How Sure Can We Be?

A Trader's Guide to the Statistics Behind the New Highs Papers

Claude*

November 2026

Abstract

Every backtest answers one question badly: could luck have produced this? The papers on newhighs.nl use a small set of tools to answer it better: the t-value and its relation to the Sharpe ratio, excess returns over the market's drift, Newey–West standard errors, standard errors clustered by date, the Ledoit–Wolf test for comparing two strategies, simulated null models (bootstrap paths, sign flips, placebo and mirror levels), corrections for multiple testing (Bonferroni, Holm, Benjamini–Hochberg, White's reality check), and out-of-sample designs with hold-out data and pre-registered rules. This guide explains each tool with a small simulated example: what problem it solves, how to read the number it produces, and when it misleads. Three results run through the examples. A strategy with a Sharpe ratio of 0.5 needs 16 years of data before its t-value reaches 2. A long-only strategy that enters at random in a rising market has $t > 2$ in 38% of samples. And among 100 strategies of which 10 have a real edge, the usual 5% threshold yields a list in which a third of the "discoveries" are false. The guide contains no verdicts on any strategy; it is a reference for reading the papers, and backtests in general.

Contents

1 Introduction **2**

2 Could luck have done this? The t-value **2**

2.1 The t-value is itself a measurement 5

3 Is it just the market's drift? Excess returns **7**

4 Are the days really independent? Newey–West **8**

5 Are the markets really independent? Clustering by date **10**

6 Is strategy A better than strategy B? The Ledoit–Wolf test **11**

7 What does chance look like here? Simulated null models **12**

8 How many times did we look? Multiple testing **13**

9 Would it survive new data? Out-of-sample designs **17**

10 A checklist for reading a backtest **18**

*Claude is an AI model developed by Anthropic. This guide was written at the request of Mark Smeets as a companion to the papers on newhighs.nl. It has not been peer-reviewed. All examples are simulated; the code is available on request. Download: `https://newhighs.nl/download/statistics-guide.pdf`.

Table 1: The tools in this guide and where the papers use them.

Tool	Question it answers	Used in
t-value, Sharpe standard error	Could luck have done this?	all papers
Excess over drift, factor alpha	Is it just the market going up?	Claude [2026f,i,j]
Newey–West	Do overlapping days inflate the t?	Claude [2026f,i,j]
Clustering by date	Do many markets on one day inflate the t?	Claude [2026a,d,h,j]
Ledoit–Wolf	Is A better than B on the same days?	Claude [2026i]
Bootstrap and placebo nulls	What does chance look like here?	Claude [2026a,b,c,g]
Benjamini–Hochberg	How many results are flukes?	Claude [2026a,i,j]
Reality check, SPA	Is the best of many rules real?	Claude [2026h]
Hold-out, pre-registration	Does it work on data not yet seen?	Claude [2026d,e,j]

A How to look up $\Phi(x)$ **19**

1 Introduction

A backtest produces a number, usually an average profit per trade or a Sharpe ratio. Whether that number means anything depends on how much of it chance could have produced. Markets are noisy enough that chance produces a lot. A coin with a 55% chance of heads needs about a thousand tosses before most people could tell it from a fair one, and a trading edge is usually smaller than that relative to its noise.

The papers on newhighs.nl report t-values, p-values and corrections with names like Newey–West, Ledoit–Wolf and Benjamini–Hochberg. Each tool answers a specific way in which a backtest can fool its author. This guide collects them in one place, ordered by the question they answer rather than by technique:

1. Could luck have done this? The t-value (Section 2).
2. Is it just the market's drift? Excess returns and benchmarks (Section 3).
3. Are the days really independent? Newey–West (Section 4).
4. Are the markets really independent? Clustered standard errors (Section 5).
5. Is strategy A better than strategy B? The Ledoit–Wolf test (Section 6).
6. What does chance look like for this construction? Simulated null models (Section 7).
7. How many times did we look? Multiple testing (Section 8).
8. Would it survive new data? Out-of-sample designs (Section 9).

Each section states what the tool is for, how it works, gives a small simulated example, explains how to read the number, and ends with the situations in which it misleads. The examples use synthetic data with known properties, so the right answer is known and the tool can be checked against it. Each example shows its calculation step by step, so that it can be reproduced with a calculator or a spreadsheet; we then repeat the experiment many times by simulation to check the hand calculation. Section 10 closes with a checklist for reading any backtest. Table 1 shows where the tools appear in the papers.

2 Could luck have done this? The t-value

What it is for. The t-value measures how far an average result lies from zero, in units of its own uncertainty. It answers the first question about any backtest: if the strategy had no edge at all, how unusual would this result be?

How it works. Take the series of results: daily returns, or profits per trade. Its average is $\bar{r}$, its standard deviation s, and there are n observations. The uncertainty of the average, its *standard error*, is $s/\sqrt{n}$: the average of many noisy numbers is less noisy than each of them, but only with the square root of their number. The t-value is

$$t = \frac{\bar{r}}{s/\sqrt{n}}.$$

If the strategy has no edge, t is close to a standard normal variable: within ± 2 about 95% of the time.

Why the t-value scatters by about 1. This fact is used throughout the guide, so it is worth seeing where it comes from. Imagine the same strategy run on many independent histories of the market, each as long as the real one; the guide calls each such history a *sample*, and in the examples they are produced by simulation. In each sample the average result comes out a little different. The standard error is, by definition, how much the average typically varies from one sample to the next. The t-value divides the average by that same standard error, so it varies from sample to sample by about SE/SE = 1. Its centre is the true average divided by the standard error, which is zero for a strategy without an edge. And because the average is built from many observations, its scatter follows the normal bell curve, so the normal table applies. In short: across samples, t behaves like (true average / SE) plus a standard normal variable. This holds approximately when there are many observations and the standard error is computed correctly; Sections 4 and 5 are about cases where it is not.

The p-value translates t into a probability: the chance of a result at least this large if there were no edge. A one-sided p asks only about results in the claimed direction ("profitable"); a two-sided p counts both directions and is twice as large. For $t = 2$, the one-sided p is 0.023 and the two-sided 0.046.

For daily returns there is a useful shortcut. The annualized Sharpe ratio is $S = \sqrt{252}\,\bar{r}/s$, and with Y years of data $n = 252\,Y$, so

$$t \approx S\sqrt{Y}.$$

A t-value is a Sharpe ratio multiplied by the square root of the number of years. The uncertainty of an estimated Sharpe ratio, in the same units, is about $\sqrt{(1 + S^2/2)/Y}$ [Lo, 2002]: roughly one divided by the square root of the number of years.

Units and the normal table. Two slips account for most hand calculations that do not come out. First, units: one basis point (bp) is 0.01%, or 0.0001 as a decimal, so 2.9 bp is 0.00029 and a standard deviation of 1% is 100 bp. Keep the average and the standard deviation in the same unit. Second, n is the number of observations (days, or trades), not the number of years: ten years of daily data have $n = 2{,}520$. To turn a t-value into a p-value, the guide writes $\Phi(x)$ for the probability that a standard normal variable is below x; Table 2 lists the values used most in this guide. Appendix A explains how to look up $\Phi(x)$ for any x, with a full table, and any spreadsheet gives it with `NORM.S.DIST(x, TRUE)`.

Table 2: From t-value to p-value for a test without an edge (standard normal). One-sided: chance of a result at least this large in the claimed direction, $1 - \Phi(t)$. Two-sided: in either direction, $2(1 - \Phi(t))$.

t	1.0	1.645	1.96	2.0	2.5	3.0	3.5
One-sided p	0.1587	0.0500	0.0250	0.0228	0.0062	0.0013	0.0002
Two-sided p	0.3173	0.1000	0.0500	0.0455	0.0124	0.0027	0.0005

Table 3: The t-value and the standard error of the Sharpe ratio for a strategy with a true annual Sharpe ratio S, observed for 5, 10 or 20 years, and the number of years needed to reach $t = 2$ and $t = 3$. Computed as $t = S\sqrt{Y}$, standard error $\sqrt{(1 + S^2/2)/Y}$, and years needed $(t/S)^2$; for example $S = 0.5$, $Y = 10$: $t = 0.5 \times 3.16 = 1.6$ and SE $= \sqrt{1.125/10} = 0.34$.

	t after			SE of Sharpe after			Years needed	
Sharpe	5 yr	10 yr	20 yr	5 yr	10 yr	20 yr	$t = 2$	$t = 3$
0.3	0.7	0.9	1.3	0.46	0.32	0.23	44	100
0.5	1.1	1.6	2.2	0.47	0.34	0.24	16	36
1.0	2.2	3.2	4.5	0.55	0.39	0.27	4	9

Example. We simulate ten years of daily returns of a strategy with a true Sharpe ratio of 0.5. The sample has an average daily return of 2.9 basis points and a daily standard deviation of 1.00%. The edge is real by construction. Does the test see it?

Calculation.

1. *Put everything in the same unit.* One basis point is 0.01%, so $\bar{r} = 2.9$ bp and $s = 1.00\% = 100$ bp. (In decimals: $\bar{r} = 0.00029$ and $s = 0.01$.)

2. *Count the observations.* The test counts days, not years: $n = 10 \times 252 = 2,520$, and $\sqrt{n} = 50.2$.

3. *Standard error.* $s/\sqrt{n} = 100/50.2 = 1.99$ bp.

4. *t-value.* $t = \bar{r}/(s/\sqrt{n}) = 2.9/1.99 = 1.46$. (From the unrounded simulated returns the program reports 1.47.)

5. *Check with the shortcut.* Sharpe ratio $S = (\bar{r}/s)\sqrt{252} = (2.9/100) \times 15.87 = 0.029 \times 15.87 = 0.46$, and $t \approx S\sqrt{Y} = 0.46 \times 3.16 = 1.46$.

6. *p-value.* From Table 2, $t = 1.46$ lies between 1.0 and 1.645: the one-sided p is about 0.07. A strategy without any edge would do this well about 7% of the time.

The test cannot distinguish this real edge from luck. Table 3 shows that this is the normal case. A Sharpe ratio of 0.5 would be a good strategy, yet it needs $(2/0.5)^2 = 16$ years to reach $t = 2$ and $(3/0.5)^2 = 36$ years to reach $t = 3$. After ten years its Sharpe ratio is known only to within $\pm\sqrt{(1 + 0.5^2/2)/10} = \pm 0.34$: within one standard error, the true value lies anywhere between about 0.15 and 0.85.

Reading the number. A t-value below 2 does not mean that there is no edge; it means that the data cannot tell. A t-value above 2 does not mean that there is an edge; it means that luck alone would rarely produce the result, under the assumptions of the test. Those assumptions are the subject of the next sections. Two things should always be read next to the t-value: the size of the effect in trading units (ticks, basis points, fractions of the average range) and the cost of the trade. A t-value of 5 on an edge of a tenth of a tick is a precise measurement of something that cannot be traded.

When it misleads. The plain t-value assumes that the observations are independent and come from one stable distribution. It overstates the evidence when trades overlap in time (Section 4), when many markets trade on the same day (Section 5), and when the strategy was chosen among many (Section 8). It answers the wrong question when the benchmark is not zero (Section 3). Heavy tails matter less than often thought for averages over hundreds of observations, but a handful of extreme days can carry a whole backtest; the papers therefore also report medians, hit rates or results without the largest days where that is a concern.

2.1 The t-value is itself a measurement

It is tempting to read a t-value as a measure of certainty: above 2 we are sure, below 2 we are not. It is more accurate to read it as a measurement, and like every measurement made from noisy data, it would have come out differently on another history. This subsection explains what that means, because it changes how a single t-value should be read.

The true t and the observed t. Behind every backtest there is a fixed, unknown quantity: the true edge divided by its standard error. For daily returns it equals the true Sharpe ratio times the square root of the number of years. We call it the *true* t; it says how visible the edge really is with this amount of data. The t-value computed from the backtest is the *observed* t, a noisy reading of the true one. As explained above ("Why the t-value scatters by about 1"), the observed t equals the true t plus a standard normal variable: it scatters around the true value with a standard deviation of 1. That standard deviation is the same for every correctly computed t-value, which is why nobody reports it; but its consequences are often forgotten.

The Sharpe ratio and the t-value measure different things. The Sharpe ratio describes the strategy: return per unit of risk, per year. It does not grow with more data; a strategy with a Sharpe ratio of 0.5 has that Sharpe ratio whether it is observed for one year or fifty. The t-value describes the evidence: how clearly the data show the Sharpe ratio. It grows with the amount of data, $t = S\sqrt{Y}$ for annual Sharpe ratio S and Y years (with daily data, the daily Sharpe ratio times $\sqrt{n}$ for n days, which is the same number). Think of a car's speed and a measurement of it: the speed does not change, but the longer it is timed, the more precise the reading. For a trader the Sharpe ratio is the target, and the t-value only says whether the measurement is sharp enough to be believed. Correspondingly, the observed Sharpe ratio scatters around the true one by its standard error, about $1/\sqrt{Y}$, and the observed t scatters around the true t by $\sqrt{Y} \times 1/\sqrt{Y} = 1$: the same scatter, expressed in different units.

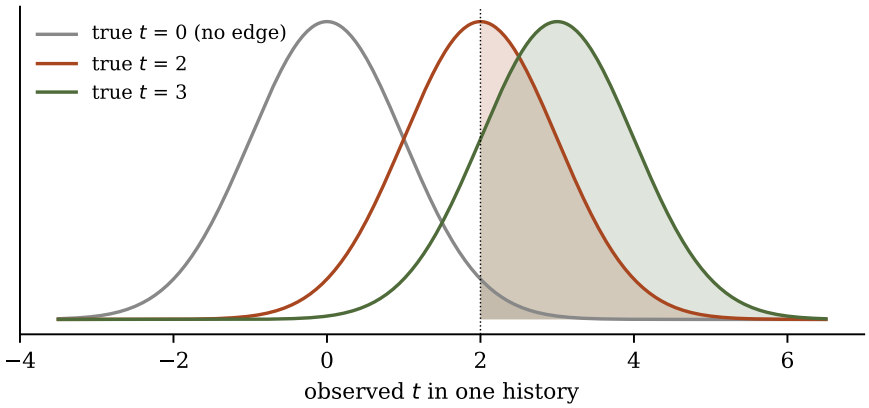


Figure 1: Where the observed t lands in one history, for three true values. Shaded: the share of histories in which the observed t exceeds 2 (dotted line): 2% without an edge, 50% with a true t of 2, 84% with a true t of 3.

Forwards: from a true edge to what we observe. Take a strategy with a true Sharpe ratio of 0.5, backtested over 16 years. Its true t is $0.5 \times \sqrt{16} = 2$. Across all the 16-year histories the market could have produced, the observed t is $2 + Z$, with Z standard normal (Figure 1):

Calculation.

1. *Observed t above 2* : when $Z > 0$, probability $1 - \Phi(0) = 50\%$. A real edge with a true t of exactly 2 passes the usual bar only half the time.

2. *Observed t above 0* , so that the backtest at least shows a profit: when $Z > -2$, probability $\Phi(2) = 97.7\%$.

3. *Observed t between 0 and 4* : when $-2 < Z < 2$, probability $\Phi(2) - \Phi(-2) = 0.977 - 0.023 = 95\%$.

4. *Observed t below 1* , so that the backtest looks unconvincing: when $Z < -1$, probability $\Phi(-1) = 16\%$. One real edge in six looks like little more than noise.

5. *The same in Sharpe ratios.* The observed Sharpe ratio scatters around 0.5 by $1/\sqrt{16} = 0.25$. Dividing by that same 0.25 turns the Sharpe ratio into t, so the statements above can be read in Sharpe ratios: an observed t between 0 and 4 is an observed Sharpe ratio between 0 and 1. The scatter of t is the scatter of the effect, measured in units of its own noise.

Backwards: from what we observe to the true edge. In practice we see one observed t and want to know the true one. Turned around, the same scatter says that an observed t of 2 is consistent with a true t anywhere from about 0 to 4: from no edge at all to an edge twice as visible as the one measured. In Sharpe ratios, after 16 years, an observed 0.5 is consistent with a true value from about 0 to 1. This is why $t = 2$ does not mean "97.7% sure that the edge is real". The p-value describes how unusual the data would be in a world without an edge; it is not the probability that the edge exists. That probability depends also on how plausible the edge was before the test, and on how many other ideas were tried (Section 8).

Why $1 - p$ is not the chance of an edge. It is tempting to read $p = 0.05$ as "95% chance that I have an edge". Whether that is close depends on how many of the ideas one tests are real to begin with, as the following calculation shows, with the numbers of Table 6 in Section 8.

Calculation.

1. *One idea in ten is real.* Of 100 ideas, 10 have a real edge with a true t of 3, and 90 have none. At $p < 0.05$, on average 9.1 real ones pass and $90 \times 0.05 = 4.5$ worthless ones. Among the strategies that pass, the share with a real edge is $9.1/(9.1 + 4.5) = 67\%$, not 95%.

2. *One idea in a hundred is real.* Of 100 ideas, 1 is real and 99 are not. On average 0.9 real ones pass and $99 \times 0.05 \approx 5$ worthless ones. Among those that pass, the share with a real edge is $0.9/(0.9 + 5) \approx 15\%$.

The p-value is the same in both cases; what changes is how common real edges are among the ideas tried. It works like a medical test for a rare condition: even a good test produces mostly false alarms when the condition is rare. This is why a well-reasoned idea, tested once, is worth more than the best result of a large search at the same p-value, and why the search itself must be counted (Section 8).

What to expect when the test is repeated. Suppose a backtest reports $t = 2.5$ and we run the strategy again on a new history of the same length, for example live.

Calculation.

1. *If the true t is the observed 2.5* , the new t is $2.5 + Z$. It falls below 2 when $Z < -0.5$: probability $\Phi(-0.5) = 31\%$. Even with the edge exactly as large as measured, one rerun in three falls short of the bar.

2. *If the result was the best of several ideas* , the true t is likely lower than the observed one (the winner's curse, Section 9). With a true t of 1.5, the new t falls below 2 when $Z < 0.5$: probability $\Phi(0.5) = 69\%$.

3. *With a shorter new history* the true t shrinks with the square root of its length. A strategy with a true t of 2.5 over 16 years has a true t of $2.5 \times \sqrt{1/16} = 0.63$ over one year; its first live year falls below $t = 2$ with probability $\Phi(2 - 0.63) = \Phi(1.37) = 91\%$, and shows a loss with probability $\Phi(-0.63) = 26\%$.

A good part of what traders experience as "it stopped working after I found it" is this scatter, before any change in the market.

A better way to report: the interval. Since the observed effect scatters by one standard error, a range of two standard errors on either side of it covers the true value in about 95% of histories. This *confidence interval* carries the same information as the t-value, but in trading

units. The example earlier in this section had a Sharpe ratio of 0.46 over ten years, with a standard error of about $\sqrt{(1 + 0.46^2/2)/10} = 0.33$. Its interval is $0.46 \pm 1.96 \times 0.33$, from about -0.2 to 1.1. Written that way, it is plain that the data allow anything from a losing strategy to an excellent one, which is exactly what $t = 1.46$ says, less readably. An interval that excludes zero corresponds to $|t| > 1.96$; how far it stays from zero, and how wide it is, show at a glance how much the backtest has really pinned down.

For a trader this is the most useful form of the result: a range for the true Sharpe ratio. Table 4 gives it for three observed Sharpe ratios and three lengths of backtest, computed as observed Sharpe $\pm 1.96 \times \sqrt{(1 + S^2/2)/Y}$. An observed Sharpe ratio of 1.0 over ten years, which would look excellent, pins the true value down only to between about 0.2 and 1.8. Narrowing the range to ± 0.25 takes about 70 years of data for a Sharpe ratio of 0.5 and about 90 years for a Sharpe ratio of 1.0.

Table 4: 95% interval for the true Sharpe ratio, given the observed Sharpe ratio and the length of the backtest: observed $\pm 1.96\sqrt{(1 + S^2/2)/Y}$. Valid for one strategy tested once on independent daily returns; for the best of several strategies the honest range lies lower.

Observed Sharpe	5 years	10 years	20 years
0.5	-0.4 to 1.4	-0.2 to 1.2	0.0 to 1.0
1.0	-0.1 to 2.1	0.2 to 1.8	0.5 to 1.5
1.5	0.2 to 2.8	0.6 to 2.4	0.9 to 2.1

Three caveats belong with every such range. First, strictly speaking the 95% describes the procedure: ranges built this way capture the true value in 95% of histories. Reading it as "95% probability that the true Sharpe ratio lies in this range" is a fair working interpretation for one idea tested once, but not otherwise. Second, if the strategy was the best of several, its observed Sharpe ratio is inflated (the winner's curse, Section 9), and the honest range lies lower than the formula says. Third, the range assumes a correct standard error and a market that stays the same; a strategy whose edge decays is not described by any range computed from its past.

In context. The rule "$t > 2$, so we are good" therefore needs four qualifications, and the rest of this guide is about them. The standard error must be right, which requires independent observations or a correction (Sections 4 and 5); otherwise the observed t scatters by much more than 1, for example by $\sqrt{10} \approx 3.2$ for overlapping ten-day trades, and the usual reading collapses. The benchmark must be right (Section 3). The t-value must not be the best of many (Section 8). And even when all three hold, $t > 2$ says only that a strategy without an edge would rarely look this good; it does not say that the edge is as large as measured, and a rerun will often come in lower.

3 Is it just the market's drift? Excess returns

What it is for. Most markets drift. Equity indices rose by several percent a year on average, and any strategy that is long some of the time collects part of that rise. Testing the strategy's return against zero then tests the market, not the strategy. The fix is to measure the strategy against what the same exposure would have earned without any timing: its *excess* return.

How it works. The simplest version subtracts, for every day, the position times the market's average return over the sample: excess = P&L $-$ position $\times \bar{r}_{\text{market}}$. A long position held 38% of the time is charged 38% of the market's drift. A random-entry benchmark does the same by simulation: enter and exit at random with the same time in the market and the same holding periods, many times, and compare the strategy with that distribution. For stocks, the

papers regress returns on the standard factors (market, size, value, profitability, investment and momentum; Fama and French, 2015, Carhart, 1997); the intercept, or *alpha*, is the return not explained by exposure to those factors.

Example. We simulate a market that rises 5 basis points a day with a daily standard deviation of 1.5% (a Sharpe ratio of $(5/150) \times 15.87 = 0.53$, roughly like the Nasdaq-100 since 1999) for 27 years. A strategy goes long at random, stays in for about five days at a time, and is in the market 38% of the time in total. It has no skill whatsoever. What t-value should we expect?

Calculation.

1. *Average daily P&L.* On 38% of days the strategy earns the market's 5 bp, on the rest nothing: $0.38 \times 5 = 1.9$ bp a day, or $1.9 \times 252 \approx 4.8\%$ a year.

2. *Daily standard deviation.* On days in the market the P&L varies like the market (1.5%), on other days it is zero. The standard deviation over all days is about $\sqrt{0.38} \times 1.5\% = 0.62 \times 1.5\% = 0.92\%$, or $0.92 \times 100 = 92$ bp.

3. *Observations.* $n = 27 \times 252 = 6,804$ days, $\sqrt{n} = 82.5$.

4. *Standard error and t.* $92/82.5 = 1.12$ bp, so $t = 1.9/1.12 = 1.69$.

5. *How often above 2?* Step 4 gives the t-value we expect. In any one 27-year history the market and the random entries come out differently, so the actual t is 1.69 plus a standard normal variable (see "Why the t-value scatters by about 1" in Section 2). It exceeds 2 when that variable is larger than $2 - 1.69 = 0.31$. From the normal table, $P(Z > 0.31) = 1 - \Phi(0.31) \approx 38\%$: in about 38% of histories, a strategy without any skill shows $t > 2$.

6. *Excess.* Charging every day in the market the market's average of 5 bp removes the 1.9 bp a day completely. The excess t then scatters around zero, and exceeds 2 in about 2.3% of samples (Table 2).

When we simulate 2,000 such 27-year histories, each with its own random entries, the median t is 1.7, the strategy earns about 4.7% a year, t exceeds 2 in 38% of the histories, and the excess t exceeds 2 in 2.2%.

Reading the number. Statements such as "the strategy returned 12% a year with only 38% time in the market" combine skill and drift. Dividing the return by the time in the market, as some newsletters do, makes it worse: it credits the strategy with drift it never had to earn. A long-only rule should be compared with a random long-only rule; a long-short rule with the drift collected by its net position.

When it misleads. Subtracting the sample's average drift is itself an estimate. A strategy that is long only when a trend filter is on (for example, price above its 200-day average) selects days with higher-than-average drift; charging it the unconditional average overstates its excess, and the fairer benchmark is the drift on days the filter allowed. A factor regression is only as good as its factors; a strategy loaded on a factor that is not in the model shows that factor's return as alpha.

4 Are the days really independent? Newey–West

What it is for. The plain standard error assumes that each observation brings new information. When a strategy holds trades for several days and opens a new trade every day, consecutive trades share most of their days. Ten overlapping five-day trades are not ten independent results; they are closer to two. The plain t-value counts them as ten and overstates the evidence. Newey–West standard errors [Newey and West, 1987] correct for this.

How it works. The variance of an average of correlated numbers includes the covariances between them. Newey and West estimate these covariances from the data, up to a chosen

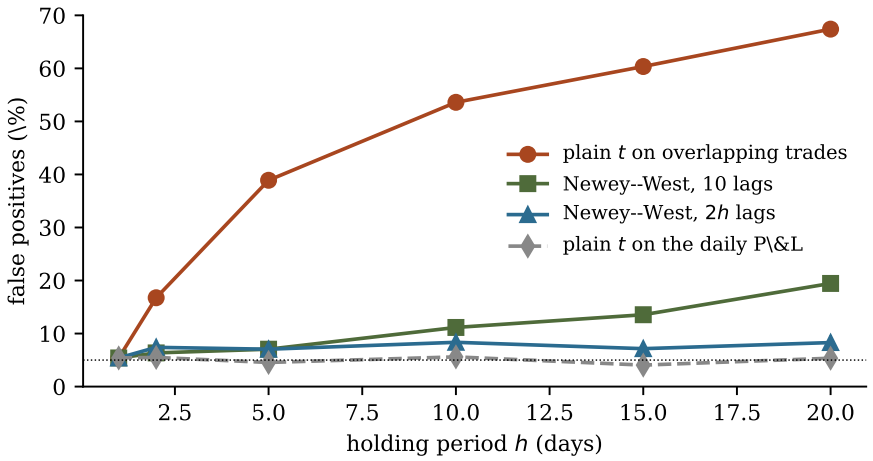


Figure 2: False-positive rate of a two-sided 5% test for a strategy with no edge that opens a new trade every day and holds it for h days, over ten years of simulated driftless daily returns (2,000 samples per point). Dotted line: the 5% the test promises.

number of lags L, give them declining weights $(1 - k/(L+1))$ for lag k, and add them to the plain variance. The result is a standard error that is valid whether or not the observations are correlated, provided the correlation dies out within roughly L observations. The papers use $L = 10$ trading days for daily series. Overlapping trades are the most common reason for correlation over time, but not the only one: event tests whose signal can fire on consecutive days ("the return over the next five days after each signal") overlap in the same way, daily returns of some markets are slightly autocorrelated (stale or illiquid prices), and trade results can cluster in regimes, winning in streaks in one period and losing in streaks in another. A cheap safeguard is to compute both t-values: if the Newey–West t is close to the plain one, nothing was hiding.

Example. We simulate ten years of daily returns with no edge and a strategy that opens a long trade every day and holds it for h days. Figure 2 shows how often a two-sided test at 5% (that is, $|t| > 1.96$) declares the strategy significant. The numbers can be worked out by hand.

Calculation.

1. *How much the plain t is inflated.* Two h-day trades that start j days apart share $h - j$ of their days, so their correlation is $(h - j)/h$. Adding up these correlations, the true variance of the average trade is h times what the plain formula assumes. The plain standard error is too small by a factor $\sqrt{h}$, and the plain t too large by the same factor: $\sqrt{5} = 2.24$ for five-day trades, $\sqrt{10} = 3.16$ for ten-day trades.

2. *False-positive rate of the plain t.* Without an edge the inflated t behaves like $\sqrt{h}$ times a standard normal variable, so $|t| > 1.96$ happens when that variable exceeds $1.96/\sqrt{h}$ in either direction. For $h = 5$: $1.96/2.24 = 0.88$, and $P(|Z| > 0.88) = 38\%$. For $h = 10$: $1.96/3.16 = 0.62$, giving 54%.

3. *What Newey–West with 10 lags recovers.* It adds the correlation at lags 1 to 10 with weights $1 - k/11$: 0.91 at lag 1, 0.55 at lag 5, 0.09 at lag 10. For ten-day trades the true factor is 10; the weighted sum recovers 7.0 of it. The t is then still too large by $\sqrt{10/7.0} = 1.20$, and the false-positive rate is $P(|Z| > 1.96/1.20) \approx 10\%$. For twenty-day trades it recovers only 9.0 of a factor of 20, which leaves about 19%.

When we simulate 2,000 ten-year histories for each holding period, the rates in Figure 2 match the hand calculation: 39% and 54% for the plain t, 11% and 19% with Newey–West at 10 lags, and between 7% and 8% with twice the holding period as lags. The grey line shows the other remedy: measure the same book as a daily P&L series, in which each day is counted once. Its plain t is correct (6% for ten-day trades) because the daily returns of a position decided in advance are not correlated with each other.

Reading the number. A Newey–West t-value reads like a plain t-value. When it is much smaller than the plain t on the same data, the observations overlap or are correlated, and the plain t was overstating the evidence. When the two are close, overlap was not a problem.

When it misleads. Too few lags: the lag length must cover the holding period, or the overlap leaks through, as the twenty-day case shows. Ten lags is a convention for daily series with horizons of up to about two weeks, not a law; automatic rules exist [Andrews, 1991, Newey and West, 1994]. Too little data: Newey–West needs many observations relative to the lag length; with one year of data and five-day trades it rejects in 9% of samples instead of 5% (7% with three years). And it corrects only for correlation over time, not across markets, which is the next section. When the series is already measured daily and the strategy does not overlap with itself, Newey–West changes little, and that is a useful check rather than a waste.

5 Are the markets really independent? Clustering by date

What it is for. Many papers pool results from twenty or thirty markets. On a day when equity markets fall together, a rule that buys weakness triggers in all equity index futures at once, and the results of those trades move together. Counting them as independent overstates the evidence in the same way as overlapping trades do, but across markets instead of over time. Clustered standard errors [Petersen, 2009, Cameron et al., 2011] correct for this.

How it works. Group the observations by date. Instead of treating each trade as one observation, sum the trades of each day into one daily result and compute the uncertainty from the variation of those daily sums. Trades on the same day may then be correlated in any way; only different days are assumed independent. The papers cluster by date, and sometimes by date and market together. An equivalent route, used for portfolios, is to average the markets' results into one daily series and test that series with Newey–West.

Example. We simulate 20 markets whose daily results have a correlation of 0.6 and no edge, over five years; each market trades on a quarter of the days.

Calculation.

1. *Effective number of markets.* The variance of the average of N results with correlation ρ is $(1 + (N - 1)\rho)/N$ times the variance of one result. If all 20 markets traded every day, that is $(1+19\times0.6)/20 = 12.4/20$: the 20 markets carry the information of $20/12.4 = 1.6$ independent ones.

2. *Inflation of the plain t.* The plain t assumes 20 independent results per day and is therefore too large by $\sqrt{12.4} = 3.5$.

3. *With a quarter of the markets trading.* On an average day about 5 markets trade. The inflation factor is then $\sqrt{1 + 4 \times 0.6} = \sqrt{3.4} = 1.84$, and a test at $|t| > 1.96$ rejects when a standard normal variable exceeds $1.96/1.84 = 1.06$: $P(|Z| > 1.06) \approx 29\%$.

When we simulate 2,000 such five-year histories, the plain pooled t is significant in 32% of them (a little more than the 29% of the hand calculation, because the number of trades per day varies). Clustering by date gives 6%, close to the promised 5%.

Reading the number. A clustered t-value reads like a plain one. A large gap between the two says that the results depend on a few common days, which is also economically useful to know: the strategy's risk is concentrated.

When it misleads. Clustering needs many clusters; with fewer than about thirty dates (or markets, when clustering by market) the clustered standard error is itself unreliable. Clustering on the wrong dimension does not help: trades that overlap over several days in one market need

clustering by market or Newey–West, not by date. And clustering corrects the uncertainty, not the average: if a few crash days dominate the average, the estimate itself is fragile.

6 Is strategy A better than strategy B? The Ledoit–Wolf test

What it is for. Two versions of a strategy (executed at the close or at the next close, with or without a filter) usually have similar Sharpe ratios, and both may be significant on their own. Whether one is better than the other is a different question, and comparing their separate t-values does not answer it. The Ledoit–Wolf test [Ledoit and Wolf, 2008] tests the difference of two Sharpe ratios measured on the same days.

How it works. The test computes the difference of the two Sharpe ratios and its standard error, taking into account how strongly the two return series move together and how they are correlated over time. The standard error comes from the delta method applied to the means and second moments of both series, with a Newey–West estimate of their joint variance (10 lags in the papers). Ledoit and Wolf [2008] also propose a bootstrap version; the papers use the Newey–West version, which is close to it in large samples. The older test of Jobson and Korkie [1981] with the correction of Memmel [2003] assumes independent normal returns and is not robust to fat tails or autocorrelation.

Example. We simulate twenty years of two strategies with Sharpe ratios of 0.50 and 0.60 whose daily returns have a correlation of 0.9.

Calculation.

1. *Each on its own.* $t \approx S\sqrt{Y}$: $0.5 \times \sqrt{20} = 2.2$ and $0.6 \times \sqrt{20} = 2.7$. Both are significant.

2. *Standard error of the difference.* For two Sharpe ratios on the same days with correlation ρ, the standard error of their difference is about $\sqrt{2(1-\rho)/Y}$. With $\rho = 0.9$ and $Y = 20$: $\sqrt{2 \times 0.1/20} = 0.10$. If the strategies were independent ($\rho = 0$) it would be $\sqrt{2/20} = 0.32$.

3. *t of the difference.* The observed difference is $0.6 - 0.5 = 0.1$, so $t = 0.1/0.10 = 1.0$. What does that mean? Suppose the two strategies were in truth equally good. The measured difference would then still wobble from history to history, with a standard deviation of 0.10, and a difference of at least 0.1 in either direction would occur with probability $2(1 - \Phi(1.0)) = 0.32$ (Table 2). One history in three would show a gap this large by chance alone, so this history does not show that B is better. Not significant.

4. *Power: how often the test would notice.* Here B really is better by 0.1, by construction. Across histories, the t of the difference scatters around its expected value of $0.1/0.10 = 1.0$ by about 1 (Section 2). The test declares B better only when $t > 1.96$, which needs the scatter to add more than $1.96 - 1.0 = 0.96$: $P(Z > 0.96) = 1 - \Phi(0.96) = 17\%$. In the other 83% of histories, the test misses a difference that is real. The *power* of a test is this probability of detecting a real effect; here it is low.

5. *Years needed.* For $t = 2$ the standard error must fall to $0.1/2 = 0.05$, so $2(1-\rho)/Y = 0.05^2$, or $Y = 2 \times 0.1/0.0025 = 80$ years. With $\rho = 0.5$: $Y = 2 \times 0.5/0.0025 = 400$ years.

One simulated twenty-year history gives individual t-values of 2.2 and 2.7, a Ledoit–Wolf t of 1.0 for the difference of 0.10, and when we simulate 1,000 such histories, the test detects the true difference in 17% of them. Pairing makes the test sharper (a standard error of 0.10 instead of 0.32), but a small difference between similar strategies still takes generations to establish.

Reading the number. The test answers one question only: which of the two versions is better? It does not say whether either of them is worth having; each strategy's own t-value answers that (Section 2).

- *A large t (above about 2)*: one version is better than the other on these days. The sign says which. In our paper on trend following [Claude, 2026i], trading at the close instead of the next close improved the Sharpe ratio with $t = 2.4$ (an upper bound, because that signal uses the closing price itself): executing earlier is better.

- *A small t*: the data cannot rank the two. That is not the same as "equally good": as the example shows, a real difference can easily give a small t. In the same paper, computing the signal fifteen minutes before the close instead of at the close gave $t = -0.3$: no measurable loss from the earlier signal. When the data cannot rank two versions, choose on other grounds, such as cost, simplicity, or the version fixed in advance.

- *Correlation helps the test.* The more similar the two versions, the smaller the standard error of their difference, $\sqrt{2(1-\rho)/Y}$. Between highly correlated versions even a small difference can be significant, and is then convincing.

Whether to *hold both* is yet another question, answered by their correlation rather than by this test. Two strategies with Sharpe ratios of 0.5 and 0.6, equal volatility and a correlation of 0.9, held half and half, have a Sharpe ratio of $\frac{(0.5+0.6)/2}{\sqrt{(1+0.9)/2}} = 0.56$: less than the better one alone. Combining pays only when the correlation is low.

When it misleads. The test requires the same days for both strategies; comparing Sharpe ratios from different periods or markets is not what it measures. It says nothing about how the two versions were chosen; if one of them was picked from many variants, the multiple-testing problem of Section 8 applies. And the Sharpe ratio is not the only criterion: two strategies with equal Sharpe ratios can have very different drawdowns or capacity.

7 What does chance look like here? Simulated null models

What it is for. Some questions have no natural zero. "Do prices bounce off trend lines?" cannot be answered by testing the bounce rate against zero, or against 50%, because a price moving at random also turns near any line drawn close to it. The right comparison is the same measurement on data in which the effect cannot exist but everything else is the same. Constructing that data is the job of a null model.

How it works. The papers use four kinds.

- *Bootstrap paths* [Efron, 1979, Brock et al., 1992]: build artificial price series by drawing real days or blocks of days at random from the market's history. The stationary block bootstrap [Politis and Romano, 1994] draws blocks of random length, which keeps volatility clustering and short-range dependence but destroys any structure tied to earlier price levels. Lines, levels and rules are then applied to these paths exactly as to the real one.

- *Sign flips*: keep the size of every price change but give it a random sign, which removes any drift and any tendency of moves to continue or reverse.

- *Placebo levels*: shift the real lines or levels by a random amount on the real price path, so that the levels have the same shape but are at prices nobody watched.

- *Mirror levels*: compare the level of interest with a level at the same distance on the other side of the starting point, on the same price path, as in the gap paper [Claude, 2026j].

Example. How often does a price that moves at random touch a level near its starting point? We simulate a driftless random walk of 390 one-minute steps (a trading day) and scale it so that its average daily range is 1, the unit the papers call ATR.

Calculation.

1. *The formula.* For a driftless random walk with standard deviation $\sigma\sqrt{T}$ over the day, the probability of touching a level at distance a is $2\,(1 - \Phi(a/\sigma\sqrt{T}))$ (the reflection principle).

2. *Convert the range.* The average range of such a walk is $2\sqrt{2/\pi}\,\sigma\sqrt{T} = 1.596\,\sigma\sqrt{T}$. A range of 1 therefore means $\sigma\sqrt{T} = 1/1.596 = 0.627$.

3. *A level 0.1 ATR away.* $a/\sigma\sqrt{T} = 0.1/0.627 = 0.16$, $\Phi(0.16) = 0.563$, so the probability is $2 \times (1 - 0.563) = 87\%$.

The simulated walk touches a level 0.1 ATR above its start on 86% of days (slightly below the formula, because one-minute steps miss touches between the steps), 0.25 away on 68%, half a range away on 43%, and one full range away on 12%. Below the start the figures are the same: 85%, 68%, 42% and 12%. A "fill rate" of 85% for small gaps is exactly what chance delivers; only the difference with the mirror level carries information.

Discreteness adds a second trap. If the same prices are rounded to ticks of 0.05 ATR, a level exactly on a tick, 0.2 ATR away, is touched on 76% of days, but a level a hair above it, which requires the next tick, only on 71%; the unrounded walk touches both on 74%. Comparing levels on and off the tick grid then finds an "effect" that is pure arithmetic. In the round-numbers paper this kind of mismatch produced z-scores near 20 until the null was matched to the tick grid [Claude, 2026c].

Doing it yourself, without the formulas

This section is the most technical in the guide, but the idea can be applied without any of its formulas. (1) Next to the level you care about, pick a level that nobody watches: the same distance on the other side of the starting point, or the real level shifted up or down by a random amount. (2) Run exactly the same measurement on both, on the same price data: the touch rate, the bounce rate, the profit of the trade. (3) Read only the difference between the two. A high rate on its own, such as 85% of gaps filling, says nothing; a clear gap between the real level and the control, in many markets and periods, is the evidence.

Reading the number. The useful number is the difference between the real data and the null, and its t-value across many null paths or many levels. A result that is large against zero but small against the null is chance, however impressive it looks.

When it misleads. A null model is only as good as what it preserves. It must keep everything except the effect under test: volatility, its clustering, the tick size, the trading hours, the drift if drift is not the question. A null that destroys too much (for example, independent normal returns instead of real days) makes ordinary features of markets look like discoveries. A null that keeps too much (for example, placebo levels shifted so little that they share the real level's position near recent prices) hides real effects. The papers therefore often use two nulls that err in opposite directions and report both.

8 How many times did we look? Multiple testing

What it is for. A 5% significance level promises that a strategy without an edge passes the test only 5% of the time. Test twenty strategies without an edge, and one of them is expected to pass. Test a hundred parameter combinations and report the best, and the best will look good regardless. Multiple-testing corrections adjust the threshold for the number of tests. Which correction is right depends on what one wants to control.

How it works. There are three families of corrections.

- *Family-wise error rate (FWER)*: the probability of even one false discovery. Bonferroni divides the significance level by the number of tests m: with $m = 20$, each test needs $p < 0.0025$ ($t > 2.8$ one-sided). Holm's procedure [Holm, 1979] is a uniformly better version: sort the p-values, compare the smallest with $0.05/m$, the next with $0.05/(m-1)$, and so on, stopping at the first that fails.

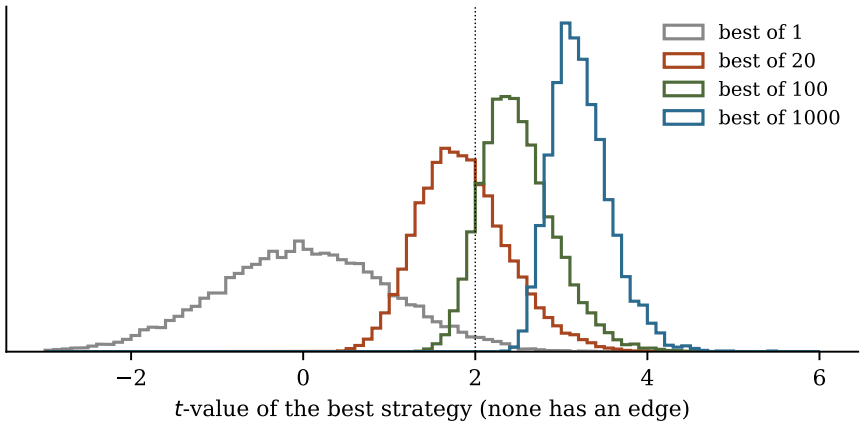


Figure 3: The t-value of the best of N strategies when none of them has an edge (independent strategies, simulated). Dotted line: $t = 2$.

- *False discovery rate (FDR)*: the expected share of false discoveries among all discoveries. Benjamini–Hochberg [Benjamini and Hochberg, 1995] sorts the m p-values and finds the largest k for which the k-th smallest is at most $q\,k/m$; the k smallest are declared discoveries. With $q = 0.10$, at most about 10% of the list is expected to be false. The procedure is valid for independent tests and for positively correlated ones, which covers most strategies on related markets. Benjamini–Yekutieli [Benjamini and Yekutieli, 2001] is valid under any dependence, at the cost of a stricter threshold.
- *The best of many*: when a single rule is chosen as the best of a grid, the question is whether the best is better than what the best of that many rules would be by chance. White's reality check [White, 2000, Sullivan et al., 1999] and Hansen's test for superior predictive ability [Hansen, 2005] answer this by bootstrapping the whole grid at once, which takes the correlation between the rules into account. Harvey, Liu and Zhu [Harvey et al., 2016] argue, from the number of factors tested in finance, that a new factor should have $t > 3$ rather than 2; the deflated Sharpe ratio [Bailey and López de Prado, 2014] adjusts a Sharpe ratio for the number of trials behind it.

Example. *The best of N.* Suppose we backtest 20 strategies, none of which has an edge, and look at the best one. Two questions: how often does the best of them reach $t > 2$, the usual bar for "significant"? And how high must the bar be so that the best of them clears it only 5% of the time, as the bar promises for one strategy? Figure 3 shows the answer by simulation; the calculation goes as follows.

Calculation.

1. *One strategy.* Without an edge, a single strategy reaches $t > 2$ with probability $1 - \Phi(2) = 0.0228$ (Table 2), about 2%.

2. *The best of 20 against the bar of 2.* The best of 20 is above 2 as soon as at least one of the 20 is above 2. That probability is easiest to find through its opposite, "none of the 20 is above 2". Each strategy is not above 2 with probability $1 - 0.0228 = 0.9772$; for 20 independent strategies, the chance that this holds for all of them is the product, $0.9772^{20} = 0.631$. So at least one, and therefore the best, is above 2 with probability $1 - 0.631 = 37\%$. For 100 strategies: $1 - 0.9772^{100} = 1 - 0.100 = 90\%$. A bar that a single worthless strategy clears 2% of the time is cleared by the best of 20 more than a third of the time.

3. *A bar for the best of 20.* To bring the chance back to 5% for the best of 20, each single strategy may clear the bar only with probability $0.05/20 = 0.0025$: twenty chances of 0.25% add up to about 5%. A probability of 0.0025 corresponds to $t > 2.8$. That is the Bonferroni bar. For 100 strategies it is $p < 0.0005$, or $t > 3.3$; for 1,000, $t > 3.9$.

Table 5: A worked example: ten sorted p-values ($m = 10$) and the threshold of each procedure. A tick means that the p-value is at or below that row's threshold. For Bonferroni, Holm and BY the ticks are the discoveries. For BH they are not: BH takes the largest rank with a tick (here rank 5) and declares every p-value up to that rank a discovery, which gives the last column. That BH and "no correction" find the same five here is a coincidence of this example (see text).

Rank k	p	None 0.05	Bonferroni 0.005	Holm $0.05/(m-k+1)$	BH test $0.10\,k/m$	BY $0.10\,k/(2.93\,m)$	BH result
1	0.001	✓	✓	0.0050 ✓	0.01 ✓	0.0034 ✓	✓
2	0.004	✓	✓	0.0056 ✓	0.02 ✓	0.0068 ✓	✓
3	0.019	✓		0.0063 (stop)	0.03 ✓	0.0102	✓
4	0.045	✓		0.0071	0.04	0.0137	✓
5	0.048	✓		0.0083	0.05 ✓	0.0171	✓
6	0.081			0.0100	0.06	0.0205	
7	0.120			0.0125	0.07	0.0239	
8	0.250			0.0167	0.08	0.0273	
9	0.470			0.0250	0.09	0.0307	
10	0.810			0.0500	0.10	0.0341	

When we simulate this many times, the best of 20 has a median t of 1.8 and exceeds 2 in 37% of cases; the best of 100 has a median of 2.5 and exceeds 2 in 90%.

Ten p-values, four procedures. Table 5 works through one family of ten tests step by step. Sort the p-values from small to large and give each its rank k.

- *No correction*: every $p < 0.05$ counts: five discoveries.
- *Bonferroni*: every $p < 0.05/10 = 0.005$ counts: two.
- *Holm*: compare the smallest with $0.05/10$, the next with $0.05/9$, the third with $0.05/8$, and stop at the first failure. Ranks 1 and 2 pass; rank 3 ($0.019 > 0.00625$) fails, so two discoveries.
- *Benjamini–Hochberg* ($q = 0.10$): compare each p with $q\,k/m = 0.01k$ and find the *largest* rank that passes. Rank 5 passes ($0.048 \le 0.05$), rank 6 and higher fail. All five smallest are discoveries, including rank 4, which failed its own threshold: the procedure steps up from the largest passing rank.
- *Benjamini–Yekutieli*: as BH, with thresholds divided by $1 + \frac{1}{2} + \cdots + \frac{1}{10} = 2.93$: $0.0034k$. Ranks 1 and 2 pass, so two discoveries.

Here Benjamini–Hochberg finds as many as no correction, because five of the ten p-values are small. With fewer small p-values the two part ways: if only 0.004, 0.035 and 0.045 were below 0.05 and the rest large, no correction would count three discoveries, while BH would compare them with 0.01, 0.02 and 0.03, find that only rank 1 passes, and count one.

A hundred strategies, ten real. Table 6 compares the procedures on 100 strategies of which 10 have a real edge (an expected t of 3) and the rest none.

Calculation.

1. *No correction.* A one-sided 5% test needs $t > 1.645$. Each of the 90 worthless strategies passes with probability 0.05: $90 \times 0.05 = 4.5$ false discoveries. A real one, whose t scatters around 3, passes with probability $1 - \Phi(1.645 - 3) = 0.912$: 9.1 real discoveries. The false share is $4.5/(4.5 + 9.1) = 33\%$.

2. *Bonferroni.* With 100 tests, each needs $p < 0.05/100 = 0.0005$, which corresponds to $t > 3.3$ (Step 3 of the previous calculation). *False discoveries*: each of the 90 worthless strategies now passes with probability 0.0005, so on average $90 \times 0.0005 = 0.05$ of them pass: almost none. *Real discoveries*: a real strategy's

Table 6: 100 simulated strategies, 10 with a real edge (expected $t = 3$), one-sided tests; averages over 2,000 simulations.

Procedure	Real edges found	False discoveries	False share
None ($p < 0.05$)	9.1	4.4	33%
Bonferroni (FWER 5%)	3.8	0.0	1%
Holm (FWER 5%)	3.9	0.0	1%
Benjamini–Hochberg ($q = 0.10$)	7.2	0.8	10%
Benjamini–Yekutieli ($q = 0.10$)	4.6	0.1	2%

t scatters around 3 by about 1 (Section 2), so it passes when the scatter adds more than $3.3 - 3 = 0.3$, which happens with probability $1 - \Phi(0.3) = 0.386$. Of the ten real strategies, $10 \times 0.386 = 3.9$ are found on average. The strict threshold removes the false ones, but loses six of the ten real ones as well.

Repeating this 2,000 times gives the averages in Table 6, which match the hand calculation. Benjamini–Hochberg sits in between: it finds 7.2 of the ten real edges with 0.8 false ones, a false share of 10%, close to the promised 10%.

The best of a grid. We simulate 50 moving-average crossover rules with random lengths on one driftless market for ten years, and repeat until the best rule looks good, as a researcher would notice it; this took 10 attempts. The reality check then works as follows.

Calculation.

1. *The observed best.* The best rule has $t = 2.56$; on its own, a one-sided p of $1 - \Phi(2.56) = 0.005$.

2. *Bonferroni.* $50 \times 0.005 = 0.26$: not significant, but Bonferroni treats the 50 rules as independent chances, while moving-average rules on the same market are strongly correlated.

3. *Resample.* Draw a new history of the same length from the real one in blocks of random length (mean 10 days), so that the 50 rules keep their correlation but any true edge is removed by centring each rule on its own average. Compute the best t among the 50 in that history. Repeat 1,000 times.

4. *Compare.* The p-value is the share of resampled histories whose best t is at least 2.56: 0.07.

The reality check is less severe than Bonferroni, because 50 similar rules are fewer than 50 independent chances, but it still says that a best-of-50 result like this is what chance produces in about one history in $1/0.07 \approx 14$.

Reading the number. Bonferroni and Holm answer "is any of these real?"; they suit a small number of tests where a single false claim is costly, such as a frozen trading rule. Benjamini–Hochberg answers "which of these are worth a closer look?"; it suits per-market results in a table, where a few false entries are acceptable if their share is controlled. The reality check answers "is the best rule of this grid better than chance?"; it suits parameter searches. A per-market table marked with Benjamini–Hochberg should be read as: the marked rows are probably real as a group, not each one with certainty.

When it misleads. The correction must count every test that was run, not only the tests that are reported. This is the core problem with published strategies. A newsletter that tested five hundred ideas and mails out the twenty-one best presents results that have been selected for small p-values; applying Benjamini–Hochberg to those twenty-one passes almost all of them, and there is no way to repair this afterwards without knowing the five hundred. The family must also be defined before the test: deciding afterwards which results belong together is another way of choosing. Finally, corrections lose power. With hundreds of tests, real but modest effects are lost; that is the price of not being fooled, and it is why pooled tests (one test on all markets together) are often preferred over many separate tests.

What if I have already tested millions of strategies in my life? Do those still count? Taken literally, Bonferroni over a million tests asks for $p < 0.05/1{,}000{,}000$, or $t > 5.3$, and the best of a million worthless independent strategies has a median t of about 4.8. Few real trading edges reach such values in a backtest. Three things make the requirement less hopeless than it sounds. First, a million strategies are not a million independent chances: most are neighbours in a parameter grid that win and lose on the same days, and a reality check over the grid accounts for that and is far less severe than Bonferroni (as in the 50-rule example above). Second, the count that matters is the number of results you would act on after looking at them; a search that is followed by a fresh test resets the count. Develop on one part of the data, pick a handful of candidates, and test only those, once, on data the search never saw (Section 9). With five candidates in the hold-out, the Bonferroni threshold is $t > 2.3$, not 5.3. Third, a rule with an economic reason, chosen before the search, is not one of the million. What resets the count is fresh data, not a new search: old searches count against any result tested on the same data they looked at, also when they inspired the new idea, but not against a test on data none of them saw, such as the period since the last look, a market never tested, or live trading. After years of searching the same history, such untouched data is the scarcest resource a researcher has, which is why post-publication periods, new markets and forward trading are so valuable. The honest conclusion is not that large searches are useless, but that their backtest t-values are not evidence on their own; the hold-out is.

9 Would it survive new data? Out-of-sample designs

What it is for. Every correction above relies on knowing how many choices were made. Out-of-sample testing does not: data that did not exist, or was not looked at, when the rule was chosen cannot have influenced the choice. It is the strongest protection the papers use.

How it works. There are four forms.

- *Sub-periods*: report results separately for, say, each decade. An effect concentrated in the early years may have disappeared, and the full-sample average hides that.
- *Post-publication data*: for a published strategy, the data after publication is a true hold-out [McLean and Pontiff, 2016]. The papers on intraday momentum and the noise area are built on this [Claude, 2026d,e].
- *Hold-out by time and by market*: develop the rule on part of the data, freeze it, and test it once on later data or on other markets that were set aside.
- *Pre-registration*: write down the rule, the test, the hold-out data and the decision criterion, with a date, before computing any result, as in the gap paper [Claude, 2026j]. Any later change is an amendment, dated and reported as such.

Example. *The winner's curse.* We simulate 50 trading ideas whose true Sharpe ratios are spread around zero with a standard deviation of 0.2, observe each for 10 years, and pick the best.

Calculation.

1. *Measurement noise.* Ten years measure a Sharpe ratio with a standard error of $1/\sqrt{10} = 0.316$ (Section 2).

2. *Spread of the backtests.* Each backtest is true value plus noise, so the 50 backtest Sharpe ratios spread with a standard deviation of $\sqrt{0.2^2 + 0.316^2} = 0.374$.

3. *The best backtest.* The largest of 50 standard normal variables is on average 2.25, so the best backtest shows about $2.25 \times 0.374 = 0.84$.

4. *How much of it is real.* Of the spread, a share $0.2^2/(0.2^2 + 0.316^2) = 0.04/0.14 = 0.29$ is true variation and the rest noise. On average the winner's true Sharpe ratio is therefore $0.29 \times 0.84 = 0.24$, and that is what it delivers out of sample.

Repeating this selection 2,000 times gives an average in-sample Sharpe ratio of 0.85 for the winner, a true value of 0.24, and 0.24 over the next 5 years: about 71% less than the backtest promised, and negative in 30% of cases (the five-year out-of-sample estimate has its own standard error of $1/\sqrt{5} = 0.45$). Nothing broke; the backtest selected the luckiest idea, and the luck did not repeat.

How long a real edge can look dead. Over a period of h years, a strategy with Sharpe ratio S has an expected return of $S\sqrt{h}$ standard deviations, so it loses money with probability $\Phi(-S\sqrt{h})$.

Calculation.

1. *Sharpe 0.5, three months.* $\Phi(-0.5\sqrt{0.25}) = \Phi(-0.25) = 40\%$.

2. *Sharpe 0.5, one year.* $\Phi(-0.5\sqrt{1}) = \Phi(-0.5) = 31\%$.

3. *Sharpe 0.5, three years.* $\Phi(-0.5\sqrt{3}) = \Phi(-0.87) = 19\%$.

4. *Sharpe 1.0, three months.* $\Phi(-1.0\sqrt{0.25}) = \Phi(-0.5) = 31\%$.

5. *Sharpe 1.0, one year.* $\Phi(-1.0) = 16\%$.

6. *Sharpe 1.0, three years.* $\Phi(-1.0\sqrt{3}) = \Phi(-1.73) = 4\%$.

Appendix A shows how to look up Φ of a negative number.

Over twenty years, the longest stretch without a new equity high for a Sharpe-0.5 strategy is typically 4.7 years in our simulations, and in one sample in ten 9.7 years or more.

Reading the number. A hold-out result is read like any other result, but it carries more weight because it is free of selection. A few months of losses after publication are not evidence that a strategy failed; it is better to place them among the strategy's own past results over equally long stretches. A rule that holds up in a hold-out of several years and several markets has passed the test that matters most.

When it misleads. A hold-out is used once. If the rule is changed after seeing the hold-out result, the hold-out has become development data, and the change must be reported as such. Hold-out samples are usually short, so their tests have little power: a negative hold-out result often means "not shown" rather than "disproved". And markets change; a rule that worked from 1990 to 2010 and fails afterwards may have been real and then arbitraged away, which a hold-out cannot distinguish from a fluke.

10 A checklist for reading a backtest

The questions below summarize the guide. They apply to the papers on newhighs.nl and to any other backtest.

1. *Size and cost.* How large is the effect in ticks, basis points or fractions of the range, and how does it compare with the cost of a round trip? What is the break-even cost?

2. *Amount of data.* How many years, and how many independent observations? Remember that $t \approx S\sqrt{Y}$: a Sharpe ratio from five years is uncertain by about ± 0.45.

3. *Benchmark.* Is the result measured against zero or against what the same exposure would have earned? Is time in the market used to inflate returns?

4. *Overlap.* Do trades overlap in time? Are many markets trading on the same days? Are the t-values Newey–West or clustered?

5. *Null model.* For patterns and levels: is the result compared with the same measurement on data where the pattern cannot exist?

6. *Search.* How many variants, parameters, markets and periods were tried? Is that number stated, and is it corrected for? Was the rule fixed before the test?

7. *Stability.* Does the result hold in each sub-period and in other markets, or is it carried by one decade or one market?

8. *New data.* Is there a hold-out or post-publication period, and what does it show?

None of these questions has a single threshold. A strategy can fail one of them and still be worth trading, and pass all of them and still disappoint. Their purpose is to make clear how much of a backtest is evidence and how much is hope.

A How to look up $\Phi(x)$

$\Phi(x)$ is the probability that a standard normal variable Z (bell-shaped, centred on 0, standard deviation 1) comes out below x. Table 7 gives it for x from 0 to 3.49. Find the row with the first decimal of x and the column with the second decimal. Four rules cover every use in this guide.

1. *Below x*: $P(Z < x) = \Phi(x)$, read directly. Example: $\Phi(0.31)$ is in row 0.3, column 0.01: 0.6217.

2. *Above x*: $P(Z > x) = 1 - \Phi(x)$. Example from Section 3: $P(Z > 0.31) = 1 - 0.6217 = 0.378$, or 38%.

3. *Negative x*: the bell curve is symmetric, so $\Phi(-x) = 1 - \Phi(x)$. Example from Section 9: $\Phi(-0.25) = 1 - \Phi(0.25) = 1 - 0.5987 = 0.401$, or 40%.

4. *Either direction*: $P(|Z| > x) = 2\,(1 - \Phi(x))$. Example from Section 4: $P(|Z| > 0.62) = 2 \times (1 - 0.7324) = 0.535$, or 54%.

The table also works backwards, from a probability to a t-value. Example from Section 8: which t does a single strategy exceed with probability 0.0025? We need $1 - \Phi(x) = 0.0025$, so $\Phi(x) = 0.9975$. That value is in row 2.8, column 0.01: $t = 2.81$. Beyond the table, $\Phi(x)$ is 0.9998 or more; for $x = 4$ the probability above is about 0.00003, and for $x = 5$ about 0.0000003.

References

Donald W. K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858, 1991.

David H. Bailey and Marcos López de Prado. The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40(5):94–107, 2014.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57 (1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

Table 7: $\Phi(x)$: the probability that a standard normal variable is below x. Row: x to one decimal; column: the second decimal. For negative x use $\Phi(-x) = 1 - \Phi(x)$.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

William Brock, Josef Lakonishok, and Blake LeBaron. Simple technical trading rules and the stochastic properties of stock returns. *Journal of Finance*, 47(5):1731–1764, 1992.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249, 2011.

Mark M. Carhart. On persistence in mutual fund performance. *Journal of Finance*, 52(1):57–82, 1997.

Claude. Do price channels hold? A test of trend-channel support and resistance in 28 futures markets at daily and intraday horizons. November 2026. Available at `https://newhighs.nl/download/price-channels.pdf`, 2026a.

Claude. Clean channels, wide bands: Do well-fitted regression channels contain the next bar?

November 2026. Available at https://newhighs.nl/download/regression-channels.pdf, 2026b.

Claude. The tick beyond the round number: Round numbers and the previous session's range in 28 futures markets. November 2026. Available at https://newhighs.nl/download/round-numbers.pdf, 2026c.

Claude. Does the last half hour still follow the day? An out-of-sample test of market intraday momentum in 28 futures markets. November 2026. Available at https://newhighs.nl/download/intraday-momentum.pdf, 2026d.

Claude. Outside the noise area: An out-of-sample test of a published intraday momentum strategy in index futures. November 2026. Available at https://newhighs.nl/download/noise-area.pdf, 2026e.

Claude. Boxes around price: Darvas's method on US stocks, futures and intraday data, 1991–2026. November 2026. Available at https://newhighs.nl/download/darvas-boxes.pdf, 2026f.

Claude. Trend following without the coin: A replication of the coin-flip trend program on 29 futures markets, 1975–2026. November 2026. Available at https://newhighs.nl/download/coin-flips.pdf, 2026g.

Claude. Breaking out of the first half hour: A broad test of opening range breakout rules in 29 futures markets. November 2026. Available at https://newhighs.nl/download/opening-range-breakout.pdf, 2026h.

Claude. The right strategy for the right market? Measuring trend and mean reversion in 38 futures markets and 3,500 US stocks. November 2026. Available at https://newhighs.nl/download/right-strategy.pdf, 2026i.

Claude. Gaps fill, but so does the other side: A pre-registered test of opening gaps in 33 futures markets, 2008–2026. November 2026. Available at https://newhighs.nl/download/gap-fill.pdf, 2026j.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1): 1–26, 1979.

Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015.

Peter Reinhard Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68, 2016.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

J. D. Jobson and Bob M. Korkie. Performance hypothesis testing with the Sharpe and Treynor measures. *Journal of Finance*, 36(4):889–908, 1981.

Olivier Ledoit and Michael Wolf. Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, 15(5):850–859, 2008.

Andrew W. Lo. The statistics of Sharpe ratios. *Financial Analysts Journal*, 58(4):36–52, 2002.

R. David McLean and Jeffrey Pontiff. Does academic research destroy stock return predictability? *Journal of Finance*, 71(1):5–32, 2016.

Christoph Memmel. Performance hypothesis testing with the Sharpe ratio. *Finance Letters*, 1: 21–23, 2003.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987.

Whitney K. Newey and Kenneth D. West. Automatic lag selection in covariance matrix estimation. *Review of Economic Studies*, 61(4):631–653, 1994.

Mitchell A. Petersen. Estimating standard errors in finance panel data sets: Comparing approaches. *Review of Financial Studies*, 22(1):435–480, 2009.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994. doi: 10.1080/01621459.1994.10476870.

Ryan Sullivan, Allan Timmermann, and Halbert White. Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Finance*, 54(5):1647–1691, 1999.

Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.