The False Strategy Theorem: A Financial Application of Experimental Mathematics

Marcos López de Prado and David H. Bailey

Abstract. The late Jonathan M. Borwein excelled at a wide range of mathematical fields. He is perhaps best known for his work in experimental mathematics and optimization. But his interests extended far beyond these two arenas, to often unexpected topics. Unbeknownst to many, he also made important contributions to mathematical finance, and additionally published studies concerned with the reproducibility of scientific discoveries in numerous different fields. In this paper we have attempted to merge some of these seemingly unrelated topics, elucidating a common thread connecting them all.

1. THE REPRODUCIBILITY CRISIS IN FINANCE Financial mathematics is one of the many research fields where Professor Jonathan M. Borwein made substantial contributions. In a series of papers, Borwein and his co-authors (including the two authors of this article) worked on providing practical solutions to various outstanding problems in the field.

Borwein's interest in finance emanated from his wider preoccupation for reproducibility of scientific discoveries. Academic finance journals routinely publish discoveries based on the historical analysis of time series (known as "backtests"). These time series are finite and available to researchers, who can test various theories looking back decades of history, both in the U.S. and international financial markets.

The reproducibility crisis in finance is primarily the consequence of three factors. First, once a theory is published, it may take many decades to collect the future ("out-of-sample") information needed to evaluate the accuracy of the proposed theory. It takes one day of markets activity to produce one day's worth of data. Second, finance is not a static system. Even if we collect enough information to debunk a theory, it is possible that the theory was correct at the time of publication, but that changes in the system since publication render it no longer valid. Or perhaps the theory was false to begin with — we may never know. Third, we cannot repeat an experiment over and over again while controlling for specific environmental variables, so as to identify a precise cause-effect mechanism. All we have is a single historical path.

In other fields, where out-of-sample time series are readily available for reproducibility testing, debunking false discoveries is relatively straightforward. For example, if someone claims to have found a new particle in high-energy physics, other laboratories will be able to confirm or refute that discovery independently, based on new and distinct evidence quickly gathered after the initial claim is published.

2. SELECTION BIAS UNDER MULTIPLE TESTING Finance is not alone in this challenge. In medical research, effects of treatments must be observed for many years, making it difficult to determine whether a discovery is true or false. If a treatment is dispensed to a thousand individuals, it is possible that a few patients improve their condition out of luck. It would be dishonest (moreover, illegal!) for a laboratory to report only those outcomes that confirm a desired result. In statistics, this kind of scientific fraud (intentional or unintentional) is known as *selection bias under multiple testing*. In order to control for this effect, the Food and Drug Administration and similar government bodies in other nations, as well as most leading journals in the biomedical-pharmaceutical field, require researchers to report the results from all trials, so that the probability of luck can be discounted from the reported results.

The standard hypothesis testing framework, first published in Neyman and Pearson [**9**], was designed for single-trial tests. To correct for selection bias under multiple testing, statisticians have developed two main approaches, known as *familywise error rate* and *false discovery rate*. Succinctly, the familywise error rate computes the probability of a single false positive, while false discovery rate estimates the portion of false test results among predicted positives.

In contrast, widely used econometrics textbooks, such as Greene [**7**] or Hamilton [**8**], tend to ignore or downplay the challenges posed by selection bias under multiple testing. What's more, even at the present date most academic finance journals do not require authors to declare the number of trials involved in a discovery, even though the authors may well have performed an extensive computer search for optimal parameters, effectively iterating over millions of possibilities. Journals operate under the demonstrably false assumption that researchers have carried out a single test. The sobering implication of this fact is that many discoveries in finance may be false, and it may be very difficult to determine which are true.

3. PERFORMANCE EVALUATION With this background, Borwein and his co-authors (including the two of us) attempted to provide a practical solution to the reproducibility crisis in finance. Investments are typically evaluated according to their risk-adjusted excess returns. Excess returns are defined as the returns of an investment in excess of a benchmark. The most popular statistic for investment performance is the *Sharpe ratio*, which is essentially the ratio between expected excess returns and the variance of the excess returns (see [**11, 12, 13**]).

More formally, consider an investment strategy with excess returns (or "risk premia") $\{r_t\},\ t = 1, \cdots, T$, which follow an independent identically distributed normal distribution, $r_t \sim \mathcal{N}[\mu, \sigma^2]$, where $\mathcal{N}[\mu, \gamma^2]$ represents a Gaussian normal distribution with mean μ and variance σ^2. The Sharpe ratio of such a strategy is defined as $SR = \mu/\sigma$. Because the parameters μ and σ are seldom known *a priori*, the Sharpe ratio is typically estimated as $\widehat{SR} = \widehat{E}[\{r_t\}] / \sqrt{\widehat{V}[\{r_t\}]}$, where $\widehat{E}[\cdot]$ is the empirical expected value and $\widehat{V}[\cdot]$ is the empirical variance. The asymptotic distribution of the Sharpe ratio is known to be Gaussian, even if the excess returns are not Gaussian. In fact, the asymptotic distribution of the Sharpe ratio is Gaussian under the rather general assumption that excess returns are stationary and ergodic (see [**1**]).

In the context of multiple testing, a researcher may carry out a large number of historical simulations (trials), and report only the best outcome (i.e., the maximum Sharpe ratio). But the distribution of the maximum Sharpe ratio is *not* the same as the distribution of a Sharpe ratio randomly chosen among the trials, hence giving rise to selection bias under multiple testing. When more than one trial takes place, the expected value of the maximum Sharpe ratio is greater than the expected value of the Sharpe ratio from a random trial. In particular, given an investment strategy with expected Sharpe ratio zero and non-null variance, the expected value of the maximum Sharpe ratio is strictly positive, and a function of the number of trials.

Given the above, the magnitude of selection bias under multiple testing can be expressed in terms of the difference between the expected maximum Sharpe ratio and the expected Sharpe ratio from a random trial (zero, in the case of a false strategy). This observation is the basis for the practical solution proposed by Borwein and his co-authors in [**2**]: The expected maximum Sharpe ratio is the hurdle or threshold that the reported Sharpe ratio must exceed. The following theorem formalizes this notion.

4. THE FALSE STRATEGY THEOREM

In the following, $\mathcal{N}$ denotes the Gaussian normal distribution, $Z^{-1}[\cdot]$ denotes the inverse of the standard Gaussian cumulative distribution function (CDF), $E[\cdot]$ denotes expected value, $V[\cdot]$ denotes variance, and γ is the Euler-Mascheroni constant $= 0.5772156649\ldots$.

Theorem 1 (False strategy theorem). *Given a sample of estimated performance statistics* $\left\{\widehat{SR}_k\right\}$, $k = 1, \cdots, K$, *with independent and identically distributed Gaussian distribution, i.e.,* $\left\{\widehat{SR}_k\right\} \sim \mathcal{N}\left[0, V\left[\left\{\widehat{SR}_k\right\}\right]\right]$, *then*

$$E\left[\max_k \left\{\widehat{SR}_k\right\}\right]\left(V\left[\left\{\widehat{SR}_k\right\}\right]\right)^{-1/2} \approx (1-\gamma)Z^{-1}\left[1-\frac{1}{K}\right] + \gamma Z^{-1}\left[1-\frac{1}{Ke}\right]. \tag{1}$$

Proof. It is known that the maximum value in a sample of independent random variables following an exponential distribution converges asymptotically to a Gumbel distribution. For a proof, see [**5**, pp. 138–147]. As a particular case, the Gumbel distribution covers the maximum domain of attraction of the Gaussian distribution, and therefore it can be used to estimate the expected value of the maximum of several independent random Gaussian variables.

To see how, suppose a sample of independent and identically distributed random variables, $y_k \sim \mathcal{N}[0,1]$, $k = 1, \cdots, K$. If we apply the Fisher-Tippet-Gnedenko theorem [**6**] to the Gaussian distribution, we derive an approximation for the sample maximum, $\max_k \{y_k\}$, leading to

$$\lim_{K\to\infty} \text{Prob}\left[\frac{\max_k \{y_k\} - \alpha}{\beta} \leq x\right] = G[x], \tag{2}$$

where $G[x] = e^{e^{-x}}$ is the CDF for the standard Gumbel distribution, $\alpha = Z^{-1}\left[1-\frac{1}{K}\right]$, $\beta = Z^{-1}\left[1-\frac{1}{Ke}\right] - \alpha$, and Z^{-1} denotes the inverse of the CDF of the standard Gaussian distribution. See [**10**] and [**5**] for a derivation of the normalizing constants (α, β).

The limit of the expectation of the normalized maxima from a distribution in the Gumbel maximum domain of attraction (see Prop. 2.1 (iii) in [**10**]) is

$$\lim_{K\to\infty} E\left[\frac{\max_k \{y_k\} - \alpha}{\beta}\right] = \gamma, \tag{3}$$

where γ is the Euler-Mascheroni constant. For a sufficiently large K, the mean of the sample maximum of standard normally distributed random variables can be approximated by

$$E\left[\max_k \{y_k\}\right] \approx \alpha + \gamma\beta \;=\; (1-\gamma)Z^{-1}\left[1-\frac{1}{K}\right] + \gamma Z^{-1}\left[1-\frac{1}{Ke}\right], \tag{4}$$

where $K >> 1$.

Now consider a set of estimated performance statistics $\left\{\widehat{SR}_k\right\}$, $k = 1, \cdots, K$, with independent and identically distributed Gaussian distribution centered at zero.

We make use of the linearity of the expectation operator to derive the expression

$$E\left[\max_k\left\{\widehat{SR}_k\right\}\right]\left(V\left[\left\{\widehat{SR}_k\right\}\right]\right)^{-1/2}\approx(1-\gamma)Z^{-1}\left[1-\frac{1}{K}\right]+\gamma Z^{-1}\left[1-\frac{1}{Ke}\right]. \tag{5}$$

This concludes the proof of the theorem. ■

5. EXPERIMENTAL ANALYSIS As Borwein emphasized in [**4**], the techniques of experimental mathematics can enhance the process of discovery and proof by (a) gaining insight and intuition, (b) discovering new patterns and relationships, (c) using graphics to suggest underlying principles, (d) testing and falsifying conjectures, (e) exploring a result to see if it is worth a formal proof, (f) replacing lengthy hand derivations, and (g) confirming analytically derived results.

In the case of the False Strategy theorem, while the result provides one with an approximation of the expected maximum Sharpe ratio, an experimental analysis can yield insights on the distribution of the approximation error and the speed of the asymptotic convergence, thus suggesting more general results. To that end, consider the following Monte Carlo experiment:

1. We generate a random array of size (SxK), where S is the number of Monte Carlo experiments, and K is number of trials among which the highest Sharpe ratio will be selected. The values in this random array are drawn from a standard normal distribution centered at zero.
2. The rows in this array are centered and scaled to match zero mean and $V\left[\left\{\widehat{SR}_k\right\}\right]$ variance.
3. The maximum value across each row, $\max_k\left\{\widehat{SR}_k\right\}$, is computed, resulting in a number S of such maxima.
4. We compute the empirical average value across the S maxima, namely $\widehat{E}\left[\left\{\widehat{SR}_k\right\}\right]$.
5. We then compute the estimation error, in relative terms, to the analytical prediction made by the theorem as $\epsilon=\widehat{E}\left[\left\{\widehat{SR}_k\right\}\right]/E\left[\left\{\widehat{SR}_k\right\}\right]-1$.
6. We repeat the previous steps R times, resulting in a set of estimation errors $\{\epsilon_r\}$, $r=1,2,\cdots,R$, and allowing us to compute the mean and standard deviation of the estimation errors associated with K trials.

Figure 1 helps to visualize the outcomes from this experiment, for a wide range of trials (in the plot, between 2 and 1000000). For $V\left[\left\{\widehat{SR}_k\right\}\right]=1$ and any given number of trials K, we simulate the maximum Sharpe ratio in $S=10000$ Monte Carlo experiments, so that we can derive the distribution of maximum Sharpe ratios. The y-axis shows that the distribution of the maximum Sharpe ratios $\left(\max_k\widehat{E}\left[\left\{\widehat{SR}_k\right\}\right]\right)$ for each number of trials K (x-axis), when the true Sharpe ratio is zero. Results with a higher probability are cast in a lighter color. For instance, if we conduct $K=1000$ trials, the expected maximum Sharpe ratio $E\left[\max_k\left\{\widehat{SR}_k\right\}\right]$ is 3.26, even though the true Sharpe ratio of the strategy is zero. As expected, there is a rising hurdle that the researcher must beat as he or she conducts more backtests. We can compare these experimental results with the results predicted by the False Strategy theorem, which are represented in the graph with a dashed line. The comparison of these two results

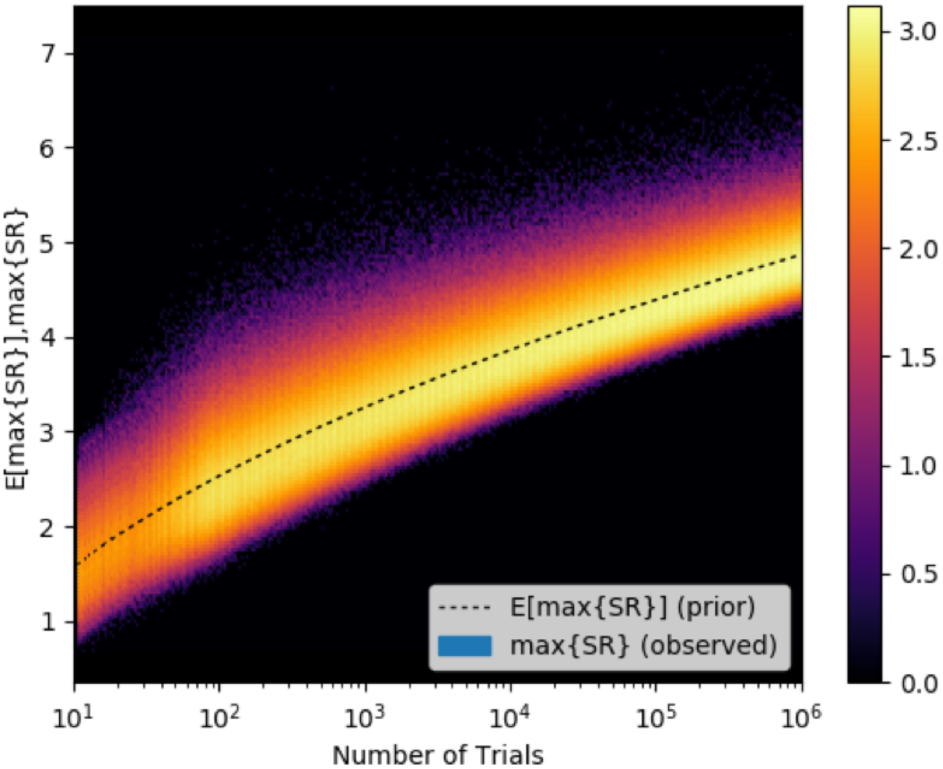

Figure 1. Comparison of experimental and theoretical results from the False Strategy theorem. Experimental maximum Sharpe ratios (for Monte Carlo trials where the true Sharpe ratio is zero) are plotted in color, Results with higher probability are cast in a lighter color. The dashed line are the results predicted by the False Strategy theorem.

(experimental and analytical) appears to indicate that the False Strategy theorem provides an accurate estimate of the expected maximum Sharpe ratio for a fairly wide range of trials studied.

We turn now our attention to evaluating the precision of the theorem's approximation. We define the approximation error as the difference between the experimental prediction (based on $S = 1000$ simulations) and the theorem's prediction, divided by the theorem's prediction. We can then re-evaluate these estimation errors $R = 100$ times for each number of trials K, and derive the mean and standard deviation of the errors. Figure 2 plots the results from this second experiment. The circles represent average errors relative to predicted values (y-axis), computed for alternative numbers of trials (x-axis). From this result, it appears that the False Strategy theorem produces asymptotically unbiased estimates. Only at $K \approx 50$, the theorem's estimate exceeds the experimental value by approximately 0.7%. The crosses represent the standard deviation of the errors (y-axis), derived for different numbers of trials (x-axis). From this experiment, we can deduce that the standard deviations are relatively small, below 0.5% of the values forecasted by the theorem, and they become smaller as the number of trials raises. These error estimates constitute upper boundaries, because the estimated errors would be smaller if we increased the number of Monte Carlo simulations.

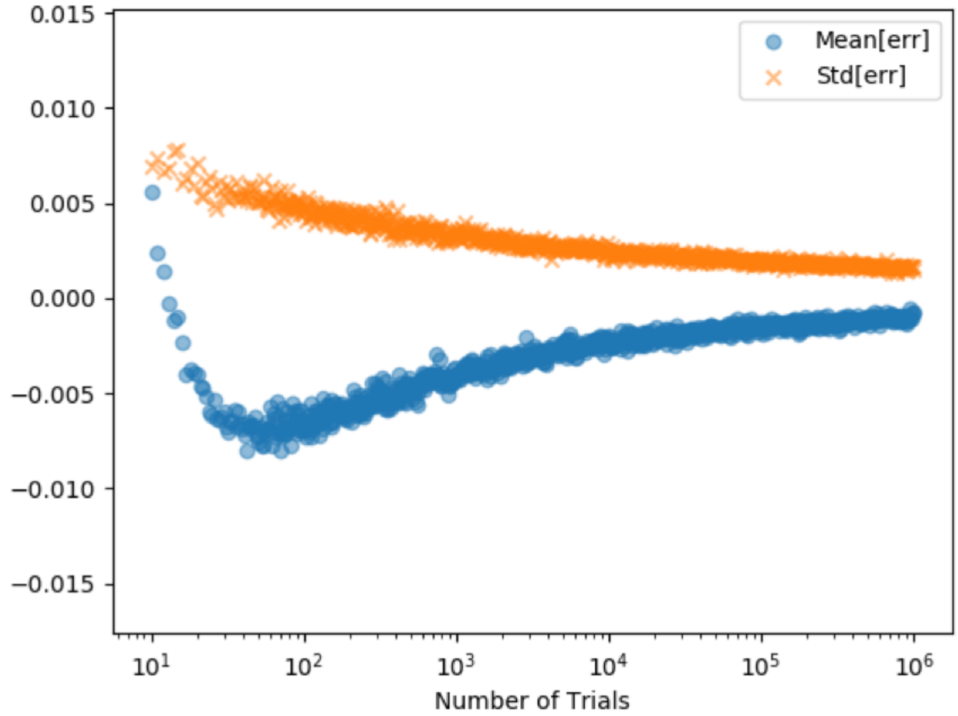

Figure 2. Statistics of the approximation errors as a function of the number of trials.

6. CONCLUSION The main conclusion from the False Strategy theorem is that, unless $\max_k \left\{\widehat{SR}_k\right\} >> E\left[\left\{\widehat{SR}_k\right\}\right]$, the discovered strategy is likely to be a *false positive*. The result can be used in connection with other techniques, such as the *deflated Sharpe ratio* [**3**], which estimates the probability of a false positive.

In real-world finance, the False Strategy theorem tells us that the optimal outcome of an unknown number of historical simulations is right-unbounded — with enough trials, there is no Sharpe ratio sufficiently large enough to reject the hypothesis that a strategy is false. Given the ease with which one can use a computer to explore many trials or variations of given strategy and only select the optimal variation, it follows that it is very easy to find impressive-looking strategy variations that are nothing more than false positives. This is the essence of *selection bias under multiple testing*.

As Jonathan Borwein and the present authors have argued [**3**], results such as the False Strategy theorem have important implications for the financial field. First, academic journals should cease to accept or publish articles that do not disclose the number of trials involved in a discovery. Second, financial regulators should withdraw their license from asset managers who publicize financial products that have not been rigorously vetted for selection bias under multiple testing. Third, investors should demand that the probability of a false discovery be reported with every product offering.

In this paper we have reflected on some of the contributions of Jonathan M. Borwein in the field of mathematical finance. This work was an extension of his broader concern for reproducibility in applied mathematics, and also was a showcase for experimental mathematics as a method to enhance the process of mathematical discovery and proof. We hope that this article unveils a common thread connecting all of these seemingly unrelated interests of this remarkable mathematician.

REFERENCES

1. D. H. Bailey and M. López de Prado, "The Sharpe ratio efficient frontier," *Journal of Risk*, vol. 15, no. 2 (2012), pp. 3-44.
2. D. H. Bailey and M. López de Prado, "The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting and non-normality," *Journal of Portfolio Management*, vol. 40, no. 5 (2014), pp. 94-107.
3. D. H. Bailey, J. .M Borwein, M. López de Prado and J. Zhu, "Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance," *Notices of the American Mathematical Society*, vol. 61, no. 5 (2014), pp. 458-471, `http://ssrn.com/abstract=2308659`.
4. J. M. Borwein and D. H. Bailey, *Mathematics by Experiment: Plausible Reasoning in the 21st Century*, Second Edition, A.K. Peters, Natick, MA, 2008.
5. P. Embrechts, C. Klueppelberg and T. Mikosch, *Modelling Extremal Events*, Springer-Verlag, 1st edition, 2003.
6. R. A. Fisher, *The Genetical Theory of Natural Selection*, Oxford University Press, Oxford, 1930.
7. W. Greene, *Econometric Analysis*, Pearson, 7th edition, 2011.
8. J. Hamilton, *Time Series Analysis*, Princeton University Press, 1st edition, 1994.
9. J. Neyman and E. Pearson, "IX. On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society*, series A, vol. 231, no. 694–706 (1933), pp. 289337.
10. S. Resnick, *Extreme Values, Regular Variation and Point Processes*, Springer, 1st edition, 1987.
11. W. Sharpe, "Mutual fund performance," *Journal of Business*, vol. 39, no. 1 (1966), pp. 119-138.
12. W. Sharpe, "Adjusting for risk in portfolio performance measurement," *Journal of Portfolio Management*, vol. 1, no. 2 (Winter 1975), pp. 29–34.
13. W. Sharpe, "The Sharpe ratio," *Journal of Portfolio Management*, vol. 21, no. 1 (Fall 1994), pp. 49–58.

MARCOS LÓPEZ DE PRADO earned a PhD in financial economics (2003), a second PhD in mathematical finance (2011) from Universidad Complutense de Madrid, and is a recipient of Spain's National Award for Academic Excellence (1999). He currently is a lecturer at Cornell University's School of Engineering, and is the Chief Executive Officer of True Positive Technologies, LLC.
Cornell University, Ithaca, NY 14853
ml863@cornell.edu

DAVID H. BAILEY earned a PhD in mathematics from Stanford University (1977), and is a recipient of the Chauvenet Prize and the Merten Hesse Prize from the Mathematical Association of America, and the Levi L. Conant Prize from the American Mathematical Society. He recently retired from the Lawrence Berkeley National Laboratory, but continues as a Research Associate at the Laboratory and also at the University of California, Davis.
Lawrence Berkeley National Laboratory, Berkeley, CA 94720
DHBailey@lbl.gov