# Type I and Type II Errors of the Sharpe Ratio under Multiple Testing

Marcos López de Prado

**Marcos López de Prado** is global head of quantitative research and development at Abu Dhabi Investment Authority in Abu Dhabi, United Arab Emirates, and a professor of practice at Cornell University in Ithaca, NY.
**ml863@cornell.edu**

## KEY FINDINGS

- Academic articles often report the Sharpe ratio associated with an investment strategy.
- Multiple testing impacts Type I and Type II errors associated with those Sharpe ratio estimates.
- This article offers analytic estimates of Type I and Type II errors for the Sharpe ratio, adjusted for multiple testing.

## ABSTRACT

Articles in financial literature typically estimate the *p*-value associated with an investment strategy's performance without reporting the power of the test used to make that discovery. In this article, the author provides analytic estimates to Type I and Type II errors for the Sharpe ratios of investments and derives their familywise counterparts. These estimates allow researchers to carefully design experiments and select investments with high confidence and power.

Financial researchers conduct thousands (if not millions) of backtests before identifying an investment strategy. Hedge funds interview hundreds of portfolio managers before filling a position. Asset allocators peruse thousands of asset managers before building a template portfolio with those candidates who exceed some statistical criteria (López de Prado 2018a). What all these examples have in common is that statistical tests are applied multiple times. When the rejection threshold is not adjusted for the number of trials (the number of times the test has been administered), false positives (Type I errors) occur with a probability higher than the test's false positive rate (see Bailey et al. 2014, Bailey and López de Prado 2014).

Empirical studies in economics and finance almost always fail to report the power of the test used to make a particular discovery. Without that information, readers cannot assess the rate at which false negatives occur (Type II errors). To illustrate this point, consider the situation in which a senior researcher at the Federal Reserve Board of Governors is tasked with testing the hypothesis that stock prices are in a bubble. The standard approach would be to apply a high significance level because only a high confidence level would justify draconian monetary policy actions. For example, at the standard 95% confidence level and assuming the null hypothesis that there is no bubble, a researcher expects that the test will falsely reject the null with a 5% probability. This modus operandi has two caveats: First, by reducing the false positives to such a low level, the test may miss a large portion of the bubbles. And yet, from the Fed's perspective, missing half of the bubbles is likely much worse

than taking a 5% risk of triggering a false alarm. Second, researchers may be getting away with this first caveat by hiding or not reporting the power of the test.

In contrast, hedge fund managers are often more concerned with false positives than with false negatives. Client redemptions are more likely to be caused by the former (an actual loss) than the latter (a missed opportunity). At the same time, it is unclear why investors and hedge funds would apply arbitrary significance levels, such as 10%, 5%, or 1%. Rather, an objective significance level could be set such that Type I and Type II errors are jointly minimized. In other words, even researchers who do not particularly care for Type II errors could compute them as a way to introduce objectivity to an otherwise subjective choice of significance level.

The purpose of this article is threefold: First, I provide an analytic estimate to the probability of selecting a false investment strategy, corrected for multiple testing. Second, I provide an analytic estimate to the probability of missing a true investment strategy, corrected for multiple testing. Third, I model the interaction between Type I and Type II errors under multiple testing.

## REVIEW OF THE RELEVANT LITERATURE

In general terms, the statistics literature on multiple testing controls for false positives in two different ways: First, the familywise error rate (FWER) is defined as the probability that *at least one* false positive takes place. FWER-based tests are designed to control for a single false positive (Holm 1979). Second, the false discovery rate (FDR) is defined as the expected value of the ratio of false positives to predicted positives. FDR-based tests are designed to generate Type I errors at a constant rate, proportional to the number of predicted positives (see Benjamini and Hochberg 1995; Benjamini and Liu 1999; and Benjamini and Yekutieli 2001). In most scientific and industrial applications, FWER is considered overly punitive, and authors prefer to use FDR. For example, it would be impractical to design a car model in which we control for the probability that a single unit will be defective. However, in the context of finance, the use of FDR is often inappropriate. The reason is that an investor does not typically allocate funds to all strategies with predicted positives within a family of trials in which a proportion of them are likely to be false. Instead, investors are only introduced to the single best strategy out of a family of millions of alternatives, due to selection bias and publication bias. Following the car analogy, in finance, a single car unit is produced per model, which everyone will use. If the only produced unit is defective, everyone will crash. For example, investors are not exposed to the dozens of alternative model specifications tried by Fama and French in their five-factor article (2015). They have only been told about the one specification that Fama and French found to be best, and they have no ability to invest in their alternative models that passed individual statistical significance tests. In the context of selection and publication bias, in which the authors or editors of a publication conceal part of the information concerning other trials conducted, FWER is more appropriate than FDR. Accordingly, the procedure explained in this article applies a FWER definition of the Type I error.

Under the assumption that returns follow an i.i.d. normal distribution, Harvey and Liu (2018) transform the Sharpe ratio estimated on $T$ observations into a $t$-ratio, which follows a $t$-distribution with $T - 1$ degrees of freedom. These authors then apply Šidák's correction (1967) to estimate the probability of observing a maximal $t$-value that exceeds a given threshold. Accordingly, their estimation of the FWER relies on the validity of two assumptions: (1) that returns are normally distributed, and (2) that trials are independent. These authors later extended their method to allow for a constant average correlation between trials.

Building on theorems from extreme value theory, Bailey and López de Prado (2014) estimated the FWER of a test of hypothesis on the Sharpe ratio while controlling for

non-normal returns, sample length, and multiple testing. Rather than assuming that returns follow an i.i.d. normal process, that framework only requires that returns are generated by stationary and ergodic processes. López de Prado (2018b) and López de Prado and Lewis (2018) introduced an unsupervised learning method to model complex hierarchical structures extracted from the trials' correlations. In doing so, that approach does not assume a constant average correlation between trials.

Harvey and Liu (2020) proposed a numerical (Monte Carlo) procedure to derive Type I and Type II errors in hypothesis testing of investment strategies. Notably, they also introduced the notion of a ratio of misses to false discoveries.

## CONTRIBUTIONS OF THIS ARTICLE

In this article, I extend the analysis of Harvey and Liu (2018) in two ways: First, I do not assume that returns follow a normal distribution. Numerous empirical studies show that investment strategies' returns exhibit negative skewness and positive excess kurtosis (among others, see Brooks and Kat 2002 and Ingersoll et al. 2007). The implication is that assuming normal returns is often unrealistic and may lead to an underestimation of Type I error probabilities. This article's results are derived under the more general assumption that returns are stationary and ergodic. Second, I do not assume a constant average correlation between trials. Researchers commonly attempt multiple implementations of several strategies, in which a subset of trials associated with a particular strategy are more correlated with each other than with the rest. The ensuing correlation patterns are better modeled as a hierarchical structure than as a constant coefficient across all pairs. Following López de Prado and Lewis (2018), it is preferable to apply an unsupervised method that learns that hierarchical structure of the trials in order to derive the number of effectively independent clusters of trials.

This article also extends Harvey and Liu (2018) by providing estimates of Type I and Type II errors that are analytic (in closed form) rather than through numerical (double bootstrap) methods. This extension is not a criticism of those authors' numerical method, which is useful and insightful in many ways (and an inspiration for this article's analytic method). I provide a closed-form solution because this enables researchers to achieve their ultimate goal, which is to optimize the performance of the statistical test.

In this article, familywise estimates of significance and power are computed under a frequentist approach rather than under a Bayesian framework. Although I appreciate the merits of the latter, the practical consideration that the frequentist approach remains the most popularly applied, in finance and elsewhere, informed this choice. The connection between the two in the context of hypothesis testing is clearly exemplified by Lindley's paradox (Shafer 1982; Robert 2014).

## FAMILYWISE ERROR RATE

Under the standard Neyman–Pearson hypothesis-testing framework, we reject a null hypothesis $H_0$ with confidence $(1 - \alpha)$ when we observe an event that, should the null hypothesis be true, could only occur with probability $\alpha$. Then, the probability of falsely rejecting the null hypothesis (Type I error) is $\alpha$. This is also known as the probability of a false positive.

When Neyman and Pearson (1933) proposed this framework, they did not consider the possibility of conducting multiple tests and selecting the best outcome. When a test is repeated multiple times, the probability that one of the positives is false is greater than $\alpha$. After a *family* of $K$ independent tests, we would reject $H_0$ with confidence $(1 - \alpha)^K$, hence the family false positive probability (or FWER) is

$\alpha_K = 1 - (1 - \alpha)^K$. This is the probability that *at least one* of the positives is false, which is the complementary to the probability that none of the positives are false, $(1 - \alpha)^K$.

## ŠIDÁK'S CORRECTION

Suppose that we set a FWER over $K$ independent tests at $\alpha_K$. Then, the individual false positive probability can be derived from the earlier equation as $\alpha = 1 - (1 - \alpha_K)^{1/K}$. This is known as the Šidák correction for multiple testing (Šidák 1967), and it can be approximated as the first term of a Taylor expansion, $\alpha \approx \frac{\alpha_K}{K}$ (known as Bonferroni's approximation).

Following López de Prado and Lewis (2018), given the time series of returns from multiple trials, we can estimate E[$K$] as the number of clusters present in the correlation matrix. While it is true that the E[$K$] clusters are not perfectly uncorrelated, they provide a conservative estimate of the minimum number of clusters the algorithm could not reduce further. Alternatively, a researcher could estimate E[$K$] as the effective number implied by the distribution of eigenvalues of the correlation matrix. With this estimate E[$K$], we can apply Šidák's correction and compute the Type I error probability under multiple testing, $\alpha_K$.

## TYPE I ERRORS UNDER MULTIPLE TESTING

Consider an investment strategy with returns time series of size $T$. We estimate the Sharpe ratio, $\widehat{SR}$, and subject that estimate to a hypothesis test, where $H_0$: $SR = 0$ and $H_1$: $SR > 0$. We wish to determine the probability of a false positive when this test is applied multiple times.

Bailey and López de Prado (2012) derived the probability that the true Sharpe ratio exceeds a given threshold $SR^*$, under the general assumption that returns are stationary and ergodic (not necessarily i.i.d. normal). If the true Sharpe ratio equals $SR^*$, the statistic $\hat{z}[SR^*]$ is asymptotically distributed as a standard normal:

$$\hat{z}[SR^*] = \frac{(\widehat{SR} - SR^*)\sqrt{T-1}}{\sqrt{1 - \hat{\gamma}_3 \widehat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\widehat{SR}^2}} \xrightarrow{a} Z \tag{1}$$

where $\widehat{SR}$ is the estimated Sharpe ratio (nonannualized), $T$ is the number of observations, $\hat{\gamma}_3$ is the estimated skewness of the returns, and $\hat{\gamma}_4$ is the estimated kurtosis of the returns. Familywise Type I errors occur with probability:

$$P\left[\max_k \{\hat{z}[0]_k\}_{k=1,\ldots,K} > z_\alpha \middle| H_0\right] = 1 - (1 - \alpha)^K = \alpha_K \tag{2}$$

For a FWER $\alpha_K$, Šidák's correction gives us a single-trial significance level $\alpha = 1 - (1 - \alpha_K)^{1/K}$. Then, the null hypothesis is rejected with confidence $(1 - \alpha_K)$ if $\max_k \{\hat{z}[0]_k\}_{k=1,\ldots,K} > z_\alpha$, where $z_\alpha$ is the critical value of the standard normal distribution that leaves a probability $\alpha$ to the right, $z_\alpha = Z^{-1}[1 - \alpha] = Z^{-1}[(1 - \alpha_K)^{1/K}]$, and $Z[.]$ is the cumulative distribution function (CDF) of the standard normal distribution.

Conversely, we can derive the Type I error under multiple testing ($\alpha_K$) as follows: First, apply the clustering procedure on the trials' correlation matrix to estimate clusters' returns series and E[$K$]. Second, estimate $\hat{z}[0] = \max_k \{\hat{z}[0]_k\}_{k=1,\ldots,K}$ on the selected cluster's returns. Third, compute the Type I error for a single test, $\alpha = 1 - Z[\hat{z}[0]]$. Fourth, correct for multiple testing, $\alpha_K = 1 - (1 - \alpha)^K$, resulting in

## EXHIBIT 1

### Type 1 Error, with Numerical Example

```
import scipy.stats as ss
#---------------------------------------------------------------
def getZStat(sr,t,sr_=0,skew=0,kurt=3):
    z=(sr-sr_)*(t-1)**.5
    z/=(1-skew*sr+(kurt-1)/4.*sr**2)**.5
    return z
#---------------------------------------------------------------
def type1Err(z,k=1):
    # false positive rate
    alpha=ss.norm.cdf(-z)
    alpha_k=1-(1-alpha)**k # multi-testing correction
    return alpha_k
#---------------------------------------------------------------
def main0():
    # Numerical example
    t,skew,kurt,k,freq=1250,-3,10,10,250
    sr=1.25/freq**.5;sr_=1./freq**.5
    z=getZStat(sr,t,0,skew,kurt)
    alpha_k=type1Err(z,k=k)
    print(alpha_k)
    return
#---------------------------------------------------------------
if __name__=='__main__':main0()
```

$$\alpha_K = 1 - Z[\hat{z}[0]]^{E[K]} \tag{3}$$

Let us illustrate the aforementioned calculations with a numerical example. Suppose that after conducting 1,000 trials, we identify an investment strategy with a Sharpe ratio of 0.0791 (nonannualized), a skewness of −3, a kurtosis of 10, computed on 1,250 daily observations (five years, at 250 annual observations). These levels of skewness and kurtosis are typical of hedge fund returns sampled with daily frequency. From these inputs, we derive $\hat{z}[0] \approx 2.4978$ and $\alpha \approx 0.0062$. At this Type I error probability, most researchers would reject the null hypothesis and declare that a new investment strategy has been found. However, this $\alpha$ is not adjusted for the E[$K$] trials it took to find this strategy. We apply the clustering algorithm on the correlation matrix of the trials' returns and conclude that out of the 1,000 (correlated) trials, there are E[$K$] = 10 effectively independent trials (again, with "effectively" independent, we do not assert that the 10 clusters are strictly independent, but that the algorithm could not find more uncorrelated groupings). Then, the corrected FWER is $\alpha_K \approx 0.0608$. Although the annualized Sharpe ratio is approximately 1.25, the probability that this strategy is a false positive is relatively high for two reasons: (1) The number of trials, since $\alpha_K = \alpha \approx 0.0062$ if E[$K$] = 1, and (2) the non-normality of the returns, since $\alpha_K = 0.0261$ should returns have been normal. As expected, wrongly assuming normal returns leads to a gross underestimation of the Type I error probability. Exhibit 1 provides the python code that replicates these results.

## TYPE II ERRORS UNDER MULTIPLE TESTING

Suppose that the alternative hypothesis ($H_1$: $SR > 0$) for the best strategy is true, and $SR = SR^*$. Then, the power of the test associated with a FWER $\alpha_K$ is

$$\begin{aligned}
P[\max_k\{\hat{z}[0]_k\}_{k=1,\ldots,K} > z_\alpha | SR = SR^*] &= P\left[\frac{(\widehat{SR} + SR^* - SR^*)\sqrt{T-1}}{\sqrt{1-\hat{\gamma}_3\widehat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\widehat{SR}^2}} > z_\alpha \middle| SR = SR^*\right] \\
&= P\left[\hat{z}[SR^*] > z_\alpha - \frac{SR^*\sqrt{T-1}}{\sqrt{1-\hat{\gamma}_3\widehat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\widehat{SR}^2}} \middle| SR = SR^*\right] \\
&= 1 - P\left[\hat{z}[SR^*] < z_\alpha - \frac{SR^*\sqrt{T-1}}{\sqrt{1-\hat{\gamma}_3\widehat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\widehat{SR}^2}} \middle| SR = SR^*\right] \\
&= 1 - Z\left[z_\alpha - \frac{SR^*\sqrt{T-1}}{\sqrt{1-\hat{\gamma}_3\widehat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\widehat{SR}^2}}\right] = 1 - \beta
\end{aligned} \tag{4}$$

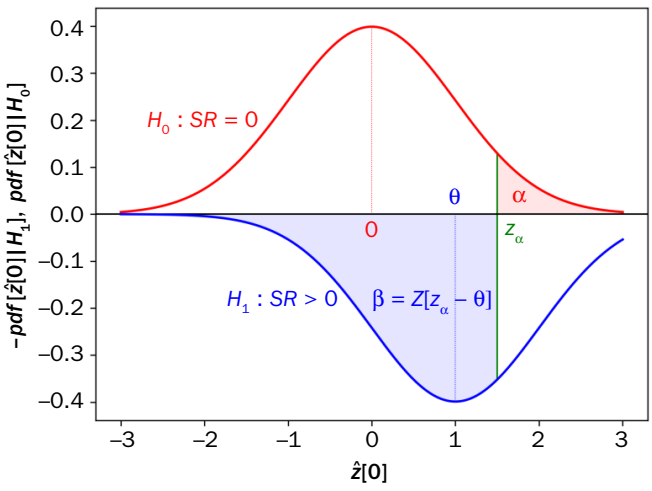
## EXHIBIT 2

### Type II Error, with Numerical Example

```
def getTheta(sr,t,sr_=0,skew=0,kurt=3):
    theta=sr_*(t-1)**.5
    theta/=(1-skew*sr+(kurt-1)/4.*sr**2)**.5
    return theta
#-------------------------------------------------------------------------
deftype2Err(alpha_k,k,theta):
    # false negative rate
    z=ss.norm.ppf((1-alpha_k)**(1./k)) # Sidak's correction
    beta=ss.norm.cdf(z-theta)
    return beta
#-------------------------------------------------------------------------
def main0():
    # Numerical example
    t,skew,kurt,k,freq=1250,-3,10,10,250
    sr=1.25/freq**.5;sr_=1./freq**.5
    z=getZStat(sr,t,0,skew,kurt)
    alpha_k=type1Err(z,k=k)
    theta=getTheta(sr,t,sr_,skew,kurt)
    beta=type2Err(alpha_k,k,theta)
    beta_k=beta**k
    print(beta_k)
    return
#-------------------------------------------------------------------------
if __name__=='__main__':main0()
```

## EXHIBIT 3

### The Interaction between α and β




where $z_\alpha = Z^{-1}[(1 - \alpha_K)^{1/K}]$. Accordingly, the *individual* power of the test increases with *SR**, the sample length, and the skewness; however, it decreases with the kurtosis. This probability $(1 - \beta)$ is alternatively known as the true positive rate, power, or recall.

We define the familywise false negative (miss) probability as the probability that *all* individual positives are missed, $\beta_K = \beta^K$. For a given pair $(\alpha_K, \beta_K)$, we can derive the pair $(\alpha, \beta)$ and imply the value *SR** such that $P[\max_k\{\hat{z}[0]_k\}_{k=1,\ldots,K} > z_\alpha | SR = SR^*] = 1 - \beta$. The interpretation is that, at a FWER $\alpha_K$, achieving a familywise power above $(1 - \beta_K)$ requires that the true Sharpe ratio exceed *SR**. In other words, the test is not powerful enough to detect true strategies with a Sharpe ratio below that implied *SR**.

We can derive the Type II error under multiple testing $(\beta_K)$ as follows: First, given a FWER $\alpha_K$, which is either set exogenously or estimated as explained in the previous section, compute the single-test critical value, $z_\alpha$. Second, the probability of missing a strategy with the Sharpe ratio *SR** is $\beta = Z[z_\alpha - \theta]$, where $\theta = \frac{SR^*\sqrt{T-1}}{\sqrt{1 - \hat{\gamma}_3\widehat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\widehat{SR}^2}}$. Third, from the individual false negative probability, we derive $\beta_K = \beta^K$ as the probability that all positives are missed.

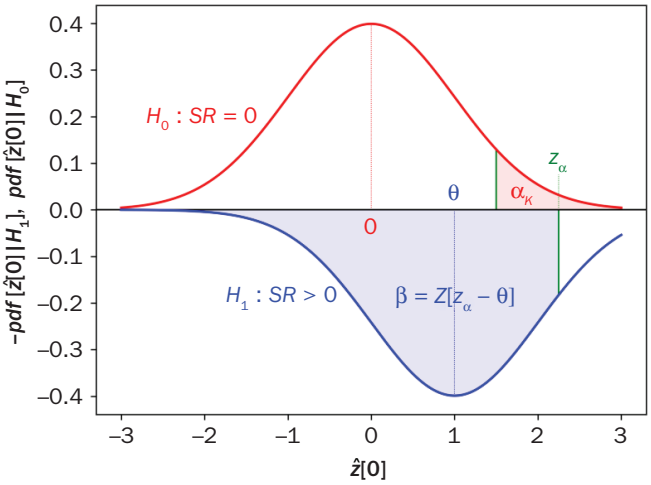
Let us apply the aforementioned equations to the numerical example in the previous section. There, we estimated that the FWER was $\alpha_K \approx 0.0608$, which implies a critical value $z_\alpha \approx 2.4978$. Then, the probability of missing a strategy with a true Sharpe ratio *SR** ≈ 0.0632 (nonannualized) is $\beta \approx 0.6913$, where $\theta \approx 1.9982$. This high individual Type II error probability is understandable because the test is not powerful enough to detect such a weak signal (an annualized Sharpe ratio of only 1.0) after a single trial. But because we have conducted 10 trials, $\beta_K \approx 0.0249$. The test detects more than 97.5% of the strategies with a true Sharpe ratio *SR** ≥ 0.0632. Exhibit 2 provides the python code that replicates these results (see Exhibit 1 for functions getZStat and type1Err).

## THE INTERACTION BETWEEN TYPE I AND TYPE II ERRORS

Exhibit 3 illustrates the relation between α and β. The red distribution models the probability of $\widehat{SR}$ estimates under the assumption that $H_0$ is true. The blue distribution (plotted upside down to facilitate display) models the probability of $\widehat{SR}$ estimates under the assumption that $H_1$ is true and, in particular, under the scenario where *SR** = 1. The sample

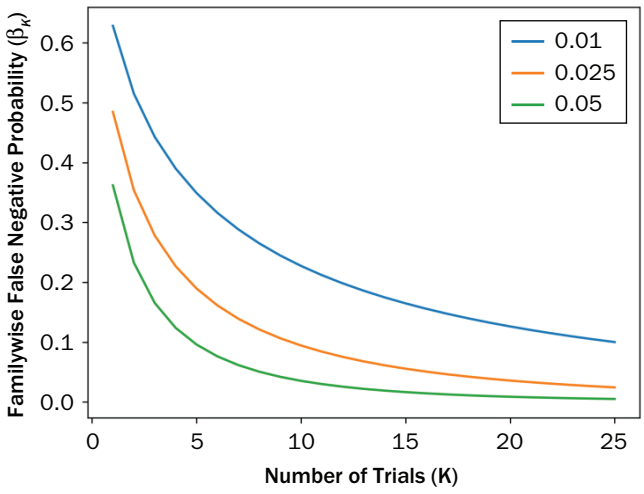
## EXHIBIT 4

**The Interaction between $\alpha_K$ and $\beta$**




## EXHIBIT 5

**$\beta_K$ as $K$ Increases for $\theta \approx 1.9982$ and $\alpha_K \in \{0.01, 0.025, 0.05\}$**




length, skewness, and kurtosis influence the variance of these two distributions. Given an actual estimate $\widehat{SR}$, those variables determine the probabilities $\alpha$ and $\beta$, in which decreasing one implies increasing the other. In most journal articles, authors focus on the red distribution and ignore the blue distribution.

The analytic solution we derived for Type II errors makes it obvious that this tradeoff also exists between $\alpha_K$ and $\beta_K$, although in a not so straightforward manner as in the $K = 1$ case. Exhibit 4 shows that for a fixed $\alpha_K$, as $K$ increases, $\alpha$ decreases, and $z_\alpha$ increases, hence $\beta$ increases.

Exhibit 5 plots $\beta_K$ as $K$ increases for various levels of $\alpha_K$. Although $\beta$ increases with $K$, the overall effect of multiple testing is a decrease of $\beta_K$. For a fixed $\alpha_K$, the equation that determines $\beta_K$ as a function of $K$ and $\theta$ is

$$\beta_K = (Z[Z^{-1}[(1-\alpha_K)^{1/K}] - \theta])^K \tag{5}$$

One interpretation of the previous equation is that multiple testing has the benefit of increasing the power of a test as long as the researcher controls for FWER.

## CONCLUSIONS

The estimated Sharpe ratio of an investment strategy under a single trial follows a Gaussian distribution, even if the strategy returns are non-normal. Researchers typically conduct a multiplicity of trials, and selecting the best-performing strategy increases the probability of selecting a false strategy. This phenomenon is called selection bias under multiple testing (SBuMT), and in this article, I have studied one procedure to evaluate the extent to which SBuMT invalidates a discovered investment strategy.

The procedure relies on Šidák's correction to derive the FWER. The FWER provides an adjusted rejection threshold on which we can test whether $\max_k\{\widehat{SR}_k\}$ is statistically significant. Researchers can use these analytical estimates of the familywise false positive probability and familywise false negative probability when they assess the significance of an investment among a multiplicity of alternatives.

## REFERENCES

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* 61 (5): 458–471.

Bailey, D., and M. López de Prado. 2012. "The Sharpe Ratio Efficient Frontier." *The Journal of Risk* 15 (2): 3–44.

——. 2014. "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-normality." *The Journal of Portfolio Management* 40 (5): 94–107.

Benjamini, Y., and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society* 57: 289–300.

Benjamini, Y., and W. Liu. 1999. "A Step-Down Multiple Hypotheses Testing Procedure that Controls the False Discovery Rate under Independence." *Journal of Statistical Planning and Inference* 82: 163–170.

Benjamini, Y., and D. Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing under Dependency." *Annals of Statistics* 29: 1165–1188.

Brooks, C., and H. Kat. 2002. "The Statistical Properties of Hedge Fund Index Returns and Their Implications for Investors." *The Journal of Alternative Investments* 5 (2): 26–44.

Fama, E., and K. French. 2015. "A Five-Factor Asset Pricing Model." *Journal of Financial Economics* 116 (1): 1–22.

Harvey, C., and Y. Liu. 2018. "Backtesting." *The Journal of Portfolio Management* 42 (1): 13–28.

——. "False (and Missed) Discoveries in Financial Economics." Working paper, Duke University and Purdue University, 2020.

Holm, S. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6: 65–70.

Ingersoll, J., M. Spiegel, W. Goetzmann, and I. Welch. 2007. "Portfolio Performance Manipulation and Manipulation-Proof Performance Measures." *The Review of Financial Studies* 20 (5): 1504–1546.

López de Prado, M. *Advances in Financial Machine Learning*, 1st ed. Hoboken, NJ: John Wiley & Sons, 2018a.

——. 2018b. "A Data Science Solution to the Multiple-Testing Crisis in Financial Research." *The Journal of Financial Data Science* 1 (1): 99–110.

López de Prado, M., and M. Lewis. 2018. "Detection of False Investment Strategies Using Unsupervised Learning Methods." *Quantitative Finance* 19 (9): 1555–1565.

Neyman, J., and E. Pearson. 1933. "IX. On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society* 231 (694–706): 289–337.

Robert, C. 2014. "On the Jeffreys-Lindley Paradox." *Philosophy of Science* 81 (2): 216–232.

Shafer, G. 1982. "Lindley's Paradox." *Journal of the American Statistical Association* 77 (378): 325–334.

Šidák, Z. 1967. "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions." *Journal of the American Statistical Association* 62 (318): 626–633.