The Pythagorean Formula Extended for Soccer

Howard Hamilton
Proprietor, Soccermetrics Blog
www.soccermetricsblog.com

October 31, 2009

Abstract

I've extended the Pythagorean formula to account for the probability of a drawn result, which makes the formula applicable to association football (soccer) and other sports that allow for a draw. The inclusion of the second term results in a much more complicated Pythagorean.

1 Introduction

The Pythagorean formula was first developed by Bill James to predict the number of wins of a baseball team from the number of runs scored and allowed during the season:

$$\frac{RS_{obs}^{\gamma}}{RS_{obs}^{\gamma} + RA_{obs}^{\gamma}} \tag{1}$$

It is typically calculated near the midpoint of the season for each team to assess whether it is performing above or below expectations. Steven Miller[1] derived from first principles the Pythagorean formula and showed with some assumptions in the statistical distribution that his result matched James' formula.

The Pythagorean has been applied to baseball, basketball and other sports leagues with varying degrees of success. However, its applications to soccer have not been as successful and have generally resulted in an underprediction of points won over a season. One reason for this is that the Pythagorean formula does not allow for the possibility of a tied result, which happens in a nontrivial percentage of soccer matches during a given season. In this short report I will derive an extension to the Pythagorean that includes the probability of a draw.

2 Statistical Assumptions

There are some initial assumptions about the statistics before going into the derivation. First, the goals scored and allowed are statistically independent random variables. This assumption is a fair one to make in soccer because of the possibility of a draw. The second assumption is that the goals scored and allowed during a season follow a three-parameter Weibull distribution:

$$f(x;\alpha,\beta,\gamma) = \frac{\gamma}{\beta}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-\left(\frac{x-\beta}{\alpha}\right)^{\gamma}},\ x \geq \beta \tag{2}$$

$$= 0,\ x < \beta \tag{3}$$

This is the same distribution that was used in the Miller paper; it is a flexible statistical distribution that

[1] "A Derivation of the Pythagorean Win-Loss Formula in Baseball", preprint at arXiv:math/0509698v4.

can take on a wide variety of shapes. The distribution parameter is defined as β, the scale parameter is α, and the shape parameter is the exponent γ, which I will call the Pythagorean exponent. The third assumption I've made isn't really as assumption; it's more of a mathematical construct. I place the data into N bins, defined

$$[-.5, .5] \cup [.5, 1.5] \cup \cdots \cup [6.5, 7.5] \cup [7.5, 8.5] \cup \cdots \cup [N - .5, N + .5] \tag{4}$$

I make this construction so that the means of the bins correspond to its center, which is where all of the data in the bins would be located (there are only integer goals in soccer), resulting in a continuous model. Therefore the translation parameter $\beta = -0.5$. This mathematical construction also allows me to calculate the probability of a draw by calculating the probability that the distribution will lie between the endpoints of a bin.

Finally I need expressions for α_{GS} and α_{GA}, which are scale parameters of the two Weibull distributions. I obtain these expressions when calculating the means of the two distributions (this is done in the Miller paper so I'm not going to repeat it here). The means are equal to the average goals scored and goals allowed – G_S and G_A, respectively – and the alpha terms are defined as the following:

$$\begin{aligned} \alpha_{GS} &= \frac{G_S - \beta}{\Gamma(1 + \gamma^{-1})} = \frac{\widehat{G_S}}{\Gamma(1 + \gamma^{-1})} \\ \alpha_{GA} &= \frac{G_A - \beta}{\Gamma(1 + \gamma^{-1})} = \frac{\widehat{G_A}}{\Gamma(1 + \gamma^{-1})} \end{aligned}$$

The $\widehat{G}_S$ and $\widehat{G}_A$ terms are the adjusted average goals scored/allowed (adjusted by the translation parameter), and theΓterm is what is called a "Gamma function". You can find more information on it on Wikipedia, but a short definition is that it allows us to calculate the factorial of a real number. Formally, it's defined as this:

$$\Gamma(s) = \int_0^\infty e^{-u} u^{s-1} du$$

It appears in the derivation of the mean of the Weibull distribution, but drops out when we calculate the probability of a win. It does *not* drop out when we calculate the probability of a draw, and that makes the calculation of the soccer Pythagorean much more difficult.

3 Derivation of Draw Probability

Now I'm ready to show the derivation for the draw probability. We want to solve for $P(X = Y = c)$, where c is the number of goals scored.

$$\begin{aligned} P(X = Y = c) &= \int_{c+\beta}^{c-\beta} \int_{c+\beta}^{c-\beta} \frac{\gamma}{\alpha_{GS}} \left(\frac{x - \beta}{\alpha_{GS}} \right)^{\gamma - 1} e^{-((x-\beta)/\alpha_{GS})^\gamma} \frac{\gamma}{\alpha_{GA}} \left(\frac{y - \beta}{\alpha_{GA}} \right)^{\gamma - 1} e^{-((y-\beta)/\alpha_{GA})^\gamma} dy\, dx \\ &= \int_{c+\beta}^{c-\beta} \frac{\gamma}{\alpha_{GS}} \left(\frac{x - \beta}{\alpha_{GS}} \right)^{\gamma - 1} e^{-((x-\beta)/\alpha_{GS})^\gamma} \left[e^{-((y-\beta)/\alpha_{GA})^\gamma} \right|_{c+\beta}^{c-\beta} dx \\ &= \int_{c+\beta}^{c-\beta} \frac{\gamma}{\alpha_{GS}} \left(\frac{x - \beta}{\alpha_{GS}} \right)^{\gamma - 1} e^{-((x-\beta)/\alpha_{GS})^\gamma} \left[e^{-((c-2\beta)/\alpha_{GA})^\gamma} - e^{-(c/\alpha_{GA})^\gamma} \right] dx \\ &= \left[e^{-((x-\beta)/\alpha_{GS})^\gamma} \right|_{c+\beta}^{c-\beta} \left[e^{-((c-2\beta)/\alpha_{GA})^\gamma} - e^{-(c/\alpha_{GA})^\gamma} \right] \\ &= \left[e^{-((c-2\beta)/\alpha_{GS})^\gamma} - e^{-(c/\alpha_{GS})^\gamma} \right] \left[e^{-((c-2\beta)/\alpha_{GA})^\gamma} - e^{-(c/\alpha_{GA})^\gamma} \right] \end{aligned}$$

Now if we substitute for $\beta = -.5$, we get the resulting expression for the probability of a drawn match at c goals:

$$P(X = Y = c) = \left[e^{-((c+1)/\alpha_{GS})^{\gamma}} - e^{-(c/\alpha_{GS})^{\gamma}}\right]\left[e^{-((c+1)/\alpha_{GA})^{\gamma}} - e^{-(c/\alpha_{GA})^{\gamma}}\right]$$

Then we sum over N to get the total probability for a draw between teams X and Y:

$$P(X = Y) = \sum_{c=0}^{N}\left[e^{-((c+1)/\alpha_{GS})^{\gamma}} - e^{-(c/\alpha_{GS})^{\gamma}}\right]\left[e^{-((c+1)/\alpha_{GA})^{\gamma}} - e^{-(c/\alpha_{GA})^{\gamma}}\right]$$

4 The Extended Pythagorean

We can now write the extended Pythagorean formula, which will give the expected points won per game:

$$3\cdot\frac{\alpha_{GS}^{\gamma}}{\alpha_{GS}^{\gamma}+\alpha_{GA}^{\gamma}}+\sum_{c=0}^{N}\left[\exp\left\{-\left(\frac{c+1}{\alpha_{GS}}\right)^{\gamma}\right\}-\exp\left\{-\left(\frac{c}{\alpha_{GS}}\right)^{\gamma}\right\}\right]\left[\exp\left\{-\left(\frac{c+1}{\alpha_{GA}}\right)^{\gamma}\right\}-\exp\left\{-\left(\frac{c}{\alpha_{GA}}\right)^{\gamma}\right\}\right] \tag{5}$$

When a practitioner is using this formula, he will have the goals scored and goals allowed statistics (which will be converted into average goals scored/allowed per game) and the league Pythagorean exponent. We would like to rewrite the above equation so that we see the terms we will actually use. Such a task is simpler for the first term because the Gamma functions cancel themselves out. It is much more difficult in the second term because the Gamma functions do not cancel.

To save space, let's define $\kappa = \Gamma\left(1 + {}^{1}\!/\!_{\gamma}\right)$. Then substitute for α_{GS} and α_{GA}in both terms to obtain the extended Pythagorean in terms of the goal averages and the Pythagorean exponent:

$$3\cdot\frac{\hat{G}_S^{\gamma}}{\hat{G}_S^{\gamma}+\hat{G}_A^{\gamma}}+\sum_{c=0}^{N}\left[\exp\left\{-\left(\frac{\kappa\,(c+1)}{\hat{G}_S}\right)^{\gamma}\right\}-\exp\left\{-\left(\frac{\kappa c}{\hat{G}_S}\right)^{\gamma}\right\}\right]\left[\exp\left\{-\left(\frac{\kappa\,(c+1)}{\hat{G}_A}\right)^{\gamma}\right\}-\exp\left\{-\left(\frac{\kappa c}{\hat{G}_A}\right)^{\gamma}\right\}\right] \tag{6}$$

The computation of the Pythagorean requires two phases. The first is to calculate the mean Pythagorean exponent for the league in a given season, then use that exponent along with goal averages to compute the Pythagorean for the following season. This is the procedure that I would take:

1. Collect match result data over a given season for all teams in the league. Arrange the data into goals scored and goals allowed columns for each team.
2. Develop a distribution of goals scored and allowed. This is basically a histogram. Divide by the total number of goals scored/allowed to obtain a probability measurement.
3. Fit the goal distribution data, both scored and allowed, to the two-parameter Weibull distribution ($\beta = -0.5$). It is necessary to minimize the sum of squares of the error between the data and the probability distribution for goals scored and allowed *simultaneously*. The nonlinear least-squares algorithm works very well, and converges quickly. The result will be $\{\alpha_{GS}, \alpha_{GA}, \gamma\}$ for each team.
4. Calculate the mean γ for the league. This is the league Pythagorean exponent. Also pre-calculateκat this stage.
5. Use the extended Pythagorean to calculate the estimated points to be won per game, and multiply by the total number of matches to get the estimated points for the season.

5 Summary

We have now extended the Pythagorean formula to include the probabilities of a drawn match, which makes it applicable to soccer and other sports that permit draws. The second term is much more

complicated than the first because of the presence of the exponential and Gamma function terms, as well as the summation term in order to consider all score possibilities in a draw. However the scores need only go up to some reasonable figure, such as 5 or 6 goals, and in practice it is usually sufficient just to consider draws up to 4-4. The Gamma function is provided in most spreadsheet packages (for example, in Excel it is GAMMALN), but alternatively it can be pre-calculated and included with the Pythagorean exponent for the league.

In the baseball sabermetrics studies, the Pythagorean has generally been accurate to within four or five games. In soccer this would mean that the extended Pythagorean should be accurate within 12-15 points. This turns out to be a very large spread when it comes to estimating total points won in a season, but it appears to do well in predicting relative league position and assessing which teams are performing at, above, or below expectations.