

多数決とスケール不変性

競馬ファンは“インテリジェント”？ それとも“コピーキャット”？

Majority Voting and Scale Invariance

Horse-racing fans are “intelligent” or “copycat” ?

守 真太郎
Mori Shintaro

北里大学 理学部 物理学科
Department of Physics, School of Science, Kitasato University
mori@sci.kitasato-u.ac.jp, <http://sharaku.sci.kitasato-u.ac.jp/mori/>

久門 正人
Hisakado Masato

スタンダード & プアーズ
Standard and Poors
masato_hisakado@standardandpoors.com

キーワード: voting, accuracy ratio, discriminative power, scale invariance, horse-racing, copycat

概要

投票結果がリアルタイムに公表される投票システムの候補者の選別能力および候補者の混合が示すスケール不変性について報告する。候補者は優秀、劣等と二値に分類される。それについては投票者は知りうる事が出来ない。投票者はその判定材料をスコアを通して知る。投票者はスコアに比例する確率で順に投票を行い、投票結果によってスコアは随時更新される。投票開始直後は、初期の二値のスコアで投票を行うので、投票者は「インテリジェント」に振る舞うが、投票が進むにつれて他の投票者の投票結果に左右される「コピーキャット」として振る舞うようになる。「インテリジェント」か「コピーキャット」かの強弱はパラメータで調節される。我々は、このモデルを分岐過程のモデルに変換し、スケールされた投票数がガンマ分布に従うことをもとに、二値の候補者の混合がスケール不変性を持つこと、投票システムの候補者の選別能力が不完全ベータ関数で与えられることを示した。また、JRAの競馬単勝市場の得票率のデータを用い、さまざまなパターンの勝馬、負け馬の組合せに対して、スケール不変性の成立と投票モデルの検証を行い、競馬単勝市場の参加者は「コピーキャット」であることを明らかにした。

1. 多数決の選別能力

ウィンストン・チャーチルの有名な言葉に『実際のところ、民主制は最悪の政治形態と言うことが出来る。これまでに試みられてきた、他のあらゆる政治形態を除けば、だが』というものがある [1]。これは、民主制が他の政治体制に打ち勝ってきた、もっとも優れた政治体制であることを彼一流の皮肉で表現したものである。その民主制の根幹をなすのが多数決であり、その透明性、公平性の確保は民主制の正統性に直結する重要事項であり、民主主義国家ではさまざまな選挙管理システムが運営されている。では、投票結果がごまかされないという透明性や選挙権がほぼ全てのひとに与えられるという公平性が確保されたとき、多数決はどの程度「正しく」機能するのだろうか？ここで、「正しい」というのは、優れた候補者と劣等な候補者を正しく区別するという意味であり、多数決の選別能力の客観的な評価を問題とする。その選別能力は投票行動のどのようなファクターで決定、影響されるのか？そこで、簡単な投票モデルで多数決の選別能力を問うてみることにする。投票者のコピーキャットの

な振る舞いが候補者の混合にスケール不変性を生み出すことを示す。競馬単勝市場での得票率のデータ解析を行い、スケール不変性や選別能力に関するモデルの結果を検証し、競馬ファンはコピーキャットであることを示す。この結果は、投票結果がリアルタイムに公表される場合、「多数決」が間違ってしまうことを意味する [2]。最後のまとめでは、インテリジェントな投票者ばかりの選別能力の高い白黒ハッキリさせるシステムよりも、コピーキャットが存在し多数決に「あいまい=スケール不変」さを持たせたシステムのほうがむしろ安全なのではという指摘を行う。

2. 投票モデル

2種類の候補者、“優秀な”と“劣等”な候補者がいるとする。これは将来の予想も含まれるので投票者は当然知りえない。候補者自身はある程度知りうるが、完璧に知ることは出来ない。候補者が知りうる場合は経済学では“情報の非対称性の問題”として扱われる。ここではモデル化にあたり便宜上決める。

一方投票者はこれらの候補者について全く無知であるわけではない。ある程度の情報は把握していると考え。これをモデル化するために、優秀な候補者はスコア s を持ち、劣等な候補者はスコア w を持つとする。概ね事前の情報が正しければ $s > w$ であり、全く情報が誤っている場合は $w < s$ 、 $s = w$ 場合は投票者は候補者の優劣について全く情報を持たない場合に対応する。以下では基本的に情報は概ね正確と仮定して議論する。情報がまったく誤っている場合については最終節のまとめでコメントする。

「インテリジェントな投票者」、すなわち情報を分析し独自に判断する投票者はこのスコアに基づいて投票するとする。ここで投票に確率的な要素を導入する。劣等な候補者が票を得る確率は $w/(s+w)$ 、優秀な候補者が得票する確率は $s/(s+w)$ とする。これは投票は情報のスコアに比例して行われるということである。この場合投票数は二項分布にしたがうので、投票者が少ない場合やスコアの差が小さい場合は、劣等な候補者が優秀な候補者より多くの票を集めることも起こるかも知れないが、投票者の数が十分大きくなると $s > w$ で正しい情報を持っている限りスコアの差 $s - w$ がどんなに小さくても必ず優秀な候補者がより多くの票を集め、選挙ならば当選することになる。つまり、この場合は優秀な候補者を確実に選別するので選別能力が 100%ということになる。投票者が全員「インテリジェント」で、投票行動が「正しい情報」に基づいて行われるならば、多少確率的な要素が入ってきたとしてもシステムは完全な選別能力を持つわけである。

投票は「正しい情報」のみに基づいて「インテリジェント」に行われるわけではない。そのため劣等な候補者も当選する。事前の情報が「正しい情報」でない場合、 $w \geq s$ も考えられる。しかしそれ以外の情報も投票行動に影響する。例えば、選挙などでは、選挙公示日までに各種世論調査が行われ、それがマスコミに発表され、人々は世論調査に協力した人々がどのように考えているのかの情報を接することになる。そうした他の投票者の考えが投票の判断に影響を全く与えないというのは考えられない。また、競馬などでは、強い馬は多くの馬券が売れ、それに伴ってオッズが下がるが、逆にオッズの下がった馬は、これから馬券を買うひとにとってみれば他の人々が強いと考えている馬であり、客観的に強いのだと判断して馬券を購入する可能性がある。このように他の投票者が投票するような候補者は優秀に違いないから投票する、とか、他の馬券購入者がたくさん馬券を買う馬は強いに違いないから馬券を買うという「コピーキャット」な投票者も存在すると考えられる。このような投票者を考慮した次のようなモデルを考えてみる。各候補者は、スコアと得票数の和に比例した相対確率で票を獲得するというモデルである [3]。これは、「Rich-get-Richer」(富めるものはますます富む)メカニズムとしてよく知られた

モデル化であり [4]、複雑ネットワークで次数分布が巾乗則に従うことを説明するのにバラバシらが導入したモデルでも採用されている [5]。

与えられたスコアを「種」(= 初期値) とし最初の投票者はスコアに比例した確率で各候補者に 1 票の投票を行う。その結果を踏まえ 2 番目の投票者は更新されたスコアに従い 1 票の投票を行う。たとえば候補者が 2 人で優秀な候補者と劣等な候補者とする。最初のスコアは優秀な候補者は s 、劣等な候補者は w である。1 番目の候補者が劣等な候補者に投票した場合スコアは優秀な候補者が s 、劣等な候補者が $w+1$ となり。2 番目の投票者は優秀な候補者へ確率 $s/(s+w+1)$ 、劣等な候補者へ確率 $(w+1)/(s+w+1)$ で投票する。このように順番に投票し時間発展するモデルである。

最初のスコアの水準、 s や w が大きい場合、最初の「種」のスコアの影響が大きいため、投票者は「種」となるスコアに基づいて投票することから上記でいう自主的な考察に基づく投票、「インテリジェント」となる。一方、最初スコアの水準が非常に小さい場合、最初の「種」の情報はすぐに投票に見えなくなってしまう。ある投票者が投票しようとする、その前の投票者による影響は非常に大きい。この場合、投票者は他者追従的、「コピーキャット」な行動をとる。

このように最初の「種」のスコアで水準の 2 つの極限ですべて「インテリジェント」になったりすべて「コピーキャット」になったりする。もうひとつのパラメータである s と w の比は「情報の正しさの強度」をあらわす。 w/s が小さいほど候補者に付随する情報は正確である。

十分長い投票時間の後、各候補者の得票数は適当にスケールするとガンマ分布に従うことが示される。スコア $\mu \in \{s, w\}$ の候補者のスケールされた投票数 u_μ の従う確率密度関数は

$$p_\mu(u) = \frac{1}{\Gamma(\mu)} e^{-u} \cdot u^{\mu-1} \quad (1)$$

となる。導出は付録の解説を参照のこと。この確率分布を $s = 2$ 、 $w = 1$ の場合にプロットしたのが図 1 である。

得票数の分布関数から、優秀な候補者のほうが劣等な候補者よりも平均して得票数が多いことが分かる。これは、ガンマ分布に従う確率変数の期待値は、 $\langle u_\mu \rangle = \mu$ となることから分かる。しかし、優秀な候補者が必ず劣等な候補者よりも多くの票を集めるわけではない。実際、グラフからも票数が少ないところで優秀な候補者の確率分布はゼロではなく、場合によっては劣等な候補者よりも少ない票数しか獲得できないこともあるし、逆に票数が多いところでも、劣等な候補者の確率分布がゼロではないので、優秀な候補者よりもより多くの票を集めることがあることが分かる。優秀な候補者が劣等な候補者に得票数で勝つ確率を計算すると、それは得票数の比

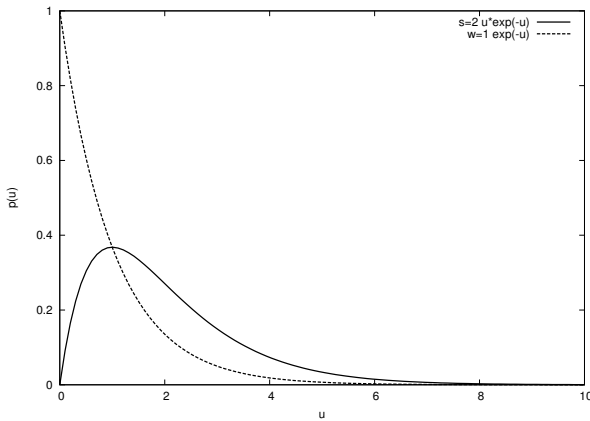


図1 スケールされた得票数が従うガンマ分布をプロットした。実線が $s = 2$ の優秀な候補者の得票数の分布関数、点線が $w = 1$ の劣等な候補者の得票数の分布関数である。

率 $u_s / (u_w + u_s)$ がベータ分布 $B(s, w)$ に従うことから、

$$\text{Prob.}(u_s > u_w) = B\left(\frac{1}{2}; s, w\right)$$

と不完全ベータ関数 [7] で与えられることが分かる。たとえば、 $s = 2, w = 1$ の場合、

$$\text{Prob.}(u_s > u_w) = \frac{3}{4}$$

となり、75%の確率で優秀な候補者の得票率が劣等な候補者を上回ることがわかる。

このように、他の候補者の投票によって決断を下す投票者が投票を行うシステムでは、最初に正しい情報をもっていたとしても、他の投票者の行動で正確な情報がぼやけてしまい、選別能力は100%から下がってしまうことが分かる。この選別能力 $\text{Prob}(u_s > u_w)$ は、AR(Accuracy Ratio) と呼ばれるものと密接に関連する。定義は、

$$\text{AR} = 2 \cdot \left(\text{Prob}(u_s > u_w) - \frac{1}{2}\right)$$

であり、識別能力がゼロの $\text{Prob}(u_s > u_w) = \frac{1}{2}$ の場合にゼロ、識別能力が100%の $\text{Prob}(u_s > u_w) = 1$ の場合に1となるように規格化されている。さきほど述べた $s = 2, w = 1$ の場合のARは50%である。このARという指標は、もともと順位付けされた対象がある性質を持つか、とか、ある事象が実現するかといった場合に、序列の精度*1を数値化したものである。ランダムな状態からの序列の完成度をあらわしている。

まとめると、投票システムが完全な選別能力を持つには、投票者が正確な情報に基づいて、他の投票者の動向に左右されない「インテリジェント」な場合である。これは、 s, w の絶対値が大きく、一票の価値が小さい場合である。その場合、投票行動は他の投票者に行動には影響されず、概ね正しい判断を下す。逆に、 s, w の絶対値

が小さく、一票の価値が大きい場合は、投票者は他の投票者の動向に左右されて「コピーキャット」になり、投票システムは選別能力が低下し、場合によっては劣等な候補者を選びだしてしまうことになる。

では、もうすこし候補者の得票数の振る舞いと選別能力の関係を見ていこう。今度は、二人ではなく多数の候補者がいる場合に、投票によってどのような性質が見えてくるのかに注目する。

3. 候補者の混合が示すスケール不変性

前セクションでは、優秀、劣等の二人の候補者に対する投票モデルを考え、スコア s, w に選別能力がどのように依存するのかを議論した。ここでは、多数の候補者が優秀、劣等の二値に分類され、前セクションと同様にスコアと累積得票数に比例する確率で次の票を得票できる場合を考える。優秀な候補者の数を N_s 、劣等な候補者の数を N_w とし、全体の候補者数 $N = N_s + N_w$ に対する割合を $r_s = N_s/N, r_w = N_w/N$ と書くものとする。そして、 N_s, N_w が十分大きい場合を考える。その場合、スケールされた得票数 u 上で N 人の候補者がそれぞれガンマ分布 $p_\mu(u)$ に従って分布することになる。前セクションと同様な議論から、得票数の大きな部分には、優秀な候補者の割合が大きい、劣等な候補者の割合もゼロではない。また、得票数の小さなところには、劣等な候補者の割合が大きい、優秀な候補者も存在する。そして、こうした二値の候補者の混合は、当然ながらスコアの差 $s - w$ が小さく、比率 $\alpha = s/w$ が1に近いと激しく、逆に差が大きかったり、比率 α が大きい場合は混合は起きず分離する。

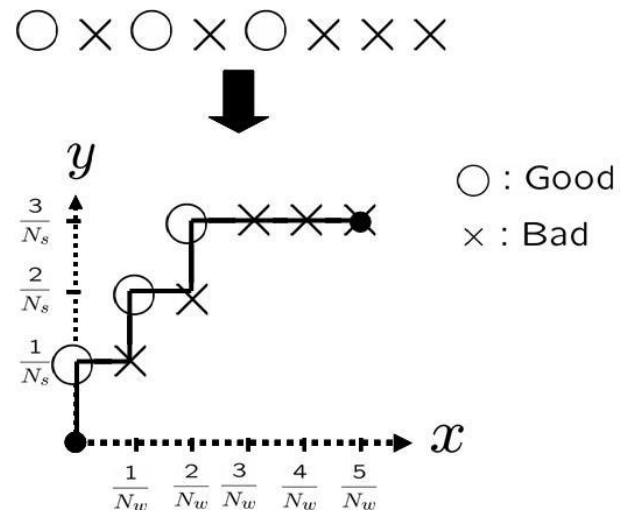


図2 ROCカーブの例。 $N_s = 3, N_w = 5$ で、得票数の多い順に並べたとき、優秀な候補者 (○) と劣等な候補者 (×) が図の上部に描かれた序列を作った場合のROCカーブを描いたもの。

二値の候補者の混合はどのようにになっているのだろうか

*1 クレジットリスクの分野なら財務諸表をもとに、債務者にスコアを付与して順序を決め、その順序通りに「デフォルト(債務不履行)」したかどうかを評価するのに用いる。

か?そこで、前セクションで導入した AR の幾何学的な表示と絡めて説明する。まず、得票数の大きさ順に候補者を並べ、その順番で二次元平面上の経路 $(x_i, y_i)_{i=0, \dots, N}$ を描く。経路の出発点は原点 $(x_0, y_0) = (0, 0)$ で、候補者が優秀なら y 軸方向に $1/N_s$ 進み、劣等なら x 軸方向に $1/N_w$ 進み、 N 番目のステップで経路は $(x_N, y_N) = (1, 1)$ に到達する。順番が若い得票数が大きい候補者は優秀である確率が高いので、経路は y 軸方向に進む確率が高いが、そのうち x 軸方向に進む確率も増加し、順番が最後のほうでは、得票数も低く、劣等な候補者である可能性が高いので、 x 軸方向に進む確率が大きくなる。このような、順序に対して描いた経路のことを ROC(Receiver Operating Characteristic) カーブと呼び、その形は上に凸になる [6]。AR は、このロックカーブ $(x_i, y_i)_{i=0, \dots, N}$ の情報を用いて、

$$\text{AR} = 2 \cdot \left(\sum_{i=1}^N y_i \cdot (x_i - x_{i-1}) - \frac{1}{2} \right) \quad (2)$$

と計算でき、AR は ROC カーブのジニ係数であることが分かる。

投票モデルにおいて ROC カーブはどのような形となるのか? また、それは二値の候補者の混合とどのような関係にあるのか? r_s, r_w を固定して N_s, N_w を無限大にしたとき、ROC カーブは滑らかな曲線となり、その媒介変数表示は次のようになる。

$$x(t) = \int_t^\infty p_w(u) du \quad (3)$$

$$y(t) = \int_t^\infty p_s(u) du \quad (4)$$

$t=0$ で $(x(0), y(0)) = (1, 1)$ 、 $t=\infty$ で $(x(\infty), y(\infty)) = (0, 0)$ を通ることが分かる。特に、 t がゼロに近い $(x, y) = (1, 1)$ の近傍では、得票数が低い二値の候補者の累積 $1-x$ と $1-y$ は次の関係式を満たすことが分かる。

$$1-y \propto (1-x)^\alpha \quad \text{with} \quad \alpha \equiv \frac{s}{w} \quad (5)$$

さらに興味深いのは、 $\alpha = s/w$ を固定したまま $s, w \rightarrow 0$ の極限をとった場合である。このとき、累積 $1-x$ と $1-y$ は $0 \leq x, y \leq 1$ の全領域で巾乗則に従う [3]。

$$1-y = (1-x)^\alpha \quad \text{with} \quad \alpha \equiv \frac{s}{w} \quad (6)$$

この極限で AR は、 $\text{AR}(\alpha) \equiv \frac{\alpha-1}{\alpha+1}$ となる。

通常、巾乗則は式 (5) のように、十分小さな得票数 u といった領域で成立するものだが [4]、この投票モデルでは $s, w \rightarrow 0$ の極限で全域 $0 \leq x, y \leq 1$ で成立する。では、この巾乗則は何を意味するのか、また、 $s, w \rightarrow 0$ の極限で巾乗則が全域で成立するのは何故なのか? まず、この巾乗則の意味だが、これは得票数が小さくなると通常、優秀な候補者の存在割合が減少し、場合によってはいなくなってしまう。その場合、ROC カーブの両対数プロットは直線にはならず、 x 座標に向かって落ちる。投

票モデルでも、確率分布関数 $p_\mu(u)$ の振る舞いは、優秀な候補者が減少することを示している。しかし、巾乗則が成立するということは、どんなに得票数が少なくなっても、優秀な候補者が存在するというスケール不変性が成立することを教えてくれる。実際、劣等な候補者の累積 $1-x$ に対する優秀な候補者の密度 ρ を計算すると、

$$\rho = \frac{d(1-y)}{d(1-x)} \propto (1-x)^{\alpha-1}$$

となることから分かる。つまり、 $1-x$ がどんなに小さくなくても、優秀な候補者の密度 ρ は有限であり、必ず存在することがわかる。

一方、なぜ投票モデルが $s, w \rightarrow 0$ の極限で全領域で厳密にスケール不変となったのかは、以下のように解釈できる。一言でいえば、投票による混合の問題が相対確率 μ でのランダムな選択の問題に帰着したからである。まず、最初の投票者の行動と、その結果について考える。彼が投じた一票を候補者が得る相対確率は μ である。票を得た候補者の相対確率は $\mu+1$ になり、他の候補者は μ のままである。いま、 $\mu \rightarrow 0$ の極限を考えているので、得票による相対確率の増加は圧倒的で、次の投票からは他の候補者を圧倒し、得票数が一気に増加する。最初に選ばれた候補者が他の候補者に得票数で負けることはなく、最終的な得票数の順位は 1 位になることが分かる。そこで、この候補者を除いた $N-1$ 人で投票を続けることにすると、ここでも最初に選ばれた候補者が残りの $N-2$ 人の候補者を圧倒し、最終的な順位は二番となる。このように、 $\mu \rightarrow 0$ の極限では一票の価値が圧倒的に大きくなり、投票者は完全なコピーキャットになり、ある候補者が最初に票を得ると、その候補者に投票が集中し、順位が決定するの繰り返しとなる。各候補者が最初に票を得る確率は μ なので、二値の候補者の累積 (x, y) の従う微分方程式は、

$$dx = (1-x)w \quad \text{および} \quad dy = (1-y)s$$

となり、これを解くことで式 (6) が得られる。

まとめると、 s, w のスケールが小さく、一票の重みが大きい場合、投票者はコピーキャットになり、結果として二値の候補者の混合は非常に広い範囲でスケール不変性を示す。逆に、 s, w のスケールが大きく、一票の重みが小さい場合は、投票者はインテリジェントで、概ね正しい判断を下し、二値の候補者の混合はあまり起きない。また、混合がスケール不変になる領域も狭くなる。では、実際の投票システムで、このようなスケール不変性は確認できるのか? 次のセクションでは、日本の競馬単勝市場の得票率データをもとに、スケール不変性や投票システムの選別能力、AR について解析を行う。

4. 競馬単勝市場のデータ解析とモデルの検証

本セクションでは、JRA の単勝馬券の得票率データをもとに、投票モデルの結果の検証を行う。得票数では

なく得票率のデータを選んだ理由は、投票時間がレースによって異なる点と投票モデルの結果がスケールされた得票数に関するものなので、各レース毎の得票率のほうが都合がよいからである。用いたデータは、1986年から2006年までの、901366頭分の単勝馬券の得票率とし、レースでの最終順位 k で分類する。着順 k に対する馬券の数を N_k 、得票率を $\{v_{k,i}\}_{i=1,\dots,N_k}$ 、平均得票率を $[v_k]$ で表す。用いたデータの着順は $1 \leq k \leq 11$ に制限した。着順が下がりすぎると、1着などの場合と比べ出走頭数の差が大きくなりすぎるからである。二値の優秀、劣等な候補者として、 $1 \leq i < j \leq 11$ の任意の着順のペア (i, j) を考え、 i 着の馬が優秀な候補者、 j 着の馬が劣等な候補者とする。対応する得票率の集合の組、 $(\{v_{i,k}\}_{k=1,\dots,N_i}, \{v_{j,l}\}_{l=1,\dots,N_j})$ に対し、前セクションで解説したモデルの結果を検証する。

4.1 スケール不変性と指数 α

二値の候補者の混合を調べるために、得票率の集合の組 $(\{v_{i,k}\}_{k=1,\dots,N_i}, \{v_{j,l}\}_{l=1,\dots,N_j})$ を大きさの順番で並べ替え、ROC カーブ $(1-x_k, 1-y_k)_{k=0,\dots,N_i+N_j}$ を描き、その原点近傍での両対数プロットの傾きから指数 α を評価する。

$$\log(1-y_k) \propto \alpha \cdot \log(1-x_k)$$

図3に(1,2)着ペア、(1,11)着ペアに対するROCカーブの両対数プロットを示した。グラフには、 $y = a \cdot x^b$ の関数形を用い、 $0 \leq x \leq 0.1$ の領域でフィットした結果も示してある。グラフから直線領域のレンジが広いことが

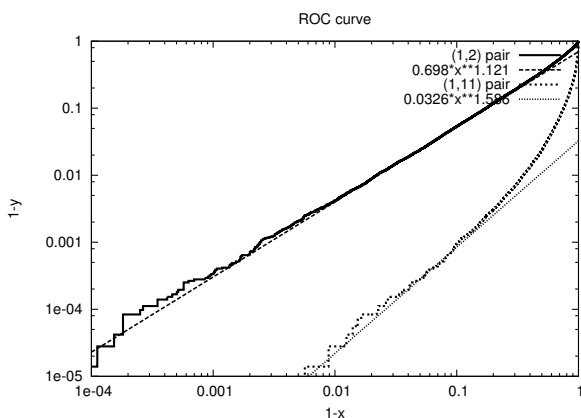


図3 (1,2)着ペア、(1,11)着ペアのROCカーブ (x, y) を両対数プロットした。 $y = a \cdot x^b$ で $[0.0, 0.1]$ の領域でフィットした結果も表示してある。

分かる。(1,2)着ペアでは、 $10^{-1} \leq 1-y \leq 10^{-5}$ の4桁のレンジで、(1,11)着ペアでも $10^{-3} \leq 1-y \leq 10^{-5}$ の二桁のレンジでスケール不変性を持つことが分かる。また、 $y = a \cdot x^b$ とのフィットから、指数 α を評価することができる。(1,2)着ペアでは $\alpha = 1.121$ 、(1,11)着ペアでは $\alpha = 1.586$ となった。

4.2 ARの振る舞い

次に、単勝市場の選別能力ARを評価する。前サブセクションで用いたROCカーブから式(2)によりARを計算すればよい。あらゆる (i, j) 着ペアに対し、ARを評価しプロットしたものが図4である。縦軸にARの値、横軸には、平均得票率の比率 $[v_i]/[v_j]$ を用いた。得票率の比率を用いた理由は、グラフからも分かる通り、ARの値がほぼ得票率の比率のみに依存しているからである。

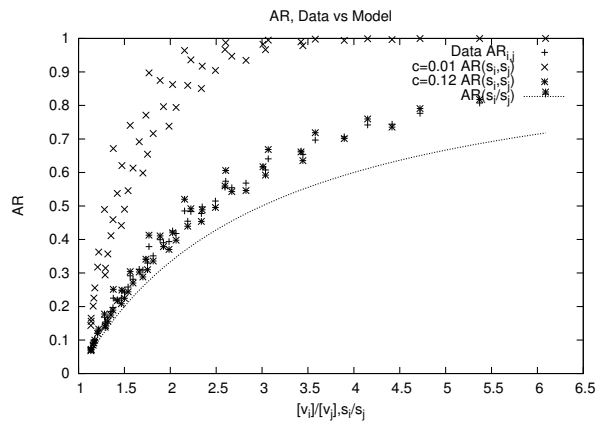


図4 得票率データから計算したAR値を平均得票率の比に対してプロットした(+記号)。投票モデルで、 $c = 0.01$ (×記号)、 $c = 0.12$ (*記号)および完全にスケール不変となる $c = \infty$ (実線)の場合の $AR(s_i, s_j)$ 値をプロットした。

モデルとデータの対応について説明する。投票モデルでは、スケールされた得票数 u_μ がガンマ分布 $p_\mu(u)$ に従い、 u_μ の期待値は $\langle u_\mu \rangle = \mu$ となる。単勝市場での得票率 $\{v_{i,k}, v_{j,l}\}_{k=1,\dots,N_i, l=1,\dots,N_j}$ は、スケールされた得票率 u_s, u_w に比例しているの、その比例係数を c と書くと N_i, N_j が十分大きい場合、次の関係式が成立する。

$$c \cdot \langle u_s \rangle = [v_i] \quad \text{および} \quad c \cdot \langle u_w \rangle = [v_j] \quad (7)$$

この比例係数は、全ての着順ペア (i, j) に対して同じ値をとらねばならない。 $c \cdot \langle u_s \rangle = c \cdot s$ 、 $c \cdot \langle u_w \rangle = c \cdot w$ より、 c の値と平均得票率 $[v_i], [v_j]$ が分かると、スコア s, w は、 $s = [v_i]/c$ と $w = [v_j]/c$ から求めることができる。こうして求まる着順 k の馬の集合のスコアを s_k と表すことにする。

比例係数 c の決定方法として、投票モデルでのARの値と得票率データから計算したARの値の差のあらゆるペアの和が最小になるものとした。つまり、 (i, j) ペアに対するARの値を $AR_{i,j}$ とスコア s_i, s_j に対する投票モデルの $AR(s_i, s_j)$ の差の絶対値の全てのペアでの和 $D(c)$ を最小にする。

$$D(c) \equiv \sum_{1 \leq i < j \leq 11} |AR_{i,j} - AR(s_i, s_j)|$$

図5に、 $D(c)$ を c に対してプロットした。このグラフから $c = 0.12$ のとき、実データのARと投票モデルのARの差 $D(c)$ がもっとも小さいことが分かる。

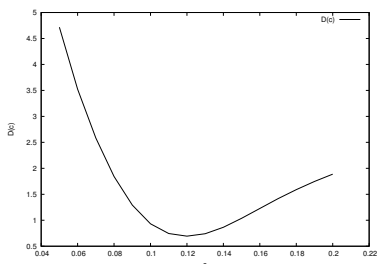


図5 (i, j) 着ペアの AR 値と、投票モデルでの AR 値の差の絶対値の和 $D(c)$ を c に対してプロットした。

また、図4には、 $c = 0.01$ と $c = 0.12$ および $c = \infty$ の場合に、AR の値 $AR(s_i, s_j)$ もプロットしている。 $c = \infty$ は、 $(s, w) \rightarrow (0, 0)$ の極限に対応するので、AR の値は $\frac{\alpha-1}{\alpha+1}$ となり、スコアの比率 $\alpha = s_i/s_j$ にのみ依存する。特に、 $c = 0.12$ の場合、得票率データの結果との一致はよく、また、この場合でも AR がほぼ s_i/s_j の比率にのみ依存することがわかる。

4.3 得票率の分布関数

前サブセクションの結果から、比例係数 c が 0.12 と定まったので、それを用いて投票モデルの得票率の確率分布 (ガンマ分布) をスケールし、実データと比較した。図6に、1着、2着、11着の場合について、実際得票率の分布関数をヒストグラム形式で、ガンマ分布を曲線で示した。

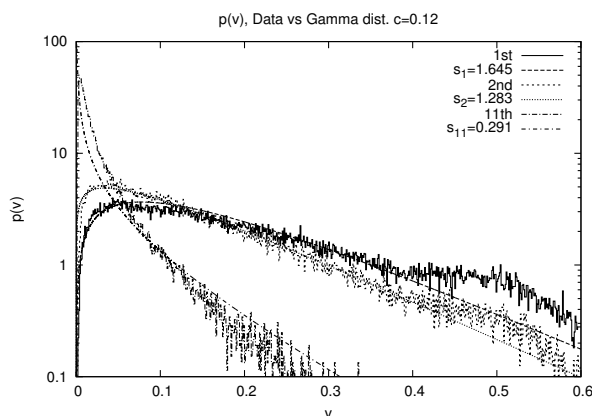


図6 得票率データの分布関数とガンマ分布関数をプロットした。1、2、11着の馬の得票率をヒストグラム形式で、平均得票率 \bar{v}_k からスコア s_k を評価し、ガンマ分布のパラメータに用いた。

1着の場合、得票率の確率分布は $v = 0.5$ の近くで膨らみを持つが、ガンマ分布でフィットできていることがわかる。2着、11着の場合も大体フィットできていることがわかる。この結果から、競馬単勝市場での競馬ファンの投票行動は、二値の候補者を導入した投票モデルで記述できることが分かる。また、ROC カーブがスケール不変となるレンジが非常に広いことから、競馬ファン

は多分に「コピーキャット」であることも分かる。

5. ま と め

今回、我々は候補者が優秀、劣等のような二値に分類し、その候補者についての情報をスコアとして投票者が知っている場合に、投票によって優秀な候補者と劣等な候補者を選別できるのかという問題を考えた。正しい情報が与えられ、他の投票者の行動に影響されずにインテリジェントな投票者として振る舞う場合は、二値の情報に基づいて投票し、優秀な候補者を選別する能力は100%となる。しかし、他の投票者の行動 (得票数) と二値の情報の和に比例した確率で投票するモデルでは、二値の候補者の混合が起こり、選別能力が低下する。また混合は得票数が少ない領域でスケール不変性を示し、どんなに得票数が少なくても優秀な候補者が存在するようになる。特に、スコア s と w の 0 極限では、投票者は完全にコピーキャットになり、混合のスケール不変性が全領域で成立することも分かった。この投票モデルは、投票者の「インテリジェント」および「コピーキャット」の二つの側面を s と w の水準でコントロールするモデルであることが分かる。二値のスコアが大きいならインテリジェントな側面が、小さいならコピーキャットな側面が強調される。投票行動の時間変化を見ると、初期段階はインテリジェントな側面が顔を出し、投票の後期段階でコピーキャットな側面が現れる。

また、以上の結果を JRA の競馬単勝市場の得票率のデータを用いて検証した。 i 着と $j (> i)$ 着の馬の集合のペアに対し、前者を優秀な候補者、後者を劣等な候補者と考え、その得票率で順序を決め、混合のスケール不変性、競馬ファンの識別能力、得票率の分布関数などを解析した。結果は、

- (1) 二値の馬の混合はスケール不変性を示す。優秀な馬の累積 $1-y$ に対し少なくとも 2 桁、場合によっては 4 桁のレンジで ROC カーブの両対数プロットは直線になる。
- (2) 得票率データから求めた AR は、得票率の比率でスケールされる。また、モデルの得票率と実データの得票率の比例係数 c を選び、投票モデルの AR と実データの AR がほぼ一致するようにできる。
- (3) 得票率の分布関数はガンマ分布に従う。競馬ファンは多分にコピーキャットである。

である。ただし、データの全ての側面を完全に再現できているわけではない。ここでは、実データでの指数 α の振る舞いと投票モデルの $\alpha = s/w$ の比較を示さなかったが、結論だけ述べると「完全に振る舞いが異なる」である。振る舞いが異なる理由は、得票率が小さい領域で累積の振る舞いが投票モデルが示唆する巾乗則には従わず、あるカットオフから急激に立ち上がる形となっているからである。カットオフの存在理由は、ある馬の得票率が

小さい場合、オッズが極端に大きくなるが、期待リターンが高いと考えてその馬券を買う投票者が現れ*2、結果として得票率が増加するからだと考えている。カットオフがあるにもかかわらず混合がスケール不変となる数理解明と指数 α の評価は今後の研究課題である。

最後に、今回の投票モデルの結果をもとに、多数決というシステムの安全性について述べる。他の投票者の投票結果が投票に影響する、すなわち「コピーキャット」が多いと、与えられた情報が正しい場合、優秀な候補者が劣等な候補者に埋没し、選別能力が低下することを先に述べた。では、この選別能力の低下はそんなに悪いことなのだろうか？投票者が独自に正しい情報に基づいて判断し投票するインテリジェントな場合は確かに選別能力100%である。しかし、もともとの二値の情報が完全に誤っている場合、優秀、劣等の区別が反転し、選別能力100%のシステムは0%のシステムになってしまう。一方、コピーキャットが存在すると、システムは二値の候補者の混合がスケール不変性を示すようになり、選別能力は低下するかもしれないが、二値の情報が完全に誤っている場合でも0%になることはない。あまり極端な結果はでないのである。このように、多数決というシステムはコピーキャットの存在により適度に誤りを含むようになり、候補者に関する情報に誤りが含まれていても、完全に間違った選択を避けるシステムであるとも言える。「白」(完全な正解)か「黒」(完全な不正解)の二元論ではなくいつも灰色な結果になることのメリットについて考えてみる必要があるのではだろうか。支持率調査、選挙予測といったさまざまな形で他の投票者に関する情報を流すことで、適度に誤りを持たせるように多数決をコントロールする。その「適度な誤り=スケール不変性」が多様性の保持を可能とし、社会が一斉にある方向を向いてしまう危険性をさけていると言えないだろうか？

謝 辞

筆者の一人(守)は E.Guitter 博士に分岐過程 [9, 10] の解析方法の教授、投票モデルのスケール不変性に関する有益な議論に感謝します。また、筆者らは競馬データの解析に対し中曽拓氏に、本予稿について建設的なコメントをいただいた二人のレフェリーの方に感謝します。レフェリーの方々コメントにより、情報の非対称性の問題(優秀、劣等は神のみぞ知ること)やモデルについての説明を追加することが出来ました。

◇ 参 考 文 献 ◇

[1] フリー百科事典『ウィキペディア (Wikipedia)』
 [2] ジェームズ・スロウイッキー (小高 尚子訳) 『『みんなの意見』は案外正しい』 角川書店 (2006)。

[3] S.Mori and M.Hisakado, *Exact Scale Invariance in the Mixing of Binary Candidates in a Voting System*, preprint <http://jp.arxiv.org/abs/0806.0185>.
 [4] M.E.J.Newman, *Contemporary Physics* **46**,323(2005).
 [5] A.-L.Barabasi and R. Albert, *Science* **286**,509 (1999).
 [6] B.Enleman, E.Hayden and D.Tasche, *Testing rating accuracy*, WWW.RISK.NET (2003).
 [7] M.Abramowitz and I.A. Stegun, eds. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, Dover, New York(1972).
 [8] Richard H.Thaler(篠原勝 訳) 『市場と感情の経済学』(ダイヤモンド社)。(1998).
 [9] P. Haccou, P.Jagers and V.A.Vatutin *Branching Processes*, Cambridge University Press (2005).
 [10] F.David, P.Di Francesco, E.Guitter and T.Jonsson, *J.Stat.Mech.*,02011(2007).

◇ 付 録 ◇

A. モデルの詳細とガンマ分布の導出

N 人の候補者のうち N_s 人が優秀、 N_w 人が劣等とし、それぞれのスコアを X_i^μ で記述する。 $\mu = s$ が優秀、 $\mu = w$ が劣等を表し、添え字 i は $i = 1, 2, \dots, N_\mu$ の値をとる。スコアの初期値は、優秀な場合 $X_i(0)^s = s$ 、劣等な場合 $X_i(0)^w = w$ とする。投票者は N 人の候補者に対し、スコアに比例した確率で投票し、スコアは票を得るごとに1増えるものとする。つまり、時刻 T での i 番目の μ 候補者のスコアが $X_i^\mu(T)$ のとき、彼が得票する確率 $P_i^\mu(T)$ は

$$P_i^\mu(T) = \frac{X_i^\mu(T)}{Z(T)}, \quad Z(T) = \sum_{\mu=s}^w \sum_{i=1}^{N_\mu} X_i^\mu(T) \quad (A.1)$$

である。もし、この i 番目の μ 候補者が得票したとすると、スコアが1増加し、他の候補者のスコアはそのままである。

$$X_i^\mu(T+1) = X_i^\mu(T) + 1 \quad (A.2)$$

$$X_j^{\mu'}(T+1) = X_j^{\mu'}(T), \quad \mu' \neq \mu \quad \text{または} \quad j \neq i \quad (A.3)$$

この離散時間 T の確率過程を取り扱いの容易な連続時間 t の分岐過程に変換する。スコア $X_i^\mu(t)$ を、時刻 t での μ 人の先祖から発生した子孫の数と考える。個々人は単位時間 dt あたり $\omega \cdot dt$ の確率で分裂し、二人の子孫となる。個々人の分裂過程は他の分裂過程とは独立に起こるので、一人の分裂過程を解析すれば、その和として最初 μ 人の先祖から出発した場合の子孫の数の時間変化を知ることができる。離散時間モデルとの関係は、分裂を一回行うことが1票を得ることに、分裂確率がその時点での子孫の数に比例することが確率法則 (A.1) に対応する。

i 番目の μ 候補者のスコア $X_i^\mu(t)$ を μ 人の祖先のスコア $x_{i,k}^\mu(t)$ の和であらわす。

$$X_i^\mu(t) = \sum_{k=1}^{\mu} x_{i,k}^\mu(t) \quad (A.4)$$

$x_{i,k}^\mu(t)$ の満たす初期条件は $x_{i,k}^\mu(0) = 1$ である。期待値 $\langle X_i^\mu(t) \rangle$ や $\langle x_{i,k}^\mu(t) \rangle$ は $e^{\omega \cdot t}$ で単調に増加する。そこで、スケールされた変数 $U_i^\mu(t), u_{i,k}^\mu(t)$ を導入する。

$$U_i^\mu(t) = X_i^\mu(t)e^{-\omega \cdot t} \quad (A.5)$$

$$u_{i,k}^\mu(t) = x_{i,k}^\mu(t)e^{-\omega \cdot t} \quad (A.6)$$

問題は、このスケールされた確率変数 $U_i^\mu(t), u_{i,k}^\mu(t)$ の確率分布関数を求めることである。前者は添え字 i に、後者は添え字 i, k や μ

*2 そうした行動は、過去の馬券の回収率から、賢い選択ではないと知られている [8]。

は依存しないので、以下この確率変数を $u_\mu(t)$, $u(t)$ と書く。確率分布関数を

$$\text{Prob.}(U_\mu(t) = u) = p_{\mu,t}(u) \quad (\text{A.7})$$

$$\text{Prob.}(u(t) = u) = p_t(u) \quad (\text{A.8})$$

それらの $t \rightarrow \infty$ の極限を

$$p_\mu(u) \equiv \lim_{t \rightarrow \infty} p_{\mu,t}(u) \quad (\text{A.9})$$

$$p(u) \equiv \lim_{t \rightarrow \infty} p_t(u) \quad (\text{A.10})$$

と定義する。

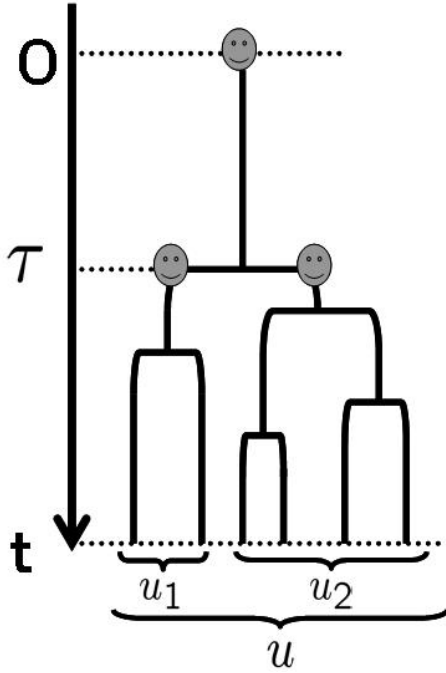


図 A.1

$p(u)$ は、最初の 1 人の祖先から次々と分裂を繰り返したときのスケールされた子孫の数の従う確率分布である。時刻 $t = \tau$ に初めて分裂し、二人の子孫が生まれたとする (図 A.1 参照)。これらの子孫のおおののから生まれるスケールされた子孫の数を u_1, u_2 と書くとする、 $t \rightarrow \infty$ の極限では u と同じ確率分布に従う。 u と u_1, u_2 の間には

$$u \cdot e^{\omega \cdot t} = (u_1 + u_2)e^{\omega \cdot (t - \tau)} \quad (\text{A.11})$$

が成立すること、時刻 τ から $\tau + dt$ の間に初めて分裂する確率は $e^{-\omega \cdot \tau} dt$ であることに注意すると、 $p(u)$ は次の積分方程式を満たすことが分かる。

$$p(u) = \int_0^\infty e^{-\omega \cdot \tau} \omega d\tau \times \int_0^\infty du_1 \int_0^\infty du_2 \times p(u_1)p(u_2)\delta(u - (u_1 + u_2))e^{-\omega \cdot \tau} \quad (\text{A.12})$$

変数変換 $X = e^{-\omega \cdot t}$ を行くと、

$$p(u) = \int_0^1 dX \int_0^\infty du_1 \int_0^\infty du_2 \times p(u_1)p(u_2)\delta(u - (u_1 + u_2)X). \quad (\text{A.13})$$

ラブラ - ス変換

$$\hat{p}(s) \equiv \int_0^\infty p(u)e^{-s \cdot u} du \quad (\text{A.14})$$

を用いて積分方程式 (A.12) を変換すると、次の方程式が得られる。

$$\hat{p}(s) = \frac{1}{s} \int_0^s dv \hat{p}(v)^2. \quad (\text{A.15})$$

両辺 s で微分すると、次の微分方程式が得られる。

$$s \frac{d\hat{p}(s)}{ds} = \hat{p}(s)^2 - \hat{p}(s). \quad (\text{A.16})$$

これを解くことにより、

$$\hat{p}(s) = \frac{1}{1 + a \cdot s} \quad (\text{A.17})$$

が得られる。ここで a は積分定数である。逆ラプラス変換により、 $p(u)$ は、

$$p(u) = \frac{1}{a} e^{-\frac{u}{a}} \quad (\text{A.18})$$

となる。確率変数 u はスケールされた子孫の数であり、その期待値 $\langle u \rangle$ は 1 となる。この条件から積分定数 a が求まり、確率分布 $p(u)$ は次式ようになる。

$$p(u) = e^{-u} \quad (\text{A.19})$$

一方、 $p_\mu(u)$ は、 μ 個の $p(u)$ を畳み込むことにより導出できる。結果を書くと、

$$p_\mu(u) = e^{-u} \frac{u^{\mu-1}}{(\mu-1)!} \quad (\text{A.20})$$

以上が、式 (1) の導出の詳細である。

最後に、この確率分布の導出を見ると、指数 μ の値は整数に限られそうだが、実際には一票の重みを導入することにより任意の実数の場合の拡張できることを注意しておく。