OR メモランダム

このコラムは，ORにかかわる概念，知識(手法，原理)，それらの図解，よい教材や問題，実学ORの実施経験，そこから得られた智恵やアドバイス，失敗談と教訓，新しい視点，視座，フレームワーク，未だ解けていない問題，面白い研究テーマなどを，“新鮮に”，しかも“コンパクトに”表現し，提示していただくものです．ユニークなアイディア，フレッシュな見方，発想，だれかと意見をたたかわせたい問題提起など，ふるってご投稿ください．(原稿は，刷り上がり，半ページから3ページに納まるようにお書きください．簡単に！ 加筆訂正をお願いする場合があります)

Jリーグサッカー得点の統計分析

繁桝 算男，植野 真臣

1. はじめに

与えられたデータから情報を読みとる点において、ベイズ的アプローチは最もわかりやすい考え方であると考える。このことを、サッカー得点の統計分析という身近な例をとおして説明することが本稿の目的である。ベイズ統計的推論は、成書にその基本的な考え方が説明されているが（例えば、繁桝 1985、Bernardo and Smith, 1994）、ここでも簡単に説明する。データ・ベクトル$\boldsymbol{x}$の発生が母数ベクトル$\boldsymbol{\theta}$を所与として、$p(\boldsymbol{x}\mid\boldsymbol{\theta})$によってモデル化され、また、$\boldsymbol{\theta}$に関する事前の知識が事前分布$p(\boldsymbol{\theta})$によって表せるものとする（ここで、$p(\cdot)$は一般的に確率密度関数であることを示す)。このとき、$\boldsymbol{\theta}$の事後分布は、

$$\begin{aligned} p(\boldsymbol{\theta}\mid\boldsymbol{x}) &= p(\boldsymbol{x}\mid\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\boldsymbol{x}) \\ &\propto \text{尤度}\times\text{事前密度} \end{aligned}$$

によって表される。ここで、$p(\boldsymbol{x}\mid\boldsymbol{\theta})$は、$\boldsymbol{x}$が与えられた場合には、$\boldsymbol{\theta}$の値の信憑性を示すものであり、尤度と呼ばれる。また、将来のデータx^*が、同じモデルによって発生されるとき、そのx^*の分布は予測分布

$$p(x^*\mid\boldsymbol{x})=\int p(x^*\mid\boldsymbol{\theta})p(\boldsymbol{\theta}\mid\boldsymbol{x})d\boldsymbol{\theta}$$

によって示される。

ベイズ的アプローチでは、全ての推論と予測がこのように簡単に示されるし、しかも、nこれは便宜的な方法ではなく、人間の心理学的な推論の規範の位置を占めるものである。(もちろん、ベイズ的推論を規範とすることに反対する考えもある。この議論に関しては、例えば、Shigemasu and Yokoyama, 1994 参照。)

しげます かずお 東京工業大学システム科学専攻
〒227 横浜市緑区長津田町 4259
うえの まおみ 東京工業大学システム科学専攻
〒227 横浜市緑区長津田町 4259

2. モデルづくり

サッカーの得点のモデルをつくる。このステップは通常の伝統的アプローチにおいてもベイズ的アプローチにおいても変わりのない重要なポイントである。統計モデルは、母数の値以外は本当であるとみなして分析を進めることになるから、言うまでもなくこれは統計分析の良し悪しに直接影響する。サッカーの得点数は、技術もさることながら、いくつかの幸運が重ならないとなかなか得点にならない。この得点は、ポアソン分布によって表現できると考える。ポアソン分布に関して、一年間に馬で蹴られて死んだプロシア軍の兵士の数を模するという例は有名であるが、サッカー得点のほうが現代的であろう。

チームjのi番目の試合の得点をx_{ji}で表すとすると、x_{ij}の密度関数$p(x_{ji})$は、

$$p(x_{ji})=\frac{\mu_j^{x_{ji}}}{x_{ji}!}e^{-\mu_j} \tag{1}$$

となる。

3. 母数に関する推論

式 (1) の母数はμ_jのみである。よく知られているようにμ_jの最尤推定量は、n 個の観測値 $x_{ji},(i=1,\cdots,n)$ の平均$\overline{x_{j\cdot}}=\sum_i x_{ji}/n$である。また、この推定量が、何回もの繰り返し（Jリーグの試合が同じ条件で、何百回も行なわれるということを前提として）において、どのような分布（標本分布）をなすかを導くことも、比較的容易である。しかし、これから述べるように、１９９３年のJリーグの途中の時点でその後の得点を予測し、ひいては、勝敗を予測するには、ベイズ的アプローチが素直に推論を進めることができるのに対し、伝統的アプローチでは、ぎこちないやり方をとらざるを得なくなる。また、事前情報を利用して、全体的に推定誤差を小さくするにもベイズ的アプローチが自然である。ただし、ベイズ的アプローチにおいては、次節で示す事前分布の設定という作業が

必要である。事前分布を、事前の知識を考慮して主観的に定めることに対しては意見の分かれるところである。積極的に擁護する立場から言えば、当面する決定を最良のものとするには、事前の知識の全てをデータの情報を活用するのは当然と見るであろう。反対する立場から言えば、科学の道具としては、なるべく主観を排するのは当然といえよう。ここでは、統計学の文法としてベイズ的アプローチをとるが、事前分布を設定することはそのために払う代価とするぐらいが適当であろうと考える。

4. 事前分布を決める

ポアソン分布が、母集団のモデル分布であるとき、相性の良い事前分布は、ガンマ分布である。ここで相性が良いとは、尤度と事前分布が親和性を持ち、しかも、事後分布が事前分布と同じ分布族に属する自然共役事前分布のことをいう。すなわち、母数μ_jの事前分布は、

$$p(\mu_j)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\mu_j^{\alpha-1}e^{-\mu_j/\beta} \tag{2}$$

である。チームjに対して、特別な知識があるならば、事前分布の母数（超母数という）は、それを反映したものになるべきであるが、できるだけ先入観を避けて予想したいというならば、全てのチームに同一の事前分布 (2) を想定することができる。各μ_jが独立に、事前分布 (2) に従うとき、Jリーグ全体のチームの得点の同時分布 $p(\mu_1,\mu_2,\cdots,\mu_p \mid \alpha,\beta)$（１９９３年の場合、$p=10$）は、添字をどのように交換しても同じである。これは、交換可能性（exchangeability）と呼ばれる分布のもっとも単純な場合であり、各チームに関して特別な知識がないことを主張している（交換可能分布の一般形は、超母数の分布による、超母数を所与とした各母数の分布の積の混合である）。

さて、ここで問題になるのは、超母数αとβの決め方である。本来は、これも未知の値であり、母数と同様に超母数も統計モデルを構成するものとして推定すべきものである。このために、*ABIC*基準（よく知られているAIC情報量基準を積分することによって母数を消去した超母数の情報量基準）を最小化することによって超母数を推定し、それを所与として母数を推定するという2段階推定（Akaike, 1980）がある。また、筆者らは母数と超母数を平等に扱い、それぞれの周辺事後分布を数値的に求める方法を開発中である（Shigemasu and Nakamura, 1994）が、ここでは、わかりやすく説明するために次のような便宜的な方法を用いる。

後述するような事後分布を考察することによって、βは、事前の知識に対する重みに関連することがわかる。そこで、$\beta^{-1}=n_0$と置き、尤度と照合することによって、n_0は事前の仮想的な観測値の数であるとみなすことができる。このとき、もうひとつの超母数αは、$n_0\times$事前の平均 (μ_0) であると解釈できる。μ_0を、x_{ij}の全体平均

$$\overline{x_{\cdot\cdot}}=\frac{1}{np}\sum_i\sum_j x_{ij}$$

と等しいとおく。この方法は直観的にも納得できるであろう。なお、事前分布として自然共役分布が便利ではあるが、最近では、スタンダードな分布の混合によって、どのような形の事前分布も扱えるようにすべきだという意見も強力である。

5. 尤度および事後分布

$\boldsymbol{x}_j=(x_{j1},x_{j2},\cdots,x_{jn})^t$を得た後の$\mu_j$の尤度は、

$$\begin{aligned} p(\boldsymbol{x}_j \mid \mu_j) &= L(\mu_j \mid \boldsymbol{x}_j) \\ &= \prod_i^n \frac{\mu_j^{x_{ji}}}{x_{ji}!}e^{-\mu_j} \\ &\propto \mu_j^{n\overline{x_{j\cdot}}}e^{-n\mu_j} \end{aligned} \tag{3}$$

となる。式 (2) と式 (3) によって、

$$\begin{aligned} p(\mu_j \mid \boldsymbol{x}_j) &\propto \mu_j^{n\overline{x_{j\cdot}}+\alpha-1}e^{-(n+\beta^{-1})\mu_j} \\ &= \mu_j^{n\overline{x_{j\cdot}}+n_0\mu_0-1}e^{-(n+n_0)\mu_j} \end{aligned} \tag{4}$$

となる。すなわち、μ_jの事後分布は、母数 $(n\overline{x_{j\cdot}}+n_0\mu_0,(n+n_0)^{-1})$ を持つガンマ分布であることがわかる。ベイズ推定値の一つEAP推定値 (μ_j^*) は、分布 (4) の期待値であり、

$$\mu_j^*=\frac{n\overline{x_{j\cdot}}+n_0\mu_0}{n+n_0} \tag{5}$$

となる。推定値 (5) は、実際の観測値$\overline{x_{j\cdot}}$と、事前平均μ_0、あるいは、全体平均$\overline{x_{\cdot\cdot}}$との加重平均であることがわかる。これは、各チームの得点を全体の平均へ近づけるという一種の縮約推定値 (shrinkage estimate) である。サッカーではなく、プロ野球でも、シーズン当初に4割の打率を記録する者も、結局は3割程度になったり、2割に低迷するバッターも実力があれば、シーズンが終ってみれば、3割近くに戻っているということがある。Effron and Morris (1975) は、このよ

表 1: 各チームにおけるμの推定値と自乗誤差

j	チーム名	$\hat{\mu}_j$	μ_j^*	e_{ML}	e_{BAYES}
1	鹿島アントラーズ	2.500	2.044	52.750	39.676
2	ジェフ市原	1.000	1.294	61.000	55.685
3	浦和レッズ	0.700	1.144	25.230	21.586
4	ヴェルディ川崎	1.500	1.544	68.750	67.337
5	横浜マリノス	1.800	1.694	52.280	51.188
6	横浜フリューゲルス	1.200	1.394	24.080	25.256
7	清水エスパルス	1.400	1.494	39.120	38.756
8	名古屋グランパス	1.600	1.594	73.120	73.085
9	ガンバ大阪	1.400	1.494	42.920	42.745
10	サンフレッチェ広島	1.200	1.394	62.680	59.579
	合計			508.89	497.93

うな野球の打率の例を用いて縮約推定値の利点を示している（縮約推定値そのものは、伝統的な統計学でも各種提案されている）。

ベイズ的に素直に導かれた推定値（BAYES）(5)を、最尤推定値（ML）$\hat{\mu}_j = \overline{x_{j\cdot}}$と比較してみよう。

１９９３年のＪリーグの成績を、最初の試合の1/4（すなわち、ステージ１の前半戦）が終った時点で、その得点平均μ_jを推定し、残りの3/4の試合の得点を予測することを考えよう。このとき、残りの3/4の試合において、自乗誤差を次のように計算する。

$$e_{ML} = \sum_{i=10}^{36} (\hat{\mu}_j - x_{ji})^2$$

$$e_{BAYES} = \sum_{i=10}^{36} (\mu_j^* - x_{ji})^2$$

表１に、$\hat{\mu}_j$、μ_j^*、および、e_{ML}、e_{BAYES}を示す。

ベイズ推定値の方が推定の自乗誤差の観点に立てば、少なくとも単純な最尤推定値よりは優れていることがわかる。しかし、ベイズ的アプローチの利点はこれだけではない。次節では、得点x_{ji}そのものの予測、及び、勝敗の予想を試みる。

6. 得点の予測

得点x_{ji}の予測分布は、ベイズ的アプローチによればごく自然に求まる。すなわち、観測された既知のx_{ji}'s に基づき、将来の試合*における得点の分布は、

$$p(x_j^* \mid \boldsymbol{x}_j) = \int p(x_j^* \mid \mu_j) p(\mu_j \mid \boldsymbol{x}_j) d\mu_j \qquad (6)$$

によって求まる。式（６）は比較的簡単に求まり、

$$p(x_j^* \mid \boldsymbol{x}_j) = \frac{\Gamma(x + \alpha_j^*) \left(\frac{\beta_j^*}{\beta_j^*+1}\right)^{x+\alpha_j^*}}{\Gamma(\alpha_j^*) \beta_j^{*\alpha_j^*} x_j^{*}!} \qquad (7)$$

となる。ここで、$\alpha_j^* = n\overline{x_{j\cdot}} + n_0\mu_0$、$\beta_j^* = (n+n_0)^{-1}$である。もし、$\alpha^*$、$\beta^*$が整数であれば、この密度関数（厳密には確率量関数）は、負二項分布を表す。

すなわち、

$$p(x_j^* \mid \boldsymbol{x}_j) = \binom{\alpha_j^* + x_j^* - 1}{x_j^*} \left(\frac{\beta_j^*}{\beta_j^*+1}\right)^{x_j^*} \left(\frac{1}{\beta_j^*+1}\right)^{\alpha_j^*} \qquad (8)$$

日程の1/4を経過した時点での将来の得点x_j^*は負二項分布である。各チームの得点の残り3/4の得点は、表２のような分布に従う（簡単のため、０点から８点及び９点以上の得点の分布を示した）。 表２を利用し、対戦チームの得点を独立とすれば、勝敗の確率を予想できる。例えば、チーム5(フリューゲルス)のチーム９（ガンバ大阪）に対する勝敗予想は、２つのチームの０点から８点及び９点以上の得点の組合せより、勝つ確率、引きわけの確率、負ける確率は、それぞれ、

$$p(x_5^* > x_9^* \mid \boldsymbol{x}_5, \boldsymbol{x}_9) = \sum_{x_5^* > x_9^*} p(x_5^* \mid \boldsymbol{x}_5) p(x_9^* \mid \boldsymbol{x}_9) = 0.4369$$

$$p(x_5^* = x_9^* \mid \boldsymbol{x}_5, \boldsymbol{x}_9) = \sum_{x_5^* x_9^* = 0}^{9} p(x_5^* \mid \boldsymbol{x}_5) p(x_9^* \mid \boldsymbol{x}_9) = 0.1443$$

と計算できる。

表 2: 各チームの得点分布

j/得点	0	1	2	3	4	5	6	7	8	9
1	0.268805	0.242902	0.188094	0.128725	0.066970	0.028236	0.010048	0.003104	0.000849	0.006058
2	0.458496	0.254015	0.160328	0.070696	0.023882	0.006584	0.001542	0.000316	0.000058	0.024075
3	0.511608	0.254068	0.148859	0.038625	0.017915	0.004454	0.000943	0.000175	0.000029	0.000260
4	0.358296	0.235445	0.177210	0.092521	0.036882	0.011962	0.003287	0.000787	0.000167	0.083410
5	0.309020	0.274306	0.236997	0.107746	0.047320	0.016882	0.005095	0.001338	0.000312	0.000920
6	0.415430	0.249193	0.170738	0.081445	0.029715	0.008834	0.002228	0.000490	0.000096	0.041814
7	0.376410	0.238723	0.173232	0.087321	0.033629	0.010543	0.002802	0.000649	0.000134	0.076533
8	0.341055	0.231929	0.180826	0.097670	0.040254	0.013491	0.003828	0.000946	0.000208	0.089751
9	0.376410	0.238723	0.173232	0.087321	0.033629	0.010543	0.002802	0.000649	0.000134	0.076533
10	0.415430	0.249193	0.170738	0.081445	0.029715	0.008834	0.002228	0.000490	0.000096	0.041814

表 3: 各チームの勝敗確率

j/j'	1	2	3	4	5	6	7	8	9	10
1	0.0000	0.7581	0.8296	0.6429	0.5778	0.7112	0.6653	0.6208	0.6653	0.7112
2	0.2419	0.0000	0.4505	0.2731	0.2251	0.3303	0.2910	0.2561	0.2910	0.3303
3	0.1704	0.5495	0.0000	0.1882	0.1512	0.2337	0.2023	0.1750	0.2023	0.2337
4	0.3571	0.7269	0.8118	0.0000	0.3499	0.4780	0.4316	0.3889	0.4316	0.4780
5	0.4222	0.7749	0.8488	0.6501	0.0000	0.5569	0.5091	0.4643	0.5091	0.5569
6	0.2888	0.6697	0.7663	0.5220	0.4431	0.0000	0.3487	0.3100	0.3487	0.3916
7	0.3347	0.7090	0.7977	0.5684	0.4909	0.6513	0.0000	0.3630	0.4045	0.4500
8	0.3792	0.7439	0.8250	0.6111	0.5357	0.6900	0.6370	0.0000	0.4580	0.5051
9	0.3347	0.7090	0.7977	0.5684	0.4909	0.6513	0.5955	0.5420	0.0000	0.4500
10	0.2888	0.6697	0.7663	0.5220	0.4431	0.6084	0.5500	0.4949	0.5500	0.0000

ただし、ご存知のように、Jリーグでは引き分けがないので、引き分けを勝ちか敗けに比例配分すると、勝つ確率は、0.5091 となる。実際の試合ではフリューゲルスの勝ちであった。このようなやり方で、残りの 3/4 試合の全体の勝率の予測、及びその予想確率と既に終った試合の勝率を統合した勝率を表 3 に示す。表中の行は各チーム jを示し、列はその対戦相手（ここでは j'と表した）を示している。この確率に基づき、最終的な勝ち数のベイズ的予測と、$\hat{\mu}_j = \overline{x_{j\cdot}}$ を代入した場合の勝ち数の予測を表 4 に示した（( ) 内は予測の自乗誤差)。ここでも、ベイズ的予測の方が優れていることがわかる。

実際の試合は、もちろん、2 つの戦うチームの得点が独立に得られるはずもなく、攻撃と防御の交互作用が常に存在するが、ここまで統計モデルを細かくすることは難しい。現実に合うために統計モデルを複雑化することは必要であろうが、それにも限度があり、実際の予想には、統計的推論に対し常識的な判断を的確に加える必要がある。そのためにも、統計的推論は、我々自身が日常的に行なう確率判断に近く、かつ、わかりやすい推論方式が望ましい。ベイズ的アプローチは、人間の心理的確率判断の規範形であり、推定、検定、予想の考え方が全てわかりやすく、我々の常識的判断との相性が良いと考えられる。

表 4: 各チームの予測勝ち数、実際の勝ち数、自乗誤差

j	実データ	BAYES	ML
1	16	13.98(4.08)	18.00(4.00)
2	9	8.02(0.96)	12.31(10.96)
3	5	6.01(1.02)	10.73(32.83)
4	20	10.86(83.54)	14.03(35.64)
5	14	13.63(0.14)	15.41(1.99)
6	11	10.26(0.55)	12.46(2.13)
7	18	16.37(2.66)	13.18(23.23)
8	8	9.63(2.66)	14.10(37.21)
9	12	11.73(0.07)	13.01(1.02)
10	13	10.11(8.35)	11.71(1.66)
自乗誤差		104.029	150.68

参考文献

[1] Akaike, H. (1980) "Likelihood and the Bayes procedure", Bernardo,J.M. et al Eds. Bayesian statistics, University Press, 143-166.

[2] Bernardo, J.M. and Smith, A.F.M. (1994) "Bayesian Theory", Wiley & Sons, Inc.

[3] Effron, B. and Morris, C. "Data Analysis using stein's estimator and it's generalization, JASA, 70(350), 311-319.

[4] 繁桝算男 (1985)「ベイズ統計入門」、東大出版会.

[5] Shigemasu, K.and Yokoyama, A. (1994) "Flexible Bayesian approach for psychological modeling of decision making", Japanese Psychological Reserch, 36(1), 20-28.

[6] Shigemasu, K. and Nakamura, T. (1994) "Bayesian marginal inference in Item Response Model using the Gibbs sampler, presented at Fifth International Meeting on Bayesian Statistics.