MPM MARKETS RESEARCH · WORKING PAPER

Testing the “Gaps Always Fill” Myth Across Six Futures Markets

An empirical study of daily gap behaviour across six futures markets, April 2019 – May 2026, with a transparent test of tradability.

Research Dataset: Nasdaq-100 Futures (NQ), S&P 500 Futures (ES), Gold Futures (GC), Silver Futures (SI), Crude Oil Futures (CL), and Copper Futures (HG)

Author: Dhaval Barot · Published by MPM Markets – Market Probability Model · mpmmarkets.com

Abstract

We test the widely-taught belief that price gaps “always fill,” using seven years of one-minute data across six liquid futures markets, and measure not just whether gaps fill but how quickly. We find that fill probability is **strongly and monotonically inverse to gap size** — but the more informative result is about timing. Small gaps fill almost immediately, usually within the same session. Large gaps behave in two phases: they tend to persist in the short term, with only about 30–47% filling on the same session, but they then revert over the following sessions, with roughly three-quarters to four-fifths — about 76–83% — filling within ten trading sessions. In other words, the short-term persistence of large gaps is real but temporary; most ultimately fill, on a longer timescale than their same-session statistics imply. This pattern is consistent across all six instruments and every calendar year, including the 2022 equity bear market. We then ask whether this regularity can be traded. Across eight distinct implementations, with walk-forward validation, permutation testing against random, and realistic costs, **we did not find a statistically reliable trading edge within the implementations tested**. We report the descriptive finding and the negative tradability result together, with full sample sizes and methodology, so that other researchers can verify, extend, or refute this work.

1. Motivation

“Gaps fill” is among the most repeated heuristics in retail trading education. It is intuitive and frequently asserted, yet rarely tested rigorously across instruments, gap sizes, and market regimes simultaneously. We set out to measure the behaviour directly and — separately — to test honestly whether any measured tendency translates into an exploitable edge. We treat those as two different questions, because a statistically real pattern and a profitable strategy are not the same thing.

2. Data and definitions

Instruments. Six CME-group futures spanning three asset classes: Nasdaq-100 (NQ) and S&P 500 (ES) equity indices; gold (GC), silver (SI), copper (HG) metals; crude oil (CL) energy.

Sample. 20 April 2019 through 24 May 2026, one-minute resolution, roughly 2.4–2.5 million bars and 1,800–2,200 trading sessions per instrument. Timezone handling was verified per file. Data source: CME-group futures minute bars, collected and normalized by the author. Methodology available upon request.

Session. To ensure direct comparability across all six futures markets, this study applies a single standardized U.S. trading session (09:00–16:00 ET, i.e. 13:00–20:00 UTC) to ES, NQ, GC, SI, CL

and HG. A gap is defined as the difference between the close of one standardized session and the open of the next, expressed as a fraction of the prior close. A gap is “filled” if price trades back to that prior close at any point within the same session (intraday, same-session only — no future data is used).

> **Session definition**
>
> To ensure direct comparability across all six futures markets, this study applies a single standardized U.S. trading session (09:00–16:00 ET) to ES, NQ, GC, SI, CL and HG. A gap is defined as the difference between the close of one standardized session and the open of the next; gap-fill behaviour is then evaluated within the same session framework. Some markets — gold, silver, crude oil and copper — have traditional floor-session hours that differ from this window. The objective of this study is *not* to replicate each market’s native floor session, but to evaluate gap behaviour using a consistent session definition across all instruments, allowing direct cross-market comparison under a common framework. This study therefore does not examine daily-candle gaps, Sunday/Globex opens, weekend gaps, or equity cash-market gaps; reported percentages are specific to this standardized-session definition.

Significance. Throughout, statistical significance refers to $p < 0.05$ unless stated otherwise. For descriptive findings we use two-proportion tests and permutation tests against a random null; across the full battery of patterns we additionally apply Benjamini–Hochberg false-discovery-rate correction so that scanning many hypotheses does not manufacture apparent winners.

3. Principal finding: larger gaps fill more slowly

Two facts emerge when fills are measured across multiple time horizons rather than within a single session. First, at every horizon, fill probability declines with gap size: small gaps fill more readily than large ones. Second, and more importantly, the difference is largely one of speed. Small gaps fill almost immediately, usually the same session. Large gaps resist same-session filling but fill steadily thereafter — roughly three-quarters to four-fifths within ten sessions (about two trading weeks).

Two phases: short-term persistence, then reversion. The same-session result is not an illusion — large gaps genuinely tend to persist in the short term, filling only 30–47% of the time on the session they occur, and remaining unfilled across the next few sessions more often than small gaps do. But this persistence is temporary, not permanent: as the window extends, fill rates climb steadily, reaching 76–83% by ten sessions. The honest description is therefore a two-phase one — large gaps show short-term persistence (a same-session and few-session tendency not to fill), followed by eventual mean reversion, with the majority filling within about two trading weeks. What appears to be continuation on the day of the gap often proves temporary when viewed over longer horizons.

Gap size (Gold, GC)	Same session	≤3 sess	≤5 sess	≤10 sess
<0.1%	94%	98%	98%	99%
0.1–0.2%	80%	90%	91%	93%
0.2–0.3%	62%	79%	84%	89%
0.3–0.5%	46%	73%	81%	85%
0.5–1.0%	28%	59%	67%	77%
>1.0%	12%	34%	47%	63%

Gap size (Gold, GC)	Same session	≤3 sess	≤5 sess	≤10 sess
Gap size (Crude Oil, CL)	Same session	≤3 sess	≤5 sess	≤10 sess
<0.1%	98%	98%	99%	99%
0.1–0.2%	93%	94%	96%	97%
0.2–0.3%	91%	94%	95%	96%
0.3–0.5%	78%	89%	91%	94%
0.5–1.0%	62%	82%	88%	92%
>1.0%	29%	55%	65%	76%

3.1 Large-gap fill rate by horizon, per instrument

Restricting to large gaps (≥0.3% of prior close), the same-session fill rate is modest everywhere, but rises sharply as the window widens. Sample sizes (N) are the count of large-gap days per instrument.

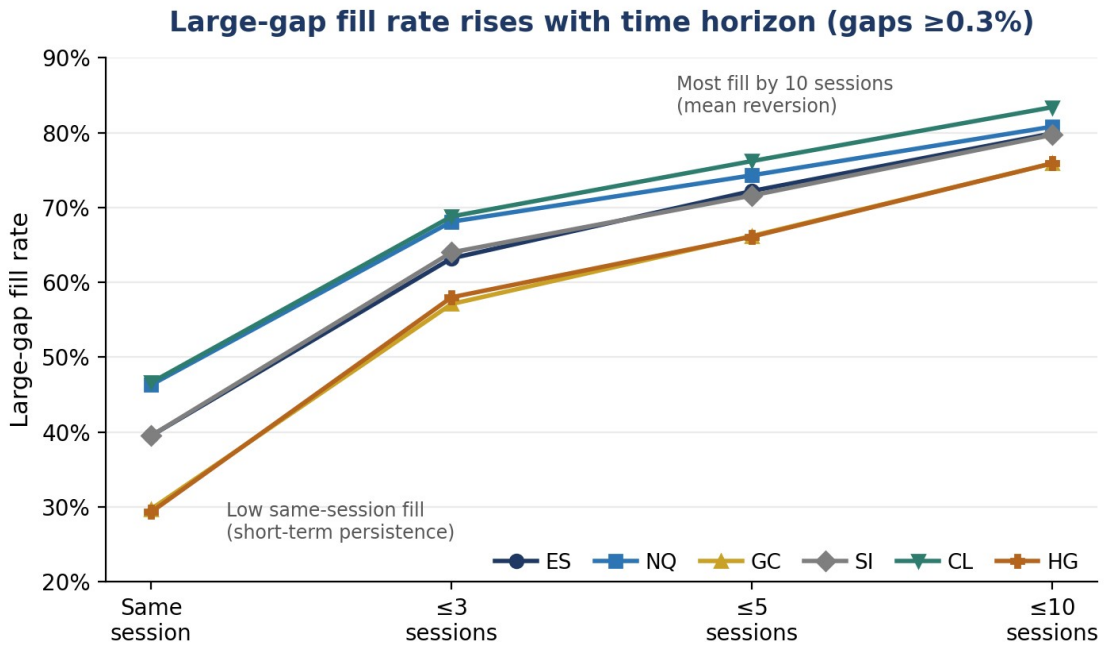

Figure 1. Large-gap (≥0.3%) fill rate by time horizon, all six instruments. Same-session fills are low (short-term persistence); rates climb to 76–83% by ten sessions (mean reversion).

Instrument	Large-gap N	Same session	≤3 sessions	≤5 sessions	≤10 sessions
ES	956	39.5%	63.2%	72.2%	79.9%
NQ	1,127	46.3%	68.1%	74.3%	80.8%
GC	1,070	29.7%	57.1%	66.2%	75.9%
SI	1,408	39.5%	64.0%	71.6%	79.7%
CL	1,485	46.6%	68.8%	76.2%	83.4%

Instrument	Large-gap N	Same session	≤3 sessions	≤5 sessions	≤10 sessions
HG	1,386	29.3%	58.0%	66.1%	75.9%

A note on the residual. Even at ten sessions, the very largest gaps (>1% of prior close) fill less completely — 63–76% across instruments, versus ~99% for the smallest gaps. So a genuine size effect remains at every horizon; it is the magnitude of the same-session gap that shrinks dramatically once time is allowed.

Long-horizon caveat. Fill probabilities are reported through ten trading sessions. Gaps remaining unfilled after ten sessions may fill later and are not classified as permanent non-fills; the study measures fill speed within a trading-relevant window, not whether gaps eventually fill given unlimited time.

3.2 Robustness checks

The finding survives the checks most likely to expose a false positive:

- Cross-instrument replication — six unrelated markets, same direction and rough magnitude.
- Year-by-year stability — the direction holds in all eight years for every instrument. The magnitude is steady for the four commodities; for the two equity indices it weakened during the 2022 bear market (see caveat above) but did not reverse. An artefact of the bull regime would have flipped the sign, which did not occur.
- Trend-regime control — under five independent definitions of trend (moving-average slope and position, rate of change, breakout regime, directional-strength), the size–speed relationship holds in up, down and sideways regimes. It is not simply a by-product of the dominant upward trend observed during much of the sample, though trend modestly modulates the same-session fill rate for the equity indices.
- Permutation testing — against randomly-placed trigger days, the observed effect is unlikely to arise by chance under the tested null models. Detailed permutation statistics for the descriptive findings are available upon request.
- Multiple-comparison control — false-discovery-rate adjustment across all patterns tested.
- **Positive control —** the same method independently recovers volatility clustering (a high-range day follows a high-range day, +23 to +30 pp across all six), a well-established effect. This confirms the method detects real structure when it is present.

4. Why a pattern is not automatically an edge

> **The central distinction of this paper**
>
> Many market behaviours are statistically real yet economically unexploitable. A tendency in *where price travels* describes the destination; a trading edge requires that the *path*, net of stops, slippage and commissions, be profitable. A real pattern can fail to produce a robust strategy once execution and risk-management constraints are applied. This study finds clear evidence that gap size governs fill speed, but — within the implementations tested — not evidence of a robust standalone trading edge: the eventual fill is real, yet its timing is too variable to exploit after costs.

5. Tradability testing

Methodology. Each implementation used next-bar execution; intrabar stop/target resolution (resolved conservatively against the position when both fall in one bar); separation of in-sample fitting from out-of-sample evaluation; walk-forward re-calibration where noted; volatility-normalised (ATR) stops, targets and position sizing so that results are invariant to the near-doubling of some price levels over the period; and results expressed in risk-multiples (R). Transaction costs of approximately two ticks per trade (commission plus slippage) were applied throughout.

Eight implementations were tested, spanning both directions, two timescales (intraday and overnight), several entry and exit constructions, and two validation regimes (in-/out-of-sample split and walk-forward). These implementations intentionally covered both continuation and reversal concepts, multiple holding periods, volatility-adjusted risk models, walk-forward re-calibration, and permutation testing against random placement:

#	Implementation tested	Validation method	Result
1	Overnight-range sweep reversal	In-/out-of-sample split	No reliable edge after costs
2	Gap continuation (directional)	In-/out-of-sample split	No reliable edge after costs
3	Gap fade (opposite direction)	Long/short P&L decomposition	No reliable edge after costs
4	Session-low reversal	Parameter sweep + IS/OOS	No reliable edge after costs
5	Gap continuation, ATR/R-scaled	Walk-forward, two windows	No reliable edge after costs
6	Pre-NY low/high reversal (1:2 R)	Walk-forward, two windows	No reliable edge after costs
7	Parameter-free gap-down short	Permutation vs random	No reliable edge after costs
8	Multi-day overnight hold	Permutation vs random	No reliable edge after costs

6. Limitations

This study has clear limitations, which also map directly onto avenues for further research:

1. **Finite implementation space.** We tested eight implementations with specific entry, exit, stop and target constructions. This is not exhaustive. Other stop models, profit-target structures, holding periods, volatility-regime filters, position-management rules, or portfolio-level combinations could produce different outcomes. Our negative result applies to the implementations tested, not to all conceivable strategies.
2. **Six markets only.** Findings are established on six CME-group futures; other instruments and asset classes may behave differently.
3. **One historical sample.** 2019–2026 contains one significant equity bear episode (2022). The descriptive finding survives it, but a single regime episode is limited evidence; future conditions may differ.
4. **Standardized session definition.** Results are based on a standardized U.S. trading window (09:00–16:00 ET) applied uniformly to all six contracts, rather than each contract's native exchange session. This was a deliberate choice for cross-market comparability. The core relationship — fill probability falling with gap size — is driven by gap magnitude and is unlikely to depend on the exact window, but the precise per-instrument percentages would shift under a different definition. Future research may evaluate whether exchange-specific

(pit-hour) session definitions, or weekly/overnight gap definitions, materially alter gap-fill probabilities.

5. **Execution model.** Backtests use one-minute bars and a conservative same-bar resolution rule; true fills, partial fills and realistic queue dynamics are not modelled.

7. What a researcher could try next

Because we want this work to be useful rather than a dead end, we list concrete extensions we did not test and which could change the tradability conclusion:

- Conditioning gap-fill setups on volatility regime (volatility clustering was the one effect that survived as strongly as the size–speed relationship here).
- Longer holding periods (multi-week), where per-trade costs are a smaller fraction of the move.
- Cross-instrument or lead-lag relationships rather than single-instrument intraday rules.
- Gap-size-conditioned sizing or selection, exploiting the monotonic fill curve directly rather than a single threshold.
- Instrument-native session definitions, particularly for the metals and energy contracts.

8. Practical takeaway

For a practitioner, the useful conclusion is informational rather than directional. Gap size carries statistically measurable information about fill timing: large gaps are unlikely to fill the same session but mostly fill within roughly two trading weeks. That context can inform expectations and patience — a large gap still unfilled at the close is normal, not a failed fill. But this is an observation about timing, not a recommendation: fill probability is not profitability, the minority of gaps unfilled within the window is a material risk, and our strategy tests did not convert this regularity into a reliable edge after costs. The value of this finding is in understanding market behaviour, *not* in a ready-made trade.

9. Conclusion

This study documents a replicated and historically robust relationship — fill probability declines with gap size at every horizon, and larger gaps exhibit slower mean reversion rather than failing to fill — that challenges a commonly repeated trading belief. Small gaps fill almost immediately; large gaps resist same-session filling but mostly fill within about ten sessions. It also demonstrates the equally important distinction between a market behaviour and a trading edge: within the implementations we tested, the regularity did not convert into a statistically reliable strategy after costs, because the timing of the eventual fill is too variable to exploit. A pattern can be observable without being exploitable. Distinguishing between descriptive market behaviour and tradable edge is one of the central goals of quantitative research.

> **Why we publish a study with no edge**
>
> Most trading research shared publicly reports only successes. We believe rigorous research should also report findings that fail to produce an edge. Publishing negative and null results reduces survivorship bias, guards against overconfidence, and gives a more honest picture of how markets actually behave. We would rather show our work — including where it did not lead

> to a tradable strategy — than present a flattering result we could not stand behind.

Appendix A. Selected permutation p-values

For the strategy tests, the permutation p-value is the probability that a random strategy, trading the same days with randomly-signed outcomes, would equal or beat the tested implementation. A small value (< 0.05) would indicate a real edge; the values below are uniformly large, confirming no implementation was distinguishable from random. The table reports the single most-favourable configuration per instrument — the lowest p-value observed across tested configurations — to give the strongest possible challenge to the null hypothesis; typical configurations were less favourable still.

Instrument	Multi-day hold (best p)	Param-free gap-down short (p)
ES	0.942	0.999
NQ	0.773	0.983
GC	0.227	0.487
SI	0.459	0.841
CL	0.673	0.997
HG	0.671	0.959

Note. The descriptive size–speed fill finding (Section 3) passed its own permutation and false-discovery-rate tests; the per-bucket and per-year sample sizes reported in Sections 3 and 3.1 provide the basis for verification and independent replication using equivalent data.

About the research

This study was conducted and authored by Dhaval Barot and published by MPM Markets – Market Probability Model. Questions, comments, replication requests, and research discussions may be submitted through the contact form at mpmmarkets.com/contact.

Suggested citation. Barot, D. (2026). Testing the “Gaps Always Fill” Myth Across Six Futures Markets: Evidence From Nasdaq-100, S&P 500, Gold, Silver, Crude Oil, and Copper Futures. MPM Markets.

First published: June 2026. Research conducted and authored by Dhaval Barot, MPM Markets.