Sharpe Ratio Inference:
A New Standard for Decision-Making & Reporting

Prof. Marcos López de Prado
Cornell University – College of Engineering
Lawrence Berkeley National Laboratory – U.S. Office of Science
ADIA Lab – UAE

Seminar’s Objective

- One of the most widely accepted measures of investment efficiency is the Sharpe ratio, which expresses excess return relative to volatility
- While the Sharpe ratio is reported ubiquitously in academic and practitioner publications, the inference done on it is often wrong
- Common mistakes include:
  - Comparing annualized Sharpe ratios, without taking into account its sampling variance
  - Using generic statistical tests that make unrealistic assumptions, such as i.i.d. Normal returns
  - Neglect minimum sample lengths and the power of the test
  - Interpreting the rejection of the null hypothesis as evidence that the null hypothesis is likely false
  - Ignoring multiple testing corrections
- **In this seminar, we propose a new standard for Sharpe ratio inference, which enables decision-making and reporting within a sound statistical framework**
  - **For additional details, find the full paper at https://ssrn.com/abstract=5520741**

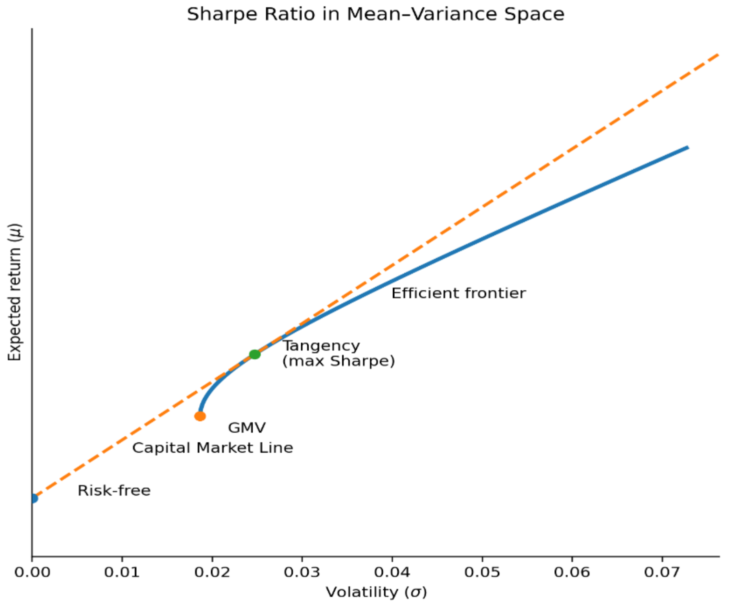
The Sharpe Ratio

Sharpe Ratio

- Consider excess returns (net of risk-free rate) drawn from a population with mean μ and variance σ^2
- These excess returns are allowed to follow a non-Normal distribution, and to exhibit serial correlation. The true (unobserved) Sharpe ratio (SR) is defined as

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

- Key property: Tangency Portfolio has max SR
- The Sharpe ratio can be interpreted as
  - measure of investment skill (signal over noise)
  - measure of investment efficiency (return on risk)



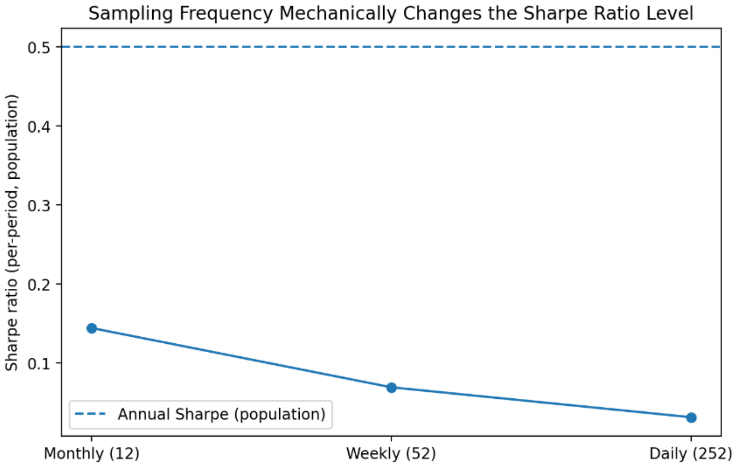
In MPT, the portfolio with the maximum SR is the tangency portfolio. All efficient portfolios are linear combinations of this portfolio and the risk-free asset. In practice, SR has become the dominant evaluation metric.

Sharpe Ratio Estimator

- Parameters μ and σ are unknown
- Consider a sample of T excess returns of an investment strategy, $\{r_t\}_{t=1,\dots,T}$, drawn from a population with mean μ and variance σ^2
- The plug-in estimator of the (observed) Sharpe ratio is

$$\widehat{SR} = \frac{\hat{\mu}}{\hat{\sigma}}$$

  where $\hat{\mu}$ and $\hat{\sigma}$ are maximum likelihood estimators of μ and σ
- **Important: Computing or using the Sharpe ratio does not imply that returns are i.i.d. Normal**



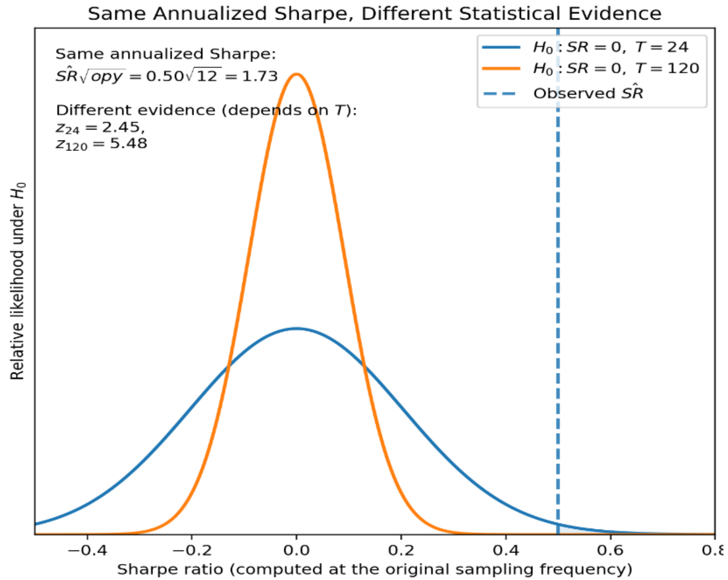
The Sharpe ratio is not invariant to sampling frequency. The level of the observed Sharpe ratio changes mechanically with how often returns are sampled.

Sharpe Ratio Comparisons

- Practitioners often compare Sharpe ratios estimated on different sampling frequencies (daily, weekly, monthly, …) by scaling them into an "annualized" equivalent ($\widehat{aSR}$)
- When returns are independent and identically distributed (i.i.d.), then

$$\widehat{aSR} = \widehat{SR}\sqrt{opy}$$

  where opy is the number of observations per year
- Annualization is a reporting convenience, not a rigorous way of comparing Sharpe ratios
- Correct comparison: Apply significance analysis


Same Annualized Sharpe, Different Statistical Evidence

Comparing annualized Sharpe ratios neglects sample length. As the above plot shows, the same annualized Sharpe ratio can be significant or not (i.e., indistinguishable from no-skill), depending on the sample length (T).

Classification: Public

Lo [2002]

- In a seminal paper, Andrew Lo derived the sampling distribution of the Sharpe ratio as
$$\widehat{SR}=\frac{\hat{\mu}}{\hat{\sigma}}\overset{a}{\sim}\mathcal{N}\left[SR,\frac{1}{T}\left(1+\frac{1}{2}SR^2\right)\right]$$
- The paper originally claimed that this expression only assumed i.i.d. returns
- In response to a reader's letter (Prof. Wolf) published in the journal, Lo later clarified that
  - "As it is written, the asymptotic distribution for the variance estimator […] does indeed require the assumption of normality."
  - "[T]he 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."

A common misconception is that Lo [2002] derived the sampling distribution of the Sharpe ratio without assuming Normality or serial independence.

First, as Lo later acknowledged, his proof still assumes Normal returns.

Second, the "Non-IID Returns" section suggests applying the general method of moments (GMM) to compute the variance of the Sharpe ratio, however it does not derive the actual closed-form expression (see his equations [14] and [A15]).

Surprisingly, despite its obvious practical usefulness, this problem appears to have remained unsolved for 24 years.

Is the “i.i.d. Normal” Assumption Realistic?

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 contribute to increasing the variance of the Sharpe ratio estimator. The following table confirms that the literature’s standard assumptions of Normality and serial independence are unwarranted and unrealistic.

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

Statistics are computed on the monthly returns 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.

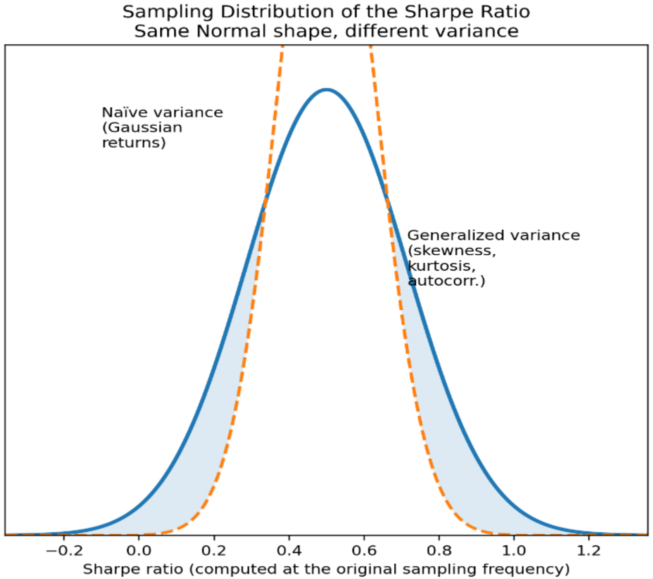
Generalized Sampling Distribution (1/3)

- Following López de Prado et al. [2025], the plug-in estimator for the Sharpe ratio ($\widehat{SR}$) is distributed as

$$\widehat{SR} = \frac{\hat{\mu}}{\hat{\sigma}} \overset{a}{\sim} \mathcal{N}\left[SR, \sigma^2\left[\widehat{SR}\right]\right]$$

$$\sigma^2\left[\widehat{SR}\right] = \frac{1}{T}\left(\frac{1+\rho}{1-\rho} - \frac{1+\rho+\rho^2}{1-\rho^2}\gamma_3 SR + \frac{1+\rho^2}{1-\rho^2}\frac{\gamma_4 - 1}{4}SR^2\right)$$

where 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 excess return's first-order autocorrelation coefficient


Sampling Distribution of the Sharpe Ratio
Same Normal shape, different variance

Levels of skewness and kurtosis typically observed in investments materially depart returns from the Normal distribution. Ignoring this departure underestimates the variance of the Sharpe ratio.

Generalized Sampling Distribution (2/3)

- 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 variance of the Sharpe ratio's estimator under the assumption that $SR = \widehat{SR}^*$ is

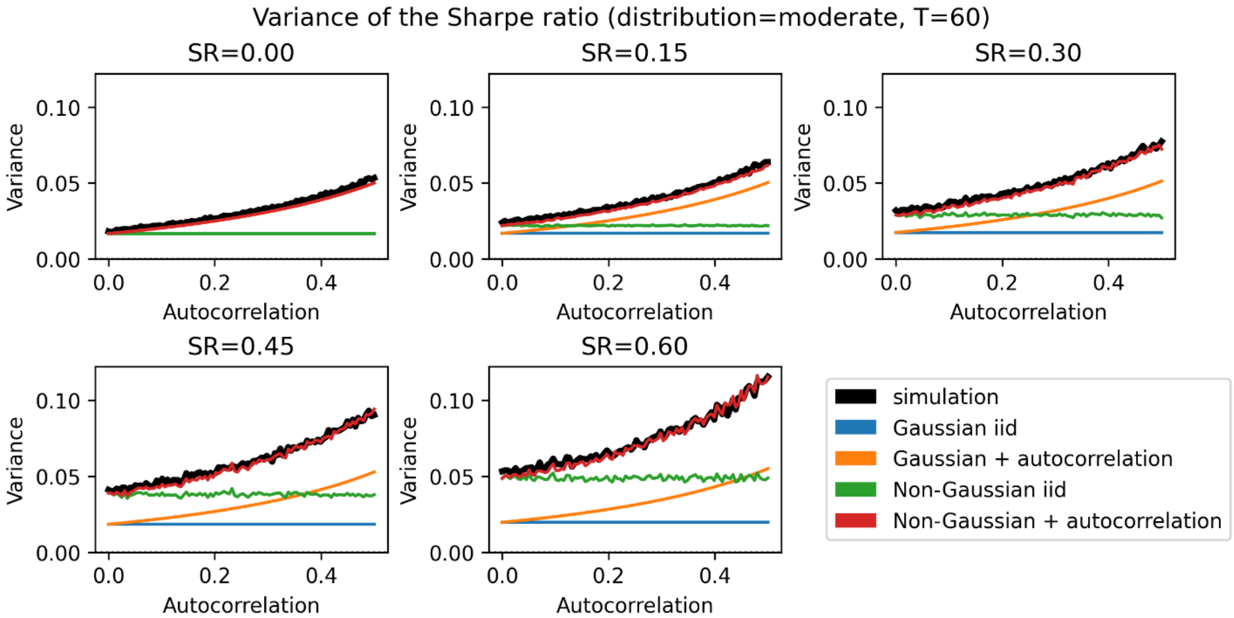
$$\sigma^2\left[\widehat{SR}^*\right] = V\left[\widehat{SR}\,\middle|\,SR = \widehat{SR}^*\right]$$
$$= \frac{1}{T}\left(\frac{1+\hat{\rho}}{1-\hat{\rho}} - \frac{1+\hat{\rho}+\hat{\rho}^2}{1-\hat{\rho}^2}\hat{\gamma}_3\widehat{SR}^* + \frac{1+\hat{\rho}^2}{1-\hat{\rho}^2}\frac{\hat{\gamma}_4-1}{4}\widehat{SR}^{*2}\right)$$

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 a standard deviation of $\sigma\left[\widehat{SR}^*\right] = 0.379$. However, assuming i.i.d. Normally distributed returns, the standard deviation would be approximately 43% smaller, $\sigma\left[\widehat{SR}^*\right] = 0.214$. 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.

Generalized Sampling Distribution (3/3)

We can confirm the accuracy of this sampling distribution with a Monte Carlo experiment. We generate 10,000 returns time series, each representing 5 years' worth of monthly observations ($T = 60$), by drawing returns from Mixtures of Gaussians with different expected Sharpe ratios (0, 0.15, 0.30, 0.45, 0.60), and different coefficients of serial correlation (between 0 and 0.5).



On the left are the results for the moderate non-Normality case: $(\gamma_3, \gamma_4) = (-1.7, 10.5)$. Even though parameters cannot be precisely estimated on small samples, the experiment demonstrates that is better to use those noisy estimates than incorrectly assuming i.i.d. Normality. 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.**

Probabilistic Sharpe Ratio (1/3)

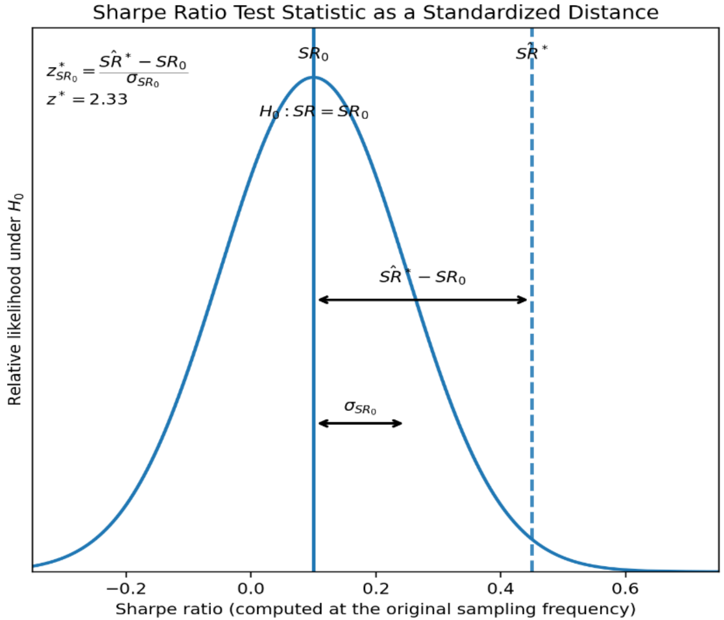
- Following Bailey and López de Prado [2012], we can assess whether $\widehat{SR}^*$ is statistically significant by testing the null hypothesis $H_0: SR \leq SR_0$ against the alternative $H_1: SR > SR_0$
- The test statistic ($z^*[SR_0]$) is

$$z^*[SR_0] = \frac{\widehat{SR}^* - SR_0}{\sigma[SR_0]} \overset{a}{\sim} Z$$

$$\sigma[SR_0] = V\left[\widehat{SR} \middle| SR = SR_0\right]$$

$$= \sqrt{\frac{1}{T}\left(\frac{1+\hat{\rho}}{1-\hat{\rho}} - \frac{1+\hat{\rho}+\hat{\rho}^2}{1-\hat{\rho}^2}\hat{\gamma}_3 SR_0 + \frac{1+\hat{\rho}^2}{1-\hat{\rho}^2}\frac{\hat{\gamma}_4 - 1}{4} SR_0^2\right)}$$

where Z is the standard Normal distribution, SR_0 reflects *the least favorable case* in the null hypothesis.



The test statistic $z^*[SR_0]$ is simply how many standard errors the observed Sharpe ratio lies above the null hypothesis. It allows investors to rank strategies in terms of their ability to overcome a hurdle SR_0.

Probabilistic Sharpe Ratio (2/3)

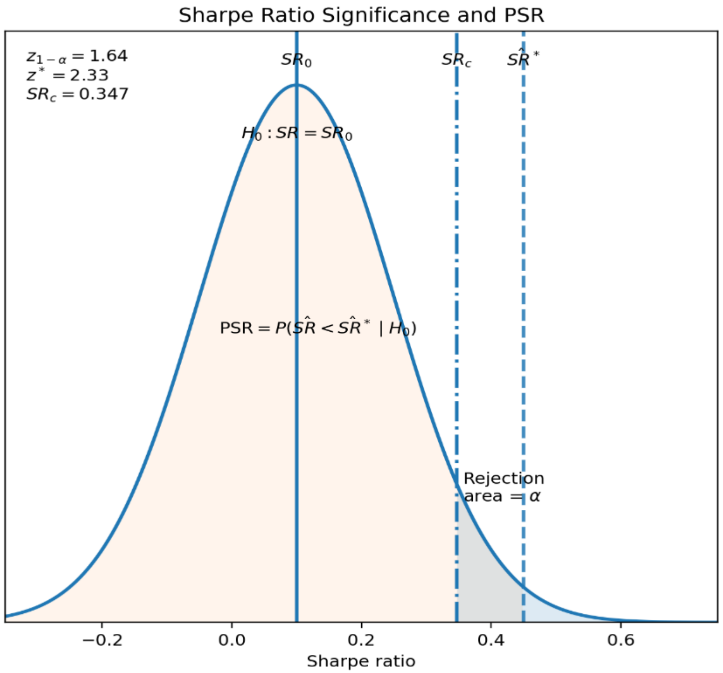
- The significance level α (false positive rate, type I error) is the probability of rejecting H_0 when it is true,

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

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

$$z_{1-\alpha} = Z^{-1}[1-\alpha]$$
$$SR_c = SR_0 + \sigma[SR_0] z_{1-\alpha}$$

- We reject H_0 with confidence $(1-\alpha)$ if $z^*[SR_0] \geq z_{1-\alpha} \Leftrightarrow \widehat{SR}^* \geq SR_c$



Relation between observed value ($\widehat{SR}^*$), least favorable case under the null hypothesis (SR_0), and rejection threshold (SR_c).

Probabilistic Sharpe Ratio (3/3)

- The Probabilistic Sharpe Ratio (PSR) is the **probability of observing a Sharpe ratio less extreme than $\widehat{SR}^*$ subject to H_0 being true,**

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

  where that probability is adjusted for sample length, non-Normality, serial correlation, etc.

- PSR may also be interpreted as the **maximum confidence with which the null hypothesis can be rejected after observing $\widehat{SR}^*$**
- **PSR's generality is a strong reason for preferring it over other tests that assume i.i.d. Normal returns**

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 $SR_0 \neq 0$.

Following with our numerical example, under the null hypothesis where $SR_0 = 0$, then $PSR = Z\left[z^*[0]\right] = Z\left[\frac{\widehat{SR}^*}{\sigma[SR_0]}\right] = 0.966$, but under the null hypothesis where $SR_0 = 0.1$, then $PSR = 0.900$.

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
$$MinTRL = \min_{T}\left\{P\left[\widehat{SR} < \widehat{SR}^{*} \middle| H_0\right] = 1-\alpha\right\}$$
- When $\widehat{SR}^{*} > SR_0$, the solution is

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

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 $(1-\alpha)$.

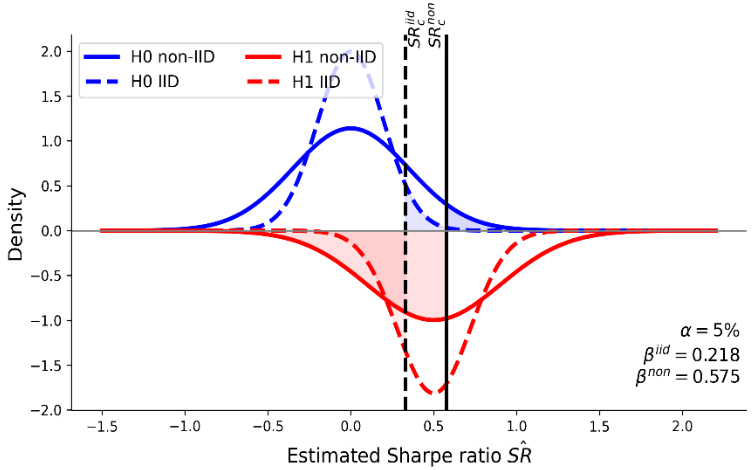
True Positive Rate (Power, Recall, Sensitivity) (1/2)

- Following López de Prado [2020], let SR_1 be the expected value of the alternative hypothesis, $H_1 \colon SR > SR_0$
  - 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,

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

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

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



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%.

True Positive Rate (Power, Recall, Sensitivity) (2/2)

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

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

This table reports precision and recall rates for the Monte Carlo experiment described earlier, where $SR_0 = 0$, with various degrees of non-Normality and values of SR_1 and ρ.

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 adjustment works as designed.

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**

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

  which is the complementary probability to Precision.

- In practice, the value of $P[H_0]$ can be estimated from the proportion of assessed strategies that have yielded negative or around zero excess returns

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 positives 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 $(\alpha, \beta, P[H_0])$, not the observed $\widehat{SR}^*$
- The Sharpe ratio's observed tail-area Bayesian false discovery rate, denoted as oFDR, is the **probability that $\boldsymbol{H_0}$ is true subject to the observed $\boldsymbol{\widehat{SR}}^*$**,

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

$$= \frac{pP[H_0]}{pP[H_0] + (1 - z^*[SR_1])(1 - P[H_0])}$$

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

Following with our numerical example, for $SR_0 = 0, SR_1 = 0.5$ and $P[H_1] = 0.1$, then the *p*-value is $P\left[\widehat{SR} \geq \widehat{SR}^* \middle| H_0\right] = 1 - PSR = 0.034$, while the oFDR is $P\left[H_0 \middle| \widehat{SR} \geq \widehat{SR}^*\right] = 0.361$.

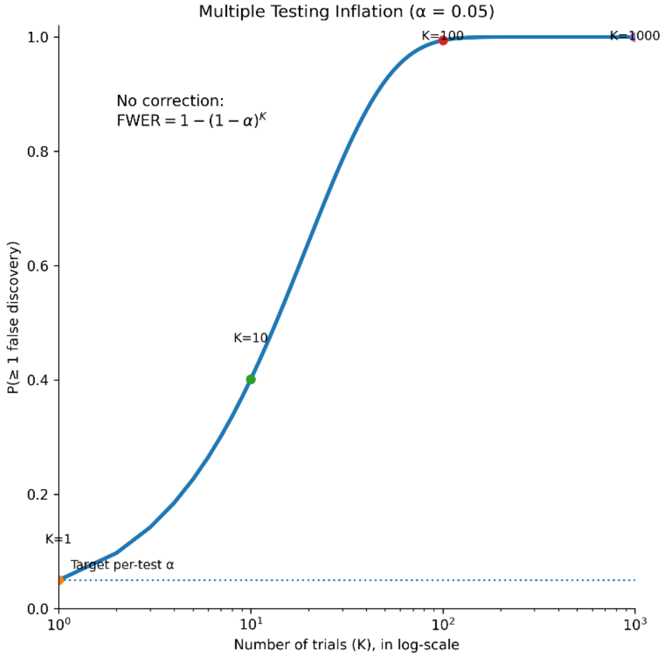
This evidences that an investment may have a statistically significant Sharpe ratio at a 3.4% *p*-value, and yet the probability that the null hypothesis is true can be relatively high (at a 36.1% level), because positives are relatively rare.

Incorrectly assuming that returns are i.i.d. Normal would yield an $oFDR = 0.165$, an underestimation of 54%.

Multiple Testing Controls

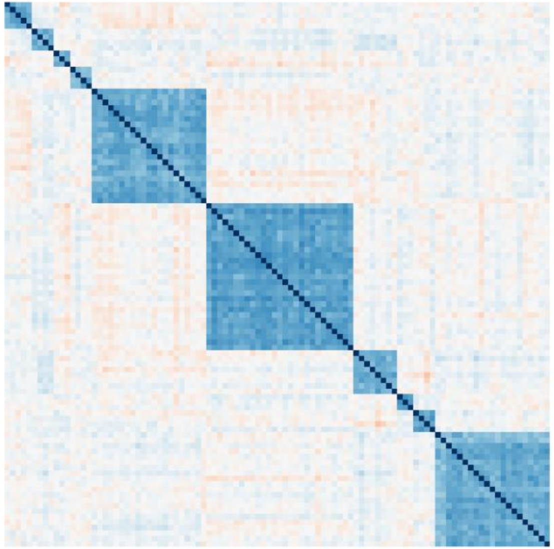
The Fundamental Problem of Multiple Testing

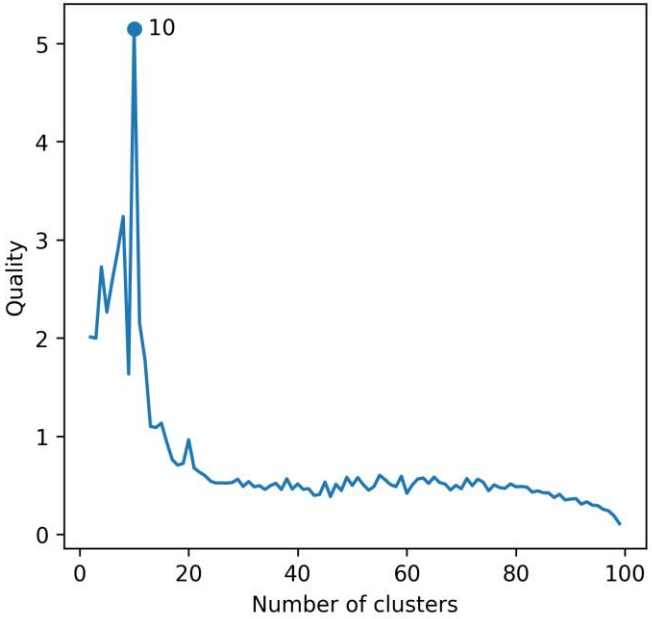
- For $K = 1$ trial, the false positive probability is set to α. However, as $K > 1$ independent trials occur, the probability that there is at least one false positive is α_K,
$$\alpha_K = 1 - (1 - \alpha)^K$$
- Two questions arise naturally:
  - what is the new rejection threshold (SR_c) for the strategy with the highest Sharpe ratio ($\max_k\{\widehat{SR}_k^*\}$), such that it controls for a given α_K?
  - what is the new rejection threshold (SR_c) such the proportion of negatives among the selected strategies (i.e., those with $\widehat{SR}_k^* \geq SR_c$) matches a given level q?
- These are two different questions that control for two different probabilities (FWER & FDR)

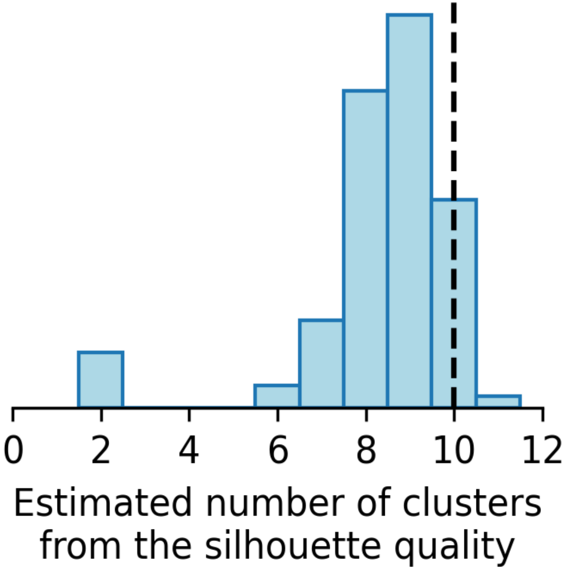


Probability that there is at least one false positive as K increases, when the rejection threshold is not adjusted.

Effective Number of Trials (K)







In practice, the trials are often dependent. When that is the case, K can be derived as the effective number of independent trials, via clustering methods. Once the effective number of trials has been derived, that value can be plugged into the equations.

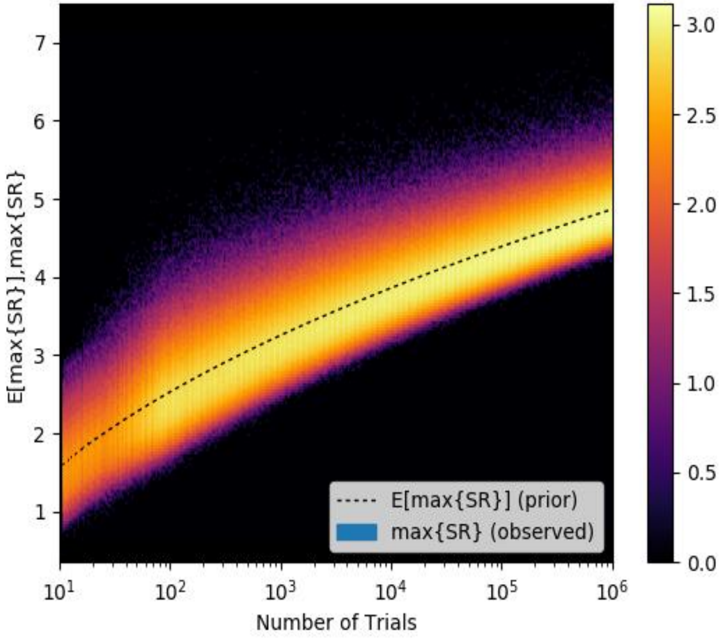
The False Strategy Theorem [2014]

- Consider a sample of K i.i.d. Normal observed Sharpe ratios, $\widehat{SR}_k^* \sim \mathcal{N}\left[SR_0, V\left[\widehat{SR}\right]\right]$
- Bailey and López de Prado [2014] derived:

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

- The standard deviation of the maximum is:

$$\sqrt{V\left[\max_k\left\{\widehat{SR}_k^*\right\}\right]}$$
$$\approx \sqrt{V\left[\left\{\widehat{SR}_k^*\right\}\right]}\sqrt{\frac{\pi^2}{6} - \frac{\gamma^2}{1+\gamma}}\left(Z^{-1}\left[1-\frac{1}{Ke}\right] - Z^{-1}\left[1-\frac{1}{K}\right]\right)$$



Distribution of the maximum Sharpe ratio as a function of K.

Control for Familywise Error Rate (1/2)

- Column Diff reports the error $P\left[\widehat{SR} \geq SR_c \middle| H_0\right] - \alpha$, computed through Monte Carlo experiments
  - FWER adjustments work as designed even for small samples sizes
  - 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
    - apply a different control (e.g., derive the correct SR_c experimentally, via Monte Carlo)

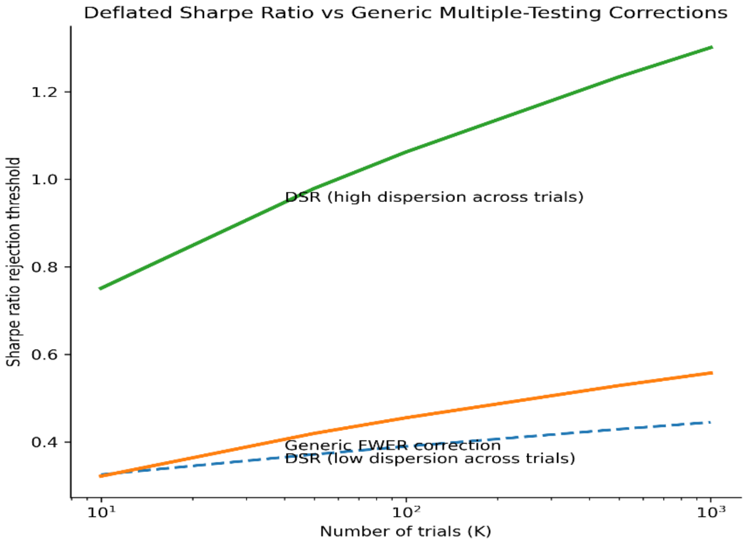
Non-Normality	Skew	Kurt	AR(1)	SR_c	Diff
gaussian	0.0	3.0	0	0.337	0.007
gaussian	0.0	3.0	0.2	0.416	0.008
mild	-0.9	5.6	0	0.340	0.044
mild	-0.9	5.6	0.2	0.416	0.034
moderate	-1.7	10.2	0	0.347	0.073
moderate	-1.7	10.2	0.2	0.418	0.068
severe	-2.3	16.1	0	0.352	0.086
severe	-2.3	16.1	0.2	0.421	0.094

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

FWER control under different processes for $\alpha = 0.05$ for $T = 60$ (top) and $T = 1{,}300$ (bottom).

Control for Familywise Error Rate (2/2)

- Applying these two adjustments to PSR (i.e., replacing SR_0 with $E\left[\max_k\{\widehat{SR}_k^*\}\right]$ and $\sigma[SR_0]$ with $\sqrt{V\left[\max_k\{\widehat{SR}_k^*\}\right]}$) gives the **Deflated Sharpe ratio (DSR)**
- One advantage of DSR over general-purpose correction methods is that it accounts for the true heterogeneity across trials $V[\{SR_k\}]$, which is often associated with overfitting
- Since $V\left[\{\widehat{SR}_k^*\}\right]$ can be much larger than $\sigma[SR_0]$ of the selected model, **DSR should be preferred over generic FWER corrections in financial applications**



Generic FWER corrections ignore cross-trial dispersion, so they under-penalize relative to the Deflated Sharpe Ratio when model search is aggressive.

Classification: Public

Control for Sequential False Discovery Rate (1/3)

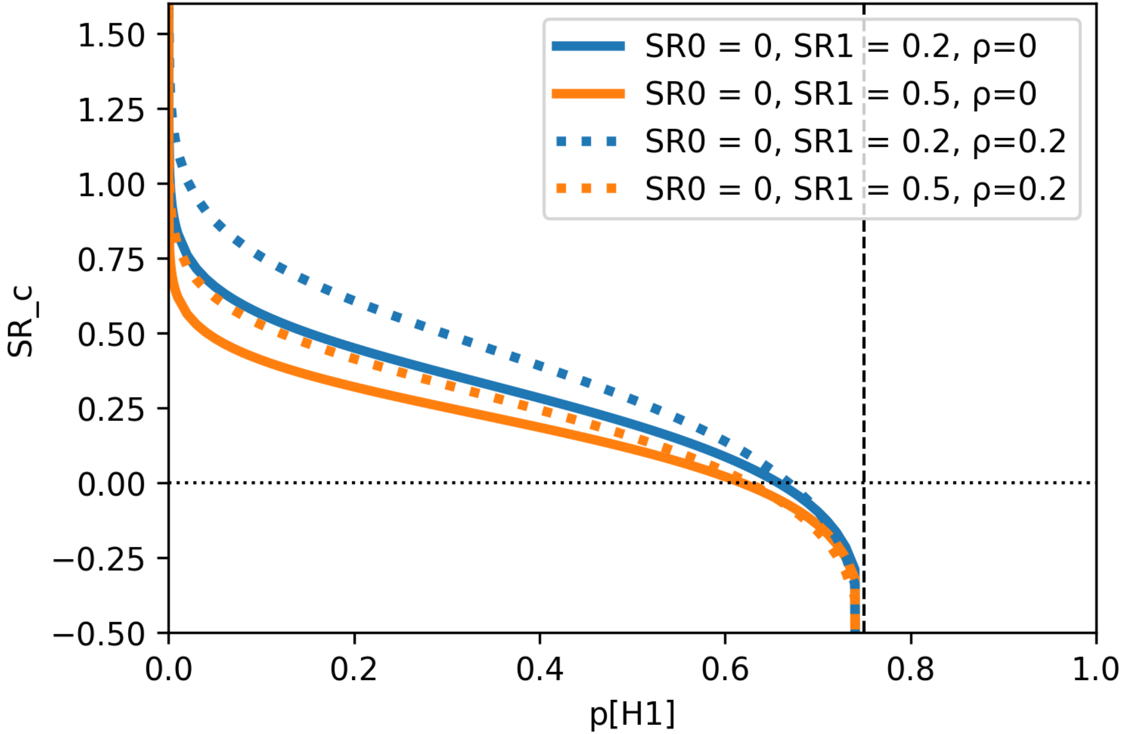
- Suppose that a researcher wishes to select strategies while ensuring that the posterior probability that *each* selected strategy is false does not exceed q
- This departs from the classical Benjamini-Hochberg framework
- The equilibrium occurs at $P\left[H_0 \middle| \widehat{SR} \geq SR_c\right] = q$

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

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 differs from typical investment 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 selections within each meeting.

We introduce the alternative sequential FDR (SFDR) formulation, which controls for the posterior probability of error in each individually approved strategy.

Control for Sequential False Discovery Rate (2/3)



As $P[H_1] \rightarrow (1-q)$, 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.

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.

Control for Sequential False Discovery Rate (3/3)

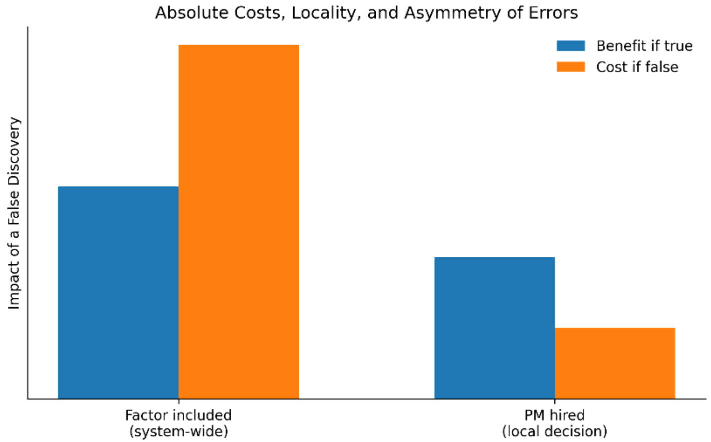
Non-Normality	Skew	Kurt	AR(1)	SR1	P[H1]	SR_c	Precision	Recall	F1	Diff
gaussian	0.0	3.0	0	0.15	0.1	0.397	0.709	0.037	0.071	0.041
gaussian	0.0	3.0	0	0.3	0.1	0.257	0.732	0.628	0.676	0.018
gaussian	0.0	3.0	0	0.45	0.1	0.234	0.731	0.962	0.831	0.019
gaussian	0.0	3.0	0	0.6	0.1	0.231	0.728	0.997	0.841	0.022
gaussian	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
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
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

Monte Carlo experiments demonstrate that SFDR adjustments work as designed even for small sample sizes. The Diff column reports the error $P\left[H_0 \middle| \widehat{SR} \geq SR_c\right] - q$.

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).

Which Control Should be Applied?

- FWER and SFDR measure different probabilities, and the choice depends on the context
- FWER corrections are more appropriate in foundational discoveries, where a false discovery propagates system-wide
  - 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
- SFDR corrections are more appropriate in industrial applications, where a false discovery has localized impact
  - recruitment of portfolio managers, or the selection and defunding of strategies by an investment committee



Several variables inform the FWER vs SFDR decision: (a) Magnitude of error costs; (b) risk of system-wide false discovery propagation; (c) asymmetry between false positive and false negative costs; (d) frequency of decisions; (e) reversibility of the error; etc.

Conclusions

Conclusions

- Valid inference requires addressing five key pitfalls:
  - Using annualized Sharpe ratios for comparison and selection
  - Ignoring sample length, non-Normality and serial correlation
  - Not reporting minTRL, test power
  - Not reporting pFDR, oFDR
  - Not applying multiple testing corrections, such as DSR, SFDR
- Experiments confirm that PSR/DSR provide more reliable inference than
  - classical t-tests
  - general-purpose multiple-testing corrections
- The choice between FWER and SFDR corrections depends on the context
  - FWER is more appropriate in settings where a single discovery supersedes the rest
  - SFDR is better suited for settings where competing discoveries are deployed simultaneously
- **Sharpe ratio remains a valuable tool only if properly adjusted and interpreted**

Conclusions: An Improved Reporting Standard

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

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	Report **DSR**, **SFDR**

Conclusions: Tools for the Strategy Lifecycle

Some inferential tools are particularly important at certain stages of the strategy's lifecycle.

Stage of Lifecycle	**Main Decision**	**Inference Tool**						
		PSR	**MinTRL**	**Power**	**pFDR**	**oFDR**	**DSR**	**SFDR**
Research / backtest (discovery)	Is this pattern signal or noise?	Primary	Useful	Primary	Useful	Useful	Useful	Useful
Replicability analysis (deflation)	Is this "discovery" real after K trials?	Useful	Primary	Primary	Rare	Primary	Primary	Rare
Embargo Analysis (validation)	Is Embargo performance consistent with backtest?	Primary	Primary	Useful	Rare	Primary	Rare	Rare
Investment Committe (sequential approvals over time)	Control long-run posterior error across approvals	Useful	Useful	Useful	Primary	Useful	Useful	Primary
Ramp-up (live testing)	Is execution/data degrading the signal?	Primary	Primary	Useful	Rare	Primary	Rare	Rare
Full deployment (monitoring)	Is alpha decaying? decommission?	Primary	Rare	Rare	Rare	Primary	Rare	Rare

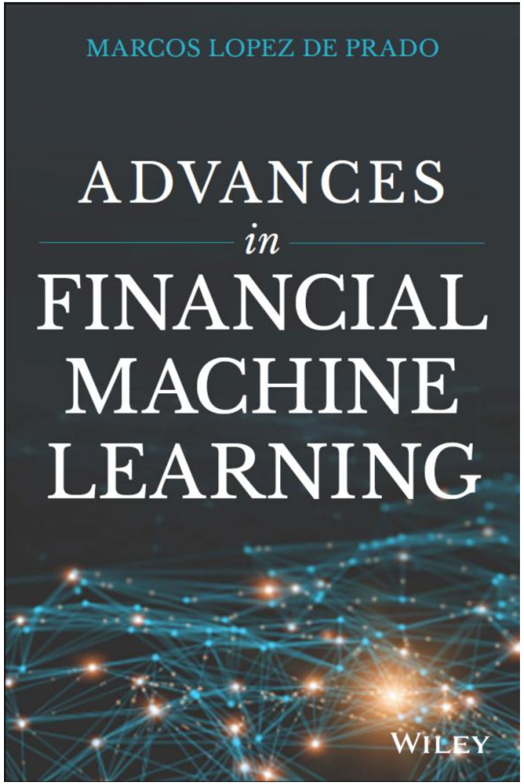
Conclusions: Review of One-Trial Literature

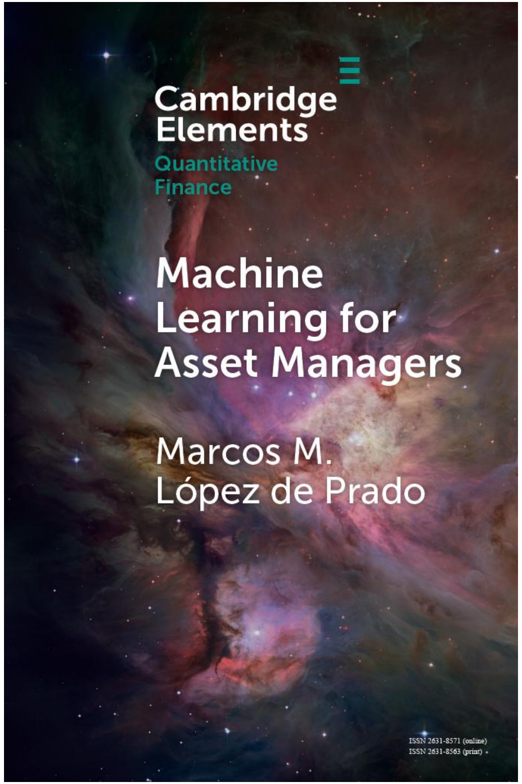
Method	Authors	Correction Type	Sharpe Specific?	Notes
Lo's Significance Test	Lo [2002]	Single-test inference	Yes	Adjusts for sample length, under Normal returns
Bootstrap Test	Ledoit & Wolf [2008]	Single-test inference	Yes	HAC standard errors and a studentized time-series bootstrap
Probabilistic Sharpe 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

Conclusions: Review of Multiple-Trials Literature

Method	Authors	Correction Type	Sharpe Specific?	Notes
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
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

For Additional Details





The first wave of quantitative innovation in finance was led by Markowitz optimization. Machine Learning is the second wave and it will touch every aspect of finance. López de Prado's Advances in Financial Machine Learning is essential for readers who want to be ahead of the technology rather than being replaced by it.
— Prof. **Campbell Harvey**, Duke University. Former President of the American Finance Association.

Financial problems require very distinct machine learning solutions. Dr. López de Prado's book is the first one to characterize what makes standard machine learning tools fail when applied to the field of finance, and the first one to provide practical solutions to unique challenges faced by asset managers. Everyone who wants to understand the future of finance should read this book.
— Prof. **Frank Fabozzi**, EDHEC Business School. Editor of The Journal of Portfolio Management.

Disclaimer

- The views expressed in this presentation are my own, and do not necessarily reflect the views of Cornell University, the Abu Dhabi Investment Authority, or ADIA Lab
- No investment decision or particular course of action is recommended by this presentation
- All Rights Reserved. © 2020 - 2026 by Marcos López de Prado

www.adialab.ae