A Backtest Measures Not the Quality of a Strategy but the Number of Attempts

— Even a strategy with no skill at all shows an annualised Sharpe ratio scattered with a standard error of **0.447214** over five years of daily data — the separator is whether the effective number of attempts is reported — [Paper 706]

Yuuki YAMAGISHI (山岸 勇輝)
Yamagishi Complexity Torus Lab, Japan
September 2026

— From thinking to syncing. —

Abstract

A backtest result is read as a measure of how good a strategy is. But with no skill at all, enough attempts always produce a good figure.How many attempts produce how much is counted here.No new theorem or law is claimed. First, even with no skill the figure scatters——an annualised Sharpe ratio measured over five years of daily data scatters with a standard error of $\mathbf{0.447214}$ (Section 2). **Second, and this is the core. Taking the best of many always yields a good number——with $N = 1000$ attempts, a strategy of no skill gives a best of $\mathbf{1.449610}$** (Section 2). **Third, the number of attempts required is startlingly small——reaching a best above 1.0 takes only $\mathbf{49}$** (Section 3). **Fourth, the growth is only logarithmic——$\mathbf{0.688147}$ at $N = 10$, $\mathbf{1.121425}$ at $N = 100$ and $\mathbf{1.722490}$ even at $N = 10000$: a thousandfold in attempts buys a factor of 2.503086** (Section 2). **Fifth, two routes to the number agree——numerical integration gives 1.449610 and 4000 random trials give $\mathbf{1.451801}$, a difference of $\mathbf{0.002191}$** (Section 3). **Sixth, the separator is whether the effective number of attempts is reported——without it, the reported figure cannot be interpreted at all** (Section 4). **Keywords**: backtest overfitting, Sharpe ratio, multiple comparison, expected maximum, effective number of trials, selection bias.

Scope of this paper (scope note): **No new theorem or law is claimed——selection bias under multiple comparison, the estimation error of a Sharpe ratio and the logarithmic growth of an expected maximum are all standard (Bailey et al. 2014, Sharpe 1994).No real strategy is treated——nothing is fitted to a real strategy or to market data; only a strategy with no skill at all is considered.The estimation error is simplified——the standard error of an annualised Sharpe ratio is fixed at the rule of thumb $1/\sqrt{\text{years}}$, with no autocorrelation and no heavy tails.Remedies are not evaluated——out-of-sample testing and the deflated Sharpe ratio are mentioned and not compared.Nothing particular is said——no claim is made about any specific fund, product or person.Relation to earlier papers**: **Paper 317 showed that a significance level is a promise per test and not per study——there the price of multiple comparison was counted on a binary verdict. Here the same price is counted on the maximum of a continuous quantity. A binary verdict is stopped by the words "not significant"; a continuous figure is simply reported as a result.**What is added is **obtaining the best Sharpe ratio for each number of attempts by numerical integration**, **obtaining the same by random trials and matching to within $\mathbf{0.002191}$**, **showing that $\mathbf{49}$ attempts suffice to pass 1.0**, **showing the bound $\sqrt{2\ln N}$ to exceed the true expectation**, and **placing the separator on whether the effective number of attempts is reported**.

1. Notation and the Question

Setting: **consider a strategy whose expected return is exactly zero. Measured over T years of daily data, its annualised Sharpe ratio scatters with a standard error of about $1/\sqrt{T}$. Take the best of N such.**

$$\mathrm{SE} = \frac{1}{\sqrt{T}}, \qquad \mathbb{E}\left[\max_{1\le i\le N} Z_i\right], \qquad \mathbb{E}\left[\max \mathrm{SR}\right] = \mathrm{SE}\cdot\mathbb{E}\left[\max Z_i\right]$$

Each Z_i is standard normal. With $T = 5$ years, $\mathrm{SE} = 0.447214$.

Question: **does a good result measure the quality of a strategy, or the number of attempts?**

2. No Skill Still Produces Good Numbers

Computation 2.1: **the expected best Sharpe ratio was obtained by numerical integration for each number of attempts.**

attempts	best Sharpe ratio	bound $\sqrt{2\ln N}$ scaled
1	-0.000004	—
10	0.688147	0.959705
100	1.121425	1.357228
1000	1.449610	1.662258
10000	1.722490	1.919410

Observation 2.2

this is the core: zero at $N = 1$, and 1.449610 at $N = 1000$.Not one strategy has changed; only the number of attempts has——the reported figure measures the number of attempts, not the quality.

Observation 2.3

the growth is only logarithmic.Raising N from 10 to 10000, a thousandfold, moves the best from 0.688147 to 1.722490, a factor of 2.503086——so hearing "we tried many" tells one nothing about how many, and the figure itself cannot say.

Observation 2.4

the bound overshoots.The scaled $\sqrt{2\ln N}$ gives 1.662258 at $N = 1000$, above the true 1.449610——estimating with the bound overstates, so the expectation itself has to be computed.

3. The Number Required Is Small

Computation 3.1: the number of attempts needed for the best to exceed 1.0 was found to be 49.

Observation 3.2

forty-nine is a number reached without trying to.A handful of indicators, a few windows and a few thresholds combine past it at once——few people believe they tried a thousand; everybody tries forty-nine.

Computation 3.3: the same quantity was obtained from 4000 random trials: 1.451801 at $N = 1000$, differing from the integration by 0.002191.

Observation 3.4

two routes were used so that the number has more than one source.A coarse grid spoils the integration and too few trials spoil the sampling, and if either were wrong they would not agree——the agreement is evidence that both are right.

Observation 3.5

hence the need for an effective number of attempts.From the reported figure alone, $N = 1$ cannot be told from $N = 1000$——one and the same 1.4 serves as evidence of skill and as evidence of counting.

4. The Separator

Observation 4.1

the separator is whether the effective number of attempts is reported.If it is, the figure can be discounted by it; if it is not, the figure cannot be interpreted at all——what is missing is not precision but one item of information.

Observation 4.2

beside Paper 317.Paper 317 showed that a significance level is a promise per test and not per study——there the price of multiple comparison landed on a binary verdict; here it lands on the maximum of a continuous quantity. A binary verdict is stopped by “not significant”; a continuous figure is simply reported, and nothing stops it. That makes this case the more dangerous.

Observation 4.3

what is not said.Nothing is fitted to a real strategy or to data, and remedies are not compared——what is shown is that attempts alone produce figures, and that few attempts suffice.

5. Roll Call and Closing

1. **standard error $\mathbf{0.447214}$** (Section 1): **five years of daily data.**
2. **$\mathbf{-0.000004}$ at $N = 1$** (Computation 2.1): **no skill.**
3. **$\mathbf{1.449610}$ at $N = 1000$** (Computation 2.1): **the core.**
4. **$\mathbf{0.688147}$ already at $N = 10$** (Computation 2.1): **it comes quickly.**
5. **$\mathbf{1.722490}$ at $N = 10000$** (Computation 2.1): **growth is logarithmic.**
6. **the bound is $\mathbf{1.662258}$** (Computation 2.1): **above the truth.**
7. **$\mathbf{49}$ attempts reach 1.0** (Computation 3.1): **reached without trying.**
8. **random trials give $\mathbf{1.451801}$** (Computation 3.3): **a second route.**
9. **differing by $\mathbf{0.002191}$** (Computation 3.3): **agreement as evidence.**
10. **is the count reported** (Observation 4.1): **the separator; Paper 317.**

Closing: **A backtest measures not the quality of a strategy but the number of attempts.Even with no skill at all, an annualised Sharpe ratio over five years of daily data scatters with a standard error of 0.447214——take the best of N such and the expectation is 0.688147 at $N = 10$, 1.449610 at $N = 1000$ and 1.722490 at $N = 10000$.Not one strategy has changed; only the number of attempts has——and reaching a best above 1.0 takes only 49.The growth being logarithmic, a thousandfold in attempts buys a factor of 2.503086, so the figure itself cannot reveal how many were tried.Numerical integration gave 1.449610 and 4000 random trials gave 1.451801, agreeing within 0.002191 ——two routes were used so that the number has more than one source.The separator is whether the effective number of attempts is reported.Placed among the earlier papers——in Paper 317 the price of multiple comparison landed on a binary verdict; here it lands on a continuous figure, and nothing stops it.To be honest——nothing is fitted to a real strategy or to data, and the estimation error is fixed at a rule of thumb.**

References

1. Bailey, D. H., Borwein, J., Lopez de Prado, M., & Zhu, Q. J. (2014). Pseudo-mathematics and financial charlatanism: the effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*.

2. Sharpe, W. F. (1994). The Sharpe ratio. *The Journal of Portfolio Management*.

3. Yamagishi, Y. (2026). *A Significance Level Is a Promise per Test, Not per Study* [Paper 317]. Zenodo. 10.5281/zenodo.22208819

— From thinking to syncing. —

Yuuki YAMAGISHI (山岸 勇輝)

© 2026 Yamagishi Complexity Torus Lab, Japan · Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)