The Deflated Sharpe Ratio in Practice: Guarding Strategy Selection Against Multiple-Testing Bias

Dmitri De Freitas

Washington University in St. Louis — B.S. Data Science & Financial Engineering

Working paper · July 2026 · Interactive companion: dmitridefreitas.com/lab/backtest-stats

Abstract. A backtested Sharpe ratio is a random variable, and the largest of many random variables is large by construction. This note assembles, implements, and stress-tests the Sharpe-ratio inference framework of Bailey and López de Prado — the Probabilistic Sharpe Ratio (PSR), the expected maximum Sharpe ratio under multiple trials, the Deflated Sharpe Ratio (DSR), and the Minimum Track Record Length (MinTRL) — in a from-scratch, dependency-light Python implementation with a pinned test suite. In a Monte Carlo experiment, the best of 200 zero-skill strategies exhibits an annualized Sharpe ratio of 2.12 and a naive PSR of 0.9999; the DSR, accounting for the 200 trials behind the selection, assigns the same track record only 0.81 — below any conventional confidence threshold. The exercise quantifies how aggressively conventional backtest statistics overstate evidence of skill, and why the number of trials belongs in the statistic itself.

1 Introduction

Quantitative strategy research is a search process: a researcher varies signals, lookbacks, filters, and universes until a backtest looks attractive. Each variation is a statistical trial, and the reported result is almost always the *maximum* over those trials. Classical inference on the winning Sharpe ratio — a t-test against zero — silently assumes a single trial, which makes it close to meaningless in practice. Harvey, Liu and Zhu (2016) argue most published factors fail once multiple testing is accounted for; Bailey and López de Prado (2012, 2014) supply a practical correction that requires only quantities the researcher already has: the moments of the candidate's returns and the number of trials behind the selection.

This note has three goals: (i) state the framework compactly with consistent notation; (ii) describe a from-scratch implementation, `backtest-statistics`, published with a pinned test suite; and (iii) demonstrate the machinery on a controlled experiment where the truth — zero skill — is known.

2 The sampling distribution of the Sharpe ratio

Let $r_1, \ldots, r_T$ denote per-period strategy returns with estimated (per-period) Sharpe ratio $SR = \hat{\mu}/\hat{\sigma}$. Following Lo (2002) and Mertens (2002), under mild conditions the estimator is asymptotically normal with standard error

$$\hat{\sigma}(SR) = \sqrt{\left[\,1 - \gamma_3 \cdot SR + \left((\gamma_4 - 1)/4\right) \cdot SR^2\right) / (T - 1)\,}\,] \tag{1}$$

where γ_3 and γ_4 are the skewness and (non-excess) kurtosis of the returns. Negative skew and fat tails — the norm for systematic strategies — inflate the standard error relative to the Gaussian case, meaning conventional t-statistics overstate significance precisely for the strategies most likely to harbor tail risk.

3 The Probabilistic Sharpe Ratio

The PSR converts (1) into the probability that the true Sharpe ratio exceeds a benchmark SR*:

$$PSR(SR^*) = \Phi\left((SR - SR^*)\cdot\sqrt{(T-1)} \,/\, \sqrt{(1 - \gamma_3\cdot SR + ((\gamma_4-1)/4)\cdot SR^2)} \right) \tag{2}$$

PSR is a strictly more honest headline number than the Sharpe ratio itself: it embeds the track-record length and the non-normality of returns. It remains, however, a *single-trial* statistic.

4 The expected maximum of N trials

Suppose a researcher runs N independent zero-skill trials whose estimated Sharpe ratios have cross-sectional variance V. Extreme-value results for Gaussian maxima give the expectation of the best observed Sharpe ratio:

$$E[\max SR] \approx \sqrt{V}\cdot \left[(1-\gamma)\cdot\Phi^{-1}(1 - 1/N) + \gamma\cdot\Phi^{-1}(1 - 1/(Ne)) \right] \tag{3}$$

with $\gamma \approx 0.5772$ the Euler–Mascheroni constant. Equation (3) is the researcher's null: it is what "the best strategy" looks like when nothing works. It grows without bound in N — roughly like $\sqrt{(2\cdot V\cdot \ln N)}$ — which is why undisclosed search intensity is the most corrosive form of backtest overfitting.

5 The Deflated Sharpe Ratio

The DSR is simply the PSR evaluated against that null instead of zero:

$$DSR = PSR\left(E[\max SR] \right) \tag{4}$$

A DSR of 0.95 says: even after granting that you kept the best of N attempts, the track record still clears the expected-maximum hurdle with 95% confidence. The two inputs beyond the candidate's own returns — N and V — are knowable by the researcher (and only by the researcher), which is why the DSR is best understood as a discipline one imposes on oneself.

6 Minimum Track Record Length

Inverting (2) in T yields the number of observations required before a track record supports a claim at confidence α:

$$MinTRL = 1 + (1 - \gamma_3\cdot SR + ((\gamma_4-1)/4)\cdot SR^2) \cdot \left(\Phi^{-1}(\alpha) \,/\, (SR - SR^*) \right)^2 \tag{5}$$

MinTRL diverges as SR → SR*: marginal edges require impractically long histories to certify, a sobering result for allocators evaluating young funds.

7 Monte Carlo demonstration

We simulate N = 200 strategies of T = 750 daily returns each, drawn i.i.d. N(0, 1%) — zero skill by construction — and select the best by in-sample Sharpe ratio. Table 1 reports the inference statistics for the selected strategy (seed 42; the experiment also runs, with different seeds, inside the repository's test suite).

Statistic	Value	Reading
Best annualized Sharpe (of 200)	2.12	looks institutional-grade
Naive PSR vs. SR* = 0	0.9999	"significant" at any level
Cross-trial variance of SR	0.00134	input to eq. (3)
E[max SR] hurdle, annualized	1.61	what luck alone produces
DSR	**0.81**	**fails a 95% bar — correctly**
MinTRL (95%, SR* = 0)	153 days	misleading if N is ignored

Table 1: The best of 200 zero-skill strategies under single-trial vs. multiple-testing-aware inference.

The contrast is the entire argument: identical data, identical returns, and the conclusion flips from "hire this PM" to "this is noise" the moment the 199 discarded trials enter the statistic. Any research process that does not log N cannot produce a DSR — which is itself an argument for logging N.

8 Implementation

The reference implementation (`github.com/dmitridefreitas-dev/backtest-statistics`) is ~160 lines of NumPy and standard library: the normal CDF via `math.erf`, the inverse CDF via Acklam's rational approximation with a Halley refinement ($\approx 10^{-15}$ accuracy), and equations (1)–(5) directly transcribed. Thirteen pinned tests cover closed-form reductions, monotonicity in T and N, PSR/MinTRL self-consistency, and the Table-1 experiment. An interactive version, including a Monte Carlo p-hacking simulator, runs at `dmitridefreitas.com/lab/backtest-stats`.

9 Conclusion

The Sharpe ratio of a selected strategy answers the wrong question. PSR fixes the distributional question; DSR fixes the selection question; MinTRL tells you how long you must wait for the answer to mean anything. All three are cheap to compute and freely available — the binding constraint is the willingness to count one's own trials.

References

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

Bailey, D. H., and M. López de Prado (2014). "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality." *Journal of Portfolio Management* 40(5), 94–107.

Harvey, C. R., Y. Liu, and H. Zhu (2016). "…and the Cross-Section of Expected Returns." *Review of Financial Studies* 29(1), 5–68.

Lo, A. W. (2002). "The Statistics of Sharpe Ratios." *Financial Analysts Journal* 58(4), 36–52.

Mertens, E. (2002). "Comments on Variance of the IID estimator in Lo (2002)." Working note, University of Basel.

Working paper — comments welcome: d.defreitas@wustl.edu · Code and interactive companion: github.com/dmitridefreitas-dev/backtest-statistics · dmitridefreitas.com/lab/backtest-stats