Nicky Sullivan

5-Page Extended Abstract

Who Wins in the MLB Playoffs?

Comparing Pythagorean record to actual record to see which is a better predictor of success in the playoffs

Every year when the MLB playoffs roll around, articles upon articles are published trying to predict who the eventual World Series champion will be. Sports fans eat it up, poring over articles to see how many experts think their favorite teams will win. It's a time-honored tradition, and yet each year it seems as if a surprise team crashes the party and shocks all the experts by making a surprising run. This past year, both the Royals and Giants made it through a win-or-go-home wildcard game and ended up making it to the World Series, upsetting favorites like the Angels, Cardinals, and Nationals along the way. All this got me wondering: what's the best way to predict who will triumph in baseball's postseason?

To try and answer this question, I decided to compare two pretty simple ways of forecasting a team's future success. The first is simply looking at their record during the regular season, arguably the simplest way to compare two teams. In theory, a team's record over baseball's notably long 162-game season should be a pretty good estimate of the team's true talent, which is what should lead to a team being successful in the postseason. The second is slightly more complicated, but still fairly simple to calculate. Over 30 years ago, Bill James first developed the concept of a team's Pythagorean record, which was based solely on the number of runs the team scored and allowed. His theory was that there were large amounts of luck in baseball, and that even a 162-game season was not enough to accurately depict each team's true talent level. So instead of simply relying on a team's win-loss record, he used the team's runs

scored and runs allowed to create a formula that would spit out a number that would represent the team's 'true' winning percentage. The formula was as follows:

$$\frac{\text{runs scored}^2}{\text{runs scored}^2 + \text{runs allowed}^2}$$

James called his formula the Pythagorean expectation, and it is still used today in baseball and has been expanded and modified for use in other sports. I wanted to test James' Pythagorean expectation, and see if it worked better at predicting the MLB playoffs than simply looking at a team's win-loss record.

I chose to look at the last eleven postseasons, excluding any wild card play-in games because there would be so much variation in such a small sample of one game series that any data I collected would essentially be useless. This left me with 77 series, 44 five-game series in the first round, and 33 seven-game series in the second round and the World Series. For each series, I tracked down data for the records of both teams involved, as well as their runs scored and runs allowed, which I used to calculate the teams Pythagorean record using Bill James' formula.

Once I had these values, I used the Bradley-Terry model to calculate the probability that the home team would win any given game in the series. To do this, I assumed that each team's actual record and Pythagorean record would represent their success against league average teams, as each team should face a fairly average schedule over the course of a season. With this assumption, I simply calculated the probability by doing: [PA/(1-PA)]*[(1-PB)/PB]=PC, and then PC/(1+PC), where PA is the probability Team A wins against league average competition and PB is the probability Team B wins against league average competition. This left me with the probability that Team A would beat Team B in any given game.

From there, I added in a basic piece to account for home field advantage. Using the fact that the home team in baseball wins about 54% of the time, and that if a division series goes the distance the home team will get 3 home games and 2 away games, I calculated 3(.54/.46)+2(.46/.54)=1.05, and 2(.54/.46)+3(.46/.54)=.98. I added those together to get 2.03 and divided 1.05 by 2.03 to get 51.59%, which is the chance that Team A would be Team B in a five game series assuming the teams are equal but Team A has home field advantage. I did the same calculations for a seven game series and got a value of 51.14%. These are only estimates of home field advantage in a series, as they assume the series goes the distance, but they should be reasonable estimates that add a level of accuracy to my predictions.

I then added these home field advantages into the probabilities I had calculated of the home team winning each individual series, giving me a new probability that took into account home field. At this point, I had 154 probabilities, two for each series (one representing the probability based on actual winning percentage and the other based on Pythagorean winning percentage). From there, I used the binomial formula to calculate the probability that the home team would win however many games they needed to win the series, in this case three for the division series and four for the championship series and World Series. I set the number of trials equal to the maximum possible number of games, and set the p value equal to the probability the home team won any given game that I had calculated earlier. This left me with the probability that the home team would end up winning each and every series, and from there I moved on to performing the actual tests.

I looked at a number of different factors to see which method performed better. First, I simply looked at how well each method predicted who would reach and win the World Series. Neither method did a very good job of predicting who would win the World Series, as the Win%

method correctly predicted two of the eleven winners while the Pythagorean% predicted just one of the eleven. When it came to predicting which teams played in the World Series, however, the results got a little more interesting. The Win% method managed to correctly predict only 5.5 of 22 teams to play in the World Series (the half game comes from when two teams had the same record, making them co-favorites to make the World Series), but the Pythagorean% method predicted 9 of the 22 teams correctly. The difference wasn't significant, but it was a step in the right direction for the Pythagorean record.

Next, I looked at how often the favorites won. Again, the Pythagorean% was slightly better, as 44 of the 77 teams it listed as the favorite won, compared to only 41 of the 77 victories for favorites as calculated by the Win% method. Again, not a statistically significant difference, but another small vote of confidence in favor of the Pythagorean method.

Next I looked at the cases where the two methods differed. There were 13 cases where the two methods listed different favorites, and in 8 of those cases, the team the Pythagorean method preferred ended up winning. There were also 15 cases where the two methods listed the home team's probability of winning the series as being at least ten percentage points different, and in those 15 cases, the team the Pythagorean method preferred won 10 times. Neither of these were significant, but they again seemed to support the Pythagorean method.

Finally, I tried to see if I could get an idea for whether the actual winning percentage matched up with the expected winning percentage I had calculated. To do this, I sorted the projected chances of winning the series into four buckets for each method, and then for each bucket calculated the actual winning percentage of the home team in those series. If the methods were accurate, I would expect to see the actual winning percentage increase as expected winning percentage increased, but this was not the case. These are the results I obtained:

Win% Method

Projected Win%	Wins	Losses	Win%
<55%	8	7	53%
55-57.5%	10	8	56%
57.5-60%	10	9	53%
>60%	14	11	56%

Pythagorean% Method

Projected Win%	Wins	Losses	Win%
<50%	8	10	44
50-57.5%	15	8	65
57.5-65%	8	11	42
>65%	11	6	65

You'll notice that the buckets are not the same for each method, this is because I was trying to get buckets of roughly equal size that were large enough that I might be able to actually tell the difference between winning less than 55% of the time and more than 60% of the time. In the end, there was simply not enough data to get an accurate picture, as there was enough variation that one or two series going in a different direction would dramatically altered the percents I calculated.

While none of the results I found ended up being statistically significant, this does not mean that there was nothing to gain from the project. In many cases, it appears that the reason there was no statistical significance was in large part because my sample size was not large enough. Due to time constraints and a lack of programming knowledge that might have made it easier to expand this project to cover a larger number of series, I was only able to do calculations going back 11 years, but the results I saw lead me to believe that further investigation may yield significant results. The Pythagorean method constantly came out either ahead or equal to the Win % method, and I believe these trends would continue if a larger study was conducted that looked at more series. Think of this as not a complete analysis that definitively proves that one method is better, but rather as a starting point from which future studies could be built upon. Based on

my results, I can't say for sure if Bill James was correct in his belief that a team's Pythagorean record would be a better representative of their actual talent than their win-loss record. But the evidence points in that direction, and the next time you want to know who will succeed in the MLB playoffs, I recommend that you look at the competing teams' Pythagorean records, not just their win-loss record in the regular season.

References

Mathletics, by Wayne Winston.

Mlb.com's standings page. http://mlb.mlb.com/mlb/standings/.

http://espn.go.com/mlb/playoffs/2014/story/_/id/11611761/why-nats-win-world-series

http://espn.go.com/mlb/playoffs/2014/story/_/id/11613081/predicting-postseason

http://en.wikipedia.org/wiki/Pythagorean_expectation