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.

4.5.1 検定の結果

図 6〜図 10 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, 4.5 節の 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 $0.7982, 0.4229, 0.6335, 0.9563, 0.4516$ 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 5 種類のデータについて正規性を仮定するが, 「3. 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.

4.6 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.

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. このとき, 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, e_i は**残差**と呼ばれる. **残差平方和**が最小となるように β_0, β_1 の推定値 $\hat{\beta}_0, \hat{\beta}_1$ を求めると,

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{5}$$

$$\hat{\beta}_1 = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{6}$$

となる.

4.6.1 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.

$$n = 108, \quad \overline{x} = 0.694, \quad \overline{y} = 557.407 \tag{7}$$

$$\sum_{i=1}^{n} x_i = 75.000, \quad \sum_{i=1}^{n} y_i = 60200, \quad \sum_{i=1}^{n} x_i^2 = 52.250, \quad \sum_{i=1}^{n} x_i y_i = 42124.95 \tag{8}$$

式 (7)〜(8) をそれぞれ式 (5), 式 (6) へ代入すると, $\hat{\beta}_1 = 1907.7$, $\hat{\beta}_0 = -767.39$ となり, 単回帰直線

$$\hat{y} = -767.39 + 1907.7x \tag{9}$$

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 11 のようになる. また, 寄与率を計算したところ $R^2 = 0.933$ となり, 式 (9) の回帰直線はデータに良く当てはまっていると言える.

第 II 部
本論

5　ピタゴラス勝率の最適化

ピタゴラス勝率はスポーツの分析に広く利用出来る式であるが, 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 1 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. そこで, 文献 [8](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.

$$\begin{cases} \omega_i & = \text{チーム } i \text{ のシーズン中の勝率} \\ \rho_i & = \dfrac{\text{チーム } i \text{ の失点数}}{\text{チーム } i \text{ の得点数}} \end{cases} \tag{10}$$

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（MAD）**は

$$\mathrm{MAD}(x) = \frac{1}{m}\sum_{i=1}^{m}\left|\omega_i - \frac{1}{1+\rho_i^x}\right| \tag{11}$$

となる.

式 (11) 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. ここでは 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

$$x^* = 1.69, \quad \mathrm{MAD}(x^*) = 0.0229$$

と評価し, 図 4 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 2011-2019 の NPB ペナントレースでチームごとに平均 2.29% しか外れていないことを意味する.

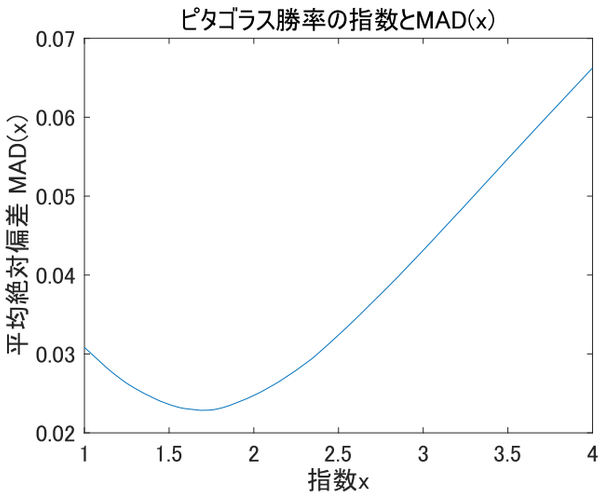

図 4　x に対する $\mathrm{MAD}(x)$ の関係

6　シミュレーション（仮想データの生成）

6.1　構成, サンプル番号の管理

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. （サンプル番号 : 1〜6000, 年度 : 1〜500）

6.2　シミュレーションの手順

1. 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
2. 上記の各組に対して **OPS**(=OBP+SLG) を計算する
3. 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
4. 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
5. **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
6. 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
7. 5 に誤差を加えたものを勝率とし, 6 とは別に順位を付ける

得点数と失点数については計算結果を整数値に丸める. また, 5 と 6 では同じチームでも勝率の値が異なるため, **順位の逆転が起こりうる**.

6.3　年間得点の再現

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．実データより，OBP と SLG の平均 μ と分散共分散行列 Σ は

$$\mu = \begin{pmatrix} 0.321 \\ 0.374 \end{pmatrix},\ \Sigma = \begin{pmatrix} 0.000177 & 0.000282 \\ 0.000282 & 0.000826 \end{pmatrix}$$

と計算され，これを基に図 5 の様な出力を得た．並行して OPS[*2]の計算も行い，4.6 で得られた次の単回帰式から得点数の推定値を求める．

$$y = -767.39 + 1907.7x \quad (x : \text{OPS},\ y : \text{得点数})$$

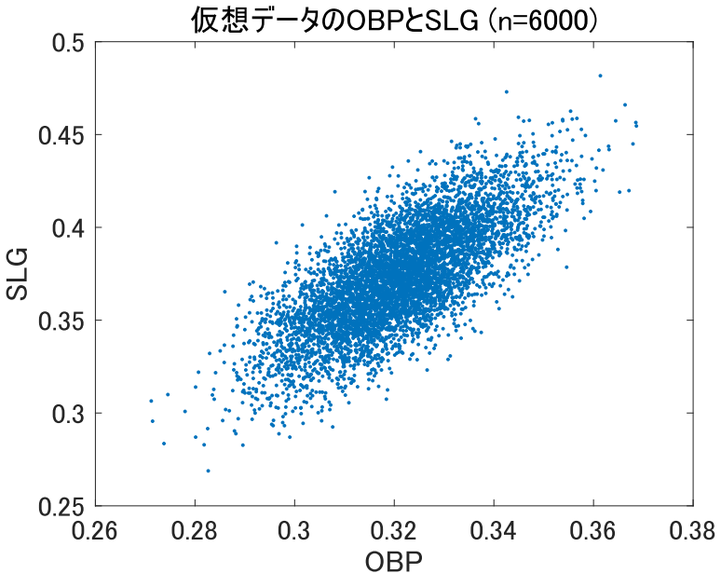

図 5　2 次元正規乱数による仮想データの OBP と SLG

6.4　年間失点の再現

6.4.1　失点数の生成と値の調節

実データによる失点数の平均 μ と分散 σ は $\mu = 557.407$，$\sigma = 76.21$ 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．

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，

$$\sum_{j=1}^{n}(\text{チーム } j \text{ の得点数}) = \sum_{j=1}^{n}(\text{チーム } j \text{ の失点数}) \tag{12}$$

という関係が成り立つ．NPB の場合，交流戦においてパ・リーグとセ・リーグの間に得失点差があるため

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

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，

$$(\text{12 球団の得点数の合計}) = (\text{12 球団の失点数の合計}) \tag{13}$$

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 のような仮想データが生成されたとする．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 (13) の条件は

$$\sum_{\substack{i=1,2 \\ 1\le j\le 6}} s_{ij} = \sum_{\substack{i=1,2 \\ 1\le j\le 6}} a_{ij} \tag{14}$$

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，

$$\alpha = \sum_{j=1}^{6}(s_{1j} - a_{1j}) = -\sum_{j=1}^{6}(s_{2j} - a_{2j}) \tag{15}$$

表 3　仮想データ（得点数と失点数）

リーグ	No.	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}
2	7	s_{21}	a'_{21}	a_{21}
	8	s_{22}	a'_{22}	a_{22}
	⋮	⋮	⋮	⋮
	12	s_{26}	a'_{26}	a_{26}

[*2] OPS = OBP + SLG

となる. このとき, a は

$$a_{in} = a'_{in} \times \left(\sum_{j=1}^{6} s_{ij} + (-1)^i \alpha \right) \div \left(\sum_{j=1}^{6} a'_{ij} \right) \qquad (16)$$

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.

6.4.2 失点数調整後も残る誤差の扱い

理論上は 6.4.1 での作業によって式 (13) 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.

例 6.1. 表 4 より,

$$(\text{リーグ 1 各球団の失点数}) = (\text{仮失点数}) \times \frac{3465 - 41}{3393.73}$$
$$(\text{リーグ 2 各球団の失点数}) = (\text{仮失点数}) \times \frac{4031 + 41}{3234.94}$$

6.5 勝率予測と順位付け

6.5.1 勝率補正なしの場合

得点数と失点数が確定した後は, ピタゴラス勝率の計算式を用いて勝率の推定値を求める. 最適化したピタゴラス勝率は Py(1.69) であるが, 今回は比較対象として Py(2) についても勝率を計算する. 勝率補正を加えない場合, 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.

表 4　年度 1 の仮失点と失点（交流戦のリーグ間得失点差 : 41）

リーグ 1				リーグ 2			
No.	得点	仮失点	失点	No.	得点	仮失点	失点
1	576	478.88	483	7	533	466.13	587
2	704	492.54	497	8	640	434.67	547
3	463	634.57	640	9	803	646.90	814
4	558	688.88	695	10	703	614.50	774
5	644	525.32	530	11	511	554.77	698
6	520	573.54	579	12	841	517.96	652
合計	3465	3393.73	3424	合計	4031	3234.94	4072

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

故に, ((ピタゴラス勝率) ∼ (勝率)) としている限りは指数 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.

6.5.2 勝率補正ありの場合

6.5.1 とは別のパターンとして, ((勝率) ∼ Py(1.69) + ERR(1.69)) として順位付けを行う. この場合の ERR(1.69) は $\mu = 0.000,\ \sigma = 0.0275$ による正規乱数から決定される. 勝率補正を加えない場合と異なり, ERR(1.69) の値によってはピタゴラス勝率で負けているチームよりも上の順位になる可能性がある.