma}_{SR_0}$. For $SR_0 = 0$, this can be simplified into

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

This equation shows 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 $\hat{\gamma}_3$ and $\hat{\gamma}_4$ are not under the control of the researcher. One possibility would be to use the above equation to derive the value of T needed to achieve a power $(1 - \beta)$, as an alternative to MinTRL.[11] Exhibit 2 reports precision and recall rates for the Monte Carlo experiment described earlier, where $SR_0 = 0$ and $SR_1 \in \{0.5, 1, 1.5, 2\}$. The results demonstrate that PSR's power does not decrease with non-Normality, evidencing that the adjustment works as designed.

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

[11] One disadvantage of such approach would be the need to set a value for SR_1, which MinTRL does not require.

Non-Normal	Annual SR1	Precision	Recall	F1
mild	0.5	0.847	0.305	0.448
mild	1	0.929	0.720	0.811
mild	1.5	0.945	0.949	0.947
mild	2	0.948	0.997	0.972
moderate	0.5	0.840	0.305	0.448
moderate	1	0.925	0.719	0.809
moderate	1.5	0.942	0.945	0.944
moderate	2	0.945	0.995	0.969
severe	0.5	0.839	0.311	0.454
severe	1	0.923	0.714	0.805
severe	1.5	0.941	0.945	0.943
severe	2	0.943	0.996	0.969

Exhibit 2 – Precision and recall of PSR test under non-Normal returns

Following with our earlier example, for $\alpha = 0.05$ and under the alternative hypothesis where $SR_1 = 0.5$, then the false negative rate is $\beta = 0.315$.

PLANNED BAYESIAN FALSE DISCOVERY RATE

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

$$pFDR = P\left[H_0 \middle| \widehat{SR} \geq SR_c\right] \tag{18}$$

Like power, pFDR is determined by test parameters, not the observed $\widehat{SR}^*$. We can compute pFDR as an application of Bayes' theorem,

$$P\left[H_0 \middle| \widehat{SR} \geq SR_c\right] = \frac{P\left[\widehat{SR} \geq SR_c | H_0\right] P[H_0]}{P\left[\widehat{SR} \geq SR_c\right]} \tag{19}$$

From the law of total probability, we know that

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

resulting in

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

In practice, the value of $P[H_0]$ can be estimated from the proportion of deployed strategies that have yielded negative or around zero excess returns during live performance. Following with our earlier example, suppose that $P[H_1] = 0.1$, $\alpha = 0.05$ and $\beta = 0.315$, then $FDR = 0.397$. This illustrates how a test with relatively high power (at a 68.5% level) can still have a high planned false discovery rate (at a 58% level) when positives are relatively rare (5% probability).

OBSERVED BAYESIAN FALSE DISCOVERY RATE

The previous equations show that pFDR is a function of the test characteristics $(\alpha, \beta, P[H_0])$,[12] not the observed $\widehat{SR}^*$. This invites the question, what is the probability that H_0 is true subject to observing $\widehat{SR}^*$? This probability, $P\left[H_0 \middle| \widehat{SR} \geq \widehat{SR}^*\right]$, denoted as oFDR, can be understood as the Bayesian posterior of the prior, $P[H_0]$, after incorporating the likelihood expressed by the *p*-value (p), $P\left[\widehat{SR} \geq \widehat{SR}^* | H_0\right]$. From Bayes' theorem,

$$oFDR = P\left[H_0 \middle| \widehat{SR} \geq \widehat{SR}^*\right] = \frac{P\left[\widehat{SR} \geq \widehat{SR}^* | H_0\right] P[H_0]}{P\left[\widehat{SR} \geq \widehat{SR}^*\right]} \tag{22}$$

From the law of total probability, we know that

$$\begin{aligned} P\left[\widehat{SR} \geq \widehat{SR}^*\right] &= P\left[\widehat{SR} \geq \widehat{SR}^* | H_0\right] P[H_0] + P\left[\widehat{SR} \geq \widehat{SR}^* | H_1\right] P[H_1] \\ &= pP[H_0] + (1 - z^*[SR_1])(1 - P[H_0]) \end{aligned} \tag{23}$$

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

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

Following with our earlier example, for $SR_0 = 0$, $SR_1 = 0.5$ and $P[H_1] = 0.1$, then the *p*-value is $P\left[\widehat{SR} \geq \widehat{SR}^* | H_0\right] = 1 - PSR = 0.013$, while the oFDR is $P\left[H_0 \middle| \widehat{SR} \geq \widehat{SR}^*\right] = 0.173$. This evidences that an investment may have a statistically significant Sharpe ratio at a 95% confidence level, and yet the probability that the null hypothesis is true can be relatively high (at a 31% level), because positives are relatively rare.

MULTIPLE TESTING CORRECTIONS

Consider a sample of K observed Sharpe ratios, $\left\{\widehat{SR}_k^*\right\}_{k=1,\ldots,K}$, independently drawn from a Normal distribution $\widehat{SR}_k^* \sim \mathcal{N}\left[0, V\left[\widehat{SR}\right]\right]$. For $K = 1$, the false positive probability is α, however selecting one result out of K has a false positive probability

$$\alpha_K = 1 - (1 - \alpha)^K \tag{25}$$

[12] In the Normal case, β is a function of T and SR_1, which are determined before the observations take place. In the non-Normal case, β is also a function of γ_3 and γ_4, which depend on the observations but not the statistic itself.

Two questions arise naturally: (a) what is the new rejection threshold for the strategy with the highest Sharpe ratio, such that it controls for α_K?; and (b) what is the new rejection threshold above which the proportion of negatives among the selected strategies is constant? These are two different questions that control for two different probabilities, and therefore have two different answers.

CONTROLLING FOR THE FAMILYWISE ERROR RATE

Suppose that from this sample of K Sharpe ratios, a researcher chooses the strategy with the highest Sharpe ratio. In order to conduct inference, we must compute the expected value of the maximum Sharpe ratio and the variance of the maximum Sharpe ratio.

Expected Value of the Maximum Sharpe Ratio

The False Strategy Theorem (Bailey and López de Prado [2014]) shows that the expected value of the selected strategy is:

$$E\left[\max_k\{\widehat{SR}_k^*\}\right] \approx \sqrt{V\left[\{\widehat{SR}_k^*\}\right]}\left((1-\gamma)Z^{-1}\left[1-\frac{1}{K}\right]+\gamma Z^{-1}\left[1-\frac{1}{Ke}\right]\right) \tag{26}$$

where $\gamma = 0.5772156649\ldots$ 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 practice, the K trials are not independent, and the effective number of independent trials can be estimated via clustering methods (López de Prado [2019]), or as the effective number derived from the eigenvalues of the trials' correlation matrix (López de Prado [2018, section 18.7], López de Prado [2020]), see Appendix 1 for details and experimental validation. This theorem proves that, when multiple trials take place, $SR_0 = 0$ is not the "least favorable case" under the null hypothesis $H_0: SR \leq 0$. Setting $SR_0 = E\left[\max_k\{\widehat{SR}_k^*\}\right]$ in the earlier equations adjusts for multiple testing by pushing SR_c to the right, thus controlling for the so-called familywise error rate (FWER).

When $SR_0 \neq 0$, the adjustment is a shift in SR_0 by the same amount, $E\left[\max_k\{\widehat{SR}_k^*\}\right]$. To see why, consider a variable $X \sim \mathcal{N}[\mu_X, \sigma_X^2]$, and a variable $Y = \mu_Y + X$. Then, by the translation invariance of the extreme-value distribution, we obtain that

$$E\left[\max_k\{Y_k\}\right] \approx \mu_Y + E\left[\max_k\{X_k\}\right] \tag{27}$$

See Leadbetter et al. [1983, chapter 1] for a proof. For the same reason, the adjustment for SR_1 also consists in a shift by $E\left[\max_k\{\widehat{SR}_k^*\}\right]$.

Variance of the Maximum Sharpe Ratio

Similarly, under multiple trials, the standard deviations $\hat{\sigma}_{SR_0}$ and $\hat{\sigma}_{SR_1}$ now represent the standard deviations of maxima of Sharpe ratios, and must be corrected accordingly. In particular, $\hat{\sigma}_{SR_0}$ and $\hat{\sigma}_{SR_1}$ must be re-scaled by the standard deviation of the maximum of K standard Normal variables,

$$\sqrt{V\left[\max_k\{\widehat{SR}_k^*\}\right]} = \hat{\sigma}_{SR_H}\sqrt{V\left[\max_k\{X_k\}\right]} \tag{28}$$

where $\{X_k\}_{k=1,\ldots,K}$ are K independent and identically distributed standard Normal variables, and $H = \{0,1\}$ depending on the hypothesis tested. The re-scaling factor $\sqrt{V\left[\max_k\{X_k\}\right]}$ can be computed as explained in Appendix 2. Exhibit 3 reports the values of re-scaling factors, from $K = 1$ to $K = 100$.

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

Applying these two adjustments to PSR gives the deflated Sharpe ratio (DSR), and applying this adjustment to MinTRL prevents an underestimation of the minimum sample length. One advantage of DSR over general-purpose correction methods (Bonferroni, Sidak, Holm, etc.) is that it takes into account the dispersion across trials ($V\left[\{\widehat{SR}_k^*\}\right]$). The variance of the Sharpe ratios across trials tends to be larger 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. The False Strategy Theorem demonstrates the value of conducting research under a theoretical causal framework, see López de Prado [2023] and López de Prado and Zoonekynd [2025].

These adjustments also enable the correct estimation of power, pFDR and oFDR.[13] Following with our earlier example, for $K = 10$, and $V\left[\{\widehat{SR}_k^*\}\right] = 0.1$, a one-trial $SR_0 = 0$ must be adjusted to $SR_0 = E\left[\max_k\{\widehat{SR}_k^*\}\right] = 0.498$, with standard deviation $\hat{\sigma}_{SR_0} = 0.200$, which results in $DSR = 0.416$ (compared to the one-trial PSR of 0.987). In words, after accounting for the multiple tests, the observed Sharpe ratio is about what would be expected from zero skill (a coin toss). As explained earlier, the one-trial $SR_1 = 0.5$ also needs to be shifted by $E\left[\max_k\{\widehat{SR}_k^*\}\right] = 0.498$ to account for multiple tests, thus the corrected values are $SR_1 = 0.998$ and $\hat{\sigma}_{SR_1} = 0.287$, which

[13] 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.

results in a corrected oFDR of $P\left[H_0\middle|\widehat{SR} \geq \widehat{SR}^*\right] = 0.844$ (compared to the one-trial value of 0.173).[14]

CONTROLLING FOR THE FALSE DISCOVERY RATE

Alternatively, suppose that a researcher wishes to select strategies while accepting that the proportion of them that are false (pFDR) is a constant level q,

$$P\left[H_0\middle|\widehat{SR} \geq SR_c\right] = \frac{\alpha P[H_0]}{\alpha P[H_0] + (1-\beta)(1-P[H_0])} = \left(1 + \frac{(1-\beta)P[H_1]}{\alpha P[H_0]}\right)^{-1} = q \tag{29}$$

Replacing α and β in the above equation, we obtain

$$q = \left(1 + \frac{\left(1 - Z\left[\frac{SR_c - SR_1}{\hat{\sigma}_{SR_1}}\right]\right)(1-P[H_0])}{\left(1 - Z\left[\frac{SR_c - SR_0}{\hat{\sigma}_{SR_0}}\right]\right)P[H_0]}\right)^{-1} \tag{30}$$

See Appendix 3 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 $P\left[H_0\middle|\widehat{SR} \geq SR_c\right] = q$. Note that under the FDR 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 was the case in the FWER framework). For this reason, SR_0 does not increase with the number of trials, like in the FWER correction.

Following with our earlier example, Exhibit 4 plots the (non-annualized) SR_c thresholds that control for a constant FDR at a level $q = 0.25$ as a function of $P[H_1]$. The orange line shows results under $SR_0 = 0$, $SR_1 = 0.5$, $\hat{\sigma}_{SR_0} = 0.204$, and $\hat{\sigma}_{SR_1} = 0.341$. For instance, under $P[H_1] = 10\%$, then $SR_c = 0.41$ controls for a constant FDR at a level $q = 0.25$. Accordingly, the strategy with an observed Sharpe ratio $\widehat{SR}^* = 0.456$ would not 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 a negative is below the tolerance for false discoveries. Exhibit 4 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 interpretation is that, when the Sharpe ratio of true strategies is lower, it is harder to discriminate between true and false strategies, and the threshold must adjust for the increased probability of a false discovery.

[14] The interpretation of a FWER adjustment on oFDR is the following: The proportion of negatives among a set of strategies, where each strategy was chosen for having the highest Sharpe ratio among a group of K strategies (with the other $K - 1$ strategies having been discarded).

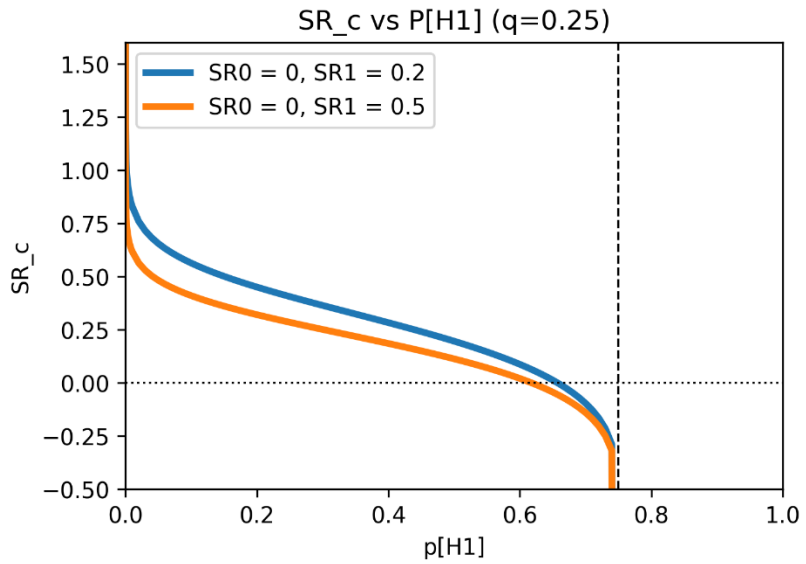

Exhibit 4 – Sharpe ratio rejection thresholds (SR_c) that control for a constant false discovery rate ($q = 25\%$) as a function of the unconditional probability of a positive ($P[H_1]$), under $SR_1 = 0.2$ and $SR_1 = 0.5$

HYBRID FWER-FDR CORRECTION

It would be incorrect to conclude from the above discussion that investment practitioners do not need to control for trials when applying the FDR correction. Suppose a hedge fund that requires each of its researchers to present at the weekly investment committee the best result they achieved after running K trials. Because of backtest overfitting, each of the Sharpe ratios presented are inflated without an upper bound. This means that SR_0 and SR_1 need to be corrected for K and $V[\{\widehat{SR}_k^*\}]$. The standard deviations $\hat{\sigma}_{SR_0}$ and $\hat{\sigma}_{SR_1}$ should also reflect the standard deviations of maxima of Sharpe ratios, and must be corrected accordingly, as explained in the FWER section. Without these corrections, the FDR rejection threshold will be overly optimistic (i.e., too low). This setup, whereby pre-selected strategies are presented on a recurrent basis, is prevalent at many large asset managers, and it calls for a combined FWER-FDR correction. The hybrid FWER-FDR correction should also prove valuable in order to account for publication bias in academia.

We can assess the effectiveness of our proposed hybrid FWER-FDR correction method with a Monte Carlo experiment. Suppose that a researcher computes the returns of $K = 10$ strategies, each with a sample length $T = 100$, all Normally distributed. To maximize his chances of receiving a funds allocation, the researcher pre-selects the best backtest out of K, thus submitting to the investment committee only the strategy with the highest Sharpe ratio. From earlier experience, the investment committee knows that true strategies have an average Sharpe ratio $SR_1 = 0.2$ once deployed, and occur with probability $P[H_1] = 0.3$, and false strategies have an average Sharpe ratio $SR_0 = 0$. The investment committee wishes to control for FDR at a level of $q = 25\%$. Exhibit 5 plots in blue the distribution of the pre-selected Sharpe ratios for false strategies, and in orange the distribution of the pre-selected Sharpe ratios for true strategies. These distributions are not centered around SR_0 and SR_1 due to the pre-selection bias introduced by the researcher. Instead, the distributions are centered around the values predicted by the False Strategy Theorem (marked in dashed white vertical lines).

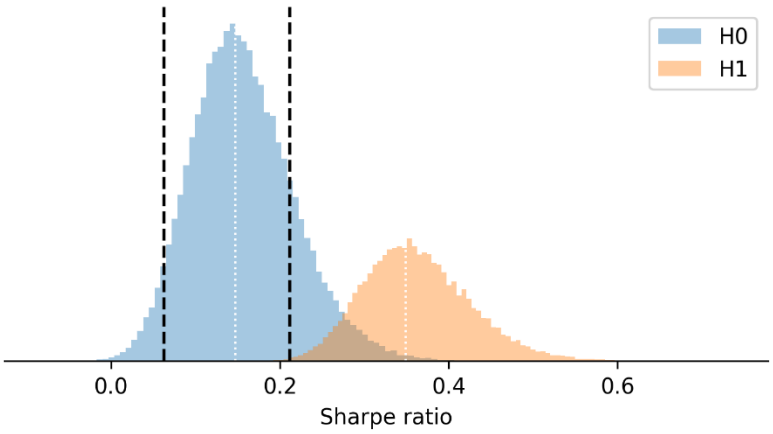

Exhibit 5 – Distribution of pre-selected Sharpe ratios for false and true strategies, with FDR and FWER-FDR rejection thresholds

If the investment committee ignores that pre-selection has taken place, they will apply the standard FDR procedure, resulting in the rejection threshold SR_c marked by the left black dashed vertical line. However, if the investment committee decides to account for pre-selection bias, they will apply our hybrid FWER-FDR procedure, resulting in the rejection threshold SR_c marked by the right black dashed vertical line. The Monte Carlo experiment shows that our hybrid FWER-FDR achieves a false discovery rate of 22.4%, very close to the targeted $q = 25\%$. Visually, this is the ratio between false positives (the blue observations to the right of the threshold) and all predicted positives (the blue and orange observations to the right of the threshold). In contrast, applying the standard FDR rejection threshold that ignores pre-selection yields a false discovery rate of 68.9%, much higher than targeted.

Exhibit 6 extends these results to Monte Carlo experiments with representative non-Normal distributions, where $K = 10$, $T = 100$, $SR_0 = 0$, $SR_1 = 0.2$, and $P[H_0] = 0.7$. In all scenarios, standard FDR thresholds produce too many false discoveries, whereas hybrid FWER-FDR thresholds produce a percentage of false discoveries close to the target.

Process	Avg Skew	Avg Ex. Kurt	q (%)	FDR (%)	FWER-FDR (%)
gaussian	0	0	25%	68.9%	22.4%
mild	-0.9	2.7	25%	69.0%	25.3%
moderate	-1.7	7.5	25%	69.3%	28.0%
severe	-2.4	13.8	25%	69.4%	28.5%

Exhibit 6 – Results of Monte Carlo experiments under Normal and non-Normal distributions

The implication is that in investing, where researchers have an economic incentive to overfitting backtests, controlling for the trials involved in a discovery is important for FWER corrections, but also for FDR corrections.[15]

[15] A reasonable counter-argument is that SR_1 could be estimated *ex-post* over multiple investment committee meetings, to reflect the necessary adjustment. This is a valid argument as long as K and $V\left[\{\widehat{SR}_k^*\}\right]$ are similar between meetings, but holding such assumption once again requires some form of control over those variables.

WHICH MULTIPLE TESTING CORRECTION SHOULD BE APPLIED?

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

Method	Authors	Correction Type	Sharpe Specific?	Notes
Lo's Significance Test	Lo [2002]	Single-test inference	Yes	Adjusts for sample length, under Normal returns
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
Bonferroni	Classical statistics	FWER	Adapted	Simple but conservative correction
Sidak	Classical statistics	FWER	Adapted	Less conservative than Bonferroni
Holm	Classical statistics	FWER	Adapted	Stepwise improvement over Bonferroni
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
Combinatorial Purged Cross-Validation (CPCV)	López de Prado [2018]	FWER	Yes	Bootstrapping of Sharpe ratio's distribution; complements DSR
Power of the Sharpe Ratio	López de Prado [2020]	FWER	Yes	Computes the type-II error associated with a Sharpe ratio rejection threshold
Benjamini-Hochberg-Yekutieli	Harvey & Liu [2015]	FDR	Adapted	Controls expected proportion of false discoveries
Double-bootstrap hurdle	Harvey & Liu [2020]	FDR	Adapted	Double-bootstrap Sharpe hurdle linked to false positives
Local FDR	Harvey, Sancetta & Zhao [2025]	FDR	Yes	Sets Sharpe/t-stat hurdle via false discovery proportion
Bayesian oFDR / pFDR	López de Prado, Lipton & Zoonekynd [2025]	FDR	Yes	Bayesian tail-area FDR
Hybrid FWER-FDR	López de Prado, Lipton & Zoonekynd [2025]	Joint FWER-FDR	Yes	Joint FWER-FDR correction, to account for pre-selection

Exhibit 7 – Methods designed and applied to perform inference on the Sharpe ratio

Controlling for FWER is important in instances where a selected model overrides the rest, therefore it is critical to control the probability that the selected model is a false positive. The stringency of FWER corrections is warranted by the fact that, should the discovery be false, the entire community will be impacted. This is the standard situation in academic publishing and scientific discovery, where the scientific community relies on a particular finding to the detriment of competing explanations. In the context of finance, FWER corrections are more appropriate in academic applications, like the discovery of factor models for risk and investing.

Controlling for FDR is important in instances where all models that satisfy a minimum threshold are applied simultaneously, therefore the focus is to control the percentage of errors (a quality control). The leniency of FDR corrections is warranted by the fact that, should a particular product be faulty, only a small fraction of users will be impacted. In the context of finance, FDR corrections are more appropriate in industrial applications, like the recruitment of portfolio managers or the graduation of investment strategies.

Finally, the hybrid FWER-FDR correction applies to instances where researchers share pre-selected Sharpe ratios, that is, Sharpe ratios that were selected among several trials, hence shifting their expected values and standard deviations. Pre-selection is standard practice in asset managers, where investment committees are only exposed to the "best ideas", but also in academia, where every published result goes through multiple layers of filtering (by authors, by editors, by readers, …). For this reason, the hybrid FWER-FDR correction likely applies to most settings in practice. Exhibit 7 summarizes the methods discussed in this article.

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) the Normality assumption; (b) neglect of statistical significance and minimum sample length assessment; (c) insufficient test power; (d) the confusion between classical *p*-values and the probability of the null hypothesis given the data; and (e) failure to correct for multiple testing.

To correct these shortcomings, we have reviewed several statistical tools. The Probabilistic Sharpe Ratio (PSR) expresses observed Sharpe ratios in probability space, adjusting for skewness, kurtosis, and sample length. The Minimum Track Record Length (MinTRL) quantifies the data requirements to achieve meaningful inference. The Deflated Sharpe Ratio (DSR) corrects for selection bias and backtest overfitting by accounting for the number and correlation of trials. Furthermore, extensions based on Bayesian inference and false discovery rate (FDR) offer practical frameworks to balance type-I and type-II errors in both academic and industrial applications.

Our Monte Carlo experiments confirm that PSR and DSR provide more reliable inference than classical t-tests and general-purpose multiple-testing corrections under non-Normal returns. In particular, PSR demonstrates superior precision and recall, while DSR achieves higher power at a controlled familywise error rate. The choice between FWER and FDR corrections depends on the context: FWER is more appropriate for academic research where a single discovery may dominate, FDR is better suited for industrial applications where multiple strategies are deployed

simultaneously, and the hybrid FWER-FDR is better suited for settings where a multiplicity of pre-selected discoveries are deployed simultaneously.

Ultimately, the Sharpe ratio remains a valuable tool only if properly adjusted and interpreted. Researchers and practitioners must move beyond raw estimates, incorporating corrections for non-Normality, small sample bias, and multiple testing. Failure to do so risks turning the Sharpe ratio from a measure of efficiency into a source of systematic error. By applying the framework laid out in this paper, investors can make sounder decisions, avoid backtest overfitting, and establish a more rigorous standard for investment evaluation.

ACKNOWLEDGEMENTS

The views expressed in this paper are the authors', and do not necessarily represent the opinions of the organizations they are affiliated with. We would like to thank our ADIA colleagues, especially Illya Barziy, Walter Distaso, Jacques Joubert, Dmitri Maksarov, Dragan Sestovic, and Blaz Zlicar. We are also grateful for useful comments provided by Patrick Cheridito (ETH Zürich), Robert Engle (New York University), Frank Fabozzi (EDHEC), Campbell Harvey (Duke University), Guido Imbens (Stanford University), Alessia López de Prado Rehder (University of Zürich), Riccardo Rebonato (EDHEC), Alessio Sancetta (Royal Holloway, University of London), Luis Seco (University of Toronto), Horst Simon (ADIA Lab), and Josef Teichmann (ETH Zürich).

APPENDIX

A.1. EXPERIMENTAL VALIDATION OF EFFECTIVE NUMBER OF TRIALS

We propose three approaches to estimate the effective number of trials (K). The first approach follows López de Prado [2019]: (i) compute the correlation matrix of the time series of returns from all trials; (ii) cluster the correlation matrix, and derive the optimal number K that maximizes the t-value of the mean Silhouette score; (iii) compute one return time series per cluster, as a weighted average of all trials included in that cluster, where the weights are the solution to the minimum variance allocation (this prevents that the most volatile trials dominate); (iv) using those derived time series, compute one Sharpe ratio per cluster; (iv) compute $V\left[\left\{\widehat{SR}_k^*\right\}\right]$ as the variance of Sharpe ratios across clusters.

The second approach is inspired by López de Prado [2018, section 18.7] and 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 is simply to compute the effective rank of the correlation matrix of the time series of returns from all trials.

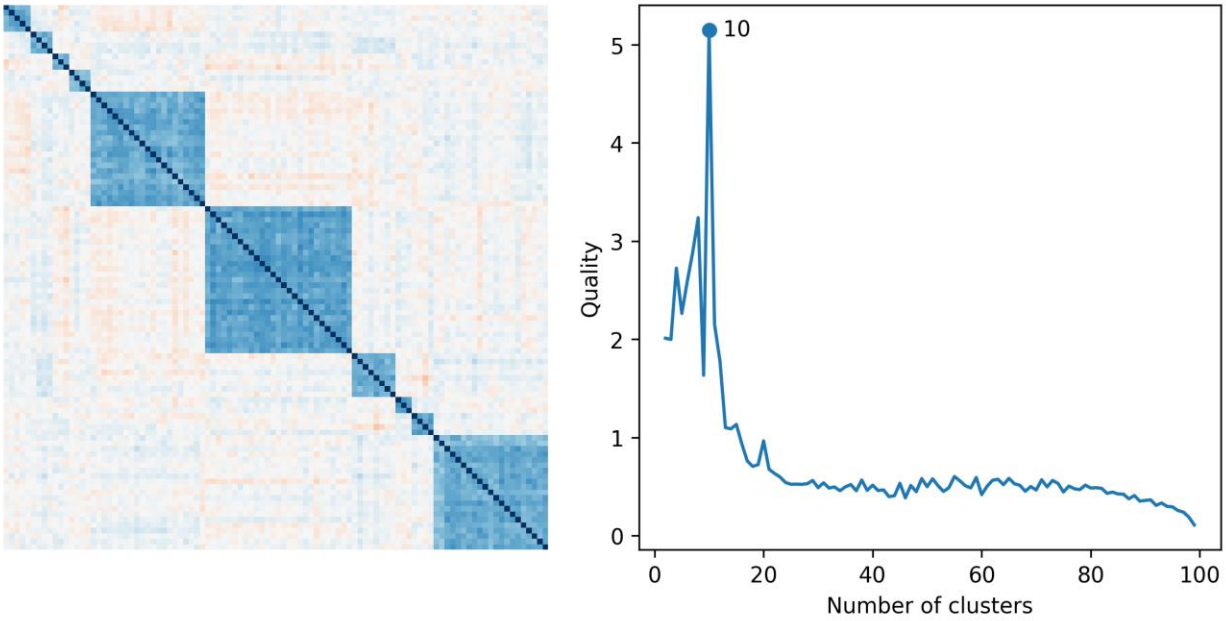

Exhibit 8 – Optimal number of clusters, using the Silhouette method

The first two approaches give an estimate of the number of independent ideas tested, but do not account for the number of variants of each idea; the third approach accounts for all those variants, and gives a much higher number. Exhibit 8 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 9 shows the distribution of the estimated effective number

of clusters with the three approaches mentioned on simulated non-Normal data, with the ground truth (10 clusters) marked as a dotted line. The first two approaches give very similar returns, where both are very close to the ground truth.

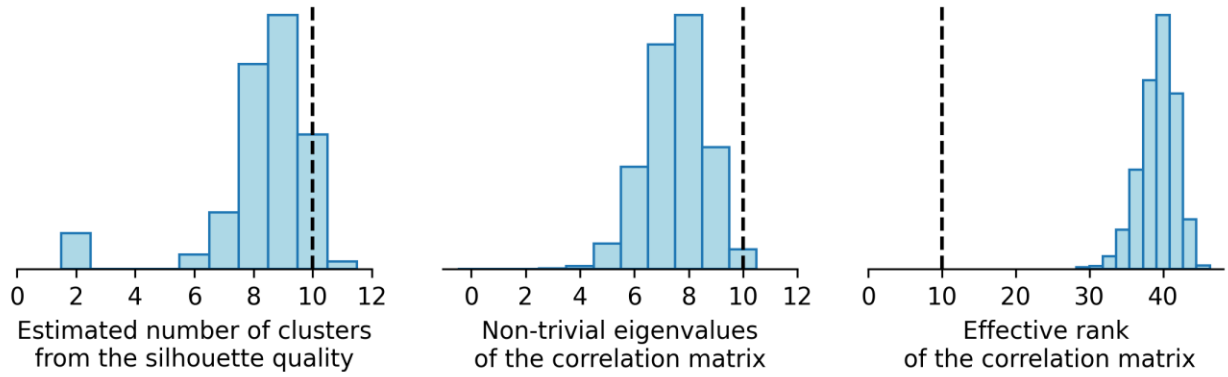

Exhibit 9 – Optimal number of clusters, using the Silhouette method

A.2. VARIANCE OF THE MAXIMUM SHARPE RATIO

Consider K independent and identically distributed standard Normal variables, $X_1, \ldots, X_K$. We would like to compute the variance of $M = max\{X_1, \ldots, X_K\}$. There is no analytical expression for this variance, but the CDF of the maximum M is

$$F[m] = P[M \leq m] = P\left[\bigcap_{k=1}^{K}(X_k \leq m)\right] = P\left[\prod_{k=1}^{K}(X_k \leq m)\right] = \Phi[m]^K \tag{31}$$

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

$$f[m] = K\phi[m]\Phi[m]^{K-1} \tag{32}$$

We can compute the moments of this distribution as

$$E[M^r] = \int m^r f[m]dm = K\int m^r \phi[m]\Phi[m]^{K-1}dm = KE[X^r\Phi[X]^{K-1}] \tag{33}$$

Using the above expression, we can compute the variance of the maximum as

$$V[M] = E[M^2] - E[M]^2 \tag{34}$$

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

A.3. REJECTION THRESHOLD THAT CONTROLS FOR FDR

Consider a random variable X, where X is drawn from a distribution $H_0: N[\mu_0, \sigma_0^2]$ with probability $P[H_0]$, and X is drawn from a distribution $H_1: N[\mu_1, \sigma_1^2]$ 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

$$q = P[H_0|X \geq c] \tag{35}$$

Let us denote the probabilities

$$\begin{aligned} \alpha &= P[X \geq c|H_0] \\ \beta &= P[X < c|H_1] \end{aligned} \tag{36}$$

We can compute these probabilities as

$$\alpha = P[X \geq c|H_0] = P\left[\frac{X-\mu_0}{\sigma_0} \geq \frac{c-\mu_0}{\sigma_0}\right] = 1 - Z\left[\frac{c-\mu_0}{\sigma_0}\right] \tag{37}$$

$$\beta = P[X < c|H_1] = P\left[\frac{X-\mu_1}{\sigma_1} < \frac{c-\mu_1}{\sigma_1}\right] = Z\left[\frac{c-\mu_1}{\sigma_1}\right] \tag{38}$$

Then, our target probability can be computed as

$$\begin{aligned} q = P[H_0|X \geq c] &= \frac{P[H_0 \cap (X \geq c)]}{P[X \geq c]} \\ &= \frac{P[X \geq c|H_0]P[H_0]}{P[X \geq c|H_0]P[H_0] + P[X \geq c|H_1]P[H_1]} \end{aligned} \tag{39}$$

We can introduce the probabilities defined earlier,

$$q = \frac{\alpha P[H_0]}{\alpha P[H_0] + (1-\beta)(1-P[H_0])} = \left(1 + \frac{(1-\beta)(1-P[H_0])}{\alpha P[H_0]}\right)^{-1} \tag{40}$$

Finally, we can compute q as a function of c,

$$q = \left(1 + \frac{\left(1 - Z\left[\frac{c-\mu_1}{\sigma_1}\right]\right)(1-P[H_0])}{\left(1 - Z\left[\frac{c-\mu_0}{\sigma_0}\right]\right)P[H_0]}\right)^{-1} \tag{41}$$

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., 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. 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.

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.

Christie, S. (2005): "Is the Sharpe Ratio Useful in Asset Allocation?", MAFC Research Papers No.31, Applied Finance Centre, Macquarie University.

Engle, R.F. (1982): "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation." *Econometrica*, Vol. 50, No. 4, pp. 987–1007.

Engle, R.F. (2002): "Dynamic Conditional Correlation: A Simple Class of Multivariate GARCH Models." *Journal of Business and Economic Statistics*, Vol. 20, No. 3, pp. 339–350.

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. 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 Statistical Tests in Financial Economics?" Working paper.

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.

Leadbetter, M., G. Lindgren, H. Rootzén (1983): Extremes and Related Properties of Random Sequences and Processes. Springer Verlag, 1st edition.

Lo, A. (2002): "The Statistics of Sharpe Ratios". *Financial Analysts Journal*, Vol. 58, No. 4, pp. 36-52.

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. 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. (2023): *Causal Factor Investing*. Cambridge University Press, 1st edition.

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

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

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.

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

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