# THE THREE TYPES OF BACKTESTS

Jacques Joubert, Dragan Sestovic,

Illya Barziy, Walter Distaso,

Marcos Lopez de Prado

This version: 29 July 2024

Jacques Joubert is a quant researcher and developer at the Abu Dhabi Investment Authority (ADIA), in Abu Dhabi, United Arab Emirates. E-mail: jacques.joubert@adia.ae

Dragan Sestovic is a lead quant researcher and developer at the Abu Dhabi Investment Authority (ADIA), in Abu Dhabi, United Arab Emirates. E-mail: dragan.sestovic@adia.ae

Illya Barziy is a quant researcher and developer at the Abu Dhabi Investment Authority (ADIA), in Abu Dhabi, United Arab Emirates. E-mail: illya.barziy@adia.ae

Walter Distaso is a lead quant researcher and developer at the Abu Dhabi Investment Authority (ADIA), in Abu Dhabi, United Arab Emirates. E-mail: walter.distaso@adia.ae

Marcos Lopez de Prado is global head of quantitative R&D at the Abu Dhabi Investment Authority (ADIA), in Abu Dhabi, United Arab Emirates, and a professor of practice at Cornell University, in New York, New York. E-mail: marcos.lopezdeprado@adia.ae

# THE THREE TYPES OF BACKTESTS


## ABSTRACT

Backtesting stands as a cornerstone technique in the development of systematic investment strategies, but its successful use is often compromised by methodological pitfalls and common biases. These shortcomings can lead to false discoveries and strategies that fail to perform out-of-sample. This article provides practitioners with guidance on adopting more reliable backtesting techniques by reviewing the three principal types of backtests (walk-forward testing, the resampling method, and Monte Carlo simulations), detailing their unique challenges and benefits. Additionally, it discusses methods to enhance the quality of simulations and presents approaches to Sharpe ratio calculations which mitigate the negative consequences of running multiple trials. Thus, it aims to equip practitioners with the necessary tools to generate more accurate and dependable investment strategies.


**Three main takeaways:**

1. Practitioners can design better experiments by understanding the advantages and disadvantages of the three main types of backtesting.
2. Practitioners can enhance backtest quality by avoiding methodological pitfalls and common biases.
3. By understanding the importance of selection bias under multiple testing and how to run statistical tests, practitioners can improve the odds of making a true discovery.


**Keywords:**
Backtesting, simulation, selection bias under multiple testing, evaluation, best practices.

**JEL Classification:**
G0, G1, G2, G11, G15, G17, G24, C58

A backtest simulates the performance of a systematic investment strategy, with the goal of assessing its expected risk and return going forward. It provides key evaluation metrics to determine the replicability of a discovered investment algorithm, and to understand the associated risk and return profile. However, backtesting is frequently misapplied; a lack of awareness about common pitfalls often results in overfitting, leading to the deployment of investment strategies that fail to perform well out-of-sample.

In this article, we explore the nuances associated with conducting effective backtests, dividing our discussion into three main sections. First, we discuss the three principal types of backtests. Next, we address common pitfalls and provide guidance on how practitioners can improve the quality of their simulations. Finally, we examine the issue of selection bias under multiple testing and demonstrate how it contributes to overfitting.

## TYPES OF BACKTESTS

There are three principal methods of conducting a backtest: walk-forward testing, resampling, and Monte Carlo simulations.

### The Walk-Forward Backtest

The most widely used backtesting method is walk-forward testing, also referred to as historical backtesting, in which the strategy is assessed against a series of subsequently observed events and asset price moves from a past period. Though the application of this method might appear simple, inaccuracies in the setup and execution can lead to biased results. Even when all the potential pitfalls are circumvented, the walk-forward approach is still path dependent and assumes that the processes and events of the past repeat in the future.

The main benefit of the walk-forward method is that the results and performance characteristics are easy to analyse, interpret, and compare between different periods. Additionally, it does not require the practitioner to identify an appropriate resampling technique, or the data generation process required by the other two backtesting methods.

One of the principal limitations of the approach is that only a single path is tested, which raises the risk of overfitting (Bailey and Lopez de Prado 2014). Another shortcoming is that the observed past performance may not be indicative of future results since the underlying relationships that the strategy aims to exploit may not repeat. This leads to the critical observation that the walk-forward method, by definition, does not require practitioners to have a comprehensive understanding of the underlying processes or knowledge of the reasons for a strategy's success or failure. Therefore, the obtained results could be based on erroneous assumptions.

### The Resampling Method

The second type of backtest encompasses variations of the inferential statistics' method known as resampling. As its names suggests, this method resamples past observations to construct new paths, thereby overcoming a primary drawback of the walk-forward approach. This method includes techniques such as cross-validation and bootstrapping. Cross-validation involves splitting

observations into groups and alternating between them for training and validation purposes, with methods including K-fold and combinatorial cross-validation, alongside the recommended practices of purging and embargo detailed by Lopez de Prado (2018). Bootstrapping, on the other hand, entails drawing observations from the sample at random using varying logic, multiple examples of which are covered by Musciotto et al. (2018). The assumption is that backtest outcomes across a spectrum of trajectories with different out-of-sample periods are indicative of a strategy's future performance.

The advantage of these methods is the availability of multiple paths to evaluate the strategy and check its robustness. Additionally, they yield a set of performance metrics rather than a single point, facilitating a more detailed analysis and comparison of strategies. Lastly, resampling methods are less prone to overfitting, as discussed by Lopez de Prado (2018).

Nonetheless, certain disadvantages identified in the walk-forward method persist. For instance, the observed performance across multiple past paths may not accurately represent the future, and again, the method does not require a reasoned understanding of why the strategy works. An added difficulty associated with resampling is that some underlying data that the strategy uses may not be in a suitable format for the bootstrap approach, or it might require data to be segmented into longer periods to preserve the data's time-dependence, which diminishes the number of groups and observed paths.

**Monte Carlo Simulations**

The third backtesting approach is the Monte Carlo method, which requires an understanding of the data generation process for its construction. This knowledge can be derived from theoretical constructs and causal relationships between the processes that the strategy aims to exploit, or through statistical analysis.

The underlying assumption is that future paths can be modelled using Monte Carlo simulations. This approach enables the generation of additional data with properties resembling those observed in the real data, thereby providing researchers with a more solid foundation for analysis compared to the resampling method. Wiese et al. (2020) and Li et al. (2020) illustrate a data generation process using neural networks, which can subsequently be employed for strategy backtesting.[1]

The main benefit of this method is that the observed performance is indeed indicative of future outcomes, provided the data generation process is unchanged and correct. To construct a backtest following this approach, it is necessary to understand why the strategy works, thereby addressing one of the main limitations inherent in the other two methods. Moreover, the risk of overfitting is mitigated when employing Monte Carlo simulations.

In addition, during live trading, the statistical properties of the live data can be monitored and compared to the hypothesised data generation process, and the detection of a structural break

[1] For a more comprehensive exploration of the techniques involved and discussions of Monte Carlo backtests, see Lopez de Prado (2018) and Fabozzi et al. (2021).

could act as an early warning system to stop trading and prevent losses. In a situation where not enough data was initially available to evaluate a strategy, additional data can also be generated to reach the required confidence levels.

The primary challenge of this method is the intricate task of creating an accurate data generation process for all the data sources utilised by the strategy. It may be difficult to replicate historical periods for extraordinary events like the 2008 financial crisis or the COVID-19 pandemic; however, one can create several data-generation processes that include novel scenarios which have not appeared previously.

## ENHANCING THE QUALITY OF HISTORICAL SIMULATIONS

In this section, we discuss how to increase the quality of a backtest by focusing on six areas known to present challenges, namely: data quality, data representativeness, statistical integrity, modelling and generalisation, costs and constraints, and performance evaluation.

### Data Quality

**Survivorship Bias:** Survivorship bias is defined as a type of selection bias that occurs when analyses are conducted only on the data that have survived a selection process while ignoring those that did not. This bias can lead to skewed results and incorrect conclusions because the non-surviving entities typically differ from the survivors in significant ways.

The impact of survivorship bias is particularly pronounced in the analysis of mutual and hedge fund performances, as well as investment strategies. For example, including only those funds that are still active at the end of the period can significantly overestimate the average fund performance, as the worst-performing funds go out of business and are not reported on.[2]

**Point-in-Time Considerations and Restated Data:** Point in time refers to ensuring that data used in analysis or backtesting is reflective of the information that would have been available to researchers at that specific point in time. This is particularly relevant in the case of restated data, which involves adjustments made to previously reported financial figures, such as earnings, revenue, or other financial metrics, due to errors, accounting changes, or compliance with new accounting standards. Restatements can significantly alter historical financial statements, affecting the accuracy of financial analysis and the models that rely on those statements. For practitioners, it is essential to manage restated data carefully; this often means using a point-in-time database to capture financial data as it was originally reported, without incorporating restatements or adjustments made after the fact.

**Incorrect and Missing Data:** Incorrect data are any errors or inaccuracies in the dataset. These inaccuracies can stem from various sources, such as data entry errors, issues with data collection methods, or problems in data transmission. Before analysis, data should be thoroughly

[2] See Carhart (1997), Malkiel (1995), Brown et al. (1992), and Elton, Gruber, and Blake (1996).

cleaned, which involves checking for and correcting errors, ensuring consistency across datasets, and verifying data accuracy against reliable sources.

Missing data occurs when information is absent from the dataset. This can happen for various reasons, such as system errors, data corruption, or when the data was never recorded or collected in the first place. Missing values can be imputed using assorted statistical techniques, such as mean imputation or regression imputation, or more sophisticated methods like multiple imputation or machine learning-based approaches. Depending on the context, forward filling may be a good option to avoid introducing additional complexity through imputation. In some cases, missing data can be augmented from alternative sources, although this approach requires careful consideration to ensure the compatibility and reliability of the augmented data.

**Dealing with Outliers:** In the development of investment strategies, it is essential to evaluate the role of outliers – that is to say, extreme values that deviate significantly from the rest of the data. Strategies that capitalise on these rare occurrences might not be sustainable, as the outliers may represent unique, non-recurring events. Consequently, relying on such anomalies could lead to strategies that perform exceptionally well under specific conditions but fail to generalise across different market environments.

When incorporating models that are inherently sensitive to outliers, such as linear regression, care must be given to the nature and impact of these extreme data points. Practitioners must ascertain whether outliers arise from errors in data collection or entry – such as misreported financial figures – or if they reflect genuine market phenomena that could offer valuable insights.

### Data Representativeness

Sample period selection bias is the primary concern in the area of data representativeness. This type of bias occurs when the time frame chosen for a backtest inadvertently influences the results, leading to conclusions that may not be robust across different market conditions or time periods. When designing all-weather investment strategies (Lopez de Prado 2019), it is essential to include data from a wide range of market conditions rather than periods of optimal performance alone. This ensures the strategy's robustness across different market cycles. However, when creating tactical investment strategies, which, as Lopez de Prado (2019a) describes, are designed for specific market conditions, it may be more appropriate to select targeted time frames. The Min-Track-Record-Length algorithm described by Bailey and Lopez de Prado (2012) can help determine the number of observations needed to validate a strategy's effectiveness across diverse market scenarios.

### Statistical Integrity

**Data Mining and Data Snooping:** The concepts of data mining and data snooping refer to the process of searching through large datasets to identify patterns, relationships, and trends that can be used to develop investment strategies. This approach to strategy development should be avoided, as it very easily leads to selection bias through the improper application of data analysis techniques to uncover patterns in data that appear to be of statistical importance. Data snooping (p-value hacking) involves researchers repeatedly probing various subsets of data or conducting

numerous tests (potentially on the same data) until they achieve a result that seems meaningful (White 2000; Sullivan, Timmermann, and White 1999).

**Accounting for Selection Bias under Multiple Testing:** Backtests need to account for the number of trials run and their statistics so that a discount measure can be applied to the performance metrics (Lopez de Prado 2018, 2020). This is covered in greater detail in the next section.

### Modelling and Generalisation

**Look-Ahead Bias:** Look-ahead bias is the mistake of using data from the future as if it were point-in-time data to make investment decisions. We separate this from the problem of not using point-in-time data as we define it because of the research process. This approach can lead to inflated performance metrics and unrealistic expectations as to the profitability of investment strategies.

Most typically, this mistake is made by not applying the appropriate lag to the indicators or signals. It also commonly occurs when computing statistics on the entire data sample rather than a rolling computation – for instance, using the entire sample when computing the mean and standard deviation of a z-score to normalise a signal.

**Introducing an Embargo Period:** An embargo period is a hold-out sample of data, a test set, typically the most recent two to three years. Practitioners should fit models and make design choices regarding their trading strategies based on the in-sample period, and once the strategy is ready to be validated, the results can be produced for the embargo period. Any market anomaly that was exploited in the training set should also be present in the embargo period.

### Costs and Constraints

An oft-overlooked step is the incorporation of trading costs and the constraints relating to short selling and liquidity.

**Transaction Costs:** Neglecting to account for transaction costs leads to a higher false positive rate. Higher turnover strategies incur greater expenses, necessitating a more substantial return to be viable. Omitting these costs from an evaluation leads to inflated performance metrics.

Borkovec and Heidle (2010) provide a good introduction to the types of costs and their impact on trading. Notably, they split costs into three types. The first, brokerage costs, are easily estimated using historical data and include commissions and fees; two examples of fees are custodial and transfer fees. The second, trading costs, consist of four components: delay costs, bid-ask spread, market volatility/trend costs, and market impact. Finally, opportunity costs are the costs of not completing the full order. The authors highlight that market impact has the highest cost (and is difficult to model), whereas commissions and spread have the lowest effect and are easily estimated. Transaction costs can be estimated either empirically, using historical data, or analytically by attempting to build a model.

**Short-Sale Constraints:** Another frequently overlooked cost is the cost of borrowing stocks to enable short selling. Estimating these costs can be challenging as brokers can group these

costs with other services (Saffi and Sigurdsson 2011). Additionally, as the demand for shorting rises, the availability of stocks to short decreases. During a financial crisis, several countries may restrict short selling, with some countries even imposing regulations such as the uptick rule (Boehmer, Jones, and Zhang 2013). Khan (2024) offers a comprehensive review paper on this topic.

**Liquidity Constraints:** A backtest must take into consideration the liquidity of the instruments being traded. Typically, trading will be capped at some percentage of the daily average trading volume for each security. This constraint is imposed to ensure that the execution of trades will not significantly impact market prices.

**Universe Selection:** Careful selection of an investment universe may help to limit the effects of both liquidity and short-sale availability; this may involve only trading assets with a market cap greater than $500 million, a median daily dollar volume greater than $5 million, a price greater than $6, and 500 days of price and volume data.

**Performance Evaluation**

**Causal Graphs:** The narrative fallacy involves a cognitive bias where practitioners create a story-like explanation for a series of events, often adding causality and meaning where there may be none. Events may be misinterpreted due to our innate desire to fit them into a familiar narrative structure. To combat this, Lopez de Prado (2023) and Lopez de Prado and Zoonekynd (2024) recommend that before running a single backtest, practitioners create a causal graph and an explanation of the sequence of events to describe the anomaly which is to be exploited for profit.

**Performance Metrics:** A popular measure of the performance of a trading strategy is the Sharpe ratio (Sharpe 1966), defined as the ratio of expected excess return over the risk-free rate to its standard deviation:

$$\text{SR} = \frac{\mu - r_f}{\sigma} \tag{1}$$

where $\mu = \text{E}(r_t)$ is the expected return of a strategy or asset, $r_f$ is the risk-free rate, and $\sigma$ is the standard deviation of returns. It thus expresses the return of a strategy per unit of risk. The unknown quantities $\mu$ and $\sigma$ are estimated using realised information via, respectively, $\bar{r} = \sum_{t=1}^{T} r_t/\text{T}$ and $\hat{\sigma} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(r_t - \bar{r})^2}$, where T denotes the sample size.

Although the formal definition of the Sharpe ratio is based on the notion of excess return, it is often reported under the implicit assumption that $r_f = 0$. Also, the Sharpe ratio is usually reported in annualized terms. This involves rescaling the numerator and denominator of (1) accordingly (e.g., in the case of daily returns, the numerator is multiplied by 252 (trading days in a year) and the denominator by $\sqrt{252}$).

The asymptotic distribution of the estimated Sharpe ratio has attracted considerable interest in the literature. Lo (2002), Mertens (2002), and Christie (2005) have derived the limiting distribution under the assumption of stationarity and ergodicity. The formula for the limiting variance derived by Mertens (2002) has been shown to hold when observations are not

independently and identically distributed Normal, see also Ledoit and Wolf (2008), Liu, Rekkas, and Wong (2012) and Qi, Rekkas, and Wong (2018) for higher-order accuracy.

Asymptotically, the distribution of the estimated Sharpe ratio is given by:

$$\sqrt{T}(\widehat{SR} - \mathrm{SR}) \xrightarrow{d} \mathrm{N}\left(0{,}1 + \frac{1}{2}\mathrm{SR}^2 - \gamma_3 \mathrm{SR} + \frac{\gamma_4 - 3}{4}\mathrm{SR}^2\right) \tag{2}$$

where N denotes a Normally distributed variate, $\gamma_3$ and $\gamma_4$ are, respectively, the raw third and fourth moments of returns. A feasible implementation of the result in expression (2) entails estimating the unknown quantities in the limiting variance by their sample counterparts. The Sharpe ratios of strategies that exhibit high excess kurtosis and/or negative skewness will be estimated with less precision.

Using the above results, Bailey and Lopez de Prado (2012) derived the Probabilistic Sharpe Ratio (PSR) as

$$PSR[SR^*] = \mathrm{Z}\left[\frac{(\widehat{SR} - SR^*)\sqrt{T-1}}{\sqrt{1 - \hat{\gamma}_3\widehat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\widehat{SR}^2}}\right] \tag{3}$$

where $Z$ is the CDF of the standard Normal distribution, $SR^*$ is the value of the null hypothesis, $T$ is the number of observations, $\hat{\gamma}_3$ is the estimated skewness, and $\hat{\gamma}_4$ is the estimated kurtosis. PSR allows researchers to express the Sharpe ratio in terms of the probability of observing performance statistically significant above a hurdle $SR^*$, after correcting for the number of observations, skewness and kurtosis. PSR is particularly useful in the context of hedge funds, where returns often exhibit negative skewness and positive excess kurtosis, which result in inflated Sharpe ratios. Furthermore, Bailey and Lopez de Prado (2012) also introduced the Minimum Track Record Length (MinTRL) statistic as

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

where $\mathrm{Z}_\alpha$ is the critical value associated with a significance level $\alpha$. The MinTRL formula shows that a longer track record will be required the smaller $\widehat{SR}$ is, or the more negatively skewed returns are, or the greater the fat tails, or the greater our required level of confidence. A first practical implication is that, if a track record is shorter than MinTRL, we do not have enough confidence that the observed $\widehat{SR}$ is above the designated threshold $SR^*$. A second practical implication is that a portfolio manager (PM) will be penalized because of her non-Normal returns, however the PM can regain the investor's confidence over time (by extending the length of her track record).

Despite its widespread use as a performance measure, the Sharpe ratio has some well-known limitations. Although it is suitable for liquid investments with Normal returns, it may be less so for highly non-Normal returns that involve the use of derivatives instruments with nonlinear payoffs. This has prompted practitioners to complement the information provided with measures that focus mostly on the left tail of the distribution of returns. Notable examples are:

1. Maximum drawdown, calculated as:

$$\max_{\tau \in (0,\tau)} \left[ \frac{\max_{t \in (0,\tau)} X_t - X_\tau}{\max_{t \in (0,\tau)} X_t} \right] \quad (5)$$

where $X_t$ is the current value of the investment.

2. The Value at Risk (VaR) at level $\alpha$ (typically 5% or 1%), often reported for regulatory purposes and defined as:

$$VaR_\alpha(X) = Min(c : P(X \leq c) \geq \alpha) \quad (6)$$

for a random variable *X* representing losses.

3. Conditional value at risk (cVaR) at level $\alpha$, calculated as follows:

$$cVaR_\alpha(X) = E(X|X \geq VaR_\alpha(X)) \quad (7)$$

4. The Calmar ratio, calculated as the ratio between average annualized return and maximum drawdown

5. The Sortino ratio, the ratio between average return and downside deviation (i.e., the standard deviation of negative returns)

To mitigate the pitfalls associated with the Sharpe ratio in the context of highly non-Normal returns, Favre and Galeano (2002) and Gregoriou and Gueyie (2003) have proposed the modified Sharpe ratio, where the denominator in equation (1) is substituted by $VaR_\alpha$. A test comparing the modified Sharpe ratios of two strategies has been developed by Ardia and Boudt (2015).

**Holistic Evaluation of Metrics:** Performance analysis, much like the analysis of financial accounting ratios, needs to be conducted holistically, highlighting that several performance metrics need to be considered, in combination to make an informed decision. A typical example of ignoring this principle would be to only look at annualized returns, without considering risk-adjusted returns or drawdowns/time-under-water.

**Peer Review:** An independent peer review process may further improve the likelihood of discovering a true positive as reviewers check the source code for bugs, validate claims made by researchers, and critique the methodology used. The reviewers may also run statistical tests to confirm that the true Sharpe ratio is statistically significant, while controlling for multiple testing (Bailey and Lopez de Prado 2014; Harvey and Liu 2015).

## SELECTION BIAS UNDER MULTIPLE TESTING

In this section, we take a closer look at the dangers of selection bias under multiple testing (SBuMT) using the following numerical example. Suppose a researcher has finished their analysis and reported a "promising" trading strategy to the PM. The backtest shows an annualised Sharpe ratio of around 1, with an annualized realized volatility of around 15%. The backtest provides daily returns for the past 5 years as an output, thus the number of observations is $T = 5 \times 252 = 1{,}260$.

The PM tries to assess how statistically significant this discovery is. They know that given the "true value" of the Sharpe ratio *SR*, a realized Sharpe ratio could be somewhere in the vicinity of the true value, with the probability defined by a Normal distribution (Bailey and Lopez de Prado 2012) and the standard deviation given by:

$$\sigma_{SR} = \sqrt{\frac{1}{T}\left(1 - \gamma_3 SR + \frac{\gamma_4 - 1}{4} SR^2\right)} \tag{8}$$

where:

- $T$ is the number of observations
- $\gamma_3$ is the skew of the distribution
- $\gamma_4$ is the kurtosis of the distribution (for a Normal distribution $\gamma_4 = 3$)

Based on this insight, the PM runs a standard statistical test of the reported value $y = 1$ against the null hypothesis that the true Sharpe ratio is actually $SR^* = 0$. To run the standard test, they need to calculate the cumulative probability:

$$F_Y(y) = P(SR \leq y) \tag{9}$$

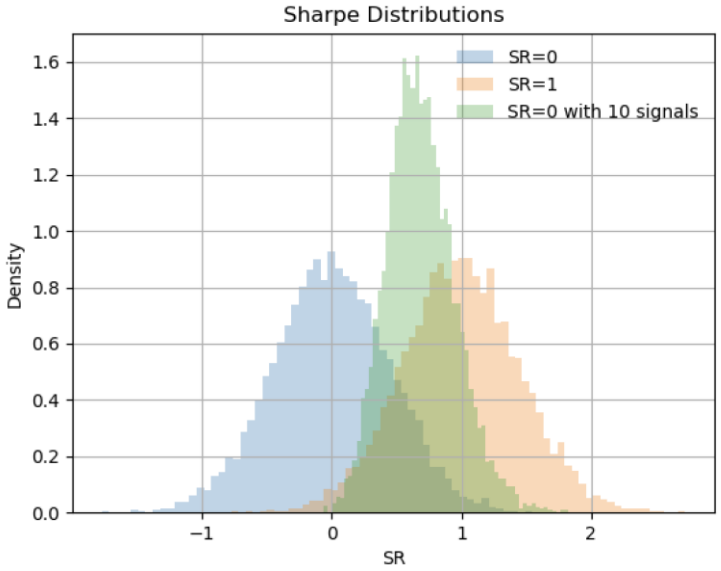
Assuming a Normal distribution of SR values and the distribution described by (8), they calculate the p-value of the test as $p = F_Y(y) = 0.02$ and the annualized cut-off value $SR_{cutoff} = 0.74$. The reported SR is well above the cutoff value, and the p-value is below 0.05, meaning there is less than a 5% chance that the null hypothesis could produce this result. The PM can reject the null hypothesis with 95% confidence and accept the researcher's report as a "new discovery".

However, there is a catch. The PM also learns that 10 different trials have been attempted and that the researcher has only reported the trials with the highest SR. This is a typical situation described by SBuMT.

For illustrative purposes, Exhibit 1 shows the histograms obtained by running 10,000 simulations. The histograms show the realized Sharpe ratios for three scenarios:

1. Blue with expected SR = 0
2. Amber with expected SR = 1
3. Green with expected maximum SR of a sample of 10 trials with expected SR = 0

Exhibit 1: Distribution of Sharpe Ratios




The mean of the green distribution is around SR = 0.7. This is quite close to the cut-off value of the standard 1-trial statistical test, which means that roughly half of the realized maximum SRs would be considered significant with the standard 1-trial statistical test.

This shows how easy it is to find an apparently significant strategy among *N* random walks with zero mean. The green distribution is narrower than distributions under the 1-trial hypothesis, which tells us that it is not appropriate to use $\sigma_{SR}$ calculated by equation (8) to test the SR significance under a multiple-trial assumption. Note also that the distribution of expected maximum SRs deviates from Normal, but not drastically.

The PM realizes that the standard test that was run previously is neither sufficient nor appropriate and that the reported Sharpe ratio should be tested against a different null hypothesis that takes these multiple trials into consideration. The new null hypothesis is that the reported Sharpe ratio is a maximum of K = 10 values obtained from zero-mean random independent trials. The cumulative distribution of interest for this test is then:

$$F_Y(y) = P(Max(\{SR_k\}) \leq y) = P\big((SR_1 < y), \ldots, (SR_K < y)\big) \tag{10}$$

There are two main approaches the PM can take at this point. The first approach is to estimate the family-wise error rate (FWER) by correcting the significance level $\alpha$. Assuming that the Sharpe ratios in the family of tests are independent, we can write:

$$F_Y(y) = P(Max(\{SR_k\}) \leq y) = P\big((SR_1 < y), \ldots, (SR_K < y)\big) = \prod_{k=1}^{K} P(SR_k < y) \tag{11}$$

The null hypothesis is rejected when $F_Y(y) < 1 - \alpha$, or equivalently when:

$$\prod_{k=1}^{K} P(SR_k < y) < (1 - \alpha_k)^K \tag{12}$$

We require that the Type 1 error rate be $\alpha = 1 - c$, i.e., that the test rejects the null hypothesis $\alpha N$ times on average if the null hypothesis is true. To achieve this goal, we need to adjust (correct) the error rate $\alpha_k$ for each single test. The correction follows from

$$1 - \alpha = (1 - \alpha_k)^K \tag{13}$$

This leads to the Sidak correction (Sidak 1967):

$$\alpha_k = 1 - (1 - \alpha)^{\frac{1}{K}} \tag{14}$$

and a simpler approximation of it, the Bonferroni correction (Bonferroni 1936):

$$\alpha_k = \frac{\alpha}{K} \tag{15}$$

With these corrections, we aim to have less than a 5% chance of rejecting the null hypothesis after testing a group of K independent trials, where the null hypothesis is correct for each strategy.

The Sidak correction leads the PM to a value of $\alpha_k = 0.0051$, which is greater than the p-value of the single test performed on the best strategy, and the cut-off SR value is 1.15, which is above the reported Sharpe ratio of 1. These results mean that the PM cannot reject the null hypothesis under the multiple trial assumption, and they conclude that the SR is not statistically significant. See Lopez de Prado (2020) for further details.

The second approach is to derive a more accurate control for FWER from Extreme Value Theory. The maximum value of the list of Gaussians $Y = Max(X_1, X_2, \ldots, X_K)$ has cumulative distribution given by:

$$F_Y(y) = P(Max(\{X_k\}) \leq y) = \prod_{k=1}^{K} F_{X_k}(y) \tag{16}$$

where $F_{X_k}(\mathrm{y})$ is the cumulative distribution function (CDF) of the process generating $X_k$. It can be shown that, under independent and identically distributed draws, the probability density function (PDF) is:

$$f_Y(y) = \frac{\partial}{\partial y} F_Y(y) = K\, \phi(y) \mathrm{Z}(y)^{K-1} \tag{17}$$

where $\phi(.)$ is the PDF of the standard Normal distribution, and $\mathrm{Z}(.)$ is the CDF of the standard Normal distribution (see Embrechts, Klüppelberg, and Mikosch 2013).

This distribution's first two moments can be calculated using numerical integration. If we assume that the SR values are drawn from a Normal distribution $\mathrm{N}(\eta, \sigma^2)$, we can conclude that:

$$E[Y] = \eta + \sigma \int_{-\infty}^{\infty} y\, f_Y(y) dy \tag{18}$$

and

$$Std[Y] = \sigma \sqrt{\int_{-\infty}^{\infty} (y - E[Y])^2 \, f_Y(y) dy} \tag{19}$$

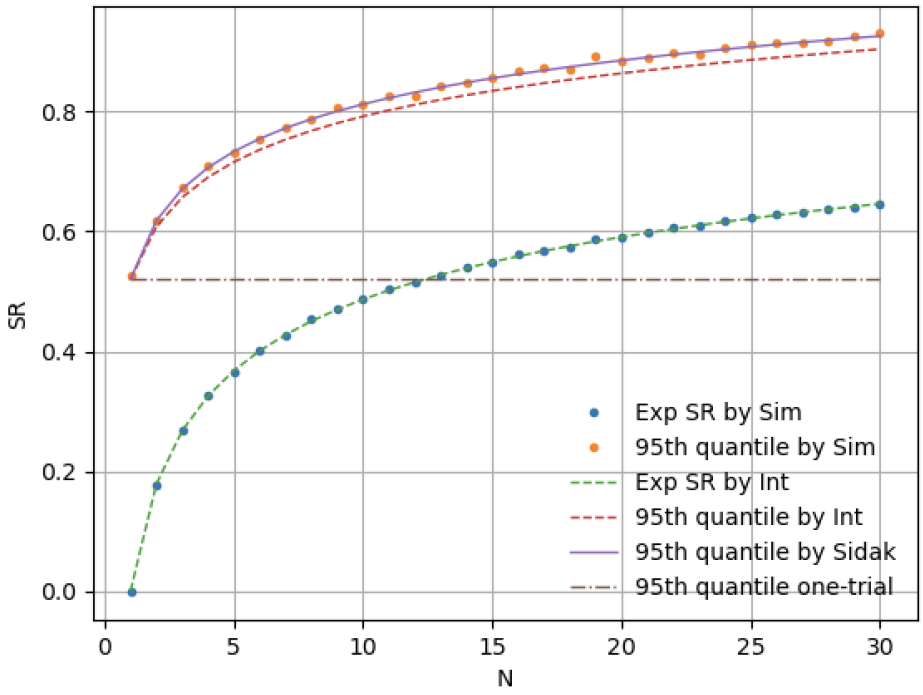
Applying Extreme Value Theory, Bailey et al. (2014) proved the *False Strategy Theorem*, namely that:

$$E[Y] \approx \eta + \sigma(1-\gamma)\mathrm{Z}^{-1}\left(1 - \frac{1}{K}\right) + \sigma\gamma\mathrm{Z}^{-1}\left(1 - \frac{1}{Ke}\right) \tag{20}$$

where K is the number of independent trials, and $\gamma$ is the Euler-Mascheroni constant, $\gamma = 0.577215665\ldots$ Bailey and Lopez de Prado (2021) confirmed the accuracy of this approximation through Monte Carlo experiments.

The PM tests the significance of SR = 1 against the hypothesis that the reported SR is the maximum SR of 10 zero-mean trials. The expected maximum SR is 0.69, and the p-value corresponding to SR = 1 is 0.88. The 95% confidence cut-off value is 1.12. The PM can therefore conclude that the reported SR is not significant, which contradicts the result of the initial standard SR significance test.

Exhibit 2: Expected SR and 95th Percentile Cutoff Values vs Number of Trials




To illustrate the effect of multiple trials on statistical testing and the performance of the different methodologies discussed above, we ran 10,000 simulations where, at each simulation step, we created 10 time series representing five years of daily returns (1,260 observations). All returns were drawn from a Normal distribution with a zero mean and an annualised volatility of 15%. The SR was calculated for each strategy, and the best one was selected. Exhibit 2 shows the expected SR values calculated by the simulation (Exp SR by Sim) and by numerical integration

(Exp SR by Int). We also show the 95th percentiles of SR calculated using several methods: simulation (95th quantile by Sim), numerical integration (95th quantile by Int), and Sidak method (95th quantile by Sidak). In the same graph, the horizontal line represents the value of the 95th quantile of the null distribution under a 1-trial assumption (a standard test).

Exhibit 2 illustrates how easy it is to inflate SR values by running multiple trials, even with time series with zero expected returns. Note that the expected maximum SR very quickly exceeds the 95th quantile obtained under the 1-trial assumption. As we can see, for practical purposes, all the applied methods give very similar results.

Note that the Sidak correction method and Extreme Value method used in this analysis are expected to work only if the trials are independent. In a real-life situation, a researcher typically tries different parameter configurations for similar trading ideas, leading to non-zero correlations between the trials. To address this concern, Lopez de Prado and Lewis (2019) and Lopez de Prado (2019b) show how to estimate the number $K$ of independent trials undertaken, using hierarchical clustering techniques.

Bailey and Lopez de Prado (2014) proposed a more accurate method that takes into account the moments of the SR distribution. The authors define the Deflated Sharpe Ratio (DSR) as a statistical distance between the reported SR value and the hypothetical $SR_0$,

$$DSR[SR^*] = Z\left[\frac{(\widehat{SR}-SR_0)\sqrt{T-1}}{\sqrt{1-\hat{\gamma}_3\widehat{SR}+\frac{\hat{\gamma}_4-1}{4}\widehat{SR}^2}}\right] \tag{21}$$

where:

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

Note that in equation (21), the denominator is the standard deviation of SR values, not the expected standard deviation of the maximum Sharpe ratio. DSR deflates the Sharpe ratio for the number of trials, the sample length, and the skewness and kurtosis of returns.

The methods presented thus far are among the FWER methodologies. FWER refers to the probability of making a Type I error among a specified group – or family – of tests. For this method type, even a single false discovery is unacceptable. This may seem too strict in some settings, e.g. in automobile production, where a certain proportion of defective cars is tolerated, as the False Discovery Rate (FDR) approach does (see Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001)). This level of tolerance may not be acceptable in investing, because the number of profitable strategies is very low, and a single false positive could impact the performance of trillions of dollars. This would be the case if, for instance, the value factor resulted to be a false

positive. Accordingly, in finance the FWER approach is arguably more relevant than the FDR approach, even if that comes at the expense of a higher number of false negatives.[3]

There is a second reason to prefer FWER over FDR. In our example the PM received only one strategy with the corresponding SR of the backtest. They did not have any information on the performance of other backtested trials. The problem of testing the "best" strategy out of $N$ backtested strategies is equivalent to the problem of testing the whole family while controlling for the FWER.

There are many challenges and limitations to the methodologies presented in this section. We have already mentioned that the presented methodologies might excessively increase Type 2 errors (rejection of true signals) while reducing Type 1 errors (acceptance of false signals). The optimal strategy should be carefully designed based on the user's requirements on a case-by-case basis. Lopez de Prado (2020, 2022), for instance, provides analytic estimates of Type 1 and Type 2 errors for the Sharpe ratios of investments and derives their familywise counterparts. Another challenge of the presented methods arises due to the non-stationarity of financial time series. Harris (2016) discusses this and other limitations of quantitative evaluations of trading strategies.

## CONCLUSION

The role of backtesting as a cornerstone technique in the development of systematic investment strategies cannot be overstated. This article has outlined three main types of backtests, namely the walk-forward, resampling, and Monte Carlo methods. Each offers unique advantages and distinct risks, which shape the techniques into specialised tools for crafting investment strategies.

We have recommended several best practices to significantly enhance the quality of the process. Noteworthy among these are the adoption of causal graphs and the use of rigorous statistical tests to validate that the Sharpe ratio significantly exceeds zero while controlling for SBuMT. Equally important are ensuring high-quality data, incorporating realistic costs, and accounting for liquidity constraints.

It should be noted that simulations leave much to be desired and that the risk of overfitting is a significant concern. This issue is exacerbated by the non-stationarity of financial markets, which undermines the reliability of past performance as a predictor of future results. Given these considerations, it is evident that while backtesting is an invaluable tool, it must be utilised with care and not as the primary driver of research but rather to validate a semi-final and well-formed investment strategy. In particular, it is highly advisable that researchers backtest only strategies that are supported by a sound causal theory (see Lopez de Prado 2023).

[3] See Harvey and Liu (2014) and Perumal and Flint (2018) for a literature overview and further analysis.

## REFERENCES


Ardia, David, and Kris Boudt. 2015. "Testing Equality of Modified Sharpe Ratios." *Finance Research Letters* 13: 97-104.

Bailey, David H., Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. 2014. "Pseudomathematics and Financial Charlatanism: The Effects of Backtest Over-Fitting on Out-of-Sample Performance." *Notices of the AMS* 61, no. 5: 458-471.

Bailey, David H., and Marcos Lopez de Prado. 2012. "The Sharpe Ratio Efficient Frontier." *Journal of Risk* 15, no. 2: 13.

Bailey, David H., and Marcos López de Prado. 2014. "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality." *Journal of Portfolio Management* 40, no. 5: 94-107.

Bailey, David H., and Marcos López de Prado. 2021. "The False Strategy Theorem: A Financial Application of Experimental Mathematics." *American Mathematical Monthly* 128, No. 9: 825-831

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57, no. 1: 289-300.

Benjamini, Yoav, and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing under Dependency." *Annals of Statistics*: 1165-1188.

Boehmer, Ekkehart, Charles M. Jones, and Xiaoyan Zhang. 2013. "Shackling Short Sellers: The 2008 Shorting Ban." *The Review of Financial Studies* 26, no. 6: 1363-1400.

Bonferroni, Carlo. 1936. "Teoria Statistica delle Classi e Calcolo delle Probabilita." *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8: 3-62.

Borkovec, Milan, and Hans G. Heidle. 2010. "Building and Evaluating a Transaction Cost Model: A Primer." *The Journal of Trading* 5, no. 2: 57-77.

Brown, Stephen J., William Goetzmann, Roger G. Ibbotson, and Stephen A. Ross. 1992. "Survivorship Bias in Performance Studies." *The Review of Financial Studies* 5, no. 4: 553-580.

Carhart, Mark M. 1997. "On Persistence in Mutual Fund Performance." *The Journal of Finance* 52, no. 1: 57-82.

Christie, Steve. 2005. "Is the Sharpe Ratio Useful in Asset Allocation?" *Macquarie Applied Finance Centre Research Paper*.

Elton, Edwin J., Martin J. Gruber, and Christopher R. Blake. 1996. "Survivor Bias and Mutual Fund Performance." *The Review of Financial Studies* 9, no. 4: 1097-1120.

Embrechts, Paul, Claudia Klüppelberg, and Thomas Mikosch. 2013. *Modelling Extremal Events: For Insurance and Finance*. Vol. 33. Springer Science & Business Media.

Fabozzi, Frank J., Francesco A. Fabozzi, Marcos López de Prado, and Stoyan V. Stoyanov. 2021. *Asset Management: Tools and Issues*.

Favre, Laurent, and José-Antonio Galeano. 2002. "Mean-Modified Value-at-Risk Optimization with Hedge Funds." *Journal of Alternative Investments* 5, no. 2: 21-25.

Gregoriou, Greg N., and Jean-Pierre Gueyie. 2003. "Risk-Adjusted Performance of Funds of Hedge Funds Using a Modified Sharpe Ratio." *The Journal of Wealth Management* 6, no. 3: 77-83.

Harvey, Campbell R., and Yan Liu. 2014. "Evaluating Trading Strategies." *The Journal of Portfolio Management* 40, no. 5: 108-118.

Harvey, Campbell R., and Yan Liu. 2015. "Backtesting." *Journal of Portfolio Management* 42, no. 1: 13.

Khan, Mostafa Saidur Rahim. 2024. "Short-Sale Constraints and Stock Returns: A Systematic Review." *Journal of Capital Markets Studies*. Vol. ahead-of-print No. ahead-of-print, pp. ahead-of-print.

Ledoit, Oliver, and Michael Wolf. 2008. "Robust Performance Hypothesis Testing with the Sharpe Ratio." *Journal of Empirical Finance* 15, no. 5: 850-859.

Li, Junyi, Xintong Wang, Yaoyang Lin, Arunesh Sinha, and Michael Wellman. 2020. "Generating Realistic Stock Market Order Streams." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 727-734.

Liu, Ying, Marie Rekkas, and Augustine Wong. 2012. "Inference for the Sharpe Ratio Using a Likelihood-Based Approach." *Journal of Probability and Statistics 2012 (Article ID 878561): 1-24*.

Lo, Andrew W. 2002. "The Statistics of Sharpe Ratios." *Financial Analysts Journal* 58, no. 4: 36-52.

Lopez de Prado, Marcos. 2018. *Advances in Financial Machine Learning*. John Wiley & Sons.

Lopez de Prado, Marcos. 2019a. "Tactical Investment Algorithms." *Available at SSRN 3459866*.

Lopez de Prado, Marcos. 2019b. "A Data Science Solution to the Multiple-Testing Crisis in Financial Research." *Journal of Financial Data Science* 1, No. 1: 99-110.

Lopez de Prado, Marcos. 2020. Machine Learning for Asset Managers. Cambridge University Press.

Lopez de Prado, Marcos. 2022. "Type I and Type II Errors of the Sharpe Ratio under Multiple Testing." *Journal of Portfolio Management*, Vol. 49, No. 1, pp. 39-46

Lopez de Prado, Marcos. 2023. Causal Factor Investing. Cambridge University Press.

Lopez de Prado, Marcos, and Michael J. Lewis. "Detection of false investment strategies using unsupervised learning methods." Quantitative Finance 19, no. 9 (2019): 1555-1565.

Lopez de Prado, Marcos, and Vincent Zoonekynd. 2024. "Why Has Factor Investing Failed?: The Role of Specification Errors." *Available at SSRN 4697929*

Malkiel, Burton G. 1995. "Returns from Investing in Equity Mutual Funds 1971 to 1991." *The Journal of Finance* 50, no. 2: 549-572.

Mertens, Elmar. 2002. "Comments on Variance of the IID Estimator in Lo." Research Note (www.elmarmertens.org).

Harris, Michael. 2016. "Limitations of Quantitative Claims About Trading Strategy Evaluation." SSRN. July 15. https://ssrn.com/abstract=2810170

Musciotto, Federico, Luca Marotta, Salvatore Miccichè, and Rosario N. Mantegna. 2018. "Bootstrap Validation of Links of a Minimum Spanning Tree." *Physica A: Statistical Mechanics and its Applications* 512: 1032-1043.

Perumal, Kovlin, and Emlyn Flint. 2018. "Systematic Testing of Systematic Trading Strategies." *Available at SSRN 3132229*.

Qi, J., M. Rekkas, and A. Wong. 2018. "Highly Accurate Inference on the Sharpe Ratio for Autocorrelated Return Data." *Journal of Statistical and Econometric Methods* 7, no. 1: 21-50.

Saffi, Pedro AC, and Kari Sigurdsson. 2011. "Price Efficiency and Short Selling." *The Review of Financial Studies* 24, no. 3: 821-852.

Sharpe, William F. 1966. "Mutual Fund Performance." *The Journal of Business* 39, no. 1: 119-138.

Šidák, Zbyněk. 1967. "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions." *Journal of the American Statistical Association* 62, no. 318: 626-633.

Sullivan, Ryan, Allan Timmermann, and Halbert White. 1999. "Data-Snooping, Technical Trading Rule Performance, and the Bootstrap." *The Journal of Finance* 54, no. 5: 1647-1691.

White, Halbert. 2000. "A Reality Check for Data Snooping." *Econometrica* 68, no. 5: 1097-1126.

Wiese, Magnus, Robert Knobloch, Ralf Korn, and Peter Kretschmer. 2020. "Quant GANs: Deep Generation of Financial Time Series." *Quantitative Finance* 20, no. 9: 1419-1440.