仮想的に再現した打撃成績に基づく プロ野球の勝率予測

芝浦工業大学　数理科学研究会
BV17057　西脇 友哉

2019 年 5 月 19 日

目次

第Ⅰ部

導入

研究背景

本研究は 2018 年度芝浦祭及び第 8 回サイエンス・インカレで発表を行った「OPS の計算式における出塁率と長打率の最適な価値配分〜一見手抜きだが便利な指標〜」の続編に位置する. 先行研究で年単位の限られたデータしか収集できなかった事にもどかしさを感じ, シミュレーションによって大量のデータを擬似的に生成したいと考えた. 最も精密な結果を得られる手段は NPB ペナントレース全試合を再現することであるが, 著者の技能では困難な上に「“野球ゲームのペナントモードをオート進行する”という形で既に実装されているのではないか」と考えたため, 簡易的な方法として, 各試合結果ではなく 1 年間の成績のみの再現を試みた. 先行研究では **OPS** や**ピタゴラス勝率**など既存の指標に多変量解析の手法を適用して勝率予測を行ったが, 本研究では, 先行研究で得られた知見と MATLAB によるシミュレーション結果を基に勝率予測を行う.

研究手順の概略

1. データの各項目について正規性（母集団が正規分布に従っている事）を確認する
2. 2 次元正規乱数によって OBP（出塁率）と SLG（長打率）の組を出力する
3. 上記の各組に対して **OPS**(=OBP+SLG) を計算する
4. OPS を説明変数, 得点数を目的変数とする単回帰式を利用して得点数を計算する
5. 正規乱数を用いて失点数を出力し, 微調整（詳細は後述）を加える
6. **ピタゴラス勝率**を用いて得点数と失点数から勝率を推定する
7. サンプル 6 個を 1 セットとしてピタゴラス勝率の降順に並べ替え, 順位を付ける

前提

- 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 ）
  本研究において, 過去の選手の成績を提示する際には慣例に従い「0」を省略して表記しているが, 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, 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)
- 紙面の都合上, NPB 所属球団の略称を以下の通りに表記する. ただし, 2011 年度以前の「横浜ベイスターズ」についても「横浜 DeNA ベイスターズ」と同様に, 「横」もしくは「DB」と表記する.

パシフィック・リーグ			セントラル・リーグ		
西	L	埼玉西武ライオンズ	広	C	広島東洋カープ
ソ	H	福岡ソフトバンクホークス	ヤ	S	東京ヤクルトスワローズ
日	F	北海道日本ハムファイターズ	巨	G	読売ジャイアンツ
オ	B	オリックス・バファローズ	横	DB	横浜 DeNA ベイスターズ
ロ	M	千葉ロッテマリーンズ	中	D	中日ドラゴンズ
楽	E	東北楽天ゴールデンイーグルス	神	T	阪神タイガース

第 1 章

用語解説

1.1 野球用語

一部の野球の指標について解説を行う．全て他の書籍や Web ページ上にも掲載されている内容であるが, 本研究は以下の記録や指標を前提知識とするため, この章の内容を先に確認することが望ましい.

各項目にはそれぞれ規定打席到達者[*1]を対象とした成績と当てはまる人数の度数分布表を掲載した．控え選手の成績が反映されないため, 厳密にリーグの平均などを表すものではないが, 各指標に対するイメージを表から掴むことは出来るはずである.

1.1.1 公式記録

打率, 打点, 本塁打, 四球, 犠打などが打撃の公式記録に該当する．これらは従来から打者の評価に用いられてきた指標であり, 特に打率, 打点, 本塁打を総称して「打撃 3 部門」とよばれるなど, 長いプロ野球の歴史において確固たる地位を築いてきた．ここでは OPS の導出に用いられる出塁率と長打率の解説を行う．これらもまた公式記録に該当するが, 成績を評価する尺度としては打率等と比べあまり注目されていなかった．現在では, 後述のセイバーメトリクスの影響により重要性が高まっている.

出塁率 (OBP) On-base percentage

出塁率は文字通り「出塁する確率」を表すが, 「アウトにならない確率」と解釈をすることもできる．打率との最大の違いは, 以下の**計算式 2** が表すように四死球による出塁が考慮に入れられている点である.

計算式 1 (打率)**.**

$$\text{打率} = \text{安打} \div \text{打数}$$

計算式 2 (出塁率)**.**

$$\text{出塁率} = (\text{安打} + \text{四球} + \text{死球}) \div (\text{打数} + \text{四球} + \text{死球} + \text{犠飛})$$

打率を「安打を打つことを試みた回数に対して実際に安打を記録した回数の割合」と捉えるならば, 出塁率は「出塁を試みた回数に対して実際に出塁した回数の割合」といえる．ここでの「出塁を試みた回数」は分母の (打数 + 四球 + 死球 + 犠飛) に相当する.

野球の攻撃においてアウトにならないことは極めて重要であり, そこに四死球の価値が見出される．かつて出塁率は打率に比べて注目度が低かったが, 近年では打撃指標としての有用性が見直されている.

なお, 犠飛は打数に含まれないため打率の計算では分母から除外されているが, 出塁率の計算においては犠飛は分母に含まれる（凡退として扱われる）ため注意が必要である．四球と犠飛の数によっては出塁率が打率よりも低くなる例が存在しうる.

[*1] 規定打席：所属球団の試合数 ×3.1（小数点以下四捨五入）と定められている．つまり, ペナントレース全 143 試合消化時点での規定打席は 443 打席である.

例 1.1. ’00 の荒木雅博選手（中）が打率 .200, 出塁率 .167 を記録している. (打席数 12, 打数 10, 安打数 2, 四死球 0, 犠飛 2)

表 1.1　規定到達者の出塁率 (OBP) の分布 (’11〜’18)

OBP		2018		2017		2016		2015		2014		2013		2012		2011	
以上	未満	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ
.400		4	9	1	0	4	4	3	2	2	4	1	4	1	1	1	0
.380	.400	3	4	1	7	3	2	3	2	5	2	9	1	1	1	1	1
.360	.380	5	7	6	7	3	4	6	5	5	7	6	4	3	4	5	2
.340	.360	4	1	7	2	7	9	7	5	11	7	5	4	9	5	4	4
.320	.340	6	4	1	7	7	4	4	3	4	4	5	5	5	5	7	10
.300	.320	5	5	11	2	3	1	3	5	1	3	5	2	7	5	6	7
	.300	2	1	0	3	1	3	4	2	3	0	2	1	4	3	3	0
計		29	31	27	28	28	27	30	24	31	27	33	21	30	24	27	24

長打率 (SLG) Slugging percentage

長打率はその字面から「長打を打つ確率」と解釈されがちであるが, 長打率が 1.000 を超えない限りは内野安打でも長打率は上昇する. 実際は「1 打数につき稼いだ進塁の数」を意味する指標であり, 計算式 3 によって求められる.

計算式 3 (長打率)**.**

$$\text{長打率} = \text{塁打} \div \text{打数}$$

計算式 4 (塁打)**.**

$$\begin{aligned}\text{塁打} &= \text{単打} \times 1 + \text{二塁打} \times 2 + \text{三塁打} \times 3 + \text{本塁打} \times 4 \\ &= \text{安打} + \text{二塁打} + \text{三塁打} \times 2 + \text{本塁打} \times 3\end{aligned}$$

表 1.2　規定到達者の長打率 (SLG) の分布 (’11〜’18)

SLG		2018		2017		2016		2015		2014		2013		2012		2011	
以上	未満	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ
.550		3	6	1	1	0	4	2	1	2	1	2	3	0	1	1	0
.500	.550	3	5	2	6	4	3	4	1	1	5	4	1	0	0	1	0
.450	.500	8	7	7	1	6	6	5	3	7	5	4	4	4	3	1	4
.400	.450	5	3	6	12	5	5	4	9	8	9	9	6	5	4	6	7
.350	.400	6	9	10	5	5	5	9	6	7	3	6	4	11	8	11	5
.300	.350	3	1	1	2	7	3	4	3	5	4	7	3	6	7	5	5
	.300	1	0	0	1	1	1	2	1	1	0	1	0	4	1	2	3
計		29	31	27	28	28	27	30	24	31	27	33	21	30	24	27	24

長打率が最大となるのは全打数で本塁打を打った場合なので最大値は 4.000 だが, NPB 歴代最高は .779（’13 W. バレンティン選手（ヤ)）である. 打率では安打はすべて等しい価値として計算されるが, 長打率の計算においては単打, 二塁打, 三塁打, 本塁打にそれぞれ異なる重み付けがなされるため, 出塁の頻度が低くても塁を稼げる（=長打力のある）打者を評価することができる.

例 1.2. ’17 に鳥谷敬選手（神）は出塁率 .390, 長打率 .377 であったのに対し, 同年の鈴木誠也選手（広）は出塁率 .389, 長打率 .547 であった．この 2 選手の成績を比較すると, 出塁率はほぼ同じであるが, 長打率は鈴木選手が鳥谷選手を大きく上回っている．この事から, 鈴木選手は鳥谷選手よりも「塁を稼ぐ効率が良い」といえる．

1.1.2 セイバーメトリクス指標

セイバーメトリクス (SABR metrics) は, アメリカ野球学会の略称である「SABR」と測定法を意味する「metrics」を足した造語であり, 1980 年にビル・ジェームズ氏により提唱された統計学的見地による野球の分析手法である．選手の成績から得点と失点への貢献度, ひいてはその選手に何勝分の価値があるかを計ることを主とし, その分析のために様々な指標を用いる．そして「OPS」は打撃成績を評価する最も基本的なセイバーメトリクス指標の 1 つである．

以下の項目は OPS 及び勝率の予測に用いられる「ピタゴラス勝率」の解説となっている．セイバーメトリクス指標は「打撃」「投球」「守備」「走塁」それぞれについて非常に多岐に渡って存在するものであり, この章で解説する指標はそれらの内のほんの一部に過ぎないという点は注意されたい．

OPS (On-base Plus Slugging)

OPS は, 前述の出塁率と長打率を同時に用いて評価することができ, いわゆる「強打者」が高い数値を示す指標である．理論上の最大値は 5.000（出塁率 : 1.000, 長打率 : 4.000）だが, NPB, MLB ともに例年リーグの平均は .700 程度である．OPS が提唱された背景には「打撃 3 部門」の指標としての欠陥があり, 特に打率に代わる評価基準として, 選手の年俸の査定や獲得選手の選考などに用いられるほどに定着した．

計算式 5 (OPS)**.**

$$\text{OPS} = \text{出塁率} + \text{長打率}$$

上記の式が示すように出塁率と長打率をそのまま足しただけという極めて単純な構造でありながら, 得点との連動性の高さに非常に優れている点が他の指標にはない長所である．文献 [11] によると, NPB の過去 30 年分のシーズン成績によるチーム OPS と 1 試合の平均得点との重相関係数は $R^2 = 0.901$ であった．それに対し出塁率と平均得点では $R^2 = 0.7284$, 長打率と平均得点では $R^2 = 0.8182$ であり, ここから OPS による評価の精度の高さがうかがえる．

欠点としては走塁能力が考慮されていないことなどが挙げられる．この論文では解説を割愛するが, リーグや年度の異なる選手同士の OPS を比較するためにリーグ平均からの傑出度を表した OPS+ や, 評価項目を更に細分化した wOBA, wRC+ などといった指標が現在のセイバーメトリクスの主流とされている．しかしそれらは非常に計算式が複雑であることから, OPS の簡易指標としての有用性は非常に高いと言える．

表 1.3　規定到達者の長打率 (OPS) の分布 (’11〜’18)

OPS		2018		2017		2016		2015		2014		2013		2012		2011	
以上	未満	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ
1.000		1	4	1	0	0	3	1	1	0	1	0	2	0	0	0	0
.900	1.000	4	6	1	3	1	1	2	1	3	3	4	1	0	1	1	0
.800	.900	9	10	5	9	10	9	7	4	8	10	9	7	4	4	3	3
.700	.800	11	7	16	12	9	10	10	12	13	9	10	7	13	7	12	10
.600	.700	2	3	4	3	8	3	9	5	7	4	9	4	9	10	11	9
	.600	2	1	0	1	0	1	1	1	0	0	1	0	4	2	0	2
計		29	31	27	28	28	27	30	24	31	27	33	21	30	24	27	24

ピタゴラス勝率

ピタゴラス勝率とは, セイバーメトリクスの考案者であるビル・ジェームズ氏が提唱した式であり, 得点と失点から勝率を推定するために用いられる. 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 6 (ピタゴラス勝率)**.**

$$\text{ピタゴラス勝率} = \frac{(\text{総得点})^2}{(\text{総得点})^2 + (\text{総失点})^2} = \frac{1}{1+\rho^2} \quad \left(\rho = \frac{\text{総失点}}{\text{総得点}}\right)$$

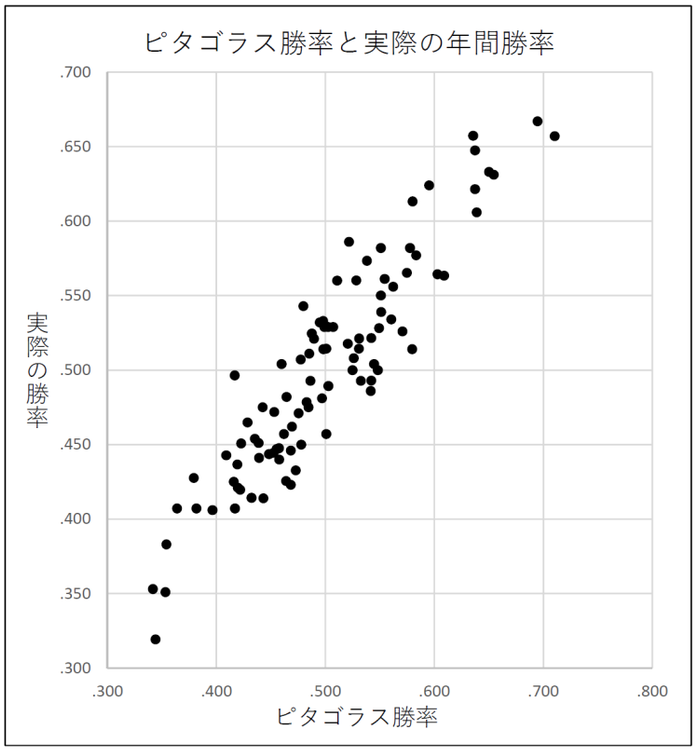

図 1.1　ピタゴラス勝率と実際の年間勝率の散布図 ('11〜'18)

一般にピタゴラス勝率は実際の試合結果に基づく勝率との相関が強いといわれている. 実際に'11〜'18 の 8 年間について相関係数を計算したところ, $r = 0.920$ (95% 信頼区間：$[0.883, 0.946]$) となり, 強い正の相関関係がみられた. 図 1.1 からも直線的な関係がうかがえる.

本来, ピタゴラス勝率は実際の勝率と比較することでチームの勝率が得点力や投手力に対して順当であるかを考察する[*2]という利用方法が一般的であるが, 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ピタゴラス勝率を利用する.

1.2　造語の定義

用語 1 (実データ)**.** 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.

用語 2 (仮想データ)**.** 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.

[*2] 例として, あるチームの勝率がピタゴラス勝率よりも極端に低い場合, 「接戦での敗戦が多い」「継投策などの采配面に問題がある」などの原因が考えられる.

第 2 章

シミュレーションの準備

2.1　収集した実データ

1.1.2 では OPS を打者個人の成績として紹介したが, 本研究の特性上, チームの記録としても扱う. 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 に掲載されている'11〜'18 のペナントレースにおける各球団の打撃成績から「チームの出塁率」「チームの長打率」などを引用した. 図 2.1 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については, 末尾の補足資料を参照されたい.

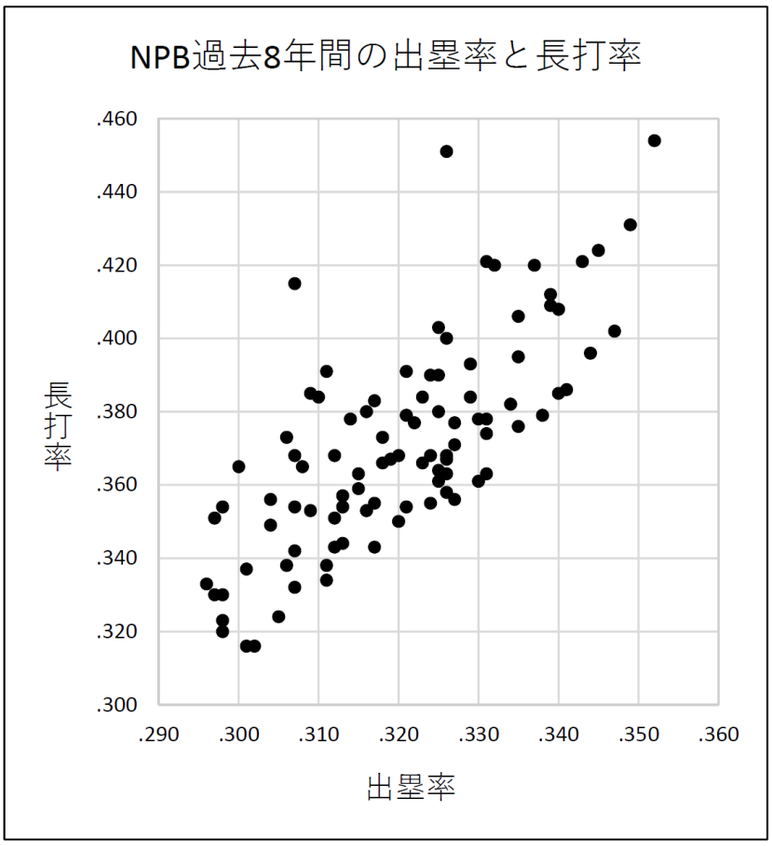

図 2.1　球団ごとの出塁率と長打率 ('11〜'18, $n = 96$)

収集した出塁率 (OBP) と長打率 (SLG) のデータから, 平均 (μ), 分散 (Var), 共分散 (Cov) を計算したところ,

$$\begin{cases} \mu(\mathrm{OBP}) & = 0.320 \\ \mu(\mathrm{SLG}) & = 0.371 \\ \mathrm{Var}(\mathrm{OBP}) & = 0.000185 \\ \mathrm{Var}(\mathrm{SLG}) & = 0.000822 \\ \mathrm{Cov}(\mathrm{OBP},\ \mathrm{SLG}) & = 0.000291 \end{cases} \tag{2.1}$$

が得られた.

2.2 データの正規性の判断

正規乱数を用いたシミュレーションを行うにあたり, 以下の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
2. SLG
3. 1 年間の失点数
4. 交流戦におけるリーグ間の得失点差（18 試合換算）

表 2.1　交流戦におけるリーグ間の得失点差 ('06〜'18)

年度	2018	2017	2016	2015	2014	2013	2012
試合数	18	18	18	18	24	24	24
パ得点	473	459	425	462	634	601	489
セ得点	409	431	359	390	576	502	469
パ得点 − セ得点	64	28	66	72	58	99	20
18 試合換算	64	28	66	72	43.5	74.25	15

年度	2011	2010	2009	2008	2007	2006
試合数	24	24	24	24	24	36
パ得点	504	701	613	624	562	841
セ得点	365	546	588	562	525	904
パ得点 − セ得点	139	155	25	62	37	-63
18 試合換算	104.25	116.25	18.75	46.5	27.75	-31.5

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 [17])

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. 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

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こと. 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.

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

正規性の検定を行う場合, 比較対象は正規分布である. 帰無仮説及び対立仮説は以下の様になり, $p < 0.05$ のとき帰無仮説が棄却される.

帰無仮説 H_0 : データの分布は正規分布と一致している
対立仮説 H_1 : データの分布は正規分布と一致していない

統計的仮説検定において帰無仮説が棄却された場合は対立仮説を採用できるが, 帰無仮説が受容されても帰無仮説を採用することは出来ない. コルモゴロフ・スミルノフ検定の帰無仮説が受容された場合に得られる結論は「データの分布が正規分布と一致しないとは言えない」という曖昧なものであり, 本来は検定結果を参考程度に捉えるべきであるが, 本研究では便宜上帰無仮説が棄却されなかったときにデータの正規性を仮定することとする.

検定の結果

図 2.2〜図 2.5 に示した Q-Q プロットから，どのプロットも直線から大きく外れてはいないことが分かる．また，先述の 1.〜4. についてコルモゴロフ・スミルノフ検定を行ったところ，p 値は順に $0.7979, 0.6561, 0.8898, 0.9961$ となり，全て 0.05 を上回っていたため帰無仮説は受容された．よって，上記の 4 種類のデータについて正規性を仮定する．ただし，「交流戦におけるリーグ間の得失点差」については実データのサンプル数が少ないため，t 分布に従う乱数を用いる．

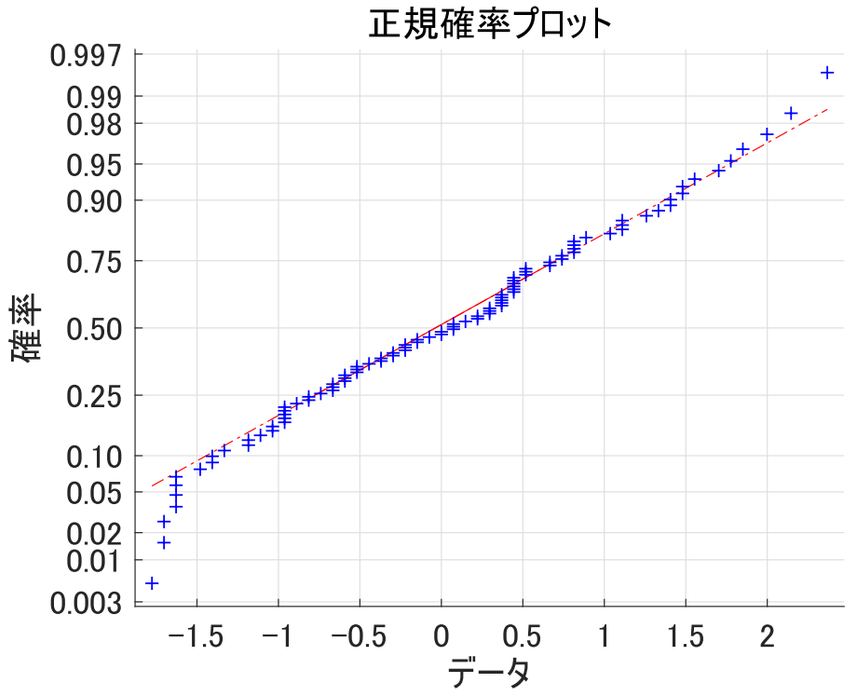

図 2.2　正規確率プロット（OBP）

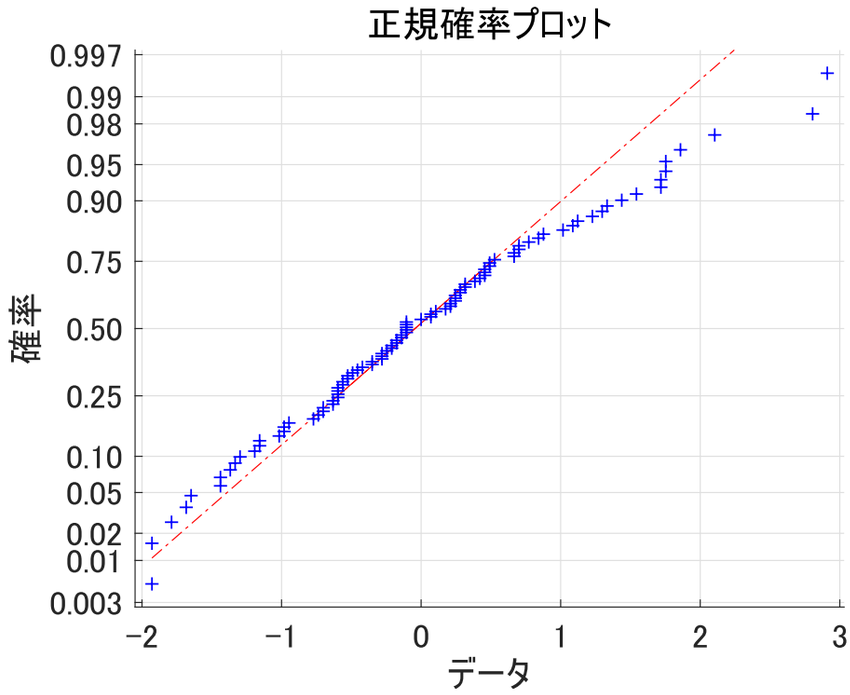

図 2.3　正規確率プロット（SLG）

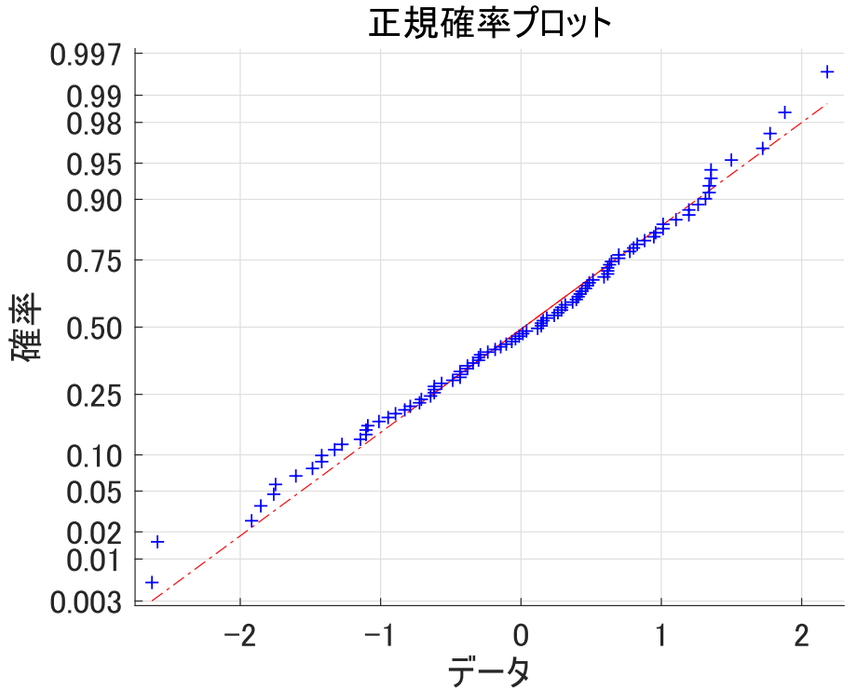

図 2.4　正規確率プロット（年間の失点数）

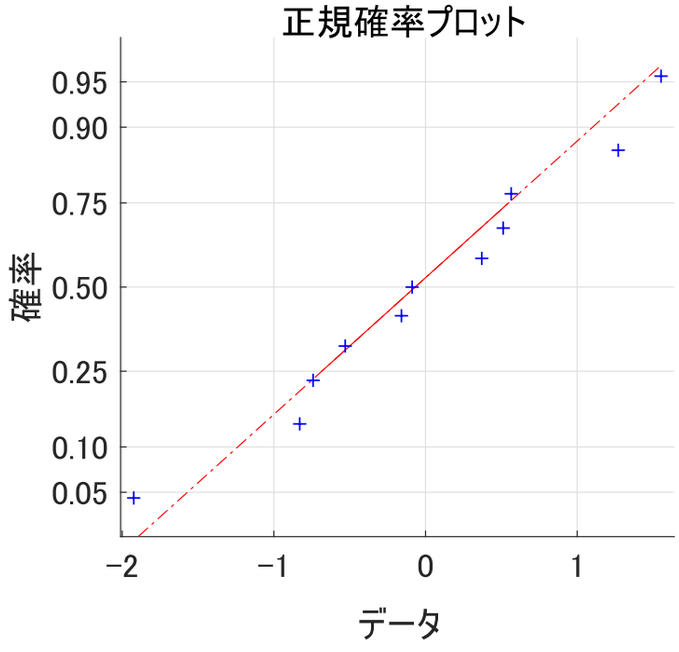

図 2.5　正規確率プロット（交流戦の得失点差）

2.3 OPS と得点の線形回帰モデル

年間総得点数を OPS に帰着させるために, OPS と得点の間にある関係を線形関数による回帰式で表す必要がある. そこで, **最小 2 乗法**を用いて**回帰直線**を導出し, 単回帰分析を行う.

2.3.1 最小 2 乗法の手順

標本サイズ n の 2 変数の観測値 $(x_i, y_i), i = 1, 2, \ldots, n$ に対して説明変数 x と目的変数 y の間に成り立つ直線関係を表す**線形回帰モデル**

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \cdots, n.$$

を考える. ここで, β_0, β_1 を**回帰係数**, ε_i を**誤差**といい, y_i の**予測値** $\hat{y}_i$ は

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{2.2}$$

で与えられる. 従って, 実測値 y_i と予測値 $\hat{y}_i$ の差は $e_i = y_i - \hat{y}_i$ で与えられ, このとき e_i は**残差**と呼ばれる.

最小 2 乗法とは, 以下に示す**残差平方和** $S(\hat{\beta}_0, \hat{\beta}_1)$ が最小となるように β_0, β_1 の推定値 $\hat{\beta}_0, \hat{\beta}_1$ を求める方法である.

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{2.3}$$

式 (2.3) は $\hat{\beta}_0, \hat{\beta}_1$ を変数とする 2 次関数であることから, S を $\hat{\beta}_0, \hat{\beta}_1$ で偏微分し 0 とおく.

$$\begin{aligned}
\frac{\partial S}{\partial \hat{\beta}_0} &= -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\
\frac{\partial S}{\partial \hat{\beta}_1} &= -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0
\end{aligned}$$

これらを移項して, **正規方程式**と呼ばれる $\hat{\beta}_0, \hat{\beta}_1$ を未知数とする線形方程式

$$\begin{cases}
n\hat{\beta}_0 + \left(\displaystyle\sum_{i=1}^{n} x_i \right) \hat{\beta}_1 & = \displaystyle\sum_{i=1}^{n} y_i \\
\left(\displaystyle\sum_{i=1}^{n} x_i \right) \hat{\beta}_0 + \left(\displaystyle\sum_{i=1}^{n} x_i^2 \right) \hat{\beta}_1 & = \displaystyle\sum_{i=1}^{n} x_i y_i
\end{cases} \tag{2.4}$$

を解くと, $\overline{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i, \overline{y} = \dfrac{1}{n} \sum_{i=1}^{n} y_i$ として, 以下の β_0, β_1 の推定値が得られる.

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{2.5}$$

$$\hat{\beta}_1 = \frac{n \left(\displaystyle\sum_{i=1}^{n} x_i y_i \right) - \left(\displaystyle\sum_{i=1}^{n} x_i \right) \left(\displaystyle\sum_{i=1}^{n} y_i \right)}{n \displaystyle\sum_{i=1}^{n} x_i^2 - \left(\displaystyle\sum_{i=1}^{n} x_i \right)^2} \tag{2.6}$$

2.3.2 数値の代入と単回帰式の計算

OPS を説明変数 x, 年間総得点を目的変数 y へそれぞれ対応させると, 以下のような計算結果を得る.

$$n = 96, \quad \overline{x} = 0.692, \quad \overline{y} = 550.990 \tag{2.7}$$

$$\sum_{i=1}^{n} x_i = 66.406, \quad \sum_{i=1}^{n} y_i = 52895, \quad \sum_{i=1}^{n} x_i^2 = 46.086, \quad \sum_{i=1}^{n} x_i y_i = 36874.46 \tag{2.8}$$

式 (2.7), 式 (2.8) をそれぞれ式 (2.5), 式 (2.6) へ代入すると,

$$\hat{\beta}_1 = \frac{96 \times 36874.46 - 66.406 \times 52895}{96 \times 46.086 - 66.406^2} = 1890.903 \tag{2.9}$$

$$\hat{\beta}_0 = \frac{1}{96}(52895 - 1890.903 \times 66.406) = -757.003 \tag{2.10}$$

となり, 単回帰直線

$$\hat{y} = -757.003 + 1890.903x \tag{2.11}$$

を得る. OPS と総得点の散布図と回帰直線を平面上に表すと図 2.6 のようになる.

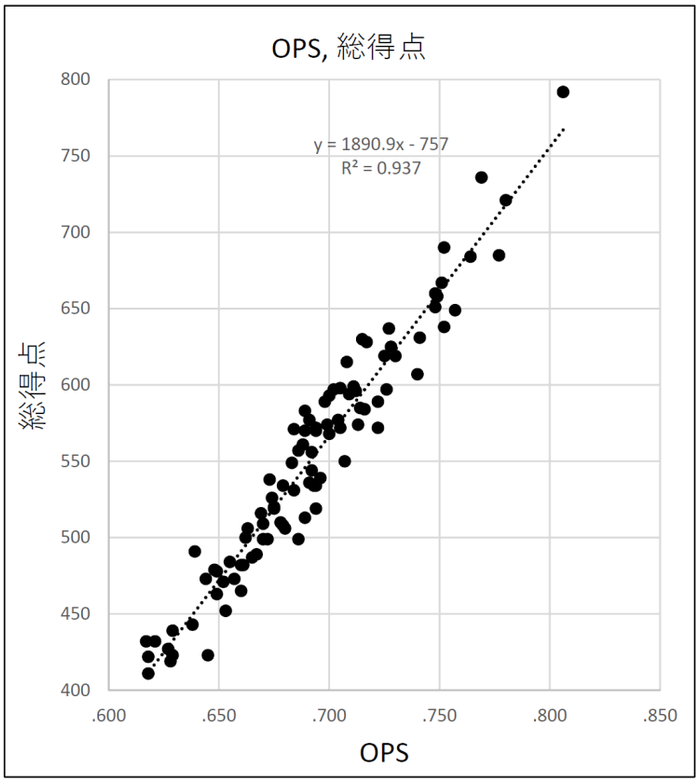

図 2.6　OPS, 総得点の散布図と回帰直線 (’11〜’18)

2.3.3 寄与率

寄与率を用いて, 推定された回帰直線の適合度を評価する.

寄与率（回帰分析）

標本サイズ n の 2 変数の観測値 $(x_i, y_i), i = 1, 2, \ldots, n$ が与えられたとき, 推定された回帰直線から計算された y_i の予測値を $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ とする. そして, 実測値 y_i, 予測値 $\hat{y}_i$ および残差 $e_i = y_i - \hat{y}_i$ のそれぞれの平方和を

$$SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2, \quad SS_R = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2, \quad SS_E = \sum_{i=1}^{n} e_i^2 \tag{2.12}$$

とするとき $\left(\text{ただし } \overline{y} \text{ は } y_i \text{ の平均値 } \overline{y} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} y_i \right)$, SS_T を**総変動**, SS_R を**回帰変動**, SS_E を**残差変動**と呼ぶ. このとき, それぞれの変動には次の関係

$$SS_T = SS_R + SS_E$$

がある. このような関係式のことを**回帰分析の変動分解**という. このとき, 回帰変動 SS_R は推定された回帰直線が当てはまっている度合いを表しており, 残差変動 SS_E は推定された回帰直線が当てはまっていない度合いを表す. 回帰変動が総変動に占める割合を計算することで, 推定された回帰直線の適合度を要約した指標が**寄与率（決定係数）**である. したがって, 寄与率は

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \tag{2.13}$$

である. 寄与率 R^2 の範囲は $0 \leq R^2 \leq 1$ であり, 1 に近づくほど良く当てはまっていると解釈される.

式 (2.12) について, データより $SS_T = 576029.0, SS_R = 539755.8, SS_E = 36237.2$ を式 (2.13) に代入し, 寄与率 $R^2 = 0.937$ を得た. ここから, 式 (2.11) の回帰直線はデータに良く当てはまっていると言える.

2.4 ピタゴラス勝率の最適化

1.1.2 で紹介したように, ピタゴラス勝率はスポーツの分析に広く利用されている式であるが, 厳密には異なるスポーツに対して異なる指数を要する. つまり, 計算式 6 において ρ^2 を ρ^x で置き換え, 分析対象のスポーツに対して最適化しなければならない. 本研究では, 文献 [15](P.53) を参考に**平均絶対偏差** (MAD, mean absolute deviation) を用いてピタゴラス勝率と実際の勝率のズレを評価する.

平均絶対偏差（MAD）

$$\begin{cases} \omega_i & = \text{チーム } i \text{ のシーズン中の勝率} \\ \rho_i & = \dfrac{\text{チーム } i \text{ の失点数}}{\text{チーム } i \text{ の得点数}} \end{cases} \tag{2.14}$$

とする. リーグに m チームいるとき, x の与えられた値に対する平均絶対偏差は

$$\mathrm{MAD}(x) = \frac{1}{m} \sum_{i=1}^{m} \left| \omega_i - \frac{1}{1 + \rho_i^x} \right| \tag{2.15}$$

となる.

式 (2.15) の指数 x の最適な値 x^* は, x の十分多くの異なる値に対して $\mathrm{MAD}(x)$ を計算し, $\mathrm{MAD}(x)$ を最小にする x を特定するという（ある意味力づくな）方法で決定される. 本研究の対象である NPB のデータに対しては, 1 から 4 まで 0.002 刻みの x に対して $\mathrm{MAD}(x)$ を計算する事で

$$x^* = 1.69, \quad \mathrm{MAD}(x^*) = 0.0230$$

と評価し, 図 2.7 にこの計算結果を表した．これは最適化されたピタゴラス勝率 $1/(1+\rho^{x^*})$ によって作られる予測が 2011-2018 の NPB ペナントレースでチームごとに平均 2.30% しか外れていないことを意味する．

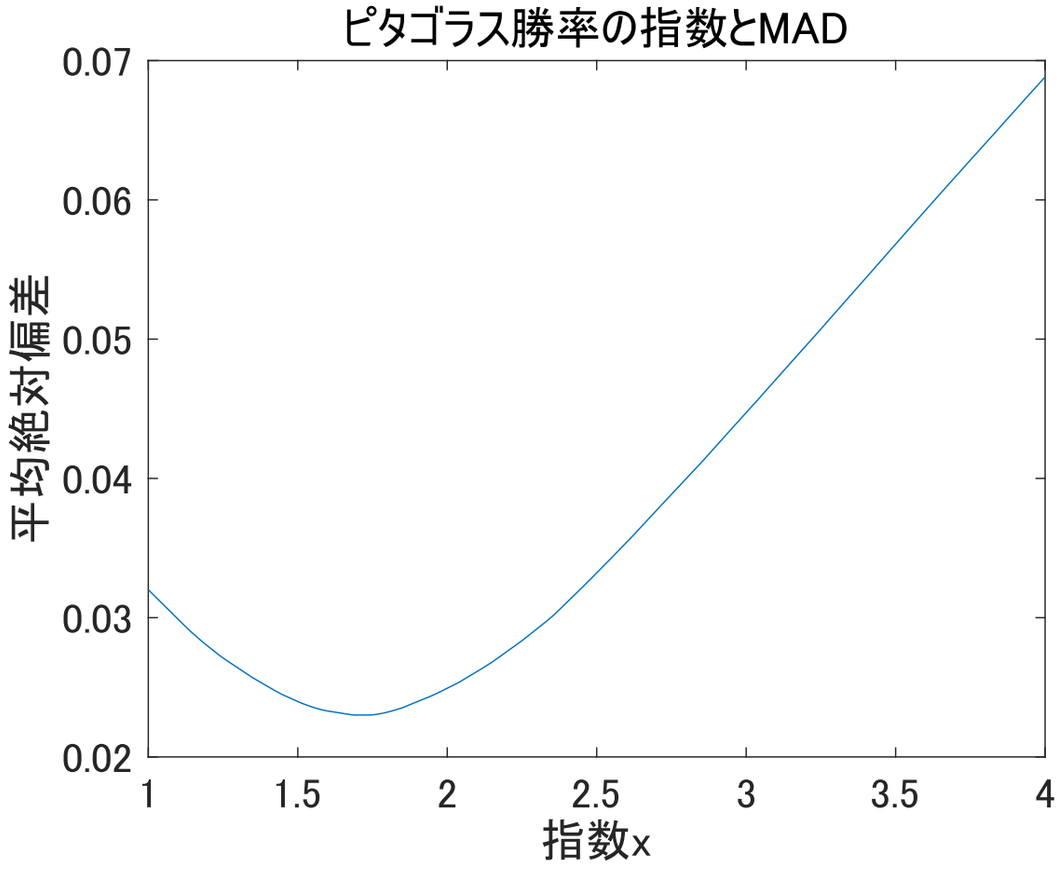

図 2.7　x に対する MAD(x) の関係

第 II 部

正規乱数を用いた打撃成績の仮想再現

第 3 章

仮想データを生成するプログラム

3.1　仮想データの構成について

シミュレーションによって延べ 6000 チーム分の仮想データを出力する．これを 12 チームずつ区切る事で年度の違いを想定し, 更に 6 チームずつに分けることで 2 リーグ制を再現する．仮想データ上においてもセ・パ両リーグの区別は行うが,[*1] 同一リーグ間のチームの区別は行わない．（サンプル番号 n 番のチームが NPB のどの球団に相当するかは考えない．）

表 3.1　サンプル番号, 年度, リーグの表示

No.	年度	リーグ	成績
1	1	1	
2			
3			
4			
5			
6			
7		2	
8			
9			
10			
11			
12			
13	2	1	
14			
⋮	⋮	⋮	⋮
5998	500	2	
5999			
6000			

3.2　年間得点

2 次元正規乱数によって OBP と SLG の組を生成する．つまり，「OBP の平均」「SLG の平均」「OBP と SLG の分散共分散行列」の数値を反映させた 2 次元正規分布に従う 6000 組の 2 次元乱数を出力する．式 (2.1) より, 2 次元正規

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

分布の平均と分散共分散行列は

$$\mu = \begin{pmatrix} 0.320 \\ 0.371 \end{pmatrix},\ \Sigma = \begin{pmatrix} 0.000185 & 0.000291 \\ 0.000291 & 0.000822 \end{pmatrix}$$

と計算され, これを基に図 3.1 の様な出力を得た.

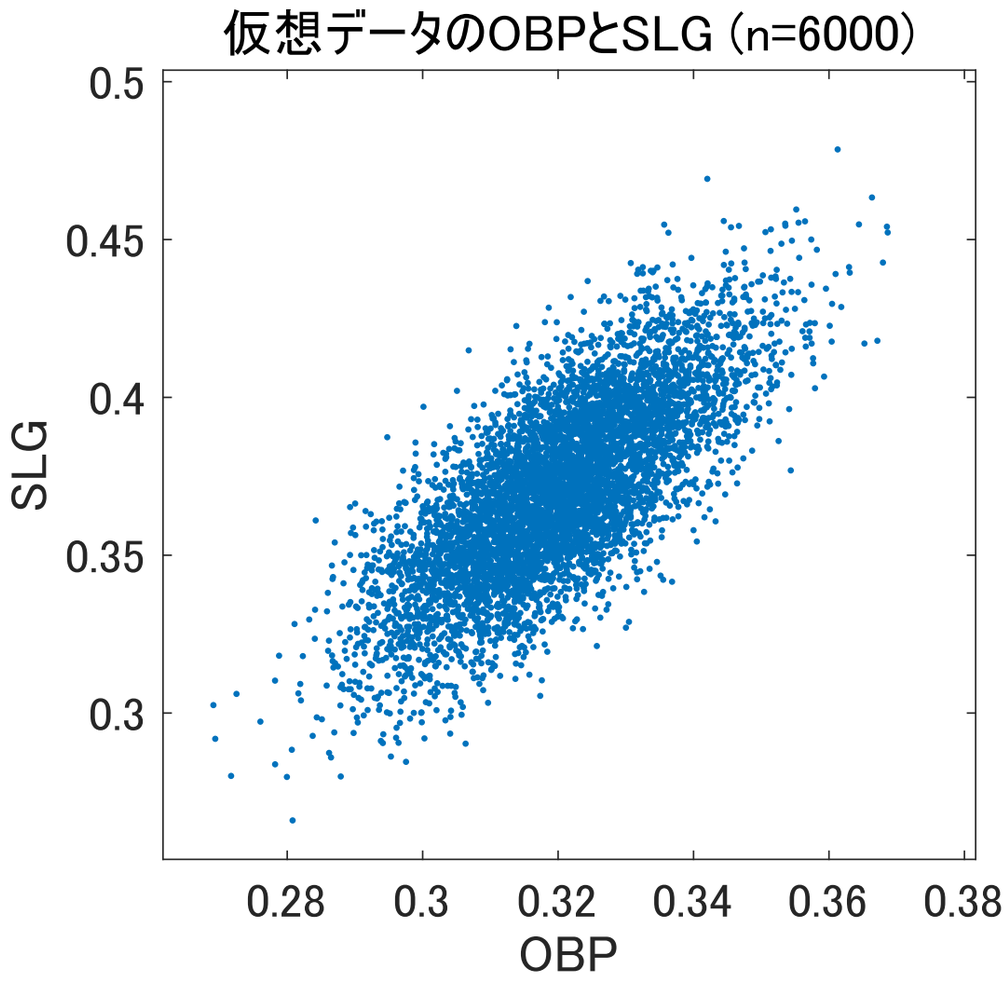

図 3.1　2 次元正規乱数による仮想データの OBP と SLG

並行して OPS[*2]の計算も行い, 2.3 で得られた次の単回帰式から得点数の推定値を求める.

$$y = -757.003 + 1890.903x \quad (x : \text{OPS},\ y : \text{得点数})$$

3.3　年間失点

3.3.1　失点数の生成と値の調節

実データによる失点数の平均と分散は

$$\mu = 550.99,\ \sigma = 76.07$$

である. これを用いて, 正規乱数から失点数のデータを生成する. ただし, 仮想データの失点数についてデータの整合性を確保するためには, 以下の性質を考慮しなければならない.

[*2] OPS = OBP + SLG

ある n チームで構成されたリーグで総当たり戦を行うとき,

$$\sum_{j=1}^{n}(\text{チーム } j \text{ の得点数}) = \sum_{j=1}^{n}(\text{チーム } j \text{ の失点数}) \tag{3.1}$$

という関係が成り立つ. NPB の場合, 交流戦におけるリーグ間の得失点差があるため

$$(\text{パ・リーグ 6 球団の得点数の合計}) = (\text{パ・リーグ 6 球団の失点数の合計})$$
$$(\text{セ・リーグ 6 球団の得点数の合計}) = (\text{セ・リーグ 6 球団の失点数の合計})$$

が常に成り立つとは限らない.(むしろ成り立つ事はほぼ無いと考えて良い.)
一方で,

$$(\text{12 球団の得点数の合計}) = (\text{12 球団の失点数の合計}) \tag{3.2}$$

は常に成り立つ.

この性質を満たす様に失点数を調整したい. ある年度について, 表 3.2 のような仮想データが生成されたとする.

表 3.2　仮想データ（得点数と失点数）

No.	リーグ	得点	失点	
			調整前	調整後
1	1	s_{11}	a'_{11}	a_{11}
2		s_{12}	a'_{12}	a_{12}
⋮		⋮	⋮	⋮
6		s_{16}	a'_{16}	a_{16}
7	2	s_{21}	a'_{21}	a_{21}
8		s_{22}	a'_{22}	a_{22}
⋮		⋮	⋮	⋮
12		s_{26}	a'_{26}	a_{26}

ここで, s, a' が既知, a が未知である状況を想定し, a の導出を試みる. 式 (3.2) の条件は

$$\sum_{\substack{i=1,2\\1\le j\le 6}} s_{ij} = \sum_{\substack{i=1,2\\1\le j\le 6}} a_{ij} \tag{3.3}$$

の成立を意味する. リーグ 1 における交流戦の得失点差を α とおくと,

$$\alpha = \sum_{j=1}^{6}(s_{1j} - a_{1j}) = -\sum_{j=1}^{6}(s_{2j} - a_{2j}) \tag{3.4}$$

となる. このとき, a は

$$a_{in} = a'_{in} \times \frac{\displaystyle\sum_{j=1}^{6} s_{ij} + (-1)^i \alpha}{\displaystyle\sum_{j=1}^{6} a'_{ij}} \tag{3.5}$$

とすれば良い. α の値は, t 分布乱数によって年度ごとに決定する.

3.3.2　失点数調整後も残る誤差の扱い

理論上は 3.3.1 での作業によって式 (3.2) が満たされるはずだが, 計算過程での丸め誤差によって左辺と右辺の数値が若干ずれることがある. ランダムに選ばれた 1 球団の失点数にその差を足す事で強引に差を解消する.

$$\sum_{j=1}^{6} s_{1j} = S_1 \text{ (リーグ 1 の合計得点)}, \qquad \sum_{j=1}^{6} a_{1j} = A_1 \text{ (リーグ 1 の合計失点)}$$
$$\sum_{j=1}^{6} s_{2j} = S_2 \text{ (リーグ 2 の合計得点)}, \qquad \sum_{j=1}^{6} a_{2j} = A_2 \text{ (リーグ 2 の合計失点)}$$

とおく. 本来 $S_1 + S_2 = A_1 + A_2$ であるが, 丸め誤差によって「リーグ 1 の合計失点」が $A_1 + \varepsilon$ になってしまったとする. このとき単純に ε を「リーグ 1 の合計失点」から引けば良いが,

$$A_1 := (A_1 + \varepsilon) + (S_1 + S_2 - (A_1 + \varepsilon) - A_2) = S_1 + S_2 - A_2$$

という演算を行うことで, プログラム上で ε を直接求める手間を省くことが出来る.
失点数に $(S_1 + S_2 - (A_1 + \varepsilon) - A_2)$ を足す対象となる 1 チームはリーグ 1 の中から一様乱数によって選ばれる.

3.4　勝率予測と順位付け

得点数と失点数が確定した後は, ピタゴラス勝率を用いて勝率の推定値を求める. 今回は, 指数部を 2（定義上の数値）とした場合と 1.69（最適解）とした場合で 2 通りの勝率を計算する. 表面上はピタゴラス勝率の序列を元にリーグの順位を決定するが, 本質的に順位を決定している要素は ρ (失点数 ÷ 得点数) である.

$$\left(\because \forall x > 0,\ \rho_1 < \rho_2 \Rightarrow \frac{1}{1+\rho_1^x} > \frac{1}{1+\rho_2^x} \right)$$

故に, 指数 x の取り方の違いによって順位が逆転することはないので, プログラム上では指数の値ごとに別々のソート関数を使う必要は無い.