Theoretical Derivation of the Extended James' Pythagorean Formula for Head-to-head Matchups in Baseball

Sabermetrics Modeling Group (SMALL 2024 REU)

Saad Waheed
sw21@williams.edu
Joint Work: Janine Wang, Raul Marquez, Prof. Steven J. Miller

AISC 2024 Conference
October 12th, 2024

The Original Pythagorean Win Percentage Formula

James' Pythagorean Formula

$$\text{Win Percentage} = \frac{\text{RS}^\gamma}{\text{RS}^\gamma + \text{RA}^\gamma}$$

- **RS** = Runs Scored.
- **RA** = Runs Allowed.
- γ: Originally taken to be 2, but empirically ≈ 1.82 for baseball.

The Original Pythagorean Win Percentage Formula

James' Pythagorean Formula

$$\text{Win Percentage} = \frac{\text{RS}^\gamma}{\text{RS}^\gamma + \text{RA}^\gamma}$$

- **RS** = Runs Scored.
- **RA** = Runs Allowed.
- γ: Originally taken to be 2, but empirically ≈ 1.82 for baseball.

How the Formula is Used

- **Extrapolation:** Predict a team's performance for the rest of the season based on mid-season data.
- **Evaluation:**
  - Assess if team is consistently over/under-performing.
  - Assess the value of adding players to team roster.

Applications of the Original Pythagorean Formula

Why is it Popular?

- Uses only two basic statistics [Runs Scored (RS) and Runs Allowed (RA)] but is fairly accurate despite simplicity.
- Examples:

Year	Team	RS	RA	Predicted Win % (Midseason)	Actual Win %
2009	Boston Red Sox	465	380	59.3	58.6
2011	Philadelphia Phillies	384	295	62.1	63.0
2014	San Francisco Giants	375	350	53.2	54.3
2019	New York Yankees	503	390	61.6	63.5
2023	Arizona Diamondbacks	449	422	52.9	51.9

Applications of the Original Pythagorean Formula

Why is it Popular?

- Uses only two basic statistics [Runs Scored (RS) and Runs Allowed (RA)] but is fairly accurate despite simplicity.
- Examples:

Year	Team	RS	RA	Predicted Win % (Midseason)	Actual Win %
2009	Boston Red Sox	465	380	59.3	58.6
2011	Philadelphia Phillies	384	295	62.1	63.0
2014	San Francisco Giants	375	350	53.2	54.3
2019	New York Yankees	503	390	61.6	63.5
2023	Arizona Diamondbacks	449	422	52.9	51.9

Extension to Other Sports

- Hockey ($\gamma \approx 2.1$, Dayaratna and Miller 2013)
- Basketball ($\gamma \approx 13.91$, Daryl Morey 1994)

Theoretical Basis for Pythagorean Formula (Miller 2006)

- RS and RA for a team modeled as independent random variables drawn from a continuous Weibull distributions.
- Weibull distributions chosen for their flexibility in modeling.

Theoretical Basis for Pythagorean Formula (Miller 2006)

- RS and RA for a team modeled as independent random variables drawn from a continuous Weibull distributions.
- Weibull distributions chosen for their flexibility in modeling.

Weibull Distribution

$$f_{\mathrm{RS}}(x;\alpha,\beta,\gamma) = \begin{cases} \frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1}\exp\left[-\left(\frac{x-\beta}{\alpha}\right)^{\gamma}\right] & \text{if } x \geq \beta \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

- α: Scale parameter (related to variance).
- β: Location parameter (translation).
- γ: Shape parameter (determines the distribution's form, whether it is more exponential or normal-like).

Shapes of the Weibull Distribution

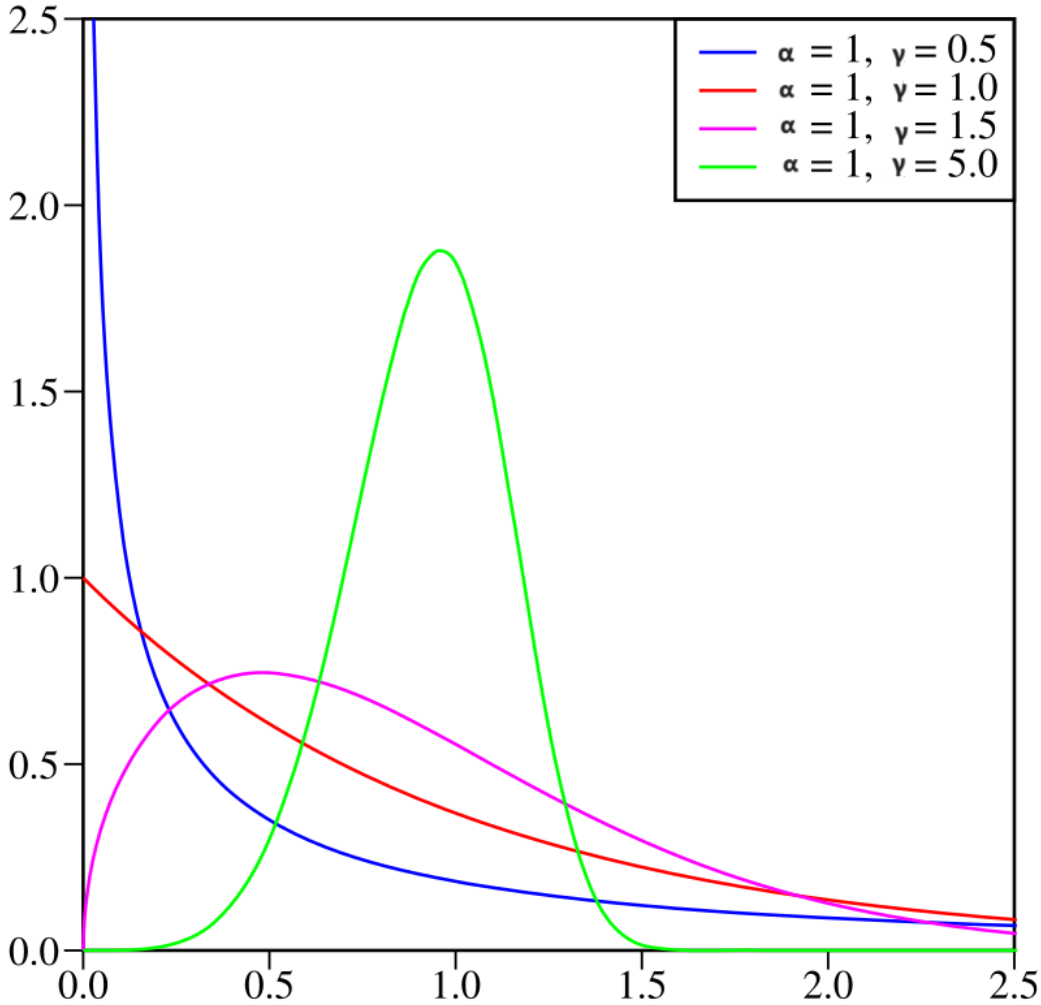

Figure: Possible Shapes of the Weibull Probability Density Function (PDF).

Weibull Distribution Expectation and Parameter Setup

Expectation of Weibull Distribution

$$\mu_{\alpha,\beta,\gamma} = \alpha\Gamma(1+\gamma^{-1}) + \beta$$

Setting Parameters for Runs Scored (RS) and Runs Allowed (RA)

- **RS**: Modeled by a Weibull$(\alpha_{RS}, \beta, \gamma)$.
- **RA**: Modeled by a Weibull$(\alpha_{RA}, \beta, \gamma)$.
- $\boldsymbol{\beta}$: Baseline for RS and RA.
- We set α_{RS} and α_{RA} so that:

$$\alpha_{RS} = \frac{RS-\beta}{\Gamma(1+\gamma^{-1})}, \quad \alpha_{RA} = \frac{RA-\beta}{\Gamma(1+\gamma^{-1})}.$$

Deriving the Pythagorean Win Percentage Formula

Key Steps in the Integration

$$
\begin{aligned}
\mathbb{P}(\text{Win}) &= \mathbb{P}(RS > RA) \\
&= \int_{x=0}^{\infty} f_{RS}(x)\,(1 - F_{RA}(x))\,dx \qquad \text{[from Weibull PDF in (1)]} \\
&= \int_{x=0}^{\infty} \frac{\gamma}{\alpha_{RS}} \left(\frac{x}{\alpha_{RS}}\right)^{\gamma-1} e^{-(x/\alpha_{RS})^{\gamma}} \left(1 - e^{-(x/\alpha_{RA})^{\gamma}}\right) dx \\
&= 1 - \frac{\alpha^{\gamma}}{\alpha_{RS}^{\gamma}} \\
&= \boxed{\frac{\alpha_{RS}^{\gamma}}{\alpha_{RS}^{\gamma} + \alpha_{RA}^{\gamma}}} \qquad \text{(Pythagorean Formula)}
\end{aligned}
$$

Drawbacks of the Original Pythagorean Formula

- **Poor Playoff Predictions**:
  - Less accurate in short playoff series due to small sample sizes.
- **Ignores Matchups**:
  - Assumes teams face an "average" opponent, ignoring head-to-head strengths.
- **No Home/Away Adjustment**:
  - Fails to account for home-field advantage or away-game disadvantages.

Extended Pythagorean Formula (Cleary and Miller 2020)

Adjusted Home Team Terms

$$\mathrm{RS}_{h,adj} = \mathrm{RS}_h \left(\frac{\mathrm{RA}_a}{\mathrm{R}} \right), \quad \mathrm{RA}_{h,adj} = \mathrm{RA}_h \left(\frac{\mathrm{RS}_a}{\mathrm{R}} \right)$$

$\mathbf{RS_h}$ $\mathbf{(RA_h)}$: Runs scored (allowed) by home team.
$\mathbf{RS_a}$ $\mathbf{(RA_a)}$: Runs scored (allowed) by away team.
$\mathbf{R}$: League average runs scored/conceded per game.

$\Longrightarrow$ Adjusted by a scaling factor of away team performance relative to league averages.

Extended Pythagorean Formula (Cleary and Miller 2020)

Adjusted Home Team Terms

$$\mathrm{RS}_{h,adj} = \mathrm{RS}_h \left(\frac{\mathrm{RA}_a}{\mathrm{R}} \right), \quad \mathrm{RA}_{h,adj} = \mathrm{RA}_h \left(\frac{\mathrm{RS}_a}{\mathrm{R}} \right)$$

$\mathrm{RS_h}$ ($\mathrm{RA_h}$): Runs scored (allowed) by home team.
$\mathrm{RS_a}$ ($\mathrm{RA_a}$): Runs scored (allowed) by away team.
R: League average runs scored/conceded per game.

$\Longrightarrow$ Adjusted by a scaling factor of away team performance relative to league averages.

Extended Win Probability Formula

$$\begin{aligned} \mathbb{P}(\text{Home Team Wins}) &= \frac{(\mathrm{RS}_{h,adj})^\gamma}{(\mathrm{RS}_{h,adj})^\gamma + (\mathrm{RA}_{h,adj})^\gamma} \\ &= \frac{(\mathrm{RS_h RA_a})^\gamma}{(\mathrm{RS_h RA_a})^\gamma + (\mathrm{RA_h RS_a})^\gamma} \end{aligned}$$

Performance Against Playoff Data (2001-2019)

Results

- **Extended Pythagorean Predictions:**
  - Higher seed predicted to win **80.18** series, lose **68.82**.

Performance Against Playoff Data (2001-2019)

Results

- **Extended Pythagorean Predictions:**
  - Higher seed predicted to win **80.18** series, lose **68.82**.
- **Observed Results:**
  - Higher seed won **80** series, lost **69**.

Conclusion

- Provides better predictions for head-to-head matchups than the original Pythagorean formula in playoffs, while maintaining simplicity.

Summer Research Summary

Goals for the Summer

- Derive an extended Pythagorean win-loss formula using a theoretical framework similar to Miller (2006).
- Improve formula further by incorporating team-specific factors into head-to-head matchups.

Summer Research Summary

Goals for the Summer

- Derive an extended Pythagorean win-loss formula using a theoretical framework similar to Miller (2006).
- Improve formula further by incorporating team-specific factors into head-to-head matchups.

Approach

- Treat $\mathrm{RS}_a, \mathrm{RA}_a$ for the away team as the observed value of a random variable with

$$\mathbb{E}[\mathrm{RS}_a] = \frac{n(m\mathrm{R}) - \mathrm{RS}_h}{n-1} \approx m\mathrm{R}$$

  where $\boldsymbol{n}$ = # of teams in league, $\boldsymbol{m}$ = # of matches so far in season.
- Use **conditional distributions** to model Runs Scored (RS) and Runs Allowed (RA) by home team, adjusted by the away team metrics $\mathrm{RS}_a, \mathrm{RA}_a$.

Method of Conditional Distributions

Conditional Weibull Distribution for RS_h

Denote X_h as the random variables for runs scored by the home team in a fixture. Then,

$$X_h \sim \text{Weibull}(\alpha_{\text{RS}_{h,adj}}, \beta, \gamma).$$

- X_h dependent only on RA_a and not RS_a.
- As before, assume RS_h and RA_h to be independent.

$$f_{X_h|\text{RA}_a}(x) = \begin{cases} \dfrac{\gamma}{\alpha_{\text{RS}_{h,adj}}} \left(\dfrac{x-\beta}{\alpha_{\text{RS}_{h,adj}}}\right)^{\gamma-1} \exp\left(-\left(\dfrac{x-\beta}{\alpha_{\text{RS}_{h,adj}}}\right)^{\gamma}\right) & \text{if } x \geq \beta \\ 0 & \text{if } x < \beta \end{cases}$$

where

$$\alpha_{\text{RS}_{h,adj}} = \frac{\text{RS}_{h,adj} - \beta}{\Gamma(1+\gamma^{-1})} = \frac{\text{RS}_h \cdot \dfrac{\text{RA}_a}{m\text{R}} - \beta}{\Gamma(1+\gamma^{-1})}.$$

Expectation of X_h and Unconditional Distribution

Expectation of X_h

$$
\begin{aligned}
\implies \mathbb{E}[X_h \mid RA_a] &= \mathrm{RS}_h \cdot \frac{\mathrm{RA}_a}{m\mathrm{R}} \\
\therefore \mathbb{E}[X_h] &= \mathbb{E}_{\mathrm{RA}_a}[E[X_h \mid \mathrm{RA}_a]] \\
&= \mathbb{E}_{\mathrm{RA}_a}\left[\mathrm{RS}_h \cdot \frac{\mathrm{RA}_a}{m\mathrm{R}}\right] \\
&= \mathrm{RS}_h \cdot \frac{\mathbb{E}[\mathrm{RA}_a]}{m\mathrm{R}} \\
&= \mathrm{RS}_h
\end{aligned}
$$

Expectation of X_h and Unconditional Distribution

Expectation of X_h

$$
\begin{aligned}
\implies \mathbb{E}[X_h \mid RA_a] &= \mathrm{RS}_h \cdot \frac{\mathrm{RA}_a}{m\mathrm{R}} \\
\therefore \mathbb{E}[X_h] &= \mathbb{E}_{\mathrm{RA}_a}[E[X_h \mid \mathrm{RA}_a]] \\
&= \mathbb{E}_{\mathrm{RA}_a}\left[\mathrm{RS}_h \cdot \frac{\mathrm{RA}_a}{m\mathrm{R}}\right] \\
&= \mathrm{RS}_h \cdot \frac{\mathbb{E}[\mathrm{RA}_a]}{m\mathrm{R}} \\
&= \mathrm{RS}_h
\end{aligned}
$$

Unconditional Weibull Distribution for X_h

$$
f_{X_h}(x) = \begin{cases} \dfrac{\gamma}{\alpha_{\mathrm{RS}_h}} \left(\dfrac{x-\beta}{\alpha_{\mathrm{RS}_h}}\right)^{\gamma-1} \exp\left(-\left(\dfrac{x-\beta}{\alpha_{\mathrm{RS}_h}}\right)^{\gamma}\right), & \text{if } x \geq \beta \\ 0 & \text{if } x < \beta. \end{cases}
$$

Deriving the Extended Pythagorean Formula

Extended Pythagorean Formula

$$
\begin{aligned}
\mathbb{P}(\text{Win}) &= \mathbb{P}(X_h > Y_h) \\
&= \int_{\beta}^{\infty} f_{X_h|\text{RA}_a}(x)\left[1 - F_{Y_a}(x)\right] dx \\
&= \int_{\beta}^{\infty} \frac{\gamma}{\alpha_{\text{RS}_{h,adj}}} \left(\frac{x-\beta}{\alpha_{\text{RS}_{h,adj}}}\right)^{\gamma-1} e^{-\left(\frac{x-\beta}{\alpha_{\text{RS}_{h,adj}}}\right)^{\gamma}} \left[1 - e^{-\left(\frac{x-\beta}{\alpha_{\text{RA}_a}}\right)^{\gamma}}\right] dx \\
&= \frac{\alpha^{\gamma}_{\text{RS}_{h,adj}}}{\alpha^{\gamma}_{\text{RS}_{h,adj}} + \alpha^{\gamma}_{\text{RA}_a}} = \frac{\left(\text{RS}_h \cdot \frac{\text{RA}_a}{m\text{R}}\right)^{\gamma}}{\left(\text{RS}_h \cdot \frac{\text{RA}_a}{m\text{R}}\right)^{\gamma} + \left(\text{RA}_h \cdot \frac{\text{RS}_a}{m\text{R}}\right)^{\gamma}} \\
&= \boxed{\frac{\left(\text{RS}_h \cdot \text{RA}_a\right)^{\gamma}}{\left(\text{RS}_h \cdot \text{RA}_a\right)^{\gamma} + \left(\text{RA}_h \cdot \text{RS}_a\right)^{\gamma}}} \quad \text{(Extended Pythagorean Formula)}
\end{aligned}
$$

Future Work

- **Analyzing Left and Right-Handed Matchups**:
  - Study the impact of batter and pitcher handedness on game outcomes.
  - Investigate how handedness affects runs scored and allowed.
  - Integrate handedness statistics into the extended Pythagorean formula.
- **Model Validation and Testing**:
  - Apply the extended Pythagorean formula to more recent data (post-2020) and compare predictive accuracy with existing metrics.

Acknowledgements

- I would like to thank Professor Steven J. Miller for his invaluable guidance and support throughout this research.
- This work was supported by Williams College and the National Science Foundation through grant DMS-2241623.

Bibliography

[1] K. D. Dayaratna and S. J. Miller, *The pythagorean won-loss formula and hockey: a statistical justification for using the classic baseball formula as an evaluative tool in hockey*, 2013.

[2] S. J. Miller, *A Derivation of the Pythagorean Won-Loss Formula in Baseball*, 2006.

[3] R. Cleary, J. Jeffries, C. Miller, S. J. Miller, J. Murray, and N. Skiera, *Extending James' Pythagorean Formula for Head-to-head Matchups*, 2020.

[4] J. Dewan and D. Zminda, *STATS Basketball Scoreboard, 1993-94*, p. 17 (STATS, Inc., Oct. 1993).