統計モデリング入門 道総研 [01]

勉強会全体の概要: 統計モデルしましょう!

久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

道総研勉強会 http://goo.gl/HQbeoh

2015–11–16

ファイル更新時刻: 2015–11–15 22:53

この時間に説明したいこと

1. **なぜ「統計モデリング入門」?**
2. **何も考えないデータ解析の問題点**
   “なんでも正規分布”とか?
3. **二日間の集中講義の概要**
   長いハナシなのでざっと全体をながめましょう

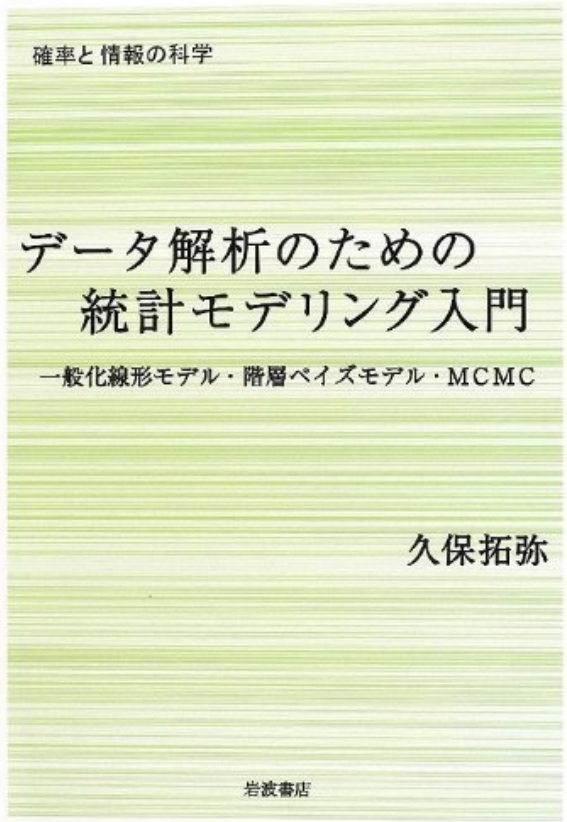
1. なぜ「統計モデリング入門」?

簡単な自己紹介: 久保拓弥 (北大・環境科学)

研究: 生態学データの統計モデリング

- 自分ではデータをとらない (野外調査・実験などをやらない) で，他のみなさんのデータ解析をすることが専門です
- これではあまりにも**寄生者**的なので，ときどきデータ解析に必要な統計モデリングの**解説ぎょーむ**などをしております……

統計モデリングの教科書も書きました!

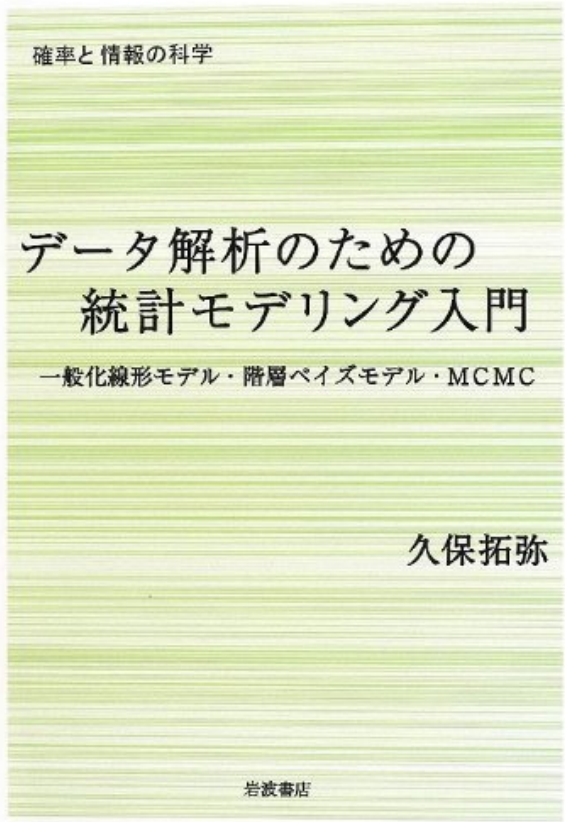


なんで，そんな本なんか書いたの?!

生態学の統計解析はあまりおもしろくなかった

- 他人の論文の method section を読んで，内容を理解しないまま同じソフトウェアを使って，$p < 0.05$ なら何でも OK といった作業になりがち
- 統計ソフトウェアが何をやっているのかわかっていないので，**誤用が多い**
- こういう発想は，計算環境が貧弱だった昔の遺物

この本ではブラックボックス統計学として批判

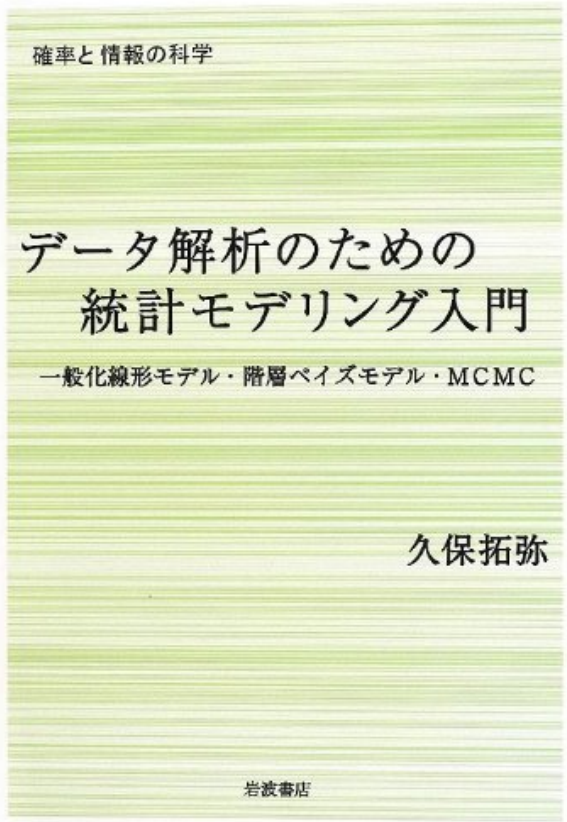


「何も考えない」データ解析はよくない

自分のデータをよくみて統計モデルを作ろう

- データはどのような確率分布にしたがうのか，あるデータのとりかたをしたときに「反復間の差」は見えるのか見えないのか?
- 現象の「背後」にあるしくみをモデル化できないか?

ちょっと　「ふつー」ではない教科書

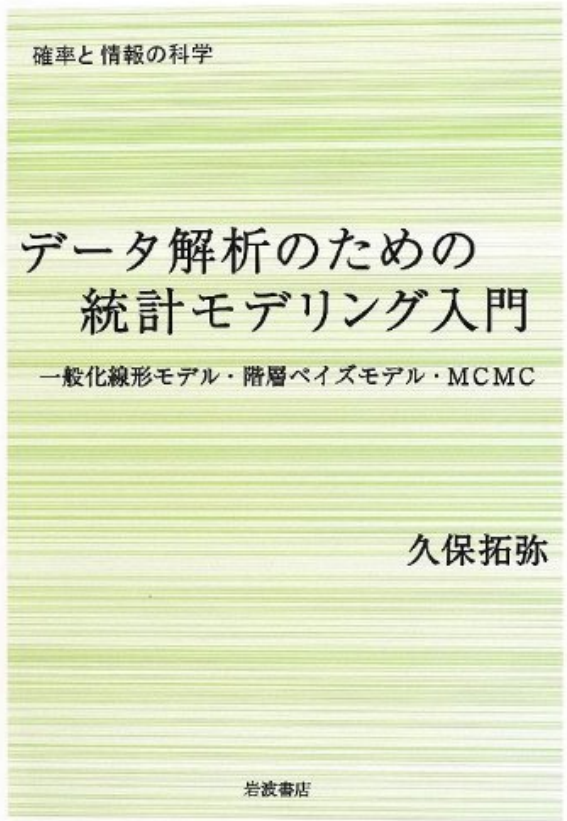


統計モデルって何?

どんな統計解析においても統計モデルが使用されている

- 観察によって**データ化された現象**を説明するために作られる
- **確率分布**が基本的な部品であり，これはデータにみられるばらつきを表現する手段である
- **データとモデルを対応づける手つづき**が準備されていて，モデルがデータにどれぐらい良くあてはまっているかを定量的に評価できる

この本では一般化線形モデルを起点に……



統計ソフトウェア R

統計学の勉強には良い統計ソフトウェアが必要!

- 無料で入手できる
- 内容が完全に公開されている
- 多くの研究者が使っている
- 作図機能が強力

この教科書でも R を
使って問題を解決する
方法を説明しています

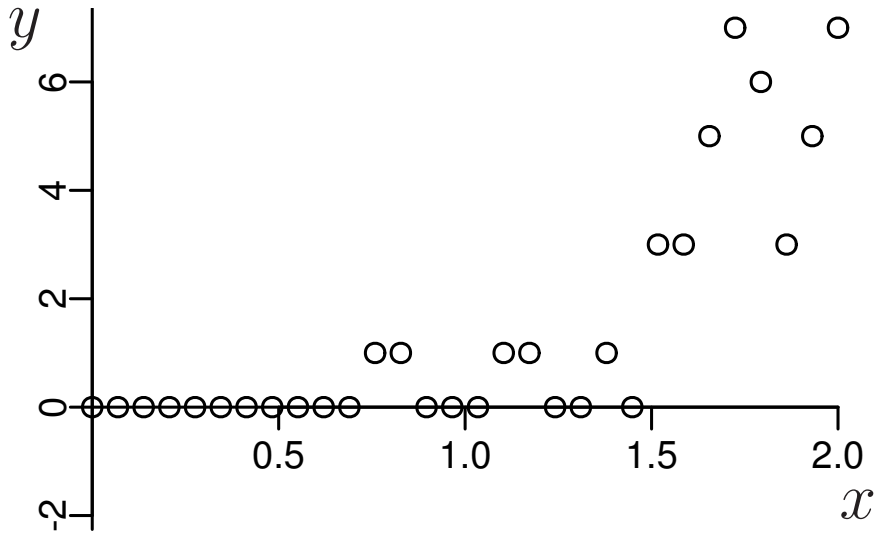
2. 何も考えないデータ解析の問題点

“なんでも正規分布”とか?

0 個, 1 個, 2 個と数えられるデータ

カウントデータ $(y \in \{0, 1, 2, 3, \cdots\}$ **なデータ**$)$

応答変数



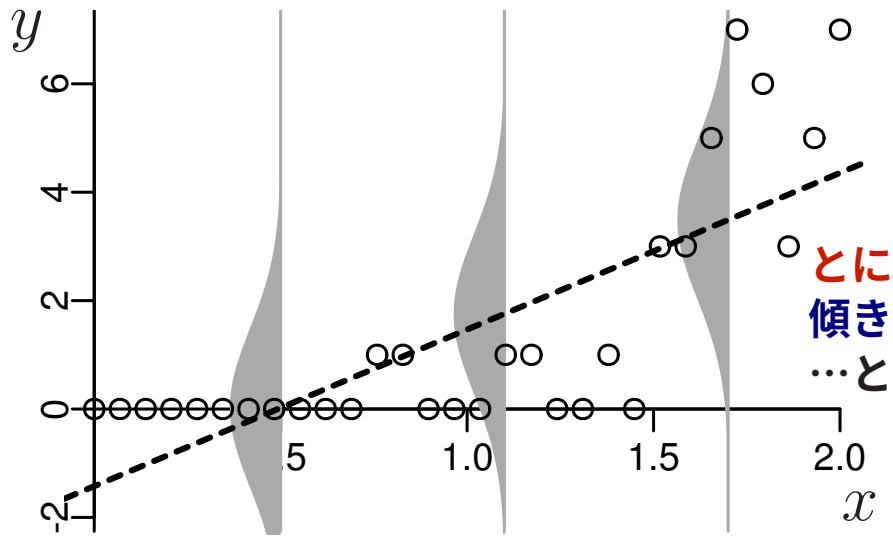
説明変数

- **たとえば** x **は植物個体の大きさ，** y **はその個体の花数**
- **体サイズが大きくなると花数が増えるように見えるが……**
- **この現象を表現する統計モデルは?**

“何でもかんでも直線あてはめ”という安易な発想……はギモン

正規分布・恒等リンク関数の統計モデル

応答変数



NO!

とにかくセンひきゃいいんでしょ
傾き「ゆーい」ならいいんでしょ
…という安易な発想のデータ解析

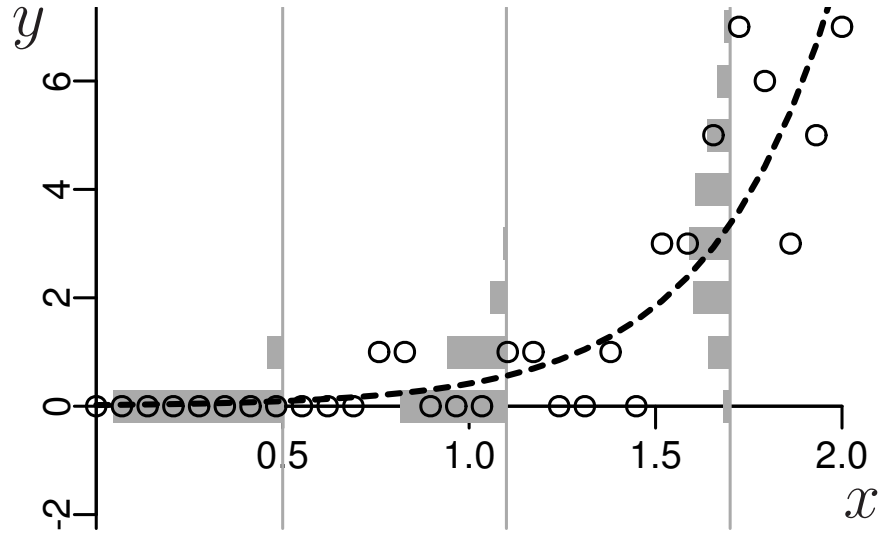
説明変数

- タテ軸のばらつきは「正規分布」なのか?
- y の値は 0 以上なのに ……
- 平均値がマイナス?

データにあわせた “統計モデル” つかうとマシかもね?

ポアソン分布・対数リンク関数の統計モデル

応答変数



YES!

説明変数

- タテ軸に対応する「ばらつき」
- 負の値にならない「平均値」
- 正規分布を使ってるモデルよりましだね

統計モデルの重要な部品: 確率分布

- データ解析をするために**統計モデル**が必要
- 統計モデルの部品として**“データにあった”確率分布**が必要
- 確率分布は**パラメーター**などを指定する必要がある
- **パラメーターの値**はデータに基づいて決めたい

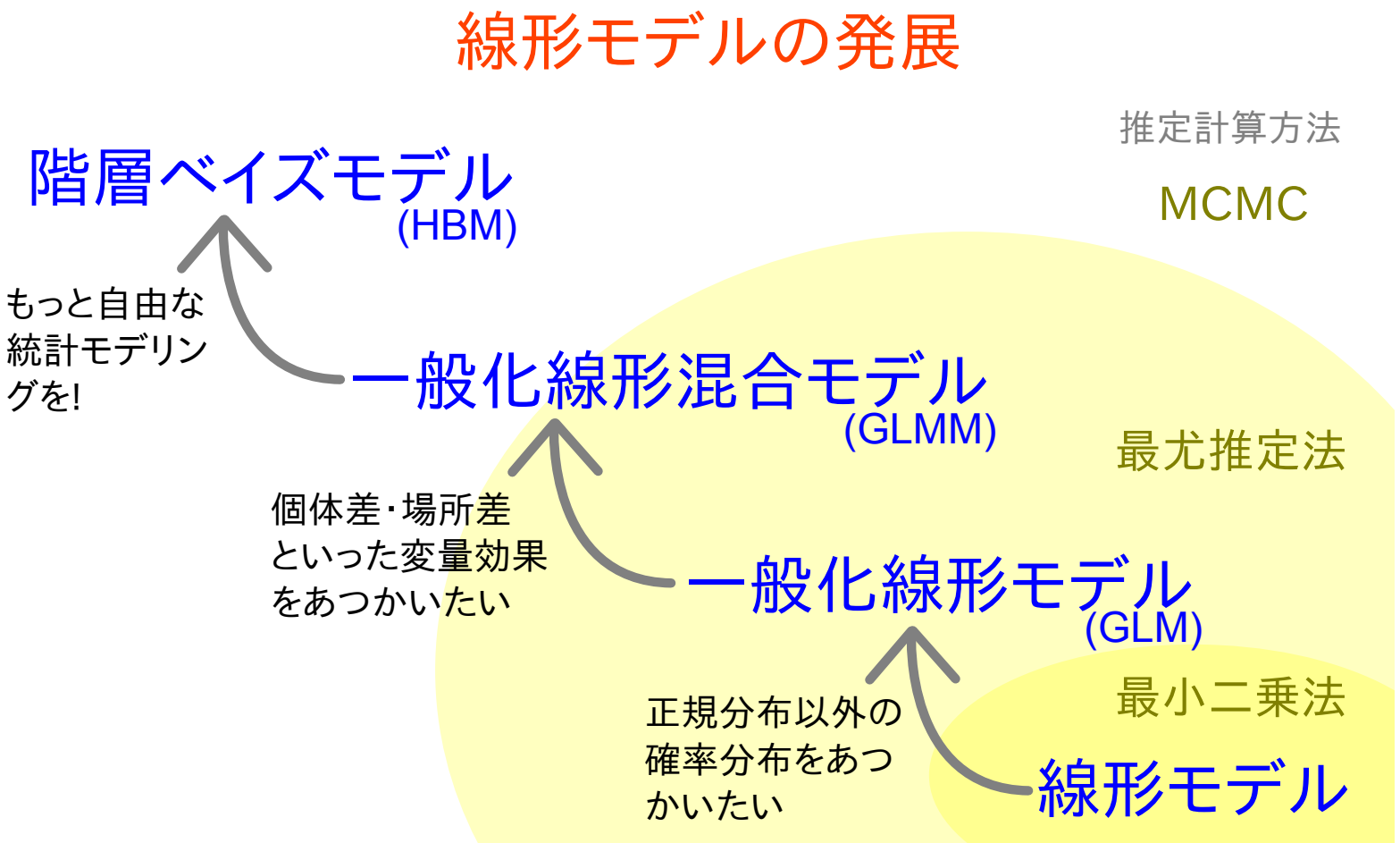
「結果 ← 原因」関係を表現する線形モデル

- 結果: 応答変数
- 原因: 説明変数
- 線形予測子 (linear predictor):

$$
\begin{aligned}
(\text{応答変数の平均}) = \quad & \text{定数 (切片)} \\
& + \quad (\text{係数 } 1) \times (\text{説明変数 } 1) \\
& + \quad (\text{係数 } 2) \times (\text{説明変数 } 2) \\
& + \quad (\text{係数 } 3) \times (\text{説明変数 } 3) \\
& + \quad \cdots
\end{aligned}
$$

(交互作用項については粕谷さんが説明してくれます)

“統計モデリング入門”に登場する統計モデル



データの特徴にあわせて線形モデルを改良・発展させる

3. 二日間の集中講義の概要

長いハナシなのでざっと全体をながめましょう

統計モデリング入門 道総研 [02]

統計モデル・確率分布・最尤推定

久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

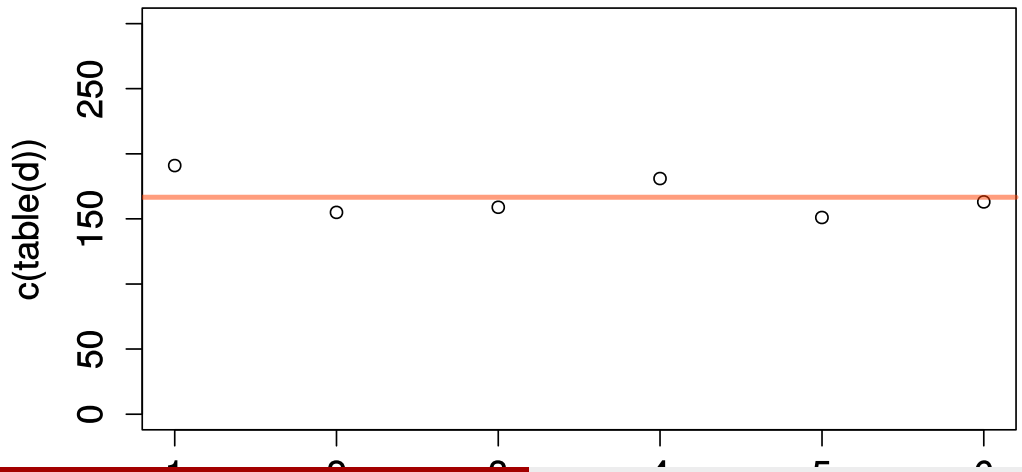
道総研勉強会 http://goo.gl/HQbeoh

2015–11–16

ファイル更新時刻: 2015–11–12 22:11

「サイコロの統計モデル」を考えよう

```
> load("dice.RData")
> length(d)
[1] 1000
> table(d)
d
  1   2   3   4   5   6
191 155 159 181 151 163
> plot(1:6, c(table(d)), ylim = c(0, 300))
> abline(h = 1000 / 6, col = "#ff400080", lwd = 3)
```



架空データ

1000回サイコロふった

1000/6 = 166.66...?

R でデータの様子をながめる

の table() 関数を使って種子数の頻度を調べる

```
> table(data)
 0  1  2  3  4  5  6  7
 1  3 11 12 10  5  4  4
```

(種子数 5 は 5 個体，種子数 6 は 4 個体 ……)

確率分布（ポアソン分布）を数式で決めてしまう

種子数が y である確率は以下のように決まる，と考えている

$$p(y \mid \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}$$

- **$y!$ は y の階乗で，たとえば $4!$ は $1 \times 2 \times 3 \times 4$ をあらわしています．**
- **$\exp(-\lambda) = e^{-\lambda}$ のこと** $(e = 2.718\cdots)$
- **ここではなぜポアソン分布の確率計算が上のようになるのかは説明しません— まあ，こういうもんだと考えて先に進みましょう**

統計モデリング入門 道総研 [03]

R の練習: 次の時間の例題データ

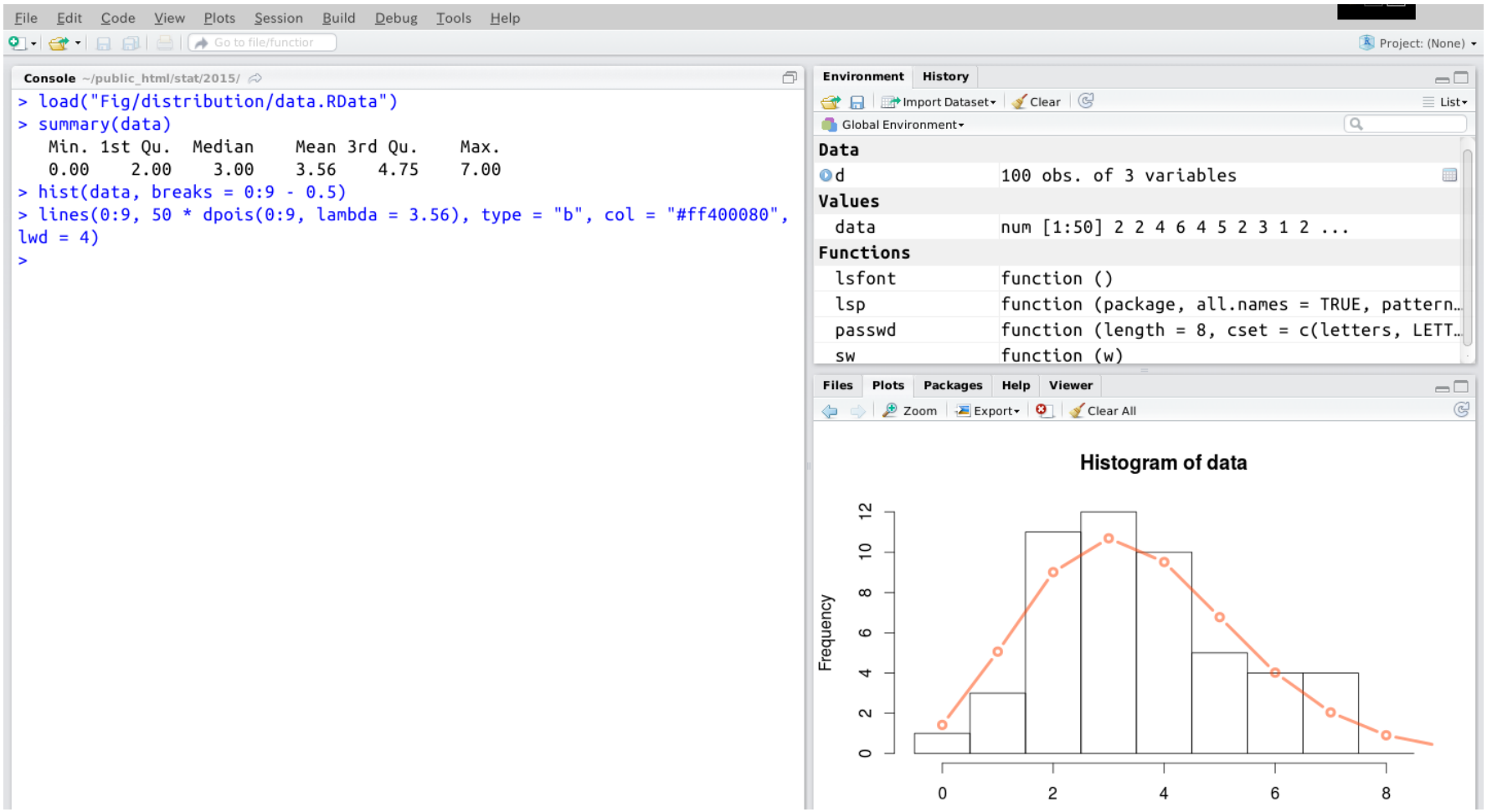
久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

道総研勉強会 http://goo.gl/HQbeoh

2015–11–16

ファイル更新時刻: 2015–11–15 22:33

RStudio 使ってみますかね?



統計モデリング入門 道総研 [04]

ポアソン分布の一般化線形モデル (GLM)

久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

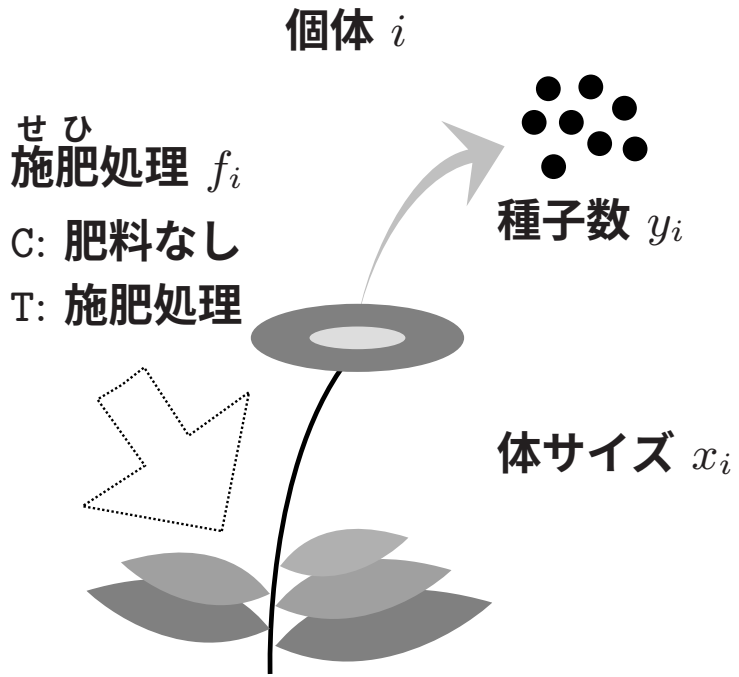
道総研勉強会 http://goo.gl/HQbeoh

2015–11–16

ファイル更新時刻: 2015–11–12 22:11

個体サイズと実験処理の効果を調べる例題

- **応答変数: 種子数** $\{y_i\}$
- **説明変数:**
  - **体サイズ** $\{x_i\}$
  - **施肥処理** $\{f_i\}$



標本数

- **無処理** ($f_i = \mathtt{C}$): 50 sample ($i \in \{1, 2, \cdots 50\}$)
- **施肥処理** ($f_i = \mathtt{T}$): 50 sample ($i \in \{51, 52, \cdots 100\}$)

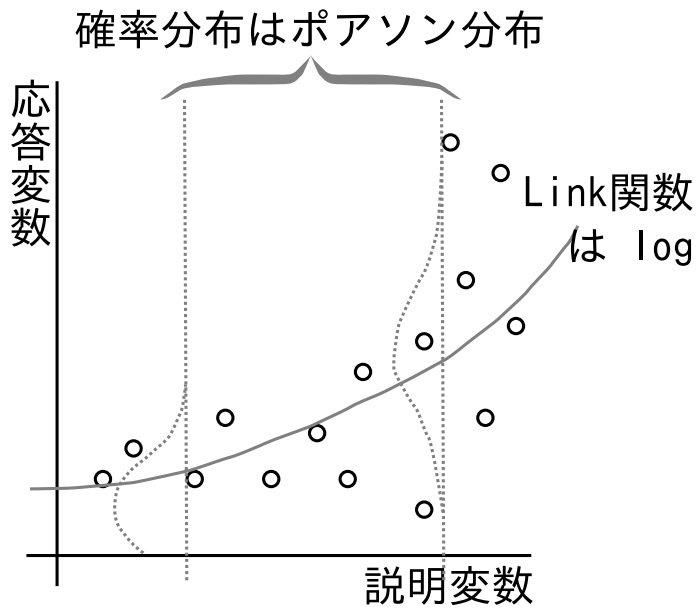
3. R で GLM のパラメーターを推定

あてはまりの良さは対数尤度関数で評価

推定計算はコンピューターにおまかせ

glm() 関数の指定を再確認

- family: poisson, **ポアソン分布**
- link **関数**: "log"
- **モデル式** (**線形予測子** z): **たとえば** y ~ x **と指定したとする**



- **線形予測子** $z = \beta_1 + \beta_2 x$
  β_1, β_2 **は推定すべきパラメーター**
- **応答変数の平均値を** λ **とすると** $\log(\lambda) = z$
  つまり $\lambda = \exp(z) = \exp(\beta_1 + \beta_2 x)$
- **応答変数 は平均** λ **のポアソン分布に従う**: $y \sim \mathrm{Pois}(\lambda)$

統計モデリング入門 道総研 [05]

モデル選択と検定

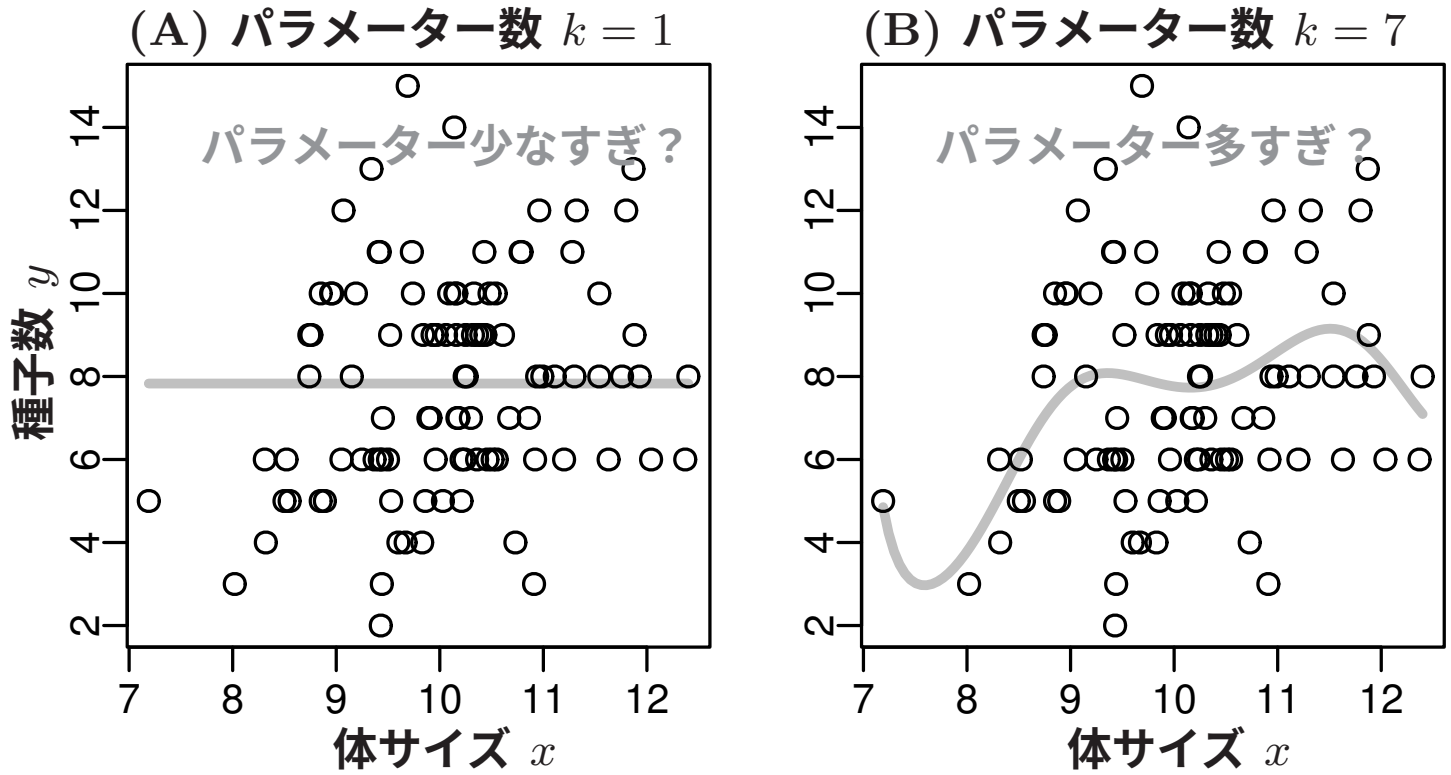
久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

道総研勉強会 http://goo.gl/HQbeoh

2015–11–16

ファイル更新時刻: 2015–11–15 21:10

パラメーター数 k は多くても少なくてもヘン?



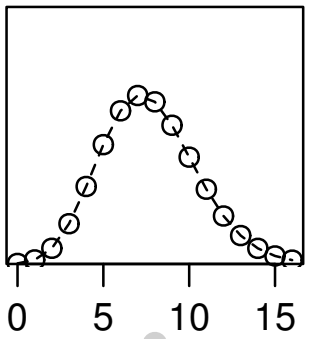
“良いモデル” とはなにか? k も重要なのか?

R の glm() は deviance を出力

```
> glm(y ~ x + f, data = d, family = poisson)

Call:  glm(formula = y ~ x + f, family = poisson, data = d)

Coefficients:
(Intercept)            x           fT
     1.2631       0.0801      -0.0320

Degrees of Freedom: 99 Total (i.e. Null);  97 Residual
Null Deviance:      89.5
Residual Deviance: 84.8         AIC: 477
```

Residual Deviance? Null Deviance? AIC?

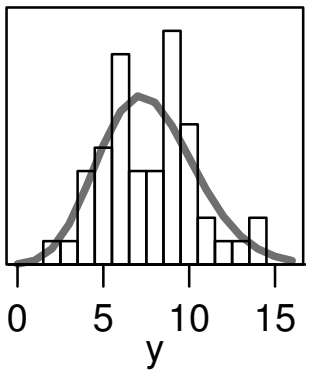
推定に使ったデータであてはまりを評価している?



観測データから
推定された constant λ
$\hat{\beta}_1 = 2.04$ のポアソン分布

推定用の観測データを使って
あてはまりの良さを評価

すると最大対数尤度
$\log L^*$ が得られる



推定用の観測データ

パラメーター推定に使った
データなのであてはまりの
良さにバイアスが生じる
(過大評価)

しかしモデル選択と検定の手順は途中まで同じ

統計モデルの検定 | **AIC によるモデル選択**

解析対象のデータを確定

⇓

データを説明できるような統計モデルを設計

(帰無仮説・対立仮説) | **(単純モデル・複雑モデル)**

⇓

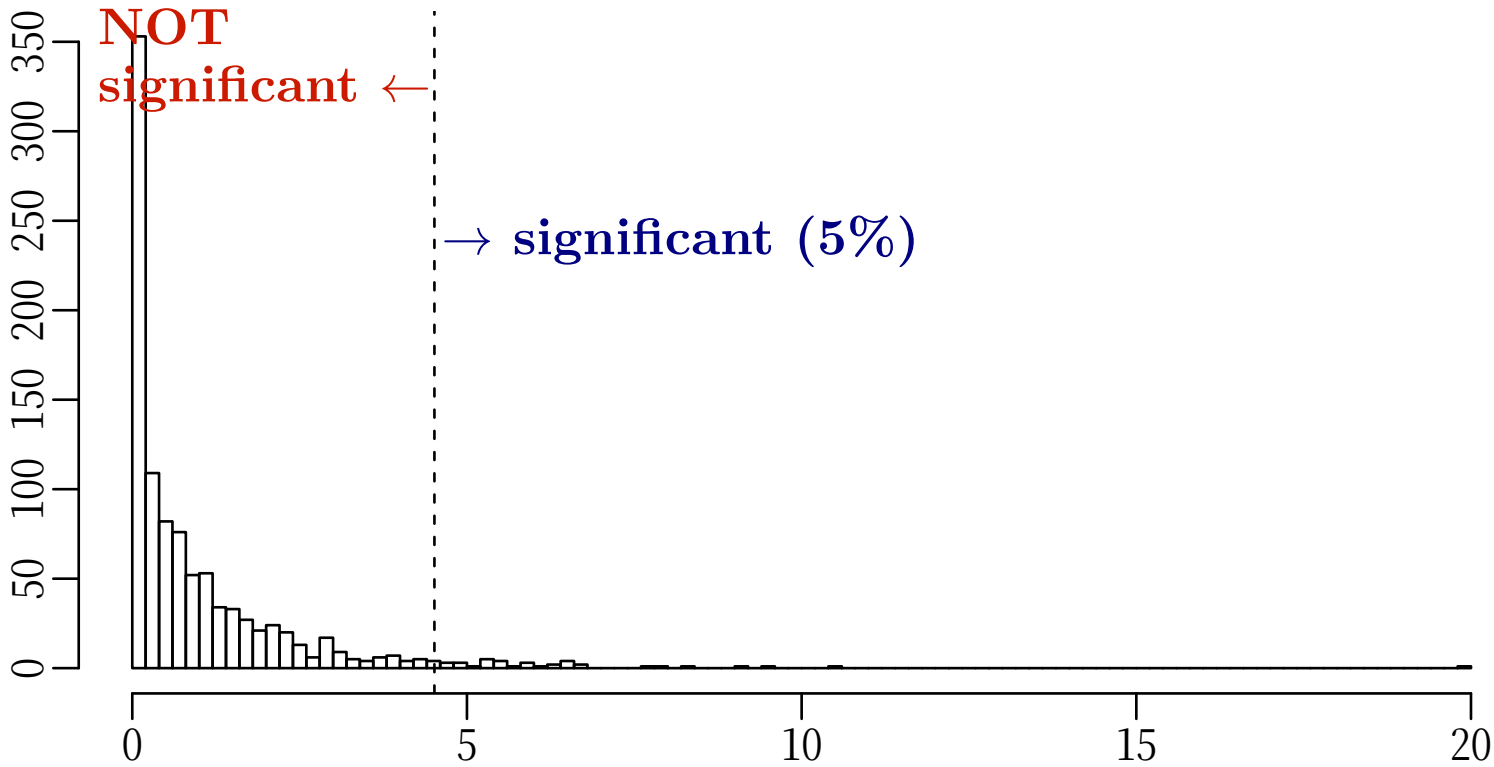
ネストした統計モデルたちのパラメーターの最尤(さいゆう)推定計算

⇓ | ⇓

帰無仮説棄却の危険率を評価 | **モデル選択規準 AIC の評価**

あらかじめ棄却域を決めておく

たとえば 5% **とか**?



統計モデリング入門 道総研 [06]

“割算” 回避のための統計モデル

久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

道総研勉強会 http://goo.gl/HQbeoh

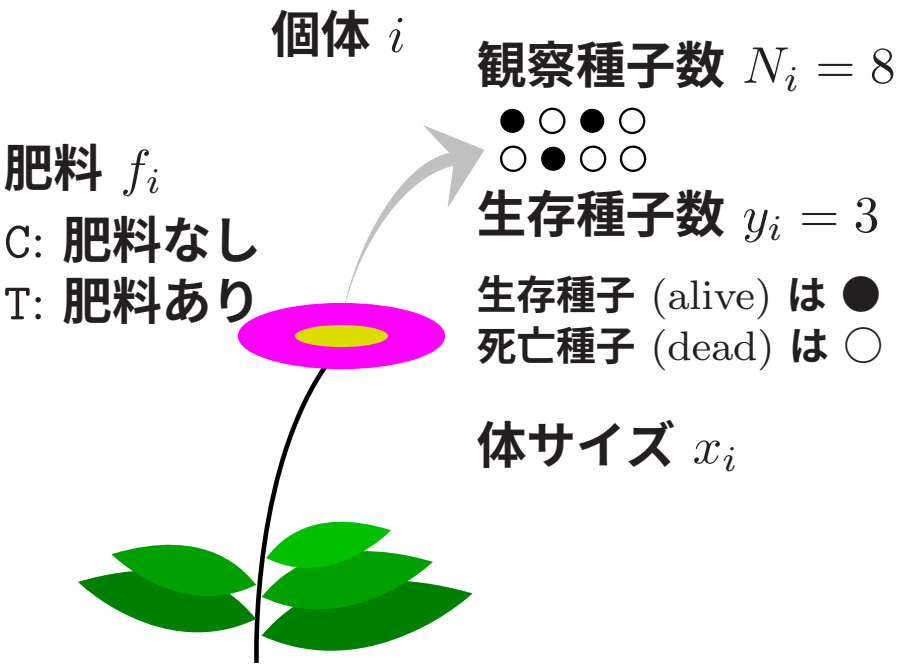
2015–11–16

ファイル更新時刻: 2015–11–12 22:11

“N 個のうち k 個が生きてる” タイプのデータ | $y_i \in \{0, 1, 2, \cdots, 8\}$

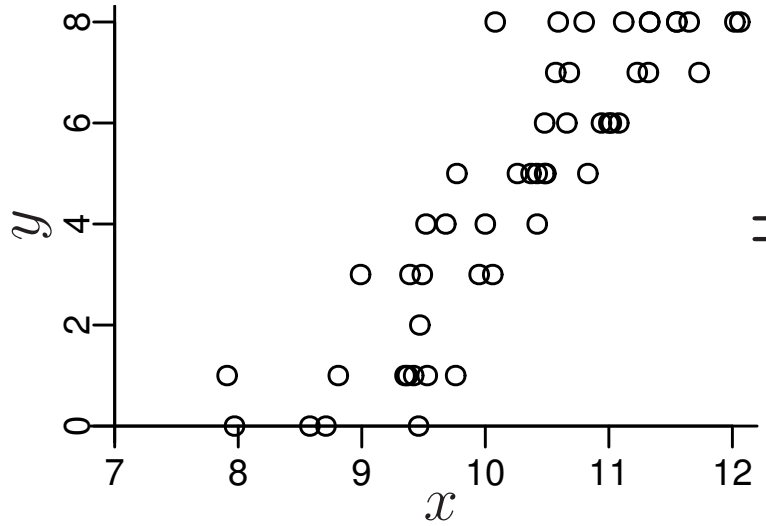
またいつもの例題? …… ちょっとちがう

8 個の種子のうち y 個が **発芽可能** だった! …… というデータ

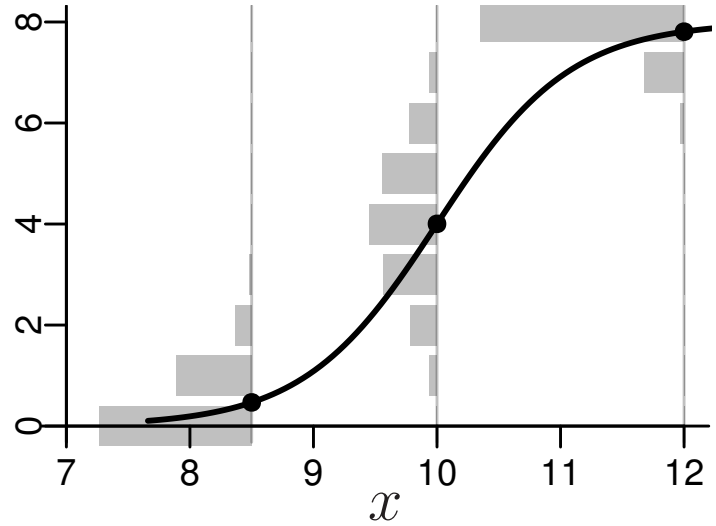


R でロジスティック回帰 — β_1 と β_2 の最尤推定

(A) 例題データの一部（f_i =C）



(B) 推定されるモデル




```
> glm(cbind(y, N - y) ~ x + f, data = d, family = binomial)
...
Coefficients:
(Intercept)            x           fT
    -19.536        1.952        2.022
```

統計モデルを工夫してわりざんやめよう

- **避けられる割算値**
  - **確率**

    例: N 個のうち k 個にある事象が発生する確率

    対策: ロジスティック回帰など**二項分布モデル**で
  - **密度などの指数**

    例: 人口密度，specific leaf area (SLA) など

    対策: **offset 項わざ** — 統計モデリングの工夫!

統計モデリング入門 道総研 [07]

一般化線形混合モデル (GLMM)

久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

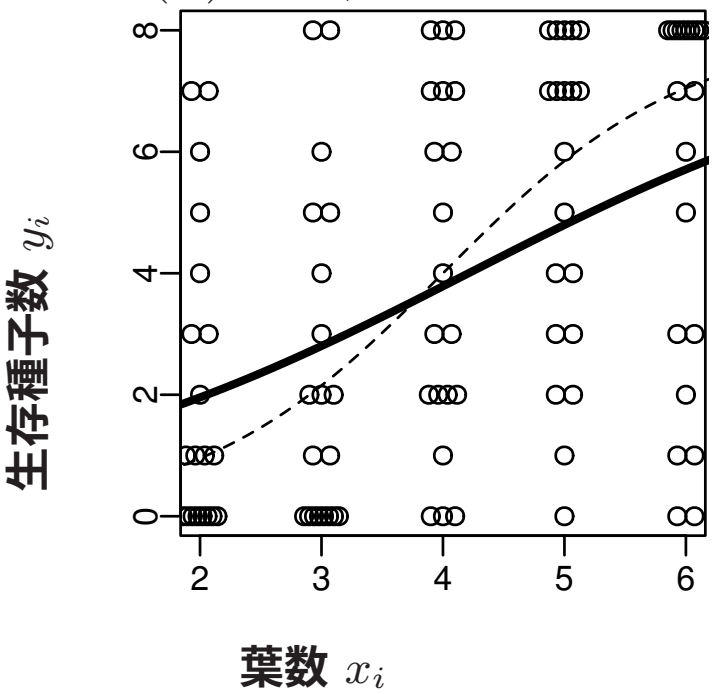
道総研勉強会 http://goo.gl/HQbeoh

2015–11–17

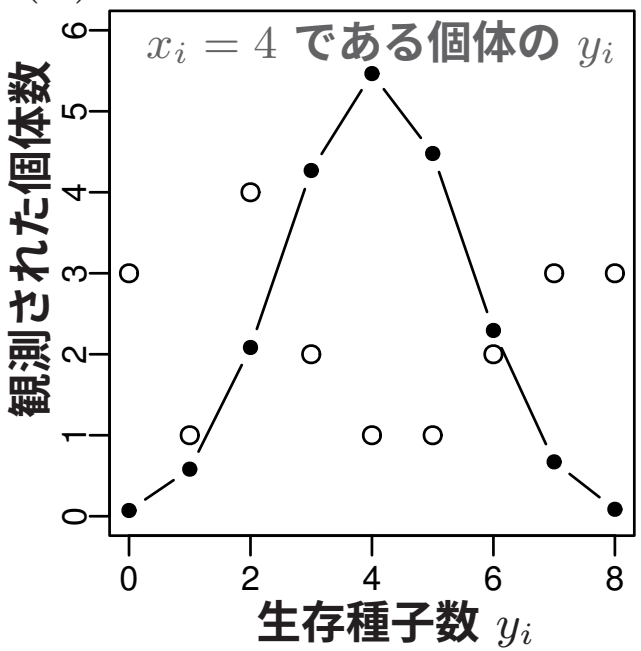
ファイル更新時刻: 2015–11–12 22:11

GLM では説明できないばらつき!

(A) 傾き β_2 が過小推定

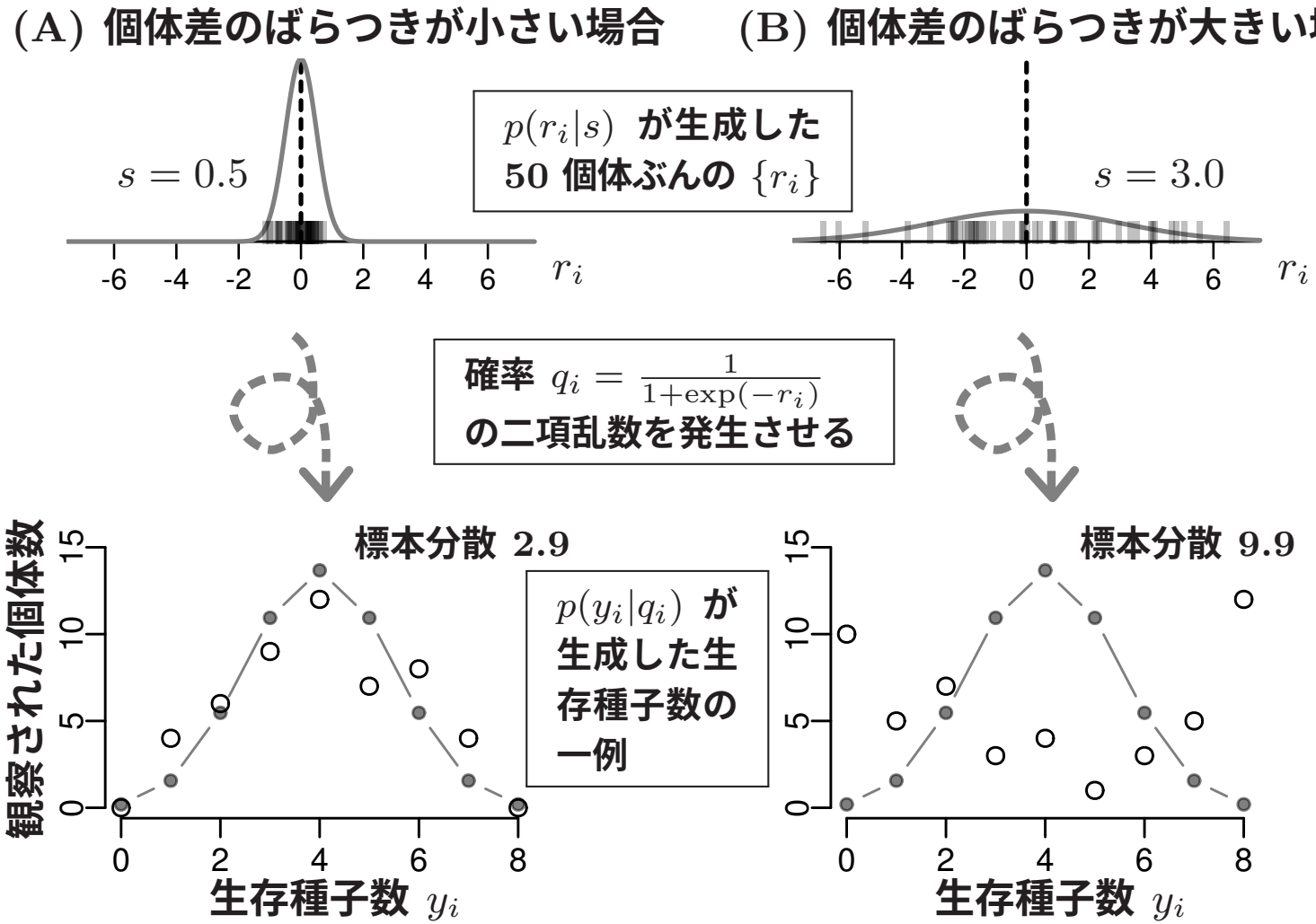


(B) ぜんぜん二項分布じゃない!

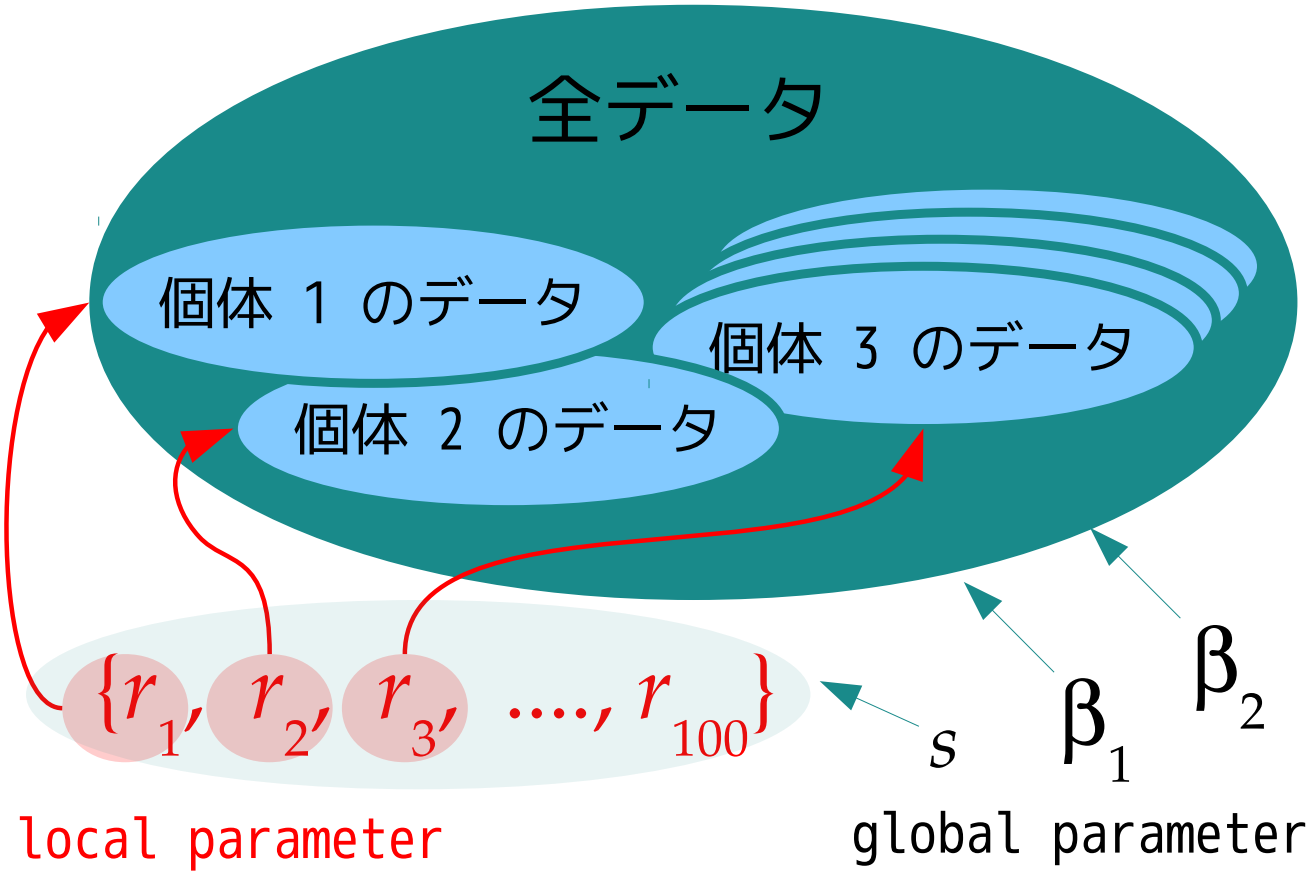


○が観測されたデータの図示

個体差 r_i の分布と過分散の関係



統計モデルの大域的・局所的なパラメーター



データのどの部分を説明しているのか?

統計モデリング入門 道総研 [08]

マルコフ連鎖モンテカルロ法

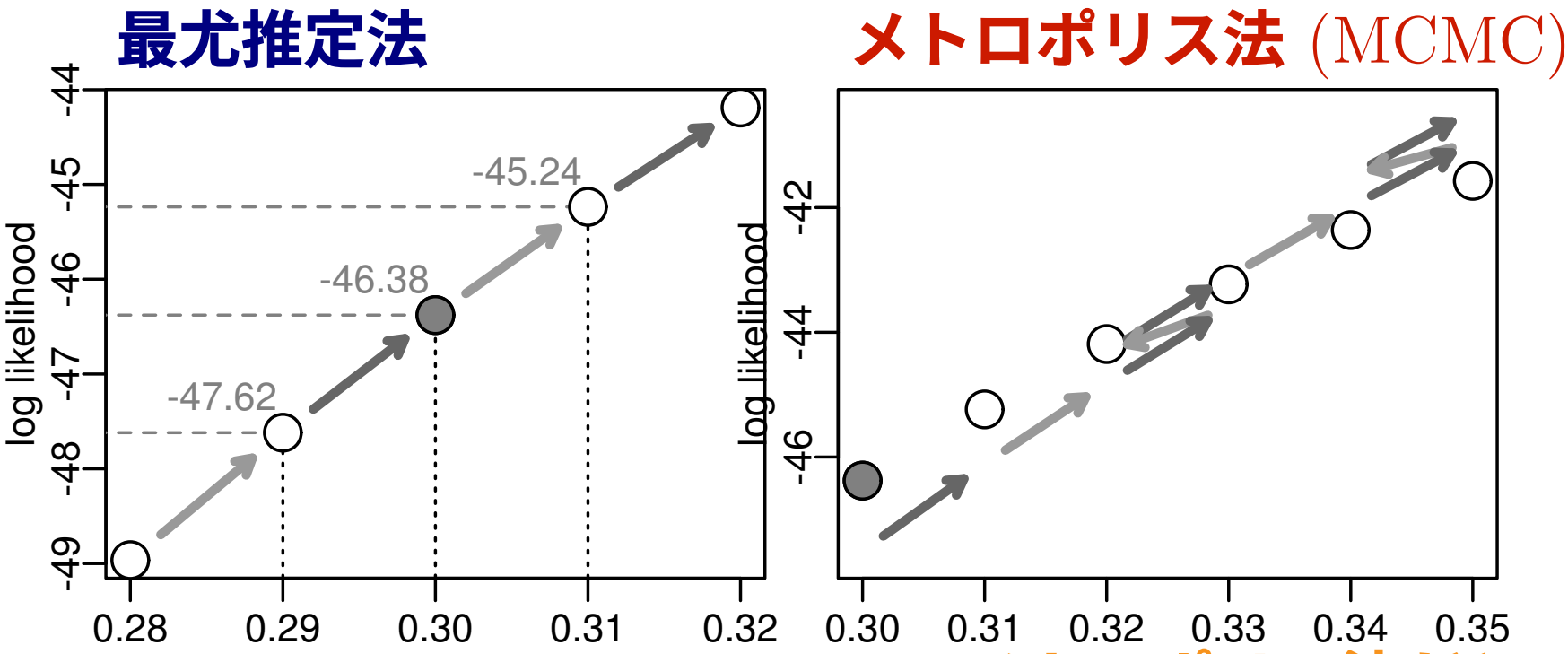
久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

道総研勉強会 http://goo.gl/HQbeoh

2015–11–17

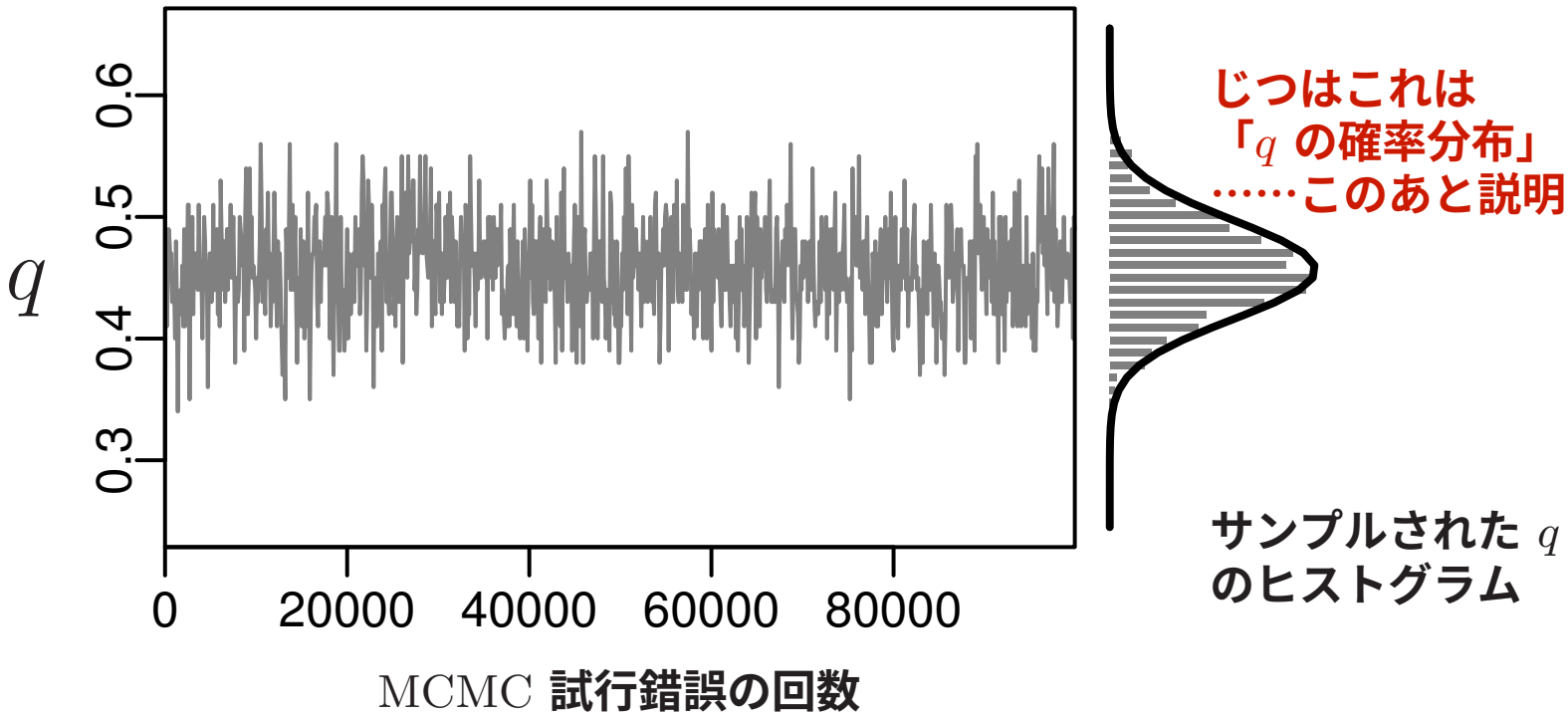
ファイル更新時刻: 2015–11–12 22:11

メトロポリス法のルールで q を動かす



メトロポリス法だと
「単調な山のぼり」にはならない

もっともっと長くサンプリングしてみる

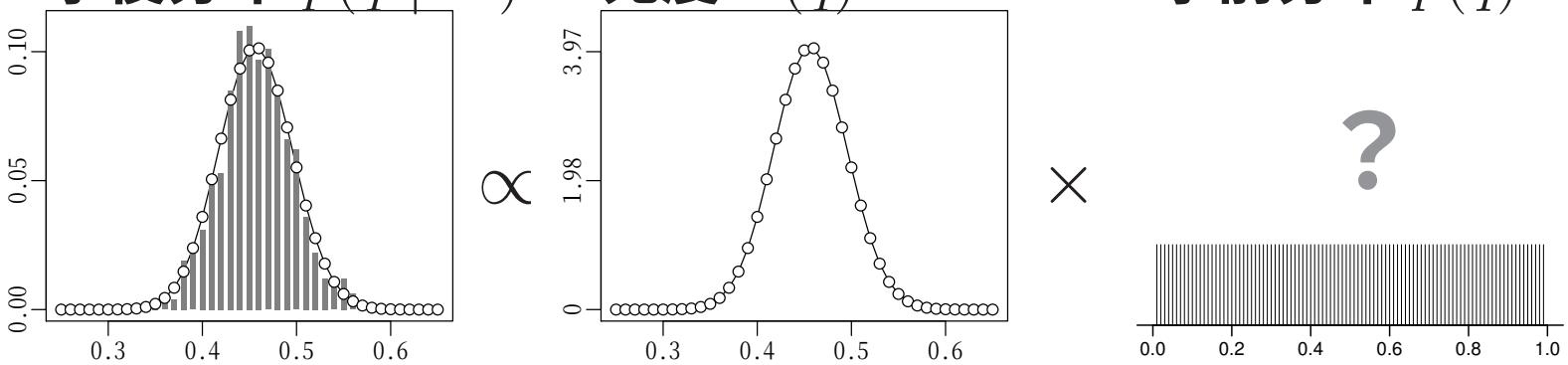


なんだか，ある「山」のかたちにまとまったぞ?

ベイズ統計にむりやりこじつけてみると?

q の事前分布は一様分布，と考えるとつじつまがあう?

事後分布 $p(q \mid \mathbf{Y})$ $\propto$ 尤度 $L(q)$ $\times$ 事前分布 $p(q)$



事前分布ってのがよくわからない……

統計モデリング入門 道総研 [09]

階層ベイズモデル

久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

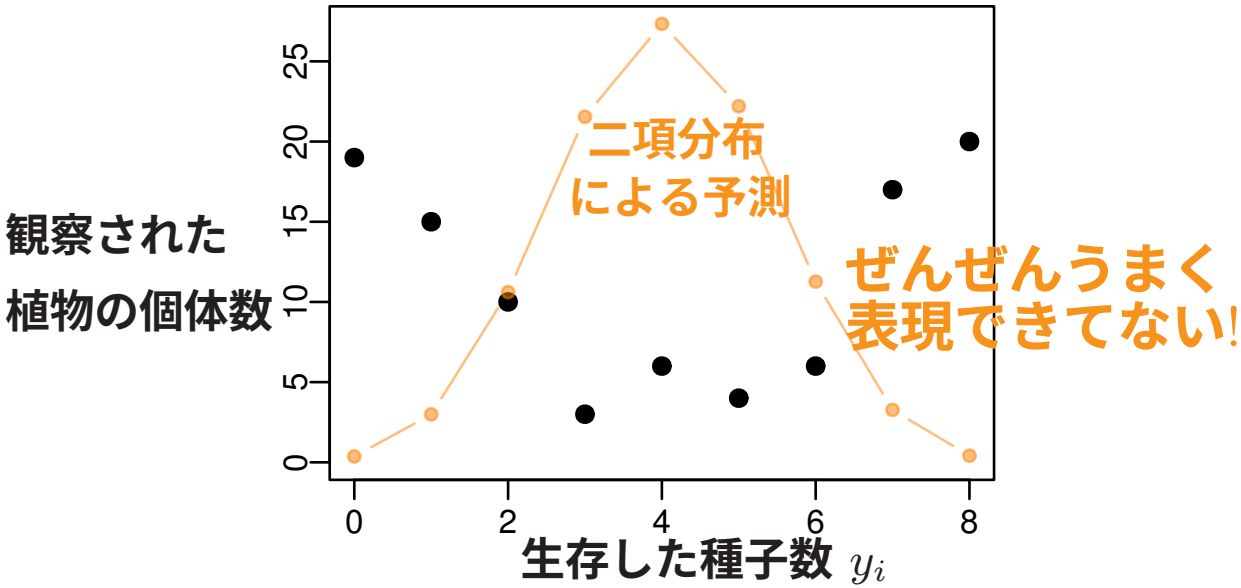
道総研勉強会 http://goo.gl/HQbeoh

2015–11–17

ファイル更新時刻: 2015–11–12 22:11

二項分布では説明できない観測データ!

100 個体の植物の合計 800 種子中 403 個の生存が見られたので，平均生存確率は 0.50 と推定されたが……

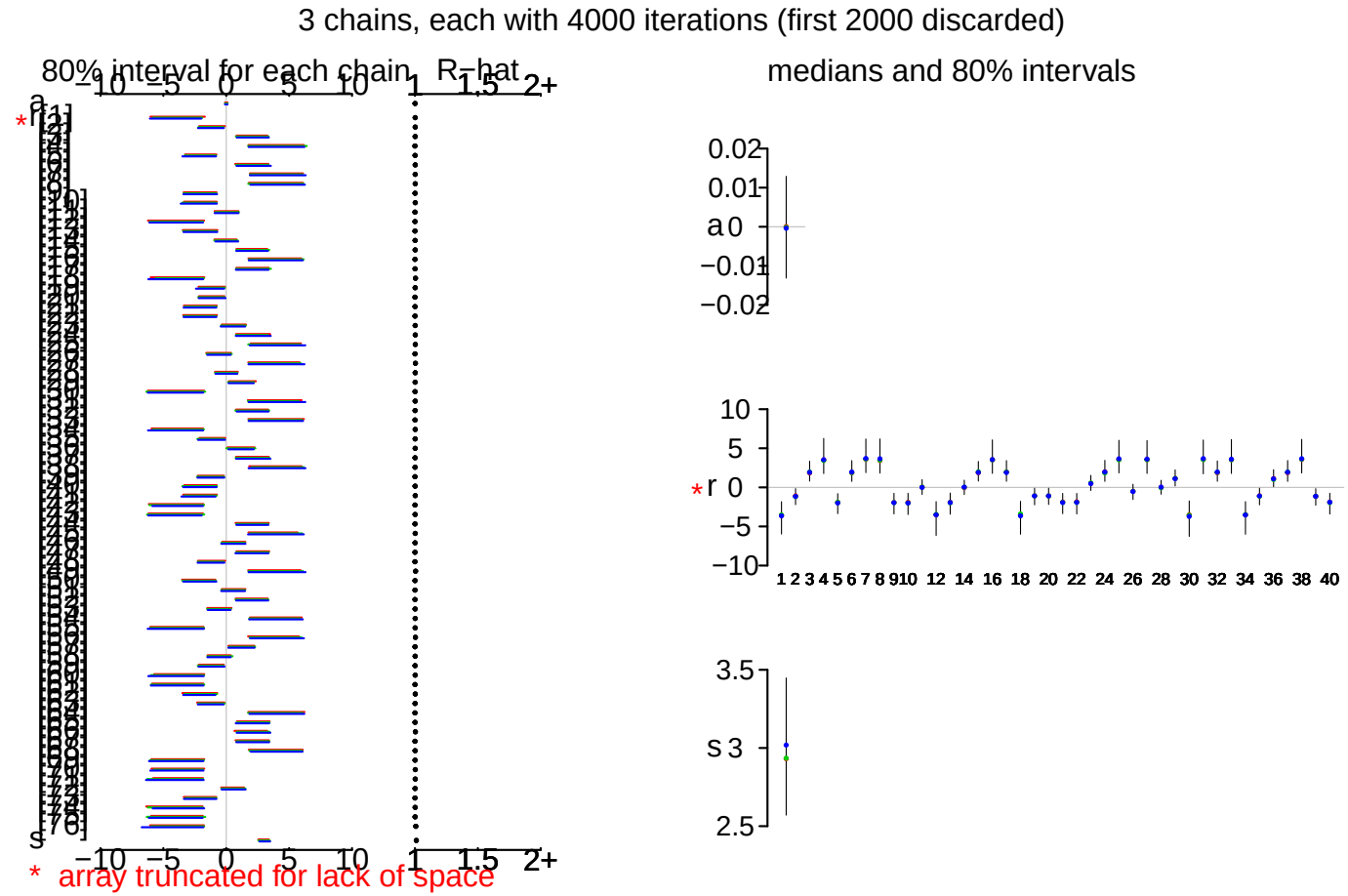


さっきの例題と同じようなデータなのに?

(「統計モデリング入門」第 10 章の最初の例題)

JAGS で得られた事後分布サンプルの要約

```
> source("mcmc.list2bugs.R")                    # なんとなく便利なので…
> post.bugs <- mcmc.list2bugs(post.mcmc.list) # bugs クラスに変換
```



bugs オブジェクトの post.bugs を調べる

- print(post.bugs, digits.summary = 3)
- **事後分布の 95% 信頼区間などが表示される**

```
3 chains, each with 4000 iterations (first 2000 discarded), n.thin = 2
n.sims = 3000 iterations saved
         mean    sd   2.5%    25%    50%    75%  97.5%  Rhat n.eff
a       0.020 0.321 -0.618 -0.190  0.028  0.236  0.651 1.007   380
s       3.015 0.359  2.406  2.757  2.990  3.235  3.749 1.002  1200
r[1]   -3.778 1.713 -7.619 -4.763 -3.524 -2.568 -1.062 1.001  3000
r[2]   -1.147 0.885 -2.997 -1.700 -1.118 -0.531  0.464 1.001  3000
r[3]    2.014 1.074  0.203  1.282  1.923  2.648  4.410 1.001  3000
r[4]    3.765 1.722  0.998  2.533  3.558  4.840  7.592 1.001  3000
r[5]   -2.108 1.111 -4.480 -2.775 -2.047 -1.342 -0.164 1.001  2300
... （中略）
r[99]   2.054 1.103  0.184  1.270  1.996  2.716  4.414 1.001  3000
r[100] -3.828 1.766 -7.993 -4.829 -3.544 -2.588 -1.082 1.002  1100
```

統計モデリング入門 道総研 [10]

階層ベイズモデル: 時系列データ解析

久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

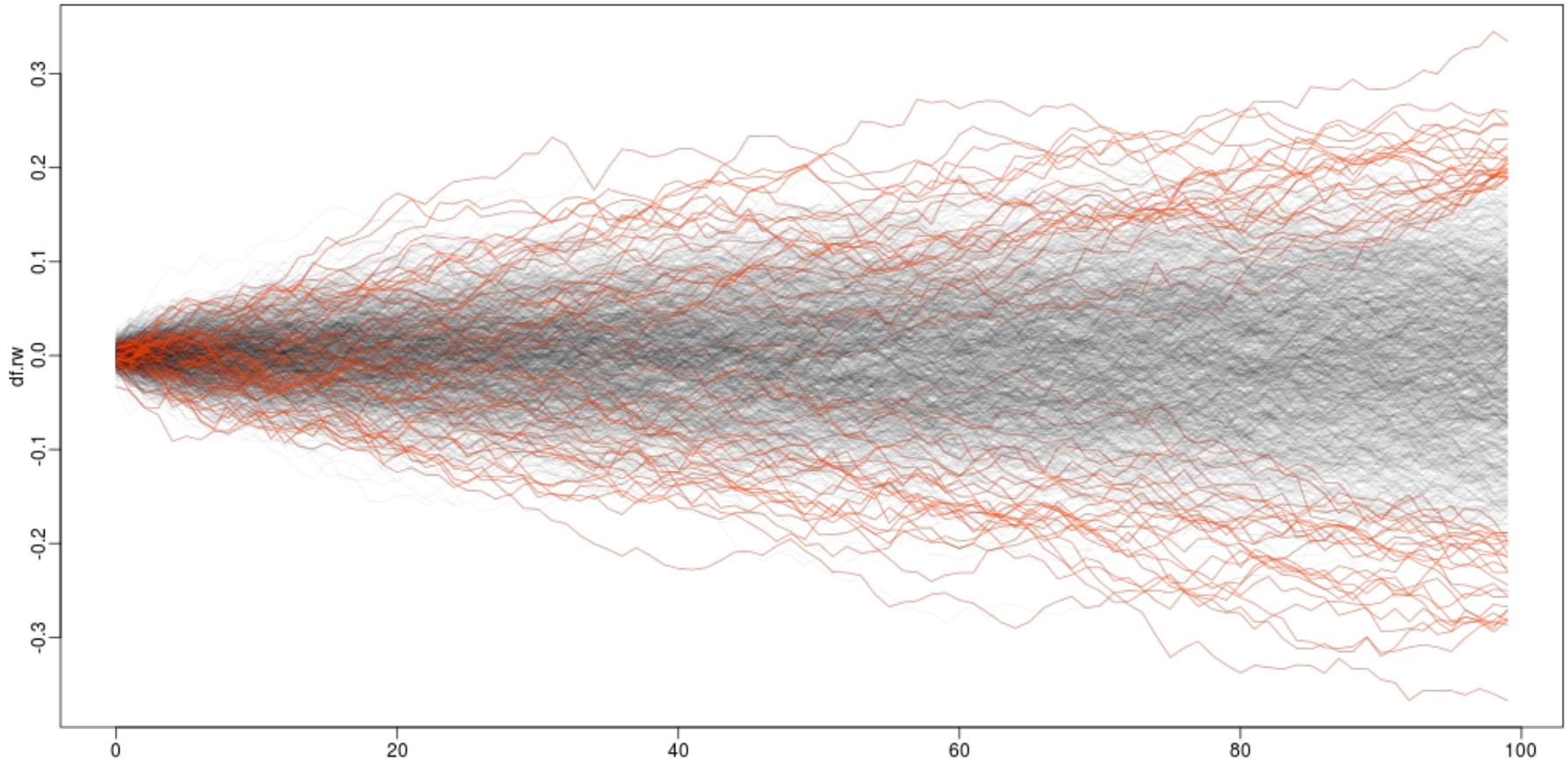
道総研勉強会 http://goo.gl/HQbeoh

2015–11–17

ファイル更新時刻: 2015–11–12 22:32

時系列データ解析でよく見る『あぶない』モデリング

久保拓弥（北海道大・環境科学）



今日の要点

「あぶない」時系列データ解析は

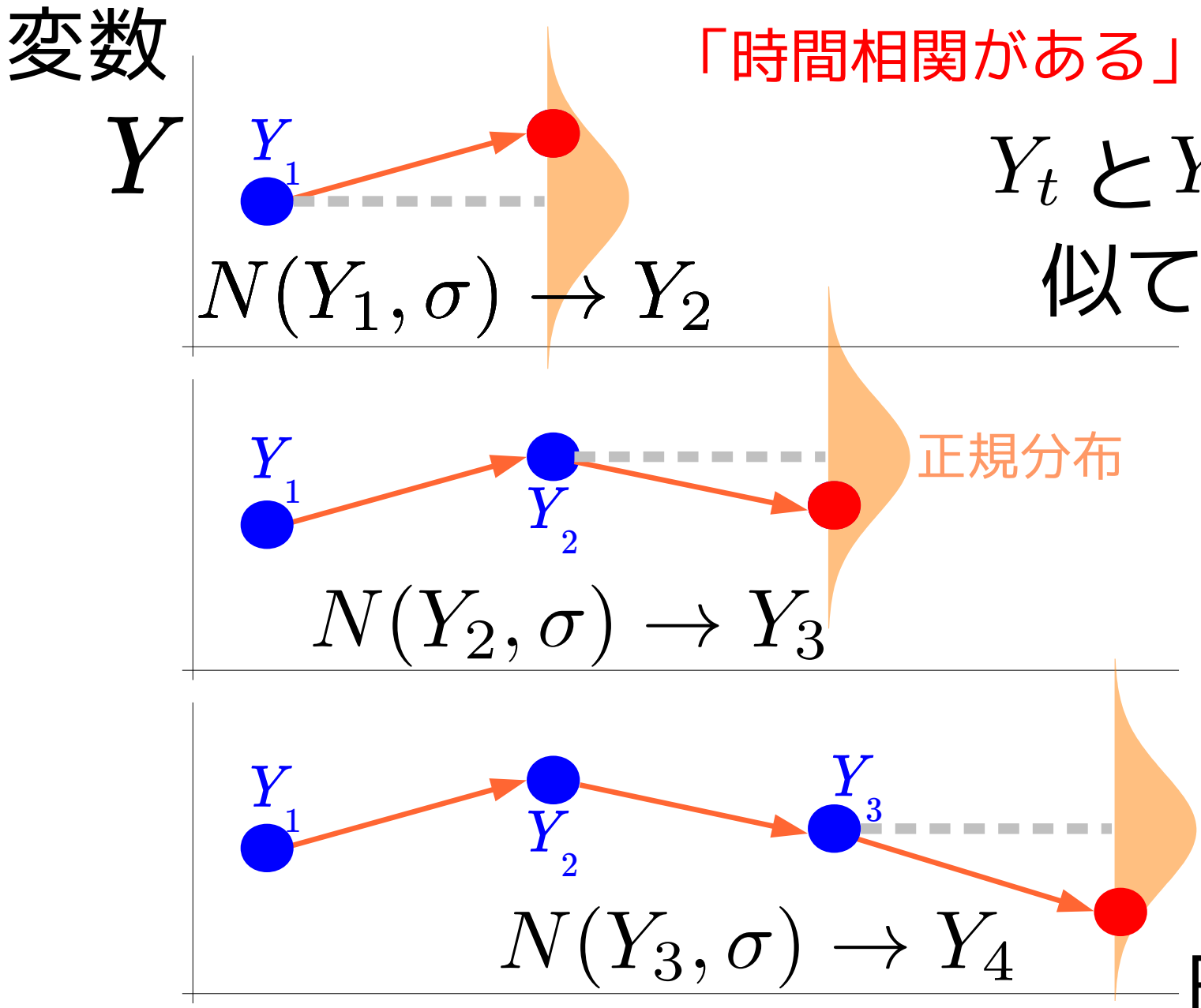
やめましょう!

統計モデル
のあてはめ

(危 1) 時系列データの GLM あてはめ

(危 2) 時系列 Y_t ～ 時系列 X_t

各時刻の個体数 ～ 気温 とか


変数
Y
「時間相関がある」とは?
Y_t と Y_{t+1} は
似ている!
Y_1
$N(Y_1, \sigma) \to Y_2$
Y_2
$N(Y_2, \sigma) \to Y_3$
正規分布
Y_3
$N(Y_3, \sigma) \to Y_4$
時間 t

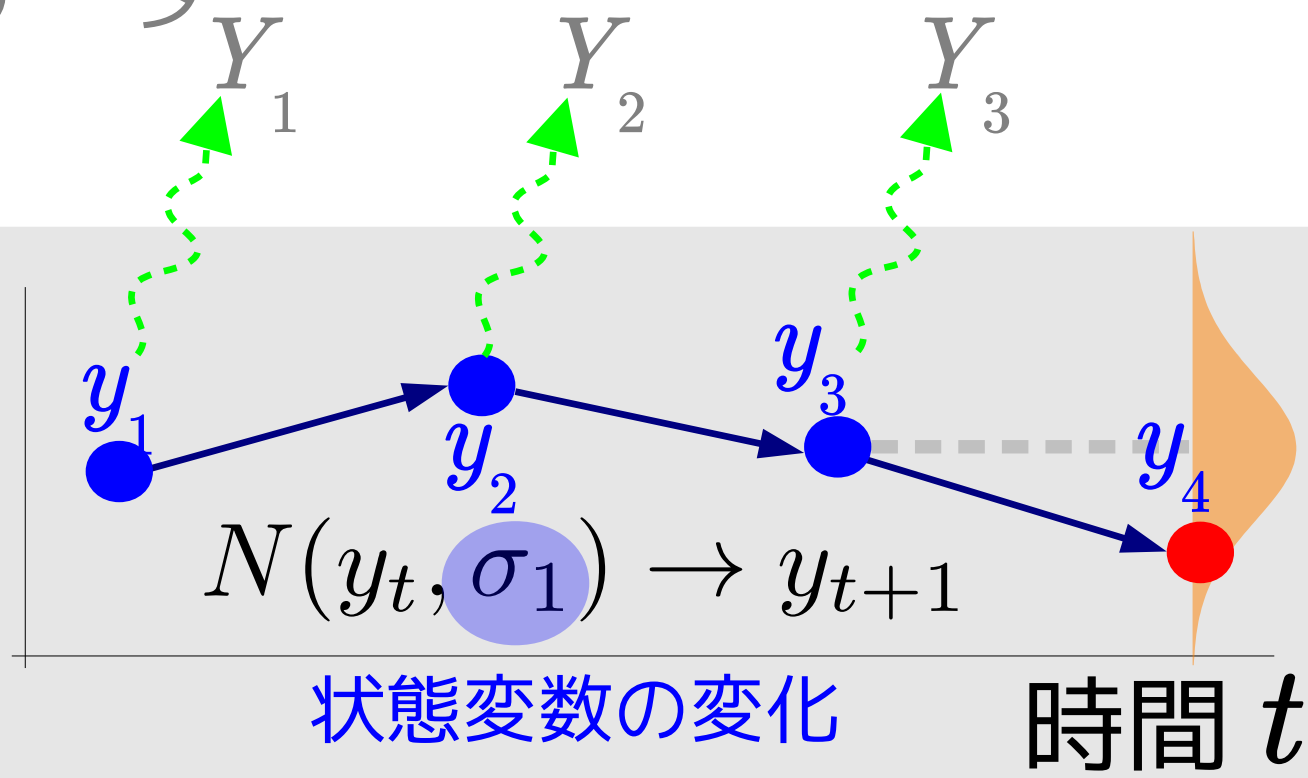
状態空間モデル

観測の誤差

$N(y_t, \sigma_2) \to Y_t$

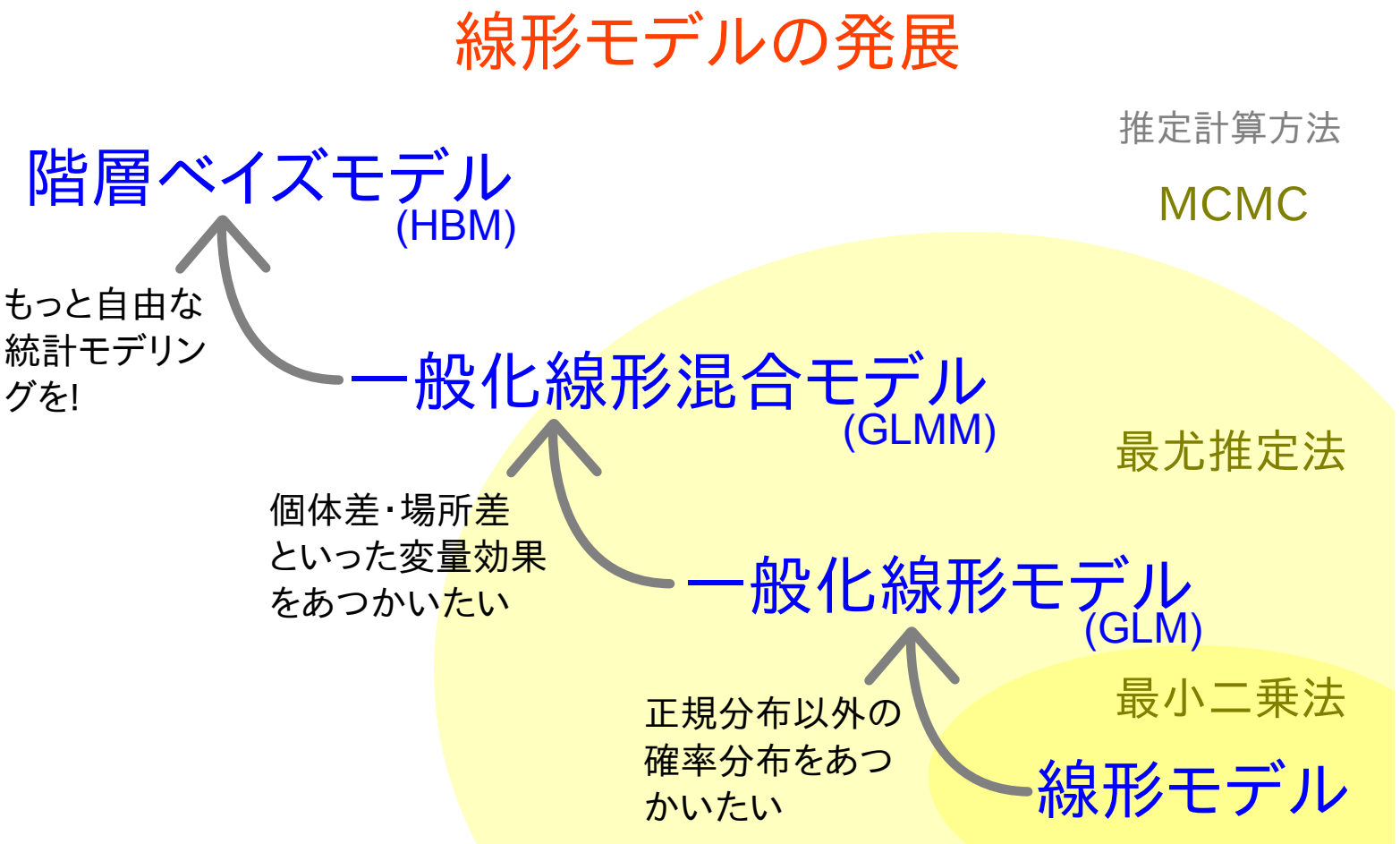
二種類のσをもつ

観測データ



観測できない世界（状態空間）

“統計モデリング入門”に登場する統計モデル



データの特徴にあわせて線形モデルを改良・発展させる