The Optimal Value and Potential Alternatives of Bill James' *Pythagorean Method of Baseball*

ABSTRACT: The nature of the relationship Bill James found between the *win/loss percentage* of a Major League Baseball (MLB) team and the number of runs the team scores and allows over the course of a season is investigated. A variety of alternatives have been considered and evaluated using the squared error criterion. In this paper we find the optimal form of James' model using the absolute error criterion and demonstrate that far more complex forms of James' model yield little in additional predictive power. We also provide empirical evidence that the relationship between *win/loss percentage* and runs scored and allowed is similar in the two leagues (National League and American League) that comprise MLB. Finally, our results suggest that annual variations in the single exponent for James' Pythagorean Method that minimizes absolute errors has no discernable pattern by league.

Keywords: Sports, Operations Research, Baseball, Mathematical Programming, Optimization, Pythagorean Method.

Corresponding Author
James J. Cochran
Ruston Building & Loan Endowed Research Professor
Department of Marketing and Analysis
PO Box 10318
Louisiana Tech University
Ruston LA USA 71272-0046
(318) 257-3445 (office),
(318) 257-4253 (fax)
jcochran@cab.latech.edu

1. INTRODUCTION

In an early volume of the *The Bill James Abstract*, Bill James (1979) postulated a strong relationship between the *win/loss percentage*

$$\frac{\text{number of wins}}{\text{number of wins + number of losses}}$$

of a baseball team and the difference between the number of runs the team scores and allows over the course of a season. James reasoned that if a team allows fewer runs than it scores, it should be expected to win more games than it looses; if a team allows more runs than it scores, it should be expected to win fewer games than it looses; and a team that scores and allows roughly the same number of runs it allows should be expected to win and loose roughly the same number of games.

Eventually James (1980) found a strong relationship between a baseball team's *win/loss percentage* and a functional form of the runs it scores and allows over the course of a season. The relationship he found is

$$\text{win/loss percentage} \approx \frac{\text{runs scored}^2}{\text{runs scored}^2 + \text{runs allowed}^2}$$

James named his approach the Pythagorean Method because of its similarity to the Pythagorean Theorem (which states that the squared length of the hypotenuse is equal to the sum of the squared lengths of the legs for any right triangle).

This estimate quickly became common in baseball literature; not only did James regularly use it in ensuing volumes of *The Bill James Abstract* (1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989), but many other books and papers on statistics and baseball did likewise. For recent examples, see Acharya (2006), Adler (2006), Perry (2006), Tango et al. (2006), and Keri (2007). Others have developed similar methods for American football/the National Football League (Schatz, 2003a;Schatz, 2003g; Gietschier, 2005; Barnwell, 2007; Melton, 2007), U.S. Collegiate basketball (Logan, 2008), and professional basketball/the National Basketball Association (Oliver, http://www.rawbw.com/~deano/helpscrn/pyth.html).

Based on empirical evidence, James ultimately revised his formula by altering the exponent assigned to *runs scored* and *runs allowed* to 1.82, i.e.,

$$\text{win/loss percentage} \approx \frac{\text{runs scored}^{1.82}}{\text{runs scored}^{1.82} + \text{runs allowed}^{1.82}}$$

Davenport and Woolner (1999) find a least squares estimate of 1.87 over all MLB teams that have played at least 120 games in a season. BaseballReference.com (http://www.baseball-reference.com/about/faq.shtml) uses an exponent of 1.83. Miller (2006) provides a theoretical justification of the functional form of James' Pythagorean under the assumption that runs scored and allowed by a team are independent Weibull random variables. He applies his results to the 2004 American League season and finds

the corresponding least squares and maximum likelihood estimates of the exponent to be 1.79 and 1.74, respectively. Each of these estimates is reasonably close to James' adjusted value of 1.82. An interesting article by Tung on using parametric and nonparametric bootstrapping to develop confidence intervals for James' Pythagorean Method is available at www.uweb.ucsb.edu/~utungd00/pythagCI.pdf.

James' Pythagorean Method has gained acceptance by MLB and its fans. MLB's official website (http://mlb.mlb.com/mlb/) reports the results of James' Pythagorean Method for an exponent of 1.82; the site refers to the value as the *Expected Won-Loss Record Based on Runs Scored and Runs Allowed* (XW-L). ESPN (the Entertainment and Sports Network, a U.S. cable television channel) also reports the results of James' Pythagorean Method, albeit based on an exponent of 2.00, and refers to it as the *Expected Winning Percentage* (ExWP) on its website (http://sports.espn.go.com/mlb/). The method is also featured on the Baseball Investor (http://baseballinvestor.com/), a website that tracks performances of MLB teams in a manner similar to stocks on the DJSE.

While many approaches have been taken to finding the optimal value of the exponent in James' Pythagorean Method, each is based on squared errors; none attempts to minimize the actual or absolute errors. Furthermore, while many have suggested alternative functional forms, none of these approaches allows for different coefficients or exponents for the three terms of the James' basic formula. Both of these issues will be addressed in this paper.

2. FINDING THE EXPONENT THAT MINIMIZES SQUARED AND ABSOLUTE ERRORS

Data collected by Cochran (2002) for each MLB team from every season from 1901 through 2000 are updated through 2004 and utilized in this analysis; this provides a total of 2,084 observations (combinations of teams and seasons). We next find the exponents that minimize the squared and absolute errors using James' Pythagorean Method over these data.

If the goal is to find the value of the exponent that minimizes the squared error committed when using James' Pythagorean Method, the formulation is

$$\begin{aligned} \min \quad & \sum_{t=1901}^{2004} \sum_{k=1}^{2} \sum_{i=1}^{n_{kt}} \left(wlpct_{ikt} - E\left[wlpct_{ikt} \right] \right)^2 g_{ikt} \\ st \quad & E\left[wlpct_{ikt} \right] = \frac{rs_{ikt}^{x}}{rs_{ikt}^{x} + ra_{ikt}^{x}} \\ & x \geq 0 \end{aligned} \tag{1}$$

where t=1901,…, 2004 is the index of the season
k=1,2 is the index of the league (National League or American League)
i=1,…n_{kt} is the index of the team
$wlpct_{ikt}$ is the win-loss percentage for the i^{th} team of league k during season t
g_{ikt} is the number of games played by i^{th} team of league k during season t

rs_{ikt} is the number of runs scored by the i^{th} team of league k during season t
ra_{ikt} is the number of runs allowed by the i^{th} team of league k during season t

The optimal objective value occurs when x = 1.8606; this is similar to the results achieved by Davenport & Woolner (1999), with a slight discrepancy due to differences in the periods of data considered. Over the data utilized in this analysis, this model has a mean error of 3.2183 games per team per season.

On the other hand, if the goal is to find the value of the exponent that minimizes the absolute error committed when using James' Pythagorean Method, the formulation is

$$\begin{aligned} min \quad & \sum_{t=1901}^{2004} \sum_{k=1}^{2} \sum_{i=1}^{n_{kt}} \left| wlpct_{ikt} - E\left[wlpct_{ikt}\right] \right| g_{ikt} \\ st \quad & E\left[wlpct_{ikt}\right] = \frac{rs_{ikt}^{x}}{rs_{ikt}^{x} + ra_{ikt}^{x}} \\ & x \geq 0 \end{aligned} \tag{2}$$

The optimal objective value occurs when x = 1.8752. The coefficient that minimizes absolute error is slightly larger that the coefficient that minimizes squared error and yields a slightly smaller mean error of 3.2177 games per team per season.

Which objective produces superior estimates of *win/loss percentage*? Obviously, if the goal is to minimize the magnitude of the total errors, minimization of absolute errors is superior by definition. However, the scatter diagrams in *Figure 1* suggest that the relationships between actual *win/loss percentages* and i) estimates produced using absolute error and ii) estimates produced using squared error are almost identical.

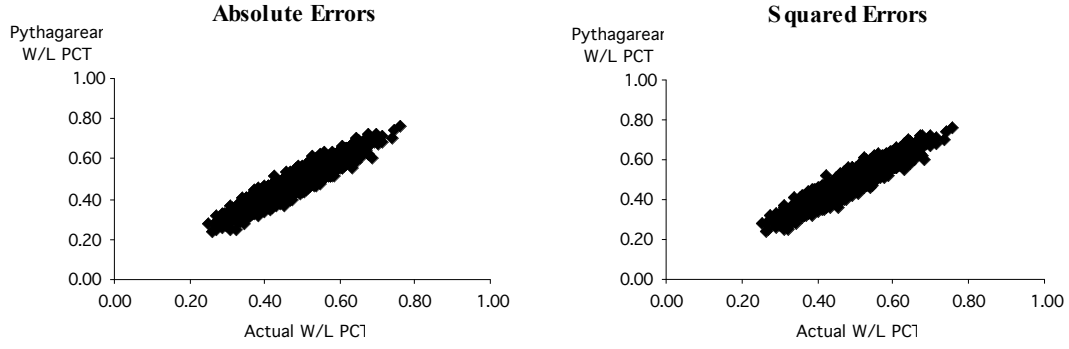

Figure 1: Scatter Diagrams of Actual and Expected *Win/Loss Percentages* Obtained using Absolute Error and Squared Error

Pearson's correlation coefficients support this conclusion. The Pearson's correlation with actual *win-loss percentages* is 0.951197 for estimates produced using squared error and 0.951224 for estimates produced using absolute error.

A frequency distribution of estimates obtained using absolute error and squared error are also very similar; results of Chi-Square goodness-of-fit tests of the frequency distribution of estimates obtained using absolute error and squared error to the actual frequencies of *win-loss percentage* are virtually identical.

Class	Actual Frequencies	Expected Frequencies	
		Absolute	Squared
>0.325	45	40	40
0.350-0.375	52	32	31
0.350-0.375	59	68	68
0.375-0.400	99	100	96
0.400-0.425	149	151	153
0.425-0.450	169	171	172
0.450-0.475	213	196	197
0.475-0.500	202	239	240
0.500-0.525	250	240	245
0.525-0.550	244	254	251
0.550-0.575	200	215	218
0.575-0.600	158	151	152
0.600-0.625	125	109	106
0.625-0.650	66	63	60
0.650-0.675	26	29	32
0.675<	27	18	15
χ^2 test statistic		31.0617	40.3495
p-Value		0.008619	0.000402

Table 1: Frequency Distributions Scatter Diagrams of Actual and Expected *Win/Loss Percentages* Obtained using Absolute Error and Squared Error

Given the expected *win/loss percentages* obtained using absolute error and squared error performances are virtually identical, the more easily explained absolute error criterion should be used.

3. ALLOWING FOR AN ADDITIVE CONSTANT, DIFFERENT COEFFICIENTS, AND DIFFERENT AND EXPONENTS

Several functional variations on James' Pythagorean Method have been proposed. Vollmayr-Lee (2002) and Jones and Tappin (2005) essentially regressed the difference between runs scored and runs allowed against *win/loss percentage* with the intercept set to 0.500 (the expected win/loss percentage when *runs scored* and *runs allowed* are equal). i.e.,

$$\text{win/loss percentage} \approx 0.50 + b(\text{runs scored - runs allowed}).$$

For the 1969-2003 seasons, Jones and Tappin report optimal slopes b by season that range from 0.00053 to 0.00078 (with a mean of 0.00065).

Davenport and Woolner (1999) compare their Pythagenport method with the Smyth/Patriot (Pythagenpat) method, each of which addresses assumption that scoring twice as many runs as your opponents always results in the same won/lost percentage regardless of the absolute difference in runs scored and runs allowed, which is implicit in James' model. These models allow the exponent to vary on the basis of the total number of runs scored and allowed per game rather than the fixed exponent utilized in traditional variations on the James' Pythagorean Method. While both models perform slightly better than variations on James' Pythagorean Method for extreme cases (teams that either win or loose an unusually large proportion of their games in a season), the authors concede that Pythagenpat outperforms the Pythagenport model.

In this section we will generalize James' method to allow direct comparisons to Vollmayr-Lee (2002) and Jones and Tappin (2005) as well as to other more complex potential variations. In order to generalize James' Pythagorean method, we add a term for runs allowed in the numerator, allow for interaction between runs scored and runs allowed in both the numerator and denominator, and allow each term to have a unique coefficient and (nonnegative) exponent. This generalization does not extend to the Pythagenpat and Pythagenport model; the advantages of these models are minimal and are outweighed by the unnecessary complexity they introduce. Using the same data (records for each team from every season from 1901 through 2004), we calculate the constant, coefficients, and exponents that minimize the squared and absolute errors using James' Pythagorean Method.

If the goal is to find the combination of additive constant, coefficients, and exponents that minimize the absolute error committed when using James' Pythagorean Method, the formulation is

$$\begin{aligned}
&\min \sum_{t=1901}^{2004} \sum_{k=1}^{2} \sum_{i=1}^{n_{kt}} \left| wlpct_{ikt} - E\left[wlpct_{ikt}\right] \right| g_{ikt} \\
&st \quad E\left[wlpct_{ikt}\right] = x_0 + \frac{x_1 rs_{ikt}^{x_2} + x_3 ra_{ikt}^{x_4} + x_5 \left(rs_{ikt} ra_{ikt}\right)^{x_6}}{x_7 rs_{ikt}^{x_8} + x_9 ra_{ikt}^{x_{10}} + x_{11} \left(rs_{ikt} ra_{ikt}\right)^{x_{12}}} \qquad (3) \\
&x_j \in \Re, j = 1,...12 \\
&\quad x_j \geq 0, j = 0,2,4,6,8,10,12
\end{aligned}$$

The optimal solution

x_0 = 1.47705 e-6	x_1 = 1.095581874
x_2 = 2.911916580	x_3 = 9.988621981
x_4 = 2.334950287	x_5 = 3,104059374
x_6 = 1.001633502	x_7 = 4.012091473

$x_8 = 2.749535061$ $x_9 = 1.685535213$
$x_{10} = 2.869184615$ $x_{11} = 3.673237039$
$x_{12} = 0.901359869$

i.e.,

$$E[wlpct_{ikt}] = 0.0 + \frac{1.1rs_{ikt}^{2.9} + 10.0ra_{ikt}^{2.3} + 3.1(rs_{ikt}ra_{ikt})^{1.0}}{4.0rs_{ikt}^{2.8} + 1.7ra_{ikt}^{2.9} + 3.7(rs_{ikt}ra_{ikt})^{0.9}} \quad (4)$$

results in a mean error of 3.186 games per team per season. This is the lower bound for all models that include an additive constant and terms for runs scored, runs allowed, and interaction between runs scored and runs allowed in both the numerator and denominator with unique coefficients and exponent for each term. Thus, this is the lower bound for the Vollmayr-Lee (2002) and Jones and Tappin (2005) models which are specific cases of (3). The improvement this model achieves over the performance of the basic James Pythagorean Method (error of 3.2177 games per team per season) with the optimal single coefficient of 1.8752 hardly justifies its additional complexity. Furthermore, the gap between the performance of this model and the James' Pythagorean Method suggests that the methods proposed by Vollmayr-Lee (2002) and Jones and Tappin (2005) do not provide substantial improvement over the James' Pythagorean Method.

4. DOES THE EXPONENT THAT MINIMIZES ABSOLUTE ERRORS DIFFER ACROSS LEAGUES?

Prior to 1901 MLB consisted of eight teams that were collectively referred to as the National League (NL). This league had several challengers but was generally accepted as the only *major* league (all other leagues were thought to be inferior). However, one of these competitor leagues, the American League (AL), rose to prominence and successfully competed with the NL for major league status. In an effort to end competitive bidding for players (and so suppress salaries), the two leagues united in 1901 to form MLB. The two leagues maintained separate identities and regular season games were not played across leagues until 1997, when limited interleague play was introduced into the regular season schedule (to increase fan interest).

Given the limited regular season play between teams from the two leagues, it is reasonable to ask if the exponent that minimizes absolute errors using James' Pythagorean Method is the same for the NL and AL. To assess this issue, the MLB data for each team from every season from 1901 through 2004 are again utilized. We resolve (1) after restricting the data to NL teams, then repeat the process using only AL teams.

For the NL the optimal exponent is 1.845965721, while the optimal exponent for the AL is 1.900180434. The resulting mean errors (in games per team per season) are 3.33247 for the NL model and 3.10042 for the AL model. While this difference is not dramatic, it does naturally lead one to ask if there is another difference between the leagues, such as a difference in how the game is played, that might explain this difference.

The two leagues play under essentially the same rules with one exception; in 1973 the AL adopted the *designated hitter* rule. This rule allows the team manager to designate a player who will bat in place of the generally weak-hitting pitcher (again, a modification designed to stimulate fan interest).

To assess the issue of whether the exponent that minimizes absolute errors using James' Pythagorean Method changes over time for the NL or AL, the MLB data for each team from every season from 1901 through 2004 are again utilized. We resolve (2) for each season from 1901 through 2004 after restricting the data to NL teams, then repeat the process using only AL teams. The graphs of optimal exponents across seasons are provided for the NL in **Figure 2** and the AL in **Figure 3**.

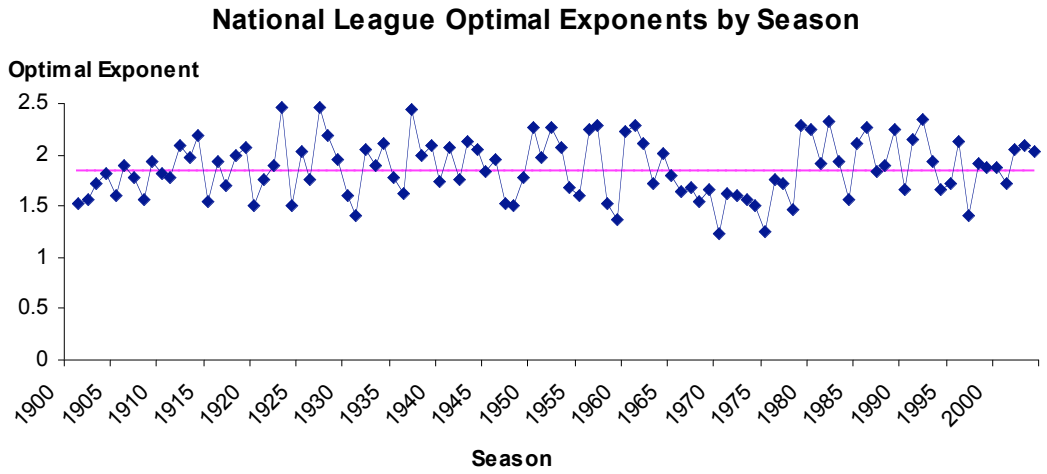

Figure 2: Optimal Exponent for the NL by Season Using the Absolute Error Criterion, 1901 – 2004 (the center line represents the optimal NL exponent of 1.845965721)

Other than an odd drop in the mid 1960s through the late 1970s, the NL graph has the appearance of a random walk, which suggests that no structural change in the relationship *win/loss percentage* and runs scored and allowed under the absolute error criterion has occurred over time in the NL.

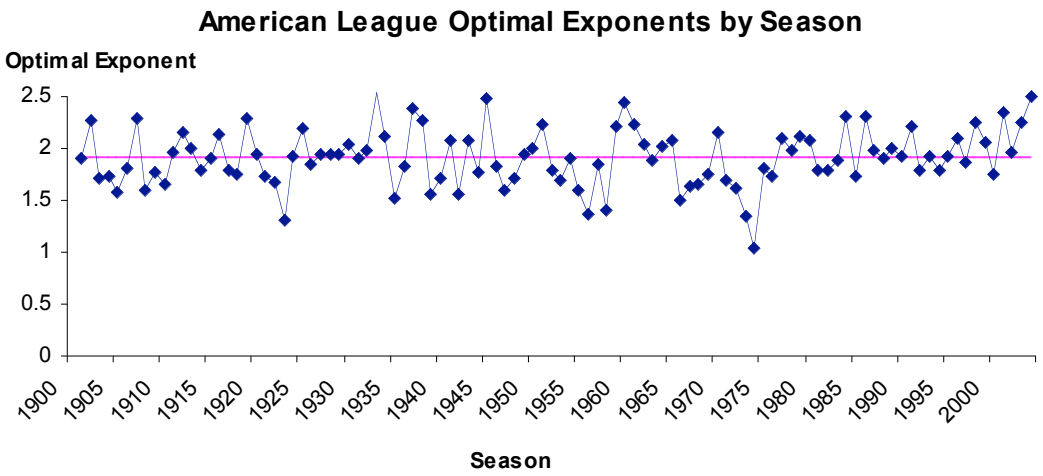

Figure 3: Optimal Exponent for the AL by Season Using the Absolute Error Criterion, 1901 – 2004 (the center line represents the optimal AL exponent of 1.900180434)

The AL graph also has the appearance of a random walk, which suggests that AL adoption of the designated hitter rule in 1973 did not result in a structural change in the relationship *win/loss percentage* and runs scored and allowed under the absolute error criterion.

5. CONCLUSIONS

Bill James' Pythagorean Method is a remarkably effective means for estimating the win/loss percentage of a MLB team. Several variations have been suggested; these variations have consistently been evaluated using squared errors. We have found the form of James' Pythagorean Method that minimizes absolute error. Although this results in a small difference in the accuracy of the final methodology, the alteration is based on a simpler justification that should be more easily understood by sports fans with a casual interest in mathematics.

The empirical evidence provided in this paper also supports the conclusion that the relationship between *win/loss percentage* and runs scored and allowed is similar in the NL and AL. Furthermore, the results indicate that annual variations in the single exponent for James' Pythagorean Method that minimizes absolute errors has no discernable pattern by league, which suggests the various changes in strategies and rules that have occurred over the period of data utilized in the analysis (1901 - 2004) have not altered the fundamental relationship between *win/loss percentage* and runs scored in either league.

Future analysis should focus on i) extending optimization to an even broader class of functional forms and ii) developing a model for other professional sports such as World football, hockey, and collegiate American football.

REFERENCES

Acharya, R.A. Brief Introduction to the "Pythagorean Theorem," *Harvard Sports Analysis Collective*, September 20, 2006, www.hcs.harvard.edu/hsac/Resources/pythagorean.pdf.
Adler, J. *Baseball Hacks*, O'Reilly Media, Sebastopol, CA, 2006.
Barnwell, B. *Does Peyton Manning Break Pythagoras?* 2007, http://www.footballoutsiders.com/2007/08/27/extra-points/5397/.
BaseballReference.com, http://www.baseball-reference.com/about/faq.shtml;
Cochran, J.J., Data Management, Exploratory Data Analysis, and Regression Analysis with 1969-2000 Major League Baseball Attendance Data, The Journal of Statistics Education, 10(2), 2002.
Davenport, C. and Woolner, K. Revisiting the Pythagorean Theorem: Putting Bill James' Pythagorean Theorem to the test, *The Baseball Prospectus*, June 30, 1999, http://www.baseballprospectus.com/article.php?articleid=342.
ESPN.com, http://sports.espn.go.com/mlb/.
Gietschier , S. A New Spin on the Pythagorean Theorem, *Sporting News.Com*, 2005, http://www.sportingnews.com/features/stats/pythagorean.html.
James, B. *The Bill James Abstract*, Ballantine Books, 1980.
James, B. *The Bill James Abstract*, Ballantine Books, 1981.
James, B. *The Bill James Abstract*, Ballantine Books, 1982.
James, B. *The Bill James Abstract*, Ballantine Books, 1983.
James, B. *The Bill James Abstract*, Ballantine Books, 1984.
James, B. *The Bill James Abstract*, Ballantine Books, 1985.
James, B. *The Bill James Abstract*, Ballantine Books, 1986.
James, B. *The Bill James Abstract*, Ballantine Books, 1987.
James, B. *The Bill James Abstract*, Ballantine Books, 1988.
James, B. *The Bill James Abstract*, Ballantine Books, 1989.
Jones, M.A. and Tappin, L.A. *The UMAP Journal* 25(2), 2005, 25-36.
Keri, J. (ed) *Baseball Between the Numbers: Why Everything You Know about the Game Is Wrong*, Publisher: Perseus Publishing, 2007.
Logan, G. *The SEC Tournament Deconstructed, and a Kudos to the Deserving*, 2008, http://www.aseaofblue.com/story/2008/3/11/19396/4687.
Melton, M. *Building a Better Mousetrap: Adjusted Pythagorean Winning Percentage*, 2007, http://leftyloon.blogspot.com/2007_07_01_archive.html.
Miller, S.J. A Derivation of the Pythagorean Won-Loss Formula in Baseball, *By the Numbers*, 16(1), 2006.
MLB.com, http://mlb.mlb.com/mlb/.
Oliver, L. D. Pythagorean 16.5 Method, http://www.rawbw.com/~deano/helpscrn/pyth.html.
Perry, D. *Winners: How Good Baseball Teams Become Great Ones (And It's Not the Way You Think)*, Wiley, Hoboken, 2006.
Schatz, A. (a) Pythagoras on the Gridiron, 2003, http://www.footballoutsiders.com/2003/07/14/ramblings/stat-analysis/4/.
Schatz, A. (b) Pythagoras Grades the Coaches, 2003, http://www.footballoutsiders.com/2003/08/07/ramblings/stat-analysis/19/.

Tango, T.M., Lichtman, M. and Dolphin, A. *The Book: Playing the Percentages in Baseball*, TMA Press, 2006.
Tung, D.D. Confidence Intervals for the Pythagorean Formula in Baseball, www.uweb.ucsb.edu/~utungd00/pythagCI.pdf.
Vollmayr-Lee, B. More than you probably ever wanted to know about the "Ppythagorean"' Method, 2002, http://www.eg.bucknell.edu/~bvollmay/baseball/pythagoras.html.