FIGURING OUT THE FIGURES

By
James Willoughby

Why we should be wary of seeing patterns in small samples of data

WHICH is a better average? Two wins from five runs, or 36 from 200?

The first ratio is 0.40 or 40%, while the second is 0.18 or 18%. These numbers are far apart, and the first is much higher. In most cases, however, a jockey or trainer or sire represented by the second statistic will be much more successful in the long run.

The importance of sample size is one of the most glaring errors in the analysis of horse racing. This is because statistical significance – the reliability of a sample of data to represent a ground truth about the world – is all but ignored. To understand why we should be cautious about records like two from five, let's use tossing a coin as an idealised model of the kind of random sequences in racing.

Say you are told that, of 20 coins in a bag, one is biased towards Heads (H). You draw a coin blindly from the bag and toss it three times, recording three Heads. Should you conclude that it is biased?

Plausible

The probability of Heads with a fair coin is ½ * ½ * ½ which is ⅛. If an event has the probability of ⅛ it happens once every eight trials, so the odds are 7-1 against. Though it becomes more plausible that the coin is biased, one cannot be certain, not least because 19 of the 20 coins placed in the bag (and most of those in the world) are known to be fair.

By contrast, if you were to toss this same coin 100 times, and the result were 59 Heads and 41 Tails (T), you could almost certainly conclude it was biased. This result seems far less compelling than getting three Heads from three trials, yet the likelihood it could happen by chance is only 3% or near 33-1 against. Bear in mind that new drugs are considered effective if there is a 5% chance they have only the placebo effect, so our coin-tossing experiment represents stronger evidence than scientific proof.

Consider this biased 59H and 41T coin once again. Because each toss is independent of the others, and 41 Tails came up in the sequence, it is entirely reasonable that three Tails could have come up in the first three tosses. In fact, there is a 17% chance (0.41 * 0.41 * 0.41) that this could happen, which equates to odds of only 5-1.

Think about this: if you tossed a coin and it came up TTT, what price would you think it is that the coin is actually biased the other way, towards Heads? I'm guessing you would say a lot more than 5-1.

While an unbiased coin has a 50% chance of coming up Heads, a horse in a race chosen at random can be said to have about a 11% chance of being the winner – if no other information is known to bias its chance one way or another. (In 2017, British Flat races had an average of nine runners, and one from nine is roughly 11%.)

So, returning to the introduction, it turns out from a standard mathematical formula called the Binomial Theorem that 36 wins out of 200 is 10 times less likely to occur at random than two wins from five runs, if the probability of each win is 11%.

The point is this: most racing statistics drawn from a small sample must be viewed as highly uncertain to represent the truth which is purported.

****

To illustrate the importance of sample-size further, let's use some data specific to horses trained by MJR. Table 1 shows the record of the stable's runners in 2016 (green) and 2017 (blue) on 36 British Flat venues, counting runners at the Rowley Mile and July course as one entity because it is desirable for the purposes of this exercise.

The rows of Table 1 are ranked by descending order of the difference between strike rates (SR) in the two seasons, with the SR rounded up or down to the nearest whole number. At the top, for instance, is Brighton, where MJR runners were 0-17 (0%) in 2016, but 7-19 (37%) in 2017. This is an absolute difference of 37%, as shown in the rightmost column. Here, it doesn't matter whether the strike-rate increased or declined, it is only the difference which is of interest.

Near the bottom of the table is the aggregated total for Newmarket's two racecourses: 15-103 (15%) in 2016 and 15-98 (15%) in 2017. Rounded to the nearest whole number, these two percentages were the same, so MJR's strike rate in 2016 proved an excellent guide to the same number in 2017.

AS the table shows, however, there is considerable variance when it comes to the other courses. Anyone regularly using racing statistics with sample-sizes of these magnitudes should be able to learn an immediate lesson. And anyone letting it influence their betting would have been in for a hiding.

If you arrange Table 1 instead by strike rate in 2016, the 24 courses where MJR runners had the highest strike rate in 2016 produced a level-stakes loss of £116.40 even at exchange odds, and only 7 courses of these 24 (29%) saw an individual profit; but the 12 courses where MJR horses had the lowest strike rate in 2016 produced a level stakes profit of £158.73 and 8 of them (67%) saw an individual profit raised. Of course, these numbers themselves are very noisy (subject to randomness) but the point is made: strike rates have little meaning in small samples (another example is so-called 'Trainer Form').

Fortunately, we can judge the efficacy of a trainer, jockey, sire or whatever much better than just considering wins and losses. Alternatives include: percentage of horses placed, percentage of rivals beaten or percentage of horses running to form. Because these numbers have more of what's called 'granularity' – they differentiate more between degrees of good and bad than just the binary cases of wins and losses – they tend to be more stable from one year to the next.

In fact, it's a general rule of thumb in sports statistics that predicting future events which are relatively rare – like wins in racing or goals in football – is better done with other statistics that happen more frequently, like places in racing or shots in football – so long as the two statistics are correlated with one another.

Course	2016 W	2016 R	2016 SR	2017 W	2017 R	2017 SR	SR diff
Brighton	0	17	0	7	19	37	37
Wetherby	1	3	33	0	5	0	33
Leicester	1	26	4	11	34	32	29
Doncaster	4	53	8	6	23	26	19
Yarmouth	4	13	31	2	16	13	18
Lingfield	9	86	10	18	65	28	17
Sandown	2	27	7	5	21	24	16
Ayr	4	18	22	2	26	8	15
Pontefract	12	42	29	8	51	16	13
Thirsk	0	18	0	1	8	13	13
Windsor	4	15	27	1	7	14	12
Ripon	11	47	23	6	52	12	12
Catterick	3	31	10	7	33	21	12
Redcar	6	24	25	4	29	14	11
Epsom	3	24	13	7	32	22	9
Chepstow	1	5	20	1	9	11	9
Salisbury	1	6	17	2	8	25	8
Chester	5	48	10	9	52	17	7
Southwell	1	20	5	2	17	12	7
Newbury	3	18	17	3	13	23	6
Musselburgh	11	53	21	8	53	15	6
Bath	0	14	0	1	18	6	6
York	3	63	5	5	51	10	5
Carlisle	6	31	19	4	26	15	4
Kempton	11	65	17	8	60	13	4
Newcastle	8	64	13	12	77	16	3
Wolver'ton	16	102	16	10	79	13	3
Nottingham	2	15	13	3	28	11	3
Ascot	4	51	8	5	48	10	3
Hamilton	5	36	14	6	52	12	2
Goodwood	9	65	14	8	52	15	2
Chelmsford	13	93	14	10	80	13	1
Newmarket	15	103	15	15	98	15	1
Beverley	9	62	15	9	59	15	1
Haydock	8	50	16	7	42	17	1
Ffos Las	0	5	0	0	4	0	0

Table 1: record of MJR runners by venue for the British Flat seasons 2016 and 2017

Premises

Want to judge a trainer's win percentage in future? Don't use his or her win percentage in the past; it takes so long to reach significance that a 'regime shift' may have taken place – the trainer may have received a fresh intake of horses or moved to new premises or signed up a new jockey, all of which will render the past a dubious guide to the future.

Instead, use the percentage of places. It is less noisy and it stabilises more quickly.

Another example: when judging a stallion or a mare, for instance, an excellent metric is the percentage of horses rated 100+ who have been produced, which are less rare than horses rated 115+ and more predictive of high-class horses than the frequency of high-class horses themselves.

A useful analogue is to think about the chance of being struck by lightning. It would be madness to judge the probability by plotting all the points on the Earth's surface where lightning has struck before, when more frequently occurring meteorological data (black clouds, storms, pressure systems) is available.

This idea is simple and extremely powerful. Humans just love to see patterns in small samples of data which really do not exist; we are so desperate to find **signal** that we tend to be fooled by **noise**. ■