An improvement to the baseball statistic "Pythagorean Wins"

Jay Heumann*

Abstract. This paper will introduce a new version of the commonly used "Pythagorean wins" statistic, which improves on the traditional method by means of a mathematical adjustment. The new statistic is called "pairwise" Pythagorean wins, and it is demonstrated, over a 30-year data set, to have a smaller overall root mean square error than the original Pythagorean formula.

Keywords: Pythagorean formula, mathematics, expectation

1. Introduction

For decades, baseball statisticians have accepted and made use of the idea, first introduced by Bill James, that a team's ratio of runs scored to runs allowed, used in the proper formula, is a better predictor of a team's future performance than its win-loss record. The idea as originally formulated (and explained in (Pythagorean Theorem of Baseball n.d.), among many other places) is roughly as follows: to generate a prediction of a team's winning percentage, all we have to know is the number of runs the team has scored (call that R) and allowed (call that r). Then the predicted winning percentage can be computed using the formula

$$P := \frac{R^2}{R^2 + r^2}$$

This formula is usually called the *Pythagorean formula*; the output P is called the *Pythagorean winning percentage*; and often P is multiplied by the number of games a team has played to obtain a number analogous to wins, called *Pythagorean wins*.

A team's Pythagorean winning percentage is supposed to represent the "true" probability that the team will win a random game it plays. As a consequence, Pythagorean wins are treated as the expected value of the number of wins a team should have. Indeed,

Pythagorean win-loss totals are often referred to as "Expected W-L" for this exact reason. (Pythagorean Theorem of Baseball n.d.) expressly claims that "a team's actual W-L record will approach the Pythagorean Expected W-L record over time, not the other way around". This clearly follows from assuming that Pythagorean wins are in fact the expected values of teams' win totals. Years of use have shown that this basic concept is a very useful notion to analyze teams' performances.

In the literature on this statistic—including (Pythagorean Theorem of Baseball n.d.) itself, and discussed more fully in (Davenport & Woolner 1999)—it is often pointed out that the original Pythagorean formula admits room for improvement. The most common claim is that, as a statistic, it is an inaccurate predictor, frequently subject to error, and it can be made more accurate by changing the exponent in the formula from 2 to a different number. If the exponent is kept constant, the "best" one is alleged to be 1.83; in (Davenport & Woolner 1999), any number between 1.8 and 1.9 is considered about the same as any other, and sources such as (Major League Baseball n.d.) use the exponent 1.83. In addition, there are other purported fixes that utilize a variable exponent (we will not discuss those here).

The purpose of this paper is to point out an entirely different way to improve the Pythagorean wins statistic, one that applies to (at least) all versions of this formula that have a fixed exponent. It is not, at its heart, a *statistical* change, but instead a *mathematical*

*Corresponding author: Jay Heumann. Tel.: +715 308 7597; E-mail: jaydotheumann@gmail.com.

change. As it is usually computed, it is mathematically impossible for the Pythagorean wins statistic to represent what it claims to represent: namely, the expected value of teams' win-loss records. It is quite simple to show the mathematics behind this assertion; however, it would appear from the current state of the literature that this has never been remarked upon.

It is generally well-known among sports statisticians that the Pythagorean formula is not meant to be an exact expectation, but instead an approximation of an expectation, and one might think, in light of the fact that this is already known, that the above observation is redundant. However, the "correctness" of the Pythagorean formula does at times seem to be entrenched in the literature in a way that seems bizarre for a formula known not to be entirely correct. There have been efforts to expand the formula to sports other than baseball—for only a few examples, see (Cochran & Blackstock 2009) and (Hamilton 2011). Another effort to enshrine the formula in the literature can be found in (Miller 2007), where the author attempts to mathematically prove the correctness of the Pythagorean formula given certain assumptions. (This paper does not attempt to contradict Miller's theorem; in fact, the ideas in that paper are easily reconcilable with the ideas in this one.)

Using a revised formula that is more in line with the mathematical notion of an expectation should have less overall error than the original Pythagorean formula. This paper offers such a formula, called the pairwise Pythagorean formula for reasons that will be made clear below. It is extremely similar to the traditional Pythagorean formula but does not fail the basic expectation property that the traditional formula does. This fact alone does not guarantee greater overall accuracy than the original—that claim can only be evaluated using data. We will do exactly this below; over a sample of 60 seasons, switching to a pairwise method does in fact show less error (measured by root mean square) than the original formula.

The organization of this paper is as follows. In Section 2, we discuss the mathematics behind the change in the formula. In Section 3, we introduce the pairwise Pythagorean formula. In Section 4, we use the new formula, with two different fixed exponents, to generate new "expected" win totals for a sample of teams throughout history (grouped by league and season)—first with one specific case to demonstrate the formula's use, then summary results from all 60 seasons in the data set. In Section 5, we conclude by briefly discussing a possible future synthesis of the

pairwise method introduced here with other previous attempts to improve the Pythagorean formula; the two are not mutually exclusive! Finally, the Appendix contains team-by-team data for five specific seasons within the total data set, including comparisons with the traditional Pythagorean formulae.

Before continuing, let us fix some notation. The variable R will always represent runs scored; the variable r will represent runs allowed. The variable P will represent the percentage that is the output of a traditional Pythagorean formula. Variables i, j, and k will always be indices—for example, R_i will mean the total runs scored by team i. The variable a will always refer to the exponent used in the Pythagorean formula; in this paper, we will only consider versions of the formula where all percentages are computed using the same exponent. We will define more notation below as needed, but all of the above are used frequently throughout, and always in the same way. Finally, throughout this paper, all random variables are discrete.

2. Basic properties of expectations

Before going into the specifics of the Pythagorean wins statistic, we first present some of the basic properties of an *expectation*. We will then see below that the Pythagorean formula does not adhere to these properties.

The definition of an expectation for a (discrete) random variable X is

$$E(X) := \sum_y y Pr(X = y)$$

where the sum is over all values y that X may take. One of the properties that follows from this definition if that if X and Y are any two random variables, then

$$E(X + Y) = E(X) + E(Y)$$

and it also follows immediately from this that the expectation of any finite sum of random variables must necessarily be the sum of their expectations.

The Pythagorean wins statistic is assumed to be an expectation by the very nature of the formula. It makes the assumption that the percentage produced by the formula is the probability that a team will win any one game it plays. (This is obviously just a simplification for estimation purposes.) Once we assume this, the next step is treating each game as a random variable. Then, by the above property, the expectation of total wins over an entire season is the sum of the

expectations of each game—which, since each game can only end in a win (1) or loss (0), are all equal to the constant percentage produced by the formula. This is why we multiply the percentage by the total number of games to produce the number of Pythagorean wins.

One key observation, and one very relevant to the discussion of the Pythagorean formula, is that the random variables being summed need not be independent for the above property to hold. For example, if we have three random variables X, Y, Z whose sum is a fixed constant C, then we must have

$$E(X) + E(Y) + E(Z)$$
$$= E(X + Y + Z) = E(C) = C$$

where obviously the last equality holds because every constant is equal to its own expectation.

The Pythagorean formula fails to adhere to this property. In particular, it fails when we consider the Pythagorean win totals of *all* teams in a given league. The expectation of the sum of their win totals must be a constant (the number of total games played in the entire league); but the sum of the Pythagorean win totals need not be equal to that same constant. A "toy" example is shown below.

2.1. An example using pythagorean wins

Suppose we have three teams, which for simplicity we will just call 1, 2, and 3. They play a round-robin of one game apiece against each other team, so each team plays 2 games total. Suppose that the results are as follows: Team 1 beats Team 2 by a score of 1-0; Team 1 beats Team 3 by a score of 2-1; and Team 3 beats Team 2 by a score of 3-1. That means we have:

$$R_1 = 3, r_1 = 1$$

$$R_2 = 1, r_2 = 4$$

$$R_3 = 4, r_3 = 3$$

We can use these numbers to compute the Pythagorean winning percentage of each team:

$$P_1 = \frac{3^2}{3^2 + 1^2} = \frac{9}{10}$$

$$P_2 = \frac{1^2}{1^2 + 4^2} = \frac{1}{17}$$

$$P_3 = \frac{4^2}{4^2 + 3^2} = \frac{16}{25}$$

Since each game results in one win and one loss for each team playing, the average winning percentage is supposed to be one-half, or .500. That means that we should expect the three fractions above to add up to 3/2. However, the actual sum is $\frac{1359}{850}$.

If we want to consider Pythagorean wins instead of percentages, all we would do is multiply each of those fractions by 2 before adding them. Since we are multiplying by the same constant, this is the same as if we multiplied the final answer by 2, which clearly shows that the total number of Pythagorean wins for the three teams is equal to $\frac{1359}{425}$. Since this fraction is greater than 3, the Pythagorean wins statistic has claimed that the three teams "should have won" more games than were actually played between all of them.

Remark 1. This was a simplified example to illustrate the math, but does it carry over into real-life computations for large samples? The data in the Appendix confirm very clearly that it does. In three out of five cases, the Pythagorean formula claims that the teams "should have won" more games than were played, and in the other two, it claims that they "should have won" fewer. It never actually recovers the exact total number of games played.

2.2. The mathematics behind the example

The example illustrates the following mathematical dilemma: suppose we are given any amount of ordered pairs

$$(R_1, r_1), (R_2, r_2), \ldots (R_N, r_N)$$

and we are told that they are related in the following way:

$$R_1 + R_2 + \ldots + R_N = r_1 + r_2 + \ldots + r_N$$

This is clearly the case if these ordered pairs represent the runs scored and runs allowed totals of baseball teams playing in the same league, because whenever one team scores a run it must be allowed by someone else, though it may be uncertain which other team allowed the run.

The problem, clearly shown in the example, is that when we fix an exponent $a \neq 0$ and take a sum of the form

$$\sum_{j=1}^{N} \frac{R_j^a}{R_j^a + r_j^a}$$

we have no idea, *a priori*, what that sum will be—but for the Pythagorean wins statistic to make sense, it *should* be $N/2$.

We of course have no assurance that this will ever be the case, and this is why the Pythagorean wins statistic runs afoul of the basic expectation property. The equal relationship between total runs scored and total runs allowed cannot be used to prove anything about what the sum must be equal to, because the individual denominators can vary so much.

3. The improvement

3.1. A special case

There is one time, however, when there is no problem at all, and we actually can prove that the sum equals $N/2$, as we hope it does. That is when there are exactly two teams, that only play each other, and four pieces of data: R_1, r_1, R_2, and r_2. In this case, it is obviously true that $R_1 + R_2 = r_1 + r_2$, and the reason it is so obvious is because in this case, we can say something even stronger: namely, that $R_1 = r_2$ and $R_2 = r_1$. This is clear because when there are only two teams, one team must allow all the runs that the other team has scored, and vice versa.

With this stronger assumption, we can now show that for any exponent a,

$$\frac{R_1^a}{R_1^a + r_1^a} + \frac{R_2^a}{R_2^a + r_2^a}$$

$$= \frac{R_1^a}{R_1^a + r_1^a} + \frac{r_1^a}{r_1^a + R_1^a}$$

$$= 1 (= 2/2)$$

Since the sum is 1, this clearly implies that in this exceptional case, the sum of the two teams' Pythagorean win totals will equal the number of games played between the two teams.

3.2. The pairwise formula

The special case above suggests a simple improvement to the formula: all we have to do is split our runs-scored data so that we are considering Pythagorean wins on a team-by-team basis, and not a cumulative basis over all teams. To be more specific: instead of using a list of ordered pairs of the form (R_j, r_j), where R_j is the total runs scored by team j and r_j is the total runs allowed, we would instead use an entire matrix
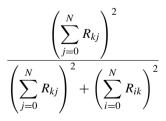
$$R = \begin{pmatrix} 0 & R_{12} & R_{13} & \cdots & R_{1N} \\ R_{21} & 0 & R_{23} & \cdots & R_{2N} \\ R_{31} & R_{32} & 0 & \cdots & R_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_{N1} & R_{N2} & R_{N3} & \cdots & 0 \end{pmatrix}$$

where each entry R_{ij} is the number of runs scored by team i against team j. This matrix includes the cumulative data: for any team k,

$$R_k = \sum_{j=0}^{N} R_{kj}$$

$$r_k = \sum_{i=0}^{N} R_{ik}$$

In other words, the sum of any row is the total number of runs scored by a team, and the sum of any column is the total number of runs allowed by a team. (This is not an original idea; matrices of this form have already been computed for most, if not all, baseball seasons in history. For example, the seasons in (Major League Baseball n.d.) contain such matrices, which is how the data in Section 4 and the Appendix were computed.)

The traditional Pythagorean wins formula is, for any team k, to take the percentage

$$\frac{\left(\sum_{j=0}^{N} R_{kj}\right)^2}{\left(\sum_{j=0}^{N} R_{kj}\right)^2 + \left(\sum_{i=0}^{N} R_{ik}\right)^2}$$

and multiply that by the total number of games played against all teams. Presenting the formula in this way shows that this formula features taking sums first, then raising those sums to a power.

The proposed improvement would do essentially the opposite: for each team k, it would take each percentage

$$p_{kj} := \frac{R_{kj}^2}{R_{kj}^2 + R_{jk}^2},$$

multiply that by the number of games team k played against team j, and *then* add up all those numbers for all other teams j to get the projected win total for team k. As a formula, if we set G_{kj} to be the number of games team k played against team j, the *pairwise Pythagorean projection* of team k's win total is

$$w_k := \sum_{\substack{j=0 \\ j \neq k}}^{N} p_{kj} G_{kj}$$

In other words, instead of the sum over all teams being done first, this method takes individual teams' Pythagorean win totals against other individual teams, and does the sum over all teams last. That explains the "pairwise" part of the name of the formula.

As shown earlier, since R_{jk} is exactly the number of runs allowed by team k against team j, and since $G_{kj} = G_{jk}$, it is necessarily the case that two teams' pairwise Pythagorean win projections against each other will add up to the total number of games played between the two teams:

$$p_{jk} G_{jk} + p_{kj} G_{kj} = G_{kj}(p_{jk} + p_{kj})$$
$$= G_{kj} \cdot 1 = G_{kj}$$

It then follows that the total number of pairwise Pythagorean wins for all teams will sum to the total number of games played between all teams. This is the property that the original formula was lacking.

Remark 2. As pointed out earlier, a key property of an expectation is that the sum of individual components is supposed to equal the total expectation. Earlier we showed that the Pythagorean formula fails this criterion for entire leagues as a whole when broken up by team; but this section implies that it also fails the same criterion for individual teams when broken up by opponent. (For examples, see every computation in the Appendix.) Viewed in this light, the pairwise formula simply replaces the "total" with the sums of the individual components.

Remark 3. We have just proved the fact that the sum of all teams' pairwise Pythagorean win totals in a single league must equal the total number of games played. This is equivalent to stating that if we look at the differences between pairwise Pythagorean wins and actual wins over all teams in a league, those differences must sum to 0. The examples in the Appendix show the sums of differences for five leagues. When we showed earlier that the traditional Pythagorean formula does not necessarily keep the

sum of Pythagorean wins equal to the total number of games played, we were also showing—as is confirmed in the Appendix—that the sum of differences is not necessarily 0, while here we have shown that it is for the pairwise method.

4. The pairwise formula in action

4.1. Pairwise computation: An example

Before delving into the entire data set, it may be illuminating to look at an example of the pairwise computations compared to the traditional formula in a specific case. In this section we will consider one specific team from the Appendix: the 1978 California Angels.

The Angels' real record was 87-75. A look at the Appendix shows that the traditional formula with exponent 2 projected them to win approximately 84 games, meaning they finished about 3 games above their projection. When this happens, it implies the $a = 2$ formula will automatically be more accurate than the traditional $a = 1.83$ formula in this case (proving this assertion is a calculus exercise that we will omit), so for this example we will only examine the pairwise formula with $a = 2$. Another look at the Appendix shows that this formula projected the Angels to win approximately 85.4 games, a higher total than the traditional formula and, because the Angels outperformed both projections, a more accurate prediction. But why exactly is the pairwise formula more accurate for this team?

The answer can be found in the following breakdown of the Angels' row and column of the 1978 AL run matrix (which can be found in (Major League Baseball n.d.)). For each opponent, the table shows their actual win-loss record, runs scored and runs allowed, their Pythagorean winning percentage (to

Opponent	W-L	R	r	P	w
BAL	6-4	33	33	0.5	5
BOS	2-9	35	59	0.2603	2.8634
CLE	6-4	48	32	0.6923	6.9231
DET	4-7	32	51	0.2825	3.1073
MIL	5-5	45	43	0.5227	5.2272
NYY	5-5	39	42	0.4630	4.6301
TOR	7-3	42	35	0.5902	5.9016
CHW	8-7	75	85	0.4377	6.5661
KCR	9-6	70	61	0.5684	8.5257
MIN	12-3	72	51	0.6659	9.9884
OAK	9-6	57	31	0.7717	11.5760
SEA	9-6	81	59	0.6534	9.8003
TEX	5-10	62	84	0.3527	5.2899
Total	87-75	691	666	–	–

4 decimal places), and their Pythagorean win total (to 4 decimal places), all against that opponent.

The last two entries are left out because computing using the total run data and adding up the pairwise win totals are not equal (so we do not choose only one of them to fill the table in). Of interest to us in this case is why the pairwise projection is more accurate, and to help see this, let us focus on two particular lines of the table: the Chicago White Sox and the Oakland Athletics.

The Angels played 15 games each against both of these teams—30 games total. We can see from the table that the Angels finished 17-13 combined against these two teams. It is also easy to read off from the table that the Angels projected to win (approximately) 6.5661 games against the White Sox and 11.5760 games against the A's, for a pairwise total of 18.1421 pairwise Pythagorean wins. This suggests that against these two teams only, the Angels were 1.1421 games below their projection.

What happens when, instead of projecting and then adding first, we combine run totals first? The Angels scored a total of $75 + 57 = 132$ runs against these two teams, and allowed $85 + 31 = 116$. Therefore they would be projected to win

$$30 \cdot \frac{132^2}{132^2 + 116^2} = 16.9275$$

games against these two teams.

Although it appears the traditional Pythagorean formula is far more accurate, this is a red herring. (Remember, overall, the pairwise formula is going to be more accurate over all teams. This is why it is a red herring.) The important fact is that the traditional formula projects the Angels to be more than a game below where the pairwise formula projects them. Why is this?

The reason can be explained as follows: *the Angels showed more dominance in games that were lower-scoring*. Against the A's, their ratio of runs scored to runs allowed is very high—almost 2. Against the White Sox, that ratio is slightly less than 1. However, the number of runs scored *overall* is different. In the 15 games in which the Angels were *not* as dominant—against the White Sox—160 total runs were scored. But against the A's, only 88 total runs were scored! It is now clear why the traditional Pythagorean winning percentage is relatively low—we have "softened" the effect of the games against the A's, in a way claiming that they should matter less because fewer total runs were scored in those games.

To see this clearly, let us imagine that instead of 57-31, the runs scored and allowed totals against the A's were 114-62. The ratio is the same; the pairwise Pythagorean projections would not be affected by this. But the traditional projection clearly will be; now the modified win projection will be

$$30 \cdot \frac{189^2}{189^2 + 147^2} = 18.6923$$

Notice that now, when the totals of runs scored are 160 and 176 instead of 160 and 88, the traditional formula is closer to, though still not equal to, the pairwise formula.

This leads to the following question: which of these should we consider as being more "correct"? Is it more correct to say that we should in fact tilt our projection towards the results in which more total runs were scored? Perhaps we might argue that the distribution of 88 runs can be more subject to "luck" than the distribution of 160 runs, and therefore it should matter less.

However, standing against this argument is the fact that both sets of run totals were amassed in sets of 15 games, the same amount of time. Therefore we might argue instead that there is something about the games between the two different pairs of teams that makes the distribution of runs—and therefore wins—different, and *we should only be combining the data at the latest stage possible*. If we believe this argument, this is a feature of the pairwise Pythagorean formula that is advantageous over the traditional formula. The evidence that the pairwise formula seems to have a tendency to be more accurate overall (see below) lends support to this line of argument as well.

4.2. Pairwise computation: Results

The previous sections attempted to make a theoretical argument that the pairwise formula should have less error than the traditional formula. However, we also have the ability to test the formula with real data. So how does the pairwise method compare to the conventional method at generating predicted win totals? This section will provide evidence towards answering this question.

The table below shows summary data for every Major League season from 1960–1990, excluding 1981 (which had fewer games than other seasons due to a player strike). Next to each league is listed the number of teams in that league, followed by four different numbers. Each of these is the root mean square of the deviations between the teams' actual wins and

their projected wins using one of four formulae. The leftmost column is the root mean square for the traditional Pythagorean formula with exponent 2; next is the root mean square for the traditional Pythagorean formula with exponent 1.83; next is the root mean square for the pairwise Pythagorean formula with exponent 2; and finally, the rightmost column is the root mean square for the pairwise Pythagorean formula with exponent 1.83. The last line contains the total number of teams and all four root mean squares when the teams are considered as a single data set. All figures are rounded to three decimal places.

A few potentially useful observations about the table:

League	Teams	Pyth. $a = 2$	Pyth. $a = 1.83$	p.Pyth. $a = 2$	p.Pyth. $a = 1.83$
1960 AL	8	3.682	4.079	3.540	4.092
1960 NL	8	3.408	3.774	3.366	3.679
1961 AL	10	2.765	3.573	2.748	3.532
1961 NL	8	5.520	5.942	5.171	5.580
1962 AL	10	3.238	3.376	3.771	3.980
1962 NL	10	4.884	5.458	4.954	5.677
1963 AL	10	4.097	3.642	3.507	3.342
1963 NL	10	4.737	4.131	4.934	4.367
1964 AL	10	3.189	3.109	3.115	3.224
1964 NL	10	3.788	3.860	4.121	4.178
1965 AL	10	3.330	3.687	3.698	4.099
1965 NL	10	3.626	3.435	3.260	3.132
1966 AL	10	4.748	4.400	4.643	4.408
1966 NL	10	4.157	3.898	3.623	3.636
1967 AL	10	5.371	4.962	4.836	4.479
1967 NL	10	2.746	2.384	2.994	2.760
1968 AL	10	3.227	2.945	2.752	2.576
1968 NL	10	4.047	3.665	3.466	3.223
1969 AL	12	3.208	2.681	2.744	2.482
1969 NL	12	4.618	4.740	4.032	4.316
1970 AL	12	3.304	3.804	3.570	4.067
1970 NL	12	6.119	5.924	5.973	5.832
1971 AL	12	3.183	3.088	3.313	3.281
1971 NL	12	3.967	3.612	3.998	3.800
1972 AL	12	5.525	5.058	5.639	5.327
1972 NL	12	5.643	5.113	5.463	5.060
1973 AL	12	4.927	4.393	4.361	4.007
1973 NL	12	3.477	3.319	3.588	3.521
1974 AL	12	5.035	4.739	4.610	4.417
1974 NL	12	4.564	3.593	3.002	2.603
1975 AL	12	3.772	3.489	3.617	3.529
1975 NL	12	5.779	5.159	5.481	5.048
1976 AL	12	3.655	3.220	3.549	3.366
1976 NL	12	4.104	3.573	3.399	3.045
1977 AL	14	3.961	4.171	3.816	4.165
1977 NL	12	4.109	4.026	3.839	3.826
1978 AL	14	3.992	3.818	3.536	3.431
1978 NL	12	5.515	5.314	4.942	4.862
1979 AL	14	3.338	3.431	3.326	3.699
1979 NL	12	3.812	4.103	2.825	3.188
1980 AL	14	4.624	4.530	4.226	4.356
1980 NL	12	3.971	4.146	4.191	4.341
1982 AL	14	2.875	2.744	3.042	3.024
1982 NL	12	4.182	4.305	4.162	4.254
1983 AL	14	3.228	3.283	3.116	3.260
1983 NL	12	3.623	3.531	3.571	3.515
1984 AL	14	2.933	3.129	3.277	3.462
1984 NL	12	6.167	6.022	5.816	5.727
1985 AL	14	4.161	4.077	3.955	3.973
1985 NL	12	3.591	3.962	4.202	4.591
1986 AL	14	2.780	2.944	2.852	3.028
1986 NL	12	5.017	5.129	5.232	5.337
1987 AL	14	3.636	3.600	3.336	3.457

League	Teams	Pyth. $a = 2$	Pyth. $a = 1.83$	p.Pyth. $a = 2$	p.Pyth. $a = 1.83$
1987 NL	12	3.756	3.847	3.886	3.988
1988 AL	14	3.019	2.805	2.692	2.668
1988 NL	12	2.421	2.571	2.086	2.451
1989 AL	14	3.046	3.175	3.133	3.369
1989 NL	12	4.024	4.100	3.819	3.847
1990 AL	14	3.778	3.773	3.686	3.749
1990 NL	12	4.371	3.851	4.209	3.812
Total	704	4.105	4.004	3.938	3.951

- When all teams are considered together and we look at the overall root mean square, *both* pairwise methods have lower root mean square error than *either* traditional method, with the root mean square for exponent 2 being slightly lower than for exponent 1.83.
- The 1969 season marks the beginning of divisional play, where each team did not play all opponents the same number of times. If we split our data by pre- and post-divisional play, we find the following:

Years	Teams	Pyth. $a = 2$	Pyth. $a = 1.83$	p.Pyth. $a = 2$	p.Pyth. $a = 1.83$
1960–1968	174	3.993	3.969	3.873	3.952
1969+	530	4.142	4.015	3.958	3.950
Total	704	4.105	4.004	3.938	3.951

So in the first 10 years of the data set, the pairwise Pythagorean formula with exponent 2 has less overall error than the other three methods. However, once divisional play begins, the traditional methods show more error while there is now little to choose from between the two pairwise methods, except that they show less error than the traditional methods do. In both parts of the data set, however, the pattern of both pairwise methods showing less error than either traditional method still holds.

- Out of the sample of 60 seasons, there were 22 seasons in which both pairwise methods showed less error than either traditional method. To contrast, there were only 12 seasons in which both traditional methods showed less error than either pairwise method. (Out of the other 26 seasons, one of the pairwise methods showed the least error in 15 of those.)

In the Appendix, five of these 60 seasons are broken down team-by-team to give the reader a sense of how the formulae compare for individual teams, and also to numerically demonstrate some of the assertions made in the previous sections.

5. Conclusion

The above has been meant to introduce and show the potential advantages of using a pairwise method to project wins based on run totals as opposed to the traditional cumulative method. Not only is it a mathematical improvement, since the traditional method cannot output a mathematical expectation for a reason that does not apply to the pairwise method, but it also seems (at least over our 30-year sample) to reduce the error in the results. However, the treatment it has received in this paper is by no means comprehensive. Many questions remain to be answered.

One such question is: if we were to use a pairwise Pythagorean formula with a fixed exponent, what exponent is most accurate? In this paper we compare the fixed exponents 2 and 1.83, and our sample seems to show that there is very little difference between the two. However, clearly there is no reason to believe the best exponent must be one of these. Since we are changing our method of projection, it may be the case that an entirely different exponent is best to use. It may in fact even be larger than 2; there is simply not enough data yet to conclude anything in this regard. This is a question that, with enough data analysis, can easily be settled in the future.

Another remaining question concerns the variations on the Pythagorean formula that do not use fixed exponents. There are some methods in which the exponent is a function of the other variables, and not constant over all teams—even all teams in the same league. In this paper we have not considered any of those, but this is another topic that is ripe for future analysis. Are they subject to the same mathematical flaw as fixed-exponent formulae? We have not answered that question here, but regardless of the answer, such formulae can also be mixed with the pairwise method introduced here. The formula

$$w_k = \sum_{\substack{j=0 \\ j \neq k}}^{N} p_{kj} G_{kj}$$

can be combined with any formula for computing the winning percentages p_{kj}. The pairwise method is meant to be a mathematical adjustment, as highlighted in this paper, and could be so for an extremely wide variety of formulae (namely any in which $p_{kj} + p_{jk} = 1$ for all indices j and k), but it is obviously still unclear whether switching to a pairwise method always reduces the error in any formula.

Those questions are beyond the scope of this paper, whose purpose was simply to highlight that Pythagorean wins cannot be a true expectation and introduce the pairwise method as a way of improving that. Now that it has been introduced, there is a great deal of future research that can possibly be done to analyze it.

A. Pythagorean projection data

Below are data from five different seasons (one league apiece) in baseball history. All five seasons predate interleague play (meaning that every team played all of its regular-season games against some other team listed). Above we gave summary data for 60 seasons; here we have chosen five of those seasons and given more complete data.

For each team, five different data points are given. The first is the team's actual win-loss record. The last four are differences between actual wins and projected wins, where each column uses a different method of projection. We first consider actual wins minus projected wins using the Pythagorean formula with exponent 2; then the Pythagorean formula with exponent 1.83; then the pairwise Pythagorean formula with exponent 2; and finally the pairwise Pythagorean formula with exponent 1.83.

In each case the differentials are shown rounded off to four decimal places. This sometimes creates a situation where the sum of the differentials of the pairwise Pythagorean projections in a single season appears to be very close to, but not exactly equal to, zero. This is caused by rounding error—we proved above that the sum must actually be 0. Rounding error is *not* the cause of the fact that the sums of the differentials of the traditional Pythagorean projections are nonzero.

All tables were computed using data from (Major League Baseball n.d.).

1960 National League

Team	W-L	Pyth. $a = 2$	Pyth. $a = 1.83$	p.Pyth. $a = 2$	p.Pyth. $a = 1.83$
CHC	60-94	−1.6465	−2.9209	−1.5393	−2.8042
CIN	67-87	−3.9971	−4.5056	−3.9896	−4.4639
LAD	82-72	−3.4414	−2.7289	−3.0884	−2.4455
MLN	88-66	3.6622	4.2826	3.8993	4.4559
PHI	59-95	−0.1929	−1.6586	0.5133	−0.8995
PIT	95-59	1.8195	3.1589	1.8329	3.0980
SFG	79-75	−2.7267	−2.3258	−3.3521	−2.9195
STL	86-68	6.1786	6.4183	5.7239	5.9788
Total	616-616	−0.3443	−0.28	0	0

1961 American League

Team	W-L	Pyth. $a = 2$	Pyth. $a = 1.83$	p.Pyth. $a = 2$	p.Pyth. $a = 1.83$
BAL	95-67	1.0379	2.1231	1.4284	2.3037
BOS	76-86	1.6986	1.1315	1.2834	0.7984
CHW	86-76	0.7655	1.1248	0.6976	1.0103
CLE	78-83	−0.8783	−1.0161	−0.7281	−0.8434
DET	101-61	2.0131	3.4974	1.9674	3.3877
KCA	61-100	−1.0056	−2.5284	−0.3820	−1.7752
LAA	70-91	−6.2882	−6.6457	−5.7184	−6.1147
MIN	70-90	−2.3676	−3.0129	−2.2437	−2.8501
NYY	109-53	4.3242	6.2339	5.1806	6.9709
WSA	61-100	−1.4832	−2.9693	−1.4852	−2.8876
Total	807-807	−2.1836	−2.0617	0	0

1962 National League

Team	W-L	Pyth. $a = 2$	Pyth. $a = 1.83$	p.Pyth. $a = 2$	p.Pyth. $a = 1.83$
CHC	59-103	–0.7282	–2.4621	–2.3515	–3.8253
CIN	98-64	4.3320	5.3933	3.6260	4.6084
HOU	64-96	–0.8592	–2.1190	–2.3515	–3.8253
LAD	102-63	4.0910	5.3738	6.0673	6.9562
MLN	86-76	–2.5320	–1.8951	–1.1693	–0.6884
NYM	40-120	–7.6087	–10.0859	–8.9865	–11.2791
PHI	81-80	6.4305	5.9280	5.3386	4.9331
PIT	93-68	2.8651	3.6772	4.2656	4.8898
SFG	103-62	0.9972	2.6000	2.0944	3.5050
STL	84-78	–9.3201	–8.2871	–7.3836	–6.5957
Total	810-810	–2.3325	–1.8770	0	0

1968 National League

Team	W-L	Pyth. $a = 2$	Pyth. $a = 1.83$	p.Pyth. $a = 2$	p.Pyth. $a = 1.83$
ATL	81-81	5.3282	4.8764	4.0596	3.7480
CHC	84-78	2.8675	2.8788	2.89	2.9022
CIN	83-79	–0.0202	0.1514	0.3976	0.5314
HOU	72-90	2.4504	1.4885	0.4927	–0.2061
LAD	76-86	1.4433	0.8976	0.7888	0.3233
NYM	73-89	–3.6698	–4.0372	–4.4306	–4.7220
PHI	76-86	5.0337	4.1886	5.1407	4.3150
PIT	80-82	–8.3944	–7.7689	–6.2947	–5.8566
SFG	88-74	–3.0146	–2.1710	–2.9328	–2.1322
STL	97-65	–0.8579	0.5383	–0.1112	1.0970
Total	810-810	1.1662	1.0425	0	0

1978 American League

Team	W-L	Pyth. $a = 2$	Pyth. $a = 1.83$	p.Pyth. $a = 2$	p.Pyth. $a = 1.83$
BAL	90-71	6.2614	6.5364	6.6958	6.7591
BOS	99-64	2.0482	3.3337	2.2984	3.4279
CLE	69-90	–3.9508	–4.5052	–2.7771	–3.3852
DET	86-76	–2.2146	–1.6042	–1.3944	–0.9821
MIL	93-69	–4.9768	–3.5630	–5.4255	–4.2730
NYY	100-63	–0.1841	1.3546	–0.4226	1.0049
TOR	59-102	–0.0731	–1.8175	–0.5278	–2.0461
CAL	87-75	3.0165	3.2699	1.6008	1.9337
CHW	71-90	1.8835	0.9273	0.4877	–0.1539
KCR	92-70	–1.7437	–0.6762	–1.1374	–0.2445
MIN	73-89	–6.5537	–6.6766	–5.8328	–5.9806
OAK	69-93	8.6016	6.9178	6.7575	5.5221
SEA	56-104	–0.2391	–2.1522	0.7152	–1.061
TEX	87-75	–1.3263	–0.7066	–1.0377	–0.5214
Total	1131-1131	0.5580	0.6382	0	0

Acknowledgments

References

Cochran, J.J., Blackstock, R., 2009. Pythagoras and the national hockey league, Journal of Quantitative Analysis in Sports 5(2). Available from: de Gruyter. [27 July 2015].

Davenport, C., Woolner, K., 1999. Revisiting the Pythagorean Theorem. Available from: <http://www.baseballprospectus.com/article.php?articleid=342>. [27 July 2015].

Hamilton, H.H., 2011. An extension of the pythagorean expectation for association football, Journal of Quantitative Analysis in Sports 7(2). Available from: de Gruyter. [27 July 2015].

n.d. Major League Baseball & MLB Encyclopedia. Available from: <http://www.baseball-reference.com/leagues/>. [27 December 2015].

Miller, S.J., 2007. A derivation of the pythagorean won-loss formula in baseball, CHANCE Magazine 20(1), 40–48.

n.d. Pythagorean Theorem of Baseball. Available from: <http://www.baseball-reference.com/bullpen/Pythagorean_Theorem_of_Baseball>. [27 July 2015].