統計モデリング入門 筑波大 (大塚) 集中講義 [01]

集中講義全体の概要: 統計モデルしましょう!

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

筑波大集中講義 http://goo.gl/HvRhXn

2015–02–28

ファイル更新時刻: 2015–02–28 09:29

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. **サイコロの統計モデル**
   “統計モデル”の構造と機能
4. **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

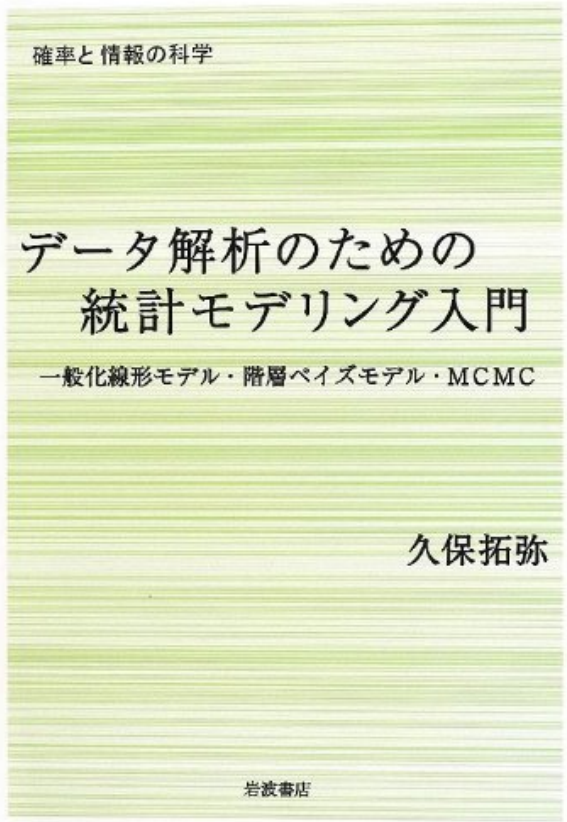
1. なぜ「統計モデリング入門」?

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!

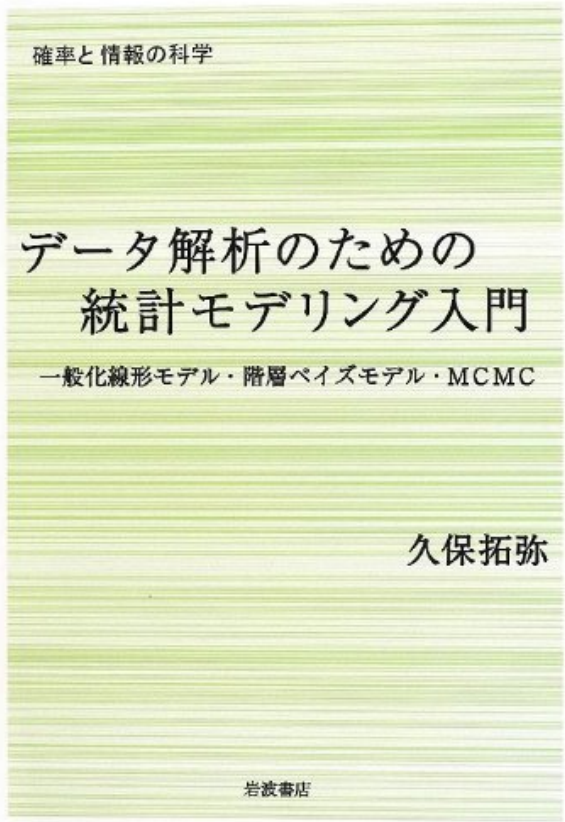


なんで，そんな本なんか書いたの?!

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

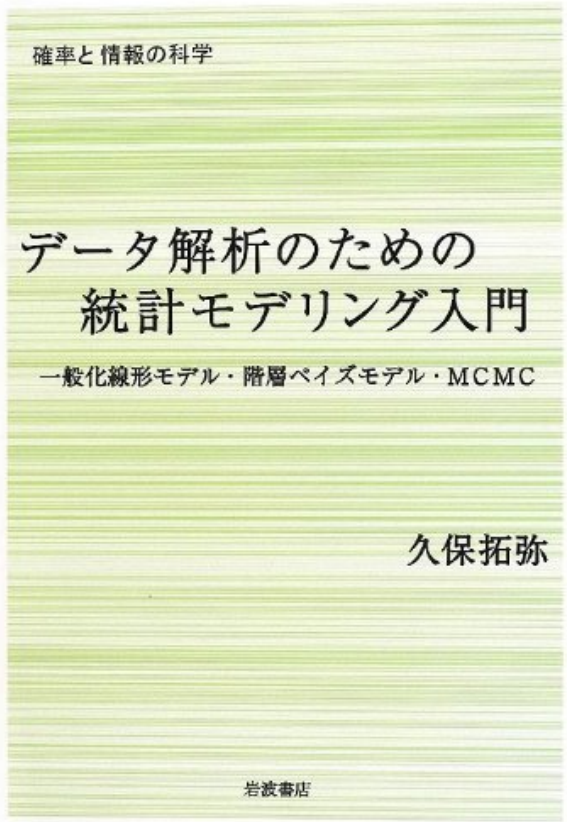


「何も考えない」データ解析はよくない

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

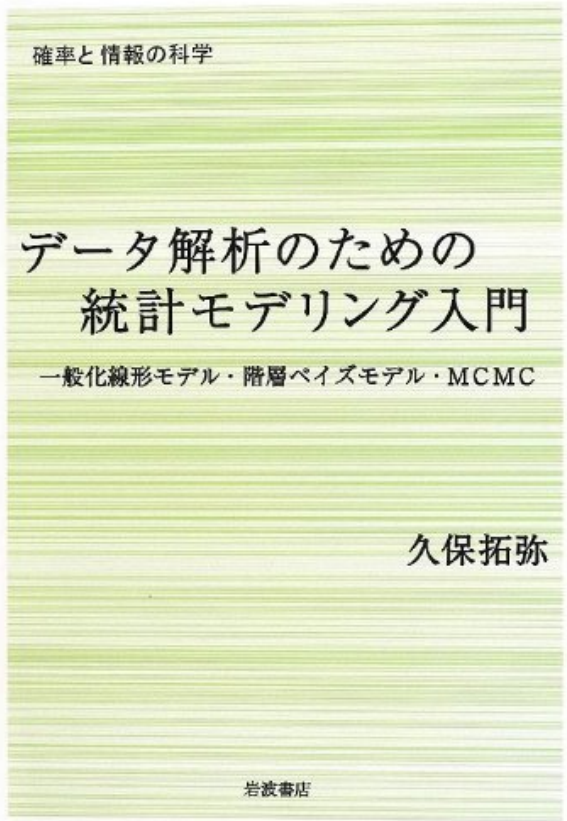


統計モデルって何?

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……



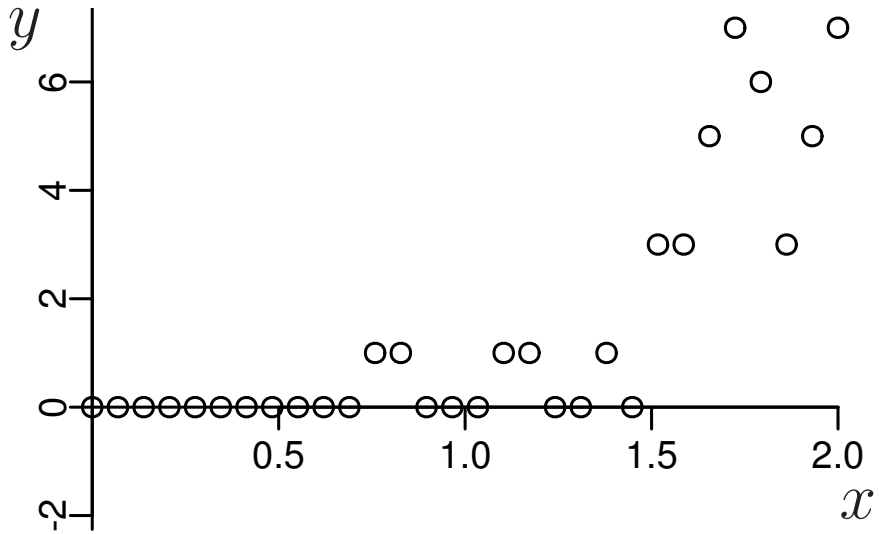
2. 何も考えないデータ解析の問題点

“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



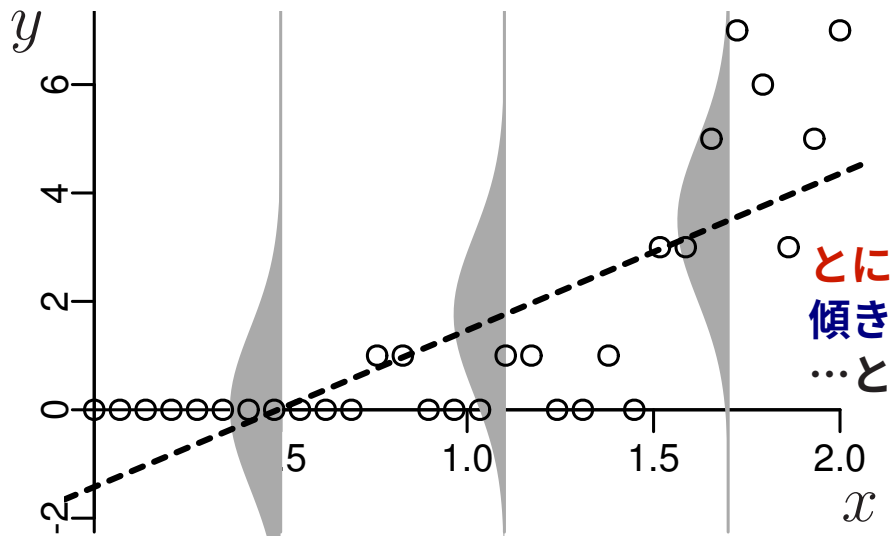
説明変数

- **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「直線回帰」「正規分布」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

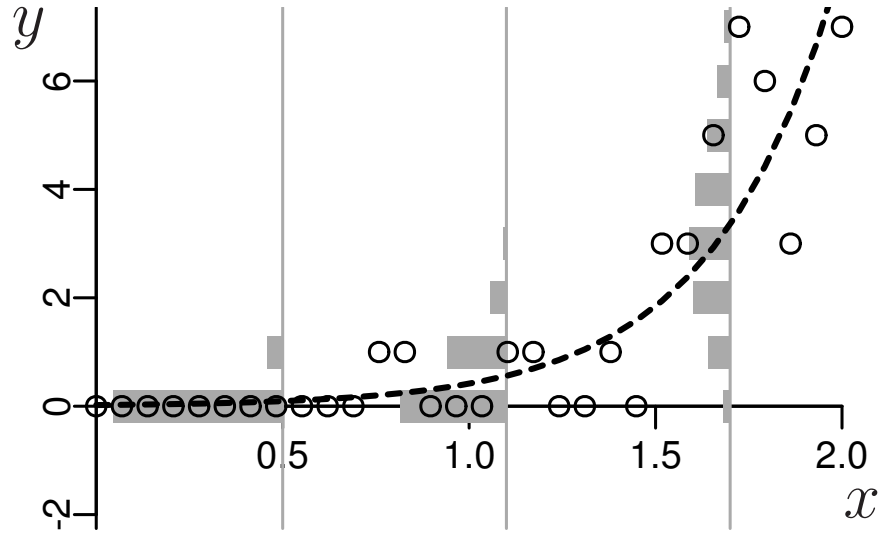
説明変数

- 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 ……

データにあわせた確率分布つかうとマシだよね

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):

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

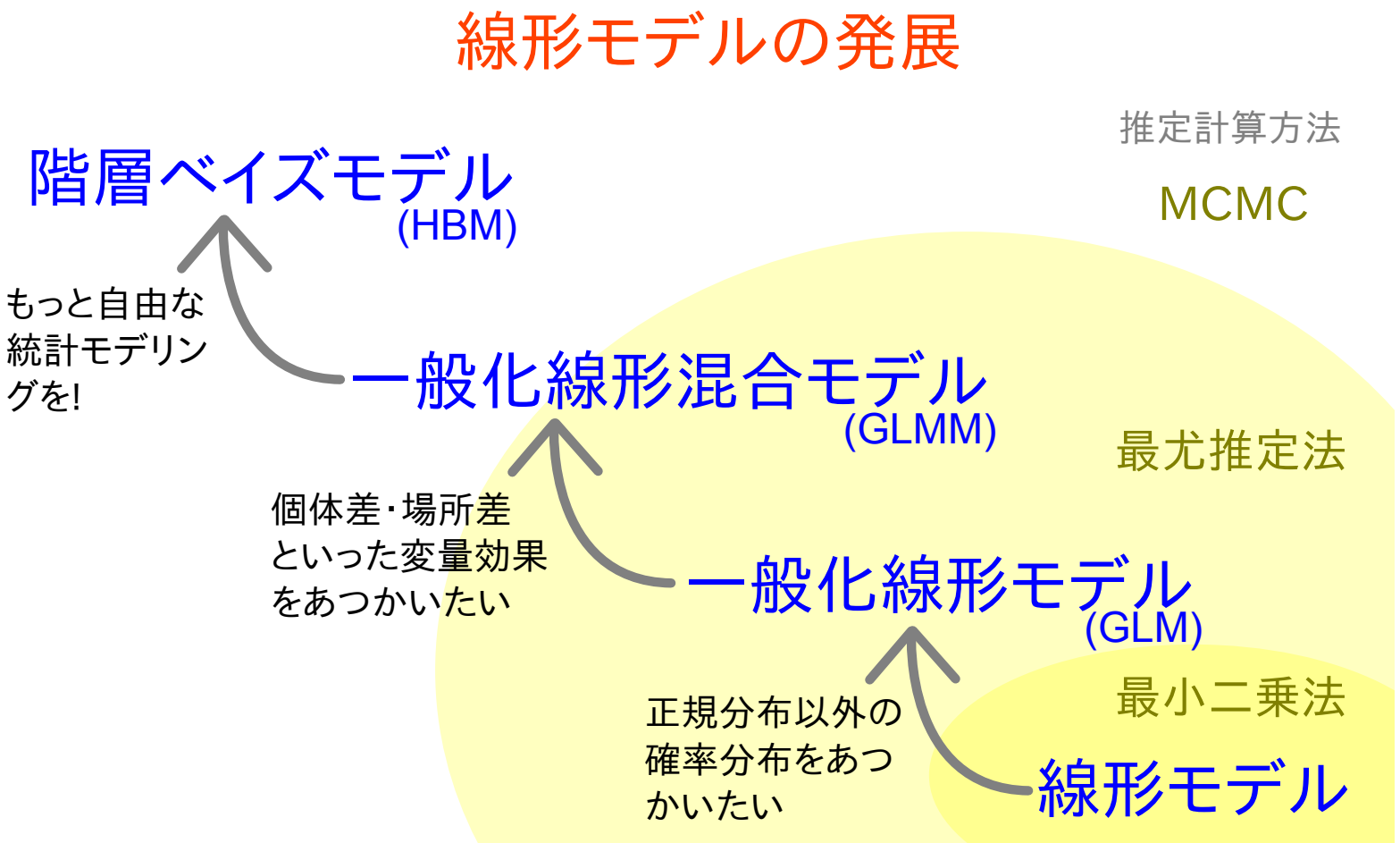
(交互作用項については粕谷さんが説明してくれます)

R で一般化線形モデル: glm() 関数

	確率分布	乱数生成	パラメーター推定
(離散)	ベルヌーイ分布	rbinom()	glm(family = binomial)
	二項分布	rbinom()	glm(family = binomial)
	ポアソン分布	rpois()	glm(family = poisson)
	負の二項分布	rnbinom()	glm.nb() in library(MASS)
(連続)	ガンマ分布	rgamma()	glm(family = gamma)
	正規分布	rnorm()	glm(family = gaussian)

- glm() で使える確率分布は上記以外もある
- GLM は直線回帰・重回帰・分散分析・ポアソン回帰・ロジスティック回帰その他の「よせあつめ」と考えてもよいかも
- 今日はポアソン回帰 を使った GLM だけ紹介します

“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. サイコロの統計モデル

“統計モデル”の構造と機能

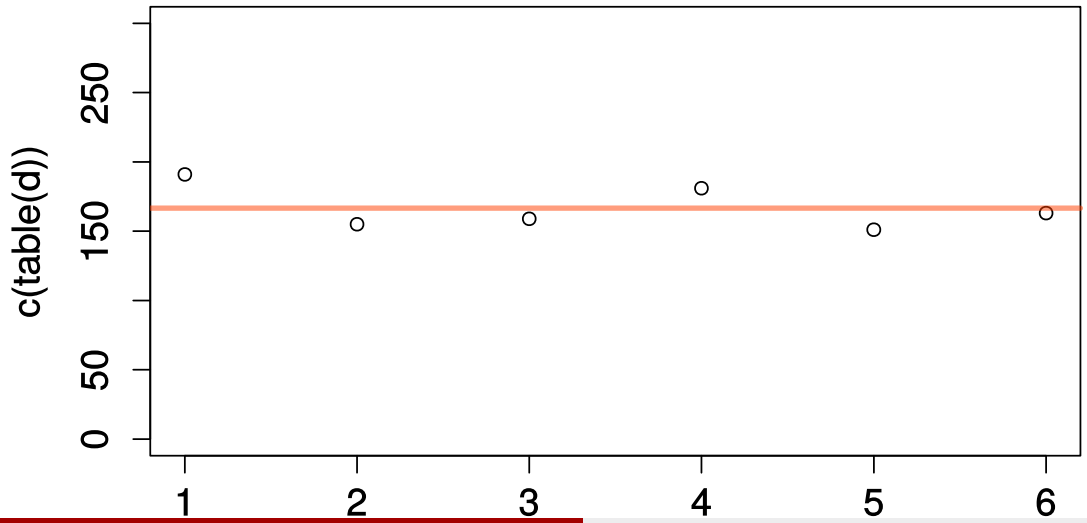
「統計モデル」とは何か?

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

「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

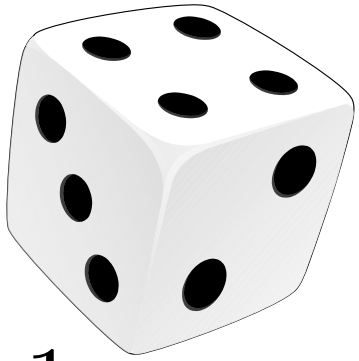
1000回サイコロふった

1000/6 = 166.66...?

「サイコロ」の確率分布は?

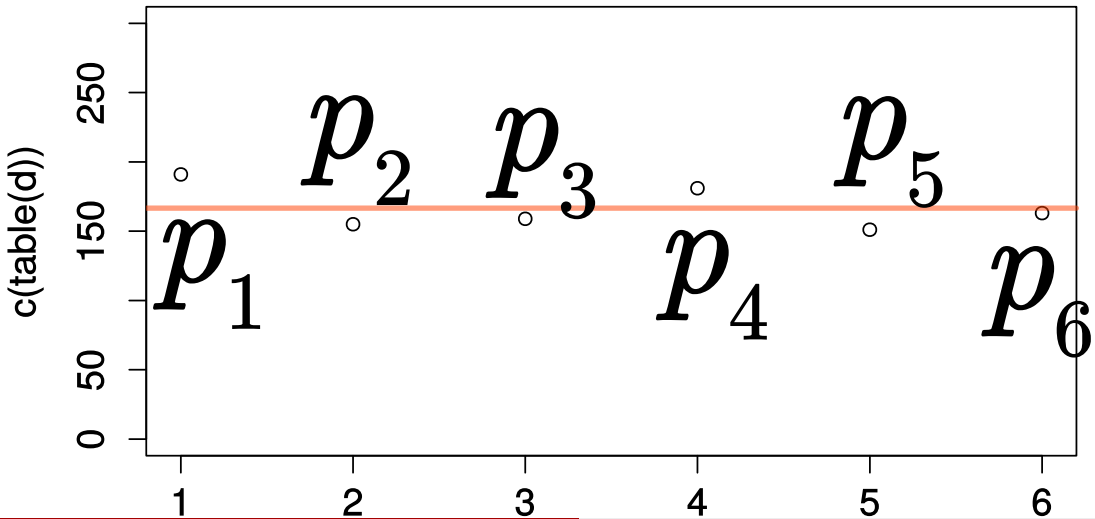
Categorical Distribution

カテゴリカル分布



$$p(Y = k) = p_k \qquad \sum_{k=1}^{6} p_k = 1$$

$$k \in \{1, 2, 3, 4, 5, 6\}$$



架空データ

1000回サイコロふった

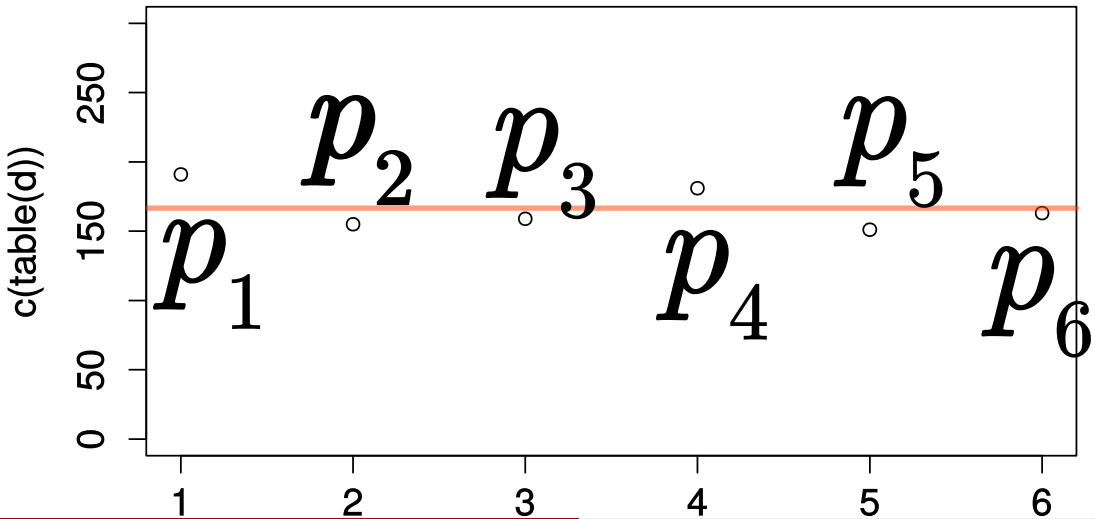
1000/6 = 166.66...?

確率分布のパラメーターは $\{p_k\}$

最尤推定量

$$\hat{p}_k = \frac{k \text{ の目が出た回数}}{1000}$$

```
> table(d)
d
  1   2   3   4   5   6
191 155 159 181 151 163
```



架空データ

1000回サイコロふった

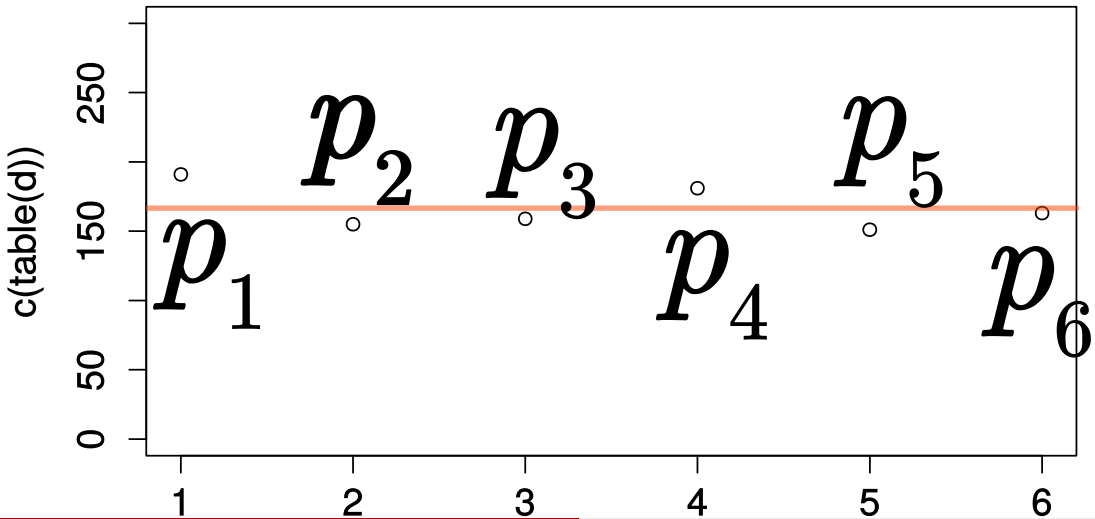
1000/6 = 166.66...?

「サイコロの統計モデル」にできること

パラメーターの推定

$$\hat{p}_k = \frac{k \text{ の目が出た回数}}{1000}$$

```
> table(d)
d
  1   2   3   4   5   6
191 155 159 181 151 163
```



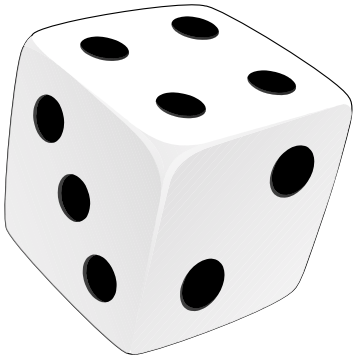
架空データ

1000回サイコロふった

1000/6 = 166.66...?

「サイコロの統計モデル」にできること

乱数発生

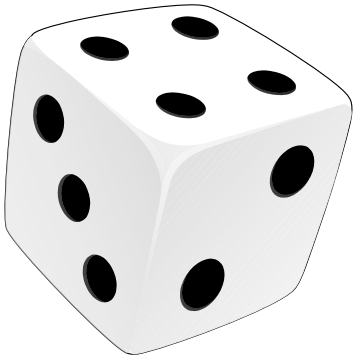


```
> v.prob <- table(d) / 1000
> replicate(8, table(sample(1:6, 1000,
+ replace = TRUE, prob = v.prob)))
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]
1	185	201	174	202	179	193	198	197
2	156	159	158	164	160	141	154	151
3	173	161	195	180	177	153	147	173
4	176	171	173	177	189	174	200	190
5	147	157	152	139	138	171	122	130
6	163	151	148	138	157	168	179	159

「サイコロの統計モデル」にできること

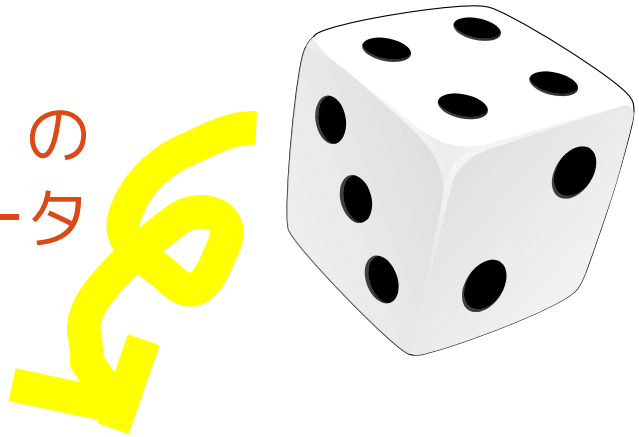
予測



```
> # サイコロ 1000 回ふりを 1000 回やる
> sim1000 <- replicate(1000,
+ table(sample(1:6, 1000, replace = TRUE,
+ prob = v.prob)))
> # 5 よりも 6 が多くでる回数は?
> sum(sim1000[5,] > sim1000[6,])
[1] 247
> # 3 よりも 4 が多くでる回数は?
> sum(sim1000[3,] > sim1000[4,])
[1] 111
> # 5 より 6 が多いときに，3 よりも 4 が多くでる回数は?
> sum((sim1000[5,] > sim1000[6,])
+ * (sim1000[3,] > sim1000[4,]))
[1] 35
```

「サイコロの統計モデル」にできること

モデル選択や検定

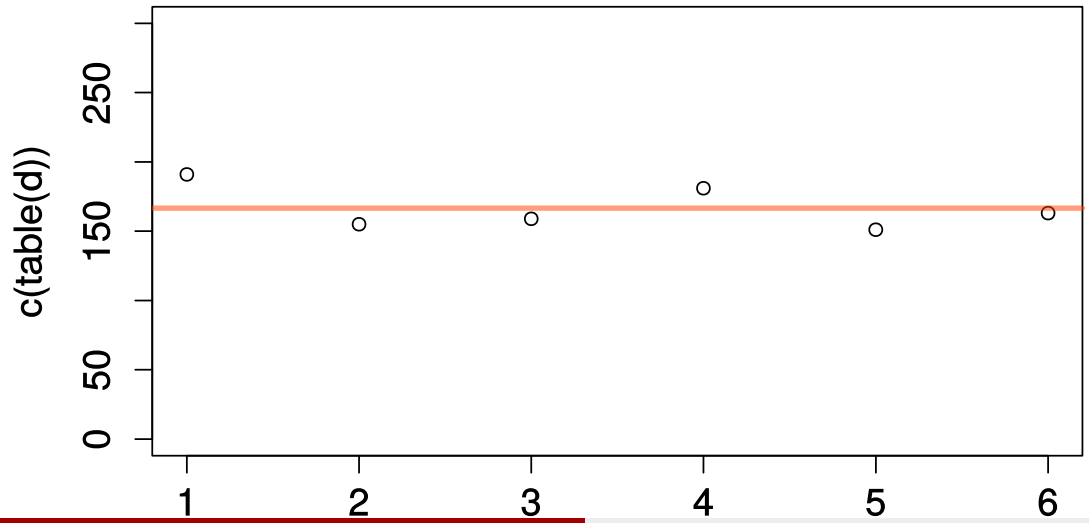


モデル選択

モデル1: p_k がすべて等しい

モデル2: p_k がすべて異なる

「予測力」の高い
モデルを選ぶ



1000回サイコロふった

1000/6 = 166.66...?

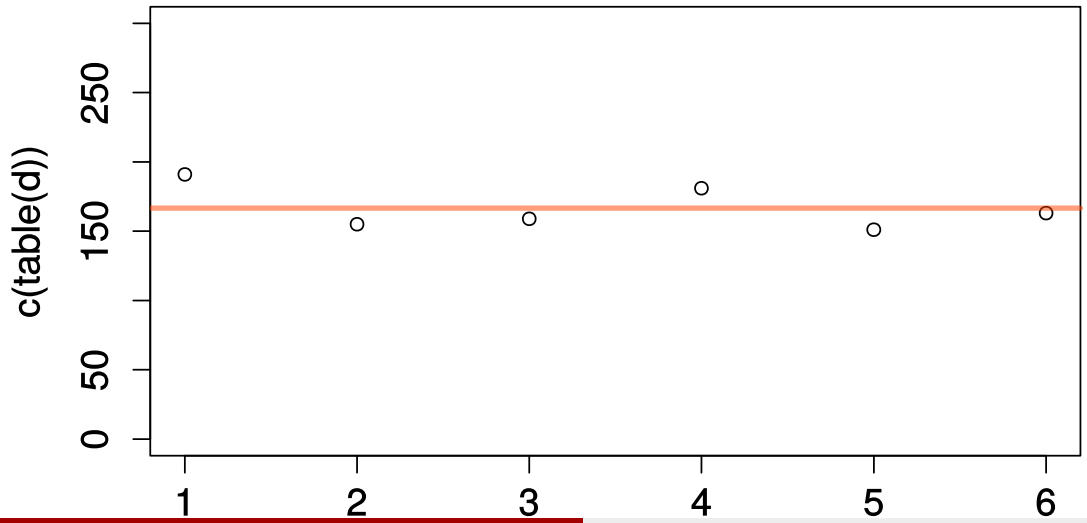
「サイコロの統計モデル」にできること

モデル選択や検定

統計学的な検定

モデル1: p_k がすべて等しい

モデル2: p_k がすべて異なる



モデル1 を安全に棄却できる確率 p だけを評価する

$p > 0.05$ なら…?

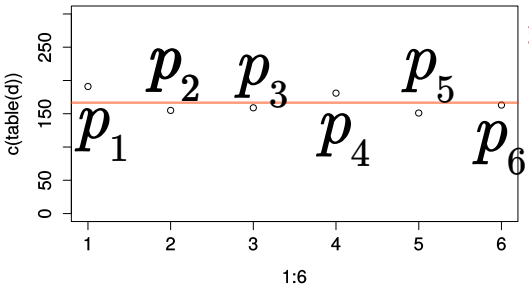
「何も言えない」と結論するのが正しい

「サイコロの統計モデル」にできること

パラメーターの推定

$$\hat{p}_k = \frac{k \text{ の目が出た回数}}{1000}$$

```
> table(d)
d
  1   2   3   4   5   6
191 155 159 181 151 163
```



架空データ

1000回サイコロふった

1000/6 = 166.66...?

「サイコロの統計モデル」にできること

乱数発生

```
> v.prob <- table(d) / 1000
> replicate(8, table(sample(1:6, 1000,
+ replace = TRUE, prob = v.prob)))
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]
1	185	201	174	202	179	193	198	197
2	156	159	158	164	160	141	154	151
3	173	161	195	180	177	153	147	173
4	176	171	173	177	189	174	200	190
5	147	157	152	139	138	171	122	130
6	163	151	148	138	157	168	179	159

「サイコロの統計モデル」にできること

予測

```
> # サイコロ 1000 回ふりを 1000 回やる
> sim1000 <- replicate(1000,
+ table(sample(1:6, 1000, replace = TRUE,
+ prob = v.prob)))
> # 5 よりも 6 が多くでる回数は?
> sum(sim1000[5,] > sim1000[6,])
[1] 247
> # 3 よりも 4 が多くでる回数は?
> sum(sim1000[3,] > sim1000[4,])
[1] 111
> # 5 より 6 が多いときに，3 よりも 4 が多くでる回数は?
> sum((sim1000[5,] > sim1000[6,])
+ * (sim1000[3,] > sim1000[4,]))
[1] 35
```

「サイコロの統計モデル」にできること

モデル選択や検定

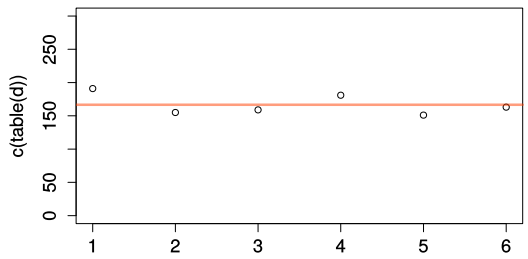
「次」のデータ

モデル選択

モデル1: p_k がすべて等しい

モデル2: p_k がすべて異なる

「予測力」の高いモデルを選ぶ



1000回サイコロふった

1000/6 = 166.66...?

4. 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 筑波大 (大塚) 集中講義 [02]

統計モデル・確率分布・最尤推定

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

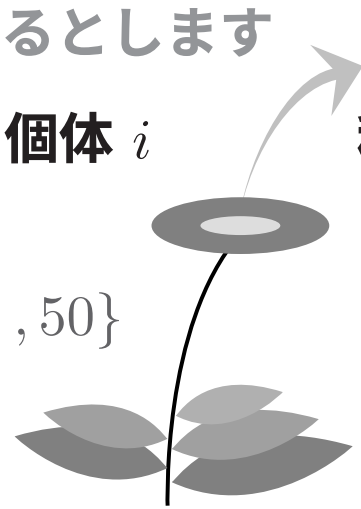
筑波大集中講義 http://goo.gl/HvRhXn

2015–02–28

ファイル更新時刻: 2015–02–27 12:49

こんなデータ (架空) があったとしましょう

まあ，なんだかこういうヘンな
植物を調査しているとします



個体 i

種子数 y_i

全 50 個体

$i \in \{1, 2, 3, \cdots, 50\}$

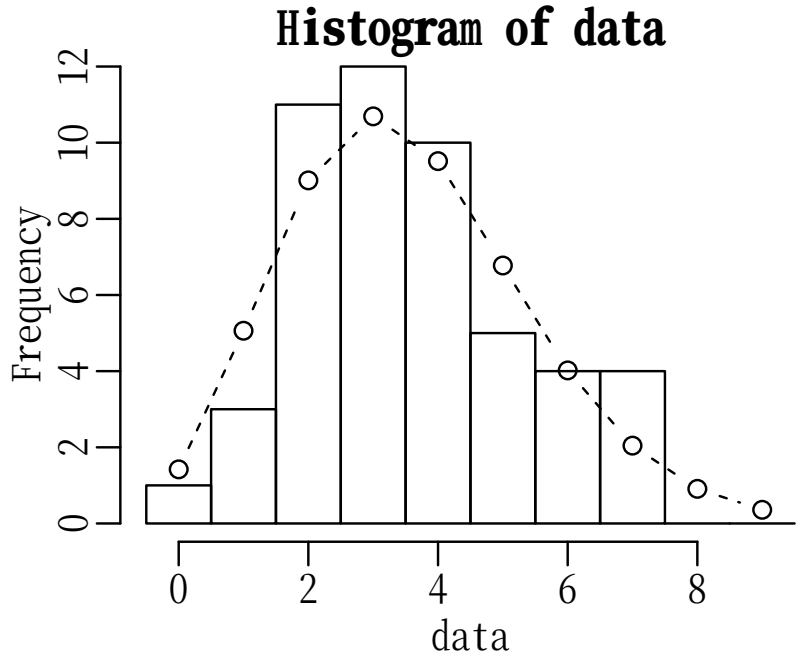
この $\{y_i\}$ が観測データ!

$\{y_i\} = \{y_1, y_2, \cdots, y_{50}\}$

このデータ $\{y_i\}$ がすでに R という統計ソフトウェアに格納されていた，としましょう

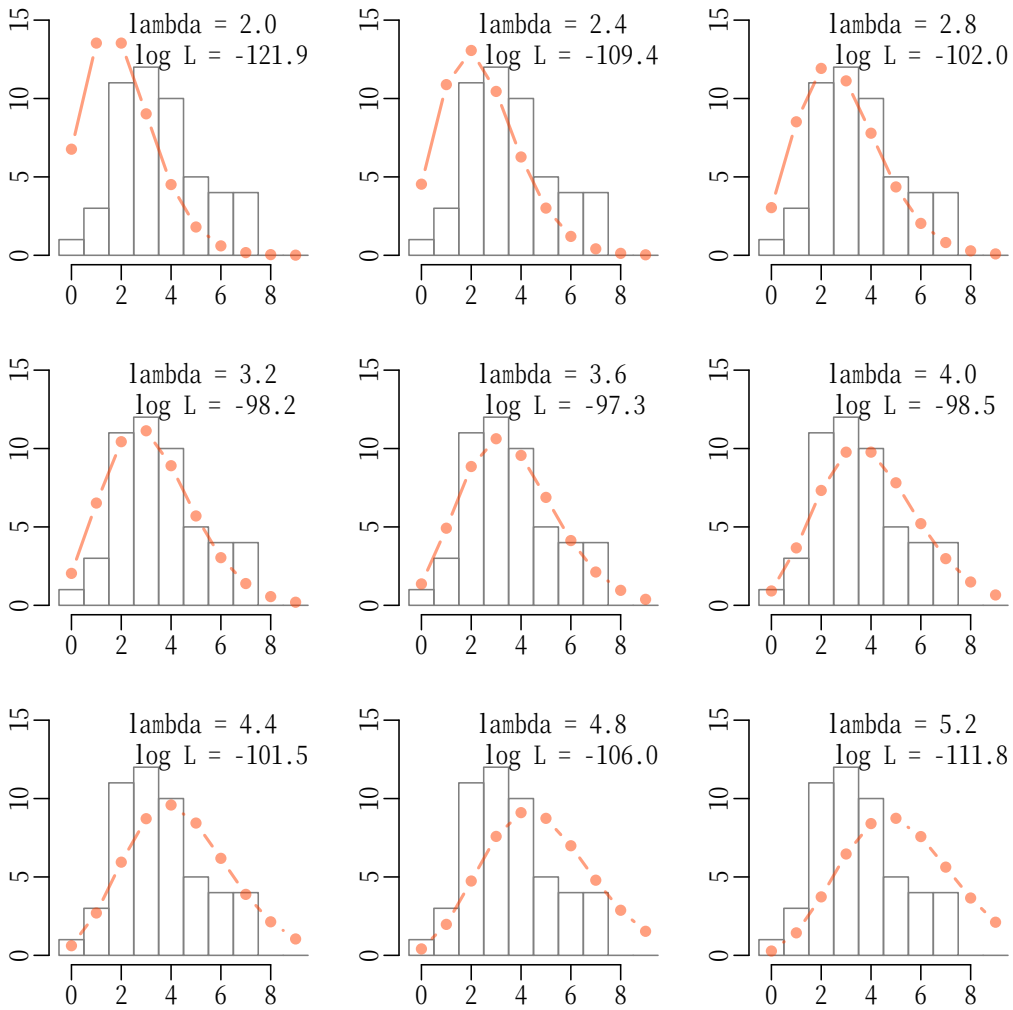
```
> data
 [1] 2 2 4 6 4 5 2 3 1 2 0 4 3 3 3 3 4 2 7 2 4 3 3 3 4
[26] 3 7 5 3 1 7 6 4 6 5 2 4 7 2 2 6 2 4 5 4 5 1 3 2 3
```

データとポアソン分布を重ね合わせる




```
> hist(data, seq(-0.5, 8.5, 0.5))      # まずヒストグラムを描き
> lines(y, prob, type = "b", lty = 2) # その「上」に折れ線を描く
```

λ を変えるとあてはまりの良さが変わる



統計モデリング入門 筑波大 (大塚) 集中講義 [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

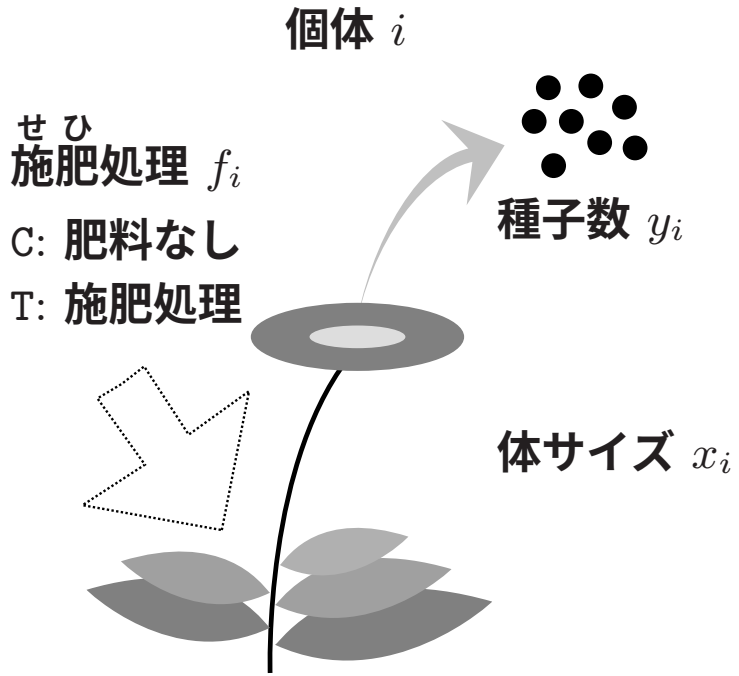
筑波大集中講義 http://goo.gl/HvRhXn

2015–02–28

ファイル更新時刻: 2015–02–28 07:58

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 ($i \in \{51, 52, \cdots 100\}$)

データファイルを読みこむ

data3a.csv は CSV (comma separated value) format file なので，R で読みこむには以下のようにする:

```
> d <- read.csv("data3a.csv")
```

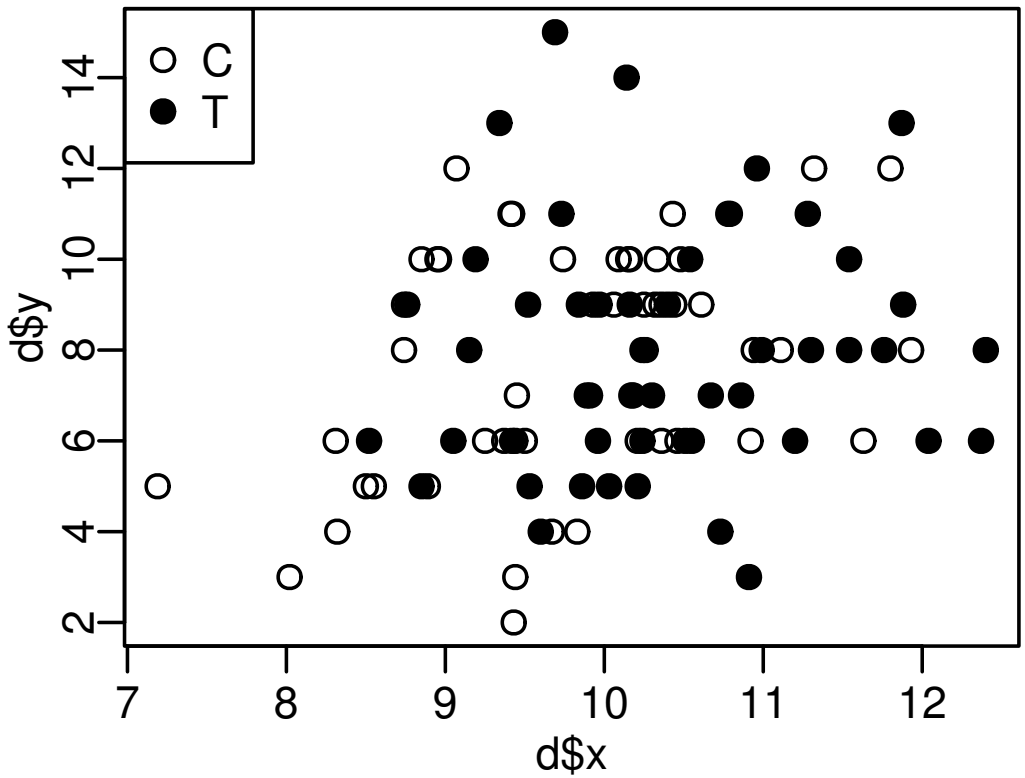
データは d と名付けられた data frame (表みたいなもの) に格納される

とりあえず data frame d を表示

```
> d
     y      x  f
1    6   8.31  C
2    6   9.44  C
3    6   9.50  C
... (中略) ...
99   7  10.86  T
100  9   9.97  T
```

データはとにかく図示する!

```
> plot(d$x, d$y, pch = c(21, 19)[d$f])
> legend("topleft", legend = c("C", "T"), pch = c(21, 19))
```



統計モデリング入門 筑波大 (大塚) 集中講義 [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

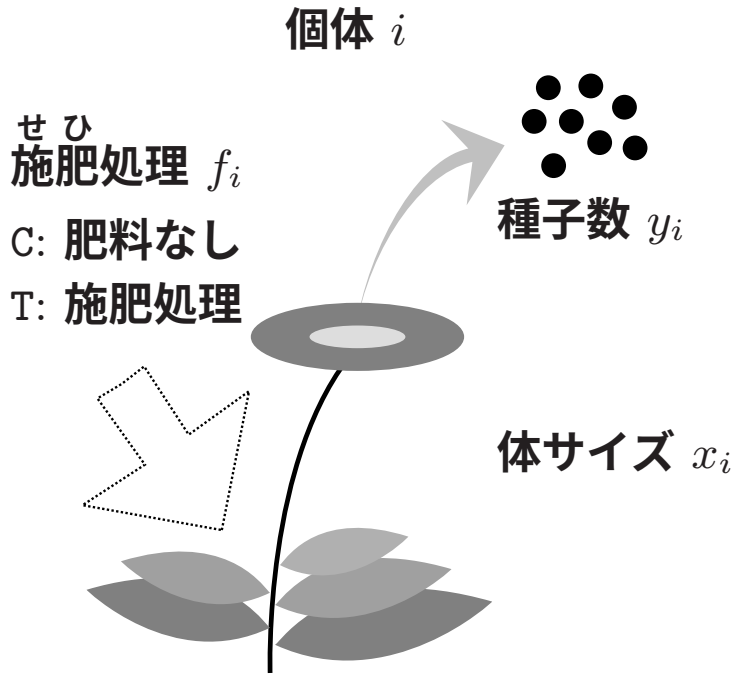
筑波大集中講義 http://goo.gl/HvRhXn

2015–02–28

ファイル更新時刻: 2015–02–28 07:50

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 ($i \in \{51, 52, \cdots 100\}$)

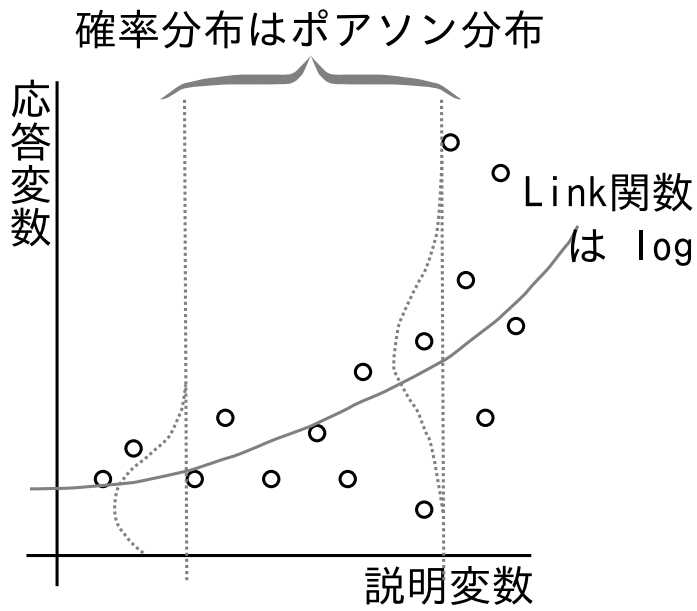
3. R で GLM のパラメーターを推定

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 筑波大 (大塚) 集中講義 [05]

モデル選択と検定

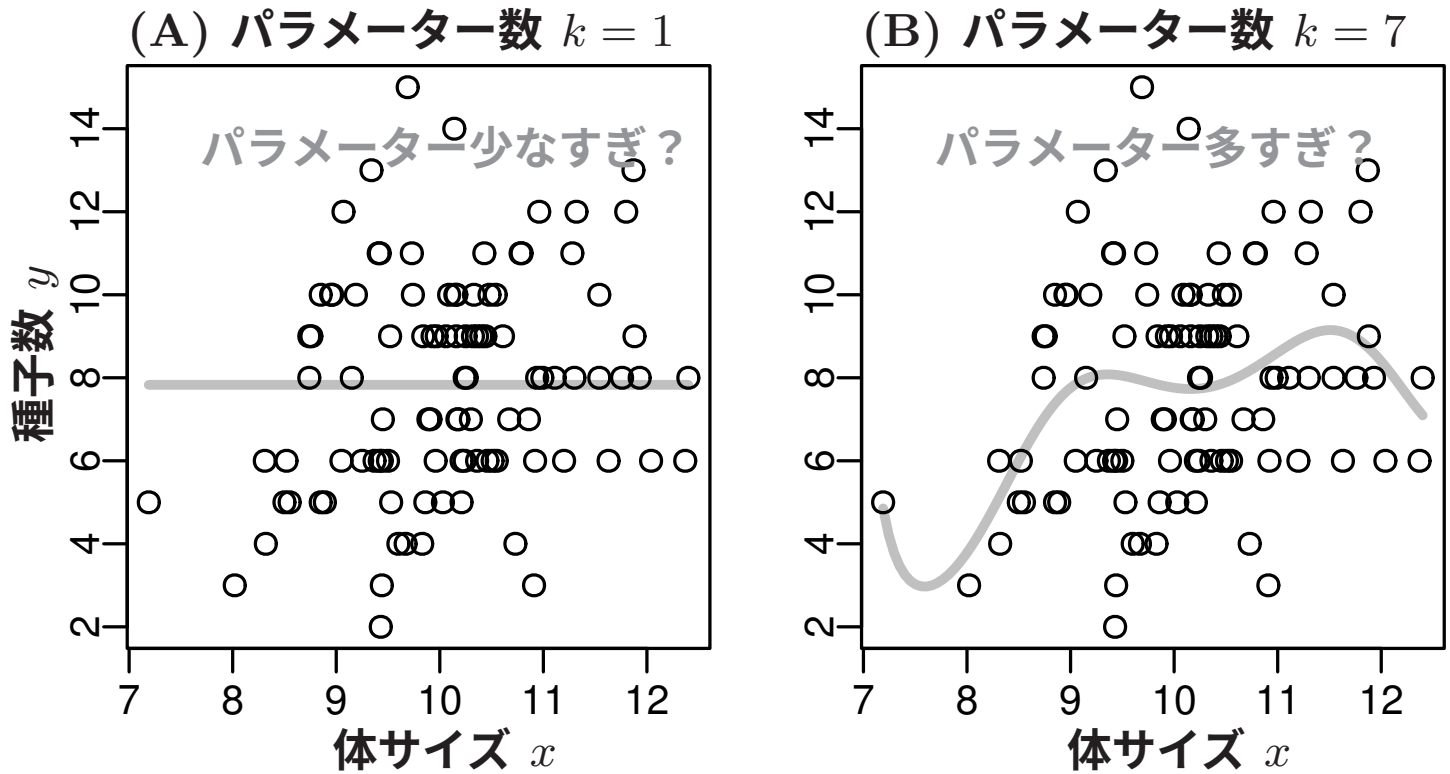
久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

筑波大集中講義 http://goo.gl/HvRhXn

2015–02–28

ファイル更新時刻: 2015–02–27 12:49

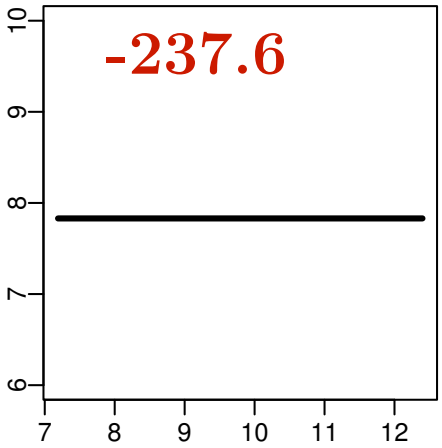
パラメーター数は多くても少なくてもヘン?



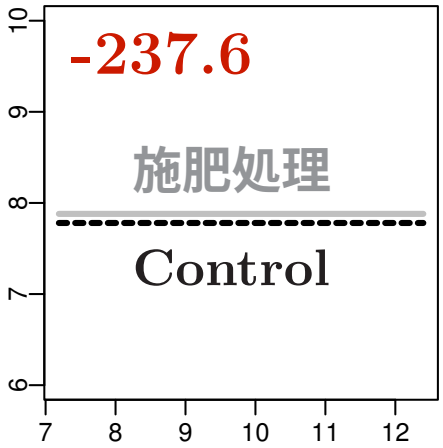
What is the **"best?"** parameter number k?

パラメーター数が多いとあてはまりが良い

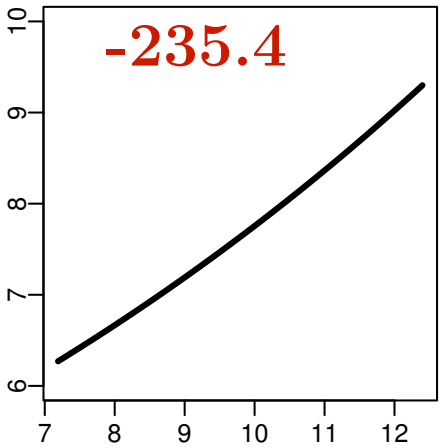
(A) constant λ ($k = 1$)



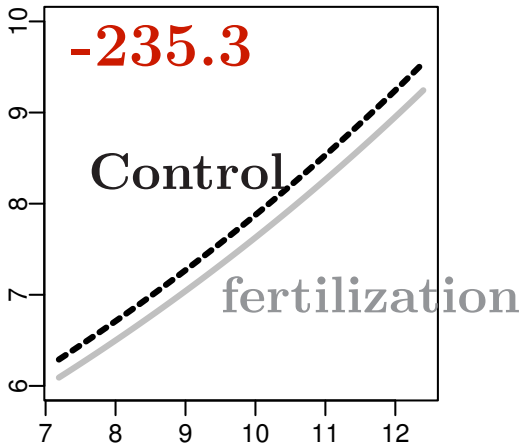
(B) f model ($k = 2$)



(C) x model ($k = 2$)



(D) x + f model ($k = 3$)



予測の悪さ: $\mathrm{AIC} = -2\log L^* + 2k$

AIC 最小のモデルを選ぶ

model	k	$\log L^*$	Deviance $-2\log L^*$	Residual deviance	AIC
constant λ	1	-237.6	475.3	89.5	477.3
f	2	-237.6	475.3	89.5	479.3
x	**2**	**-235.4**	**470.8**	**85.0**	**474.8**
x + f	3	-235.3	470.6	84.8	476.6
saturation	100	-192.9	385.8	0.0	585.8

AIC: A (or Akaike) information criterion

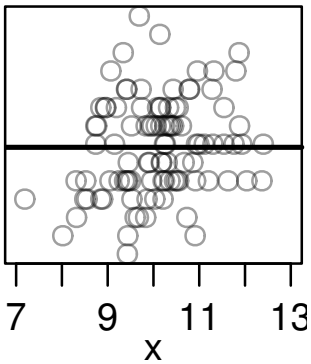
モデル選択 と統計学的検定 は

その目的がぜんぜんちがう

$\Delta D_{1,2}$ の分布を生成: ブートストラップ尤度比検定

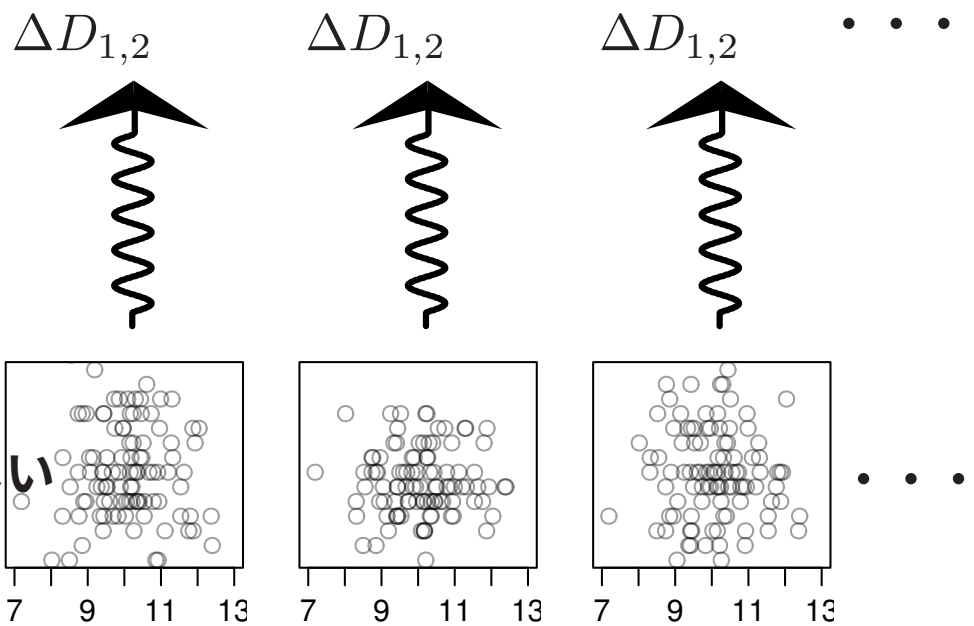
帰無仮説 が真のモデルであるとしよう!

帰無仮説が真の統計モデル
ということにしてしまう
($\hat{\beta}_1 = 2.06$ のポアソン分布)



帰無仮説のモデルから新しい
データをたくさん生成する

評価用データに constant λ と x model
をあてはめて逸脱度差 $\Delta D_{1,2}$ の分布を予測



あてはまりの良さ評価用のデータ（多数）

統計モデリング入門 筑波大 (大塚) 集中講義 [06]

“割算” 回避のための統計モデル

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

筑波大集中講義 http://goo.gl/HvRhXn

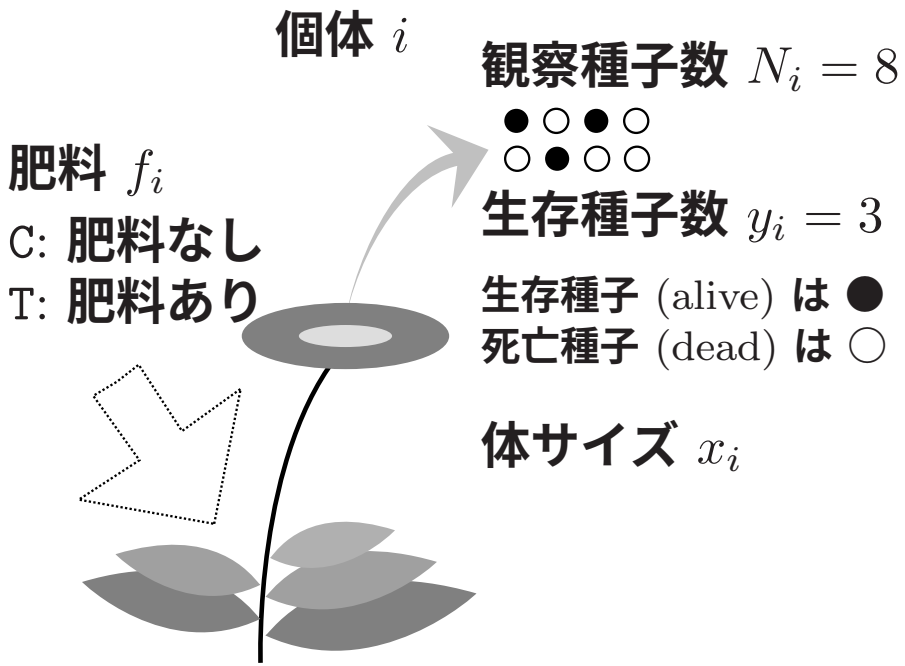
2015–03–01

ファイル更新時刻: 2015–02–27 12:50

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

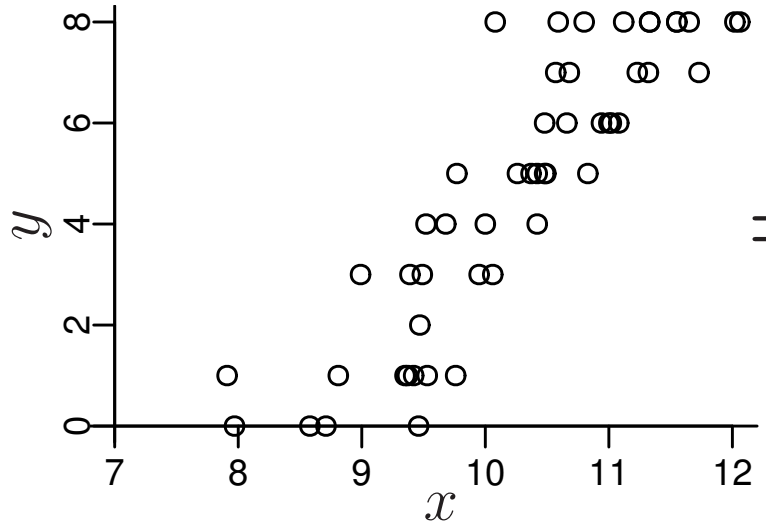
またいつもの例題? …… ちょっとちがう

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

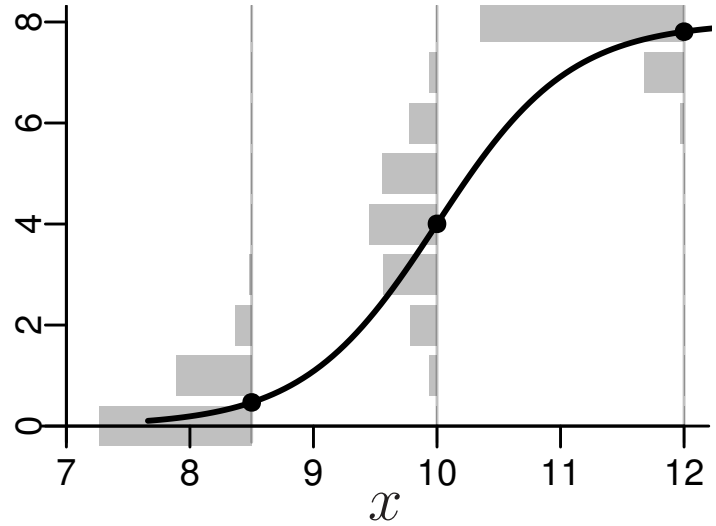


R でロジスティック回帰 — β_1 と β_2 の最尤推定

(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 筑波大 (大塚) 集中講義 [07]

一般化線形混合モデル (GLMM)

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

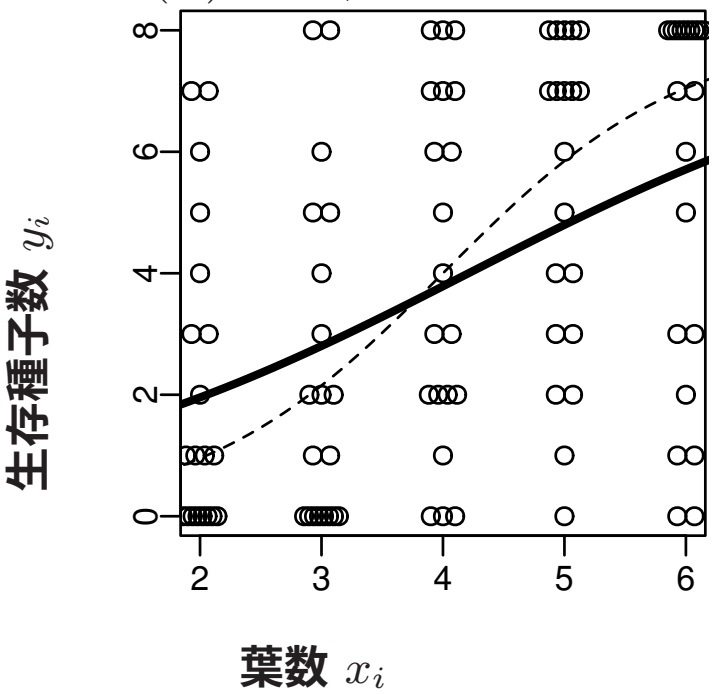
筑波大集中講義 http://goo.gl/HvRhXn

2015–03–01

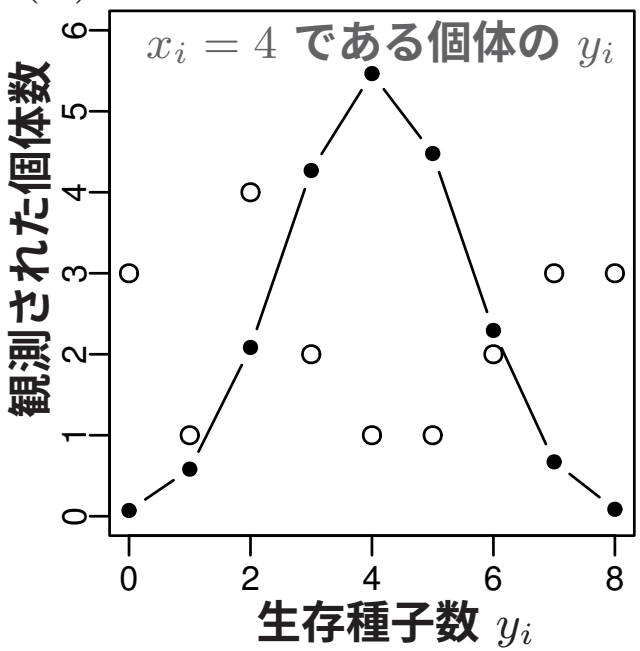
ファイル更新時刻: 2015–02–27 12:50

GLM では説明できないばらつき!

(A) 傾き β_2 が過小推定

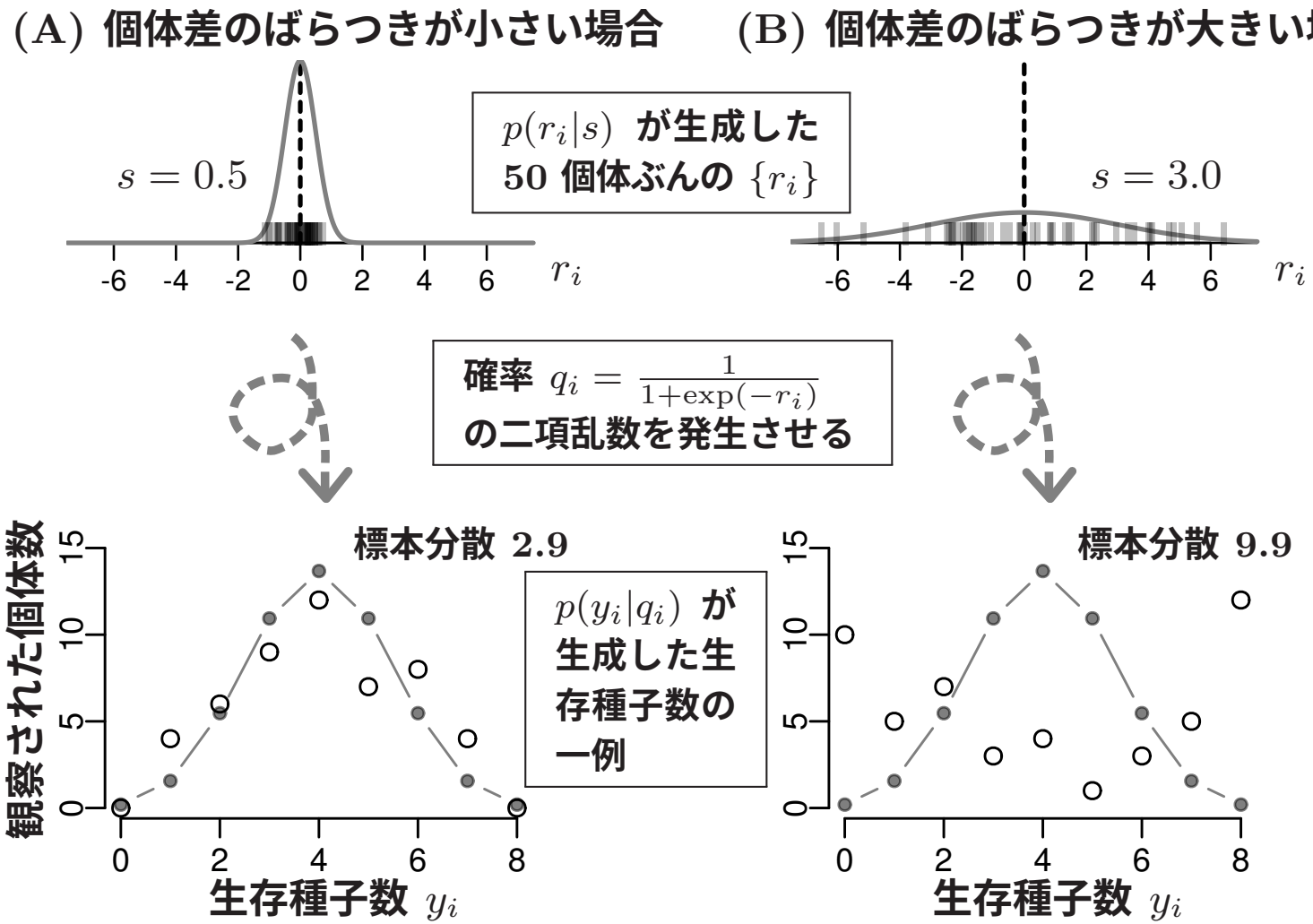


(B) ぜんぜん二項分布じゃない!

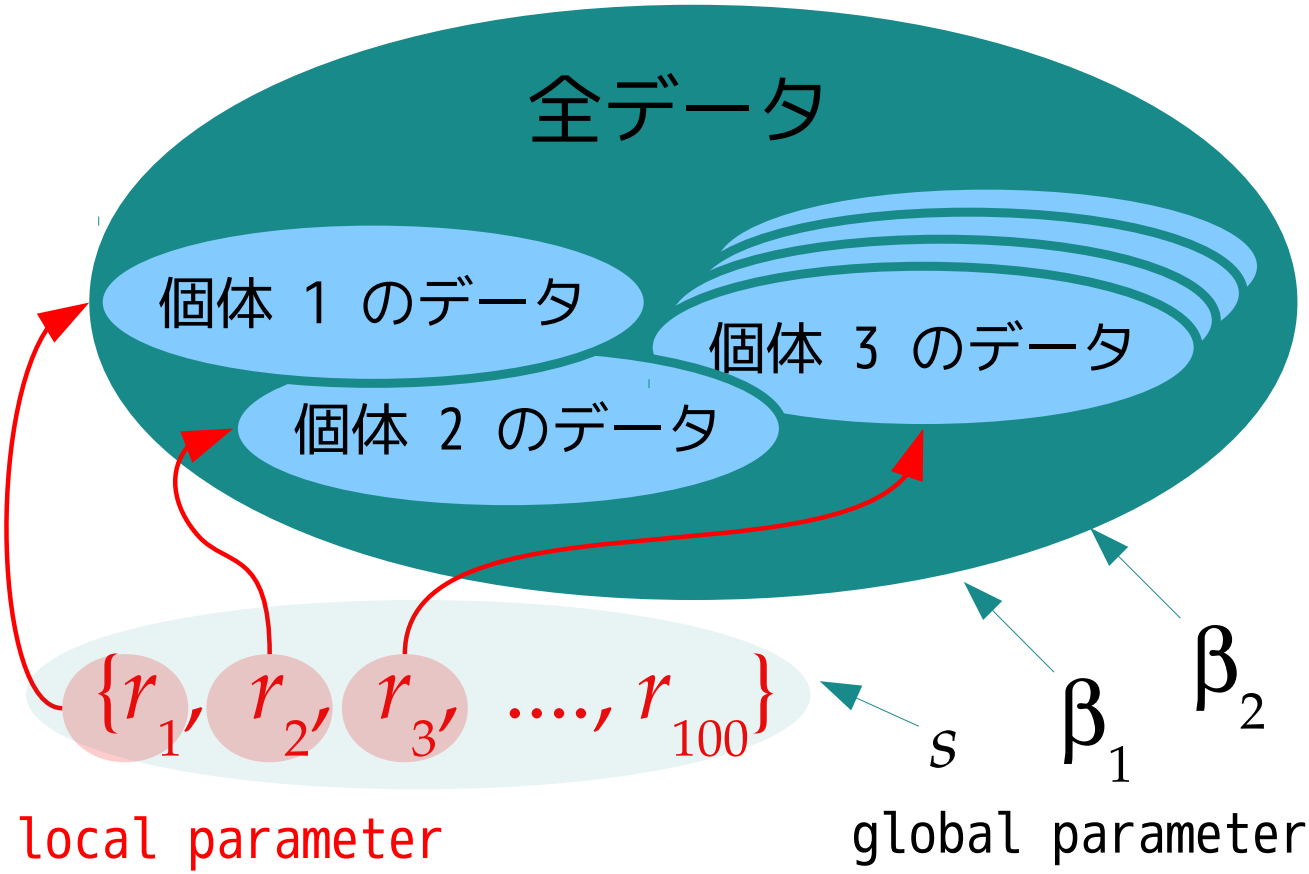


○が観測されたデータの図示

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 筑波大 (大塚) 集中講義 [08]
マルコフ連鎖モンテカルロ法

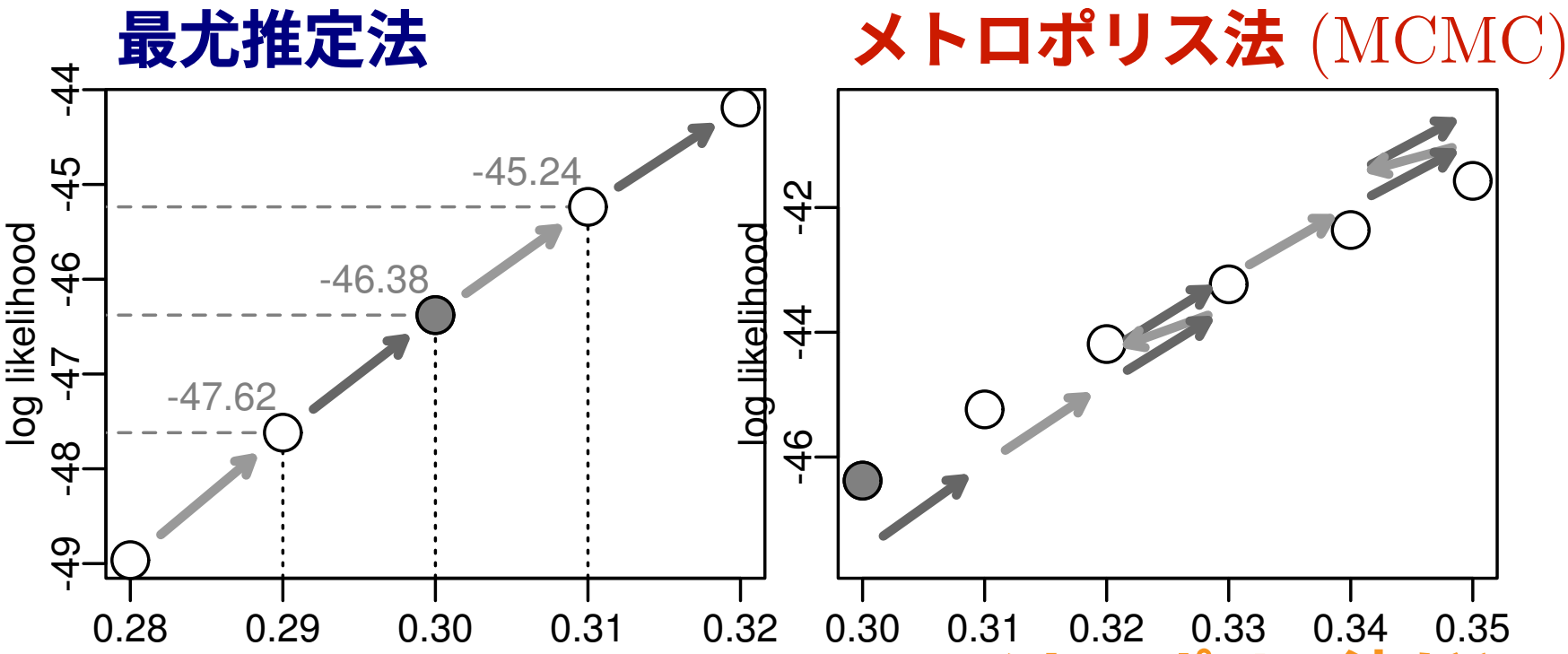
久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

筑波大集中講義 http://goo.gl/HvRhXn

2015–03–01

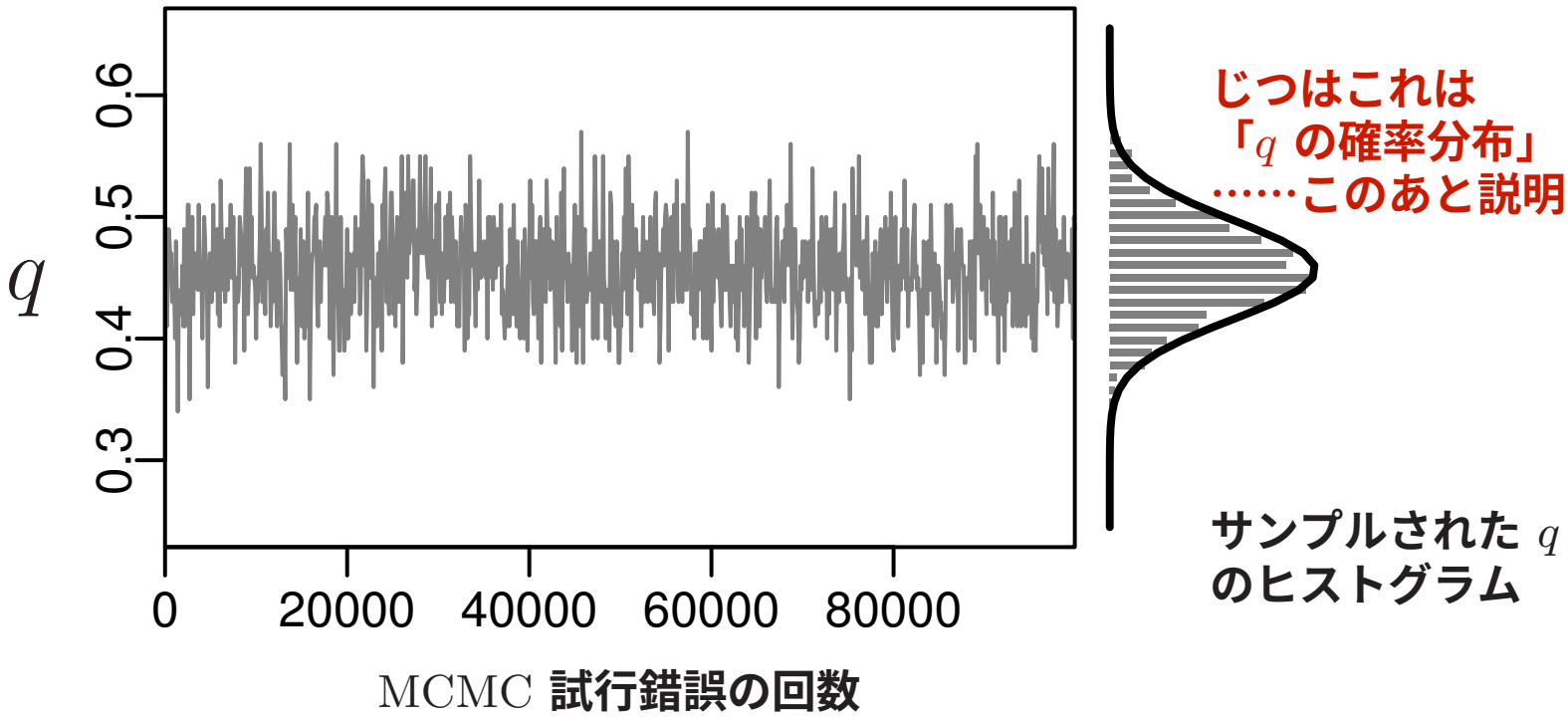
ファイル更新時刻: 2015–02–27 12:50

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

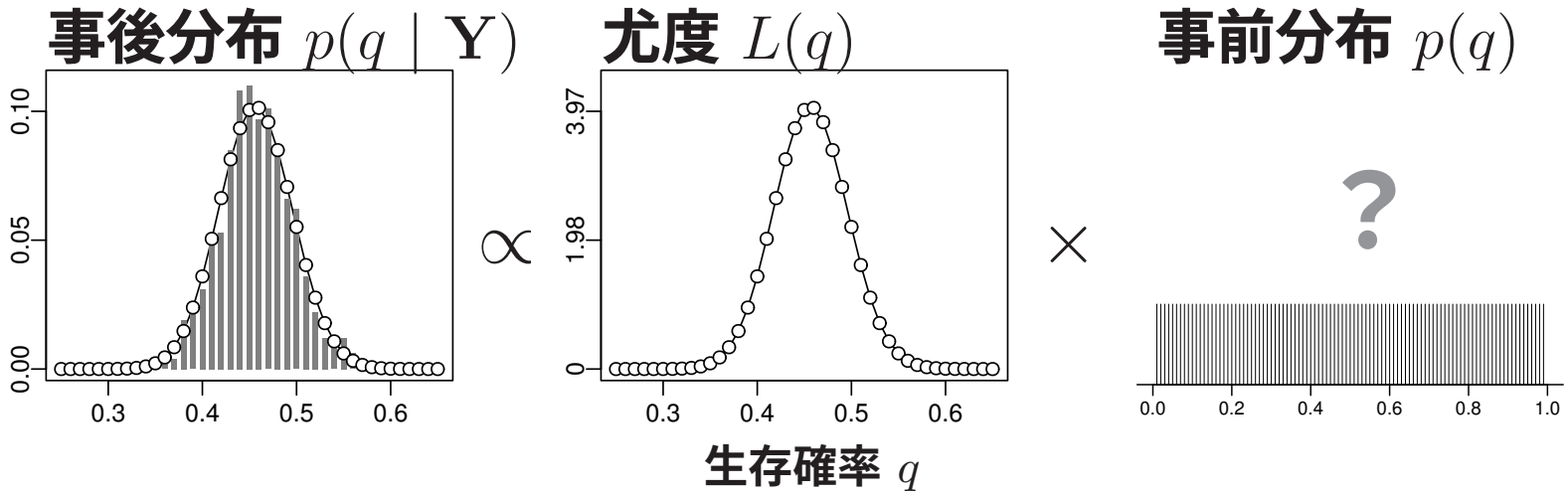
もっともっと長くサンプリングしてみる



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?



事前分布ってのがよくわからない……

統計モデリング入門 筑波大 (大塚) 集中講義 [09]

階層ベイズモデル

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

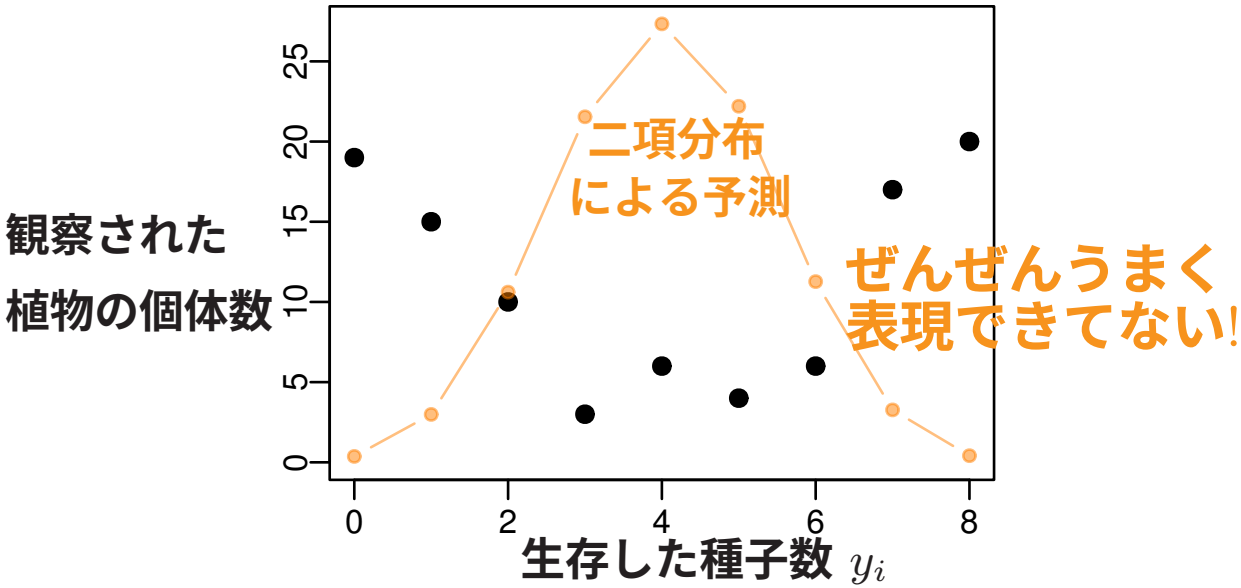
筑波大集中講義 http://goo.gl/HvRhXn

2015–03–01

ファイル更新時刻: 2015–02–27 12:50

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……

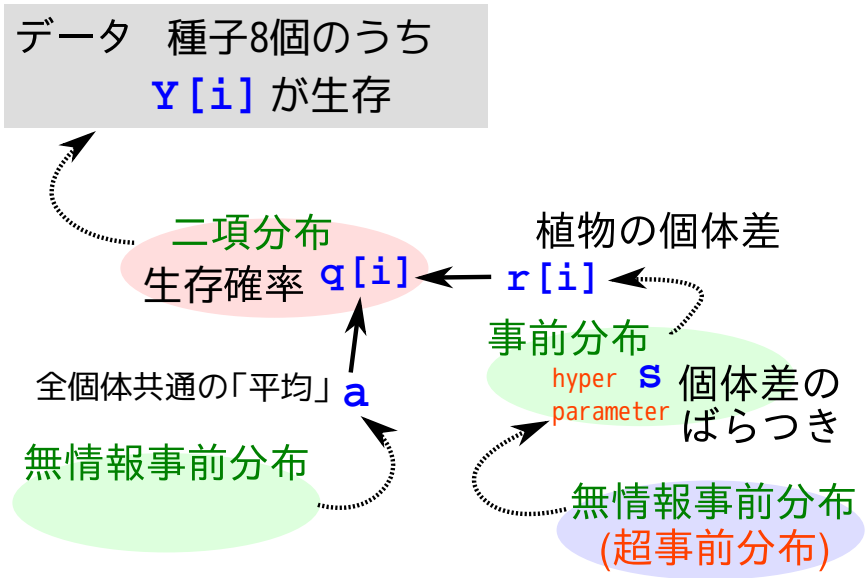


さっきの例題と同じようなデータなのに?

(「統計モデリング入門」第 10 章の最初の例題)

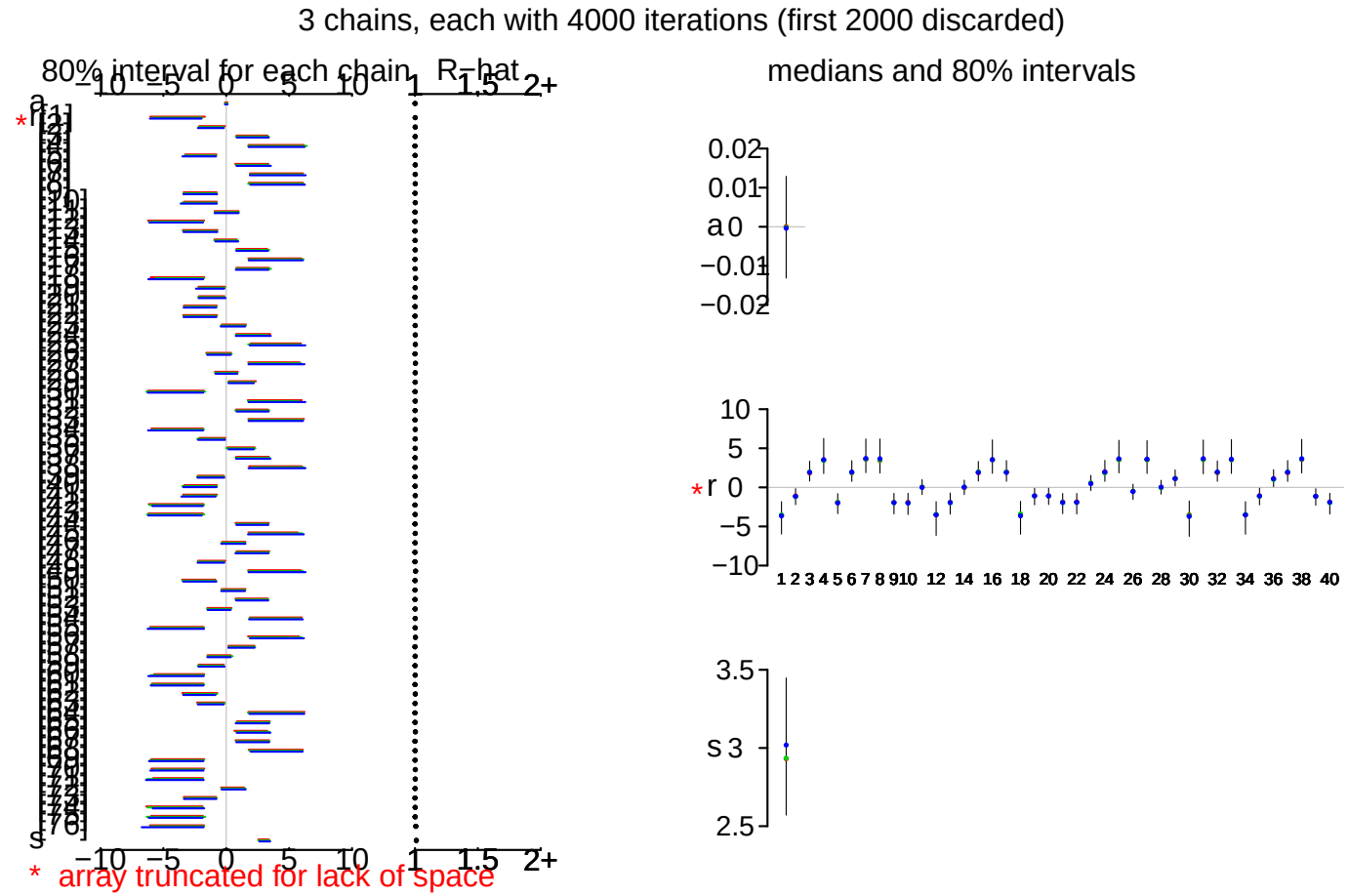
階層ベイズモデルを BUGS コードで記述する

```
model
{
  for (i in 1:N.data) {
    Y[i] ~ dbin(q[i], 8)
    logit(q[i]) <- a + r[i]
  }
  a ~ dnorm(0, 1.0E-4)
  for (i in 1:N.data) {
    r[i] ~ dnorm(0, tau)
  }
  tau <- 1 / (s * s)
  s ~ dunif(0, 1.0E+4)
}
```



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
```



統計モデリング入門 筑波大 (大塚) 集中講義 [10]

階層ベイズモデルの応用例など

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

筑波大集中講義 http://goo.gl/HvRhXn

2015–03–01

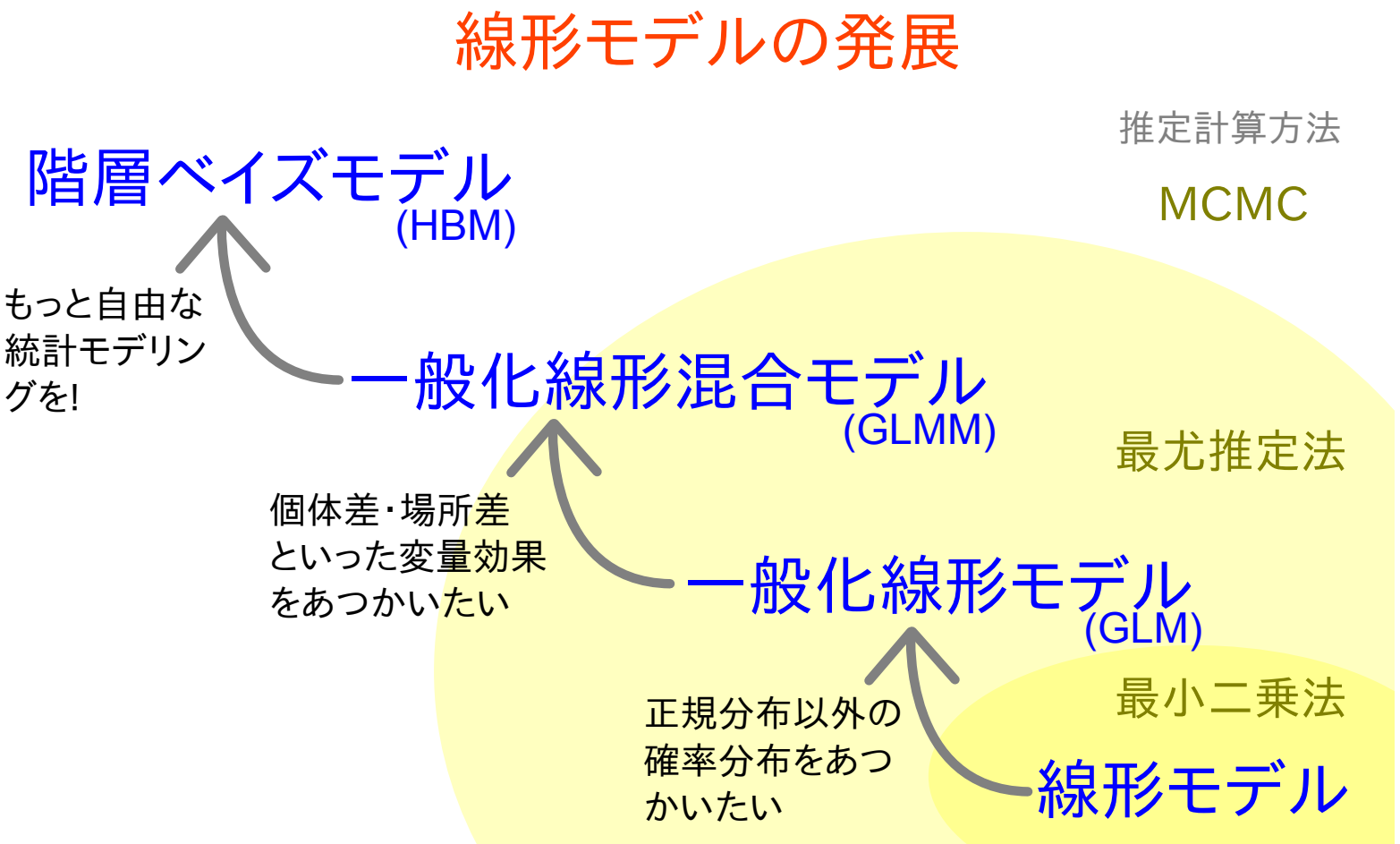
ファイル更新時刻: 2015–02–28 01:18

階層ベイズモデルの例題を何か紹介したい……

どれにするかはまだ決めてません

1. 分割表の問題と統計モデリング
2. 選択・勝敗の階層ベイズモデル
3. 連続値「脱」割算値の統計モデリング
4. 時系列データの階層ベイズモデル?

“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