2011–03–09

第 58 回日本生態学会大会 自由集会 (W15)

データ解析で出会う統計的問題 ー 選択や勝負の統計モデル

選択・勝敗の階層ベイズモデル

久保拓弥 kubo@ees.hokudai.ac.jp

http://goo.gl/eDZLE

今日あつかう話題

- 闘争のモデル: Bradley-Terry モデルと Burczyk モデル
- 「個体差」と階層ベイズ統計モデル
- 個体の「内部状態」をあつかう一般化状態空間モデル

複雑な意思決定

状況・文脈に依存する動物の応答

個体の記憶?

「複雑だ」をこえた理解 (粕谷さん)

一般的には「内部状態」の変化

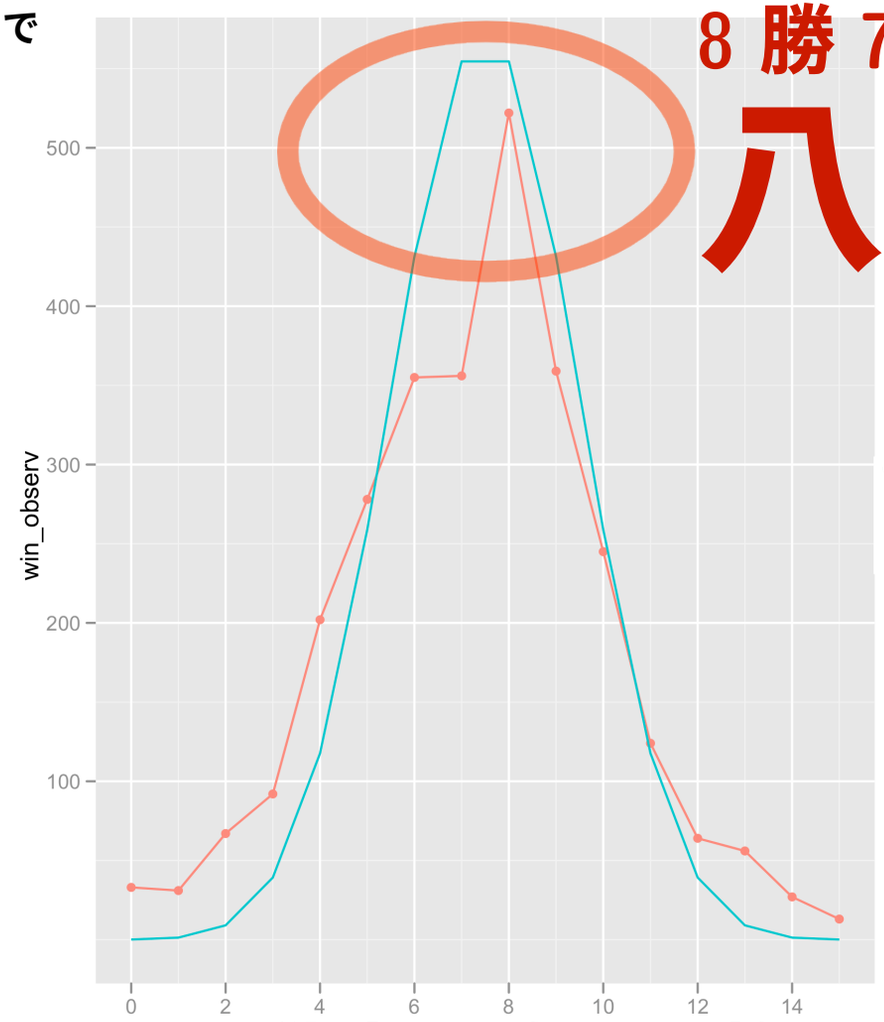
まずは「個体差」のあつかい

とりあえず

2011–03–09　　(2011–03–09　19:32 修正版)

1999 年から 2010 年 9 月までの大相撲の勝敗分布

柏野雄太氏の「実践！ R で学ぶ統計解析の基礎: 第 7 回 大相撲のアノーマリー (1)」からの引用



15 戦中の勝ち星数

8 勝 7 敗が多すぎる?

八百長?!

……といったハナシはここでは検討しませんが……

観測データ

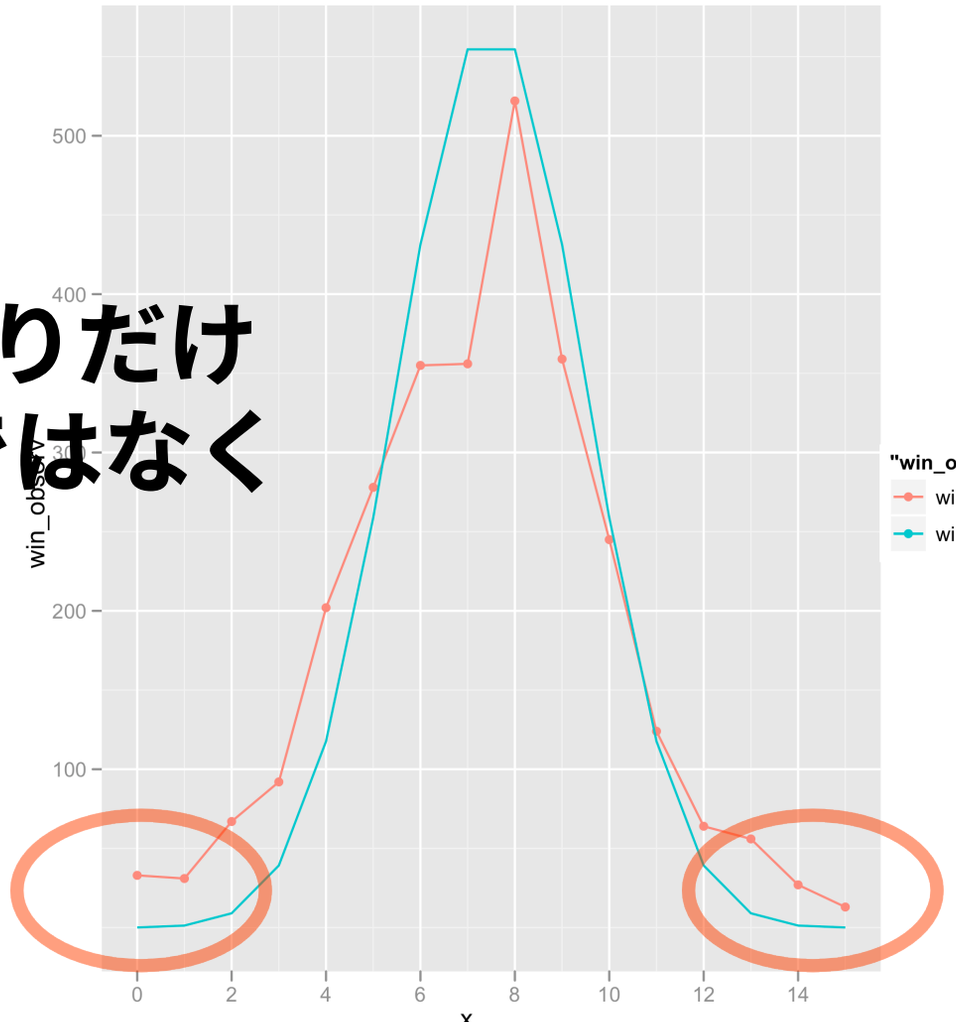
二項分布による予測

いろいろとおもしろいモデリング可能でしょう

http://www.atmarkit.co.jp/fcoding/articles/stat/07/stat07a.html

生態学研究者はスソのずれにも注目してみたい

「平均」まわりだけ
注目するのではなく



観測データ

二項分布による予測

二項分布の予測とくらべると

全勝・全敗 なんかが多すぎる?

そんなのあたりまえでしょう

強い力士もいれば弱い力士もいる

「勝つ確率」が全員 0.5 のわけない

「個体差」は直接観測できない

実験・観測と統計モデルあてはめ

によって推定可能な数量

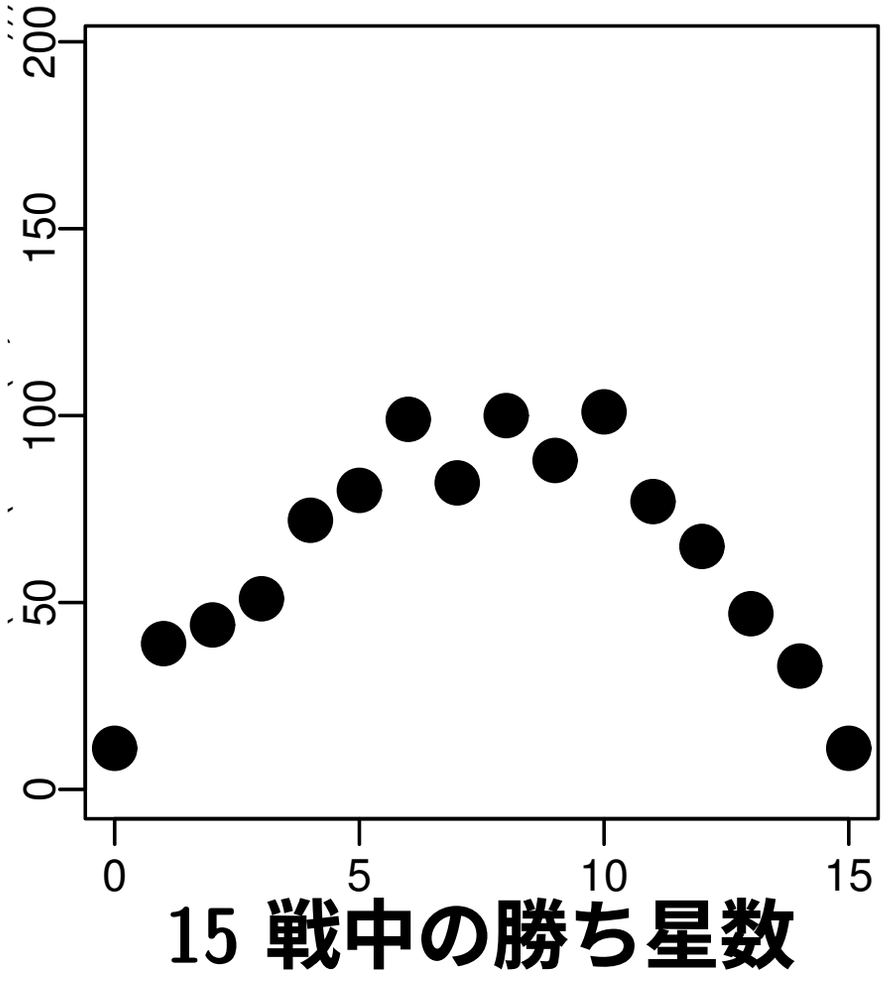
「個体差」の要因は特定できないかもしれない

- fixed effects: 体重とか? — 人間が測定可能な説明変数
- **random effects**: 体重は同じなのに強さに差がある!

今日は後者に注目

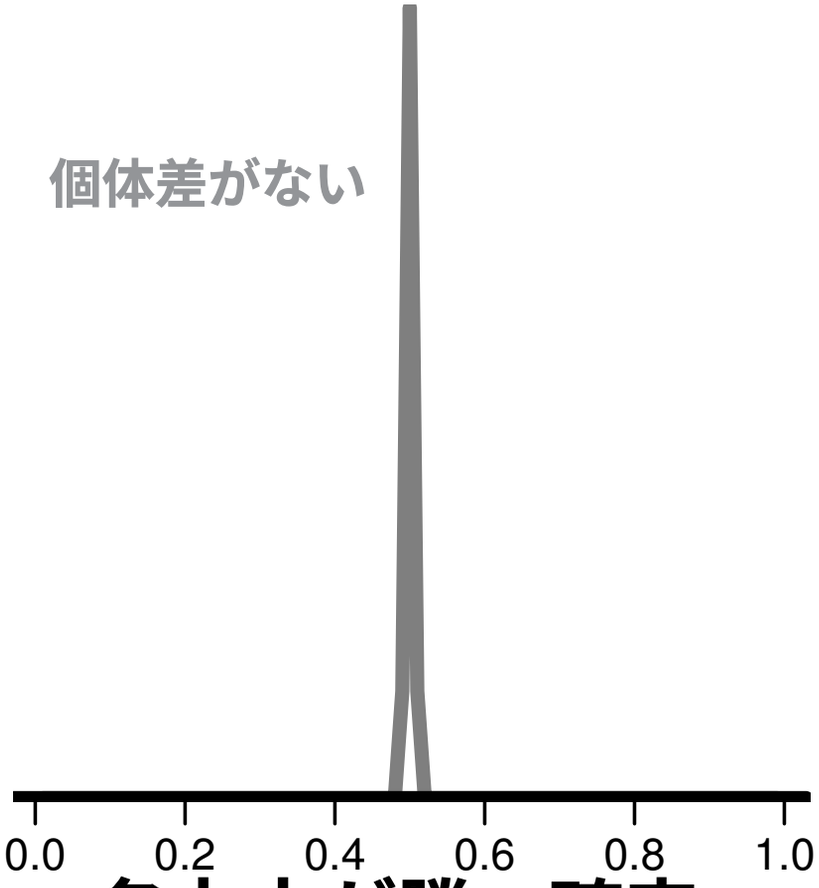
架空データ: 体格同じ架空力士 1000 人

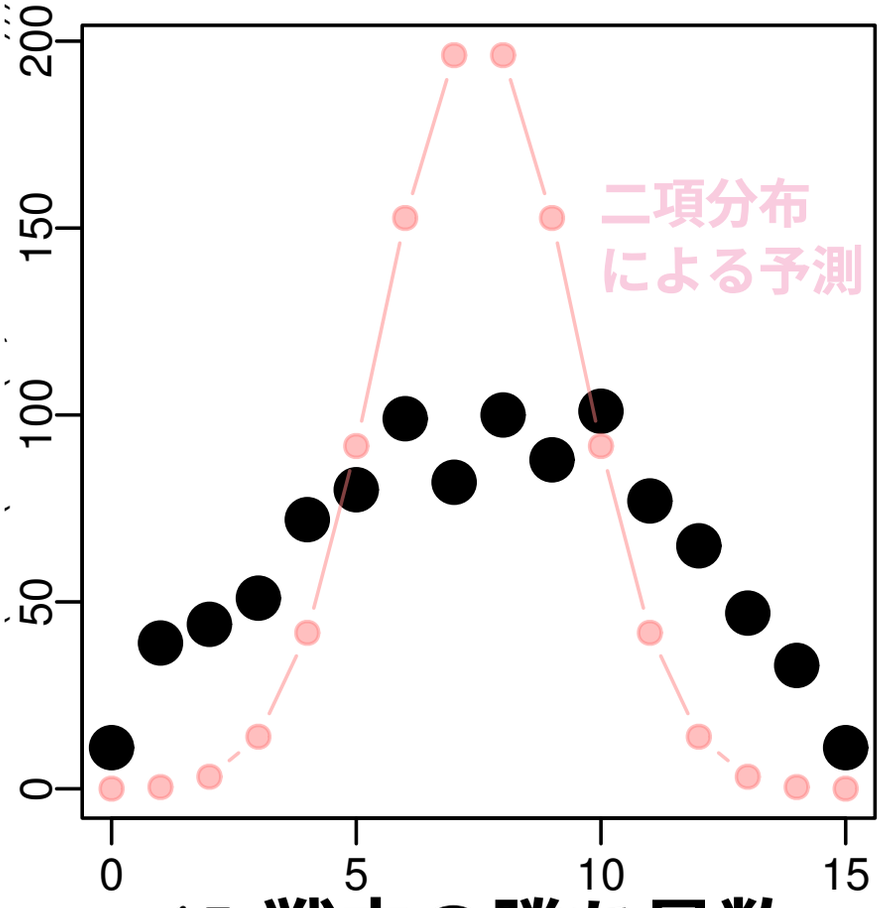
力士たちの平均勝率は 5 割



生起確率 0.5 の二項分布ではだめだめ?

15 戦中の y 勝の分布だから二項分布のはず……?

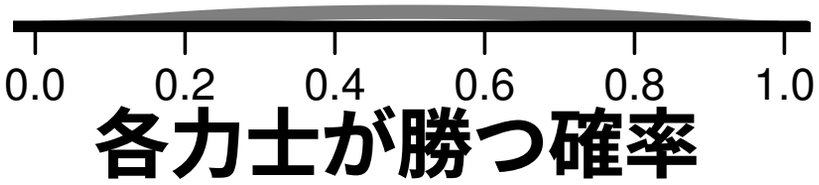




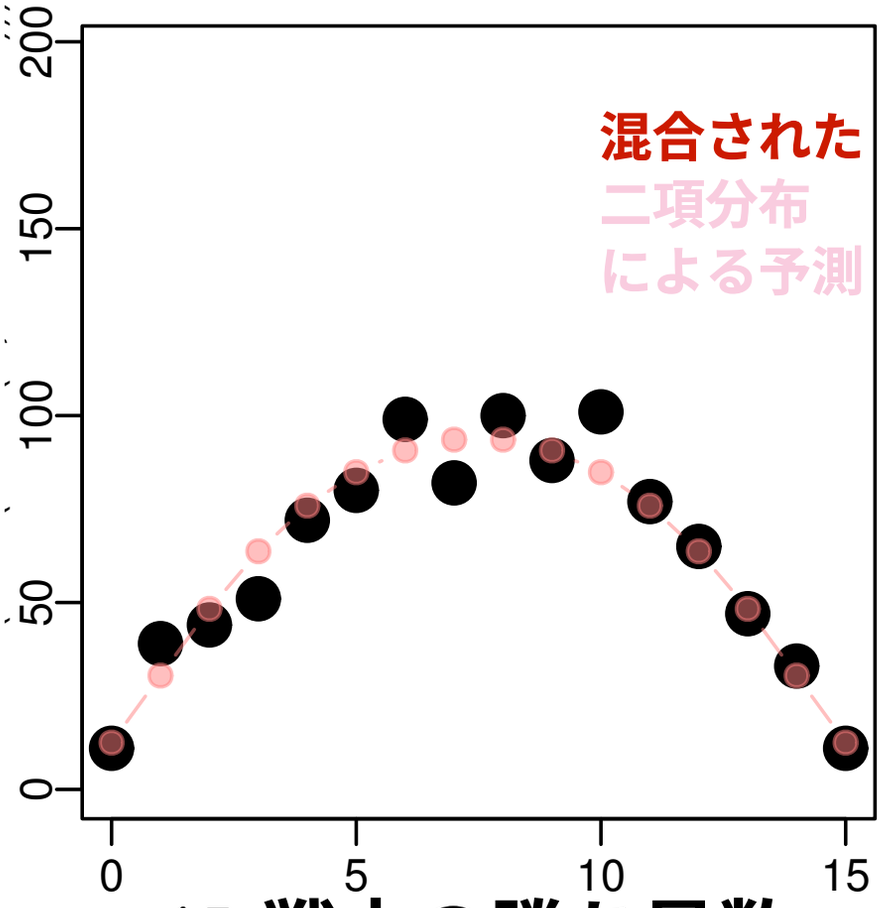
個体差を考慮するとばらつきを説明できる

勝つ確率のばらつきを仮定してみる

個体差がある



各力士が勝つ確率



15 戦中の勝ち星数

一般化線形混合モデル (GLMM) あるいは階層ベイズモデル

階層ベイズモデルの BUGS code

```
for (i in 1:N) {
  Win[i] ~ dbinom(q[i], 15) # 15 戦のうち何勝
  logit(q[i]) <- re[i]
  re[i] ~ dnorm(0, tau) #  「個体差」
}
tau <- 1 / (sigma * sigma)
sigma ~ dunif(0, 1000) # 好きな値にしろという指定
```

1. 勝ち星数 $\sim$ 二項分布 (q_i, 15 戦)
2. 勝つ確率 $q_i \leftarrow \frac{1}{1+\exp(-r_i)}$
3. 個体差 $r_i \sim$ 正規分布 (平均ゼロ, ばらつき σ)
4. ばらつき $\sigma \sim$ 正の値なら何でもいい

体重差が影響する場合の BUGS code

```
for (i in 1:N) {
  Win[i] ~ dbinom(q[i], 15) # 15 戦のうち何勝
  logit(q[i]) <- beta1 + beta2 * Size[i] + re[i]
  re[i] ~ dnorm(0, tau) # 「個体差」
}
# beta1, beta2: 好きな値にしろという指定
beta1 ~ dnorm(0, 1.0E-4) # 切片
beta2 ~ dnorm(0, 1.0E-4) # 傾き
（以下略）
```

さっきとのちがい

- **戦闘力** $\text{logit}(q_i) \leftarrow$ **(切片)** $+$ **(傾き)** $\times$ **(体格)** $+ r_i$

Bradley-Terry モデルのベイズモデル化

階層ベイズモデル化

2011–03–09　　(2011–03–09　19:32 修正版)

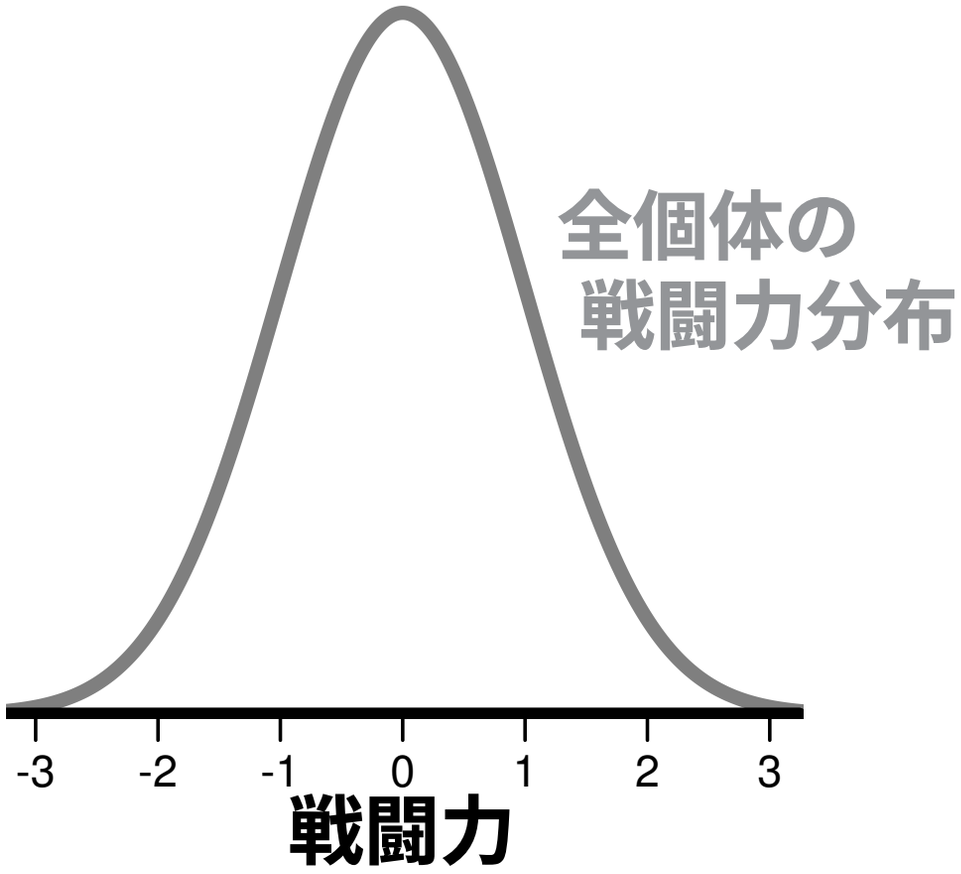
「個体差」があると認めよう

しかしそれって

実験の結果とかに何か影響あるの?

単なるばらつきでは?

個体差: 個体間の「戦闘力」に差があるとしよう



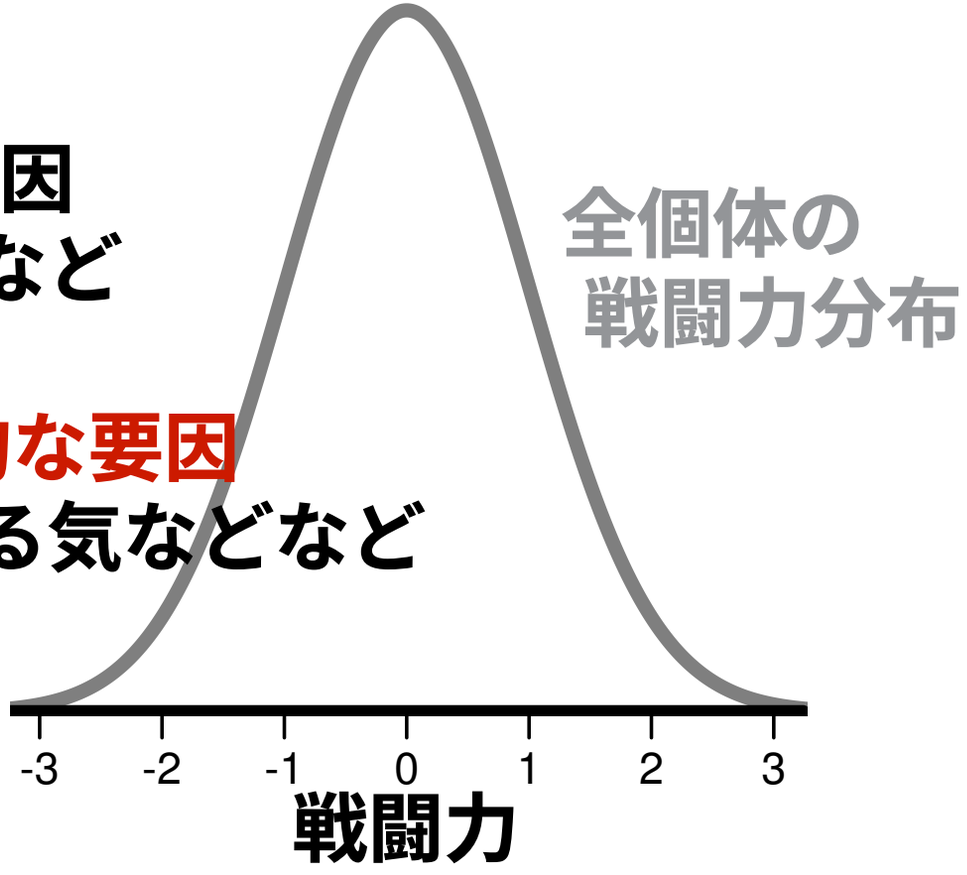
戦闘力って何? どらごんぼーる?

2011–03–09 (2011–03–09 19:32 修正版)

Bradley-Terry model: 勝敗の統計モデルのひとつ

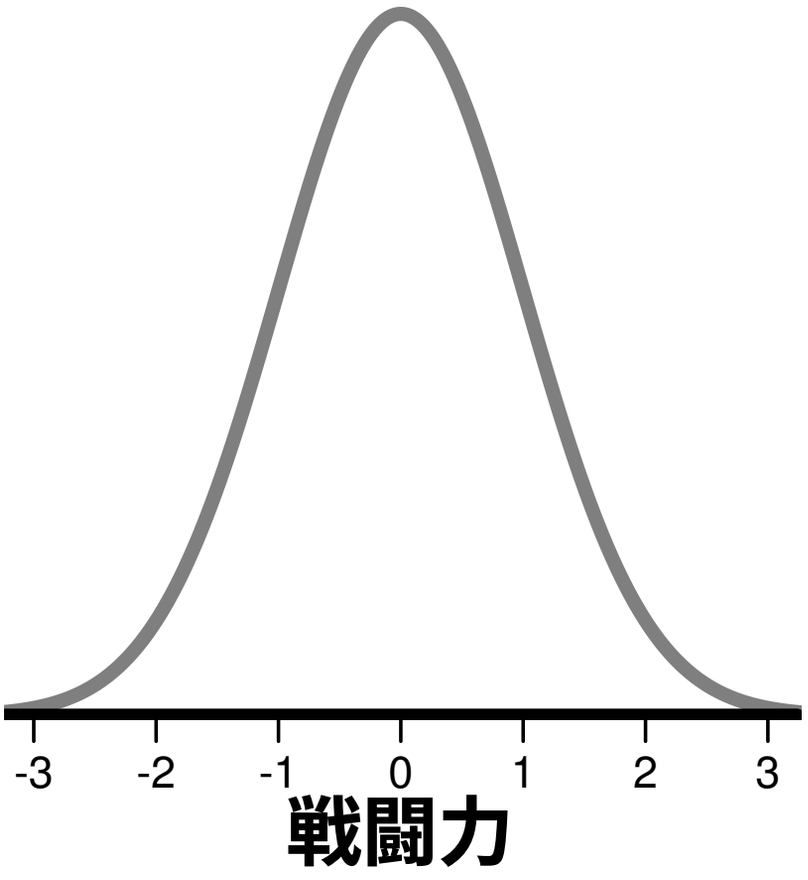
fixed effects 的な要因
　体格・年齢などなど

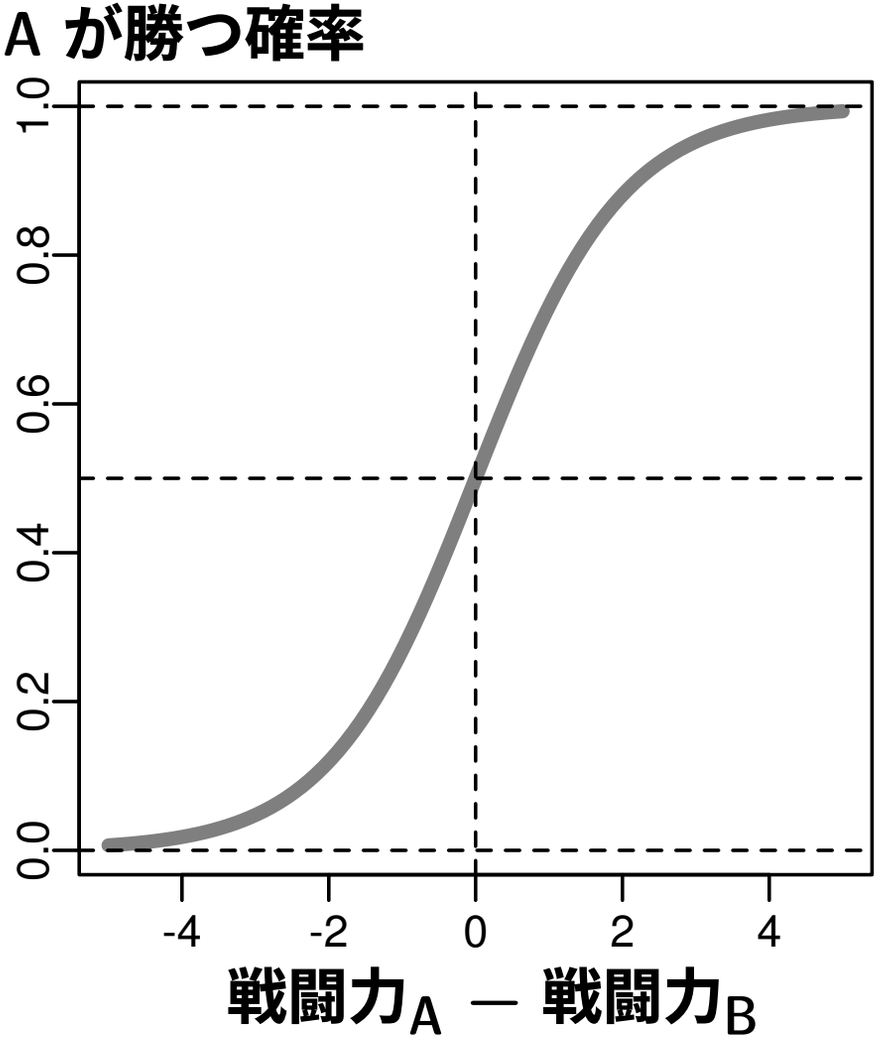
random effects 的な要因
　過去の記憶・やる気などなど



$$(\text{A が勝つ確率}) = \frac{\exp(\text{戦闘力}_A)}{\exp(\text{戦闘力}_A) + \exp(\text{戦闘力}_B)}$$

Bradley-Terry モデル





(戦闘力のモデリング例) = (切片) + (サイズなど fixed effects 的なもの)

+ (人間には直接観測が難しい random effects 的なもの)

Random effects 的な項は必要か?

従来の Bradley-Terry モデルでは個体差はどのようにあつかわれているか

- 個体差を因子型変数 (`factor` 型) としてあつかう (パラメーターを最尤推定)
  - うまくいく場合: 個体あたりの対戦数が多いとき
  - うまくいかない場合: 個体あたりの対戦数が少ないとき

階層ベイズモデル (or GLMM) 化すると?

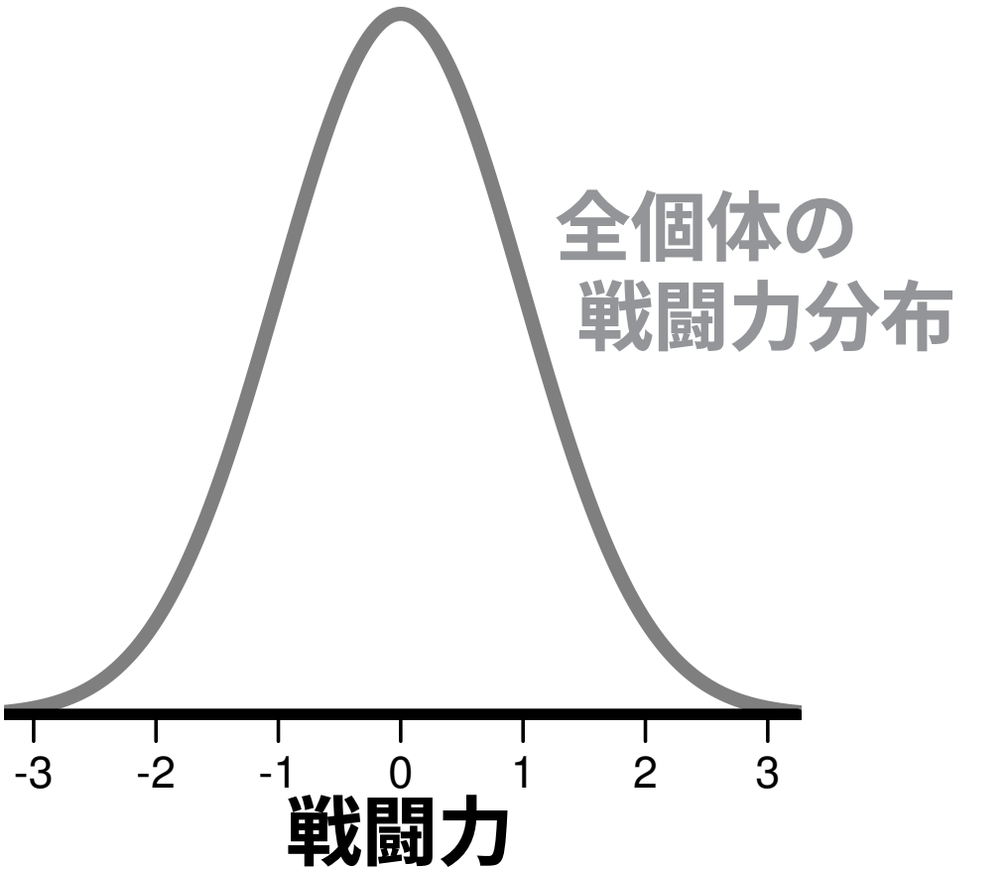
- 個体差が確率分布 (**事後分布**) として得られる
- 安全?

「個体差」の事後分布

研究者による介入の影響?

2011–03–09 (2011–03–09 19:32 修正版)

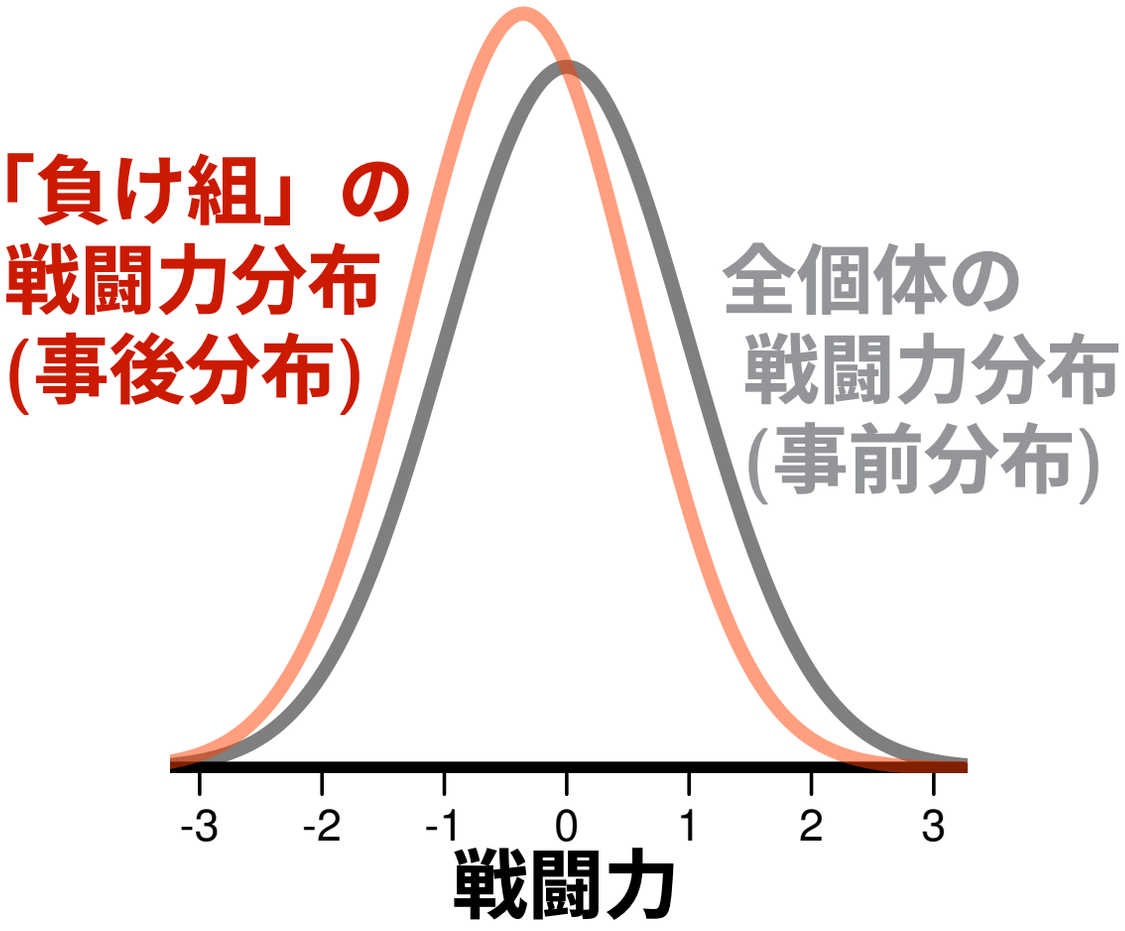
ランダムに二個体を選ぶ，戦わせるを繰り返してみる



個体間の闘争解析ではよくある実験設定，

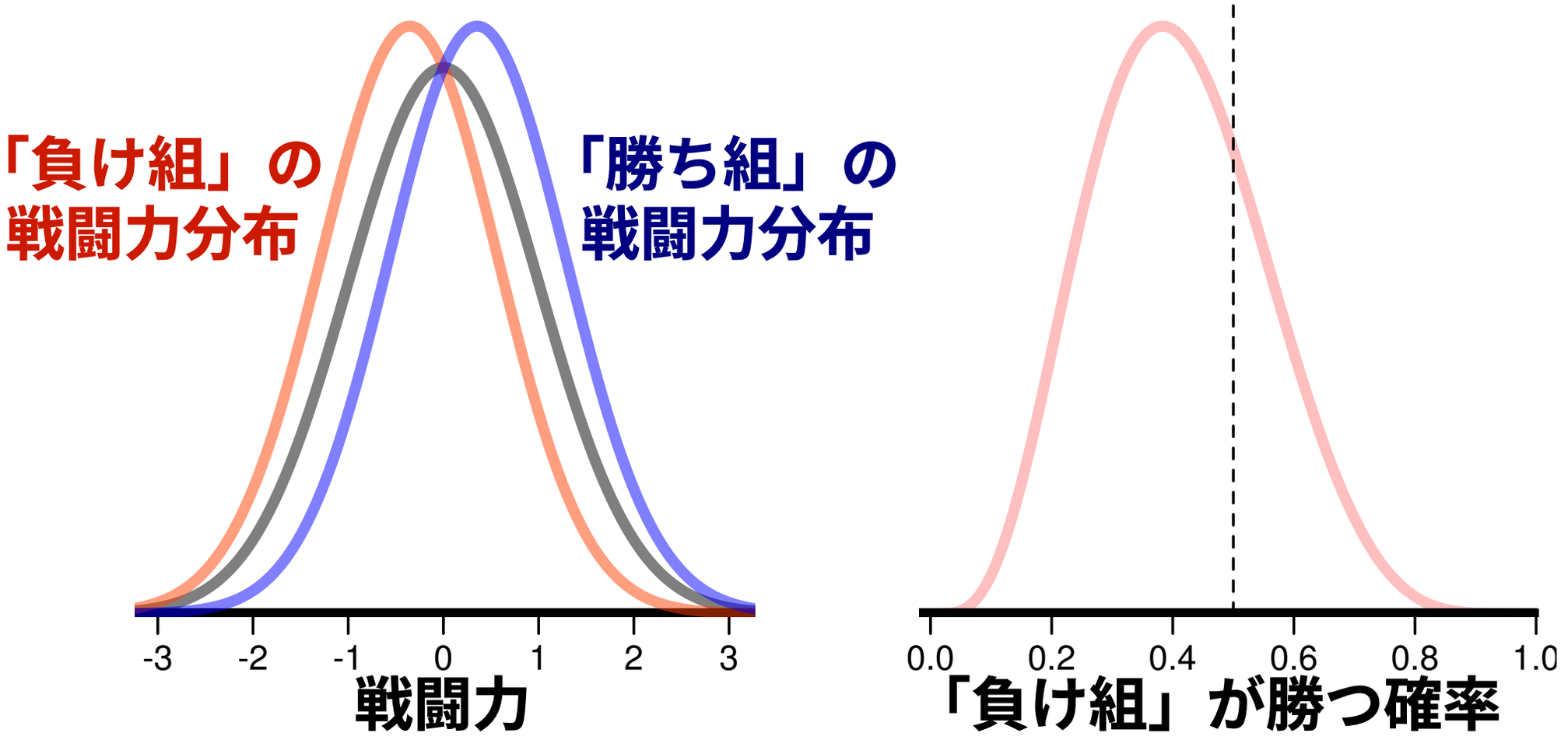
戦わせてから「勝ち組」「負け組」に分離 → 再選，「負けぐせ」の推定

「負け組」の戦闘力の事後分布



勝敗の結果をもとにグループわけするとこうなる

「負け組」が「勝ち組」に勝つ確率



アーティファクトな「負けぐせ」？ グループわけよくない??

このベイズ推定で仮定していることは何?

一回目の勝負・二回目の勝負

どんなときにも個体の

戦闘力 (内部状態) は変化しない

これってホントらしいのか?

むしろ内部状態の変化を知りたいのでは?

「変化しない」と仮定するモデルは

おもしろくない?

時間変動する「個体差」
―般化状態空間モデル

2011–03–09　　(2011–03–09　19:32 修正版)

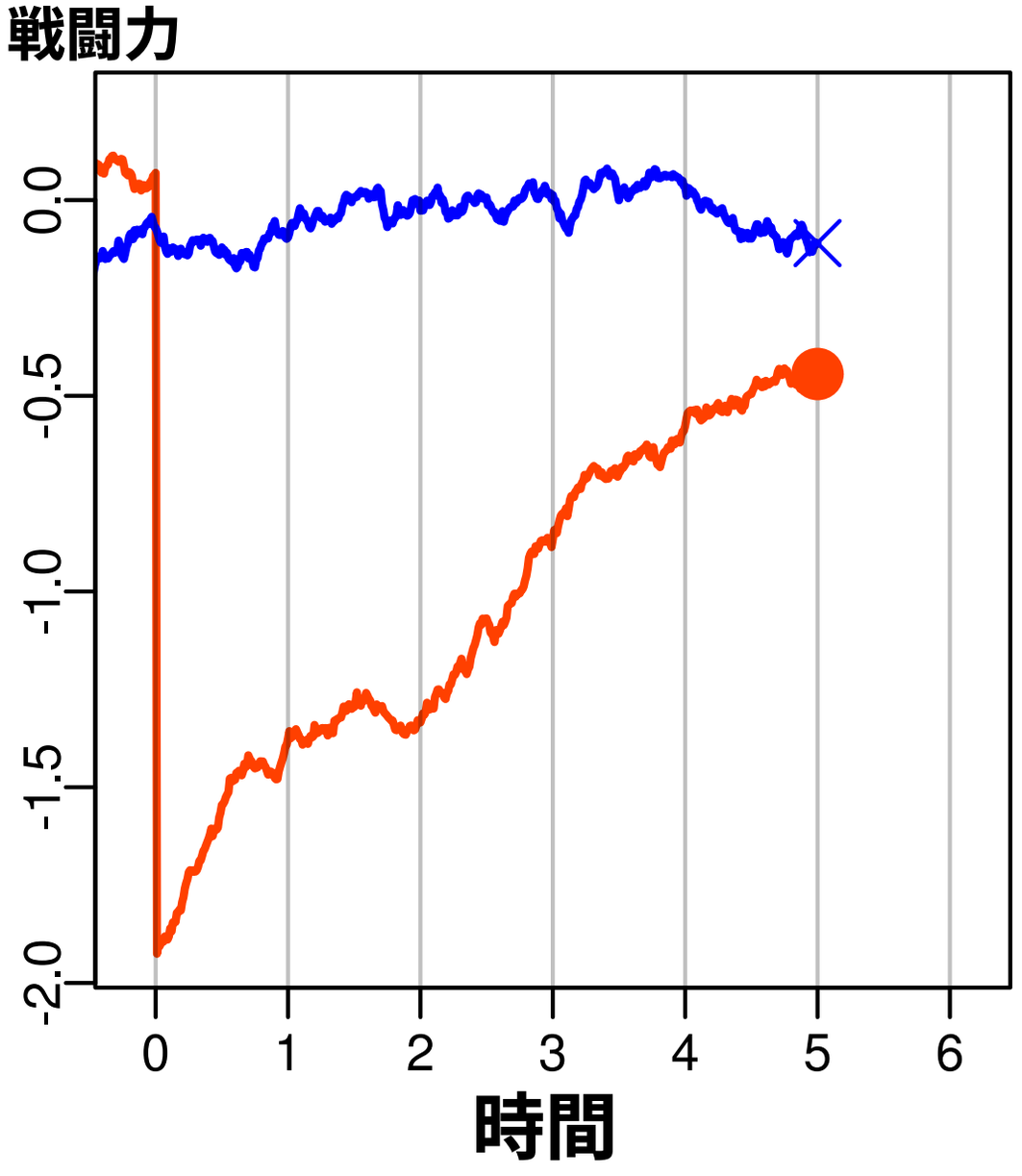
内部状態が時間変化する潜在変数

一般化状態空間モデルともよばれる

階層ベイズモデルの一種

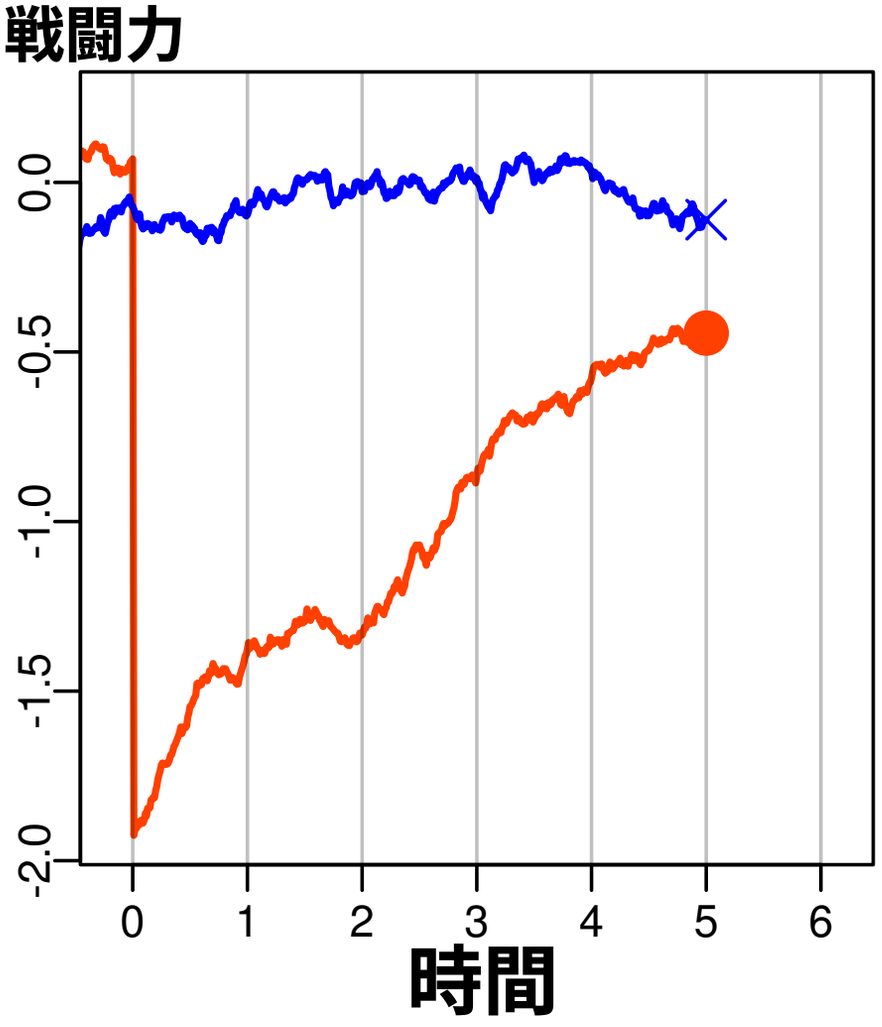
2011–03–09 (2011–03 –09 19 :32 修正版)

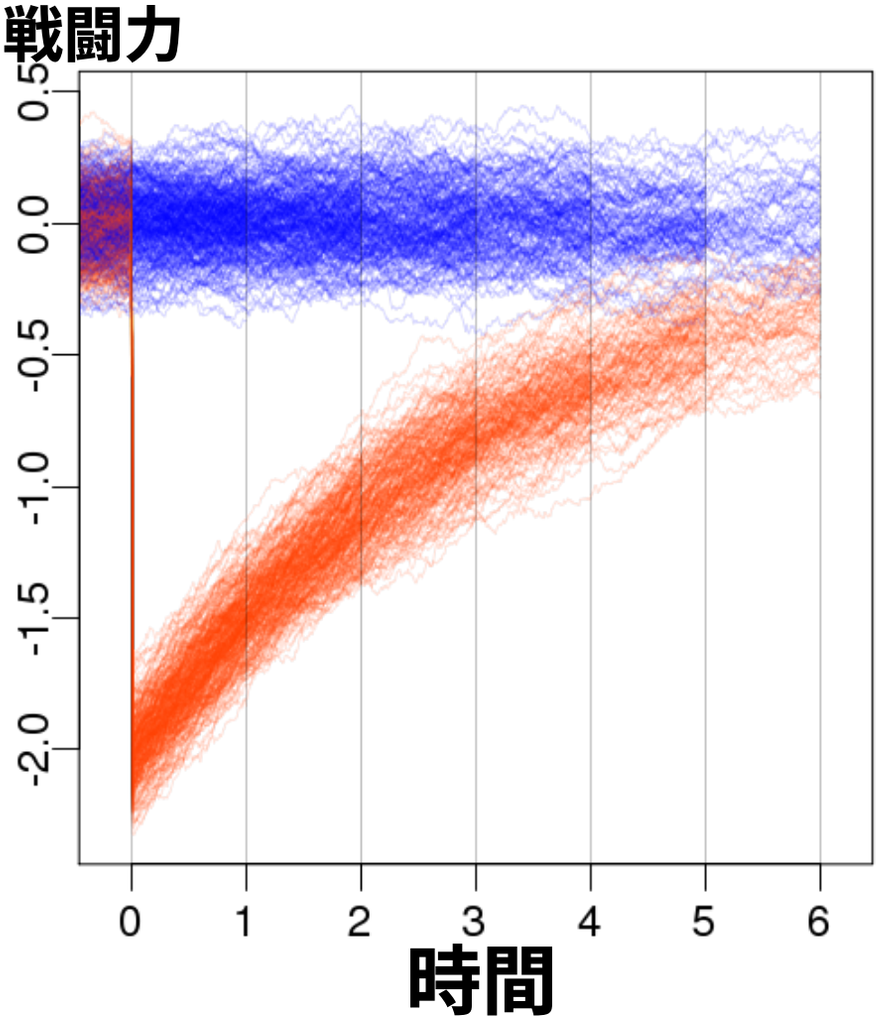
個体たちの戦闘力が時間変動するモデル



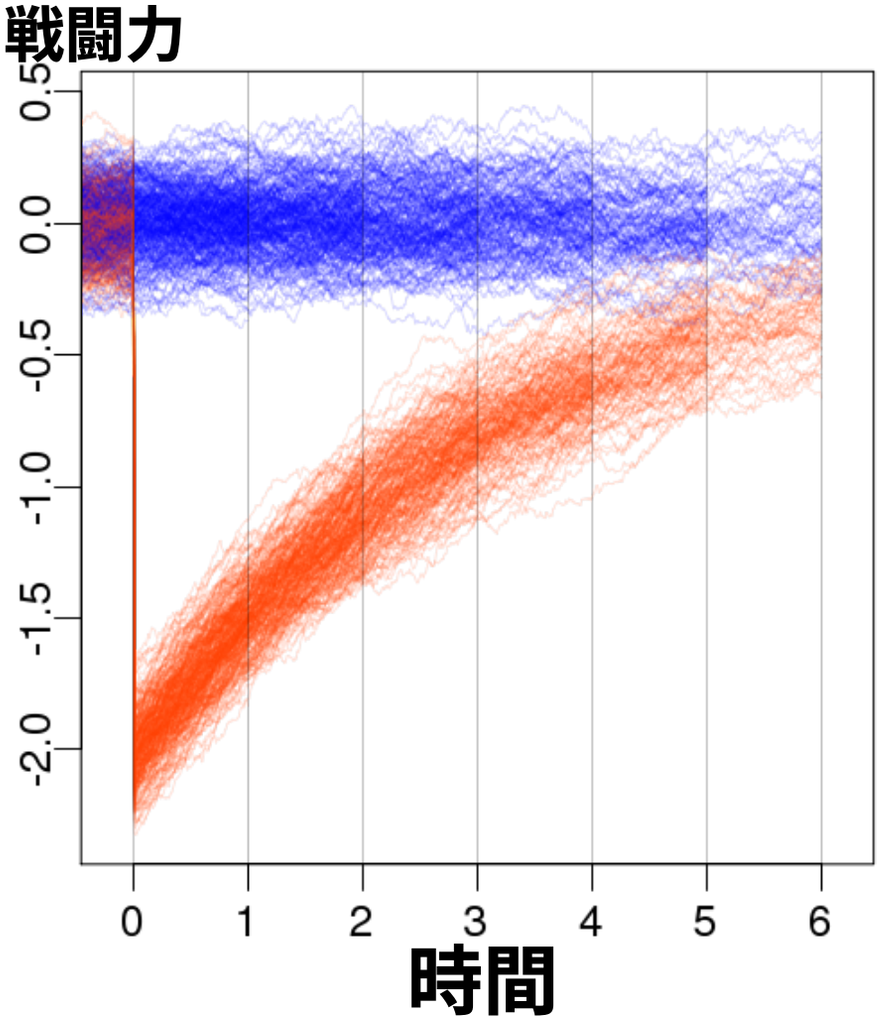
外部からのチカラで**内部状態**が変動したり

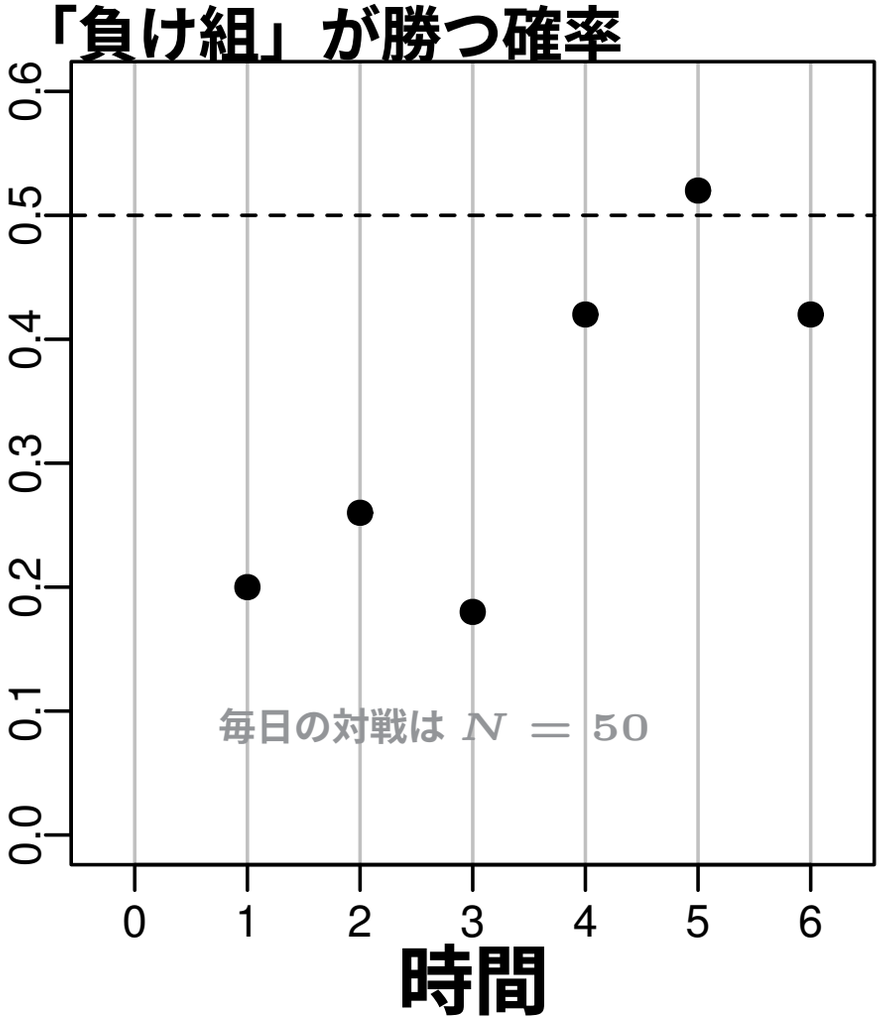
たくさんの個体を使って実験すればこうなる?





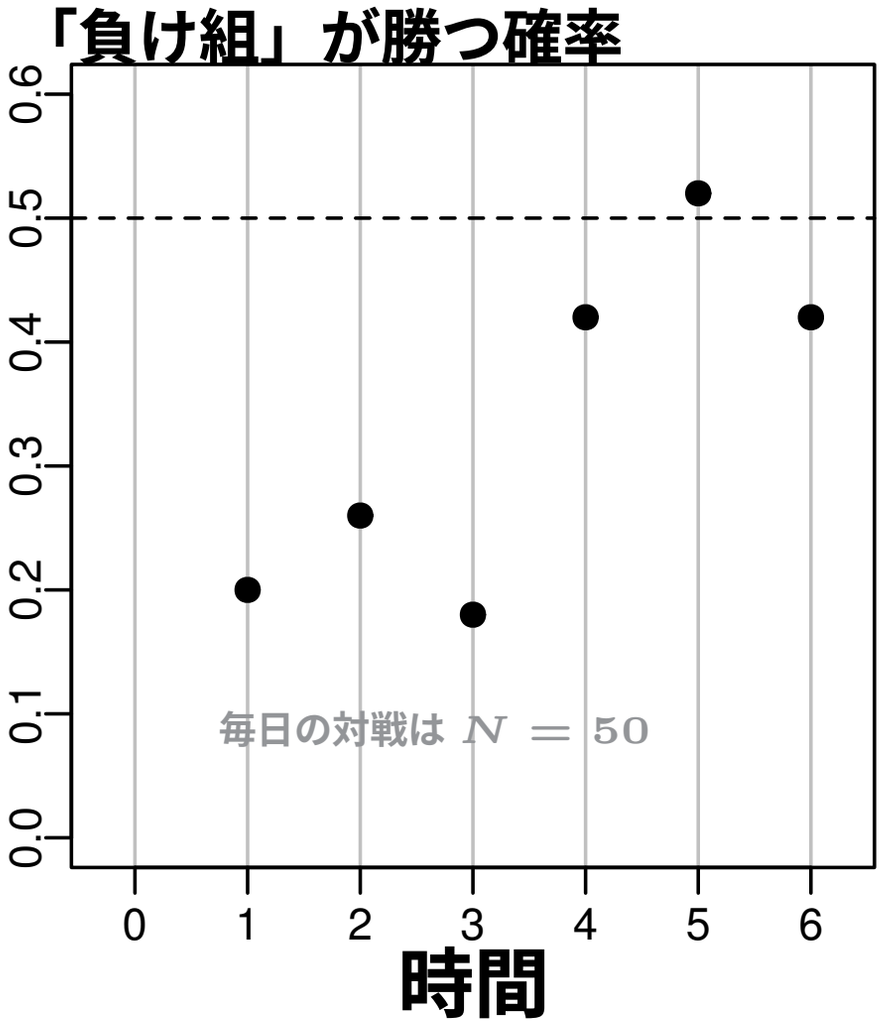
しかし観測できるのは勝敗の結果だけ……

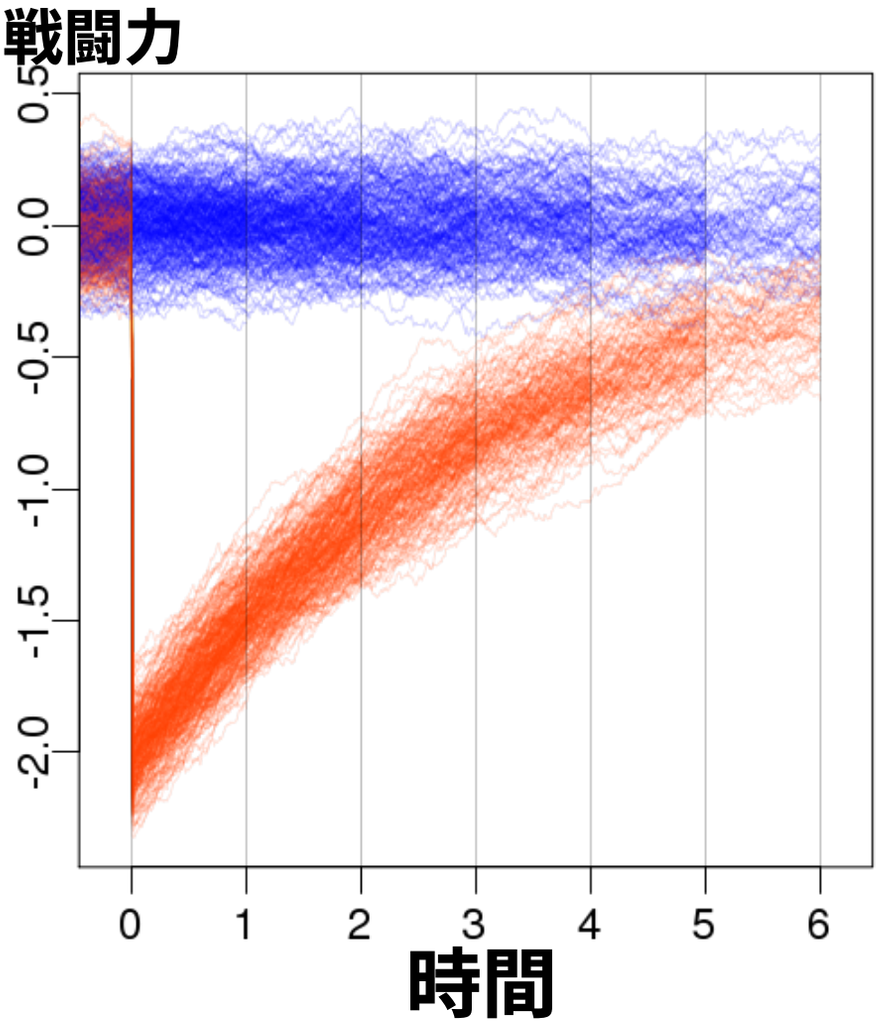




実験結果はけっこうばらつく

観測結果から個体の内部状態の変化を知りたい!





推定できるのか?

難しい……

一個体から 2 回 しか観測してないので

統計モデルにいろいろ仮定をおかざるをえない

「時間ともにどうばらつきが増大するか」とか

内部状態はゆっくり変化するといった仮定も重要

階層ベイズモデルの BUGS code

```
for (j in 1:N.loser) {
  B1 ~ dbern(win1[j])      # 1 回目の勝負（負け）
  B2[j] ~ dbern(win2[j])   # 2 回目の勝負（勝てるかも）
  # B-T model で決まる勝敗
  logit(win1[j]) <- x[T.b1, Loser[j]] - x[T.b1, Winner1[j]]
  logit(win2[j]) <- x[T.b2[j], Loser[j]] - x[T.b2[j], Winner2[j]]
  for (t in 1:T.max) {
    x[t, Loser[j]] ~ dnorm(meanL[t], 20) # ばらつき小さいと仮定
    x[t, Winner1[j]] ~ dnorm(0.0, 20)
  }
}
```

（続きは次のページ）

重要な点

- logit(勝つ確率) ← (戦闘力の差)

階層ベイズモデルの BUGS code

```
# グループ平均の時間変化
meanL[T.b1] <- 0.0
meanL[T.b1 + 1] <- ab
for (t in 3:T.max) {
  meanL[t] <- meanL[t - 1] * a # ゼロに近づく効果
}
```

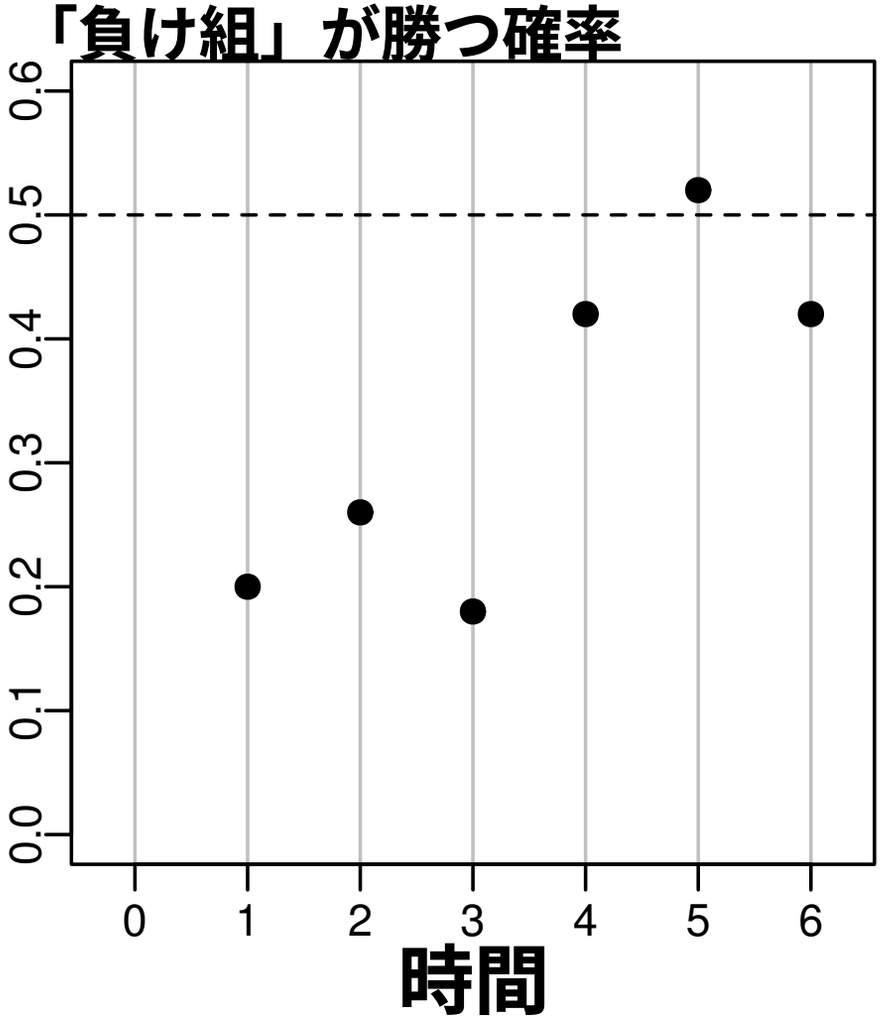
（以下略，パラメーターの事前分布の設定）

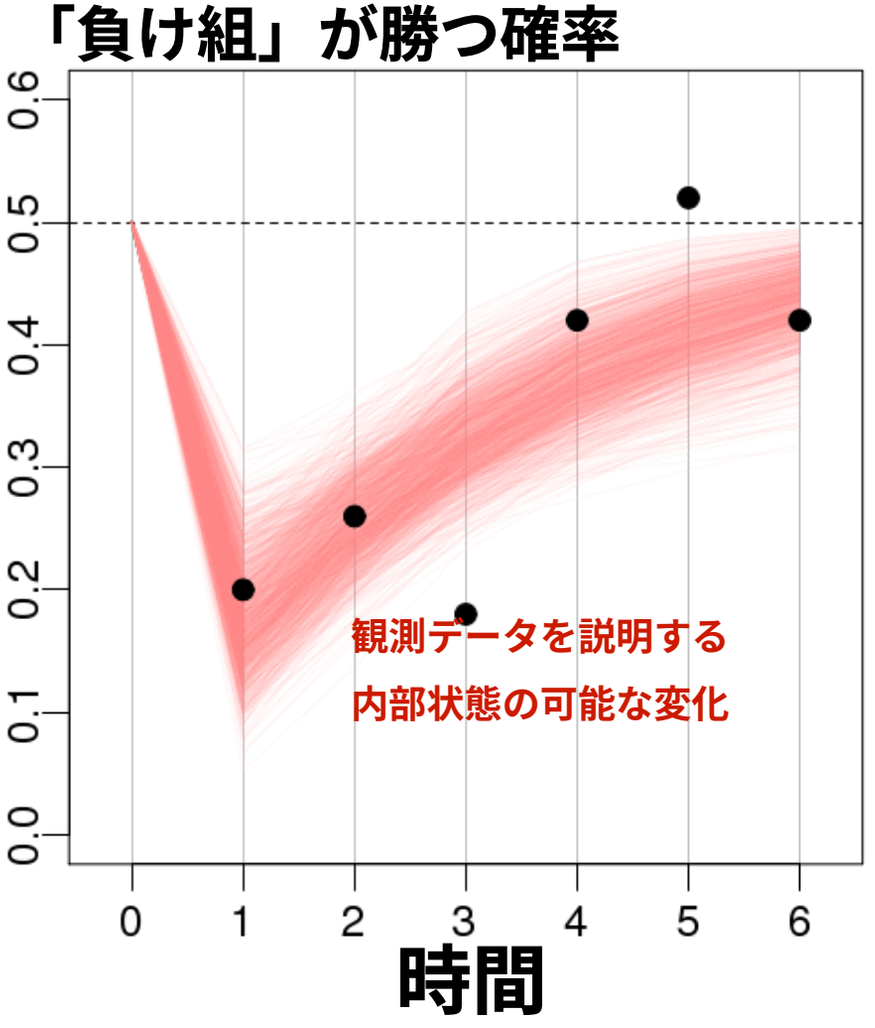
一個体から何度もくりかえし観測

「擬似反復」とよばれる状況をあつかう統計モデル

2011–03–09 (2011–03 –09 19 :32 修正版)

平均的な個体の内部状態の時間変化を推定

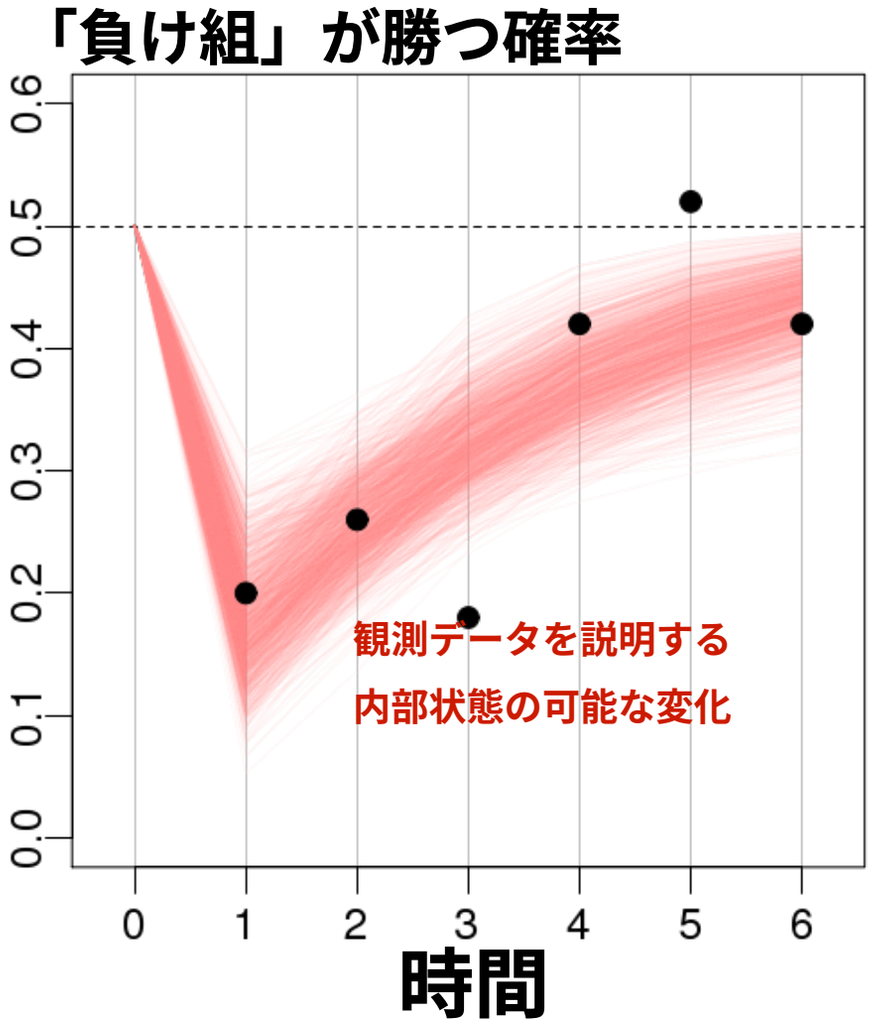


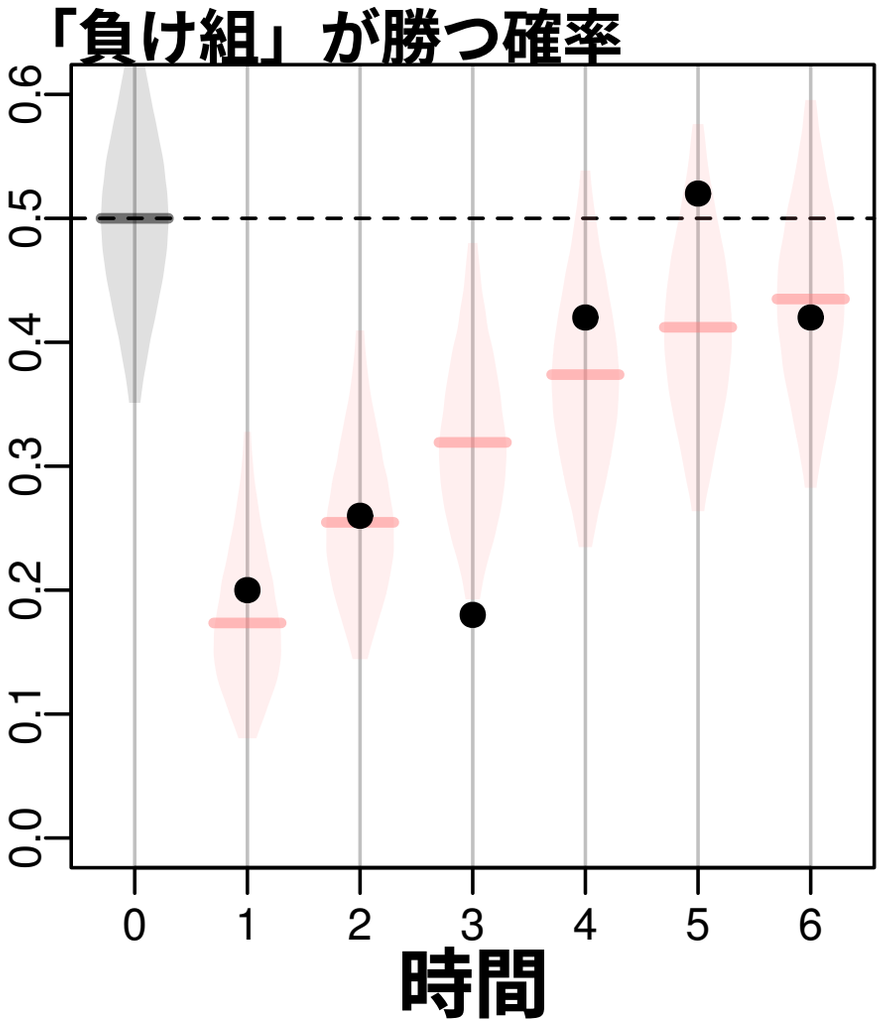


「どれぐらい落ちるか」「どう回復するか」を推定させた

ベイズ推定によって「可能性の高い」内部状態の変化がわかる

さらに個体差も考慮した「逆転勝ち」の確率





「個体差」も考慮するとさらにばらつく → 解釈に影響?

ヒヨコの内部状態の階層ベイズモデル

Anim Cogn (2010) 13:431–441
DOI 10.1007/s10071-009-0293-1

ORIGINAL PAPER

Subjective value of risky foods for individual domestic chicks: a hierarchical Bayesian model

Ai Kawamori · Toshiya Matsushima

Kawamori, A. and Matsushima, T. (2010)

- 二種類の feeder への選好性が内部状態
- 内部状態が時間変化する
- feeder 選択実験の結果を解析するためのモデル

ちょっと補足説明

B-T モデルの拡張: Burczyk モデル

2011–03–09 (2011–03 –09 19 :32 修正版)

N 個体いる場合の B-T モデルを考えよう

- 3 個体の場合

$$(\text{A が勝つ確率}) = \frac{\exp(\text{戦闘力}_\text{A})}{\exp(\text{戦闘力}_\text{A}) + \exp(\text{戦闘力}_\text{B}) + \exp(\text{戦闘力}_\text{C})}$$

- N 個体の場合

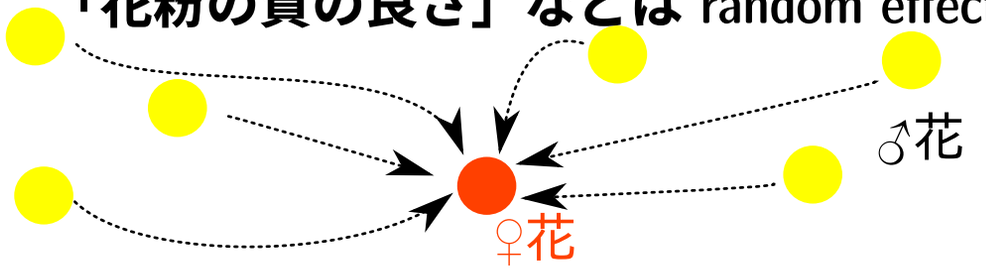
$$(\text{A が勝つ確率}) = \frac{\exp(\text{戦闘力}_\text{A})}{\sum_i^N \text{個体} \exp(\text{戦闘力}_i)}$$

例: 多数のオス花間の花粉競争

観察されること: 種子の父親の分布

オス花の「戦闘力」: 花粉量，メス花への距離，遺伝的距離……

「花粉の質の良さ」などは random effects 的かも



おわりに

勝ち負けデータ解析の雑感 (1)

- Bradley-Terry モデルや Burczyk モデルのように，統計モデルベイスで考えるのが見とおしが良いだろう
  - それに対して，データどうしの割り算値である勝率と，形質との順位相関をみるといった方法 (素朴な方法) はよくなさそう

勝ち負けデータ解析の雑感 (2)

- これらのモデルのパラメーターを最尤推定すべきか
  - まあ，とりあえずやってみたらよいかも
  - しかし random effects を導入した階層ベイズモデル化が，結局のところ便利ではないかと思います
  - とくに個体あたりの対戦数が少ないときは安全かも

    (今日のところはよい架空例を示せまんでした)
  - 相性問題,「飽和モデル」の回避?
  - うまいデータセット + モデリングがそろえば，「全勝個体」なんかもあつかえるかも

自由集会に参加していただき

ありがとうございました

今日の発表ファイルは後日

http://goo.gl/eDZLE

(プログラム参照) で配布します

自由集会に参加していただき

ありがとうございました

今日の発表ファイルは後日

http://goo.gl/eDZLE

(プログラム参照) で配布します

森本さんの「ドラえもん」は無理か?!

2011–03–09 (2011–03 –09 19 :32 修正版)