Asian Journal of Social Science Studies; Vol. 1, No. 2; 2016
ISSN 2424-8517 E-ISSN 2424-9041
Published by July Press

Comparing Post Positions Using the Chi Square Distribution

Roger L Goodwin[1]

[1] 1989 Leetown Road, Summit Point, WV, USA

Correspondence: Roger L Goodwin, 1989 Leetown Road, Summit Point, WV, USA.

Received: June 16, 2016 Accepted: July 4, 2016 Online Published: July 9, 2016

doi:10.20849/ajsss.v1i2.47 URL: http://dx.doi.org/10.20849/ajsss.v1i2.47

Abstract

In horse racing, at the start of a race, horses line-up in starting gates. The rack track assigns horses and their jockeys their starting gates in advance of a race. We examine the published data from a racetrack to determine if a starting gate is preferred over another. We compare the number of wins across the post positions for sprint races and long distance races using two chi-square distributions. Given the post positions and the number of wins for a given race track, it can be determined that the inner-most and outer-most track positions tend to be preferred over the center track positions.

Keywords: starting gate, post position, stakeholders, risk, Bernoulli distribution, normalized statistic, chi-square distribution

1. Introduction

This paper answers the question, should a stakeholder in the outcome of a race prefer one post position over another? This paper examines post position data from the Charles Town Race Track (West Virginia) to determine if a person placing a wager would prefer one post position to another. The post position is the position of the stall in starting gate from which a horse starts. The post position of a racehorse on a flat track is numbered by its position relative to the inside barrier of the track.

We begin this analysis by assuming a reasonable distribution for the given data. The two data sets in Section 2 give the number of starting horses and the number of wins. The Bernoulli distribution seems reasonable for the given data sets. The question remains, do all of the post positions have the same distribution?

This paper presents a chi-square distribution for the wins at Charles Town Race Track for the post positions across all of their race tracks for the time January 1, 2014 to May 23, 2014.

2. The Data

On May 24, 2014 the Charles Town Race Track published the data in Table 1 for *sprint races.* The table contains the number of races, the number of wins, and the percentage number of wins. The percentage is simply a numeric calculation based on the previous two columns. For the percentage number of wins, Charles Town Races rounded their table entries to one decimal to the right of the decimal point. For analytic purposes, we need more precision. Four decimal places should do.

Table 1. Sprint Race Data January 1 to May 23, 2014

Post Position	Starts	Number Wins	Percent Wins
1	**587**	**95**	**0.161840**
2	**587**	**76**	**0.129472**
3	**587**	**81**	**0.137990**
4	**587**	**71**	**0.120954**
5	**583**	**81**	**0.138937**
6	**559**	**62**	**0.110912**
7	**465**	**64**	**0.137634**
8	**339**	**32**	**0.094395**
9	**215**	**21**	**0.097674**
10	**100**	**6**	**0.060000**

Similarly, on May 24, 2014 the Charles Town Race Track published the data in Table 2 for *distance races*. The table contains the number of races, the number of wins, and the percentage number of wins. The percentage is simply a numeric calculation based on the previous two columns. For the percentage number of wins, Charles Town Races rounded their table entries to one decimal to the right of the decimal point. For analytic purposes, we need more precision. Four decimal places should do.

Table 2. Distance Race Data January 1 to May 23, 2014

Post Position	Starts	Number Wins	Percent Wins
1	**89**	**16**	**0.179775**
2	**89**	**13**	**0.146067**
3	**89**	**12**	**0.134831**
4	**89**	**10**	**0.112360**
5	**89**	**9**	**0.101124**
6	**89**	**13**	**0.146067**
7	**73**	**11**	**0.150685**
8	**54**	**3**	**0.055556**
9	**32**	**1**	**0.031250**
10	**18**	**1**	**0.055556**

Given the data in Tables 1 and 2, should a stakeholder in the outcome of a race prefer one post position to another? Stakeholders include jockeys, racehorse owners, and people placing a wager.

3. Wagers

Charles Town Races separates their races into two categories: 1) sprint races and 2) distance races. A person can place a wager on either type of race in the following ways:

- Win --- Your horse must win.
- Place --- Your horse must finish first or second.
- Show --- Your horse must finish first, second, or third.
- Across the board --- Wager the same amount on one horse to win, place, and show.
- Exacta --- Pick the first two horses in exact order to finish.
- Trifecta --- Pick the first three horses in exact order to finish.
- Superfecta --- Pick the first four horses in exact order to finish.
- Daily Double --- Pick the winners of two consecutive races.
- Pick Three --- Pick the winners of three consecutive races.
- Pick Four --- Pick the winners of four consecutive races.

Given the above phrases, assume that the following phrases have the same meaning (excepting plurality).

- Must win
- Must finish first
- Pick the winners

Other phrases suggest ways that a person can win a wager. However, those are the most commonly used. A horse can only "win" the race by crossing the finish line first without, somehow, being disqualified.

Other factors can account for the outcome of a horse race. Those other factors include racetrack number, the jockey weight, type of horses racing, race track conditions, horse gender, and so on.

4. The Models

We define the Bernoulli random variable in Equation (1) as a starting basis for comparing the starting post positions of the horses.

$$X = \begin{cases} 0, & \text{If horse does not win a race.} \\ 1, & \text{If horse wins a race.} \end{cases} \quad (1)$$

We add the index *i* to denote the post position *i = 1, 2, ..., 10* since these are ten Bernoulli random variables and ten binomial distributions each with their own expectations.

$$X_i = \begin{cases} 0, & \text{If horse does not win a race from post position } i. \\ 1, & \text{If horse wins a race from post position } i. \end{cases} \quad (2)$$

for the i^{th} post position *i=1, 2, ...,10.* Tables 1 and 2 give the probabilities for each post position in the right-most column for the Bernoulli random variables in Equation (2).

$$\frac{X_i}{P(X_i = x_i)} \Bigg| \frac{0 \quad 1}{p_i \quad 1 - p_i}$$

Summing the Bernoulli random variables in Equation (2) gives ten binomial distributions for sprint races and ten binomial distributions for distance races. From this we can test to see if each of the distributions are the same or if one post position is preferred over another (vise-versa one post position is shunned over another).

(Hogg and Tanis 1993, pages 511-521; Hogg and Craig 1995, pages 116-123) present a multinomial probability distribution. The multinomial distribution has the following assumptions:

- The experiment has *k* possible outcomes that are mutually exclusive and exhaustive, say $A_1, A_2, ..., A_k$.
- *n* independent trials of this experiment are observed.
- The random variable X_i is equal to the number of times A_i occurs in the *n* trials, *i = 1, 2, ..., k.*

Horses can run under multiple post-positions. This is because the given data is over a six-month period. Not that the same horse started in two or more different gates in the same race, it is the case that the same horse started in multiple gates in multiple races over the six-month period.

Since we cannot guarantee that the probabilities sum to 1, we rule out the multinomial distribution as a plausible model. We can still use a chi-square distribution to compare the binomial distributions.

We can model the data using the quadratic formula. First, we normalize the test statistic, and then square it. This gives a $\chi^2(1)$ distribution. The quadratic theorem allows us to add *r* independent $\chi^2(1)$ distributions by squaring the normalized test statistic. The quadratic theorem will be demonstrated next.

5. Sprint Races

Under a fair assignment of the starting post positions, we would expect the percentage of wins to be one-tenth for each post. However, this does not take into account the number of races run. A more accurate measure would take into account the number of races run. We obtain the fair, self-weights p'_i by dividing the number of starts for each post position *i* by the total of all start positions 4,609.

Under the null hypothesis, we have ten binomial distributions $b_{H0}(X'_i, n_i, p'_i)$. Under the alternative hypothesis, we have ten binomial distributions $b_{H1}(X_i, n_i, p_i)$. In the statistical model, we wish to test the probabilities p'_i under H_0 against those probabilities under H_1. We calculate the probabilities p'_i under the null hypothesis H_0 as follow:

$$p'_i = \frac{n_i}{\sum_{i=1}^{10} n_i} = \frac{n_i}{n}$$

The expected values under the first model are simply the fair, self-weights times the number of races. The expected values must be integers since they represent the number of wins or the number of horses that crossed the finish line first. Equation (3) gives the expected number of wins for each post position *i* under the fair, self-weighting model for sprint races.

$$E(X'_i = x'_i) = n_i p'_i, i = 1,2, \ldots ,10 \quad (3)$$

To test that these ten binomial distributions have the same distribution, we use the chi-square distribution with 10 degrees of freedom.

(Hogg and Craig 1995, page 249) discuss random sampling from a distribution that is binomial *b(1, p).* (Hogg and Craig 1995, page 481-485) discuss quadratic forms of random variables, the chi-square distribution and the degrees of freedom. We normalize ten binomial test statistics. Each normalized test statistic is approximately *N(0,1).* The distribution of the square of a normalized test statistic is $\chi^2(1)$. The quadratic theorem allows us to add these ten statistics to obtain the distribution $\chi^2(10)$.

Equation (4) gives the test statistic.

$$Q_{10} = \sum_{i=1}^{10} \frac{\left(X_i - n_i p_i'\right)^2}{n_i p_i'(1-p_i')} \approx \chi_\alpha^2(10) \quad (4)$$

where α is the desired significance level of the test. Reject the null hypothesis if $Q_{10} \geq \chi^2_\alpha(10)$. We arbitrarily set α = 0.05. Using the data in Table 1, we test

$$H_0: p_i' = \frac{n_i}{\sum_{i=1}^{10} n_i}, i = 1, 2, \ldots, 10.$$

versus

$$H_1: p_i = \frac{X_i}{n_i}, i = 1, 2, \ldots, 10.$$

The critical value for the hypothesis test is $\chi^2_{\alpha=0.05}(10) = 18.3$. Since $Q_{10} = 37.08 \geq 18.3$, we reject the null hypothesis. Post positions are significant in the sprint races. The next section will determine which post-positions a stakeholder prefers.

5.1 Sprint Race Post Position Analysis

Table 3 shows the individual chi-square tests for each post position. Which post positions are preferred? The cut-off value for a chi-square test with one degree of freedom is $\chi^2_{0.05}(1) = 3.84$.

Table 3. Post Position Chi-Square Tests for Sprint Races

Post Position	Starts n_i	Number Wins X_i	Percent Wins p_i	Self Weighted p'_i	Q_i	Comment
1	587	95	0.161840	0.1274	6.279	Reject
2	587	76	0.129472	0.1274	0.024	Accept
3	587	81	0.137990	0.1274	0.597	Accept
4	587	71	0.120954	0.1274	0.217	Accept
5	583	81	0.138937	0.1265	0.817	Accept
6	559	62	0.110912	0.1213	0.564	Accept
7	465	64	0.137634	0.1009	6.921	Reject
8	339	32	0.094395	0.0736	2.161	Accept
9	215	21	0.097674	0.0466	12.588	Reject
10	100	6	0.060000	0.0217	6.912	Reject
	4,609		1.19	1.0	37.080	

Horses starting from post positions 1, 7, 9, and 10 are preferred in sprint races.

6. Distance Races

We perform a similar analysis as in Section 5.

Table 4. Post Position Chi-Square Tests for Distance Races

Post Position	Starts n_i	Number Wins Y_i	Percent Wins p_i	Self Weighted p'_i	Q_i	Comment
1	89	16	0.179775	0.1252	2.423	Accept
2	89	13	0.146067	0.1252	0.355	Accept

3	89	12	0.134831	0.1252	0.076	Accept
4	89	10	0.112360	0.1252	0.133	Accept
5	89	9	0.101124	0.1252	0.470	Accept
6	89	13	0.146067	0.1252	0.355	Accept
7	73	11	0.150685	0.1027	1.827	Accept
8	54	3	0.055556	0.0759	0.320	Accept
9	32	1	0.031250	0.0450	0.141	Accept
10	18	1	0.055556	0.0253	0.667	Accept
	711		1.11	1.0	6.77	

Equation (5) gives the expected numbers of wins for each post position *i* under the fair, self-weighted model for distance races.

$$E(Y_i' = y_i') = n_i p_i', i = 1,2, \dots ,10 \quad (5)$$

Equation (6) gives the test statistic for comparing the distributions for the long distance races of post positions.

$$Q_{10} = \sum_{i=1}^{10} \frac{(Y_i - n_i p_i')^2}{n_i p_i'(1-p_i')} \approx \chi_\alpha^2(10) \quad (6)$$

Using the data in Table 4, we test

$$H_0: p_i' = \frac{n_i}{\sum_{i=1}^{10} n_i}, i = 1, 2, \dots, 10.$$

versus

$$H_1: p_i = \frac{Y_i}{n_i}, i = 1, 2, \dots, 10.$$

The critical region for $\chi^2_{\alpha=0.05}(10) = 18.3$. Since $Q_{10} = 6.77 \leq 18.3$, we accept the null hypothesis. Post positions are not important in distance races.

7. Concluding Remarks

We developed statistical models for both sprint horse races and distance horse races. The multinomial distribution could not be used to model the data because the required assumptions did not hold true. We used ten normalized binomial distributions to fit the data. Each normalized test statistic has a chi-square distribution with one degree of freedom.

The models show that for sprint races the innermost and outer-most track positions tend to be preferred over the middle track positions. For the distance races, track positions are not a significant variable towards winning.

Acknowledgements

Equibase Company LLC (2014) published the data in Tables 1 and 2.

References

Hollywood Casino at Charles Town Races. *Racing Program*, Saturday, May 24, 2014.

Hogg, R. V., & Craig, A. T. (1995). *Introduction to Mathematical Statistics* (5th ed.). Upper Saddle River, New Jersey, Prentice Hall.

Hogg, R. V., & Tanis, E. A. (1993). *Probability and Statistical Inference* (5th ed.). Englewood Cliffs, New Jersey, Prentice Hall.