プロ野球の「順位逆転現象」を
シミュレーションで紐解く
～2019 年のホークス 2 位は当たり前じゃない!?～

芝浦工業大学　数理科学研究会
BV17057　西脇 友哉

2019 年 11 月 1 日

第 I 部
序論

研究背景

プロ野球の**福岡ソフトバンクホークス**は毎年優勝争いに加わる強豪チームである．ここ 10 年の間に 3 位以内に入らなかった年は 1 回しかなく, 野球ファンの間でもホークスの強さは周知の事実である．同チームは 2019 年も強さを発揮し, パ・リーグ 2 位からの日本シリーズ優勝という好成績を残した．一方で, **ピタゴラス勝率**と呼ばれる指標による順位はリーグ内で 4 位であり, 例年と比較すると圧倒的な強さではなかったことが伺える．このような背景から「ピタゴラス勝率 4 位のチームがシーズン 2 位になる事は珍しいのではないか」と考えて研究を行うことにした．「ピタゴラス勝率の順位が実際の順位と異なる」という現象についてデータを集計したところ, 2019 年のホークスと同じ例は過去 9 年間で 4 度起こっていることが分かった (表 1 参照) が, サンプル数が少なく, 集計結果から何かを推察することが難しい．そこで, このような現象を**順位逆転現象**と名付け, シミュレーションによって仮想再現した大量のデータから「2019 年のホークスの凄さ」を検証するという発想に至った．

表 1　ピタゴラス勝率の順位とシーズン順位 ('11〜'19)

		シーズン順位						計
		1	2	3	4	5	6	
ピタゴラス勝率	1	14	2	1	1	0	0	18
	2	3	7	5	2	1	0	18
	3	1	5	7	5	0	0	18
	4	0	**4**	3	6	4	1	18
	5	0	0	1	2	10	5	18
	6	0	0	1	2	3	12	18
計		18	18	18	18	18	18	

ピタゴラス勝率

ピタゴラス勝率 (2) は, 得点と失点から勝率を推定する計算式である．式の形がピタゴラスの定理（三平方の定理）と似ていることからピタゴラス勝率と命名されたと言われている．

計算式 1 (勝率, ピタゴラス勝率)**.**

$$(\text{勝率}) = \frac{(\text{勝利数})}{(\text{勝利数}) + (\text{敗戦数})} \tag{1}$$

$$\begin{aligned}(\text{ピタゴラス勝率}) &= \frac{(\text{得点数})^2}{(\text{得点数})^2 + (\text{失点数})^2} \\ &= \frac{1}{1+\rho^2} \quad \left(\rho = \frac{\text{失点数}}{\text{得点数}}\right)\end{aligned} \tag{2}$$

ピタゴラス勝率 (2) は実際の試合結果に基づく勝率 (1) との相関が強く, 実際に'11〜'19 の 9 年間のデータに対する相関係数は $r = 0.916$ (95% 信頼区間 : $[0.880, 0.942]$) であった．図 1 からも直線的な関係がうかがえる．

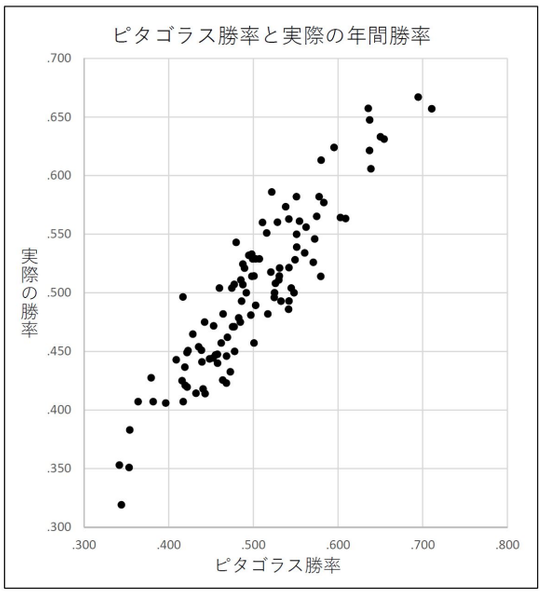

図 1　ピタゴラス勝率と実際の年間勝率の散布図 ('11〜'19)

1　新規性・独創性

ピタゴラス勝率は実際の勝率と強い正の相関を持つため, 得点と失点のみで勝率を予測出来る指標として非常に有用である．しかし, ピタゴラス勝率と実際の勝率を完全に等しいものとして扱った場合, 勝率の決定要素は得点と失点のみとなり, 「ピタゴラス勝率が最も高いチームは必ずリーグ優勝出来る」という状況になってしまう．表 1 からも分かる様に, これは現実にそぐわない．そこで, これまであまり議論されていなかった「ピタゴラス勝率と実際の勝率の差異（勝率補正)」に着目し,

$$(\text{勝率}) \sim (\text{ピタゴラス勝率}) + (\text{誤差})$$

という概念を新たに導入した. そして, 「得点と失点をベースに勝率を予測するが, 完全に依存している訳ではない」という設定のもと, 前述の**順位逆転現象**を再現した.

また, 「テーマに対して必要なデータ（年間のチーム成績）だけを生成出来れば十分であり, 公式戦全試合を無理に再現する必要は無い」と考え, 仮定や近似によって成績の決定要素を簡略化したことで, **野球ゲームにおけるペナントモードのオート進行**等とは違った方向性でシミュレーションを実装した.

これらの点において本研究は新規性, 独創性を有する.

2 勝率補正

本研究では実際の勝率をピタゴラス勝率で近似し, 得点数と失点数から勝率を予測するという目的で利用する. ピタゴラス勝率の計算式は指数部を 2 ではなく x と置くことで一般化できるため, 本研究では式 (3) の通りに表記する. また, ピタゴラス勝率の最適化を行い, 最適な指数 x^* を特定する. (5 章)

最適化を行うとピタゴラス勝率と実際の勝率の誤差を減らす事が出来るが, 0 にはならない. この誤差を**勝率補正**と名付け, 式 (4) の通りに定義する.

$$\mathrm{Py}(x) := \frac{1}{1+\rho^x} \quad \left(\omega : \text{勝率}, \rho : \frac{\text{失点}}{\text{得点}}\right) \tag{3}$$

$$\mathrm{ERR}(x) := \omega - \mathrm{Py}(x) \tag{4}$$

3 研究の大まかな流れ

1. プロ野球のチーム成績に関するデータを収集する
2. 正規乱数を使用するための準備として, データの正規性を確認する（4.5 節）
3. ピタゴラス勝率の最適化を行う（5 章）
4. シミュレーションを行うためのプログラムを MATLAB で組む（詳細 : 6 章）
   (a) 得点数を出力する
   (b) 失点数を出力する
   (c) 勝率補正を乱数で出力する
   (d) 4a〜4c の結果を利用して勝率を計算する
   (e) 勝率の順に仮想データをソートし, 順位を付ける
5. プログラムを実行し, 結果を考察する

4 前提

4.1 表記

- 野球の指標のうち, 打率, 守備率, 勝率など, 割合を表すものを表記する際には 1 の位の「0」を省略することが一般的である.（例 : 打率 2 割 8 分 6 厘→ .286） 本研究の計算式中において指標を用いる際には, 数値としての側面を意識して 1 の位の「0」を省略せずに表記することがある.
- 文章中では年度を西暦の下 2 桁で表すことがある. (例 : 2018 年→'18)
- 本研究で使用するデータのうち, 実際のプロ野球の試合結果に基づくものを**実データ**と呼ぶことにする. 一方, 実データの数値を元にシミュレーションを行い, 擬似的に生成したデータを**仮想データ**と呼ぶことにする.

4.2 野球の指標

本研究で用いる野球の指標を以下に記し, 指標間の関連性のイメージを図 2 に表した.

計算式 2 (各種打撃の指標)**.**

$$
\begin{aligned}
\text{打率} &= \text{安打} \div \text{打数} \\
\text{出塁率 (OBP)} &= \frac{(\text{安打} + \text{四球} + \text{死球})}{(\text{打数} + \text{四球} + \text{死球} + \text{犠飛})} \\
\text{長打率 (SLG)} &= \text{塁打} \div \text{打数} \\
\text{OPS} &= \text{出塁率} + \text{長打率}
\end{aligned}
$$

それぞれの指標が持つ意味合いについては参考資料を参照されたい.

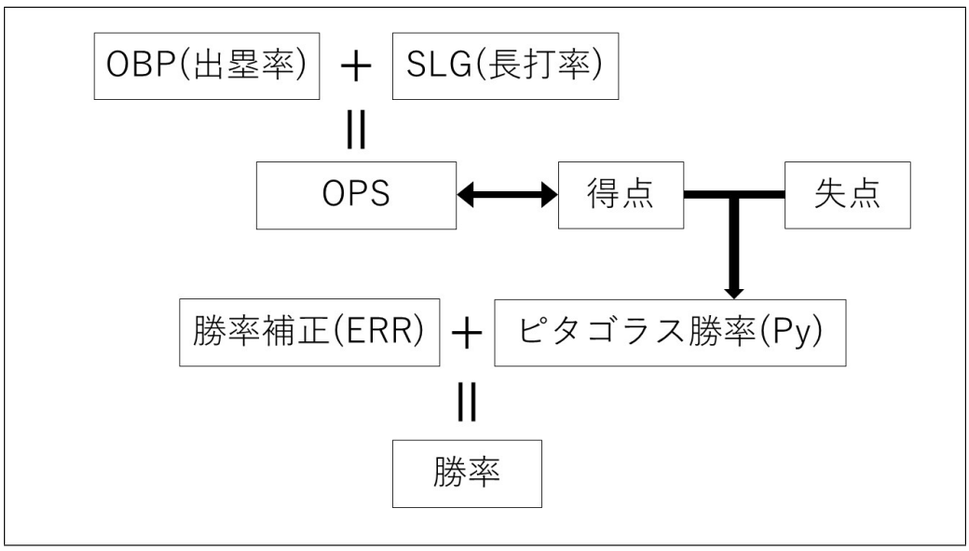

図 2　勝率を決定する過程と指標間の関連性のイメージ

4.3　仮定するもの

以下は本研究を進める上で仮定した事項の一覧である.

1. データの正規性（4.5 節を参照）
2. ペナントレースにおけるチームの年間成績のみを再現する. 各試合結果, 個人成績, ポストシーズンの結果などは完全に無視する.
3. チームの得点数は, OPS を説明変数, 得点数を目的変数とする単回帰式で求められるものとする. 得点数と失点数の計算結果は小数を含む可能性があるが, 整数値に丸めることで対処する.
4. チームの年間勝率は「ピタゴラス勝率」または「ピタゴラス勝率 + 誤差」で近似できるものとする.
5. 年数経過によるチームの戦力の変化を考慮しない. また, 各年度は独立しているものとする. （年度 $n+1$ は年度 n の 1 年後ではない）
6. チームの攻撃力と守備力は互いに独立しているものとする.
7. 仮想データ上においてもセ・パ両リーグの区別は行う. [*1]
8. 同一リーグ間のチームの区別は行わず, サンプル番号 n 番のチームがどの球団に相当するかは考えない.

7, 8 より, 「リーグ間の格差」は考慮するが, 「チーム間の格差」は考慮しないという状況になる.

[*1] プログラム上では「リーグ 1」「リーグ 2」と表示するが, リーグ 1 はパ・リーグ, リーグ 2 はセ・リーグに対応する.

表 2　交流戦におけるリーグ間の得失点差 ('06〜'19)

年度	2019	2018	2017	2016	2015	2014	2013
試合数	18	18	18	18	18	24	24
パ得点	476	473	459	425	462	634	601
セ得点	449	409	431	359	390	576	502
パ得点 − セ得点	27	64	28	66	72	58	99
18 試合換算	27	64	28	66	72	43.5	74.25
年度	2012	2011	2010	2009	2008	2007	2006
試合数	24	24	24	24	24	24	36
パ得点	489	504	701	613	624	562	841
セ得点	469	365	546	588	562	525	904
パ得点 − セ得点	20	139	155	25	62	37	-63
18 試合換算	15	104.25	116.25	18.75	46.5	27.75	-31.5

4.4　収集した実データ

正規乱数を使用するために必要なデータとして, NPB の公式 HP から「チームの出塁率」「チームの長打率」などを引用した. 図 3 は出塁率と長打率の関係を表した散布図である. 実データの具体的な数値は参考資料に掲載した.

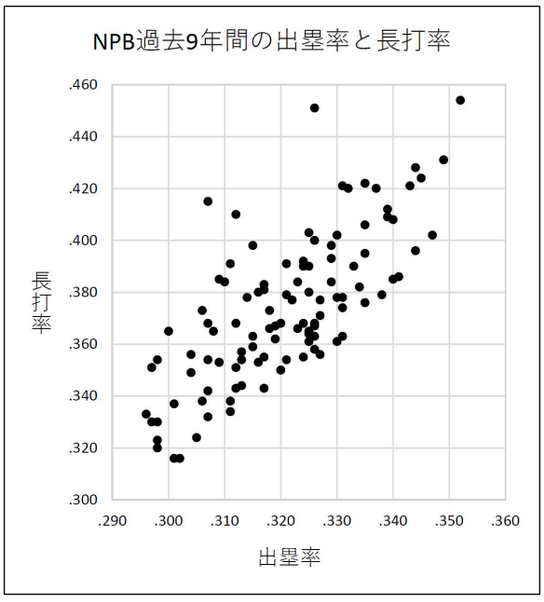

図 3　球団ごとの出塁率と長打率 ('11〜'19, $n = 108$)

4.5　データの正規性の判断

正規乱数を用いて仮想データを作成するためには, 以下の実データについて正規性（母集団が正規分布に従っている事）を確認する必要がある.

1. OBP, SLG（2 次元正規分布に従うことを確認する）
2. 1 年間の失点数
3. 交流戦におけるリーグ間の得失点差（18 試合換算）
4. (勝率) − Py(1.69)

そこで, **Q-Q プロット**による目視での判断と, **コルモゴロフ・スミルノフ検定**による正規性の仮定を試みる.

正規 Q-Q プロット (文献 [11]) とは, 観測値が正規分布に従う場合の期待値を Y 軸にとり, 観測値そのものを X 軸にとった確率プロットである. 観測値を昇順に並べた順位からパーセンタイル（累積確率）を求め, 正規分布の確率密度関数の逆関数を用いて期待値を予測する. プロットが一直線上に並べば, 観測値は正規分布に従っていると考えられる.

1 標本コルモゴロフ・スミルノフ検定とは, ある標本の母集団の確率分布が帰無仮説で提示した確率分布と一致しているかどうかを検定するものであり, KS 検定とも言う. 正規性の検定データの累積確率分布と正規分布の累積確率の差の絶対値の最大値を検定統計量 D とし検定を行う.

$$D = \max_x |F(x) - F_n(x)|$$