Longshot Bias Analyses

Hong Kong racing: closing Win and Place odds versus realised outcomes

Study window: 11 September 2022 – 14 October 2026 | Prepared: 20 October 2026 | Community-sharing edition

Abstract

Do larger potential payouts compensate for how rarely longshots win or place? Across 42,463 horse-runs in 3,485 races, the answer is no on average in this sample. Independent, approximately equal-frequency 20-bin analyses show materially worse odds-based level-stakes returns at the high-odds end of both the Win and Place pools. The pattern persists in each original file and across all four completed seasons. These are historical, data-implied estimates—not independently reconciled official settlement returns.

Pool	Group	N	Hits	Hit rate	Est. ROI	95% CI %
Win	Short	8,798.0	2,024.0	23.01%	-13.65%	-16.1 to -11.1
Win	Long	8,378.0	67.0	0.80%	-44.23%	-58.1 to -29.5
Place	Short	9,177.0	4,847.0	52.82%	-16.25%	-17.7 to -14.9
Place	Long	7,358.0	276.0	3.75%	-41.34%	-48.4 to -34.1

Main finding: longshots lose more per unit staked. This does not mean favourites make money: all 20 Win bins and all 20 Place bins have negative estimated ROI.

Three takeaways

1. Relative mispricing matters: high odds are not automatically good value. The high-minus-low estimated ROI gap is –30.59 percentage points for Win and –25.09 points for Place.

2. Hit rate and investment value are different. Lower-odds horses win or place more often, but their shorter payouts still leave negative average returns.

3. This is an aggregate calibration study, not a betting system or a causal test of bettor psychology. An individual longshot can still be underpriced or overpriced.

Low-odds group = bins 1–4; high-odds group = bins 17–20, defined separately for each pool. Win: odds ≤6.5 versus >45. Place: odds ≤2.1 versus >10.

Format follows the research-led structure used in Sigma Quant's Odds Drift Research and Money Flow and "Following Smart Money": separate predictive information from price value, and state limitations explicitly.

1. Data assembly and outcome rules

The historical file contains 21,158 rows from 11 September 2022 to 13 October 2024. The recent file contains 21,305 rows from 16 October 2024 to 14 October 2026. Both have the same 104 columns. There are no overlapping dates, no duplicate Date–Heat–Horse_Number keys and no missing or non-positive Win/Place odds. The supplied files end on 14 October, not the report date.

Two combinations are supplied. The literal horizontal merge places the ig-2 file's 104 columns on the left and the recent file's 104 on the right, aligned by row index; it has 21,305 rows and 208 columns, with 147 blank historical-side rows. It is an archive view, not race-matched data. The analysis uses a chronological row append: ig-2 first, then recent, yielding 42,463 observations and the original 104 columns. No unrelated records are treated as one horse-run.

Outcome definitions, exactly as requested

A race is a unique Date–Heat pair. Field size is the count of supplied horse records in that pair. Win = 1 only when Rank = 1. Placed = 1 when Rank ≤3 in a race with at least seven horses, or Rank ≤2 with six or fewer. All supplied ranks are positive integers. The data contains 32 races with 5–6 horses and 3,453 races with 7–14 horses.

This gives 3,490 winning horse-runs and 10,437 placed horse-runs. These exceed one winner / the standard slot count in some races because of tied or otherwise non-standard ranks. The main analysis preserves the user's definitions; a sensitivity check removes affected races. HKJC's published local Place rule uses declared starters, which is not necessarily identical to the CSV record count.

2. Binning: balance sample size, preserve readable cutoffs

The target is 2,123.15 horse-runs per bin. We start from 5%, 10%, …, 95% empirical odds quantiles and choose nearby readable boundaries, without looking at outcomes. Win and Place use separate boundaries; identical odds are never split across bins. The final count is exactly 20 bins per pool. This is a transparent, manually rounded quantile design, not a claim of mathematically optimal equalisation.

All internal intervals are (lower, upper]: the lower endpoint is excluded and the upper included. The first bin includes everything up to its cutoff; the last includes everything above its cutoff. Win bin sizes range from 1,309 to 2,958; Place from 1,421 to 3,085. Coarse odds ticks make exact equal counts impossible without splitting tied odds.

Win upper cutoffs: 3, 4.5, 5.5, 6.5, 7.5, 9, 10, 11, 13, 15, 17, 20, 25, 30, 35, 45, 60, 80, 115.

Place upper cutoffs: 1.4, 1.6, 1.8, 2.1, 2.3, 2.6, 2.9, 3.2, 3.5, 4, 4.5, 5, 6, 7, 8, 10, 12, 15, 20.

The workbook includes raw quantiles, selected cutoffs and actual cumulative sample shares, so each rounding choice can be inspected.

3. Win odds: all 20 bins

Binned by final Win odds. Both average odds and both realised outcome rates are shown. “Mean 1/o” uses each horse’s Win odds; ROI is the odds-based Win estimate. All rates and intervals below are percentages.

Bin	Odds range	N	Avg Win	Avg Place	Win %	Place %	Mean 1/o %	Est. ROI %	95% CI %
1	<= 3	1,927	2.35	1.24	39.70	69.12	44.87	-11.37	-16.6 to -6.1
2	(3, 4.5]	2,811	3.81	1.57	21.99	55.53	26.57	-16.47	-21.8 to -10.8
3	(4.5, 5.5]	2,007	5.06	1.83	16.94	47.58	19.82	-14.18	-22.2 to -6.4
4	(5.5, 6.5]	2,053	6.05	2.05	14.66	40.57	16.57	-11.40	-19.7 to -2.6
5	(6.5, 7.5]	1,767	7.04	2.27	12.11	37.13	14.22	-14.77	-24.8 to -4.2
6	(7.5, 9]	2,547	8.29	2.52	9.38	33.22	12.10	-22.33	-30.9 to -13.5
7	(9, 10]	2,958	9.76	2.90	6.73	27.89	10.26	-34.52	-43.2 to -25.4
8	(10, 11]	1,309	11.00	3.20	7.64	25.97	9.09	-15.97	-31.8 to 0.4
9	(11, 13]	2,351	12.46	3.50	6.72	24.37	8.04	-16.29	-28.1 to -3.1
10	(13, 15]	2,085	14.49	3.91	5.37	22.16	6.91	-22.25	-35.3 to -8.6
11	(15, 17]	1,730	16.47	4.26	4.34	19.36	6.08	-28.50	-44.3 to -12.5
12	(17, 20]	2,042	18.96	4.80	3.72	18.22	5.28	-29.68	-45.7 to -13.0
13	(20, 25]	2,700	22.90	5.51	3.11	14.81	4.38	-29.48	-44.2 to -13.4
14	(25, 30]	1,886	27.89	6.51	2.70	11.77	3.59	-24.71	-45.4 to -1.6
15	(30, 35]	1,551	32.92	7.44	2.84	11.03	3.04	-7.09	-33.9 to 19.7
16	(35, 45]	2,361	40.26	8.80	1.99	8.17	2.50	-19.86	-40.9 to 2.9
17	(45, 60]	2,339	52.49	10.92	1.33	6.41	1.92	-32.11	-54.0 to -10.3
18	(60, 80]	1,984	69.59	14.12	1.16	5.39	1.45	-20.01	-51.2 to 13.1
19	(80, 115]	1,997	96.16	18.70	0.40	3.41	1.05	-62.19	-85.9 to -33.0
20	> 115	2,058	158.03	28.67	0.24	1.75	0.66	-63.95	-92.9 to -29.6

Reading the extremes: odds ≤3 win 39.70% of the time, versus a mean inverse-odds break-even benchmark of 44.87%; estimated ROI is –11.37%. Above 115, only 5 of 2,058 horses win (0.24%), versus a 0.66% benchmark; estimated ROI is –63.95%.

The extreme >80 group combines the final two bins: 13 wins from 4,055 runs, 0.32% hit rate and –63.08% estimated ROI. Within the >115 bin, 1 / observed win rate is 411.6, compared with mean offered odds of 158.03; that is a descriptive historical comparison, not a fair-price estimate for each horse.

The curve is not monotonic. The (30,35] bin returns –7.09%, much better than neighbouring bands, but remains negative. Do not convert a noisy local bump into a new betting rule. Confidence intervals are pointwise date-cluster bootstrap intervals; see Section 6.

4. Place odds: all 20 bins

Binned independently by final Place odds. “Placed” follows the specified field-size rule, not an unconditional top-three definition. “Mean 1/o” now uses Place odds; ROI is the odds-based Place estimate.

Bin	Odds range	N	Avg Win	Avg Place	Win %	Place %	Mean 1/o %	Est. ROI %	95% CI %
1	<= 1.4	2,480	2.79	1.25	35.36	67.62	81.02	-16.53	-18.8 to -14.3
2	(1.4, 1.6]	1,951	4.12	1.55	21.32	54.74	64.53	-15.11	-18.2 to -11.9
3	(1.6, 1.8]	1,989	5.12	1.75	16.94	47.26	57.19	-17.35	-21.0 to -13.8
4	(1.8, 2.1]	2,757	6.39	2.00	14.18	42.15	50.15	-16.02	-19.6 to -12.4
5	(2.1, 2.3]	1,652	7.80	2.25	11.86	36.50	44.51	-17.98	-22.9 to -13.1
6	(2.3, 2.6]	2,339	8.54	2.50	9.53	32.24	40.11	-19.58	-23.8 to -15.3
7	(2.6, 2.9]	2,154	9.98	2.80	7.85	28.51	35.77	-20.32	-25.4 to -15.1
8	(2.9, 3.2]	2,086	11.26	3.10	6.57	25.12	32.30	-22.24	-27.9 to -16.7
9	(3.2, 3.5]	1,849	12.75	3.40	6.76	25.37	29.45	-13.91	-19.9 to -7.9
10	(3.5, 4]	2,557	14.71	3.79	5.98	22.64	26.42	-14.35	-21.0 to -8.0
11	(4, 4.5]	2,126	17.13	4.29	4.33	21.07	23.32	-9.82	-17.5 to -2.3
12	(4.5, 5]	1,767	19.85	4.80	4.30	16.53	20.87	-20.86	-29.5 to -12.3
13	(5, 6]	2,717	23.58	5.52	3.24	15.31	18.18	-15.72	-22.6 to -8.7
14	(6, 7]	2,029	28.74	6.52	2.86	12.12	15.37	-21.26	-29.1 to -12.7
15	(7, 8]	1,567	34.46	7.54	2.55	8.68	13.29	-34.66	-44.5 to -24.8
16	(8, 10]	3,085	43.65	9.23	2.01	7.55	10.89	-30.51	-38.4 to -22.2
17	(10, 12]	1,421	56.57	11.47	1.34	6.33	8.74	-28.08	-41.8 to -13.3
18	(12, 15]	1,666	68.96	13.89	0.84	5.10	7.22	-29.35	-42.4 to -15.7
19	(15, 20]	1,745	90.30	17.78	0.63	3.21	5.66	-43.72	-57.5 to -28.5
20	> 20	2,526	143.64	28.14	0.24	1.78	3.73	-55.07	-67.7 to -41.8

Odds ≤1.4 place in 67.62% of runs, below the 81.02% mean inverse-odds benchmark; estimated ROI is –16.53%. Above 20, only 45 of 2,526 horses place (1.78%), against a 3.73% benchmark; estimated ROI is –55.07%.

Across the broad low- and high-odds groups, estimated ROI worsens from –16.25% to –41.34%. The two extreme bins (>15) combine to 101 places in 4,271 runs, a 2.36% hit rate and –50.43% estimated ROI.

The (4,4.5] Place bin is the least negative at –9.82%; this does not establish an exploitable sweet spot. Place payouts depend on pool and settlement structure, so inverse Place odds are used only as a payout-based break-even benchmark, not a uniquely identified true placing probability.

5. What comparison actually diagnoses bias?

Average odds are not the same thing as average probability

Return convention: treat the recorded Win/Place odds as gross payout multipliers inclusive of the stake. A hit at 4.0 returns 4 units on a 1-unit stake, a net gain of 3; a miss returns zero. This is a dataset convention for estimation, not independent dividend certification.

For horse i, let o_i be the supplied odds multiplier and y_i the win/place indicator. We report mean odds = sum(o_i)/N, observed rate = sum(y_i)/N, and mean inverse odds = sum(1/o_i)/N. In general, mean(1/o_i) is not equal to 1/mean(o_i). Using only the reciprocal of average odds can distort wide-bin comparisons.

The odds-based level-stakes estimate is ROI = sum(o_i × y_i)/N – 1. It uses each winning horse's own odds. Multiplying mean odds by the bin's hit rate is generally wrong because odds and outcomes are associated. A raw observed/expected ratio = sum(y_i)/sum(1/o_i) is a diagnostic, not ROI.

Separate the payout hurdle from the market probability baseline

The sum of 1/Win across a race averages 1.22749 (122.75%). Consequently, a fair comparison cannot treat every raw reciprocal as a takeout-free probability. This observed reciprocal sum is not a direct estimate of a statutory takeout rate.

For Win only, define q_i = (1/Win_i) / sum_race(1/Win). This gives a race-normalised market baseline summing to one. For this diagnostic alone, a race with k Rank=1 horses assigns each a realised outcome of 1/k. Main win rates and ROI retain the original indicators.

Win band	Normalized baseline	Observed win rate*	Observed / expected	Est. ROI
<= 3	36.54%	39.67%	1.086	-11.37%
(5.5, 6.5]	13.51%	14.66%	1.085	-11.40%
(20, 25]	3.57%	3.09%	0.866	-29.48%
(30, 35]	2.48%	2.84%	1.144	-7.09%
(80, 115]	0.86%	0.40%	0.468	-62.19%
> 115	0.54%	0.24%	0.448	-63.95%

*Tie-adjusted only in this table. The shortest band records about 8.6% more wins than its normalised baseline, while the two longest bands achieve only about 46.8% and 44.8% of baseline expected wins. This supports a relative favourite–longshot pattern beyond a common payout haircut; the baseline is not ground truth.

> Favourite–longshot bias means that longshots have worse average returns than favourites—not simply that most bets lose. The literature documents this distinction; the present evidence is historical and market-specific.

We do not normalise inverse Place odds to sum to two or three and call the result true placing probabilities: the Place settlement structure does not justify that shortcut.

6. Uncertainty and independent-period checks

Uncertainty is estimated by resampling 363 complete race dates with replacement, 2,000 times, using seed 20261020. All horses and races on a selected date travel together; this preserves within-race and same-day dependence. Bin boundaries stay fixed. The reported 95% intervals are bootstrap percentile intervals, not independent-horse binomial errors.

Pool	Group	N	Hits	Hit rate	Est. ROI	95% CI %
Win	Short	8,798.0	2,024.0	23.01%	-13.65%	-16.1 to -11.1
Win	Long	8,378.0	67.0	0.80%	-44.23%	-58.1 to -29.5
Place	Short	9,177.0	4,847.0	52.82%	-16.25%	-17.7 to -14.9
Place	Long	7,358.0	276.0	3.75%	-41.34%	-48.4 to -34.1

Contrast	Estimated gap (pp)	95% CI (pp)
Win: high minus low	–30.59	–45.01 to –15.28
Place: high minus low	–25.09	–32.23 to –17.44

Both aggregate intervals lie below zero. That is evidence of lower historical returns in the high-odds bands, under this sampling design. It does not establish that every adjacent bin differs, nor does it provide simultaneous significance across 40 bins. Boundary selection uncertainty and dependence across separate dates are not captured.

Earlier and recent files tell the same directional story

Source period	Pool	Short ROI	Long ROI	Long minus short (pp)
Earlier	Win	-11.02%	-44.27%	-33.26
Recent	Win	-16.30%	-44.20%	-27.91
Earlier	Place	-15.75%	-37.95%	-22.20
Recent	Place	-16.76%	-44.09%	-27.34

Earlier = 11 September 2022–13 October 2024; recent = 16 October 2024–14 October 2026. The high-odds disadvantage appears in both original files for both pools. Boundaries are derived from the full sample, so this is a stability check, not a prospective out-of-sample trading test.

The same direction also holds at both venues (Sha Tin and Happy Valley). Venue and field-size summaries are supplied in the workbook, but small field-size cells should not be treated as independent proof of a universal effect.

7. Cross-season stability and data-quality sensitivities

Season	N	Win short ROI	Win long ROI	Place short ROI	Place long ROI
2022/23	10,104	-11.05%	-37.76%	-16.26%	-37.48%
2023/24	9,901	-11.64%	-48.67%	-15.56%	-38.53%
2024/25	10,341	-17.30%	-57.74%	-17.40%	-46.66%
2025/26	10,797	-15.04%	-36.84%	-16.13%	-44.34%
2026/27	1,320	-8.58%	-14.29%	-13.22%	-14.55%

Low = Win ≤6.5 / Place ≤2.1; high = Win >45 / Place >10. Seasons are assigned from September to August. The four completed seasons all show worse high-odds returns in both pools. The 2026/27 sample contains only 1,320 horse-runs through 14 October: its smaller gaps are too early to call a structural change.

Outcome anomalies: preserve first, then test robustness

Five races have two Rank=1 horses; 14 other races have four horses satisfying the requested placed rule. These 19 races contain 230 rows. Some tied ranks are not coded in standard competition order. We do not silently rewrite the user's outcome rule or claim that every qualifying row received an official dividend.

Excluding all 19 outcome-anomaly races leaves 42,233 rows. The high-minus-low ROI gap is still –31.62 points for Win and –25.12 for Place. Excluding only the 76 Rank=15 runners gives –30.68 and –25.15 points. Rank=15 comments include non-finishers; they remain losses in the primary analysis rather than being automatically removed.

Stored returns versus odds-based estimates

The main tables use supplied odds × outcome as a return proxy, not certified HKJC payouts. Win_Return differs from this proxy on 10 rows and Place_Return on 24, consistent with additional adjustments around tied results. Some positive stored Place_Return values are below 1.0 (minimum 0.65), so their exact settlement interpretation needs audit.

Return convention	Win overall ROI	Place overall ROI
Odds × specified outcome	–25.32%	–22.92%
Stored return columns	–25.56%	–23.02%

Using stored returns also leaves the broad longshot disadvantage intact: Win high-minus-low = –30.80 points; Place = –25.03 points. Neither series is independently reconciled to official dividends. HKJC results illustrate that displayed Win odds and paid dividends can differ in precision; rebates, refunds and settlement conventions can also matter.

8. Interpretation, limits and community message

What the evidence supports

In these supplied Hong Kong records, the compensation offered by long odds is too small, on average, relative to the rarity of success. The short-to-long deterioration is economically large, survives broad period and season checks, and remains after removing anomalous outcome races. This is evidence consistent with favourite–longshot bias in both pools. It does not identify whether probability misperception, preferences for lottery-like payoffs, informed trading or another mechanism caused it.

What it does not support

It does not support “always back favourites”, “never back a longshot”, a guaranteed profitable band, or importing one universal odds-to-probability correction into every race. All 40 displayed bands lose on the historical odds-based estimate. A relative advantage is not a positive absolute expectation.

The existing Odds Drift report treats market signals as overlays rather than proven betting edges; the Money Flow report likewise distinguishes a high hit rate from positive expected value. This historical dataset adds an empirical odds-level check, not validation of those papers’ simulations or every theoretical claim.

Limitations to retain when sharing

• The final odds are supplied by the user’s dataset; timestamps, full race coverage, official starters, refunds and dividends have not been independently reconciled. A closing snapshot is not necessarily a price available when a real bet is placed.

• Odds-based returns use recorded multipliers and requested rank flags. Dead heats, rounded quotes, compacted ranks, minimum dividends and any eligible rebates can change realised cash returns. Existing odds already reflect pool pricing; no extra takeout haircut is applied.

• The 20-bin design is rounded and full-sample descriptive. It preserves tied odds but does not provide exact equal counts. Local peaks invite selection bias; the extreme Win tail has few winners and wide uncertainty.

• Date resampling is not proof of causal behaviour or future stationarity. There is no own-bet market impact, execution model, staking optimisation or prospective test in this study.

Shareable summary

> A bigger payout is not automatically a better bet. In 42,463 Hong Kong horse-runs, the broad high-odds groups returned about –44% for Win and –41% for Place, compared with roughly –14% and –16% for their low-odds counterparts. Longshots were worse value on average—but favourites still lost. The key is probability relative to price, not how attractive the payout looks.

For analysis and education only; not betting advice and no returns are guaranteed. Racing involves risk. Research context: Sigma Quant Research.