The Economics Of The Duration Of The Baseball World Series

Alexander E. Cassuto, (E-mail: alexander.cassuto@csueastbay.edu), California State University, Hayward
Franklin Lowenthal, (E-mail: frabklin.lowenthal@csueastbay.edu), California State University, Hayward

ABSTRACT

This note examines some statistical features of the major league baseball World Series. We show that, based upon actual historical data, we cannot reject the hypothesis that the two World Series teams are evenly matched. Yet, we can also calculate the relative strengths of the teams that would best match the actual outcomes, and we find that those relative strengths are not equal. Including the home field advantage in the calculations indicates that the differential in relative strength between the competing teams can be explained by this advantage.

We present the relative team strengths that would maximize the probability of four, five, six and seven game series. We find that a six or seven game series is most likely when the two teams are evenly matched, a four game series is most likely when the probability of the stronger team winning is one, while the probability of a five game series is maximized if one team has a relative strength of 0.789. We also show that, on average, the expected number of World Series games will be between 4 and 5.81, depending upon the relative strengths of the teams and the home field advantage. Contracts that do not consider the likelihood of less than seven game Series will create windfall gains to MLB and marginal economic losses to broadcasters.

INTRODUCTION

Relative team strength is a factor determining the number of games played in baseball’s World Series. This paper examines some statistical aspects of expected outcomes using the binomial probability distribution and data from the initial World Series in 1903 to the 2003 Marlins-Yankees Series. By extending the analysis we can also determine the revealed relative strength of World Series teams based upon the actual outcome of the ninety-three World Series played as best of seven series. The four World Series played as best of nine (1903, 1919-1921) were not included in the results and games played to ties (1907, 1912, 1922) were not considered although the actual World Series results were included.

DATA AND ANALYSIS

The formula: \[ P(r) = \frac{n!}{r!(n-r)!} p^r q^{n-r} \], where \( r \) is the number of successes (in this case 4), \( n \) the number of trials (games played) and \( p \) the probability of one team winning a game (\( q=1-p \)), describes the probabilities of all possible outcomes of a World Series. There are a total of seventy possible outcomes, not \( 2^7 = 128 \). To see this outcome recall that for team A (American League) to win in \( M \) games they must win exactly 3 of the first \( M-1 \) games and the \( M^{th} \) game. There is only one way that team A can win in 4 games while there are \( 4!/3!1! = 4 \) ways to win in 5 games, \( 5!/3!2! = 10 \) ways to win in 6 games and \( 6!/3!3! = 20 \) ways to win in 7 games. The number of ways that team N (National League) can win is the same (although one of the authors believes that team Y (Yankees) should always win).

The probabilities for each outcome are presented below assuming each team has an equal probability of winning each game (\( p=q \)).
Examining the actual results from all the World Series that have been played we find that seventeen series were decided in four games, twenty-one were decided in five games, twenty-two were decided six games while thirty five went the full seven games.

The theoretical outcomes, based upon the probabilities in Table 1 are 11.875 for a four game series, 23.75 for a five game series and 29.6875 for both six and seven game series. Using the null hypothesis that the World Series teams are evenly matched the computed value of Chi-Square is routinely found to be 5.47. Using the $\chi^2$ test, at 0.05 permissible type one error and three degrees of freedom, we do not reject this hypothesis.

One of the more interesting aspects of this analysis emerges as we minimize the computed $\chi^2$ in order to find the revealed relative strengths of the teams given the actual results. In this sense we are finding the expected probabilities that best match the actual outcome over all World Series played. Although we cannot reject the null hypothesis of equality between the teams, we find, instead, that the probabilities that minimizes $\chi^2$ at a value of 5.08 are approximately $p = 0.569$ and $q = 0.431$. Based on actual outcomes utilizing the $\chi^2$ the two teams entering the series do not have the same probabilities of emerging victorious in terms of the revealed team strengths.

The minimum Chi-Square can be found by calculus; the detailed calculations are extremely lengthy so we instead used the SOLVER feature in EXCEL. A word of caution is in order. SOLVER finds only a local minimum, one that depends upon the initial value of the independent variable, here the probability of the superior team winning a given game. However, we believe that we have made a sufficient number of different initial values to justify our claim that this is a global minimum.

One of the most surprising aspects to the actual outcome data is that while Table 1 indicates that if the two teams are equally matched, six and seven game series are equally probable, there have been only 22 six game series but 35 seven game series. Under the assumption that the teams are evenly matched we compute the probability that of the 57 series that lasted more than 5 games, 35 or more would last the full 7 games to be only .0427. Here we used the normal distribution as an approximation to the binomial distribution. However, if we assume that a two-tail test is appropriate the probability doubles to .0854. For a one-tail test we can reject the null hypothesis at the 95% confidence level while for a two-tail test we cannot reject it at the 95% confidence level, but do reject it at the 90% confidence level. Clearly the probability of a series that is not decided in 5 games ending with the sixth game should increase if the teams are not evenly matched i.e. if $p$ is greater than .50. In fact, the probability that the series will end in 6 games given that the series stands at 3 to 2 is given by $p^2 + (1 - p)^2$. This is because the probability that the “better” team is ahead is $p$ (it won the extra game) and their probability of winning the 6th game is $p$. Likewise, the “worse” team has a probability of leading of $(1-p)$ and the same probability of $(1-p)$ of winning the 6th game. The function $p^2 + (1 - p)^2$ has a minimum at $p = .5$ and is, of course, maximum at the end points. If we use the value of $p = .569$ which minimizes the chi-square statistic, we find the probability of the series ending with the 6th game when one team leads 3 to 2 is .510. Under this assumption, we compute the probability that of the 57 series that lasted more than 5 games, 35 or more would last the full 7 games to be only .0307 . However, if we assume that a two-tail test is appropriate the probability doubles to .0614. For a one-tail test we can reject the null hypothesis at the 95% confidence level while for a two-tail test we can reject it at the 90% confidence level, but not at the 95% confidence level. We offer a non-statistical explanation for this statistical anomaly. The team that is behind must play to win game at all costs; thus it may change its rotation to start its best pitcher, use its best reliever for more innings than usual etc., while the team that is ahead will formulate its strategy so as to win one more game, not necessarily the sixth game.
Finally we can use the theoretical outcomes to see what relative team strengths are most likely to bring about each possible outcome in terms of games played. In other words, for what value of p are the probabilities in Table 1 above maximized? Our results are summarized in Table 2 below.

<table>
<thead>
<tr>
<th>Number of Games in Series</th>
<th>Probability that Maximizes</th>
<th>Maximum Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 games</td>
<td>p = 1 or p = 0</td>
<td>1.000</td>
</tr>
<tr>
<td>5 games</td>
<td>p = .789 or p = .211</td>
<td>.3333</td>
</tr>
<tr>
<td>6 games</td>
<td>p = .50</td>
<td>.3125</td>
</tr>
<tr>
<td>7 games</td>
<td>p = .50</td>
<td>.3125</td>
</tr>
</tbody>
</table>

The four game outcome is obvious; a relative team strength of 1 and 0 will bring about a probability of 100% that the series will end in four games. We could also verify this by examining the function

\[ f(p) = p^4 + (1-p)^4 \]

that represents the probability that one of the two teams wins all 4 games (p is the probability that a given team wins one game—we assume throughout that this probability does not change from game to game). The maximum occurs at the endpoints 0 and 1; to find the minimum, compute the first derivative

\[ df/dp = 4p^3 - 4(1-p)^3 = 8p^3 - 12p^2 + 12p - 4 = 4(2p-1)(p^2 - p +1) \]

This is zero if and only if p = ½ since the quadratic factor has no real root. The minimum value for the probability of a sweep is the value 1/8 = .125 given in Table 1. According to Table 2 if the relative team strength is 0.789 and 0.211 the probability of a five game series is maximized at 33.3%. To verify this we must examine the function

\[ f(p) = 4p^4 + 4p(1-p)^4 \]

which represents the probability that one of the two teams wins a 5 game series (p is the probability that a given team wins one game). The minimum of zero occurs at the endpoints 0 and 1 (a 5 game series is impossible because there is a sweep); to find the maximum, compute the first derivative

\[ df/dp = 16(p^3 - 4p^4 + 4(1-p)^3 - 16p(1-p)^3 = 4(2p-1)(-6p^2 +6p - 1) \]

This is zero if and only if either p = ½ or the quadratic factor is zero. These roots are just ½ + sqrt(12) /12 = .788675 and ½ - sqrt(12)/12 = .211325. It is easy to see that p = ½ leads to the relative minimum value of .25 listed in Table 1 while the solutions of the quadratic lead to the absolute maximum value of 1/3. The relative minimum at p = ½ is what we would expect since for evenly matched teams it is likely that the series will last at least 6 games. To summarize, at the endpoints p = 0 and p = 1, a sweep is certain; as the probability of a sweep decrease the probability of a 5 game series increases, reaching a maximum of 1/3 at the symmetric points p = .789 and p = .211. But then as the teams become still more evenly matched the probability steadily decrease to the relative minimum value of ¼ achieved at p = ½.

In the recent World Series between the New York Yankees and the Arizona Diamondbacks a sports columnist predicted a five game series with the Yankees as winners. When is the probability of a specified team winning in 5 games maximized? This is a slightly different problem than the one we addressed above. The probability of such an event is represented by only the first term of f(p), i.e. 4(p^4)(1-p). The derivative of this function is routinely found to be zero at p = 4/5 = .80. In that case we obtain the maximum probability of a 5 game Yankee win (which sadly did not occur) to be (4/5)^5 = .32768; interestingly this is only slightly less than the maximum probability of 1/3 of either the Yankees or the Snakes winning in 5 games. The explanation of why there is so little difference between
these maximum probabilities of .3333… and .32768 is that at p = .8 the probability of the weaker team winning in 5 games is just .00512.

An equal relative strength (p = 0.5) will bring about the highest probability of both a seven game series as well as a six game series with probabilities are listed in Table 1. For a six game series the function that we wish to maximize is

\[ f(p) = 10(p^4)(1-p)^2 + 10(p^2)(1-p)^4 \]

To find the maximum, compute the first derivative and simplify to obtain the 5th degree polynomial

\[ \frac{df}{dp} = 20p(1-p)(-6p^3 + 9p^2 - 5p + 1) \]

This is zero if and only if either p = 1 or p = 0 or the cubic factor is zero. The cubic factors into the product \((2p-1)(-3p^2 + 3p - 1);\) the first factor yields the root \(p = \frac{1}{2}\) which gives the absolute maximum probability of .3125 listed in Table 1 while the quadratic factor is irreducible.

For a seven game series the function that we wish to maximize is

\[ f(p) = 20(p^4)(1-p)^3 + 20(p^3)(1-p)^4 = 20p^3(1-p)^3 \]

To find the maximum, compute the first derivative and simplify to obtain the 5th degree polynomial

\[ \frac{df}{dp} = 60p^2(1-p)^3 - 60p^3(1-p)^2 = 60p^2(1-p)^2(1-2p) \]

This is zero if and only if either p = 1 or p = 0 or \(p = 1/2\). Again the root \(p = \frac{1}{2}\) gives the absolute maximum probability of .3125 listed in Table 1.

**HOME FIELD ADVANTAGE**

Home field advantage may play an important part in this analysis. For the purpose of this paper we define home field advantage as the advantage gained by the team playing at home for the first game of the Series. There are at least two alternative definitions of home field advantage. One is that there is only home field advantage if there is a final deciding seventh game (ninth game in those World Series that were best 5 out of 9). There have been 35 such series with the home team winning the final deciding game 18 times and the visiting team winning 17 times. Clearly this is not statistically significant. However, an interesting fact is that the last 8 such series were all won by the home team, so perhaps there is now a home field advantage due to more rabid fans e.g. the Angels rally monkey. Another definition of home field advantage is that it exists in any series with an odd number of games. The advantage belongs to the team with more home games (not necessarily the same as the team that is home for the final game). There has been a total of 57 such World Series with the home team winning only 27 of them (less than half) while the visiting team has won 30 times.

Now we return to our operational definition of home field advantage, namely the team with the first game at home. This team should have the advantage of getting off to a good start, even if the actual series ends up with both teams having an equal number of home games.

A perusal of the actual outcomes shows that the team with the home field advantage ultimately won the World Series 53 times while the visiting team won 42 times, for a winning percentage of 55.8%. We can test the statistical significance of this result at the .05 or 5% significance level. A binomial distribution with parameter .50 and 95 (42+53) may be approximated by a normal distribution with mean 47.5 and standard deviation 4.87. If we assume that there certainly is not a home field disadvantage, a one tailed test is appropriate. The probability that, under the null hypothesis of no home field advantage (and that the teams are evenly matched), our test statistic \(\frac{53 - 47.5}{4.87} = 1.13\) equals about .13 or 13%. As this exceeds the 5% significance level or tolerable type one error, we cannot
reject the null hypothesis that there is no home field advantage in the World Series. Further if we exclude series were both teams were from the same geographical area (Yankees-Giants, Yankees-Dodgers, Cubs-White Sox, Browns-Cardinals, Giants-A’s) the outcomes change to 42 Series where the team with the home field advantage won and 37 where it lost, the home field percentage drops to 53.2%. Counting all games played (but excluding ties) the home team won 304 games while the visiting team won 246, a home field percentage of 55.3%. Again we can test the statistical significance of these figures. This time the approximating normal distribution has a mean of 275 and a standard deviation of 11.72. The probability that, under the null hypothesis of no home field advantage (and that the teams are evenly matched), our test statistic $t = (304 - 275) / 11.72 = 2.47$ equals about .0068 or 0.68%. As this is less than the 5% significance level or tolerable type one error (in fact, it is less than 1%), we can reject the null hypothesis that there is no home field advantage in the World Series in an individual game. We are left with a paradox: while there is a home field advantage in an individual game there is no home field advantage in the World Series as a whole. If we remove the games played in World Series where the opposing teams were from the same geographical area the results change to 256 won, 202 lost and a home field victory percentage of 56.0. This leads to a statistic of 2.52 and the same conclusion that there is a home field advantage in an individual game.

If the actual home field advantage is added into the analysis the two teams entering the World Series are actually evenly matched, but one gains approximately a 5%-6% higher probability of winning by virtue of its home field advantage. This would account for the differences revealed by the $\chi^2$ analysis from the previous section.

The home field advantage seems to have changed markedly after WWII in terms of game victories, but not of series victories. For all World Series games played up to and including 1941 the home team had won 95 games while the visiting team won 91 (51.1%). The percentages change dramatically after 1945; the home team has won 201 games while losing 140 (58.9%) since then. These results, however, are not mirrored in the World Series outcomes: pre WWII, the team with the home field advantage won 19 of 34 (55.8%), post WWII the team with the home field advantage won 30 of 57 (52.6%).

As a further note on home field advantage the initial World Series were not standardized. The 1905 Series featured the following home team sequence: AL, NL, AL, NL, NL. In 1906 the sequence was NL, AL, NL, AL, NL, AL, while 1907 had NL (tie), NL, NL, AL, AL. The seven game series (1909, 1912) gave 4 home games to the team that did not have the opening game (NL, NL, AL, AL, NL, AL, AL). The Series of 1918, played during WWI, had NL, NL, NL, AL, AL, AL. The sequence we have currently began with the 1924 World Series between the Senators and the Giants, abandoned temporarily during the war years 1942-45, and resumed permanently thereafter. Leagues alternated initial home fields from 1905 onward.

**HOME FIELD DETERMINATION**

The current proposal, advanced by the Commissioner, awarded home field advantage, and the commensurate higher probability of winning the World Series, to the league that wins the All-Star game. The effect on all-star voting remains to be seen. Imagine fans from the National League voting for the strongest players in their League on their all-star ballots, while voting for the weakest all-stars on the American League ballot. All-star managers have limited possibilities to correct fan voting of position players; their selection strength is in the choice of pitchers. It may well be the case that the league with the strongest individual players gains a long run advantage in World Series outcomes through home field advantage. Many will argue that this is an untenable situation. After all the American League has won the last seven all-star games played to a conclusion, and won 12 of the first 16 games. Having a World Series home field advantage determined on the basis of All-star game victories may leave one league out in the cold for many consecutive years. This does not seem equitable in view of the demonstrated advantages of home field advantage in World Series play.

**CONTRACTS AND OUTCOMES**

Major League Baseball has a long history of contractual relationships with the broadcast medium. The first national broadcast of the World Series occurred in 1922 and no fees were received by MLB for the rights. The first rights were given to the Ford Motor Company in 1934 for $100,000. The current contract calls for $2.5 billion
payments by Fox to MLB for the rights to World Series, league and divisional championships, regular season and All-Star games for six years, an average of $417 million per year.

Clearly the rights include an estimate of the value of each of the components since past contracts between MLB and broadcasters have often divided the four components listed above among more than one broadcaster. In fact, from 1976 on, at least two networks were broadcasting national baseball games. The World Series is the highlight of the season and carries with it the highest per game value in terms of advertising fees. Based upon past contracts the authors estimate current fee for the World Series broadcast right to be approximately $115 million for a per game average of approximately $15 million. The average assumes that the value of all games is equal, an assumption that is dubious at best. Certainly the seventh game of a World Series played on a Sunday night in prime time is worth considerably more than any of the other games.

Table 3 shows the relationship between the strength of the stronger team and the expected number of games to be played in the World Series. If the relative team strengths are 1 and 0 then the stronger team will sweep the series. If the balance of strength narrows somewhat to 0.75 and 0.25, an average of 5.163 games will be played in the World Series. As we have shown above, if the teams are of equal relative strength, 0.5 and 0.5 we would expect the greatest number of games, an average of 5.81, until one team won four. If the networks are estimating a value of $15 million per game and expect a seven game World Series the networks winning the bid would lose almost $20 million, on average, even if the teams were equally matched. A four game sweep would cost the broadcaster $45 million unless some provision for lowered payments were to be included in the contract.

<table>
<thead>
<tr>
<th>Expected Number Of Games</th>
<th>Relative Strength Of Strongest Team Including Home Field Advantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1.00</td>
</tr>
<tr>
<td>4.21</td>
<td>0.95</td>
</tr>
<tr>
<td>4.439</td>
<td>0.90</td>
</tr>
<tr>
<td>4.68</td>
<td>0.85</td>
</tr>
<tr>
<td>4.927</td>
<td>0.80</td>
</tr>
<tr>
<td>5.163</td>
<td>0.75</td>
</tr>
<tr>
<td>5.378</td>
<td>0.70</td>
</tr>
<tr>
<td>5.56</td>
<td>0.65</td>
</tr>
<tr>
<td>5.7</td>
<td>0.60</td>
</tr>
<tr>
<td>5.78</td>
<td>0.55</td>
</tr>
<tr>
<td>5.81</td>
<td>0.50</td>
</tr>
</tbody>
</table>

One example of the effect of short series on earnings was reported in CNNSI.com [2]. In 1998 the Yankees swept the Padres in four games. “Fox never had the benefit of a drama-building long series and ended up losing an estimated $15 million because of the sweep. The network needed a five-game series to break even.” Since the contract between Fox and MLB was substantially lower in 1998 the estimated figures cited above for the value of a World Series game are in line with the increase value of the new contract.

CONCLUSION

Using the binomial distribution and the actual World Series outcomes we have been able to determine the revealed team strengths as well as the most likely relative strengths for each of the possible outcomes of a seven game series. We attribute this seeming inequality to the home field advantage enjoyed by one of the teams. We have also shown that a typical World Series has an expected number of games of 5.81. Since each game has economic value to the contractual broadcaster, contracts that do not consider the probability of less than seven game series are likely to cause marginal economic losses to the networks and windfall gains to MLB.
REFERENCES

1. CNNSI.com, Fox Lands Exclusive TV Rights to Postseason Baseball, 
2. CNNSI.com, Fox Posts Lowest Rated Series Ever, 
4. Staudohar, Paul D., The Symbiosis Between Baseball and Broadcasting, in Alvin L. Hall. editor, The 
5. Summer 1997: 75 Years of National Baseball Broadcasts, 
   http://roadsidephotos.com/baseball/nationalbroadcast.htm