Pythagorean expectation

Pythagorean expectation is a sports analytics formula devised by Bill James to estimate the percentage of games a baseball team "should" have won based on the number of runs they scored and allowed. Comparing a team's actual and Pythagorean winning percentage can be used to make predictions and evaluate which teams are over-performing and under-performing. The name comes from the formula's resemblance to the Pythagorean theorem.[1]

The basic formula is:

$$\text{Win Ratio} = \frac{\text{runs scored}^2}{\text{runs scored}^2 + \text{runs allowed}^2} = \frac{1}{1 + (\text{runs allowed}/\text{runs scored})^2}$$

where Win Ratio is the winning ratio generated by the formula. The expected number of wins would be the expected winning ratio multiplied by the number of games played.

Contents

Empirical origin

Empirically, this formula correlates fairly well with how baseball teams actually perform. However, statisticians since the invention of this formula found it to have a fairly routine error, generally about three games off. For example, in 2002, the New York Yankees scored 897 runs and allowed 697 runs. According to James' original formula, the Yankees should have won 62.35% of their games.

$$\text{Win} = \frac{897^2}{897^2 + 697^2} = 0.623525865$$

Based on a 162-game season, the Yankees should have won 101.01 games. The 2002 Yankees actually went 103–58.[2]

In efforts to fix this error, statisticians have performed numerous searches to find the ideal exponent.

If using a single-number exponent, 1.83 is the most accurate, and the one used by baseball-reference.com.[3] The updated formula therefore reads as follows:

$$\text{Win} = \frac{\text{runs scored}^{1.83}}{\text{runs scored}^{1.83} + \text{runs allowed}^{1.83}} = \frac{1}{1 + (\text{runs allowed}/\text{runs scored})^{1.83}}$$

The most widely known is the Pythagenport formula[4] developed by Clay Davenport of Baseball Prospectus

$$\text{Exponent} = 1.50 \cdot \log\left(\frac{R + RA}{G}\right) + 0.45$$

He concluded that the exponent should be calculated from a given team based on the team's runs scored (R), runs allowed (RA), and games (G). By not reducing the exponent to a single number for teams in any season, Davenport was able to report a 3.9911 root-mean-square error as opposed to a 4.126 root-mean-square error for an exponent of 2.[4]

Less well known but equally (if not more) effective is the Pythagenpat formula, developed by David Smyth.[5]

$$\text{Exponent} = \left(\frac{R + RA}{G}\right)^{.287}$$

Davenport expressed his support for this formula, saying:

> After further review, I (Clay) have come to the conclusion that the so-called Smyth/Patriot method, aka Pythagenpat, is a better fit. In that, $X = ((rs + ra)/g)^{0.285}$, although there is some wiggle room for disagreement in the exponent. Anyway, that equation is simpler, more elegant, and gets the better answer over a wider range of runs scored than Pythagenport, including the mandatory value of 1 at 1 rpg.[6]

These formulas are only necessary when dealing with extreme situations in which the average number of runs scored per game is either very high or very low For most situations, simply squaring each variable yields accurate results.

There are some systematic statistical deviations between actual winning percentage and expected winning percentage, which include bullpen quality and luck. In addition, the formula tends to regress toward the mean, as teams that win a lot of games tend to be underrepresented by the formula (meaning they "should" have won fewer games), and teams that lose a lot of games tend to be overrepresented (they "should" have won more).

"Second-order" and "third-order" wins

In their Adjusted Standings Report,[7] Baseball Prospectus refers to different "orders" of wins for a team. The basic order of wins is simply the number of games they have won. However, because a team's record may not reflect its true talent due to luck, different measures of a team's talent were developed.

First-order wins, based on pure run differential, are the number of expected wins generated by the "pythagenport" formula (see above). In addition, to further filter out the distortions of luck, Sabermetricians can also calculate a team's *expected* runs scored and allowed via a runs created-type equation (the most accurate at the team level being Base Runs). These formulas result in the team's expected number of runs given their offensive and defensive stats (total singles, doubles, walks, etc.), which helps to eliminate the luck factor of the order in which the team's hits and walks came within an inning. Using these stats, sabermetricians can calculate how many runs a team "should" have scored or allowed.

By plugging these expected runs scored and allowed into the pythagorean formula, one can generate second-order wins, the number of wins a team deserves based on the number of runs they should have scored and allowed given their component offensive and defensive statistics. Third-order wins are second-order wins that have been adjusted for strength of schedule (the quality of the opponent's pitching and hitting). Second- and third-order winning percentage has been shown to predict future actual team winning percentage better than both actual winning percentage and first-order winning percentage.

Theoretical explanation

Initially the correlation between the formula and actual winning percentage was simply an experimental observation. In 2003, Hein Hundal provided an inexact derivation of the formula and showed that the Pythagorean exponent was approximately $2/(\sigma\sqrt{\pi})$ where σ was the standard deviation of runs scored by all teams divided by the average number of runs scored.[8] In 2006, Professor Steven J. Miller provided a statistical derivation of the formula[9] under some assumptions about baseball games: if runs for each team follow a Weibull distribution and the runs scored and allowed per game are statistically independent, then the formula gives the probability of winning.[9]

More simply, the Pythagorean formula with exponent 2 follows immediately from two assumptions: that baseball teams win in proportion to their "quality", and that their "quality" is measured by the ratio of their runs scored to their runs allowed. For example, if Team A has scored 50 runs and allowed 40, its quality measure would be 50/40 or 1.25. The quality measure for its (collective) opponent team B, in the games played against A, would be 40/50 (since runs scored by A are runs allowed by B, and vice versa), or 0.8. If each team wins in proportion to its quality, A's probability of winning would be 1.25 / (1.25 + 0.8), which equals 50^2 / (50^2 + 40^2), the Pythagorean formula.The same relationship is true for any number of runs scored and allowed, as can be seen by writing the "quality" probability as [50/40] / [50/40 + 40/50], and clearing fractions.

The assumption that one measure of the quality of a team is given by the ratio of its runs scored to allowed is both natural and plausible; this is the formula by which individual victories (games) are determined. [There are other natural and plausible candidates for team quality measures, which, assuming a "quality" model, lead to corresponding winning percentage expectation formulas that are roughly as accurate as the Pythagorean ones.] The assumption that baseball teams win in proportion to their quality is not natural, but is plausible. It is not natural because the degree to which sports contestants win in proportion to their quality is dependent on the role that chance plays in the sport. If chance plays a very large role, then even a team with much higher quality than its opponents will win only a little more often than it loses. If chance plays very little role, then a team with only slightly higher quality than its opponents will win much more often than it loses. The latter is more the case in basketball, for various reasons, including that many more points are scored than in baseball (giving the team with higher quality more opportunities to demonstrate that quality, with correspondingly fewer opportunities for chance or luck to allow the lower quality team to win.)

Baseball has just the right amount of chance in it to enable teams to win roughly in proportion to their quality, i.e. to produce a roughly Pythagorean result with exponent two. Basketball's higher exponent of around 14 (see below) is due to the smaller role that chance plays in basketball. And the fact that the most accurate (constant) Pythagorean exponent for baseball is around 1.83, slightly less than 2, can be explained by the fact that there is (apparently) slightly more chance in baseball than would allow teams to win in precise proportion to their quality. Bill James realized this long ago when noting that an improvement in accuracy on his original Pythagorean formula with exponent two could be realized by simply adding some constant number to the numerator, and twice the constant to the denominator. This moves the result slightly closer to .500, which is what a slightly larger role for chance would do, and what using the exponent of 1.83 (or any positive exponent less than two) does as well. Various candidates for that constant can be tried to see what gives a "best fit" to real life data.

The fact that the most accurate exponent for baseball Pythagorean formulas is a variable that is dependent on the total runs per game is also explainable by the role of chance, since the more total runs scored, the less likely it is that the result will be due to chance, rather than to the higher quality of the winning team having been manifested during the scoring opportunities. The larger the exponent, the farther away from a .500 winning percentage is the result of the corresponding Pythagorean formula, which is the same effect that a decreased role of chance creates. The fact that accurate formulas for variable exponents yield larger exponents as the total runs per game increases is thus in agreement with an understanding of the role that chance plays in sports.

In his 1981 Baseball Abstract, James explicitly developed another of his formulas, called the log5 formula (which has since proven to be empirically accurate), using the notion of 2 teams having a face-to-face winning percentage against each other in proportion to a "quality" measure.His quality measure was half the team's "wins ratio" (or "odds of winning"). The wins ratio or odds of winning is the ratio of the team's wins against the league to its losses against the league. [James did not seem aware at the time that his quality measure was expressible in terms of the wins ratio. Since in the quality model any constant factor in a quality measure eventually cancels, the quality measure is today better taken as simply the wins ratio itself, rather than half of it.] He then stated that the Pythagorean formula, which he had earlier developed empirically, for predicting winning percentage from runs, was "the same thing" as the log5 formula, though without a convincing demonstration or proof. His purported demonstration that they were the same

boiled down to showing that the two different formulas simplified to the same expression in a special case, which is itself treated vaguely, and there is no recognition that the special case is not the general one. Nor did he subsequently promulgate to the public any explicit, quality-based model for the Pythagorean formula. As of 2013, there is still little public awareness in the sabermetric community that a simple "teams win in proportion to quality" model, using the runs ratio as the quality measure, leads directly to James's original Pythagorean formula.

In the 1981 Abstract, James also says that he had first tried to create a "log5" formula by simply using the winning percentages of the teams in place of the runs in the Pythagorean formula, but that it did not give valid results. The reason, unknown to James at the time, is that his attempted formulation implies that the relative quality of teams is given by the ratio of their winning percentages. Yet this cannot be true if teams win in proportion to their quality, since a .900 team wins against its opponents, whose overall winning percentage is roughly .500, in a 9 to 1 ratio, rather than the 9 to 5 ratio of their .900 to .500 winning percentages. The empirical failure of his attempt led to his eventual, more circuitous (and ingenious) and successful approach to log5, which still used quality considerations, though without a full appreciation of the ultimate simplicity of the model and of its more general applicability and true structural similarity to his Pythagorean formula.

Use in basketball

American sports executive Daryl Morey was the first to adapt James' Pythagorean expectation to professional basketball while a researcher at STATS, Inc.. He found that using 13.91 for the exponents provided an acceptable model for predicting won-lost percentages:

$$\text{Win} = \frac{\text{points for}^{13.91}}{\text{points for}^{13.91} + \text{points against}^{13.91}}.$$

Daryl's "Modified Pythagorean Theorem" was first published in *STATS Basketball Scoreboard, 1993-94*.[10]

Noted basketball analyst Dean Oliver also applied James' Pythagorean theory to professional basketball. The result was similar

Another noted basketball statistician, John Hollinger, uses a similar Pythagorean formula, except with 16.5 as the exponent.

Use in pro football

The formula has also been used in pro football by football stat website and publisher Football Outsiders, where it is known as **Pythagorean projection**. The formula is used with an exponent of 2.37 and gives a projected winning percentage. That winning percentage is then multiplied by 16 (for the number of games played in an NFL season), to give a projected number of wins. This projected number given by the equation is referred to as Pythagorean wins.

$$\text{Pythagorean wins} = \frac{\text{Points For}^{2.37}}{\text{Points For}^{2.37} + \text{Points Against}^{2.37}} \times 16.$$

The 2011 edition of *Football Outsiders Almanac*[11] states, "From 1988 through 2004, 11 of 16 Super Bowls were won by the team that led the NFL in Pythagorean wins, while only seven were won by the team with the most actual victories. Super Bowl champions that led the league in Pythagorean wins but not actual wins include the 2004 Patriots, 2000 Ravens, 1999 Rams and 1997 Broncos."

Although *Football Outsiders Almanac* acknowledges that the formula had been less-successful in picking Super Bowl participants from 2005–2008, it reasserted itself in 2009 and 2010. Furthermore, "[t]he Pythagorean projection is also still a valuable predictor of year-to-year improvement. Teams that win a minimum of one full game more than their Pythagorean projection tend to regress the following year; teams that win a minimum of one full game less than their Pythagoerean projection tend to improve the following year, particularly if they were at or above .500 despite their underachieving. For example, the 2008 New Orleans Saints went 8-8 despite 9.5 Pythagorean wins, hinting at the improvement that came with the next year's championship season"

Use in ice hockey

In 2013, statistician Kevin Dayaratna and mathematician Steven J. Miller provided theoretical justification for applying the Pythagorean Expectation to ice hockey. In particular, they found that by making the same assumptions that Miller made in his 2007 study about baseball, specifically that goals scored and goals allowed follow statistically independent Weibull distributions, that the Pythagorean Expectation works just as well for ice hockey as it does for baseball. The Dayaratna and Miller study verified the statistical legitimacy of making these assumptions and estimated the Pythagorean exponent for ice hockey to be slightly above 2.[1][12]

See also

- Baseball statistics
- Sabermetrics
- Football Outsiders

Notes

1. "The Game Designer: Pythagoras Explained" (http://thegamedesignerblogspot.com/2012/05/pythagoras-explained.html). Retrieved 7 May 2016.
2. "2002 New York Yankees" (https://www.baseball-reference.com/teams/NYY/2002.shtml). *Baseball-Reference.com* Retrieved 7 May 2016.
3. "Frequently Asked Questions" (https://www.sports-reference.com/blog/baseball-reference-faqs/). *Baseball-Reference.com*. Retrieved 7 May 2016.
4. "Baseball Prospectus- Revisiting the Pythagorean Theorem" (http://www.baseballprospectus.com/article.php?articleid=342). *Baseball Prospectus*. Retrieved 7 May 2016.
5. "W% Estimators" (http://gosu02.tripod.com/id69.html) Retrieved 7 May 2016.
6. "Baseball Prospectus- Glossary" (http://baseballprospectus.com/glossary/index.php?mode=viewstat&stat=136) Retrieved 7 May 2016.
7. "Baseball Prospectus- Adjusted Standings" (http://www.baseballprospectus.com/statistics/standings.php). Retrieved 7 May 2016.
8. Hundal, Hein. "Derivation of James Pythagorean Formula (Long)" (http://groups.google.com/group/rec.puzzles/browse_thread/thread/3be0e6ad49631ddb/bfb52d16b12955ac?q=hein+hundal+pythagorean&fwc=1)
9. Miller (2007). "A Derivation of the Pythagorean Won-Loss Formula in Baseball". *Chance*. **20**: 40–48. arXiv:math/0509698 (https://arxiv.org/abs/math/0509698). Bibcode:2005math.....9698M (http://adsabs.harvard.edu/abs/2005math.....9698M) doi:10.1080/09332480.2007.10722831 (https://doi.org/10.1080/09332480.2007.10722831).
10. Dewan, John; Zminda, Don; STATS, Inc. Staff (October 1993). *STATS Basketball Scoreboard, 1993-94*. STATS, Inc. p. 17. ISBN 0-06-273035-5.
11. *Football Outsiders Almanac 2011* (ISBN 978-1-4662-4613-3), p.xviii
12. Dayaratna, Kevin; Miller, Steven J. (2013). "The Pythagorean Won-Loss Formula and Hockey: A Statistical Justification for Using the Classic Baseball Formula as an Evaluative Tool in Hockey" (http://web.williams.edu/Mathematics/sjmiller/public_html/math/papers/DayaratnaMiller_HockeyFinal.pdf) (PDF). *The Hockey Research Journal 2012/13*. **XVI**: 193–209.

External links

- Miller (2007) [2005]. "A Derivation of the Pythagorean Won-Loss Formula in Baseball". *Chance Magazine*. **20** (1): 40–48. arXiv:math.ST/0509698. Bibcode:2005math.....9698M. doi:10.1080/09332480.2007.10722831
- Current Major League Baseball Pythagorean expectation.
- Adjusting football's Pythagorean Theorem