presented at the MIT Sloan
Sports Analytics Conference
on March 6, 2010

Beyond Pythagorean expectation: How run distributions affect win percentage

Kerry Whisnant

Department of Physics and Astronomy

Iowa State University

Ames, IA 50011

Abstract

Although the Pythagorean expectation formula does well at predicting win percentage, the shape of the run distribution can also be a factor. Given two baseball teams with the same average runs per game, the team with the narrower run distribution tends to win more games. Modified formulas that take into account both the runs per game and the shape of the run distributions are presented. Also, slugging percentage has an inverse correlation with the width of the run distribution. A team slugging percentage .080 above average is worth about one extra win compared to the simple prediction using only runs scored and runs allowed.

1 Introduction

The Pythagorean expectation formula [1] provides a reasonably good estimate of the win percentage of a baseball team using the number of runs scored and runs allowed by that team. Improvements on the formula, such as the Pythagenport [2] and the Pythagenpat [3], allow for variation of the Pythagorean exponent and give a very good estimate for win percentages over a wide range of run environments. This paper looks at possible improvements on the Pythagorean expectation formulas that take into account not only the run environment, but also the shapes of the run distributions. There is a clear pattern that teams with a higher slugging percentage score runs more consistently, i.e., their run distributions have a smaller standard deviation, and they tend to win more games than their Pythagorean expectation. This paper also examines how metrics that use runs for evaluating teams and players might be adjusted to account for these effects.

Derivations of the Pythagorean expectation formula have been made under certain assumptions using continuous probability distributions. Hein Hundal showed that if run distributions are independent log-normal distributions, then the Pythagorean exponent can be approximated by a particular function of the standard deviation and mean of a typical run distribution [4]. Steven Miller showed that if the run distributions are Weibull distributions, then win percentages are given exactly by the Pythagorean formula, where the Pythagorean exponent is a parameter used to fit the run distributions to data [5]. In both cases the Pythagorean exponent inferred from run distribution data is very close to the empirical value of 1.8 to 1.9.

In the two derivations described above, the shape of the run distribution is determined for a given average runs per game (RPG), i.e., they are single-parameter distributions once the run environment has been determined. Consequently, a team with a higher RPG is always predicted to win more games. Actual run distributions depend on many variables, such as the rates for walks, singles, doubles, triples and home runs. A team with the same runs scored and runs allowed will not necessarily have a win percentage exactly equal to .500 if the shapes of those run distributions are different. This paper reports on an investigation of these effects and discusses their potential consequences:

- How the shapes of run distributions affect win percentage will be examined using four different sources for the distributions: (i) actual runs scored data, (ii) a log-normal distribution, which has two parameters that can be taken as the mean and standard

deviation of the distribution, (iii) a toy model where a given team hits only one type of base hit (single, double, or home run), and (iv) a Markov chain model where a team can have a mixed batting profile. In each case the more consistent a team is in scoring runs (i.e., the narrower the run distribution, or the smaller the standard deviation), the better its win percentage, for fixed RPG.

- Modifications of the Pythagorean expectation formula[1] are proposed that fit all of these data sets significantly better than the standard forms (including the Pythagenpat version).

- Using the results from above, the implications for evaluating players and building a team will be discussed.

The outline of the paper is as follows. In Section 2 it is shown in some extremely simplified cases how the shape of the run distribution can in principle dramatically affect a team's win percentage. Also discussed is how run distributions characterized by a single parameter are not sufficient to determine which team will win in a head-to-head matchup. In Section 3 actual run distributions are used to show that teams with a smaller standard deviation tend to win more games for fixed RPG. In Section 4 two different simple models are used to show that the standard deviation of a run distribution depends on the underlying intrinsic run distribution and is not just an artifact of random noise. In Section 5 a Markov chain analysis is used to provide a more realistic model for run distributions, and modifications of the Pythagorean expectation formulas are proposed that take into account the shape of the run distribution as well as the RPG. Finally, in Section 6 the implications for player evaluation are discussed.

2 Run distributions matter

Pythagorean expectation formulas use only the RPG ratio to determine a team's win percentage. The original version took the form [1]

$$\mathrm{WPct} = \frac{R_1^2}{R_1^2 + R_2^2}\,, \tag{1}$$

[1]In this paper I will use this term for all Pythagorean expectation formulas, including those that adjust for the run environment.

where R_1 is the number of runs scored by a team and R_2 is the number of runs allowed. An alternate form is

$$\frac{W}{L} = \left(\frac{RPG_1}{RPG_2}\right)^{\alpha}, \tag{2}$$

where runs are replaced by RPG, and for the original version $\alpha = 2$. Subsequent improvements also accounted for the run environment, with, e.g., $\alpha = 1.5\log(RPG_1 + RPG_2) + 0.45$ for the Pythagenport version [2] and $\alpha = (RPG_1 + RPG_2)^{.287}$ for the Pythagenpat version [3]. Since only the RPG ratio is used, a team that scores more RPG than they allow is necessarily predicted to win more games than they lose. In particular, a team that scores the same number of runs as they allow is predicted to have a win percentage of exactly .500.

However, the shape of the distribution can also have a large effect on the expected win percentage. To show this, consider three teams, each of which averages 5.0 RPG: Team A, which always scores 5 runs, Team B, which scores 4 runs two-thirds of the time and 7 runs one-third of the time, and Team C, which scores 6 runs two-thirds of the time and 3 runs one-third of the time. In none of these cases do the teams split their games evenly in head-to-head matchups, as predicted by the Pythagorean expectation. It is easy to show that Team A beats Team B 6 games out of 9, Team B beats Team C 5 games out of 9, and Team C beats Team A 6 games out of 9!

This rock/scissors/paper example shows that which is the better team is a non-transitive property. It immediately follows that no single parameter (e.g., RPG) is sufficient to determine which team is better in a head-to-head matchup, since if there was such a parameter (call it P), then $P_A > P_B$ and $P_B > P_C$ would necessarily imply that $P_A > P_C$, where greater than may be interpreted as "wins more games than in a head-to-head matchup."

This simplistic example clearly shows that in principle we must go beyond RPG to determine which team is better. Since a single parameter is insufficient, one or more new parameters must be introduced that include effects due to the shape of the run distribution. While completely general run distributions might in principle require a large number of extra parameters to properly describe them, baseball distributions are all qualitatively similar, and it will be shown that only one additional parameter is needed to significantly improve the predictive power.

3 Run distributions for 1999-2008

To try to understand how run distributions can affect win percentage, actual major league run distributions were tabulated from 1999-2008 [6]. For each year of the data, there were 435 possible head-to-head matchups between teams, which gave 4350 total pairings of teams. For each pairing, the win percentages were determined assuming that the runs scored distributions of the teams were independent from each other. Therefore if team i has probability $P_i(n)$ of scoring n runs in a game, then its win/loss ratio against team j with probabilities $P_j(n)$ is

$$\frac{W}{L} = \left[\sum_{n=1}^{\infty} P_i(n) \left(\sum_{m=0}^{n-1} P_j(m)\right)\right] / \left[\sum_{m=1}^{\infty} P_j(m) \left(\sum_{n=0}^{m-1} P_i(n)\right)\right] , \tag{3}$$

where it is assumed that ties are decided with the same ratio. This is the same procedure used to determine the expected win percentage using the Randomized Wins method [7], which has been shown to provide a better prediction for win percentages than the Pythagorean expectation[2].

To isolate the effect of the shape of the run distribution, initially only pairs of teams with approximately the same RPG (within 1%) were compared; it was found that the win percentage varies from .500 by as much as .035, compared to the maximum deviation of about .005 predicted by the Pythagorean expectation for two teams with RPG within 1%.

The next step is to show that much of the deviation is due to the different shapes of the distributions. As mentioned before, completely general distributions might in principle require many parameters to describe them. One such set would be the moments about the mean; here I define the nth moment to be

$$\sigma(n) = \left[\frac{1}{N}\sum_{i=1}^{N} (R_i - RPG)^n\right]^{1/n} . \tag{4}$$

The second moment is just the standard deviation (which will be referred to as simply $\sigma(2) \equiv \sigma$), while the third moment is related to the skew.

In fact baseball run distributions are all qualitatively similar, in the sense that they all tend to have more probability below the mean (i.e., there are more games with runs below the

[2]It is not surprising that the Randomized Wins method gives better predictions since it is using information about the shape of the run distribution that is ignored by the Pythagorean expectation, which uses only RPG.

average than above the average), which means they all have a positive skew. Furthermore, the skew is strongly correlated to the standard deviation; in the 1999-2008 data sample, the correlation was 0.79. Therefore σ may effectively be used as a measure of distribution shape.

In the 1999-2008 data sample, among head-to-head matchups in which the RPG were within 1%, a strong correlation (with correlation coefficient 0.92) was found between the win percentage and the inverse ratio of the standard deviations of the two teams (see Fig. 1). This relationship means that a more consistent team, i.e., one with the smaller standard deviation, tends to do better than the Pythagorean expectation, while an inconsistent team (larger standard deviation) does worse.

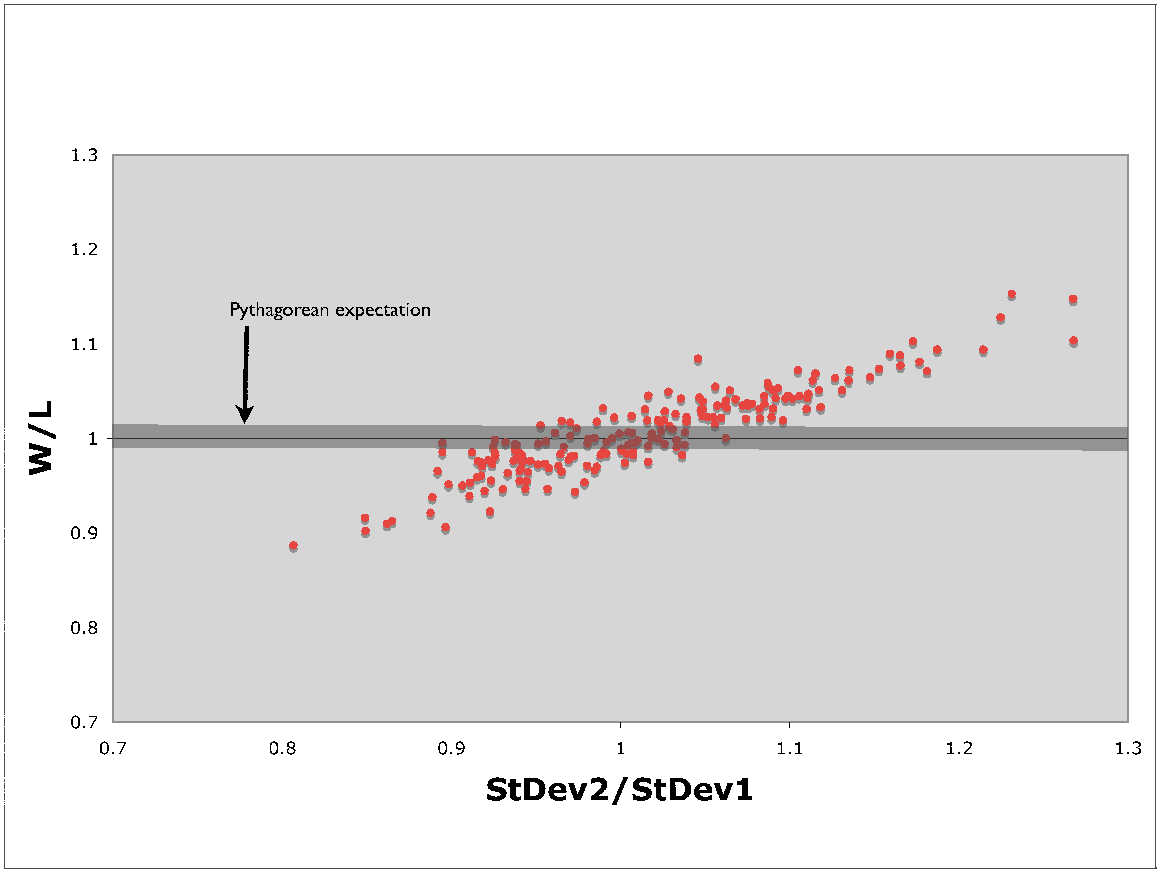

Figure 1: Win/loss ratio versus inverse standard deviation ratio for head-to-head matchups between teams with RPG within 1% for the years 1999 to 2008. The Pythagorean expectation is shown by the dark gray band.

There is a similar spread in win percentages and correlation with inverse standard deviation ratios for teams with different RPG. Dividing the pairs of teams into groups with approximately the same RPG ratio shows that the teams with the smallest (largest) stan-

dard deviation within a group tend to win the most (least) games (see Fig. 2). The degree of correlation between the win percentage and standard deviation ratio goes down as the RPG ratio deviates from one, but generally remains above 0.80.

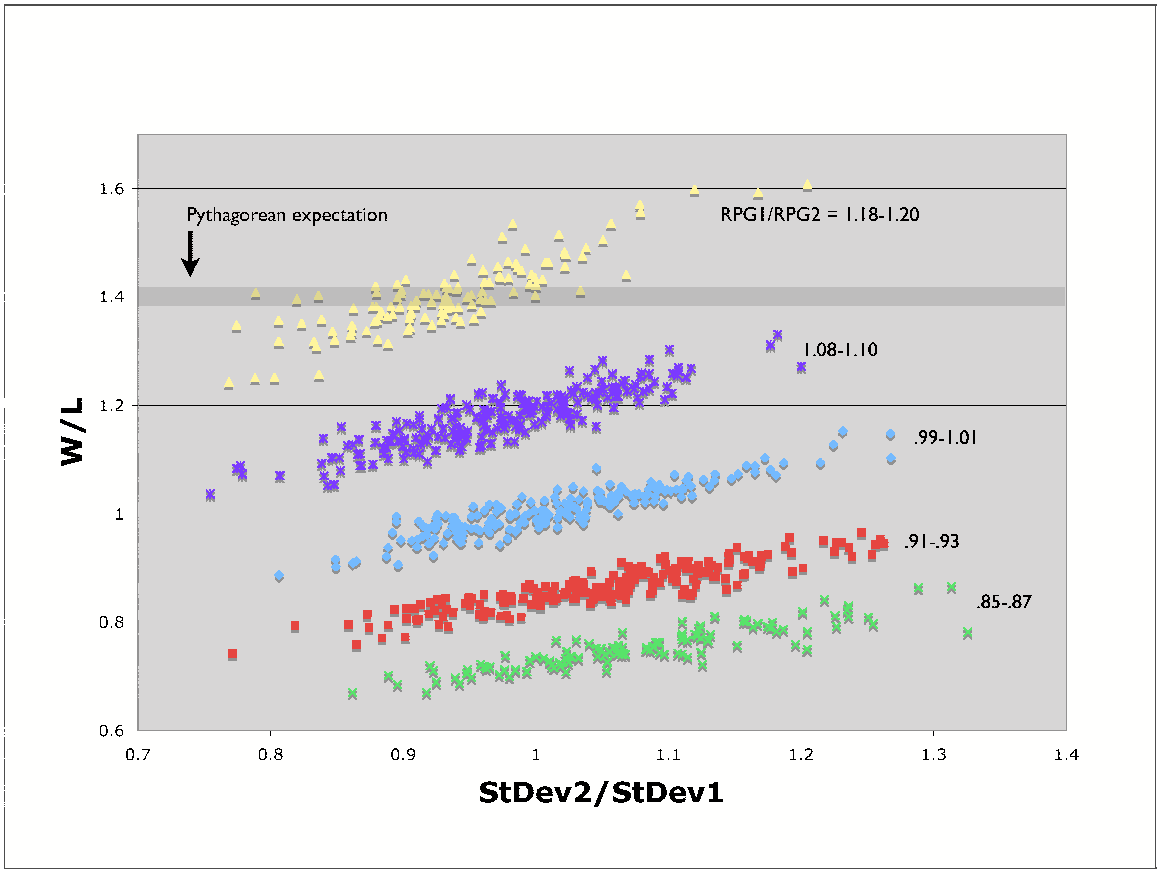

Figure 2: Win/loss ratio versus inverse standard deviation ratio for head-to-head matchups between teams with different fixed RPG ratios for the years 1999 to 2008. The Pythagorean expectation for $RPG_1/RPG_2 = 1.18$ to 1.20 is shown by the dark gray band.

There is also a mild dependence on the run environment (defined as the total RPG for both teams). A good fit to the 1999-2008 data set, with root mean square (RMS) error .0046 in win percentage, for the win/loss ratio (for Team 1 playing Team 2) is

$$\frac{W_1}{L_1} = \left(\frac{RPG_1}{RPG_2}\right)^{\alpha} \left(\frac{\sigma_2}{\sigma_1}\right)^{\beta}, \tag{5}$$

where RPG_i and σ_i are the RPG and standard deviation for team i, and

$$\alpha = 1.313\ (RPG_1 + RPG_2)^{.214}, \qquad \beta = 1.020\ (RPG_1 + RPG_2)^{-.356}. \tag{6}$$

This can be compared with an RMS error of 0.0109 when using the Pythagenpat formula, $W_1/L_1 = (RPG_1/RPG_2)^{\alpha}$ with $\alpha = (RPG_1 + RPG_2)^{.287}$, on the same data set.

The question now arises as to whether the differences in standard deviation are entirely due to random fluctuations, or are due partly to the different shapes of the intrinsic run distributions of the teams[3]. This will be studied in the next section.

4 Modeling run distributions

4.1 Log-normal distribution

First, we investigate how intrinsic run distributions might give the effect seen in the previous section by using a log-normal distribution. Log-normal distributions are well-suited for this since they have two independent parameters, which can be taken as the mean and standard deviation. They also have a positive skew, similar to the actual run distributions. A sample of 5000 pairs of teams were randomly generated with RPG in the range 4.0 to 6.0 and standard deviations in the range 2.0 to 3.0, which are typical values for baseball teams. Then for each team a run distribution for a 162-game season was randomly generated, to simulate the 1999-2008 data set. Then win percentages in the head-to-head match-ups were calculated as described in the previous section.

As with the actual data, a strong correlation was found between the win percentage and inverse standard deviation ratio for teams with approximately the same RPG ratio (see Fig. 3). However, although they are qualitatively similar, the win percentage did not have exactly the same relationship to RPG ratio and standard deviation ratio; the best linear fit of the W/L ratio to inverse standard deviation ratio gives slopes that are about 20-30% lower for the log-normal distribution than for the actual data. This is not surprising since log-normal distributions are continuous, not discrete, and the shapes of the intrinsic run distributions are not necessarily similar to log-normal, but it does show that different shapes of the intrinsic run distribution can affect win percentage in a way similar to that seen in the actual data.

[3]By intrinsic run distribution we mean the run distribution in the limit of an infinite number of trials, i.e., without random noise.

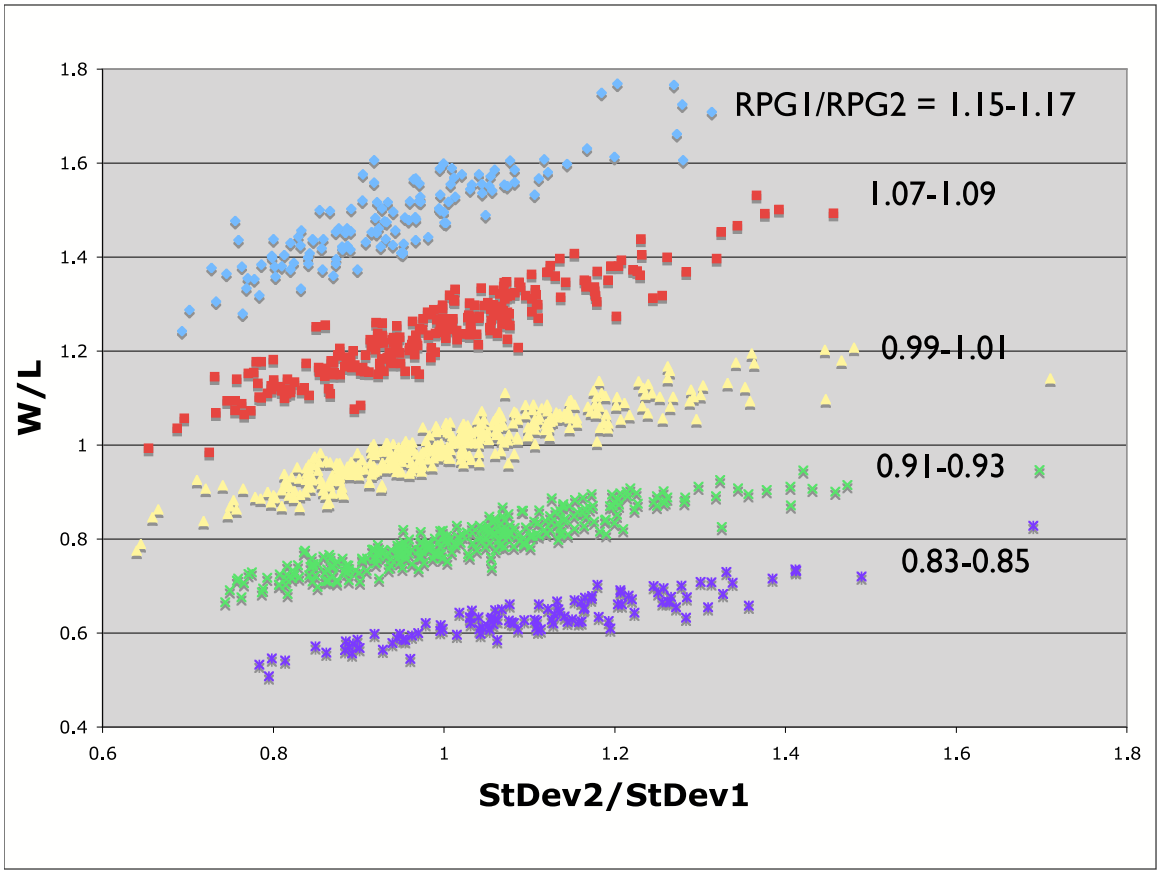

Figure 3: Win/loss ratio versus inverse standard deviation ratio for head-to-head matchups between teams with different fixed RPG ratios assuming log-normal distributions.

4.2 Toy model

Next, a simple toy model was used that attempts to mimic the way runs are actually scored. To keep it simple, it is assumed that a given team has only one type of base hit: single with runners advancing one base (denoted as the 1B team), single with runners advancing two bases (the 1B+ team), double (the 2B team), or home run (the HR team). It is also assumed that each batter on the team has the same batting average, i.e., all the players on the same team are identical to each other.

The HR team naturally scores a single run on every hit, the 2B team scores a run on the second hit in an inning and on each hit thereafter, the 1B+ team scores a run on the third hit in an inning and on each hit thereafter, and finally the 1B team scores a run on the fourth hit in an inning and on each hit thereafter. Therefore these four cases cover all the possibilities given these simplifying assumptions.

It is not too hard to show that the probability of getting n hits in an inning is

$$P(n) = \frac{1}{2}(n+2)(n+1)\ AVG^n\ (1-AVG)^3\,, \tag{7}$$

where AVG is the team batting average. The runs per inning (RPI) distributions follow directly from the $P(n)$, and the RPG distributions[4] may then be determined from the RPI distributions[5]. Then using the same method as before the win percentages for head-to-head match-ups may be calculated, assuming the distributions are independent.

In the previous calculations of win percentages only RPG distributions were available and ties were assumed to be decided by the same ratio as games that did not go into extra innings. However, for the toy model ties may be correctly decided in extra innings by using the RPI distributions. Since the RPI distributions are different from RPG distributions, the probability of winning in extra innings is not necessarily the same as the overall probability of winning the game.

The results for teams with RPG = 5.0 are displayed in Table 1, which shows the win percentages for each possible match-up, as well as the standard deviation of a team's run distribution and the prediction using the empirical formula from Eqs. 5 and 6.

Clearly the team with the more consistent offense (lower standard deviation) wins more, *even though the RPG are identical*. The additional wins per 162-game season can be as many as eight in the most extreme case (the HR team versus the 1B team). Also, the empirical formula derived from actual data, Eqs. 5 and 6, gives fairly good predictions for these win percentages.

The run distributions are shown in Fig. 4. The HR team is the most consistent, with more probability peaked near the average, while the 1B team is the least consistent, with more probability at high and low run values. It is evident from looking at the figure that

[4]Strictly speaking these are runs-per-27-outs distributions.

[5]Although the RPG distributions are not simple expressions of AVG, the average RPG are:

$$\begin{aligned}
RPG(HR) &= \frac{27\ AVG}{1-AVG} \\
RPG(2B) &= \frac{9\ AVG^2\ (6-4\ AVG+AVG^2)}{1-AVG} \\
RPG(1B+) &= \frac{9\ AVG^3\ (10-10\ AVG+3\ AVG^2)}{1-AVG} \\
RPG(1B) &= \frac{9\ AVG^4\ (5-6\ AVG+2\ AVG^2)}{1-AVG}\,.
\end{aligned}$$

Table 1: Win percentages for the four teams in the toy model in head-to-head matchups. Also shown is the predicted win percentage using the modified Pythagenpat expectation formula in Eqs. 5 and 6.

			Actual W Pct vs.				Predicted W Pct vs.			
Team	AVG	St.Dev.	HR	2B	1B+	1B	HR	2B	1B+	1B
HR	.15625	2.434	.5000	.5216	.5369	.5491	.5000	.5244	.5391	.5495
2B	.28367	3.024	.4784	.5000	.5160	.5290	.4756	.5000	.5148	.5253
1B+	.37344	3.449	.4631	.4840	.5000	.5133	.4609	.4852	.5000	.5105
1B	.44077	3.787	.4509	.4709	.4867	.5000	.4505	.4747	.4895	.5000

teams that are more consistent have the peak of their distribution at a higher run value, which is why they tend to win more games.

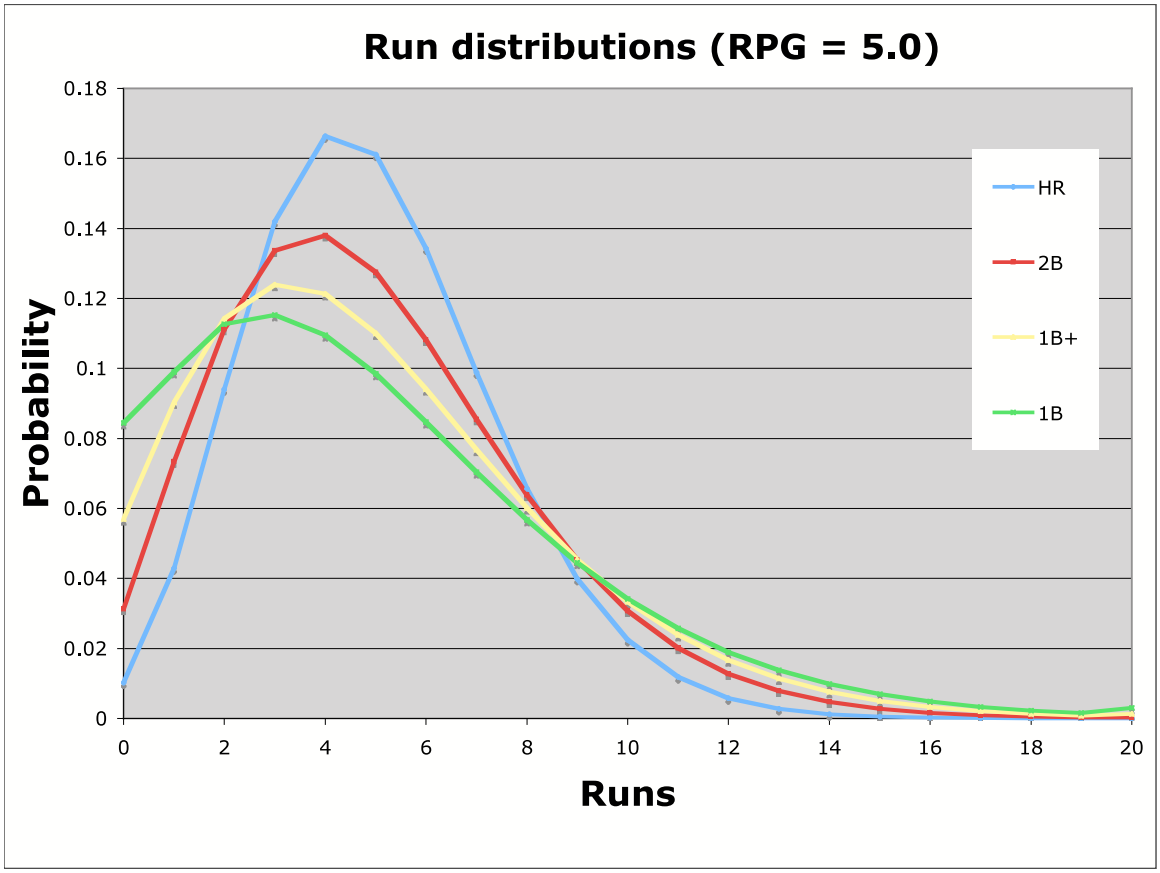

Figure 4: Run distributions for the four teams in the toy model.

The extra-innings win percentages are even more different from the Pythagorean expec-

tation. The results for teams with RPI = 5/9 (i.e., RPG = 5.0) are shown in Table 2. The typical excess win percentage in extra innings is 2.35 to 2.50 times the overall excess win percentage. For example, the HR team beat the 1B team about 62% of the time in extra innings, primarily due to the fact that the HR scores in 40% of innings while the 1B team scores in only 24% of innings[6].

Table 2: Win percentages in extra innings for the four teams in the toy model.

			Extra-innings W Pct vs.			
Team	RPI	St.Dev.	HR	2B	1B+	1B
HR	0.5556	0.8114	.5000	.5541	.5925	.6226
2B	0.5556	1.0088	.4459	.5000	.5391	.5701
1B+	0.5556	1.1535	.4075	.4609	.5000	.5313
1B	0.5556	1.2729	.3774	.4299	.4687	.5000

5 Markov chain analysis

The toy model in the previous section provides a more realistic description of run distributions than the log-normal distribution, and likely defines the largest possible scoring distribution effect for baseball. However, baseball teams are not one-trick ponies that only have one type of hit. To make a more realistic model we need to allow for a more general batting profile with individual probabilities for getting a walk, single, double, triple or home run per plate appearance (denoted x_0, x_1, x_2, x_3 and x_4, respectively). The parameter x_0 also includes hit by pitch, which has the same effect on the field as a walk.

Furthermore, runners sometimes (but not always) advance one more base than the batter does. To include this effect we define six advancement probabilities: y_{13}, the probability of a runner advancing from first to third on a single, y_{24}, the probability of a runner scoring from second on a single, y_{14}, the probability of a runner scoring from first on a double, y_{12}, the probability of a runner advancing from first to second on an out, y_{23}, the probability of a

[6]Of course the 1B team tends to score more runs in the innings in which they do score, since the overall RPG are the same. The feast or famine experienced by the 1B team is not surprising considering they must have four hits in an inning before they score a run. However, once they do score, the probability of subsequent scores is much higher due to their much higher batting average.

runner advancing from second to third on an out, and y_{34}, the probability of a runner scoring from third on an out. The latter case would include, for example, sacrifice flies, and the y_{12} and y_{23} coefficients would include sacrifice bunts, although in all three cases other plays with runner advancement are also possible. The advancement coefficients were determined from 2008 major league baseball play-by-play data taken from Retrosheet [6]: $y_{13} = 0.246$, $y_{24} = 0.548$, $y_{14} = 0.389$, $y_{12} = 0.159$, $y_{23} = 0.300$ and $y_{34} = 0.367$.

To reduce the size of the (five-dimensional) batting profile parameter space, the triples rate was set at 0.5% (about the MLB average in 2008) and the doubles rate was set to 31% of the singles rate (also about the MLB average). Triples were frozen because they are relatively rare, and the doubles-to-singles ratio was fixed because it has the smallest variation of all of the possible hit ratios, at least for team totals. Then the batting profile is uniquely determined by the walk, single and home run rates (x_0, x_1 and x_4), which can be converted to the traditional slash stats, AVG/OBP/SLG, or vice versa[7].

The batting profiles and run advancement parameters are then used in a Markov chain analysis to determine the RPI distributions as follows [8]. Each possible base-out-score situation (i.e., which bases are occupied, the number of outs and how many runs the team has scored in the current inning) is considered to be a different state. The batting profile and run advancement parameters then define the probability for moving from one state to another in a given plate appearance. The state-to-state transition is implemented as a matrix acting on the state vector, where the initial state for a given inning has a probability of one for having no men on base, no outs and no runs scored. The matrix is applied to the state vector many times until all states with less than three outs have negligible probability (generally 100 applications is more than sufficient). The coefficient of the state with three

[7]These conversion are:

$$OBP = x_0 + x_1 + x_2 + x_3 + x_4\,, \quad AVG = \frac{x_1 + x_2 + x_3 + x_4}{1 - x_0}\,, \quad SLG = \frac{x_1 + 2x_2 + 3x_3 + 4x_4}{1 - x_0}\,,$$

and, letting $x_2 = cx_1$,

$$\begin{aligned} x_0 &= \frac{1 - OBP}{1 - AVG}\,, \\ x_1 &= \frac{1}{3 + 2c}\left[\left(\frac{1 - OBP}{1 - AVG}\right)(4\ AVG - SLG) - x_3\right]\,, \\ x_4 &= AVG\left(\frac{1 - OBP}{1 - AVG}\right) - x_3 - \frac{1 + c}{3 + 2c}\left[\left(\frac{1 - OBP}{1 - AVG}\right)(4\ AVG - SLG) - x_3\right]\,, \end{aligned}$$

outs and n runs scored is then the probability of scoring n runs in an inning.

The RPG distribution can then be calculated from the RPI distribution by a similar process. Inning-score states are defined as having a particular score after a given inning. The initial state has a probability of one for starting the first inning with no runs scored. The RPI distribution determines the transition matrix for moving from one inning-score state at the beginning of an inning to an inning-score state at the end of the inning. Once this transition matrix is applied nine times, the state vector then gives the RPG distribution.

A random sample of 100 teams was generated assuming the following ranges: .220 to .300 for AVG, 0.050 to 0.100 for (OBP − AVG), and 0.080 to 0.170 for (SLG − AVG). The differences (OBP − AVG) and (SLG − AVG) were used, rather than OBP and SLG directly, since they tended to give more realistic slash stats. These ranges cover most of the values seen for teams since 1900 (although the extremes for AVG are actually about .200 and .320, and .050 and .200 for (SLG − AVG))[8].

From these 100 teams, the win percentage for each of the possible 4950 head-to-head matchups was determined using the RPG and RPI distributions, where the latter were needed to determine the winner in extra innings. Since the Markov chain is an exact calculation, there is no noise in these distributions. A least-squares fit to the modified Pythagorean W/L formula of Eq. 5 was made, with best fit parameters

$$\alpha = 1.251\ (RPG_1 + RPG_2)^{.216}\,, \qquad \beta = 1.328\ (RPG_1 + RPG_2)^{-.430}\,. \tag{8}$$

The RMS error in the win percentage was 0.00077, compared to 0.00234 for the Pythagenpat formula with $\alpha = (RPG1 + RPG2)^c$, where c was allowed to vary to give the best fit. The RMS error of the parameters from Eqs. 6 and 8 are shown in Table 3 for both the actual 1999-2008 data and the Markov chain simulated "data."

The parameters derived from the Markov chain data set provide a good fit to the actual data, while the parameters that fit the actual data do not do nearly as well on the Markov chain data set (with an RMS error ten times the best fit). This suggests that the parameters derived from the actual data are fitted to some extent to noise, and are not good representations of the effect of the intrinsic run distribution. On the other hand, the parameters derived from the Markov chain analysis do almost as well in describing the actual data as

[8]These extremes were found by looking at team slash stats in the years 1908, 1930, 1968 and 2000, the low and high water marks for offense since 1900.

Table 3: RMS error in win percentage for the best fits to the 1999-2008 and Markov chain simulated data sets.

	RMS error in WPct	
	1999-2008 data	Markov "data"
Parameters from 1999-2008 data	.0046	.00778
Parameters from Markov "data"	.0062	.00077

the best fit to the actual data, which strongly suggests that they *are* a good representation of the run distribution effect.

One problem with using the standard deviation as a measure of the shape of the distribution is that there is a lot of uncertainty in determining its value, given that there are only 162 games in a season. It is also not a statistic that is regularly quoted for baseball teams. Therefore it would be nice to be able to quantify the run distribution effect using a typical baseball statistic. Alternatively, perhaps some other property of the distribution (e.g., skew) would provide a better description of the run distribution effect. To test this, the basic modified form

$$\frac{W_1}{L_1} = \left(\frac{RPG_1}{RPG_2}\right)^{\alpha} \left(\frac{P_1}{P_2}\right)^{\beta}, \tag{9}$$

was assumed, where P is the parameter being used to describe the shape of the run distribution.

First, skew (defined here as $\sigma(3)$ from Eq. 4) was tried as the parameter P. It gave an RMS error in win percentage of .00080 for the Markov chain data set, about the same as the error using standard deviation. This is not surprising considering the strong correlation (correlation coefficient 0.99) between the standard deviation and skew for Markov chain simulated distributions.

Since in the toy model the HR team did best and the 1B team worst, it appears that high SLG is good and high OBP is bad given the same number of runs per game. This suggests that either SLG or OBP^{-1} could be used for P. Or perhaps P could be chosen as the HR rate itself. The HR team leaves no men on base, so another option for P would be the fraction of runners that score, $FRS = RPG(1 - OBP)/(27\ OBP)$. Table 4 shows the RMS error for all of the different choices for P considered here. Also shown is the best-fit Pythagenpat expectation with $\beta = 0$ (i.e., only RPG are used).

Table 4: RMS error in win percentage for different choices of the parameter P used to represent the shape of the run distribution. Also, shown is the error of the best fit Pythagenpat formula.

P	SLG	StDev^{-1}	Skew^{-1}	FRS	OBP^{-1}	HR	Pyth'pat
RMS error	.00068	.00077	.00080	.00090	.00092	.00136	.00234

Clearly choosing $P = SLG$ gives the best fit to the win percentages derived from the Markov chain simulated data set, which is more than three times better than the Pythagenpat expectation that uses only RPG. The most accurate modified Pythagorean expectation formula found was

$$\frac{W_1}{L_1} = \left(\frac{RPG_1}{RPG_2}\right)^{\alpha} \left(\frac{SLG_1}{SLG_2}\right)^{\beta}, \tag{10}$$

where

$$\alpha = 0.723\ (RPG_1 + RPG_2)^{.373}, \qquad \beta = 0.977\ (RPG_1 + RPG_2)^{-.947}, \tag{11}$$

were the best-fit values for α and β.

It is interesting to see how well the modified Pythagorean expectation formulas fit to the simulated Markov chain data (Eqs. 5 and 8 with standard deviation and Eqs. 10 and 11 with SLG) predict the win percentages for the toy model. Table 5 shows that although the standard deviation formula works fairly well even in this extreme case, the SLG formula does not. Therefore the SLG formula apparently provides a good proxy for the distribution shape only for realistic baseball run distributions. It is not surprising that the standard deviation formula still works since it is a direct measure of the shape.

Table 5: Predicted win percentages for the four teams in the toy model in head-to-head matchups using the best-fit solutions with standard deviation and SLG. These may be compared to the actual win percentages for the toy model in Table 1.

			StDev predicted W Pct vs.				SLG predicted W Pct vs.			
Team	SLG	St.Dev.	HR	2B	1B+	1B	HR	2B	1B+	1B
HR	.62500	2.434	.5000	.5267	.5429	.5543	.5000	.5027	.5142	.5096
2B	.56734	3.024	.4733	.5000	.5162	.5277	.4973	.5000	.5115	.5069
1B+	.37344	3.449	.4571	.4838	.5000	.5115	.4858	.4885	.5000	.4954
1B	.44077	3.787	.4457	.4723	.4885	.5000	.4904	.4931	.5046	.5000

6 Discussion of player evaluations

6.1 The additional value of SLG

Equations 10 and 11 may now be used to determine how much a higher SLG is worth *given the same RPG*. From the derivative of the W/L formula it is easy to show that

$$\Delta WPct \approx WPct\ (1 - WPct) \left[\alpha\ \frac{\Delta R_1}{R_1} + \beta\ \frac{\Delta SLG_1}{SLG_1} \right] , \tag{12}$$

where terms due to the variation of α and β are small and have been ignored. From this equation it is immediately obvious that the extra SLG (for fixed RPG) needed to give the same increase in win percentage as one additional run is

$$\Delta SLG = \frac{\alpha}{\beta} \frac{SLG}{R_1} , \tag{13}$$

where R_1 is the number of runs scored in a season. For the current run environment ($RPG =$ 4.61 per team, so that $\alpha = 1.64$ and $\beta = 0.122$), this gives $\Delta SLG \approx 0.008$.

This means that a team with an average RPG and SLG 0.080 higher than average should on average win as many games as a team with average SLG that scored ten runs more than average. Using the usual equivalence that about ten runs equals one win, such a team should on average win about one more game per season than their traditional Pythagorean expectation. Similarly, a player that has a SLG .072 higher (lower) than average would increase (decrease) his team's SLG by about .008, so he should be rated one run better (worse) than he would be otherwise.

The consequences for building a team follow directly from these results. If two teams have the same value in runs (using whatever runs metric you prefer), and one team has a SLG that is .080 higher, then that team should expect to win about one more game a season, *even though it has the same RPG as the other team*. Therefore one should choose players with higher SLG if run values are equal, and in some cases a higher SLG even if their value in runs is slightly less.

Player values usually include both an offensive and defensive contribution. Then both R_1 and R_2 in Eq. 10 are affected, and

$$\Delta WPct \approx WPct\ (1 - WPct) \left[\alpha\ \frac{\Delta R_1}{R_1} - \alpha\ \frac{\Delta R_2}{R_2} + \beta\ \frac{\Delta SLG_1}{SLG_1} \right] , \tag{14}$$

where ΔR_1 is the offensive contribution and ΔR_2 is the defensive contribution. Although the relative values of the offensive and defensive contributions may vary from team to team,

for an average team

$$\Delta WPct \approx WPct\ (1 - WPct) \left[\frac{\alpha}{R} (\Delta R_1 - \Delta R_2) + \beta\ \frac{\Delta SLG_1}{SLG_1} \right], \tag{15}$$

where R is the average RPG for both offense and defense. Then $\Delta R_1 - \Delta R_2$ is just a player's total contribution as measured in runs, such as that determined during the calculation of WARP by Baseball Prospectus or WAR by Fangraphs, and once again $\Delta SLG = .008$ is equivalent to one additional run.

6.2 A tale of two teams

As an example of how large the SLG effect can be, several full-time players were chosen for each position, some with low SLG and others with high SLG. All possible lineups were considered for both the low-SLG and the high-SLG team. The one pair of lineups with the same WARP for each team *and* the largest difference in team SLG is shown in Table 6 (WARP and SLG values are all from the 2009 season).

Table 6: High-SLG team (Team A) and low-SLG team (Team B) that have the same WARP values.

Pos.	Team A	WARP	SLG	Team B	WARP	SLG
C	Miguel Olivo	1.9	.490	Russell Martin	3.3	.329
1B	Kendry Morales	2.8	.569	Nick Johnson	2.4	.405
2B	Jose Lopez	0.2	.463	Nick Punto	0.0	.284
3B	Mark Reynolds	3.3	.543	Chone Figgins	6.0	.393
SS	Troy Tulowitzki	6.1	.552	Marco Scutaro	5.6	.409
LF	Raul Ibanez	4.5	.552	Nyger Morgan	4.9	.388
CF	Cody Ross	1.2	.469	Tony Gwynn, Jr.	2.5	.344
RF	Michael Cuddyer	1.7	.520	Randy Winn	1.6	.353
DH	Jason Kubel	3.3	.539	Pat Burrell	-1.3	.367
	Total	25.0	.522	Total	25.0	.364

Although the two teams have the same total WARP, if we assume the same pitching strength for both teams, then because the high-SLG team has a SLG .158 higher than the low-SLG team, they should therefore win on average two more games a season. Why the big difference? Just because the teams have the same WARP does *not* mean they have the same

RPG – in fact the high-SLG team (not surprisingly) has an offensive value that is about 12 wins better than the low-SLG team. Of course the low-SLG team has a correspondingly better defensive value, so that their WARP values are the same. But this shows that the most effective way to capitalize on the run distribution effect is to choose high-offense, poor-defense players.

6.3 The extra win value of base advancement

We saw in the toy model that the 1B+ team would win more games in a head-to-head matchup with the 1B team, even though they had the same RPG. This makes sense since taking an extra base effectively adds to your total bases, and therefore is similar to increasing your SLG. But how many wins might a good baserunning team be worth (once again assuming the same RPG)? By varying the base advancement parameters y_{ij}, a similar Markov chain analysis can be done. For example, if a team increases y_{13}, y_{24} and y_{14} by .075 (e.g., going from first to third on a single 32.5% of the time instead of 25% of the time) without changing its RPG, it would be worth the equivalent of about one extra run, or 0.1 wins, per season. This is a much smaller effect than having a different batting profile, but it is consistent with the earlier findings that the 1B+ team won more games than the 1B team, even though they had the same RPG.

6.4 Run distribution effect for pitching

How about pitching evaluations? What is good for the offense must be bad for the defense – for the same expected RPG allowed, the pitcher with the *lower* SLG allowed will help his team win more games. If one set of pitchers had a SLG allowed .080 less than another set of pitchers for the same RPG allowed (presumably because they allow more walks and singles but not as many extra base hits), the first set of pitchers would be worth about one extra win per season.

However, in looking at pitchers with similar runs allowed averages, there was a much smaller difference than for the offensive players in Table 6. A big reason for this is that there is no fielding term to offset a difference in offensive values, and it is harder to get a different SLG allowed for two pitchers with the same RPG allowed, at least for realistic run distributions. In principle pitcher fielding can provide an offset to RPG allowed, but since there are fewer fielding chances for pitchers in a typical game, this offset would not be as

big for pitchers. The pitching side of the run distribution effect deserves more study, but it appears to be smaller than for position players.

6.5 Using the run distribution effect

Although an additional two wins would certainly be significant, and could mean the difference between making the playoffs or not playing in October, it is only about 0.2 wins per player. Uncertainties in WARP or WAR values are generally not quoted, but they are certainly larger than this, especially given the large differences in fielding evaluations between different wins metrics. Therefore the run distribution effect might be hard to use in practice, and would probably not change the relative ranking of most players.

However, even though these effects are not large for a single player, a concerted effort by a team to fill its roster with offensive players that have a higher SLG would mean they would be getting perhaps as many as two additional wins per season compared to the standard win metrics such as WARP or WAR. Also, if a team was trying to choose between two players who had the same WARP rating (and were otherwise equivalent in terms of age, etc.), they might want to pick the one with the higher SLG since that player would tend to give a narrower run distribution for the team and hence be worth a little bit more in terms of wins.

Alternatively, if a team uses a wins-based metric (where wins are derived from runs) to help determine the monetary value of a player, then they might be willing to pay a little more for a high-SLG player and a little less for a low-SLG player. Fangraphs in 2008 claimed that one WAR point was worth \$4.5M, where M stands for million. However, that number was based on free-agent salaries. If the total salary pool was divided by the total number of WAR points, the number is about \$2.5M per WAR point. Even so, an additional 0.1 win for a high-SLG player due to the run distribution effect might mean a team would be willing to pay \$0.25M more for that player (or \$0.25M less for a low-SLG player), or perhaps even more using the free-agent value of a win. Similarly, it could sway the decision in an arbitration case if the arbitrator was one the fence between the club offer and the player request.

Of course there are many factors that go into evaluating how much a team is willing to pay for a player. High-SLG players might already be valued at a premium if it is perceived that fans favor the long ball. Star power and popularity with the fans also must also be considered. The run distribution effect, which leads to a slightly higher evaluation for SLG, is just one piece of a very complicated puzzle.

References

[1] For an early discussion of Pythagorean expectation, see B. James, *1982 Bill James Baseball Abstract*, p. 18, "The Pythagorean Method."

[2] C. Davenport and K. Woolner, "Revisiting the Pythagorean Theorem," http://www.baseballprospectus.com/article.php?articleid=342.

[3] See http://gosu02.tripod.com/id69.html for a history of the Pythagenpat formula developed by D. Smyth and B. Heipp, and other discussions of win percentage estimators.

[4] H. Hundal, "Derivation of James's Pythagorean Formula," http://groups.google.com/group/rec.puzzles/browse_thread/thread/3be0e6ad49631ddb/bfb52d16b12955ac?q=hein+hundal+pythagorean&fwc=1

[5] S. Miller, "A Derivation of the Pythagorean Won-Loss Formula in Baseball," http://arxiv.org/pdf/math/0509698.

[6] The game scores and play-by-play information used in this paper were obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at www.retrosheet.org.

[7] T. Roux, "Randomized Wins: Predicting team wins using game run totals," http://www.diamond-theory.com/index.php?option=com_content&view=article&id=137

[8] Markov chain models have a long history in baseball analysis; for a brief introduction, see M. Pankin, "Baseball as a Markov chain,", http://www.pankin.com/markov/intro.htm and "Markov chain models: theoretical background," http://www.pankin.com/markov/theory.htm.