Research on information propagation analyzing odds in horse racing

Shingo Ohta[1] *–Masato Yamanaka*[2] *–Joe Sato*[3] *–Takaaki Nomura*[4]

We focus on the odds in horse racing to study the information propagation to get an idea for fluctuation on information propagation. Analyzing past data of horse racing, and constructing a mathematical model of winning probability, we find a correlation between odds and results of races.
As a first step, we focus on two ways of betting called „WIN" and „EXACTA." We confirm that the winning probability derived from odds of each horse is mostly in accord with actual data. Then, comparing the result with a stochastic model of winning probability constructed with EXACTA odds, we find out fluctuations between them. Finally, we consider where is the origin of the fluctuations.

Keywords: information propagation, stochastic model, horse racing

1. Motivation and Introduction

Our motivation is to find a "better way" to share information. Especially, we focus on the case that people share the information. The "better way" means the way to convey information more correctly, more rapidly. To find it, we must take into account the "fluctuations" caused by human errors, since the "fluctuations" cause information errors. Therefore the information considered in this research should be formed by numerous people. In this sense, the horse racing is desirable to analyze.

[1] Shingo Ohta, Bachelor of Science, Ph.D student, Saitama University, Faculty of Science, Japan.
[2] Masato Yamanaka, Ph.D, post doctoral fellow, University of Tokyo, Institute for Cosmic Ray Research, Japan.
[3] Joe Sato, Ph.D, Associate Professor, Saitama University, Faculty of Science, Japan.
[4] Takaaki Nomura, Ph.D, post doctoral fellow, Saitama University, Faculty of Science, Japan.

2. The analysis of *WIN* odds

2.1. Odds and WIN

There are several ways to bet in horse racing. For the first step, we focus on *WIN*, which is the simplest way to bet. Before explaining our analysis, we explain briefly "odds" and "*WIN*".

The odds are the values which change according to betting, and dividends are defined by these values. WIN is a way to bet that people guess which horse will win. The odds therefore reflect popularity of each horse. WIN odds of a horse is defined as the ratio between total money bet on a race and the money bet on a horse:

$$\frac{\text{total money bet on a race}}{\text{money bet on a horse}} .$$

2.2. χ^2 test of goodness-of-fit

In the analysis of WIN odds, what we want to know is whether the WIN odds reflect strength of each horse.Then we consider that how well the results reflect the expectations. We define N(O) as the number of horses which have odds O and P(O) as the probability that the horses which have odds O win. Then we consider that "with what probability P(O), N horses with odds O win?"

For this analysis, we construct a stochastic model and test the goodness-of-fit between the theoretical value of the number of winner, N(O)P(O), and the observed value of number of winner, n(O). Note that we can get data of the odds O, N(O) and n(O). To test the goodness-of-fit, we should define the criterion of the fit.

We adopt the χ^2 test of goodness-of-fit for this analysis.We split whole region of odds, give a number to the each odds bin like O_i, and sum up the random variable corresponding to each odds bin. Then we derive the following χ^2 by assuming that N is large enough to satisfy the Stirling's formula. Using this criterion, we test the goodness-of-fit with a significance level:

$$\chi^2 = \sum_{i}^{\#\text{ of bin}} \frac{(n(O_i) - \bar{n}(O_i))^2}{\bar{n}(O_i)(1 - P(O_i))} \ ; \ \bar{n}(O) = N(O)P(O) \ .$$

2.3. Stochastic model

As the last preparation for the analysis, we set a stochastic model of winning probability. We assume that people know a winning probability of horses correctly. Then, it is equal to a share of a bet, namely, it is given by

$$P(O_i) = \frac{0.788}{O_i - 0.1} \ .$$

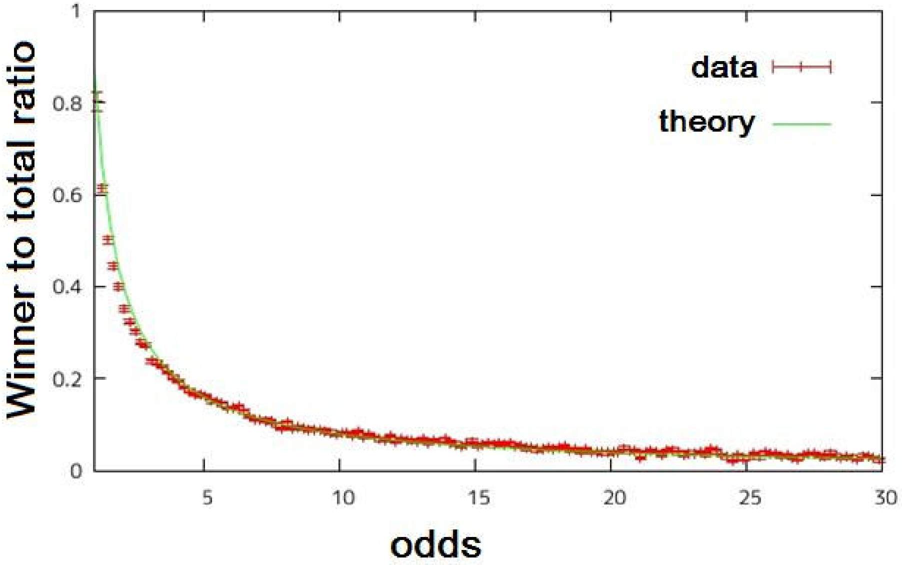
The right hand side is originally a share of a bet derived according to the definition of dividends by Japan Racing Association(JRA).

Note that the width of each odds bin should be wide enough. To analyze more precisely, there should be enough data in each odds bin. In Figure 1, we divided odds region from odds equal 1 to odds equal 31 into 150 bins. Then, we draw the theoretical value P(O_i) and plots the actual results n(O_i)/N(O_i), where the horizontal axis means the odds of horses, and the vertical axis means the probability of winning. At a glance, they coincide with each other very clearly.

2.4. Quantitative check of the stochastic model

To test the goodness-of-fit, we show the χ^2 values for each year from 2003 to 2008 with corresponding significance level (Table 1). We thus confirm the good fitting quantitatively.

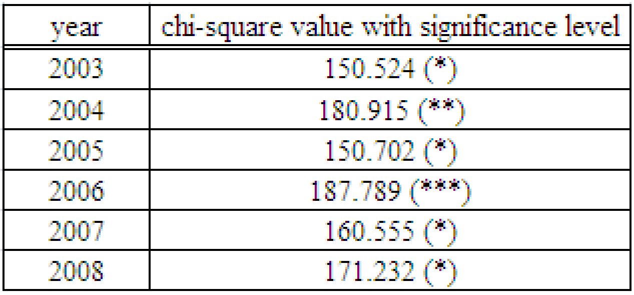
Figure.1 The result of WIN odds analysis(degrees of freedom are 150).



Source: own creation

Most of data plots are in accord with the theoretical line.(JRA official data 1986-2009, free database soft PC KEIBA Database for JRA-VAN Data lab)

Table.1 χ^2 values from data in each year from 2003 through 2008. We apply the usual notations, * for significant at 10%, ** significant at 5%, *** significant at 1%.



year	chi-square value with significance level
2003	150.524 (*)
2004	180.915 (**)
2005	150.702 (*)
2006	187.789 (***)
2007	160.555 (*)
2008	171.232 (*)

Source: own creation

3. Comparison with EXACTA odds

3.1. What is EXACTA

Next, we compare the winning probability constructed from WIN odds with that of EXACTA. This is a way to bet that people guess the winner and the next in order. The reason we choose EXACTA is that the process to guess is almost same as that of WIN. Though the two processes are similar, there may be a little difference between them. For instance, we may expect that even if two horses' WIN odds are small, it is not always true that its EXACTA correspondense.

3.2. Stochastic model and indicator of the "fluctuation"

As we did in the analysis of WIN odds, we construct a stochastic model for the analysis of EXACTA odds. We have to construct a probability that horse α win and β finish next.

One of such a probability is given by WIN odds, O_α and O_β:

$$P(p_\alpha, p_\beta) = P(\alpha)P(\beta|\alpha) = p_\alpha \frac{p_\beta}{1-p_\alpha} \; ; \; p_\sigma = \frac{0.738}{O_\sigma - 0.1} \; (\sigma = \alpha, \; \beta)$$

The right equation is the stochastic model for WIN odds. The p_α is the probability for the horse α win, and the fact that the horse β is the second winner is same as the

fact that the horse β win without horse α. Therefore the probability is given by a conditional probability p_β /(1- p_α), and we get the above equation.
The other probability is given by EXACTA odds, $O_{\alpha\beta}$:

$$P_{\alpha\beta} = \frac{0.738}{O_{\alpha\beta} - 0.1} \ .$$

We assume that “people know the probability that horse α win and the horse β is the next”, as WIN odds. Then, the stochastic model for EXACTA odds is similar with that for WIN odds. The difference is the numerator factor, which comes from the difference of definition between two types of dividends.

Next, to compare these values, we define the indicator of the “fluctuations”. We denote two types of probabilities as follows:

$$P_{\alpha\beta} = \frac{0.738}{O_{\alpha\beta-0.1}} \longrightarrow P^a_{\text{actual}}, \ P(p_\alpha, p_\beta) = p_\alpha \frac{p_\beta}{1 - p_\alpha} \longrightarrow P^a_{\text{theo}}.$$

We give a serial number „*a*” for all conditions of horses α and β from 1 to m. In each race the number of combination is given by m_i and hence $m=\Sigma_{i=1}{}^{M}$ m_i with total M races. Thus the super script „*a*” runs from 1 to m. Then, we define the correlation coefficients as the indicator of the “fluctuations”:

$$r = \frac{\sum_{a=1}^{m}(P^a_{\text{actual}} - \bar{P}_{\text{actual}})(P^a_{\text{theo}} - \bar{P}_{\text{theo}})}{\sqrt{\sum_{a=1}^{m}(P^a_{\text{actual}} - \bar{P}_{\text{actual}})^2}\sqrt{\sum_{a=1}^{m}(P^a_{\text{theo}} - \bar{P}_{\text{theo}})^2}},$$

where

$$\bar{P}_{\text{actual}} = \sum_{a=1}^{m} P^a_{\text{actual}}/m \ , \ \bar{P}_{\text{theo}} = \sum_{a=1}^{m} P^a_{\text{theo}}/m,$$

and also check the slope of regression line for these data,

$$b = \frac{\sum_{a=1}^{m}(P^a_{\text{actual}} - \bar{P}_{\text{actual}})(P^a_{\text{theo}} - \bar{P}_{\text{theo}})}{\sum_{a=1}^{m}(P^a_{\text{actual}} - \bar{P}_{\text{actual}})^2}.$$

3.3. A fluctuation between two kinds of probabilities

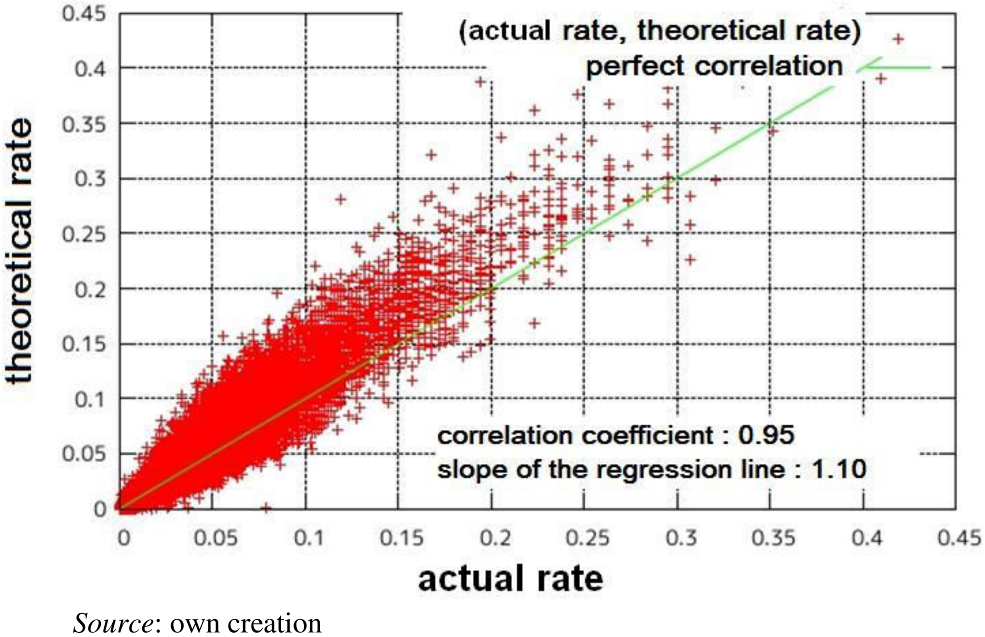
In Figure 2, we plot $(P^a_{\text{actual}}, P^a_{\text{theo}})$ for all races in year 2007. The horizontal axis is the probability given by EXACTA odds and the vertical axis is that of given by WIN odds. Note that if there is a perfect coincidence between them all data points must lie on the green line. We find that the data points are dense along the line and most of them are along the line. Indeed, they have large correlation coefficient, that is, they have quite strong correlation.

However, in whole region, they have a little larger slope than that of the green line, in other words, we find that they have different distributions though they should have same ones because of their propaties. Especially, in the region with high probability, most of the data points are over the green line. In lower probability region, the situation is quite opposite.

Figure 3 is a magnified figure in the region with low probability. Contrast to whole region, the regression line has a little smaller slope than that of green line. This means that the pairs of horses with high EXACTA odds look more attractive than expected by WIN odds, which represent the strength of each horse. This is the fluctuation caused by people to bet who are eager to get much money at one bet.

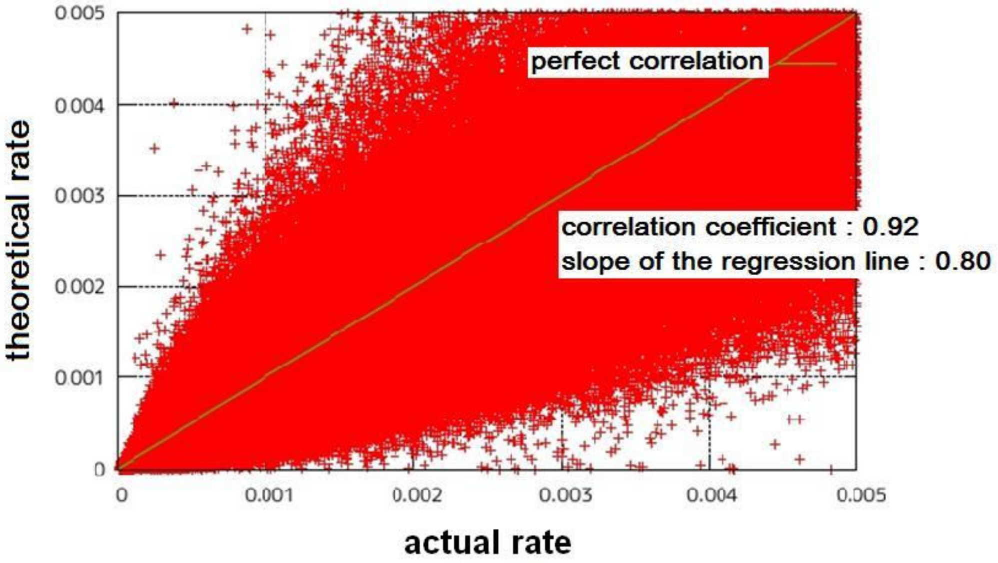
Hence we conclude that there is definitely the fluctuation between WIN odds and EXACTA odds.

Figure 2 The comparison between two kinds of probabilities we saw in section 3.2(for all races in 2007).



Source: own creation

Figure 3 The magnified figure of Fig.2 in low probability region. The slope of regression line is lower than that of perfect correlation line.



Source: own creation

4. The origin of fluctuations

4.1. Stochastic model and Indicator of the "fluctuations"

In the last step, we focus on the origin of the fluctuations. From the previous two analyses, we saw that the expectations in EXACTA deviate from the WIN expectations though they should be same. Thus we investigate where is the origin of the deviations.

For the investigation, we compare two types of probabilities.One is the probability we saw in the WIN odds analysis,

$$P_\alpha = \frac{0.788}{O_\alpha - 0.1} \ ,$$

which contains the information purely about the expectation of winner.
The other is defined as the sum of probabilities for EXACTA,

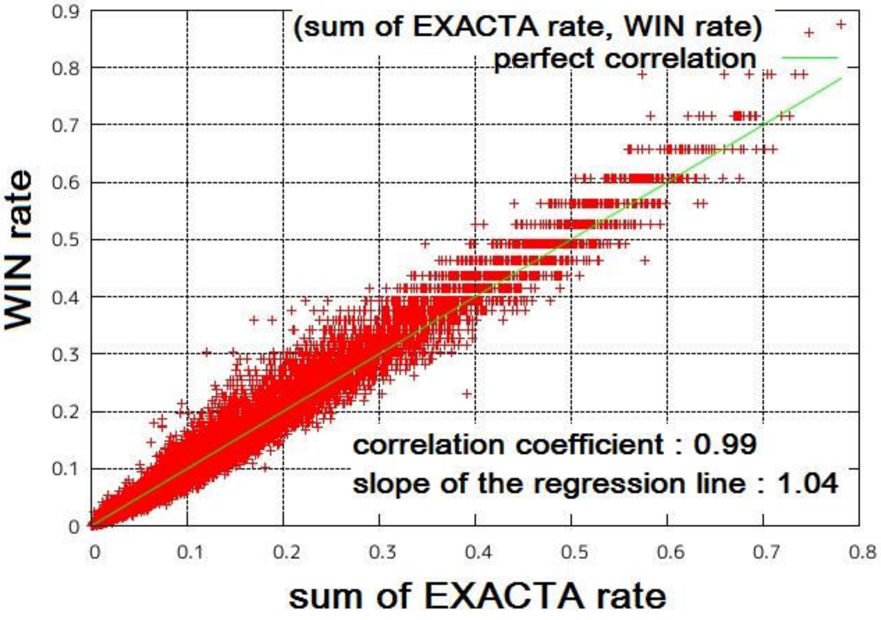
$$P_{1\alpha} = \sum_{\beta \neq \alpha} P_{\alpha\beta} \ ; \ P_{\alpha\beta} = \frac{0.738}{O_{\alpha\beta} - 0.1} \ ,$$

which contains the information of the expectation for the second horse. Then, comparing these two values, we can find out the deviations of the EXACTA odds from the WIN odds. In this analysis, we adopt the correlation coefficient and the slope of regression line as the indicators of deviations as in the previous analysis.

4.2. The Deviation from the probability for WIN odds

In Figure 4, we plot $(P_{1\alpha}, P_\alpha)$ for all races in year 2007. The horizontal axis is the probability given by the EXACTA odds, and the vertical axis is that of given by WIN odds. There is quite strong correlation between these two values. However, as we can see, they are not exactly same each other. Therefore the main origin of the deviation in the previous analysis is seen in this analysis. Thus we conclude that the deviation is in the expectation of the second horse.

Figure 4 The comparison between two kinds of winning probability(for all races in 2007).



Source: own creation

5. Summary

In the first analysis, we confirmed that the result of each race reflects the WIN odds. In the next, we confirmed the existence of the "fluctuations" in EXACTA odds. Then, from these results, we investigated the origine of the deviations of the EXACTA odds from the WIN odds. Finally, we concluded that there is fluctuation in the expectation of the second horse.

References

[1] K. Park and E. Domany, Power law distribution of dividends in horse races, Europhy Lett, 53(4)pp.419-425.

[2] T. Ichinomiya, Physica A, Volume 368, Power-law distribution in Japanese racetrack betting, Issue 1, 1 August 2006, Pages 207-213

[3] D. A. Harville. Assigning Probabilities to the Outcomes of Multi-Entry Competitions, pp.213-217 in Efficiency of racetrack betting markets 2008 edition D. B. Hausch, V. SY Lo and T. Ziemba

[4] Plackett, R. L. (1975). The analysis of permutations. Apl. Statist., 24, 193-202