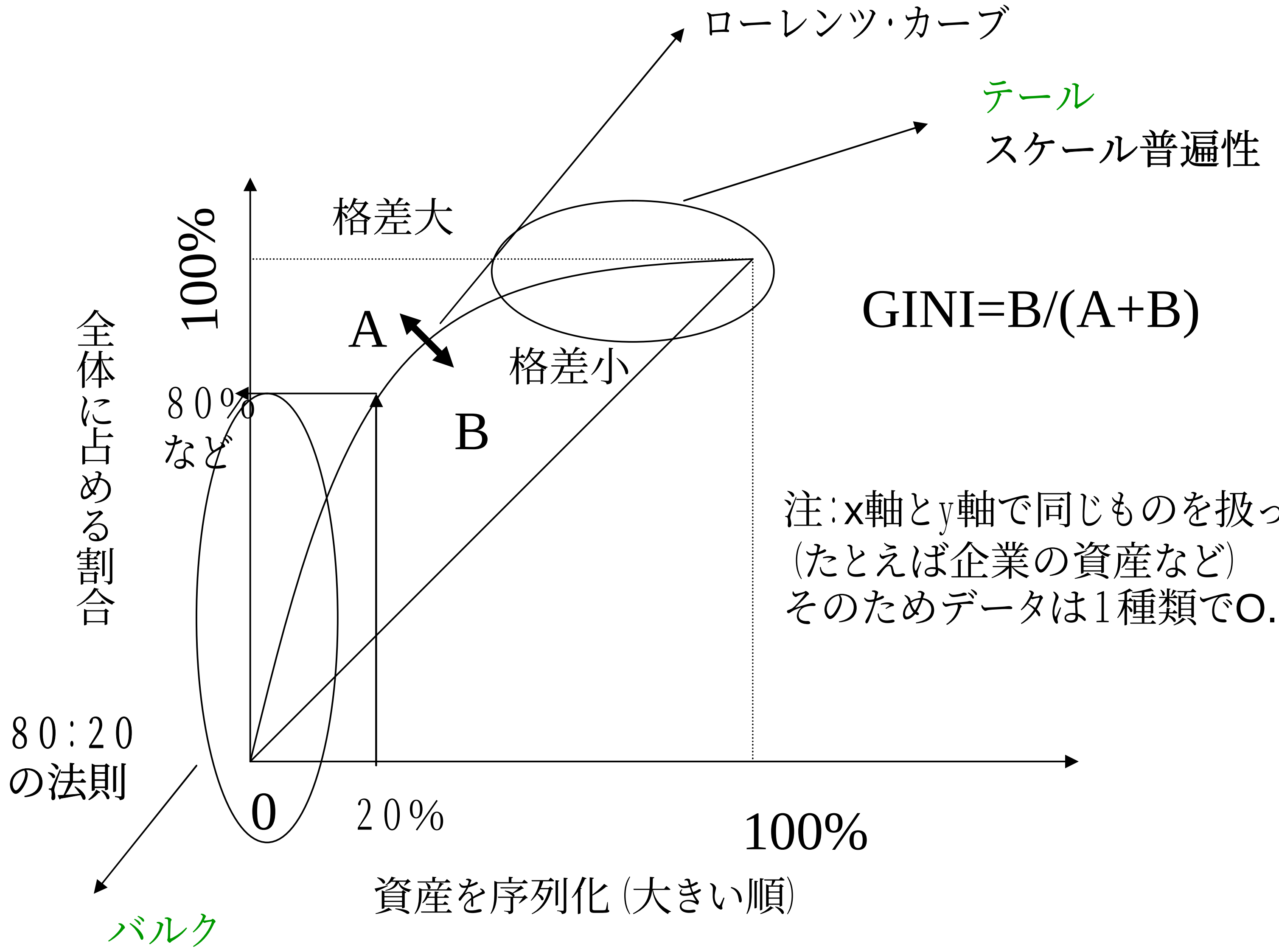
0と1の世界のスケール普遍性

ローレンツカーブ

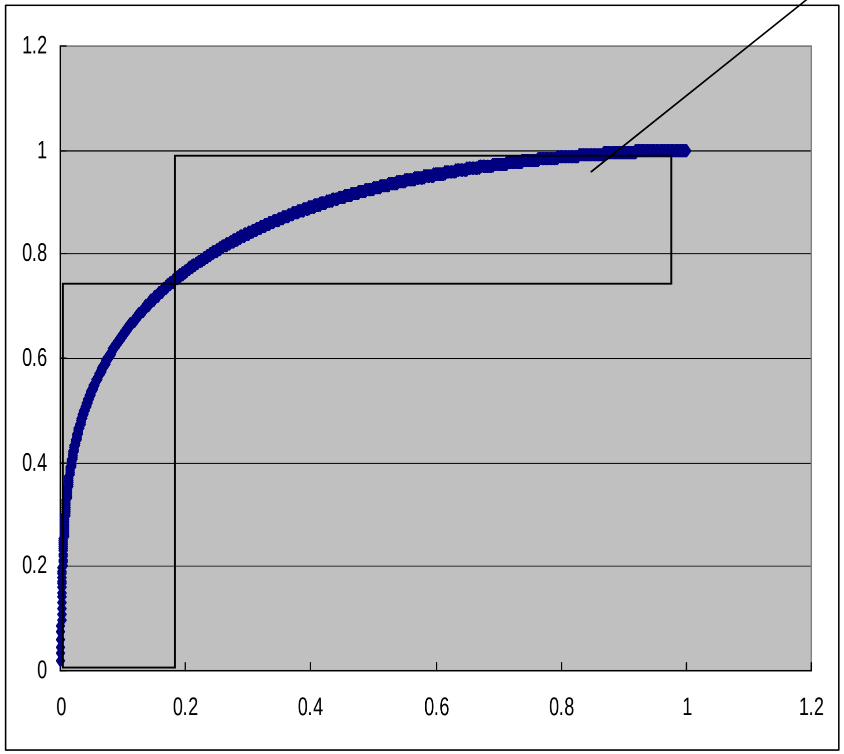


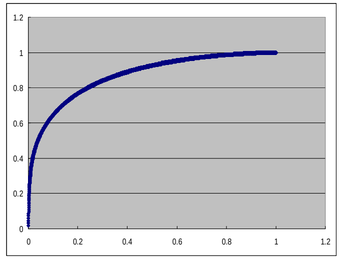
GINI=B/(A+B)

注:x軸とy軸で同じものを扱っている。
(たとえば企業の資産など)
そのためデータは1種類でO.K.

企業の資産のローレンツ曲線

ここを拡大





メガバンクの融資先中小企業の資産の分布

スケール普遍性

ローレンツカーブの特徴。

1（0，0）と（1，1）を通る。

2上に凸

3全域でスケール普遍性が成り立っている場合

Y＝1－（1－x）^α

⇔

1－y＝（1－x）＾α

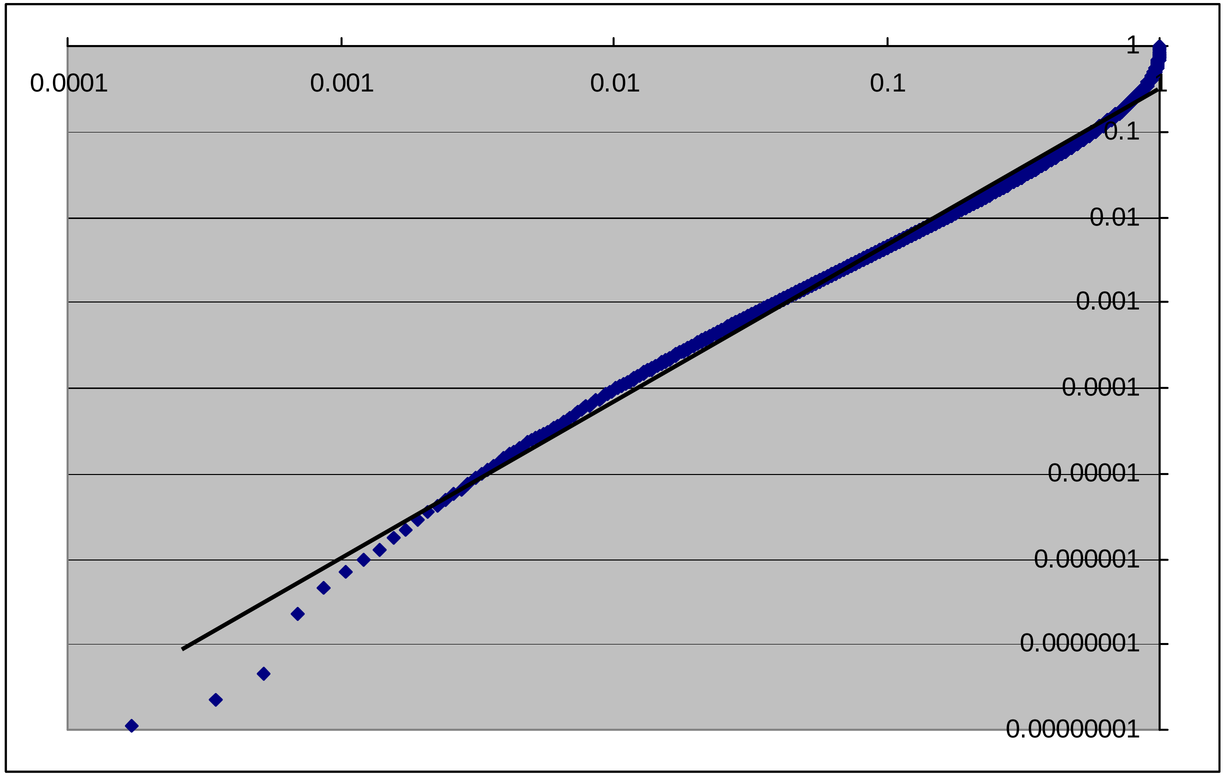
反転しLog-Logでプロットした場合直線！

注：８０：２０の法則の場合α＝4

すべて均等の場合α＝1

GINI係数以前はこのαが指数として使われていた。しかしバルク部分を評価する現在のGINI係数が次第に使用されるようになってきた。

企業の資産の分布のスケール普遍性



メガバンクの中小企業の資産の分布

クレジット・スコアリング・モデル

財務データからデフォルト確率を予想するモデル。

パフォーマンスの検証

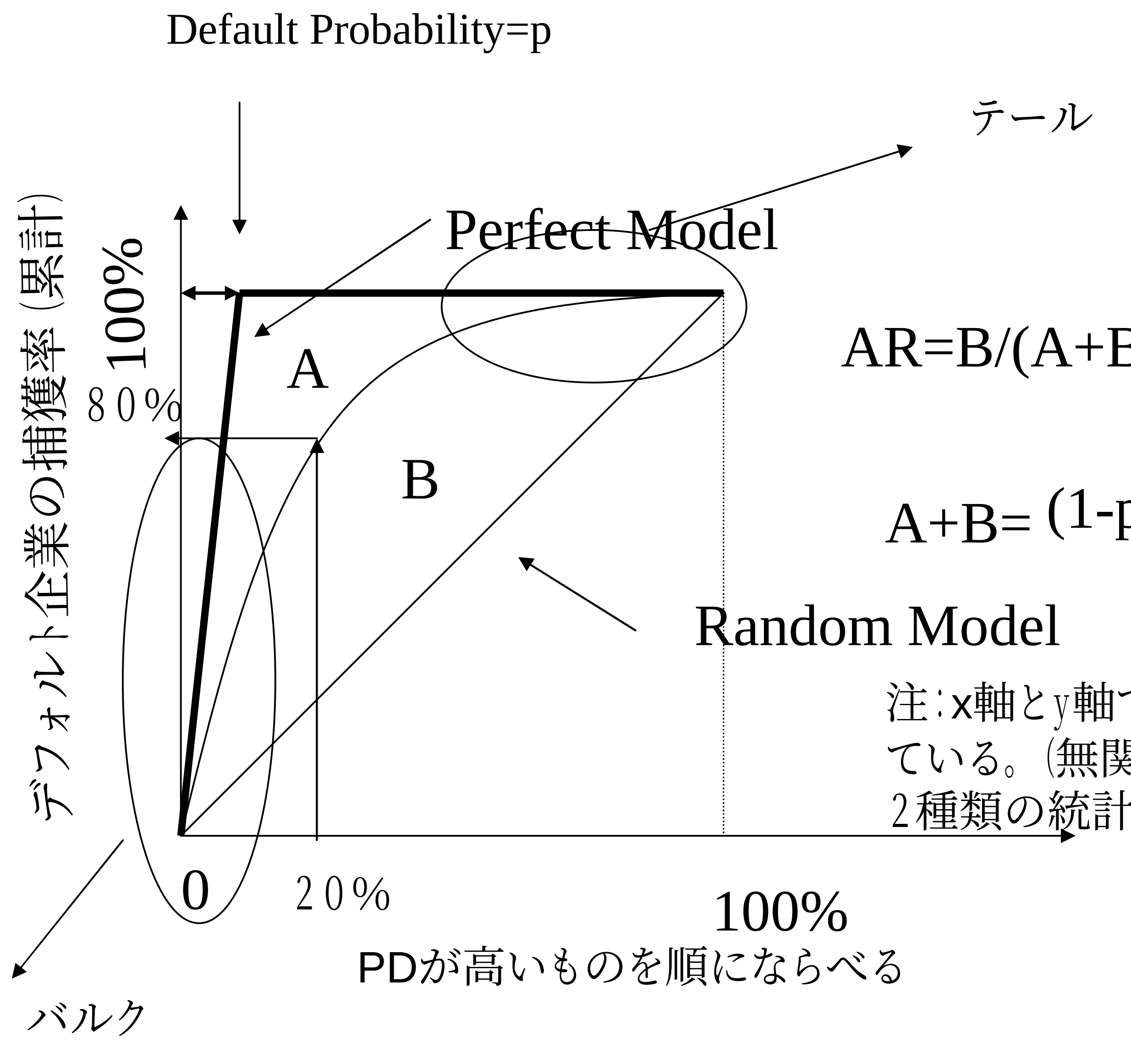
☆デフォルト確率（PD）と1年後デフォルトしたか（1）、しなかったか（0）の関係性を確認。

☆最もメジャーな指標としてAR（Accuracy Ratio)

序列 ↓

PD	デフォルトしたかどうか？
４０％	×
２０％	×
１０％	０
３％	×
２％	○
１％	○
０．１％	×
０．０２％	○

AR（Accuracy Ratio)

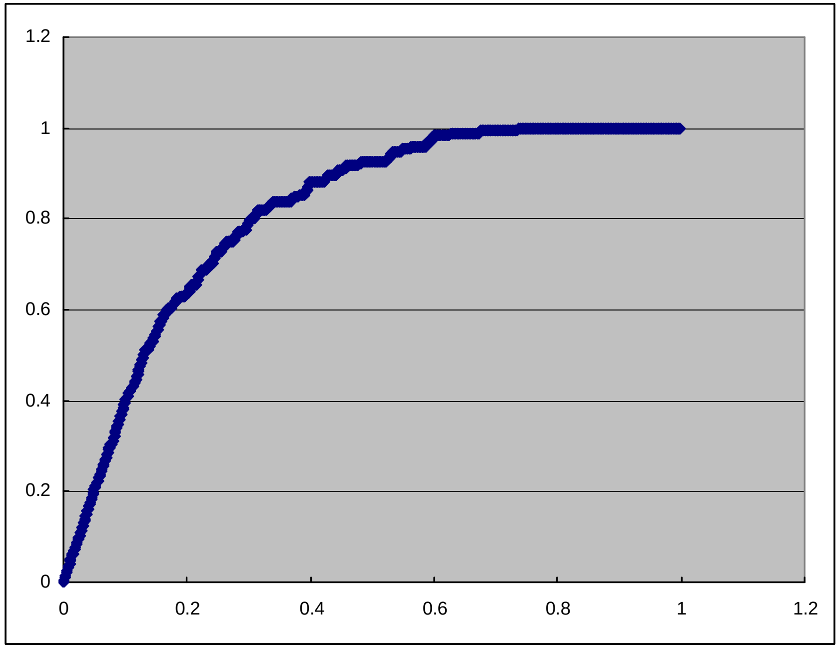


AR=B/(A+B)

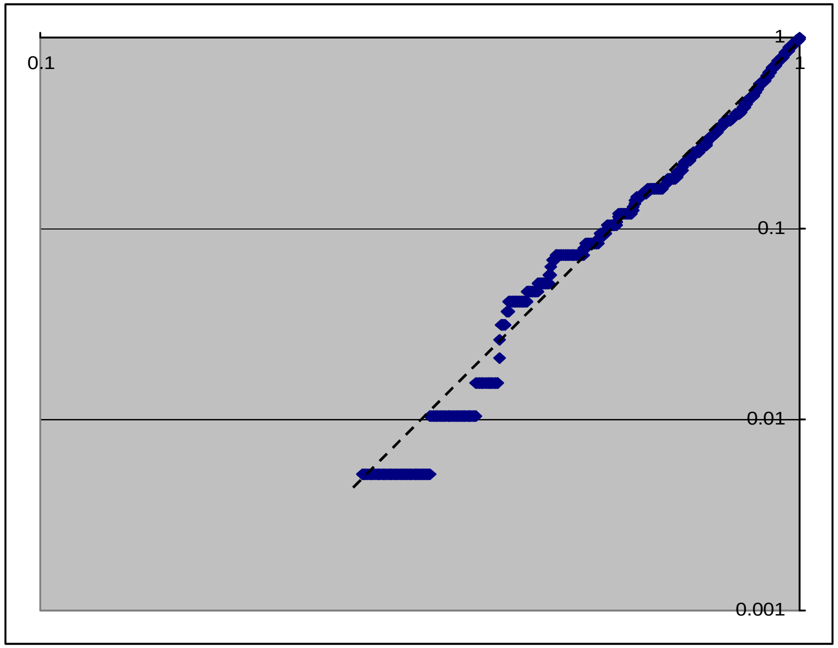
A+B= (1-p)/2

注：x軸とy軸で異なる量を扱っている。（無関係ではない。）
2種類の統計量が必要。

クレジット・スコアリング・モデルのローレンツカーブ

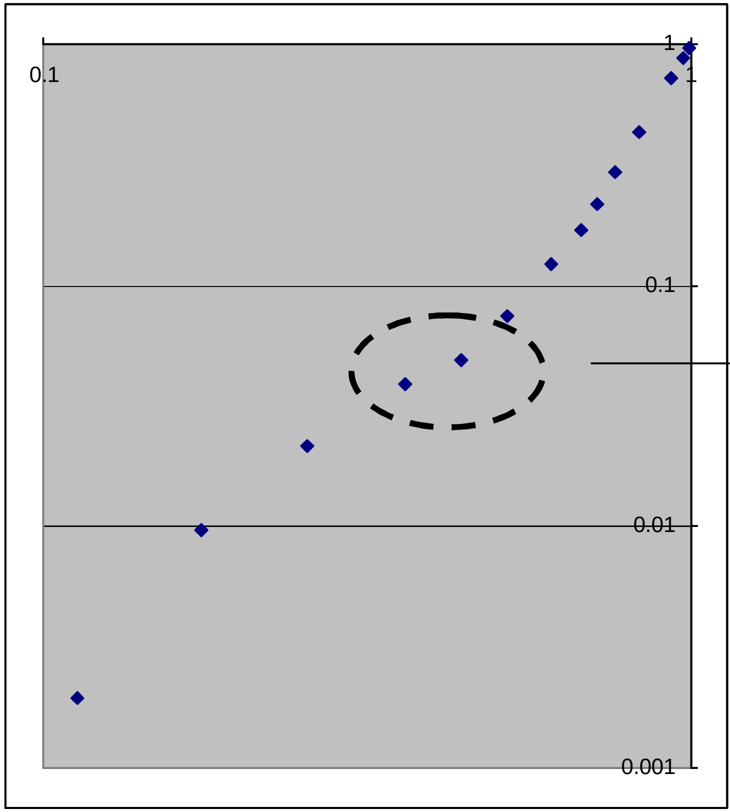


ローレンツカーブ



反転してLog-Logプロットしたもの

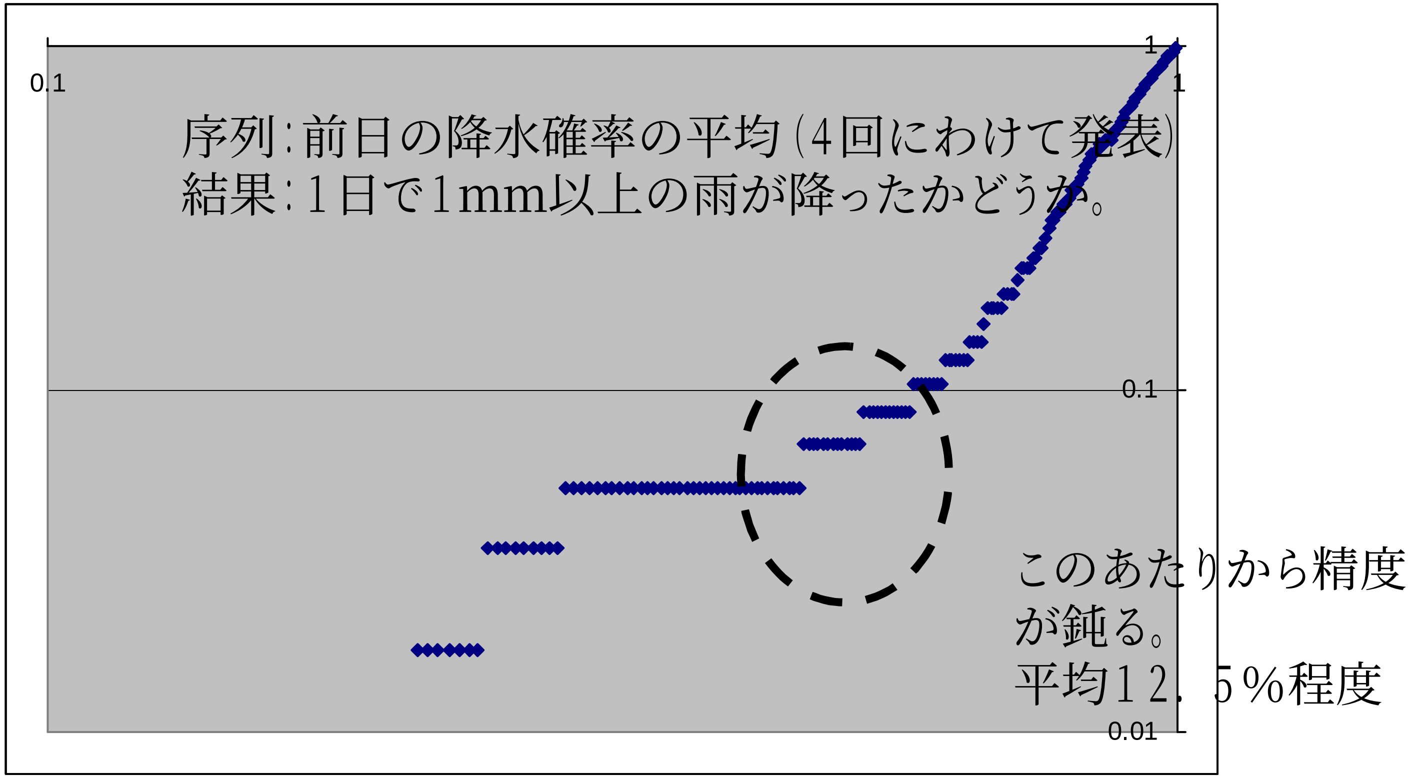
S&Pの格付けのローレンツカーブ



BBBあたりで
曲がっている。
→
スケール普遍な精度
にはなっていない。

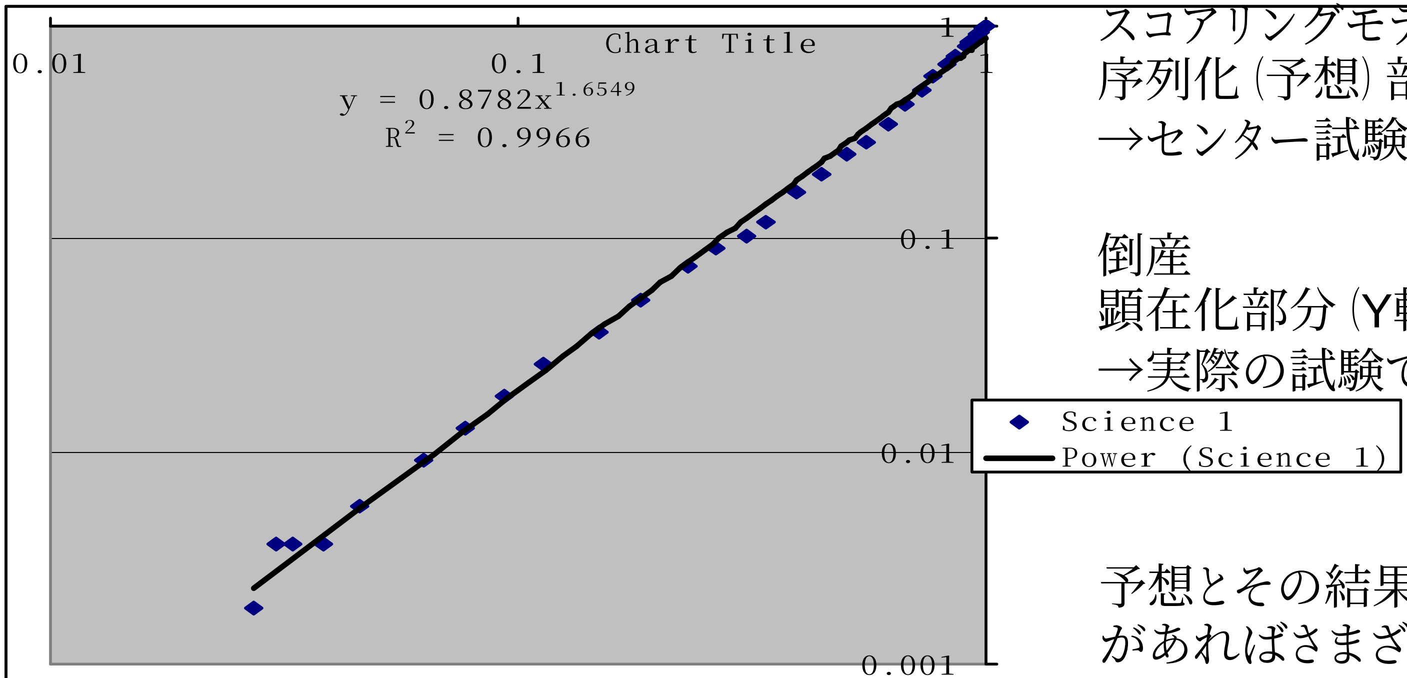
１９９８年から２００７年の１０年間のデータを使用

天気予報



２００８年半年分

大学入試



スコアリングモデル
序列化（予想）部分（X軸）
→センター試験

倒産
顕在化部分（Y軸）
→実際の試験で合格

予想とその結果
があればさまざま
なものに応用可能！

0,1世界のローレンツ曲線とは？

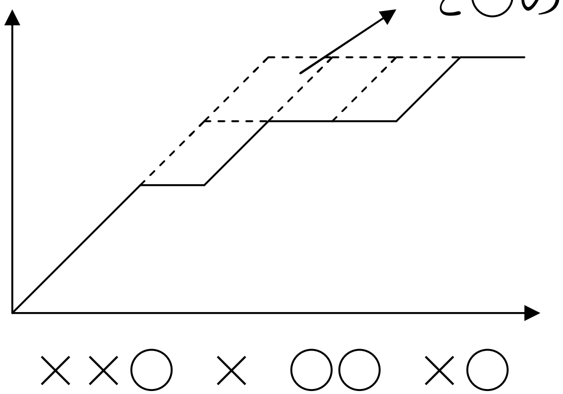
例

××○×○○×○

予想が顕在化した結果

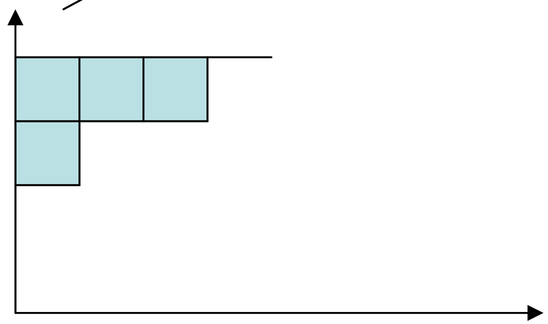
序列化

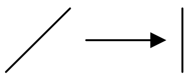
この面積は×と○の交換に相当



ローレンツ曲線

ヤングダイアグラムに解釈可能





ROCカーブ
PDの順にデフォルトなら上
ノンデフォルトなら右に進んでできたカーブ。

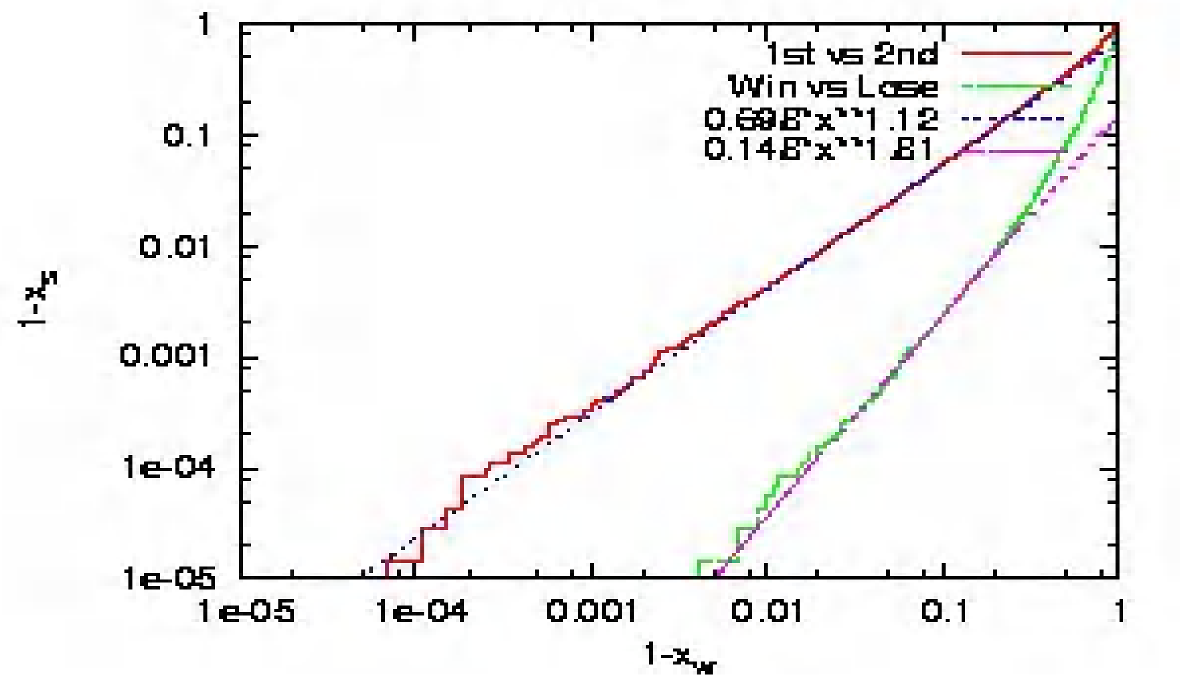
ARとは？（まとめ）

1　ローレンツカーブもしくはROCカーブでの面積比。
2　×と〇がランダムな状態からどの程度完成しているかを示したもの。
3　ケンドールの順位相関（順位付けの似具合）の一種。

１２３４５６７
１４２５３７６　　相関１３/２１
０１０１０１１　　AR　　０．５

ケンドールの場合は両方順位なので対称性があったがARの場合対称でないので注意が必要。

なぜ0と1の世界でスケール普遍性がみられるのか？



競馬のROCカーブ

1着なら上2着なら右
1着なら上それ以外なら右

何の順にならべたものか？

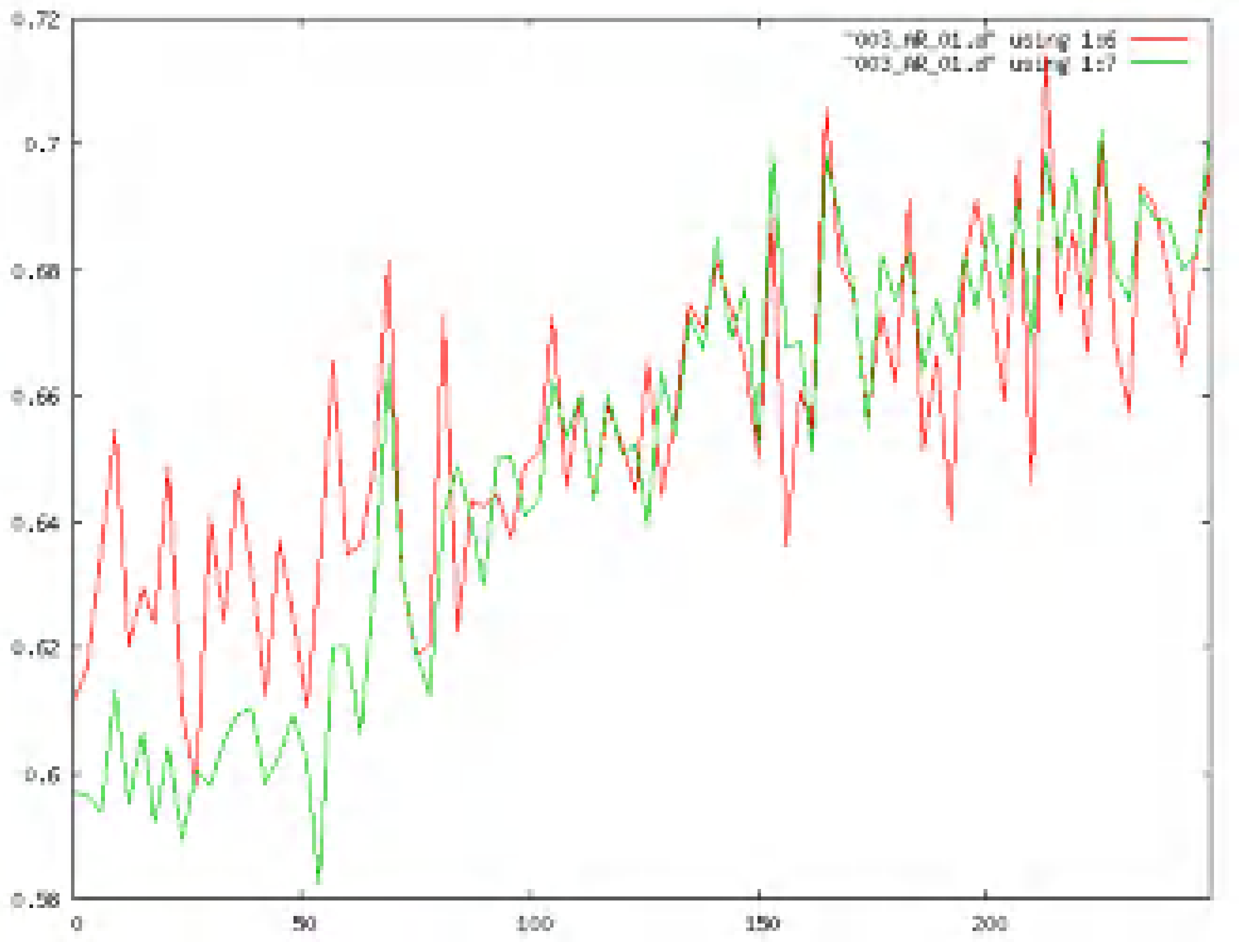
競馬について

★投票者：勝ち馬投票券（単勝馬券）を購入
★倍率はオッズを通して投票者にフィードバックされる。
★倍率は投票数の逆数に比例する。
→
☆オッズから投票割合が逆算可能！
これを序列化に用いる（クレジット・スコアリング・モデルではPDに相当。X軸）
→投票数が多いものほど勝つ可能性が高いと仮定。
☆0，1（○×）は勝ち馬か負け馬かに対応させる。（Y軸）

→
多数決で正解がわかる数少ない例とも考えられる。

注：投票者は必ずしも勝ち馬を当てるためだけに馬券を買っているわけではないが、
単勝馬券（1着のみ払い戻し）は比較的その性格が強いと考えられる。例外としては応援馬券があるが、G1レースなどは全体の中で占める割合は十分小さい。

競馬（オッズ）のAR



赤：期待
緑：実際

現在
約２０年分

多数決のモデル

×勝ち馬
〇負け馬
たとえば2レース分
××〇〇〇〇〇〇
★投票者が知りえるのは投票数（スコア）
★最初の投票数を　s、w
★×はs、〇はwとする。
★馬によって変わらず同じ。
★一般的には$s > w > 0$、ただしこれは整数とは限らない。

投票の約束
1　1人ずつ1票投票。
2　投票は投票数（スコア）に比例する確率で投票。
3　投票数は一人が投票するごとに更新。
4　時間は∞まで。

モデルの特徴

☆パラメータは2つ

①sとwの比

比が大きいほど正解がえられやすい。

→ARは高くなる。

②sとwの水準（投票を1票ずつとしているので）

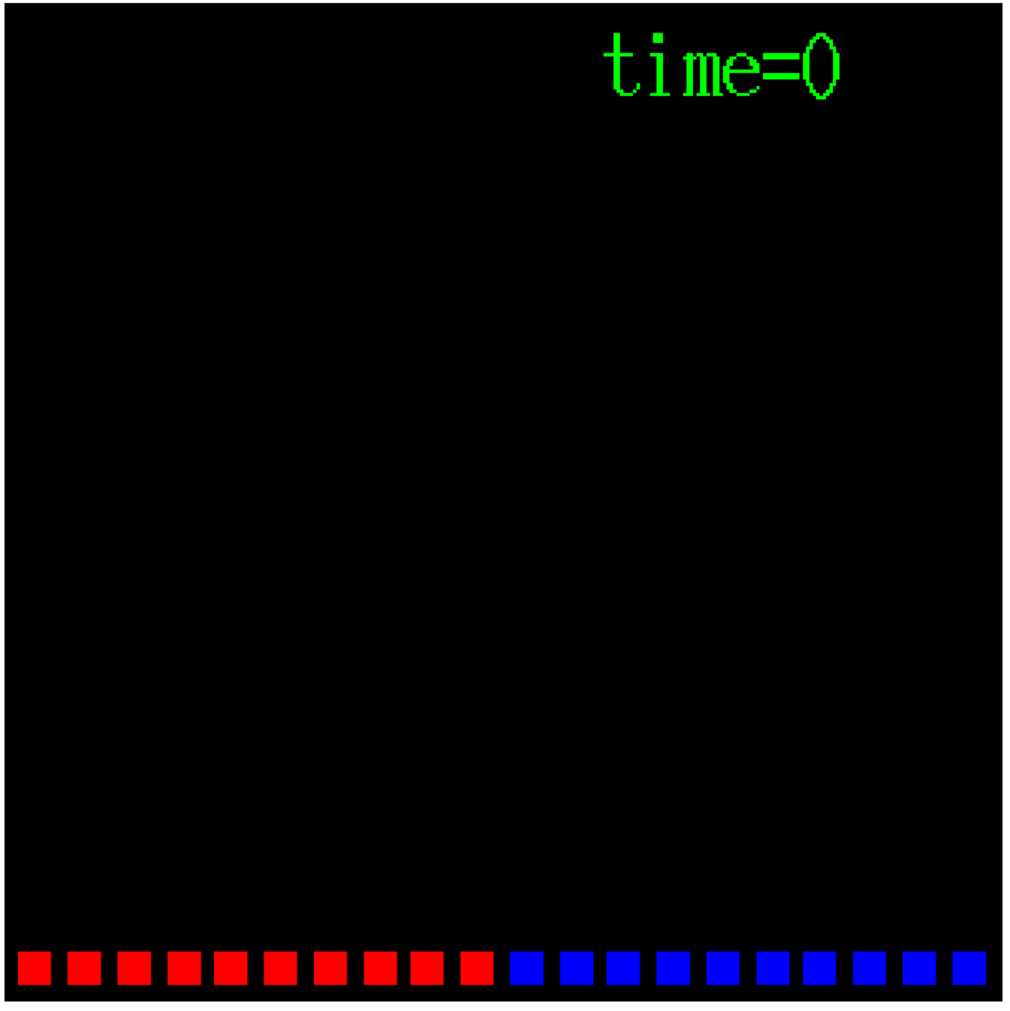
水準が高いほど他者の投票（フィード・バック）の影響は少なくなる。

→コピーキャット（他人の投票を真似し投票をする人）が少なくなる。これはインテリジェント（他人の影響を受けずに投票する人）の割合が増える。

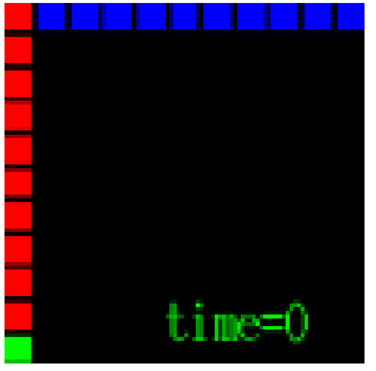
注：S、W→0のリミットがコピーキャット・リミット＝100％追随！

シミュレーション

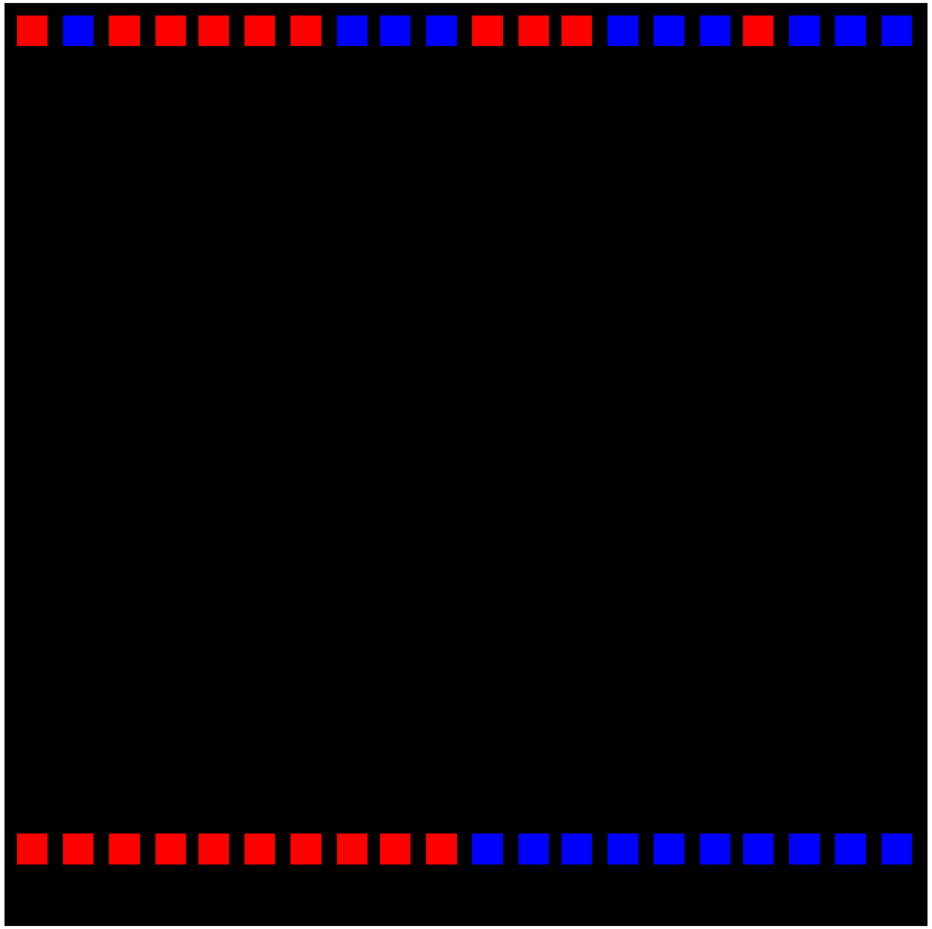
$$N = 20, N_s = 10, N_w = 10, s = 2, w = 1$$



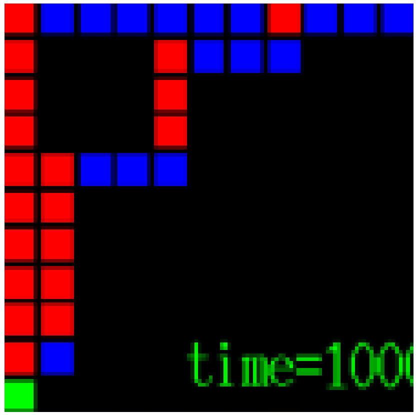
ROC カーブ



投票結果



ROC カーブ



得票数
124
110
94
92
88
79
71
67
56
43
43
33
21
20
17
16
15
6
5
0

解析結果

分岐過程（連続時間モデル）にマップすると厳密に解ける。

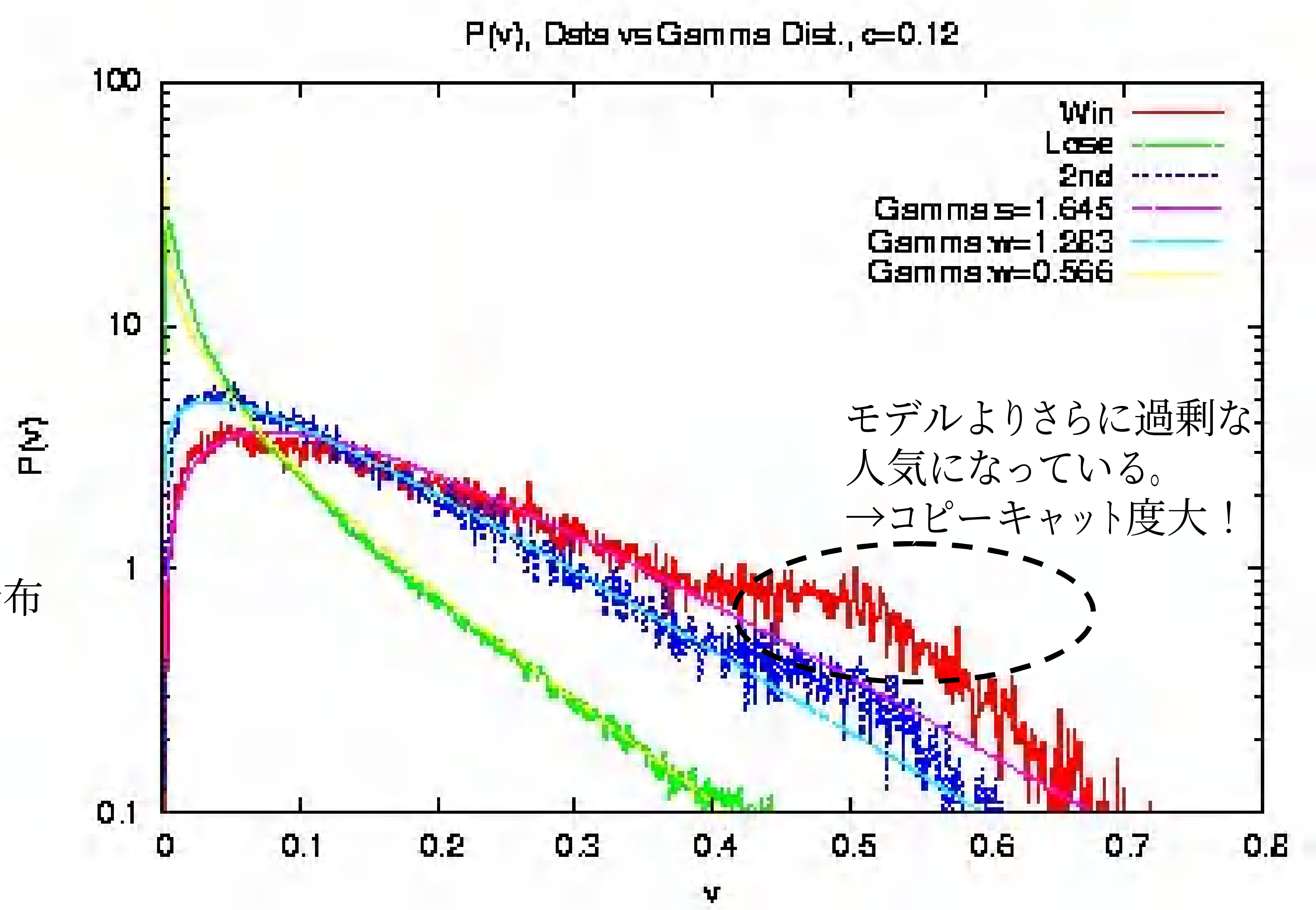
投票数はガンマ分布に従う。

$$\begin{cases} 1 - x_w(t) = \dfrac{1}{\Gamma(w)} \cdot \gamma(w, t) \\ 1 - x_s(t) = \dfrac{1}{\Gamma(s)} \cdot \gamma(s, t) \end{cases}$$

$\gamma(a, t) \equiv \int_0^t e^{-u} \cdot u^{a-1} du$ **: Incomplete Gamma Function of the 1st Kind**

得票率の分布はガンマ分布？

P(v)



分布

得票率

スケール不変性

(x,y)=(1,1) 近傍(テール部分)でのスケール不変性

$$1 - x_s \sim (1 - x_w)^{\alpha} \quad \text{with} \quad \alpha = \frac{s}{w}$$

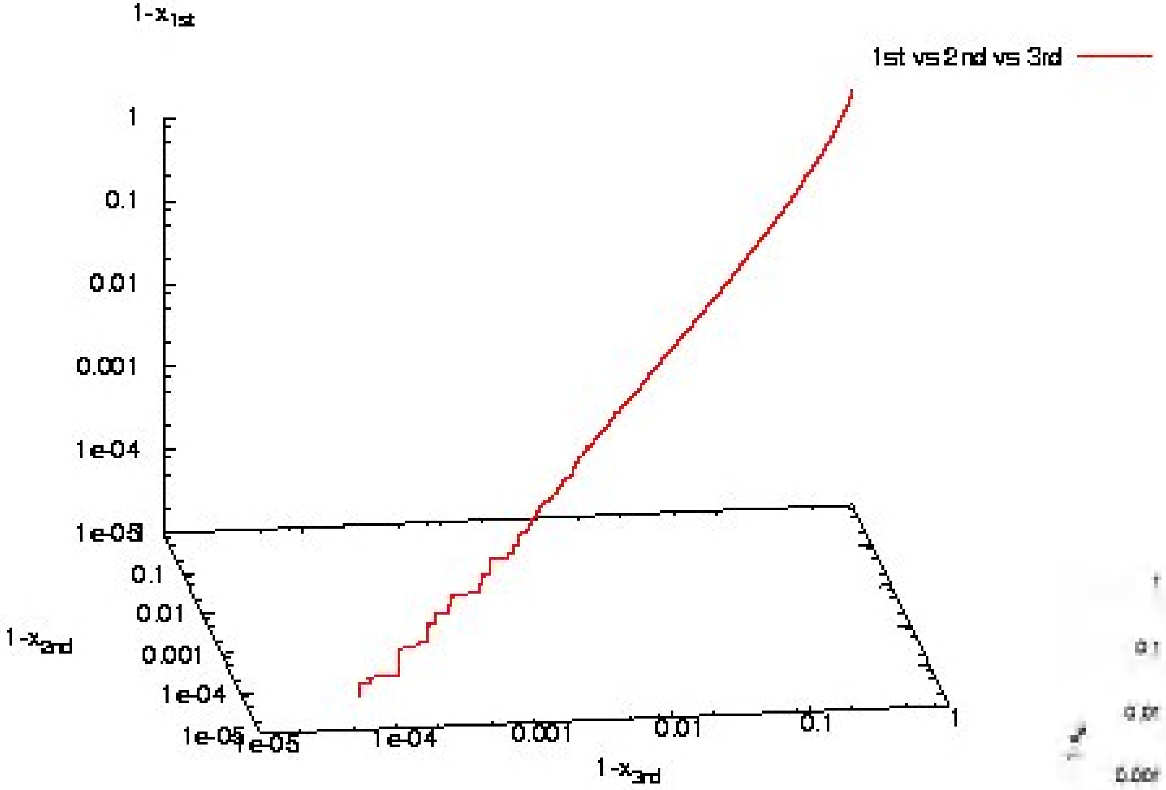
$$(s, w) \to (0, 0) \quad \text{with} \quad \alpha = \frac{s}{w} \quad \text{fixed}$$

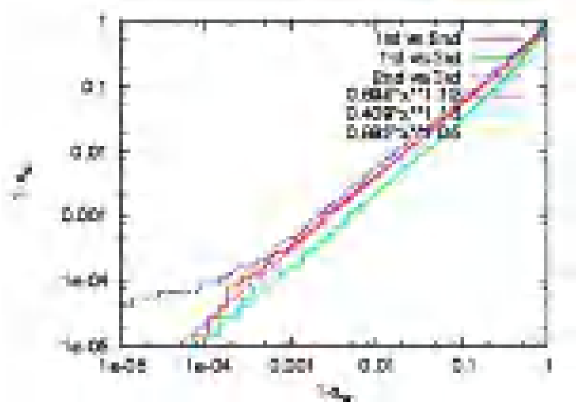
コピーキャット・リミット

$$1 - x_s = (1 - x_w)^{\alpha}$$

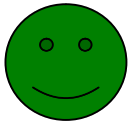
コピーキャットが多くなるにつれて直線の領域は広くなる！

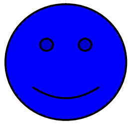
スケール不変性＝3種の場合






1着


2着


3着

まとめ

競馬について

☆モデルは簡単であるがまあまあ競馬を記述しているのではないか。

☆コピーキャットはかなりの数がいる。

多数決について

☆コピーキャットが多い場合、正解の率は低いものの情報が根本的に間違っていた場合（s＜wの場合）などに比べて安定性は高い。