

Crash Replay: Telling Skill from Luck in Market-Crash Decisions with a Permutation Test

Parveen Kumar Naresh Kumar
University of South Florida
github.com/parveen06naresh/crash-replay

Abstract

A good result in one market episode says little about skill, because a path that is rising or falling does most of the work. We describe Crash Replay, an open-source simulator in which a player manages \$10,000 through seven real equity drawdowns (1973 to 2022) one trading day at a time, and an engine that scores the result against buy-and-hold, a trend-following quant, and a reference population of 2,000 random traders who make the same number of trades. The percentile of the player among those traders is a permutation test. We check the test three ways. It is calibrated: on traders drawn from the null itself it flags 5.0% at the 95th percentile (range 3.5%–6.3% over 14 settings, 400 draws each). It has power: a two-trade player whose timing is within ± 10 trading days of the true peak and trough is detected 95–100% of the time in every crash, but that falls to 19–93% at ± 63 days. And it separates three common retail rules: a panic seller in Black Monday is worse than 99.5% of random traders, while a buy-the-dip rule in the 2020 rebound beats essentially all 2,000. The engine has no look-ahead and is tested in CI against historic days.

1 Introduction

Anyone who sold in September 2008 and bought back in March 2009 looks brilliant, and anyone who did the opposite looks foolish. Both stories are told after the fact, with the path known. The question a skeptical reader should ask is the one quantitative researchers ask of every backtest: how often would a trader with no information have done as well?^[5,6] The answer is a number between 0 and 1, and it depends heavily on how many trades were made and how long the episode lasted.

Crash Replay turns that question into something a person can play in about three minutes, then answers it with a permutation test. The contribution is not a trading strategy. It is a small, fully reproducible measurement instrument, and a study of that instrument: is it calibrated, how much power does it have, and does it say sensible things about behaviour people actually exhibit? The code, data pipeline and every experiment in this paper are in the repository and run in under three minutes on a laptop.

2 Design

2.1 Data and accounting

The simulator uses 76 years of daily S&P 500 closes and a Nasdaq series, 3-month Treasury bill yields from FRED, and S&P dividend yields from Robert Shiller's data.^[9] A portfolio holds an exposure e between 0 and 1 to the index, with the rest in cash at the T-bill rate. Dividends are added to the S&P return; the Nasdaq series is price-only. A decision is made at a close and earns the next day's return, so no decision ever sees the day it is paid for. Each change of exposure costs 5 basis points of the value moved:

$$V_{t+1} = V_t (1 + e_t r_{t+1}^m + (1 - e_t) r_{t+1}^f), \quad V \leftarrow V - 0.0005 |e' - e| V \text{ on a trade.}$$

Both properties are unit-tested: tomorrow's return cannot change today's value, and a trade costs exactly 5 bp. The tests also pin the data to the record (Black Monday is -20.47% , 29 September 2008 is -8.81% , and the 2008 and 1973 to 1974 index drawdowns are 56.8% and 48.2%).

2.2 Seven crashes

Windows are chosen to start before the peak and end after the recovery begins (Table 1). They differ a lot: Black Monday falls 34% in about three months, the dot-com bust takes 78% off the Nasdaq over three years, and the 2011 debt-ceiling episode is a 19% fall that fully recovers within the window.

2.3 The quant opponent

The opponent follows two rules used by trend-following funds: be invested only while the close is above its 200-day average^[2,3], and size the position as $\min(1, 15\% / \text{realised 20-day volatility})^{[4]}$, rounded to 5-point steps and only re-sized when the target moves by 10 points. Nothing about it is tuned to any of the seven windows.

2.4 The skill test

If a player makes k trades, the reference population is 2,000 traders who also make exactly k trades, on days drawn uniformly without replacement, each moving to a position drawn uniformly from $\{0, 0.5, 1\}$ other than the current one. Everyone starts fully invested. The player's score is the share of reference traders whose final value is lower, called the percentile q . This is a Monte Carlo permutation test of the null "trade timing is exchangeable with random timing",^[8] and $1 - q$ is its one-sided p-value. Conditioning on k matters: a trader who trades twice is not compared with one who trades forty times, and the reference includes the cost of its own trades.

3 Results

3.1 A worked example

Figure 1 is the run shown in the project's demo video: sell everything at the Lehman close (15 September 2008) and go all in on 18 March 2009. It returns +15.3% against -22.0% for buy-and-hold and +2.9% for the quant, with a maximum drawdown of 24.8% against 55.3%. Against 2,000 random two-trade traders it ranks at the 92.8th percentile (Figure 2; the in-game number can differ by about a point because each game draws its own seed, and the Monte Carlo standard error at 2,000 draws is 0.6 points). The one-sided p-value is about 0.07. That is good, but it is not evidence of skill at the 5% level: with one run in one crash, a lucky pair of dates is the more economical explanation, and the game says so to the player.



Figure 1. The worked example. Starting value \$10,000, 2007-06-01 to 2009-12-31.

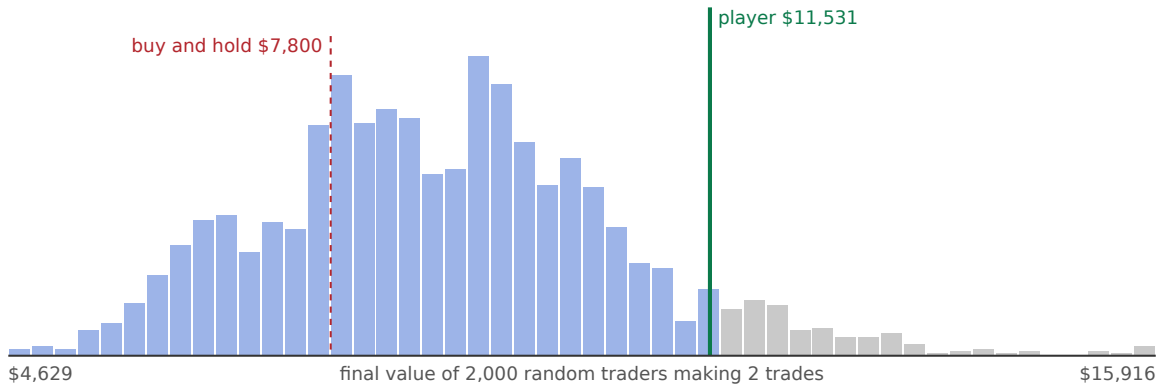


Figure 2. Reference distribution for the worked example. Blue bars are traders the player beats (92.8%).

3.2 The seven crashes and the quant

Crash	Days	Index DD	Buy and hold			Trend + vol target			
			Ret	MaxDD	Sharpe	Ret	MaxDD	Sharpe	Trades
1973 Oil Shock	689	48%	-4.7%	45%	-0.39	+31.9%	9%	0.55	17
1987 Black Monday	211	34%	-8.6%	33%	-0.29	+7.7%	10%	0.36	4
2000 Dot-com Bust	814	78%	-51.1%	78%	-0.42	+22.3%	16%	0.33	34
2008 Financial Crisis	651	57%	-22.0%	55%	-0.19	+2.9%	16%	0.04	38
2011 Debt-Ceiling Scare	228	19%	+4.9%	19%	0.34	-3.5%	12%	-0.34	11
2020 COVID Crash	272	34%	+22.3%	34%	0.71	+7.7%	14%	0.59	20
2022 Rate Shock	458	36%	-11.2%	36%	-0.20	+10.1%	9%	0.35	19

Table 1. The seven windows. Returns are total returns over the window, net of costs, with dividends and cash interest.

The quant has a higher return than holding in 5 of 7 windows and a higher Sharpe ratio in 5 of 7; it loses in 2011, where it was whipsawed (11 trades) by a sharp August fall and rebound, and in the 2020 rebound. Its maximum drawdown is 9%–16% against 19%–78% for holding. This should be read as a property of the scenario selection, not of the rule. The windows were picked because they contain crashes, and trend-following is built to do well in sustained falls;^[3] a fair evaluation would also include long bull markets, where the same filter lags. The quant is an opponent, not a result.

3.3 Calibration

A test that flags skill too often is worse than none. We drew 400 reference-style traders per setting ($k = 2$ and $k = 6$, all seven windows), scored each with an independent 1,000-draw test, and counted how often the percentile reached 0.95. The false-positive rate is 5.0% on average (range 3.5%–6.3%), and 9.9% at the 0.90 threshold, within Monte Carlo error of the nominal 5% and 10% (the standard error at $N = 400$ is 1.1 points). The mean percentile is between 0.48 and 0.52 everywhere. This confirms the implementation rather than the realism of the null, which is a modelling choice discussed in Section 5.

3.4 Power

Power depends on the effect: how much better than random timing a player is. We use a controlled one. A noisy oracle sells within $\pm J$ trading days of the true peak and buys within $\pm J$ days of the true trough (two trades), with J drawn uniformly, and we report how often its percentile reaches 0.95 (200 oracles per cell). Figure 3 shows the

result. At $J = 10$, about two weeks, detection is 95–100% in every crash. At 21 days it is 64–100%, at 63 days 19–93%, and at 126 days 6–60%.

The spread across crashes is the useful finding. Long, deep windows (2008, 1973, 2000) keep power above 0.8 at 42 days because a mistimed exit still avoids most of the fall; Black Monday, which is over in about three months, drops below 0.4 at the same error. The practical reading is that a high percentile is strong evidence of good timing in a long crash, and a middling one in a short crash tells you almost nothing, because there was little to time. The game reports the percentile with the sentence "better than most, though luck could still explain it" for exactly this reason.

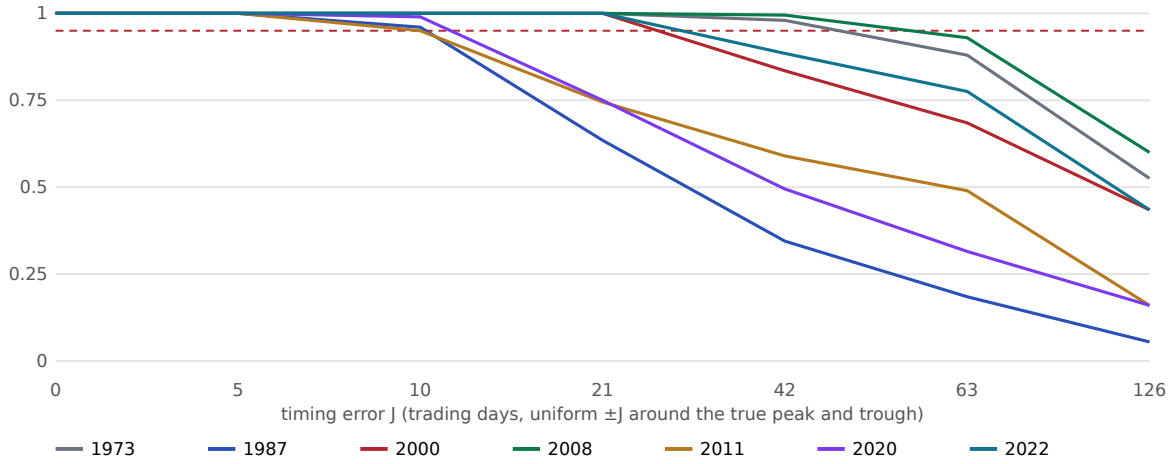


Figure 3. Detection rate (percentile ≥ 0.95) of a two-trade player with timing error J . Dashed line: 0.95.

3.5 Three common behaviours

To check that the test says sensible things about real behaviour, we ran three mechanical rules with thresholds fixed in advance (not tuned). **Panic seller**: sell everything when the index is 20% below its running peak, and buy back only when it regains that peak. **Rebound buyer**: sell at the same trigger, buy back after a 20% bounce off the post-sale low. **Dip buyer**: start half invested, go all in at -20% from peak. An oracle that sells at the peak close and buys at the trough close shows the ceiling. Table 2 gives returns with the percentile in brackets.

Crash	Oracle	Panic seller	Rebound buyer	Dip buyer
1973	+96.9%	+0.5% (56)	+25.7% (84)	+4.7% (64)
1987	+38.6%	−19.5% (1)	−23.1% (0)	+5.3% (75)
2000	+144.2%	+51.4% (92)	−51.3% (6)	−57.3% (0)
2008	+78.1%	−16.8% (51)	+5.9% (86)	−12.4% (67)
2011	+28.9%	no trade	no trade	no trade
2020	+84.6%	−11.3% (3)	+10.0% (44)	+37.5% (100)
2022	+42.3%	−12.4% (60)	−11.0% (36)	−2.7% (82)

Table 2. Return and percentile among random traders with the same trade count (in brackets). The oracle ranks 100 everywhere by construction. "No trade": the -20% trigger never fired in 2011, whose fall was 19%.

The panic seller's results are mixed, and instructively so. It is catastrophic in Black Monday (-19.5% , percentile 0.5: worse than 99.5% of random single-trade traders), because the -20% trigger first fires at the close of the crash day itself, near the low, and the rule then sits in cash through the rebound. It is also poor in 2020 (-11.3% , percentile 3). In the dot-com bust it ranks 92 ($+51.4\%$): the Nasdaq never regained its 2000 peak inside the window, so staying in cash was right. In 1973 and 2008 it is unremarkable. The dip buyer is the clearest split:

+37.5% in 2020 (percentile 100, the V-shaped rebound rewarded buying early) and −57.3% in 2000 (percentile 0.3, buying a falling Nasdaq at −20% two years early). Neither is skill or stupidity. They are one rule meeting two different paths, which is precisely what a single-crash anecdote cannot show, and why the game lets a player try several crashes.

4 Related Work

Barber and Odean show that active retail trading lowers returns on average,^[1] which motivates conditioning the null on trade count. Faber and Moskowitz, Ooi and Pedersen document trend and time-series momentum rules of the kind the quant uses,^[2,3] and Moreira and Muir study volatility-targeted sizing.^[4] The statistical frame is the standard one for backtest overfitting and data snooping: White's reality check,^[6] and the deflated Sharpe ratio and the warning against tuning many variants on one path.^[5,7] We apply the same logic to a single player's decisions, with a simulation null instead of an analytic one.

5 Limitations

The null is a model. Random timing to random exposures is a convenient reference, not a model of a skilled trader or of a typical investor; a different null (for example, uniform over exposures in $[0, 1]$, or timing drawn near turning points) changes the percentile. **Seven windows chosen for being crashes** bias every comparison with buy-and-hold, as noted in Section 3.2. **One index at a time**, daily closes only, no taxes, a flat 5 bp cost, and exposure in only three levels for the reference. **Few trades.** With one or two trades the test has few distinct reference outcomes, and a single percentile has Monte Carlo error of about 0.6 points. **No multiple-comparison correction** across replays: a player who tries all seven crashes and reports the best has inflated their own result (the game's per-crash percentile does not account for this). **The behavioural rules** in Section 3.5 are three simple stand-ins, not a study of investors, and we did not tune them. **Educational use only**; none of this is investment advice.

6 Reproducing

Everything is in the repository: `node --test tests/` runs the 7 engine tests, `node paper/experiments.mjs` regenerates every number in this paper from the same `web/engine.js` the game runs (seeded, deterministic, about two minutes), and `node paper/build.mjs` rebuilds the figures and tables. The game is playable at parveenkumar.site/crash-replay.

References

1. B. Barber and T. Odean. Trading is hazardous to your wealth: the common stock investment performance of individual investors. *Journal of Finance* 55(2), 2000.
2. M. Faber. A quantitative approach to tactical asset allocation. *Journal of Wealth Management* 9(4), 2007.
3. T. Moskowitz, Y. Ooi and L. Pedersen. Time series momentum. *Journal of Financial Economics* 104(2), 2012.
4. A. Moreira and T. Muir. Volatility-managed portfolios. *Journal of Finance* 72(4), 2017.
5. D. Bailey, J. Borwein, M. López de Prado and Q. Zhu. Pseudo-mathematics and financial charlatanism: the effects of backtest overfitting on out-of-sample performance. *Notices of the AMS* 61(5), 2014.
6. H. White. A reality check for data snooping. *Econometrica* 68(5), 2000.
7. D. Bailey and M. López de Prado. The deflated Sharpe ratio. *Journal of Portfolio Management* 40(5), 2014.
8. P. Good. *Permutation, Parametric and Bootstrap Tests of Hypotheses*, 3rd ed. Springer, 2005.
9. R. Shiller. *Irrational Exuberance* (online data), Yale. S&P composite prices and dividends. Also: FRED, Federal Reserve Bank of St. Louis, 3-month Treasury bill; Yahoo Finance index histories.