By the Numbers

Volume 14, Number 2 The Newsletter of the SABR Statistical Analysis Committee May, 2004

Summary

Academic Research: HBPs and Retaliation

Charlie Pavitt

The author reviews a recent academic paper on whether hit batsmen come in retaliation against a team's pitcher, against the team's batters, both, or neither.

This is one of a series of reviews of sabermetric articles published in academic journals. It is part of a project of mine to collect and catalog sabermetric research, and I would appreciate learning of and receiving copies of any studies of which I am unaware. Please visit the Statistical Baseball Research Bibliography at its new location www.udel.edu/communication/pavitt/biblioexplan.htm . Use it for your research, and let me know what is missing.

Gregory A. Trandel, Hit by Pitches: Moral Hazard, Cost-Benefit, Retaliation, or Lack of Evidence? Journal of Sports Economics, February 2004, Volume 5 Number 1, pp. 87-92

Upon reading this research note, I learned that a debate has been occurring since 1997 in the economics literature concerning whether hit by pitches are primarily directed against the offending pitcher or one of the pitcher's teammates. Advocates of the former view cite evidence that the DH rule has led to fewer HBPs in the American League, while supporters of the latter argue that the former was only true for the first few years of the DH and that pitchers are themselves hit less than position players or DHs. Both sides, however, presume that retaliation is at the core of many HBPs, but the present study leads to doubt about even that presumption. Trandel found absolutely no correlation during the 1960 to 2002 period between the number of HBPs thrown by a given team's pitchers and the number of HBPs suffered by that team's batters. Along with evidence from an older study cited by Trandel showing that National League pitchers who hit a lot of batters have not been particularly likely to be hit themselves, the implication is either that retaliation is not as big a factor in HBPs as presumed, or that a different method (perhaps including some measure of brush-back pitches) is required to find it.

In this issue

Call For Help

I have been unable as yet to find copies of the first two volumes of the Journal of Sports Economics. If anyone in SABRland has access to this journal and can get me copies of sabermetric articles (not including those solely concerned with MLB finances), I would be most appreciative and happy to reimburse expenses.

Charlie Pavitt, 812 Carter Road, Rockville, MD, 20852, chazzq@udel.edu ♦

Receive BTN by Internet Subscription

You can help save SABR some money, and me some time, by downloading your copy of *By the Numbers* from the web. BTN is posted to http://www.philbirnbaum.com in .PDF format, which will print to look exactly like the hard copy issue.

To read the .PDF document, you will need a copy of Adobe Acrobat Reader, which can be downloaded from www.adobe.com.

To get on the electronic subscription list, visit http://members.sabr.org, go to "My SABR," and join the Statistical Analysis Committee. You will then be notified via e-mail when the new issue is available for download.

If you don't have internet access, don't worry – you will always be entitled to receive BTN by mail, as usual.

Informal Peer Review

The following committee members have volunteered to be contacted by other members for informal peer review of articles.

Please contact any of our volunteers on an as-needed basis - that is, if you want someone to look over your manuscript in advance, these people are willing. Of course, I'll be doing a bit of that too, but, as much as I'd like to, I don't have time to contact every contributor with detailed comments on their work. (I will get back to you on more serious issues, like if I don't understand part of your method or results.)

If you'd like to be added to the list, send your name, e-mail address, and areas of expertise (don't worry if you don't have any - I certainly don't), and you'll see your name in print next issue.

Expertise in "Statistics" below means "real" statistics, as opposed to baseball statistics - confidence intervals, testing, sampling, and so on.

Member	E-mail	Expertise
Jim Box	im.box@duke.edu	Statistics
Keith Carlson	kcarlson2@mindspring.com	General
Rob Fabrizzio	rfabrizzio@bigfoot.com	Statistics
Larry Grasso	l.grasso@juno.com	Statistics
Tom Hanrahan	HanrahanTJ@navair.navy.mil	Statistics
John Heer	jheer@walterhav.com	Proofreading
Dan Heisman	danheisman@comcast.net	General
Keith Karcher	karcherk@earthlink.net	Statistics
Chris Leach	chrisleach@yahoo.com	General
John Matthew IV	john.matthew@rogers.com	Apostrophes
Nicholas Miceli	nsmiceli@yahoo.com	Statistics
Duke Rankin	RankinD@montevallo.edu	Statistics
John Stryker	johns@mcfeely.interaccess.com	General
Tom Thress	TomThress@aol.com	Statistics (regression)
Joel Tscherne	Joel@tscherne.org	General
Dick Unruh	runruhjr@dtgnet.com	Proofreading
Steve Wang	scwang@fas.harvard.edu	Statistics

Brief Reviews

Phil Birnbaum

The author reviews two recent sabermetrically-oriented articles in recent magazines – one with an interesting take on clutch hitting, and one that includes feedback from Bill James.

Sports Illustrated on Sabermetrics

As further evidence of the influence of *Moneyball*, this year's *Sports Illustrated* Baseball Preview issue devotes an unheard-of 14 pages to the "new stat-crazy culture that turns the numbers inside-out."

The feature is devoted to stats, rather than sabermetrics per se; the field's most famous results – Voros, Pythagoras, MLEs, etc. – are given only passing mention. Instead, a couple of pages are devoted to which stats are "in" (e.g. OPS, pitcher strikeouts) and which are "out" (batting average, RBIs), and Keith Woolner's VORP gets a special mention. But there's not much on the logic behind the stats. For instance, OPS is praised because it "tells the story of how well a hitter gets on base and how much damage he does with his hits." Er, yes, that's the definition, but *why* is that stat so valuable? SI doesn't tell us.

But there's a bit of new sabermetric knowledge, courtesy of Paul Depodesta. SI quotes Tim McCarver as suggesting that AL pitchers throw fewer fastballs than NL pitchers, especially when behind in the count. During his Oakland years, Depodesta checked. It turns out there is almost no difference: NL pitchers throw 65% fastballs as compared to the AL's 64%; on hitters' counts, it's NL 75%, AL 73%.

Perhaps the most interesting of the section's several featurettes is the one entitled "Does Clutch Hitting Truly Exist?" SI quotes a few sabermetricians who explain the answer is no – and also a bunch of baseball people who say the answer is yes. Emphatically.

"That [clutch hitting doesn't exist] could be the dumbest thing I've ever heard in my life," says Reggie Jackson. "You can take those stats guys and throw them out the window," adds Derek Jeter.

Jeter is positioned in the article as an example of a clutch hitter – "one of [Reggie's] disciples in the church of October." For the record, Jeter hits for his highest OPS (.853) with the bases empty. By contrast, he's 838 with runners on; .832 with runners in scoring position; and .850 with runners in scoring position with two outs. But he has hit for a .415 average in late-inning pressure situations in the post-season – the only Jeter clutch stat the article specifically mentions.

The article quotes a few major league players and executives on both sides of the issue – suprisingly, Theo Epstein is equivocal, admitting to "having a foot in both camps." But to me, the most intriguing angles come from Jason Giambi and Derek Jeter himself.

Giambi argues that "Derek Jeter is the reason I believe [clutch hitting exists] … I know in big situations he takes better at-bats. He shrinks his strike zone. That's what makes him so good."

And Jeter adds: "Over a long season, early in games, your mind has a tendency to wander. But not with the game on the line."

Which raises the obvious questions: If Derek Jeter can have better at-bats by shrinking his strike zone, why doesn't he do it all the time? And if clutch hitting is a result of your mind wandering in the early innings, is it really something to brag about?

"Welcome to the New Age of Information," <u>*Sports Illustrated*</u>*, April 5, 2004, pp. 52-65*

The Actuarial Game of Baseball

Another recent article on baseball's newfound fascination with statistics appears in "Contingencies," a magazine published by the American Association of Actuaries. John Dewan gets mention as Sabermetrics' most famous actuary, and actuaries are touted as potential major-league sabermetricians ("who understand how to unleash the predictive power of numbers … who can see the concrete in the abstract").

The piece covers the usual ground, with one happy exception – some original material from Bill James. In a feature box of a few hundred words, James explains that it's difficult to explain attendance because there are so many contradicting forces acting on it. There's a "new team" effect, where expansion teams draw well for awhile, and an "old team" effect where multiple generations of fans are available to fill the seats. A big park can hold more fans, but, paradoxically, a small park may sell more tickets, as fans fear sellouts and therefore buy their tickets early. And so forth.

The online version of the magazine has an extra treat. The magazine's subscribers – that is, actuaries – were invited to submit ideas for new statistics, which would be submitted to Bill James for critique. Alas, there's not much new of interest in the submissions.

For instance, take the "Ultimate Baseball Statistic," from Spencer M. Gluck – "with one amazing statistic, the value of each player's contribution can be measured." Mr. Gluck's new stat credits batters with the difference in win probability between the initial and final states. This has been done several times before, including by Eldon and Harlan Mills in 1970.[1]

Also, two contributors coincidentally come up with a similar stat, the percentage of total bases advanced (so a walk with a runner on first counts as 2-for-7, since two bases were advanced by the runners (combined) and seven were possible (via home run). Another contributor declares that a team's performance correlates better to its performance three years ago than to its performance last year. This would be revolutionary if true, but James checks the numbers and finds it false.

The most reasonable statistic of the bunch comes from Damian Birnstihl, who suggests charging relievers with half a run each for inherited runners who score, and half a run each for his own runners who score after he's left the game. (That is, charge one-half run each instead of zero runs for the former and one run for the latter.)

And, finally, the most interesting piece is from Aryeh Bak, who analyzes a problem of matching starters. Is it better to match up your number one starter with the other pitcher's ace? Or is it better to throw your number five starter against him, on the assumption that you're going to lose anyway, and save your better pitchers for the games where you have a better chance of winning?

Mr. Bak does a bit of calculation, and comes up with the result that under ideal conditions -- if the opposing rotation is fixed, you know exactly how good the individual pitchers are, everyone pitches complete games, pitchers can start under any number of days' rest, and all rotations are similar to the Red Sox in 2003 (one ace and four average pitchers) – a team can save two wins a year by starting the worst pitcher against the ace. From this, he decides that the strategy should be used, a conclusion with which Bill James quite rightly disagrees, on the grounds that real life whittles those ideal conditions down to almost zero.

Mr. Bak's article contains an interesting derivation of the log5 formula. Suppose you have a .600 coin and a .400 coin (so named because they land heads 60% and 40% of the time, respectively). Suppose you flip both coins, and the "winner" of the contest is the coin that lands heads while the other one lands tails. (If both land heads or both land tails, reflip both until there's one of each.) What is the chance the 60% coin will win?

The exact formula to answer the coin flipping question can be mathematically shown to be

$$p = \frac{.6(1-.4)}{.6(1-.4)+.4(1-.6)}$$

which is exactly the log5 formula for the chance of a .600 team beating a .400 team.

That the coin-flipping formula has been shown to work for baseball games[2] suggests that baseball can be modelled similar to the coin-flipping contest. Bill James comments seem to imply that's how he originally figured it, but this is the first I've heard of it.

Steve Sullivan, "Stat of the Art: The Actuarial Game of Baseball," *Contingencies, May/June 2004, pp. 34-40 and online;*
"Major League Stats from the Actuarial Bullpen," online only. Website: *http://www.contingencies.org/mayjun04/index.html*

Phil Birnbaum, birnbaum@sympatico.ca ♦

[1] See also "*New Stat from Yale*" (BTN, August, 2003, p. 3).
[2] See Dan Levitt, "*The Batter/Pitcher Matchup*" (BTN, February, 1999), and Tom Hanrahan, "*Does Good Hitting Beat Good Pitching?*" (BTN, August, 2001).

Resolving the Probability and Interpretations of Joe DiMaggio's Hitting Streak

Dan Levitt

What is the probability that a player of Joe Dimaggio's caliber will embark on a 56-game hitting streak? The answer depends on the interpretation of the question, and the assumptions made when figuring the solution. Here, the author examines two of ways the question has been interpreted in the past, and discusses the methods that can be employed to calculate each answer.

Last summer (2003), SABR-L enjoyed a lively discussion on the probability of Joe DiMaggio's 56-game hitting streak. Several methods were suggested for calculating this probability. This is a surprisingly complicated mathematical question so it should not be unexpected that no consensus emerged as to which, if any, method was correct. In this paper I summarize and integrate several approaches that clarify and resolve the question.

Pieces in the two most recent Baseball Research Journals (numbers 31 and 32) framed the discussion by using as a point of reference an article in BRJ 23 by Charles Blahous ("The DiMaggio Streak: How Big a Deal was It?") that provided an incorrect solution. In BRJ 31 Michael Freiman, in a piece titled "56-Game Hitting Streaks Revisited," thoughtfully pointed out the inaccuracy of Blahous' formula. He then offered his own solution for the probability of a streak of length s, in a season of n games and a probability p of at least one hit in any particular game (This interpretation of the question I am labeling the "Probability Method"):

$$pr(n) = p(n-1) + \left(1 - p(n-s-1)\right) \cdot (1-p) \cdot p^{S}$$

(The rationale behind this formula is reasonably intricate, and a number of the formulas which will be presented below are even more so. I intend to provide correct formulas that the reader can use to evaluate streak probabilities on his own; please see the original sources (which I will reference) for further understanding of the mathematics behind the formulas.)

Freiman's formula can be categorized as recursive, meaning that it is calculated by calling itself. For example, the probability of a 56-game hitting streak in 139 games (DiMaggio's games in 1941) equals the probability of such a streak in 138 games plus some additional computations; the probability of a streak in 138 games then equals the probability of such a streak in 137 games plus some additional computations; et cetera until the number of games is less the length of the streak. As it turns out, Freiman's formula is correct, but in general recursive formulas have two disadvantages. First, they can be tricky to understand, although in the present case it turns out that the non-recursive expression of the probability formula is at least as complex and difficult to comprehend. Second, recursive formulas use a lot of computer memory. For instance Freiman's formula locks up my computer if trying to find the probability of a 10 game streak.

In BRJ 32 authors Bob Brown and Peter Goodrich offered another approach. After deconstructing at length Blahous' solution, the two authors determined the probability by essentially simulating the physical situation. With probability problems that are difficult to solve analytically, probabilistic simulation offers another approach to arriving at an approximate solution[1]. The two authors first estimated DiMaggio's probability of getting at least one hit in a game at 0.817[2], which they carried forward as 817. They then generated a random number between 0 and 999 to determine if DiMaggio had a hit in a particular game (0–816 represented a game with a hit, 817–999 a no-hit game). By looking at a run of 139 games they could check if there had been a hit in 56 consecutively. The authors ran this simulation 1,000,000 times and found DiMaggio hit in 56 or more consecutive games in 222 of them, giving a probability of DiMaggio's feat of .000222 (or approximately one in 5000). Freiman's formula calculates this probability to be 197 per 1,000,000; thus the simulation by Brown and Goodrich corroborates Freiman's formula.

[1] See Paul J. Nahin, *Dueling Idiots and Other Probability Puzzlers,* for a further discussion of tackling difficult probability problems with probabilistic simulation.

[2] Obviously the probability of hit changes from game to game. The most evident variable is the pitcher, but other factors such as the number of at bats a player receives is also crucial. Nonetheless, to simplify the problem down to manageable proportions, a constant hit probability from game to game is assumed. The assumption seems reasonable enough for a general approximation of the probability of a streak. As an aside, a changing probability from game to game reduces the likelihood of a streak when compared to using the overall average.

Bob Harris, an active participant in the SABR-L discussion, posted the streak question to the newsgroup alt.math.recreational.[3] During the SABR-L debate, Harris posted a recursive solution equivalent to Freiman's, while also noting that the problem could also be interpreted differently. In his newsgroup post he presented the problem under this alternative interpretation. Rather than looking at the probability of an event, Harris proposed the question as one of combinations. For example, if a team wins 100 out of 162 games what is the probability it wins at least 20 in a row? Or in a DiMaggio example, if one reasonably assumes DiMaggio had at least one hit in 115 out of his 139 games, what is the probability he hit in at least 56 consecutively at some point? Robert Israel,[4] a professor in the math department at the University of British Columbia, accepted the challenge. Using some fairly high-test mathematics he derived a formula to calculate the probability of such a streak. (This interpretation of the question I am labeling the "Combination Method.")

$$pr(n) = \frac{C(n,r) - \sum_{i=0}^{\min(n-r+1,\, floor(r/s))} -1^{i} \cdot C(n-r+1,i) \cdot C(n-si, r-si)}{C(n,r)}$$

(Where n is the total number of games; r is the number of successes, such as games with hits or wins by a team; and s is the length of the streak. C(n,r) is the number of combinations of r objects chosen from a set of n objects. Mathematically, C(n,r) equals n!/((n-r)! * r!).)

Interestingly, Israel's formula calculates a probability of .000022 (about ten times less than Freiman's), a substantial difference when viewed as 1 per 50,000 seasons versus the 1 per 5,000 resulting from Freiman's formula. That the probabilities generated under the two possible definitions of the question are materially different at first caused me some uncertainty over Israel's proposed solution. The formula is not recursive which is advantageous, but if not correct it is obviously of little value. Israel's web based credentials are impressive, so it seemed unlikely that his calculation was flawed.

Out of curiosity I created my own simulation of the problem under both interpretations to help verify the possible solutions.[5] I simulated the Probability Method 1,000,000 times which generated 204 streaks, confirming Freiman, Harris, Brown and Goodrich. I simulated the Combination Method 2,000,000 times (because of the extremely low probability) and found 34 streaks (17 per 1,000,000) corroborating Israel's formula. As a further check of Israel's method, I compared his formula result to a simulation assuming a 40 game streak because of the greater probability. The simulation again substantiated the formula: the simulation generated 2,867 such streaks; the formula computes a probability of 2,861 per 1,000,000.

Upon further reflection, I concluded that this difference between the two approaches to the problem seems reasonable. The Combination Method requires a fixed number of "successes," e.g. games with at least one hit, and then evaluates the probability of a streak. The Probability Method does not require a fixed number of successes. If one assumes a probability of .817 of getting at least one hit in a game, then in 139 games the average number of games in which the player will get a hit is 113.6. But, and this is the key, in a number of seasons, a player will hit in many more games than the average. For example, in 5,363 of the simulations at least one hit was registered in 125 or more games and at least 127 hits in 1,078. These seasons with considerably more games with hits dramatically increases the probability of a 56 game streak. By way of Israel's formula, the probability of a 56 game hitting streak in 139 games, assuming at least one hit in 127 games, is nearly 2%.

Further research on the web uncovered a direct (non-recursive) formula for calculating the probability of a streak using the Probability Method. In the newsgroup rec.puzzles a streak challenge derived from billiards was posted.[6] The problem formulation mirrored that for the Probability Method of DiMaggio's streak. With secondary support from another participant, a poster named Jim Ferry[7] from the Center for Simulation of Advanced Rockets at the University of Illinois offered the following formula:

[3] http://groups.google.com/groups?hl=en&lr=&ie=UTF-8&selm=bet25v%24k0f%241%40nntp.itservices.ubc.ca

[4] http://www.math.ubc.ca/~israel

[5] Please contact me if you wish the VBA code for either simulation.

[6] http://groups.google.com/groups?hl=en&lr=&ie=UTF-8&threadm=mumfordF7sCG8.7IM%40netcom.com&rnum=2&prev=/groups%3Fq%3Dprobability%2Bgroup:rec.puzzles.*%26hl%3Den%26lr%3D%26ie%3DUTF-8%26group%3Drec.puzzles.*%26selm%3DmumfordF7sCG8.7IM%2540netcom.com%26rnum%3D2

[7] http://www.uiuc.edu/ph/www/jferry/

$$pr(n) = \sum_{i=1}^{(n+1)/(s+1)} -1^{i-1} \cdot p^{is} \cdot q^{i-1} \cdot \left[\left(p \cdot C(n-is, i-1) + q \cdot C(n-is+1, i)\right)\right]$$

(Where n = the total number of games; s = the length of the streak; p = the probability of a "success," e.g. at least one hit; and q = 1-p. C(n,r) = n!/((n-r)! * r!) as described above.)

This formula provides the same results as the recursive formula when tested over a number of possible streak scenarios, thus substantiating its validity.

In sum, at least two ways exist to define DiMaggio's hitting streak. Of the two, the Probability Method is the more typical characterization of the problem. Nonetheless, to some the Combination Method may be more compelling, particularly if evaluating the probability of a winning streak by a team as opposed to a hitting streak by an individual player.

Interested readers may contact the author for Microsoft Excel (VBA) computer code that implements the formulas shown in the text.
Dan Levitt, danrl@attglobal.net ♦

Submissions

Phil Birnbaum, Editor

Submissions to *By the Numbers* are, of course, encouraged. Articles should be concise (though not necessarily short), and pertain to statistical analysis of baseball. Letters to the Editor, original research, opinions, summaries of existing research, criticism, and reviews of other work are all welcome.

Articles should be submitted in electronic form, either by e-mail or on PC-readable floppy disk. I can read most word processor formats. If you send charts, please send them in word processor form rather than in spreadsheet. Unless you specify otherwise, I may send your work to others for comment (i.e., informal peer review).

If your submission discusses a previous BTN article, the author of that article may be asked to reply briefly in the same issue in which your letter or article appears.

I usually edit for spelling and grammar. (But if you want to make my life a bit easier: please, use two spaces after the period in a sentence. Everything else is pretty easy to fix.) If you can (and I understand it isn't always possible), try to format your article roughly the same way BTN does.

I will acknowledge all articles within three days of receipt, and will try, within a reasonable time, to let you know if your submission is accepted.

Send submissions to:
Phil Birnbaum
18 Deerfield Dr. #608, Nepean, Ontario, Canada, K2G 4L1
birnbaum@sympatico.ca

Get Your Own Copy

If you're not a member of the Statistical Analysis Committee, you're probably reading a friend's copy of this issue of BTN, or perhaps you paid for a copy through the SABR office.

If that's the case, you might want to consider joining the Committee, which will get you an automatic subscription to BTN. There are no extra charges (besides the regular SABR membership fee) or obligations – just an interest in the statistical analysis of baseball.

The easist way to join the committee is to visit http://members.sabr.org, click on "my SABR," then "committees and regionals," then "add new" committee. Add the Statistical Analysis Committee, and you're done. You will be informed when new issues are available for downloading from the internet.

If you would like more information, send an e-mail (preferably with your snail mail address for our records) to Neal Traven, at beisbol@alumni.pitt.edu. If you don't have internet access, we will send you BTN by mail; write to Neal at
4317 Dayton Ave. N. #201, Seattle, WA, 98103-7154.

Using Calculus to Relate Runs to Wins: Part III

Ralph Caola

How many extra runs does it take to create an additional win? In the first two parts of this study, the author showed how the answer can be derived from the Pythagorean Projection formula. Here, in the final part of the study, he shows how to derive an alternative to Pythgoras, this time an approximate one.

Bill James' Pythagorean Theorem predicts winning percentage (W%) as a function of runs scored (Rs) and runs allowed (Ra). The formula is:

$$W\% = \frac{Rs^2}{Rs^2 + Ra^2}$$

or, in terms of wins (W),

$$W = G\frac{Rs^2}{Rs^2 + Ra^2}$$ where G is the number of games played.

The Pythagorean Theorem can be written in terms of differential runs (Rd) and total runs (Rt) using

`Rd = Rs – Ra` and
`Rt = Rs + Ra`

Substituting these expressions into the Pythagorean Theorem yields

$$W\% = \frac{(Rt + Rd)^2}{(Rt + Rd)^2 + (Rt - Rd)^2}$$

Rt and Rd can be expressed in runs or runs per game.

I found that the graph of winning percentage vs. Rd is nearly a straight line. It was not immediately obvious to me why this should be, so I investigated further. I expanded the expressions in parentheses and found

$$W\% = \frac{\left(Rt^2 + 2RtRd + Rd^2\right)}{2\left(Rt^2 + Rd^2\right)}$$

Then, I divided the numerator and denominator by Rt^2, and got

$$W\% = \frac{1 + 2\frac{Rd}{Rt} + \left(\frac{Rd}{Rt}\right)^2}{2 + 2\left(\frac{Rd}{Rt}\right)^2}$$ (Equation 1)

Equation 1 gives the Pythagorean Theorem expressed in terms of the ratio of differential runs to total runs (Rd/Rt).

If we assume $Rt^2 >> Rd^2$, which it is for most reasonable cases, we can justify ignoring the term $(Rd/Rt)^2$, and get

$$W\% = \frac{1 + 2\frac{Rd}{Rt}}{2} = \frac{1}{2} + \frac{Rd}{Rt} \text{ or}$$

$$W\% = .500 + \frac{Rd}{Rt} \qquad \text{(Equation 2)}$$

Equation 2 is the approximate formula for the Pythagorean Theorem. The approximation plots as a straight line with a y-intercept of 0.500 and a slope of Rd/Rt.

I thought it was interesting that this simple expression is a direct consequence of the Pythagorean Theorem. I suppose equation 2 is slightly easier to use than the Pythagorean Theorem since no squaring is involved. It is definitely easier if Rd and Rt have already been calculated.

Equation 2 has the same form as Palmer's 10 run rule, which states

$$W\% = .500 + \frac{Rd}{10}$$

Therefore, the "ten-run rule" is just equation 2 with Rt = 10.

Next, I rearranged equation 2 and got

$$Rd = Rt(W\% - .500)$$

To get incremental differential runs per win, I took the derivative with respect to winning percentage and got

$$\frac{dRd}{dW\%} = Rt$$

This is the same approximation I derived in Part I of this study (BTN, November, 2003).

Comparison of Approximate Formula and Pythagorean Theorem

I ignored Rd^2 because, for many cases, it is small compared to Rt^2. This suggests the approximate formula will be more accurate at high values of Rt and low values of Rd and less accurate at low values of Rt and high values of Rd.

This is exactly what happens. Consider the extreme case in which a team's runs scored is Rs and runs allowed are zero (Ra = 0). For this case,

```
Rt      = Rs + Ra
        = Rs + 0
        = Rs
```

and

```
Rd      = Rs – Ra
        = Rs – 0
        = Rs
```

and equation 2 becomes

```
W%      = 0.500 + Rd/Rt
        = 0.500 + Rs/Rs
        = 1.500
```

The Pythagorean Theorem yields the correct winning percentage of 1.000 for this case. Obviously, the approximate formula produces a ridiculous result.

A similar thing occurs when Rs = 0.

So, the approximate formula breaks down as the absolute value of Rd/Rt increases. We would have expected as much, because our rationale for ignoring Rd was that it was small compared to Rt.

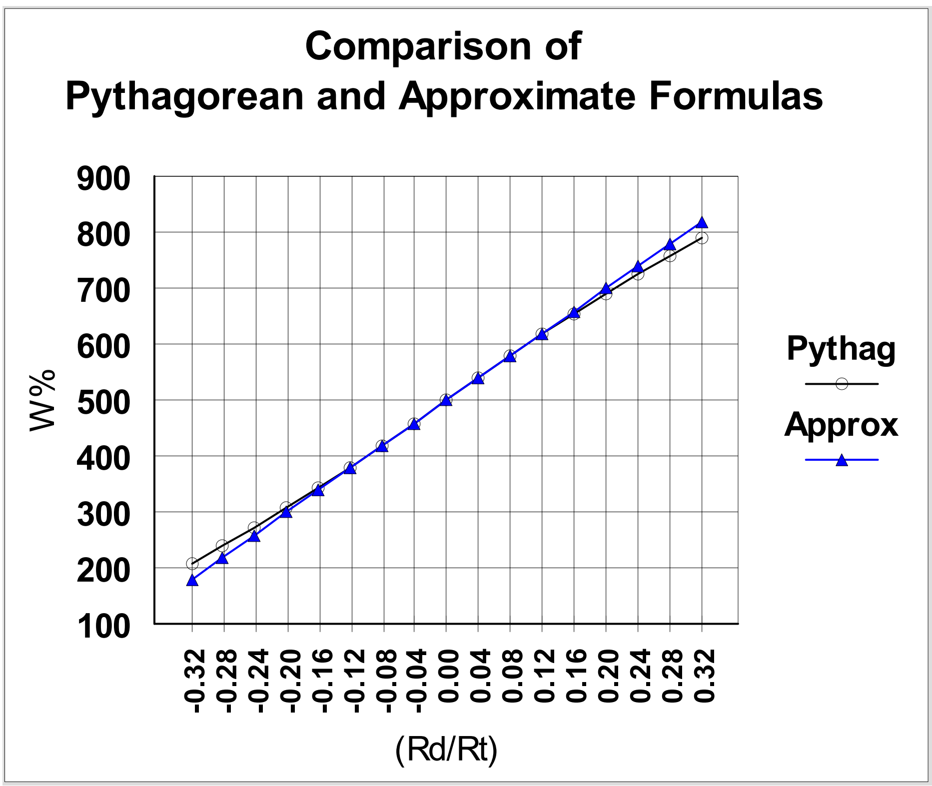
Even so, the approximate formula produces results very close to the Pythagorean Theorem for practical values of winning percentage. For winning percentages between 0.300 and 0.700 (-0.2 < Rd/Rt < 0.2), the maximum absolute difference between the two is 0.008 and the percent difference is less than 3% (see Table 1). More than 98% of all AL and NL teams, since 1901, have had winning percentages between 0.300 and 0.700.

Table 1 and the accompanying graph show how the approximate formula compares with the Pythagorean Theorem.

On this scale, for winning percentages between 0.300 and 0.700, there is almost no discernable difference between the Pythagorean Theorem and the approximate formula. For this range of winning percentages, the Pythagorean Theorem is effectively linear when winning percentage is plotted versus (Rd/Rt). I included winning percentages beyond this range to illustrate how the equations eventually diverge.

Table 1 – Comparison of Approximate and Pythagorean Formulas

Rd/Rt	W% Approx	W% Pyth	Absolute Diff	% Diff
-.30	.200	.225	-.025	-11.0
-.20	.300	.308	-.008	-2.5
-.16	.340	.344	-.004	-1.2
-.12	.380	.382	-.002	-0.4
-.08	.420	.421	.001	-0.1
-.04	.460	.460	.000	0.0
.00	.500	.500	.000	0.0
.04	.540	.540	.000	0.0
.08	.580	.579	.001	0.1
.12	.620	.618	.002	0.3
.16	.660	.656	.004	0.6
.20	.700	.692	.008	1.1
.30	.800	.775	.025	3.2



Approximate Formula From General Form of Pythagorean Theorem

I started with the Pythagorean Theorem expressed in terms of the general exponent (n)

$$W\% = \frac{Rs^n}{Rs^n + Ra^n}$$

Then I converted to

$$W\% = \frac{(Rt + Rd)^n}{(Rt + Rd)^n + (Rt - Rd)^n}$$

Next, I divided the numerator and denominator by $(1/Rt)^n$. The result is

$$W\% = \frac{\left(1 + \frac{Rd}{Rt}\right)^n}{\left(1 + \frac{Rd}{Rt}\right)^n + \left(1 - \frac{Rd}{Rt}\right)^n}$$

Next, I used the binomial expansion, which says

$$\left(1 + \left(\frac{Rd}{Rt}\right)\right)^n = 1 + n\left(\frac{Rd}{Rt}\right) + \frac{n(n-1)}{2!}\left(\frac{Rd}{Rt}\right)^2 + \frac{n(n-1)(n-2)}{3!}\left(\frac{Rd}{Rt}\right)^3 \ldots$$

and

$$\left(1 - \left(\frac{Rd}{Rt}\right)\right)^n = 1 - n\left(\frac{Rd}{Rt}\right) + \frac{n(n-1)}{2!}\left(\frac{Rd}{Rt}\right)^2 - \frac{n(n-1)(n-2)}{3!}\left(\frac{Rd}{Rt}\right)^3 \ldots$$

Again, since Rt >> Rd, I ignored terms containing $(Rd/Rt)^2$ and higher powers, so

$$\left(1 + \left(\frac{Rd}{Rt}\right)\right)^n = 1 + n\left(\frac{Rd}{Rt}\right)$$

and

$$\left(1 - \left(\frac{Rd}{Rt}\right)\right)^n = 1 - n\left(\frac{Rd}{Rt}\right)$$

and so

$$W\% = \frac{1 + n\left(\frac{Rd}{Rt}\right)}{2} = .500 + \frac{n}{2}\left(\frac{Rd}{Rt}\right) \quad \text{(Equation 3)}$$

When the common value of n = 2 is used, the result is the approximate formula I derived earlier:

$$W\% = .500 + \frac{Rd}{Rt} \qquad \text{(Equation 4)}$$

Finally, I ran a regression in which (Rd/Rt) was the independent variable and winning percentage the dependent variable. The regression included all AL and NL teams from 1901- 2003. The following equation results:

$$W\% = .500 + .912\left(\frac{Rd}{Rt}\right) \qquad \text{(Equation 5)}$$

The standard error is 0.0260 and r^2 is 0.904.

Bill James found that the Pythagorean Theorem produces better results with an exponent of 1.83, rather than 2. If we return to equation 3, and plug in n = 1.83, equation 3 becomes

$$W\% = .500 + .915\left(\frac{Rd}{Rt}\right)$$

which is strikingly close to the regression equation.

From this analysis, it appears the equation which best fits actual data is equation 5, above. The approximate formula in equation 4 doesn't fit the data best, but is only marginally worse and is a simpler expression.

Finally, the average total runs (Rt) for all major league teams from 1901-2003 is 8.8 runs per game. Notice what happens when that value is inserted in the regression equation:

$$W\% = .500 + .912\left(\frac{Rd}{8.8}\right) = .500 + \left(\frac{Rd}{9.7}\right)$$

Therefore, Palmer's ten run rule,

$$W\% = .500 + \frac{Rd}{10}$$

is a very good approximation to use, especially if Rt is not available.

References and Related Work

An article similar to this one is:

More Than You Ever Wanted to Know About the Pythagorean Method by Ben Vollmayr-Lee, available at: http://www.eg.bucknell.edu/~bvollmay/baseball/pythagoras.html .

This fine article also presents a linear approximation of the Pythagorean Theorem along with a graph similar to the one above. The author derives the approximation using a Taylor Series expansion of the Pythagorean Theorem. He goes on to examine several implications of the Taylor Series expansion. He compares the accuracy of the Pythagorean form to a linear equation, a cubic equation and equations that incorporate total runs (Rt) as a variable. He derives an equation the same as my Equation 3 above. In conclusion, Mr. Vollmayr-Lee writes that he prefers this approximation because it is both reasonably accurate and very simple.

Other references are “Linearization of Win Formulas” by Mike Mehl, available at http://home.comcast.net/~mikemehl/baseball/linwin/ , and “Revisiting the Pythagorean Theorem” by Clay Davenport with Keith Woolner, available at http://baseballprospectus.com/news/19990630davenport.html .

Ralph Caola, RJCSB25@aol.com ♦

Are Traded Players "Lemons"?

Phil Birnbaum

A famous theorem in economics predicts that where sellers know more about the item they are selling than buyers, the quality of the goods available on the market will be low. Does this apply to baseball players "sold" – traded – from one team to another? That is, does a team trade a player because they know something the other team does not?

In 2001, economist George Akerlof won a Nobel Prize for Economics for his work on the theory of the "market for lemons." The idea is this: suppose in the population of otherwise-identical used Chevrolets, some will be defective "lemons," while some will be very reliable "cherries." The owner of the car knows, from his repair bills, whether his car is a lemon or a cherry. But the prospective buyer does not.

The theory of the "market for lemons" shows that in this situation, where buyer and seller have asymmetric information, there will be a much larger proportion of lemons for sale. The reason: anyone with a lemon will have incentive to sell it, since the buyer won't know it's a lemon and may pay more than what it's worth. Similarly, the owner of a cherry has no way of proving to the buyer that it is indeed worth more than average, and is therefore more likely to hang on to it.

The result is that when sellers have more information than buyers, the overall quality of goods for sale is lowered, since there is no possibility of getting full value for high-quality goods.

Which leads to the question: if this is true for Chevrolets, is it also true for baseball players? A team that trades a player can be assumed, perhaps, to have more information about the player than the team trading for him. While both have access to the player's season and career statistics, the trading team might know his arm has been hurting lately, or that his attitude isn't what it should be, or that he's not keeping himself in shape the way he used to. According to the lemon theory, we should find that, on average, traded players won't turn out to be as good as their statistics suggest.

This study is an attempt to see if that's really the case.

The Study

How can we tell, in retrospect, if a traded player wasn't as good as we thought he was? I tried to measure this by using Bill James' "Favorite Toy." The Favorite Toy (TFT) is a method for estimating the probability of a player achieving a certain future goal (for instance, 3000 hits). If the lemon theory applies to ballplayers, then traded players as a group should be less likely to reach a TFT goal than players who were not traded.

So what I did was this: for all players from 1901 to 1975, I found qualifying players who changed teams during or after one of those years; 1975 was chosen because prior to 1975, it can safely be assumed that players would change teams only as a result of being traded or sold.

For each player, I calculated TFT's "projected remaining runs created" for his career as of the end of that season. I then checked how many players achieved their projections plus 50%. The TFT formula predicts that the chance of a player achieving his projection plus 50% is 1 in 6, or 17%.

I added the 50% because Bill James stressed that TFT is only valid for exceptional players attempting to achieve a difficult goal. I assume that a goal with a 17% chance of success counts as difficult.

And also, because the players must be "exceptional," I limited the study to players who had accumulated 1,000 runs created in their careers so far.

I also eliminated players who did not play at all after their qualifying season, for whatever reason (injury, retirement, etc.)

Example – Hank Aaron

Take Hank Aaron in 1969. After the 1969 season, TFT projected Hammerin' Hank to create an additional 369 more runs by the end of his career[1]. Our test sees if he will create 150% of that, or 552 runs. In fact, Aaron outdid even the 150% goal, creating 577 more runs before retiring in 1976.

Aaron was not traded after the '69 season, but he *was* traded after the 1974 season. At that time, he was expected to create 116 more runs. He failed to achieve 150% of that (and he even failed to achieve 100% of that), creating only 87 runs over the next two seasons before retiring.

I repeated this calculation for every player from 1901-1975 who had 1,000 runs created under his belt at the time. I then separated those players into those who were traded (like Aaron 1974) and players who weren't (like Aaron 1969).

Results

Including both traded and non-traded cases, there were 678 qualifying player-seasons in the study.[2]

Of these 678 seasons, 148 had the player began the next season with another team (by changing teams during or after the previous season). Of those 148, 10 of them, or 7%, reached the goal.

In the other 530 seasons, the player was not traded. Of those 530, 109 of them, or 21%, achieved the goal.

It certainly looks like there's a large lemon effect here: players not traded were three times more likely to achieve the goal as traded players.

Table 1 – Players with at least 1000 career RC who played again after the current season

	Total number of players	Players achieving 150% of projection	Percentage
Only non-traded players	530	109	21%
Only traded players	148	10	7%
All qualifying players	678	119	18%

Test 2

Now, as mentioned, TFT is supposed to work only on "exceptional" players. Perhaps we got the results we did because the traded players weren't "exceptional." Perhaps they accumulated their 1000 runs created, but then were in the twilights of their careers, playing part-time, or pinch hitting. That might make them different from the non-traded players, which might skew the results.

To test this, I tightened the criteria for inclusion. In addition to having 1000 runs created, I limited the sample to players who

(a) had at least 90 runs created in the previous year; and
(b) had a TFT "established level" (weighted average of the past three years) of at least 80 RC.

I ran the study again. The results (see table 2) were even more dramatic: None of the traded players – zero – achieved the goal.

[1] Techincal notes: The basic version of the Runs Created formula was used throughout the study. The player's age was taken as of December 31 of the year before starting the next season with the new team. The "16% chance of reaching 150%" does not consider the Favorite Toy's "97%" rule.

[2] Raw data for each test can be found, until October, 2004, at http://www.philbirnbaum.com/lemondata.txt .

Table 2 – Players with at least 1000 career RC who played again after the current season, with at least 90 RC that season and an established level of 80 RC, achieving 150% of their Favorite Toy projection

	Total number of players	Players achieving 150% of projection	Percentage
Only non-traded players	170	42	25%
Only traded players	19	0	0%
All qualifying players	189	42	22%

Test 3

Now, 22% is already a small proportion of successful seasons. Also, those 42 successful seasons aren't by 42 different players – some players have multiple successes, often in consecutive seasons. For example, Hank Aaron accounted for 13 of the 42, from 1962 to 1973, plus 1975. (He was recorded as successful in the first 9 of those 13 seasons.) Perhaps, then, it's just by chance that the lemon effect seems so large, due to the list's domination by the most successful players.

To check that, I changed the study so that instead of requiring 150% of the expected RC to count as successful, only 100% would be required. This should bring the success rate, in theory, to 50%.

Here are the results (Table 3):

Table 3 – Players with at least 1000 career RC who played again after the current season, with at least 90 RC that season and an established level of 80 RC, achieving their Favorite Toy projection

	Total number of players	Players achieving 100% of projection	Percentage
Only non-traded players	170	86	51%
Only traded players	19	3	16%
All qualifying players	189	89	47%

Again, there's a strong effect here: players whose teams kept them were more than three times as likely to achieve expectations than traded players. Looked at another, equally shocking way – only 1 out of 6 traded players achieved the expectations you'd expect from a player of that age and performance. You'd expect 50% of those players to be below what was expected, but a full 84% had careers that fizzled out early.

Test 4

Perhaps, I thought, it's an age thing: maybe the traded players were old, and TFT just doesn't work as well for old players. But the results turned out to be similar. Here are the results for the young players, younger than 35:

Table 4 – Players with at least 1000 career RC who played again after the current season, with at least 90 RC that season and an established level of 80 RC, who were 34 or younger in the current season

	Total number of players	Players achieving 100% of projection	Percentage
Only non-traded players	105	53	50%
Only traded players	13	1	8%
All qualifying players	127	52	41%

And here are the players 35 or older:

Table 5 – Players with at least 1000 career RC who played again after the current season, with at least 90 RC that season and an established level of 80 RC, who were 35 or older in the current season

	Total number of players	Players achieving 100% of projection	Percentage
Only non-traded players	65	33	51%
Only traded players	6	2	33%
All qualifying players	71	35	49%

Conclusions

It does seem that there's a lemon effect for baseball players – traded players do seem to be damaged goods, at least compared to what you'd expect (or what TFT would expect) from their statistics.

However, TFT is a blunt tool, and using TFT limited the sample to certain types of players. A more comprehensive study would use a different projection method, one that can apply to all traded players, not just outstanding ones.

One possibility is to use Bill James' Brock2 method. Fed a career so far, Brock2 projects the player's career totals, and can be used for all players (although I think the player must have at least three years' worth of major league statistics). The disadvantage is that the algorithm is difficult to implement without a spreadsheet (although if anyone has a VB implementation they are willing to share, please drop me a line).

Alternatively, the editors of *Baseball Prospectus* talk about their "Pecota" method, which projects a player's career by finding retired players most similar at the same age, and averaging their stats. However, the BP editors do not explain their proprietary method in full.

If the difficulties of either of these methods can be resolved, it would be fairly easy to do a full test of all players, and see to what extent, if any, the lemon effect continues to appear.

Phil Birnbaum, birnbaum@sympatico.ca ♦

Bias Against the Home Team in the Pythagorean Theorem

Thomas Thress

While the Pythagorean Theorem is generally accurate, it turns out that it is incorrect for home games or road games considered separately. That's because it doesn't consider that in games it wins, the home team's runs come over fewer innings, which means its runs scored per game, and thus its projected wins, is understated. Here, the author explains, and comes up with a new version of Pythagoras that can be used with accuracy for both home and road teams.

In looking at predictions of won-lost records using Bill James's Pythagorean Theorem over the past decade or so, I discovered what I at first thought was a rather unusual result. The Pythagorean Theorem consistently under-estimates the true home-field advantage in Major League Baseball.

Table 1 shows that the Pythagorean Theorem consistently underestimated the true home-field advantage in Major League Baseball by an average of 4 games per 162.

Also, in 1994, the Pythagorean Theorem suggested that Home teams should have lost more games than they won. It found no home-field advantage in either 1999 or 2001.

In fact, the result is overwhelming. Looking at all American and National League teams from 1900 through 2003, the Pythagorean Theorem underestimated the number of home wins for 72.8% of all teams.[1] Clearly, the Pythagorean Theorem is not an unbiased estimator for wins by the home team (or road team, of course).

Table 1 – Home Won-Lost Records for Major League Baseball, 1990 – 2003

Year	Actual Wins	Predicted Wins	Error
2003	89	83	-6
2002	88	84	-4
2001	85	81	-4
2000	88	84	-4
1999	84	81	-3
1998	87	83	-4
1997	88	83	-5
1996	87	84	-3
1995	86	82	-4
1994	83	80	-3
1993	87	84	-3
1992	90	86	-4
1991	87	82	-5
1990	88	83	-5

Why Does the Pythagorean Theorem Underpredict Home Wins?

Once you think about it, this really has a pretty simple answer. When the home team wins a game, they bat in fewer innings than the visiting team. In most cases, the home team will not bat in the bottom of the ninth inning. Hence, they have one-ninth fewer opportunities to score than the visiting team. Even when the home team bats in the ninth or extra innings, they will have only a partial inning (i.e., less than three outs) in their final inning.

If the home team won 50% of the time and all games went exactly nine innings, then, in every two games, the visiting team would bat in 18 innings, while the home team would only bat in 17 innings. Hence, one would expect the visiting team to outscore the home team by nearly 6% (18/17 – 1).

Look at it another way. The Pythagorean Theorem says that, if a team scored as many runs at home as it allowed at home, that team would win 50% of the time. But, in its victories, this team would not necessarily have scored as many runs as they might have had they had as many opportunities as their opponents. The fact that this team was able to score as many runs as its opponents in fewer opportunities indicates that this team was probably better, on average, than its opponents.

From 1900 to 2003, a total of 12 teams allowed exactly as many runs as they scored at home. Nine of these 12 teams had a winning record at home and two of the three teams with losing home records lost exactly one game more than they won. Overall, the average home record of these 12 teams was 40–37.

[1] Pythagorean projections use an exponent of 2. Tie games are treated as one-half wins and one-half losses throughout this paper. All win totals are normalized to a 162-game season.

On the other hand, from 1900 to 2003, 14 teams allowed exactly as many runs as they scored on the road. Only four of these 14 teams had a winning record on the road (one team won as many as they lost) and two of those four teams won exactly one more game than they lost. Overall, the average home record of these 12 teams was 37–39 (and this record is actually greatly exaggerated because the 1917 St. Louis Cardinals amazingly went 44–32 on the road despite both scoring and allowing 265 runs).

To correct for the fact that the home team will have fewer expected opportunities to score than the visiting team (the home team can never have more opportunities and will have fewer opportunities whenever they win), therefore, the home team's runs scored should be adjusted upward by a multiplier before the Pythagorean Theorem is applied.

Adjusting Basic Pythagorean Equation for Home Team Bias

To correct for this bias, I re-specified the basic Pythagorean equation as follows:

$$Win\% = \frac{(a \cdot hRS + rRS)^e}{(a \cdot hRS + rRS)^e + (hRA + a \cdot rRA)^e}$$

where hRS = runs scored at home, rRS = runs scored on the road, hRA = runs allowed at home, rRA = runs allowed on the road.

I then estimated the variables *a* and *e* using a nonlinear least squares procedure in the statistical software package EViews over a sample period which included all home teams and all road teams from 1900 – 2003 (American and National League). This provided a total of 4,108 observations (2,054 teams, home and road).

Results are presented in Table 2 below with and without the home-team multiplier. This table shows both mean errors (which should be close to zero if the estimate is unbiased) as well as root mean-squared errors (which are minimized by the nonlinear least squares procedure) for Home Wins, Road Wins, and Total Wins. Total wins were estimated in two ways. First, they were estimated based upon total runs

Table 2 – Basic Pythagorean Equation Results

No Home Team Adjustment

	Coefficient	Standard Error
Home-Team Adjustment (a)	1.000000	
Pythagorean Exponent (e)	1.836418	0.015421

	Mean Error	% Errors > 0	Root Mean Square Error	% of Deviation Explained
Home Team	1.85	73.76%	3.37	56.76%
Road Team	-1.89	25.51%	3.37	55.63%
Overall Record (est.)	-0.03	50.00%	4.01	71.12%
Home + Road	-0.04	49.71%	4.02	71.11%

With Home Team Adjustment

	Coefficient	Standard Error
Home-Team Adjustment (a)	1.055726	0.001462
Pythagorean Exponent (e)	1.834842	0.012670

	Mean Error	% Errors > 0	Root Mean Square Error	% of Deviation Explained
Home Team	-0.02	50.10%	2.81	63.97%
Road Team	-0.02	51.07%	2.78	63.36%
Overall Record (est.)	-0.03	49.90%	4.02	71.04%
Home + Road	-0.04	49.76%	4.02	71.11%

scored and total runs allowed (adjusted as appropriately). This is shown in the row titled "Overall Record (est.)". Total Wins were then estimated a second way as the sum of Home and Road wins. The results here are shown in the row titled "Home + Road."

Two results are fairly apparent in Table 2. First, the home-team adjustment significantly improves the prediction of wins by the home team and by the road team. Both root mean-squared errors are reduced by approximately 17% in these cases, and an additional 17-18% of the variation in Home/Road team wins can be explained by including this multiplier.

Second, the home-team adjustment has very little, if any, effect on the accuracy of predicting total wins for teams that play half of their games at home and half on the road. This isn't terribly surprising, as the adjustment on runs scored at home will, on average, offset the adjustment on runs allowed on the road. Nevertheless, I might have expected a somewhat more clear advantage to the newer method.

Adjusting PythagoPat Equation for Home Team Bias

A somewhat more advanced variation of the Pythagorean Theorem is the so-called PythagoPat equation, which expresses the Pythagorean exponent as a function of the average runs per game. This equation is the following:

$$Win\% = \frac{RunsScoredPerGame^{RE}}{RunsScoredPerGame^{RE} + RunsAllowedPerGame^{RE}}$$

where

$$RE = (TotalRunsPerGame)^{e}$$

I modified this, as above, by introducing a multiplier for runs scored by the home team, or, the following:

$$Win\% = \frac{\left(\frac{a(HomeRS) + RoadRS}{Games}\right)^{RE}}{\left(\frac{a(HomeRS) + RoadRS}{Games}\right)^{RE} + \left(\frac{a(HomeRA) + RoadRA}{Games}\right)^{RE}}$$

where RS=Runs Scored, RA= Runs Allowed, and

$$RE = \left(\frac{a(HomeRS) + RoadRS + HomeRA + a(RoadRA)}{Games}\right)^{e}$$

As above, values for a and e were estimated using nonlinear least squares in EViews. Results are presented in Table 3.

Table 3 – PythagoPat Equation Results

No Home Team Adjustment

	Coefficient	Standard Error
Home-Team Adjustment (a)	1.000000	
PythagoPat Exponent (e)	0.279757	0.003883

	Mean Error	% Errors > 0	Root Mean Square Error	% of Deviation Explained
Home Team	1.88	74.20%	3.36	56.91%
Road Team	-1.91	25.22%	3.36	55.71%
Overall Record (est.)	-0.03	50.34%	3.98	71.34%
Home + Road	-0.03	49.90%	3.98	71.35%

With Home Team Adjustment

	Coefficient	Standard Error
Home-Team Adjustment (a)	1.055686	0.001437
PythagoPat Exponent (e)	0.278830	0.003147

	Mean Error	% Errors > 0	Root Mean Square Error	% of Deviation Explained
Home Team	-0.02	49.81%	2.77	64.38%
Road Team	-0.01	51.46%	2.77	63.58%
Overall Record (est.)	-0.02	50.54%	3.99	71.29%
Home + Road	-0.03	50.15%	3.98	71.38%

Comparing Table 2 and Table 3, one can see that the PythagoPat formula yields a slightly more accurate fit. The effect of introducing the home-adjustment multiplier is comparable to in Table 2 above.

The best fit here for overall team records is the sum of predicted home wins plus predicted road wins using PythagoPat with the home team adjustment multiplier. Of course, the overall advantage of adding the home adjustment multiplier here is to explain an additional 0.03% of total deviation. In other words, in terms of predicting overall team records, the advantage is trivial.

Conclusion

So, is the added complication of adjusting the runs scored by the home team worth it? If you're focusing only on full seasons in which teams played the same number of games at home and on the road, the benefit is minimal at best. On the other hand, when evaluating expected winning percentages within a season – as a means of predicting expected future wins, for example – if the number of home versus road games differs for a team, then such a consideration may well represent a significant improvement.

Another area where this might be of interest, although I have not thought extensively about this, may be in comparing home versus road splits and, for example, in measuring park effects.

In any event, even if this analysis has no practical implications, I still found it interesting.

Thomas Thress, tomthress@aol.com ♦