生態学基礎論 (生物多様性論 II)

5. 生物多様性解析法：統計モデリングの基礎

全部で 2 回講義の 2

random effects
「個体差」を階層ベイズモデルであつかう

個体差・ブロック差の random effects

http://hosho.ees.hokudai.ac.jp/~kubo/stat/2006/

(ここから授業で使った PDF ファイルなどをダウンロードできます)

講釈: 久保拓弥 kubo@ees.hokudai.ac.jp

この 2 回だけの統計学授業でやること

- 自然科学の データ解析 に統計学は必要不可欠
- しかし多くのユーザーは よくわからん 状態で使ってる
- この授業の目的はその「わからん度」を少しだけ下げる こと

- **第 1 回**: 2007-01-22 (**月**)
  「数えられる」データの統計解析・統計モデリング

  観測データを一般化線形モデル (GLM) 化しよう
- **第 2 回**: 2007-01-24 (**水**)
  「個体差」を階層ベイズモデルであつかう

  個体差・ブロック差の random effects

前回までのあらすじ: logistic 回帰を最尤推定で

1. 「統計学って何 ?」を理解する

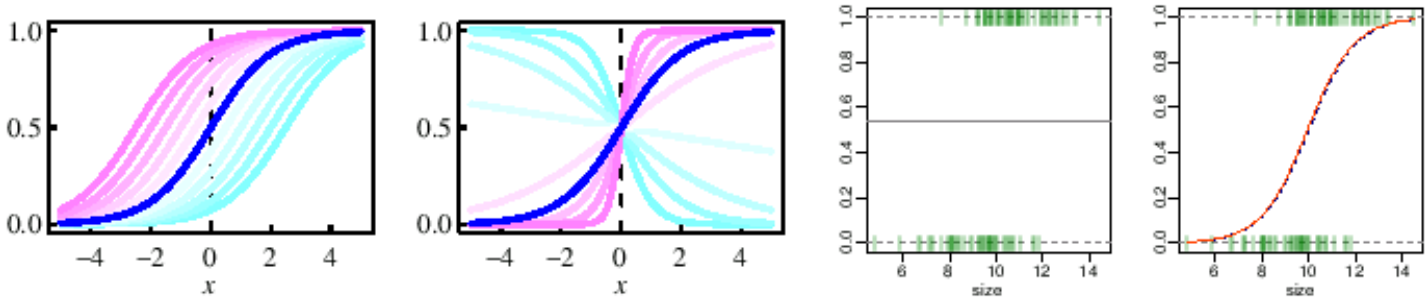
データ解析とはモデリングによる情報圧縮

2. 最尤推定法とロジスティック回帰

カウントデータは，まず `glm()` で!

3. さらに強めるロジスティック回帰わざ

割り算値解析しない，データにあわせたばらつき (確率分布) を



今日のハナシ: データ中の「個体差」にいどむ

1. `glm()` がうまくいかない状況?

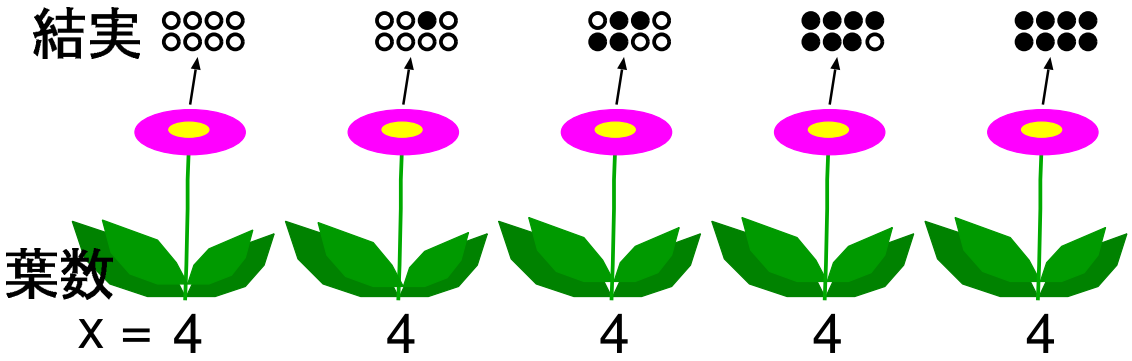
   原因: 「個体差」による過分散 (overdispersion)

2. 屋久島照葉樹 22 樹種のデータ解析

   たくさんの「種差」のモデリング

3. 階層ベイズモデルを MCMC 計算であつかう

   「種差」「個体差」そして環境の影響を同時に推定



エンドユーザーからみた統計学ツール「含有関係」

(一般化) 線形モデル的に現象を表現する場合

[尤度をあつかうモデル]

「すべてのパラメーターは確率分布」とする Bayes 統計学

階層 Bayes モデルなどなど

[最尤推定法 であつかうモデル]

パラメーター (fixed + random effects) は特定の値

経験 Bayes 法や一般化線形混合モデル (GLMM) などなど

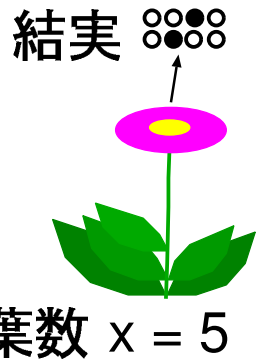
[一般化線形モデル (GLM)]

指数関数族の確率分布 + 線形モデル, fixed effects のみ

[最小二乗法 であつかうモデル]

等分散正規分布 + 線形モデル

直線回帰，いわゆる「分散分析」などなど



すごく単純化した状況なのに

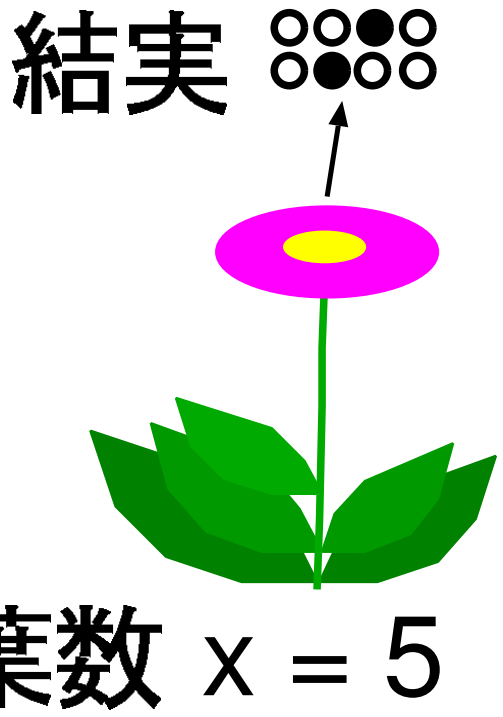
1. glm() がうまくいかない状況

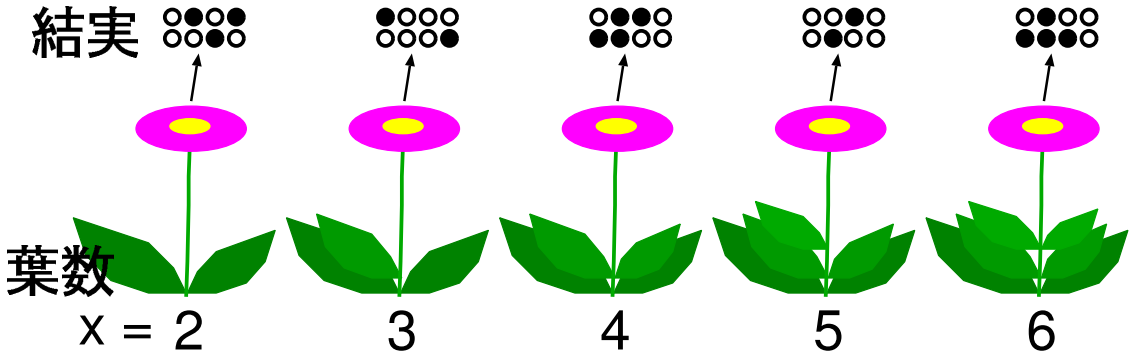
データを何回とりなおしてもダメ?

架空植物: 胚珠が種子になる確率を知りたい

[架空植物の性質あれこれ]

- 一個体にひとつの花
- 花の胚珠数 (最大種子数) は 8
- 結実率 p: ある胚珠が種子になる確率
- 「個体差」(?) とやらが大きいらしい (しばらく ? で表現)

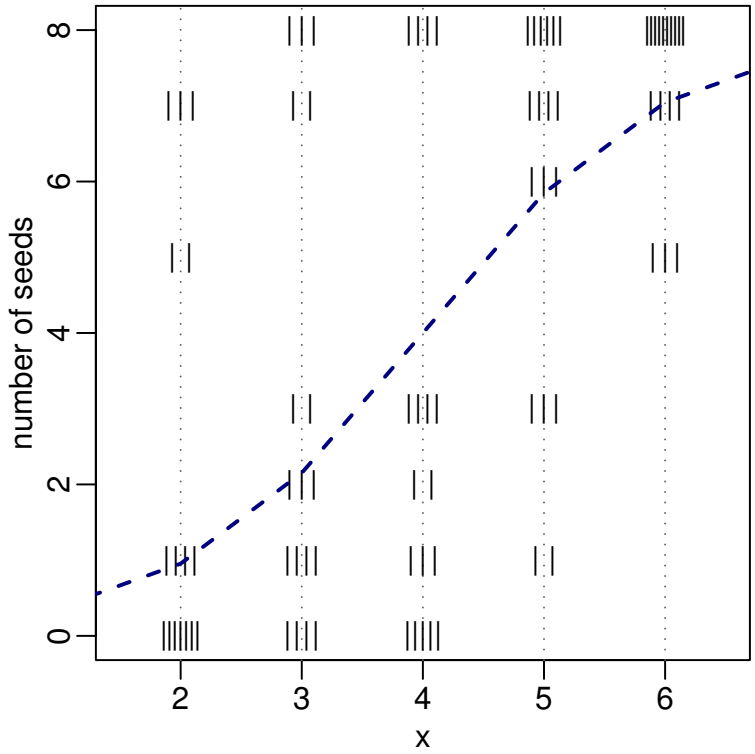


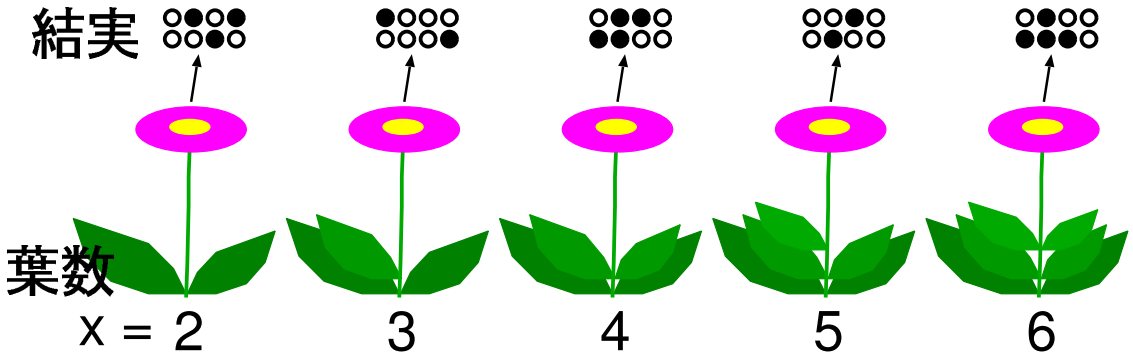


- 葉数 x は個体ごとに 2-6 枚
- 葉数が結実率を決めるらしい

問: 結実率 p は葉っぱの枚数 x でどう変わるか?

葉数 x と種子数 (標本個体数 100)





["神" の立場で知ってるコト]

- 葉数 x 大きいほど結実率が高い

$$p(a,b) = \frac{1}{1+\exp(-(a+bx+?))}$$

$a = -4$ かつ $b = 1$ である

[観測者 (人間) が知りたいコト]

葉数パラメーター $b = 1$ を正しく推定したい

「個体差」とやらは
とりあえず無視

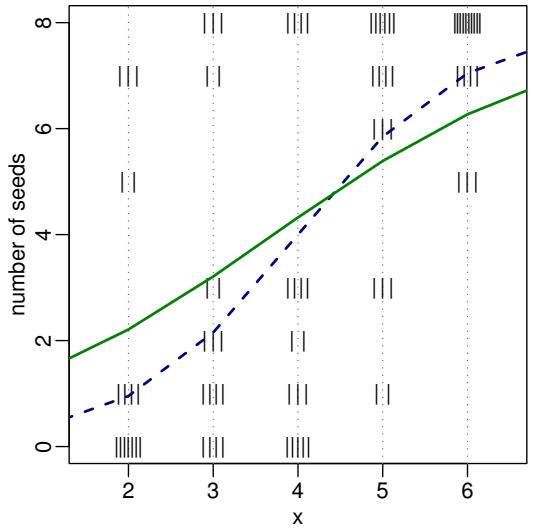
(あとで解説する)

こんなのロジスティック回帰で簡単に……?

```
> d <- read.csv("d.csv") # データファイル d.csv を読みこむ
> summary(glm(n.seed ~ 1 + x, family = binomial, data = d))
Call:
glm(formula = n.seed ~ 1 + x, family = binomial, data = d)
...(略)...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0888     0.2359   -8.86   <2e-16
x             0.5627     0.0569    9.89   <2e-16
...(略)...
```

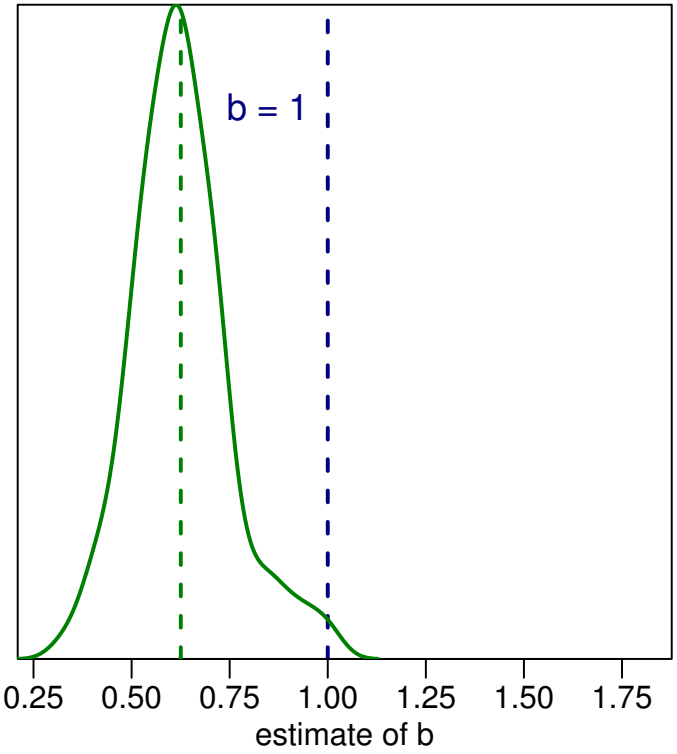
なンか葉数パラメーターの推定値がずれてるんですけど…… $\hat{b} = 0.5627$

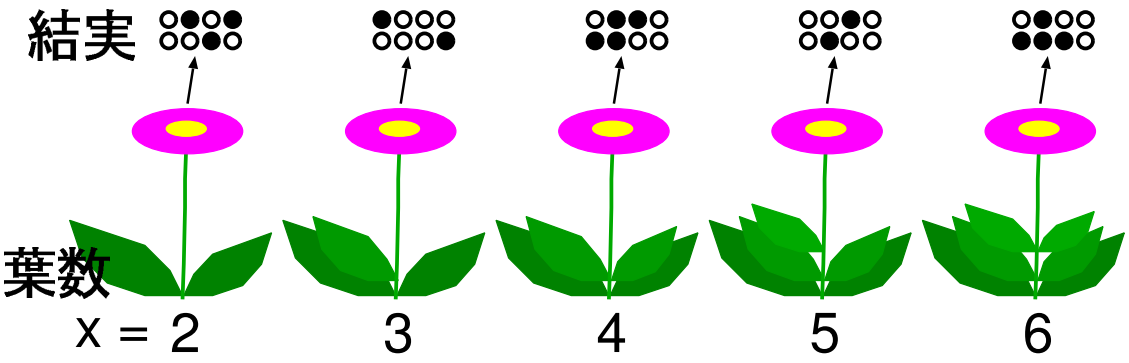
(真の値は $b = 1$)



ダメな推定は何回データとりなおしてもダメ

- 「データが悪い」と思ってまたべつの集団から種子データとりなおしてみた
- `glm(n.seed  1 + x, ...)` やりなおしてみた (標本数 100 個体 × 8 胚珠)
- これを何度も何度も……
  100 回ほど繰りかえしてしまった
- `glm()` では何回やりなおしてもダメ
  …… 偏り (bias) のある推定方法だ





どうして問題はこんなに簡単なのに，$b = 1$ から偏ってしまうのか?

そもそもこの観測データ，ばらつきすぎでは?

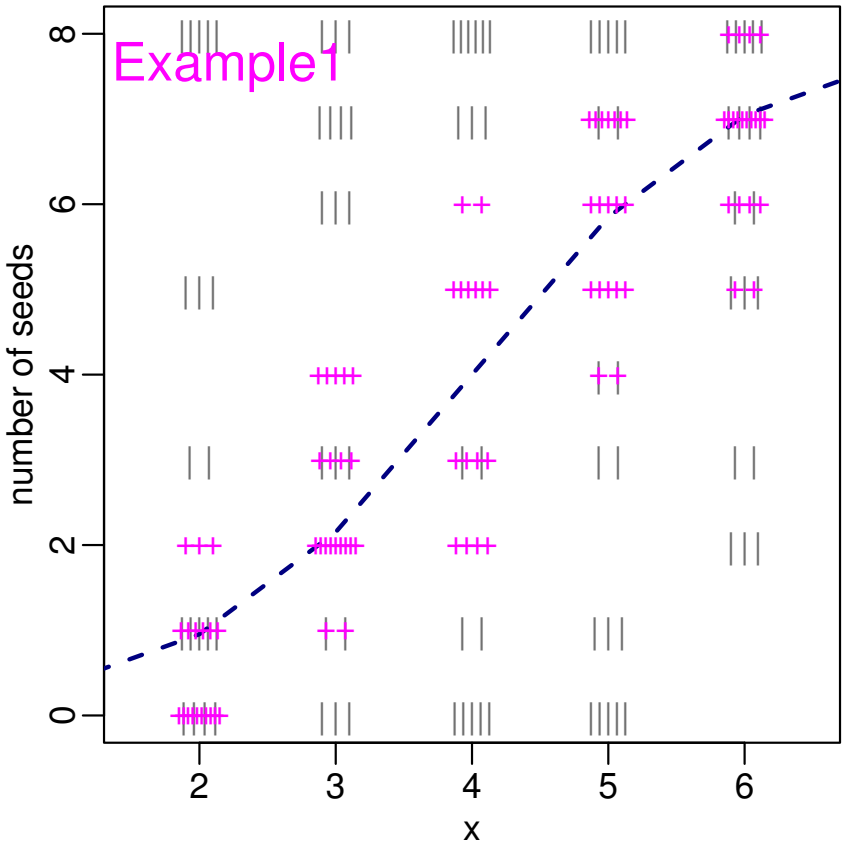
- 気をとりなおして原点から考えなおす
- そもそも logistic 回帰ではデータが**二項分布** (binomial distribution) になることが前提
- 8 個の種子のうち y 個が結実する確率は

$$\frac{8!}{y!(8-y)!}p^y(1-p)^{8-y}$$

(注) $8! = 8 \times 7 \times \cdots \times 2 \times 1 = 40320$

- [R による実験で調べる] 結実率を ……

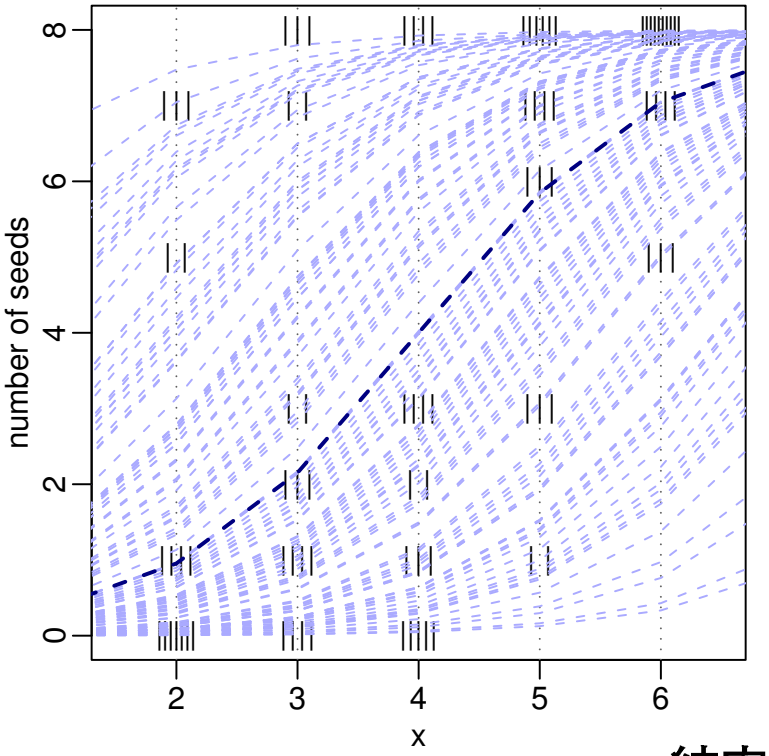
$$p = \frac{1}{1+\exp(-(-4+1\times x)}$$



- ……とおいて R で二項乱数を発生させる `rbinom(100, 8, prob = p)`
- 観測データの図のうえに発生させた二項乱数でシミュレイトした**種子数**を表示させた
- これは「ホントの期待種子数」 (**破線**) まとわりついているのに……?

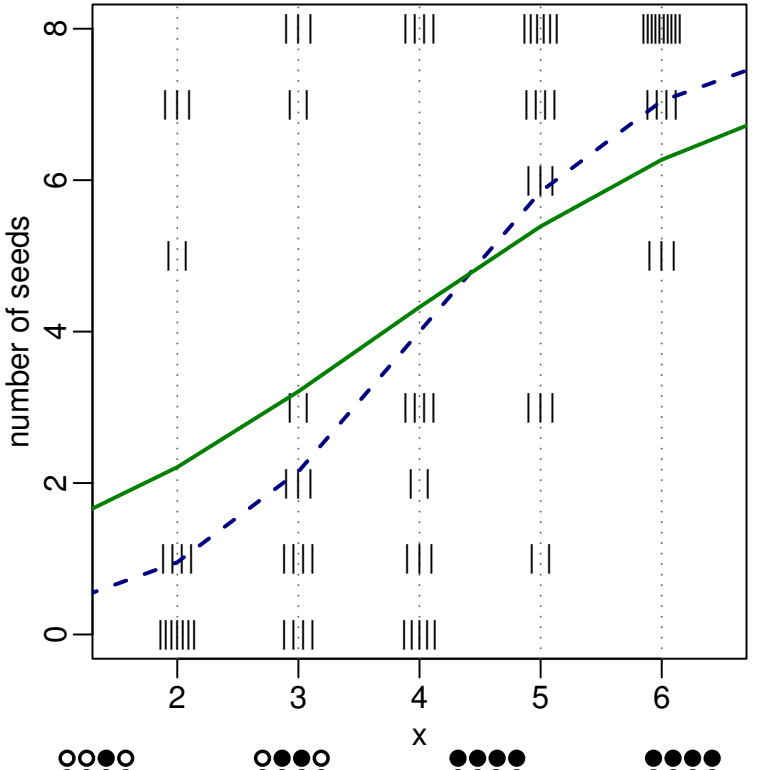
個体ごとの「ずれ」が b の推定を偏らせた

「傾き」 (b) には「個体差」ない
「切片」 (a) には「個体差」ある



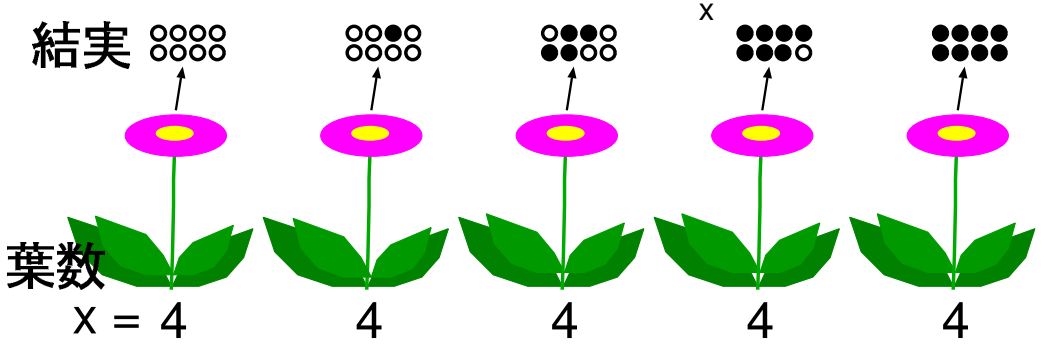
→

「個体差」を考慮してない `glm()` による「なだらかな」推定結果



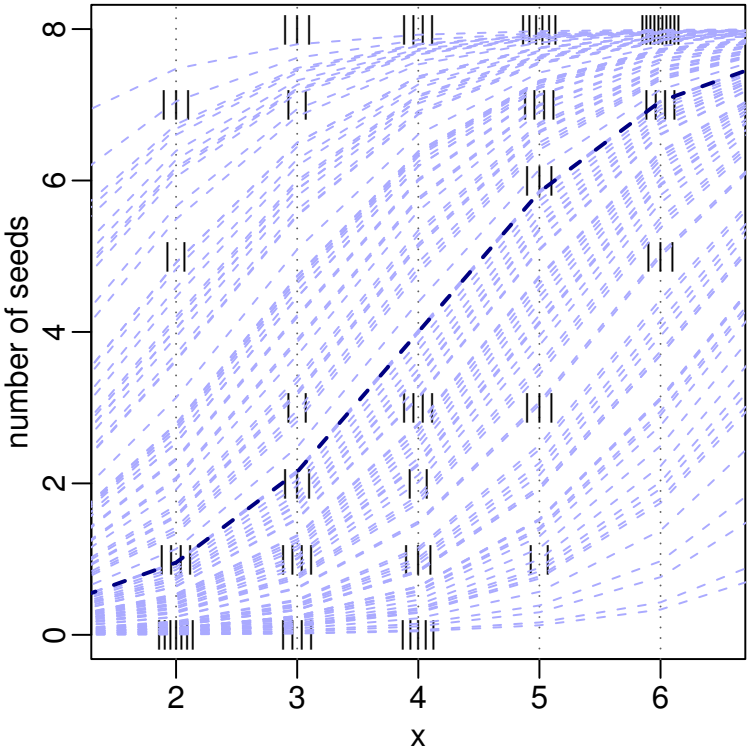
葉数 x は同じでも個体ごとの結実率は異なる!

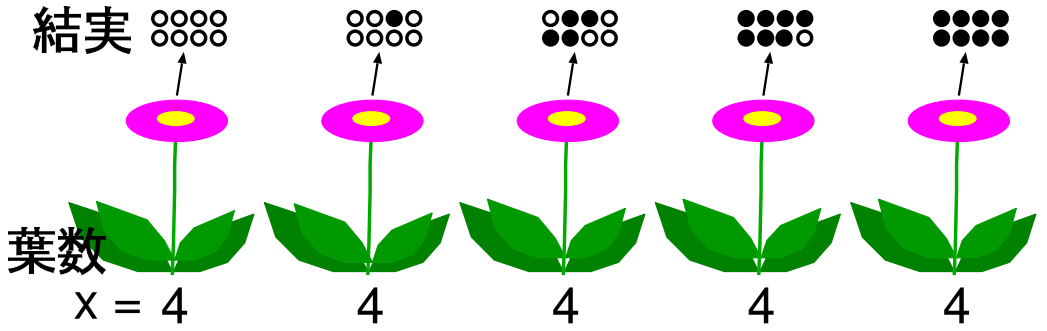
しかし集団全体で平均すると……?



ここでいう「個体差」とは何か? (生物学的側面)

- 結実率の曲線の「ずれ」は観測者が**観測していない**そして「興味のない」量である．ここではこのずれを**「個体差」**とよぶ
- 葉数 x なども個体ごとに異なっているけれど，これは観測した (そして興味のある) 数量なので，ここでは「個体差」とは**よばない**





「個体差」生じる生物学的原因

遺伝的要因，土壌中の栄養塩類，
日あたり，訪花昆虫の努力……
などなど (原因不明なことも多い!)

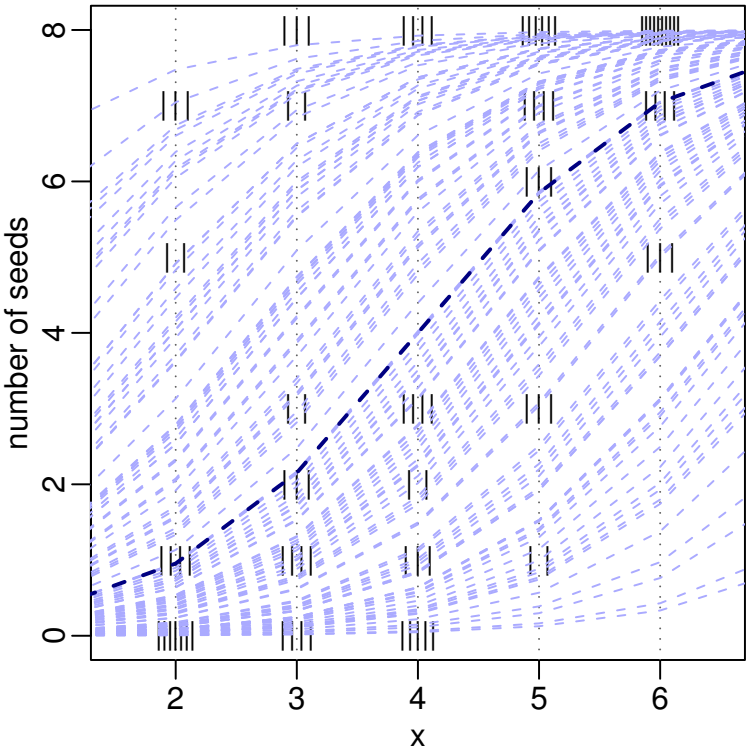
「個体差」とは何か? (統計モデリング的側面)

- 結実率を表現する logistic 曲線

$$p(a, b) = \frac{1}{1 + \exp(-(a + bx + ?))}$$

線形部分 $a + bx + ?$ に注目する

- $a + bx$ は p の平均値を変化させている → "fixed effects"



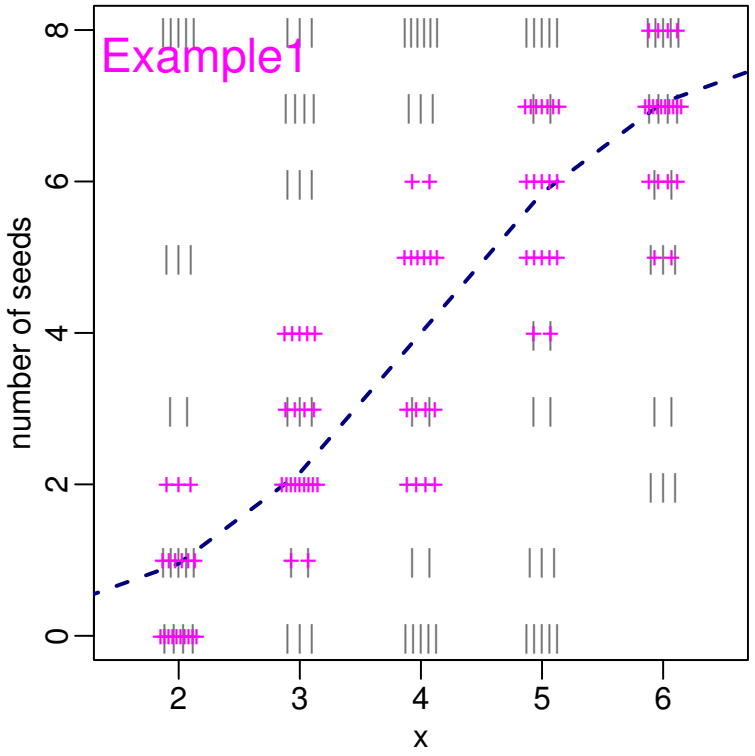
- ? は p の平均値を変化させず，ばらつきだけを変えている
  → "random effects"
- fixed effects と random effects を両方ふくむ統計モデルを混合モデル (mixed model) とよぶ

Random effects がもたらす過分散 (overdispersion)

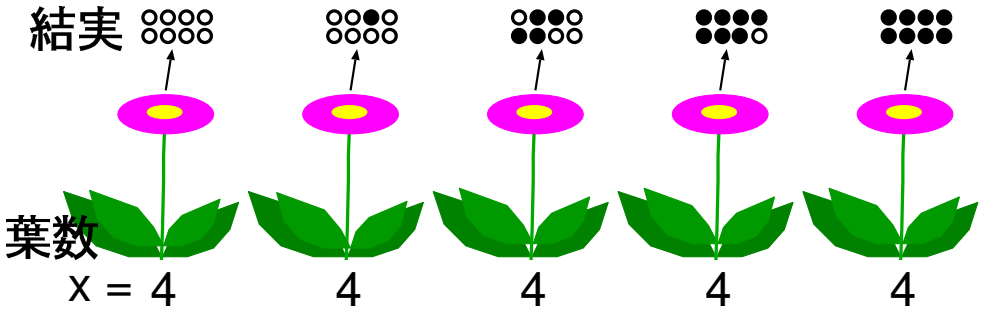
- Random effects なしの結実率モデル

$$p(a,b) = \frac{1}{1+\exp(-(a+bx))}$$

を仮定して二項乱数を生成させると + のようなデータが得られる



- しかしながら架空植物からの観測データ ||| は「二項乱数ではありえない」ばらつきを示している (← random effects)
- これを**過分散** (overdispersion) という
- 過分散を発見したら,「個体差」無視できないと考える

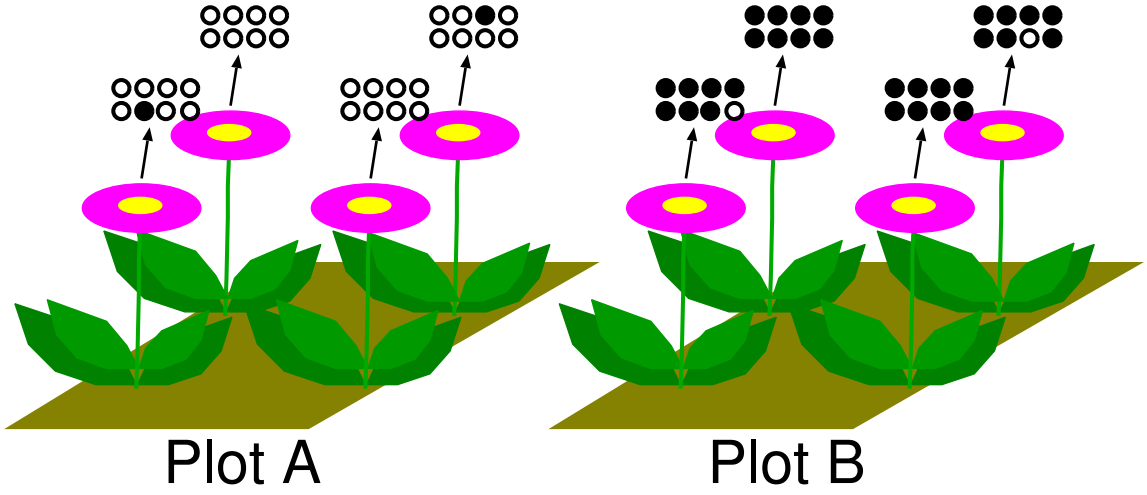


個体差考慮した GLMM のモデリング・推定計算は?

- R の glmmML() 関数など使って推定計算する
- 今日は解説しない (脱力)
  - 理由: GLMM の「その先」にススむから
- 昨年の授業で詳しく解説
  - http://hosho.ees.hokudai.ac.jp/~kubo/stat/2005/
- 昨年の生態学会大会でも詳しく解説
  - http://hosho.ees.hokudai.ac.jp/~kubo/ce/2006/
- この問題をさらに詳しく解説した PDF ファイル
  - http://hosho.ees.hokudai.ac.jp/~kubo/ce/GlmmEsj2006.html

Random effects ... 「個体差」だけじゃない!

- ここで「個体差」と呼んでる random effects が表現できることは「各個体で観測されなかった差位」だけではない
- たとえば下の図のような「ブロック差」もモデル化できる



- さらに「ブロック差ある中のブロック内個体差」もモデル化できる (推定計算はすごくしんどい)
- さらにこの考えかたは空間相関ある場合の推定にも応用できる

エンドユーザーからみた統計学ツール「含有関係」

(一般化) 線形モデル的に現象を表現する場合

[尤度をあつかうモデル]

「すべてのパラメーターは確率分布」とする Bayes 統計学

階層 Bayes モデルなどなど

[最尤推定法 であつかうモデル]

パラメーター (fixed + random effects) は特定の値

経験 Bayes 法や一般化線形混合モデル (GLMM) などなど

[一般化線形モデル (GLM)]

指数関数族の確率分布 + 線形モデル, fixed effects のみ

[最小二乗法 であつかうモデル]

等分散正規分布 + 線形モデル

直線回帰，いわゆる「分散分析」などなど

今日でてくるベイズ用語の整理

(事後分布) $\propto$ (尤度) $\times$ (事前分布) $\times$ (超事前分布)

- **階層ベイズモデル** $p(\beta, \alpha|y) \propto p(y|\beta)p(\beta|\alpha)p(\alpha)$
  - 推定計算方法: Markov Chain Monte Carlo (MCMC) 法
    - MCMC 計算わざ 1: Metropolis-Hastings 法
    - MCMC 計算わざ 2: Gibbs sampler

    (上のふたつは本質的には同じもの)

- **経験ベイズ法** $p(\beta, \alpha|y) \propto \int p(y|\beta)p(\beta|\alpha)\mathrm{d}\beta$
  - 推定計算方法: α, β の点推定 (最尤推定)
    - 例: 一般化線形混合モデル (GLMM)
  - 簡単化した階層ベイズモデル，と考えるべきか?

(参照: 石黒ほか. 2004. 階層ベイズモデルとその周辺)

あつかう例題をちょっと変えて……

2. 屋久島照葉樹 22 樹種のデータ解析

たくさんの「種差」のモデリング

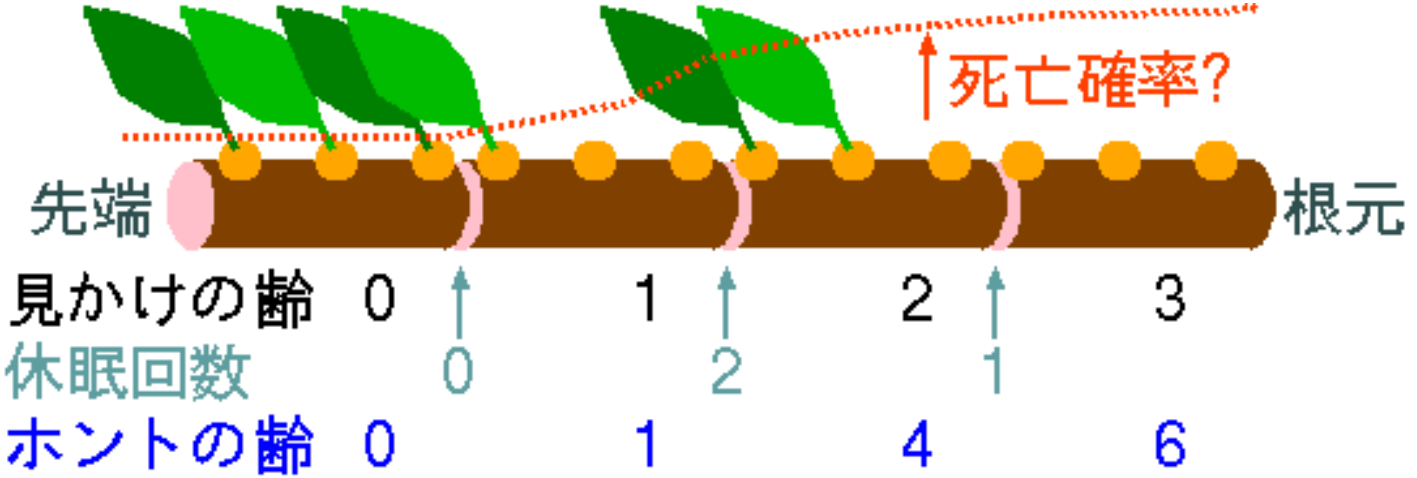
昨年度の
修士論文発表
(牛原阿海さん)

葉寿命と明るさ・葉内窒素濃度の関係を調べたい

- 調査地は屋久島の西部林道川原と花山歩道
  - それぞれの場所で $\{\text{open, close}\}$ の両方あり
- 対象樹種: 22 樹種 — 芽鱗痕明瞭な高木，亜高木，低木
- サンプリング対象個体: 樹高 5m ぐらいまでのもの
- 測定項目
  - **葉齢分布**: 1 樹種 3 個体　1 個体 1 シュート，年枝ごとに測定
    - 葉数, 葉痕数, 葉乾重, 葉面積, CN 比, 枝長, 枝直径
    - 2004 年 10 月
  - **二度伸び・休眠の調査**: 1 樹種数シュート (樹種による)
    - マーキングにより二度伸びと休眠調査
    - 2004 年 10 月と 2005 年 3, 7, 10 月

問題: 葉っぱの死亡確率は何で決まるか?

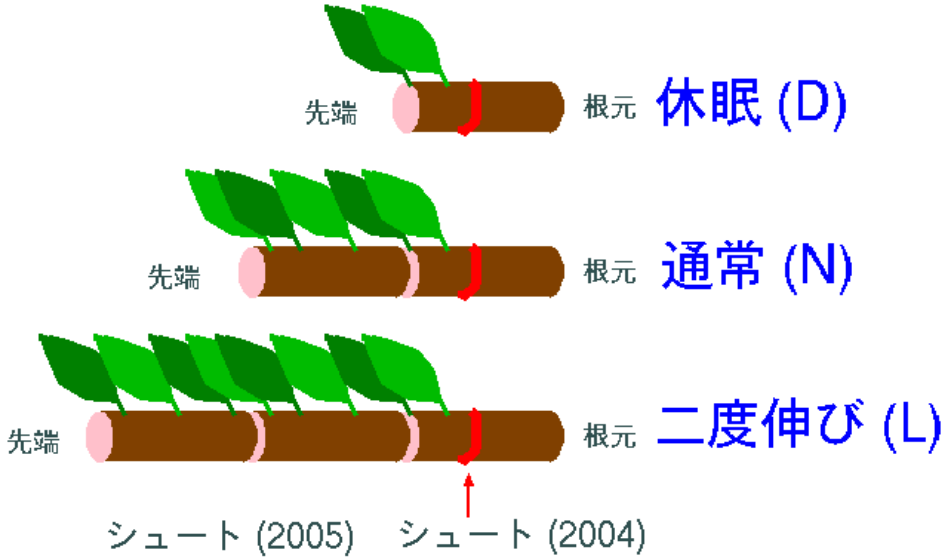
要因: 葉齢? 明るさ? 葉内窒素濃度? 樹種差? 個体差?



- しかし今日は葉寿命問題は**あつかわない** (脱力)
  - 理由: 複雑で難解だから

今日の問題: シュート伸長は何で決まるか?

要因: 明るさ? 樹種差? 個体差?



問題: シュート伸長の休眠・二度伸びは何で決まるか?

多項 logistic モデル，p: 休眠確率 (年 $^{-1}$)，q: 二度伸び確率 (年 $^{-1}$)，

$$p = \exp(\beta_{\mathrm{CD}} + \beta_{\mathrm{LD}} L)/Z, \qquad L: \text{暗} \to 0, \text{明} \to 1.$$

$$q = \exp(\beta_{\mathrm{CL}} + \beta_{\mathrm{LL}} L)/Z,$$

$$Z = 1 + \exp(\beta_{\mathrm{CD}} + \beta_{\mathrm{LD}} L) + \exp(\beta_{\mathrm{CL}} + \beta_{\mathrm{LL}} L).$$

観測データ: 樹種・場所ごとに {D, N, L}

(species)	(dark)	(light)
adeku	DDDNNNN, NNNNN, NN	NL, DDNNNN, DDNNNNN
baribarinoki	NNN, DN, NNN	DNNNNNNNNNN, DNNLLL
hisakaki	DDDNNL, DDDDNNNN, DDDDN	NNNL, NNNLLL, DDNNNLL
hosobatabu		NN, DNNNNN
inugashi	DDDDNNN, DDDDNN, DDDDNNN	DNNNNN
kurobai	DDNNNNN, NNNNNNLL, DNNL	NNNN, NNNNNN
kuroganemochi		DNNNNL, NNNNN, NNNN
kuroki	NNNNNN, DNNNN, DDDNN	NNNNN, DNNNN, NNNNN
matebashii	DDDD	DDDN, NNNN
mimizubai	NNN, DNNN, DNNNNN	
nagi	DDDDD, DDDNNN, DDNNNN	
onikuroki	DNNNN, DNN, DDNN, DNNNN	DNNN, NNNL, NNNNNNN
sakaki	DNNNN, DDDDNN, NNNL, DDNNNN	NNNNN, NNNNNNNNN
sakuratutuzi	NNNNNNN, NNNNNNN, NNNNNN	NNNNNN, NNN, NNNNNN
sanngozyu	DNNNNL, NNNNN, DDDL, NNNNN	NNN, DN
sazannka	DDDNNN, DDNNN, DDDNNN	DNNNNNN, NNNNNN, NNNLL, NNNN
sikimi	DNNNNN, NNN, DNNNNN	NNNNNNNNNNLL, NNNNL, NNNNNL
sudazii		DD
tabunoki		DDDNNNN, DDD, NNN
taiminntachibana	NNNNN, NNNNNN, NNNNNN, NNNNNL	NNNNLL, NNNNNL
tubaki	DDDDDDDDDN, DDDDDNN, DDDDDNN	DDN, NNNNN, DNNN
urazirogashi	DDDNN, DDDNN, DDDDD	DDDNNN, DDDNNNNL

D: dormancy, N: normal, L: lammas shoot

- R の library(grid) による作表
- 562 シュート / 105 個体 / 22 樹種
- 観測地
  - dark : 林床
  - light : 林道わき
- シュートの種類
  - D: 休眠
  - N: 通常
  - L: 二度伸び
- 暗い場所では “D” が増えていそう?

面倒な多樹種各少標本 ……「甲山さんぷりんぐ」

- 葉齢分布: 1 樹種 3 個体　1 個体 1 シュート，年枝ごとに測定
  - ...
- 二度伸び・休眠の調査: 1 樹種数シュート (樹種による)
  - ...

もし各樹種独立にパラメーター推定やったら……

信頼できない推定値が多量にでてくる!

- 「甲山さんぷりんぐ」のウラをかくしかない
  - なぜ「多樹種各少標本」なのか?
  - 「あの場にいる木はどれも似ている部分がある」からなのか?
  - ならばそのように統計モデルを作ればよい
  - 各樹種独立の推定計算なんてことはやらない

Nest した階層ベイズモデル: 超樹種 - 樹種 - 個体

各パラメーターの分割: $\beta_x = \beta_x^{\text{Hyperspecies}} + \beta_x^{\text{Species}} + \beta_x^{\text{Individual}}$

(Hyperspecies)

- 全樹種に共通する傾向

(Species)

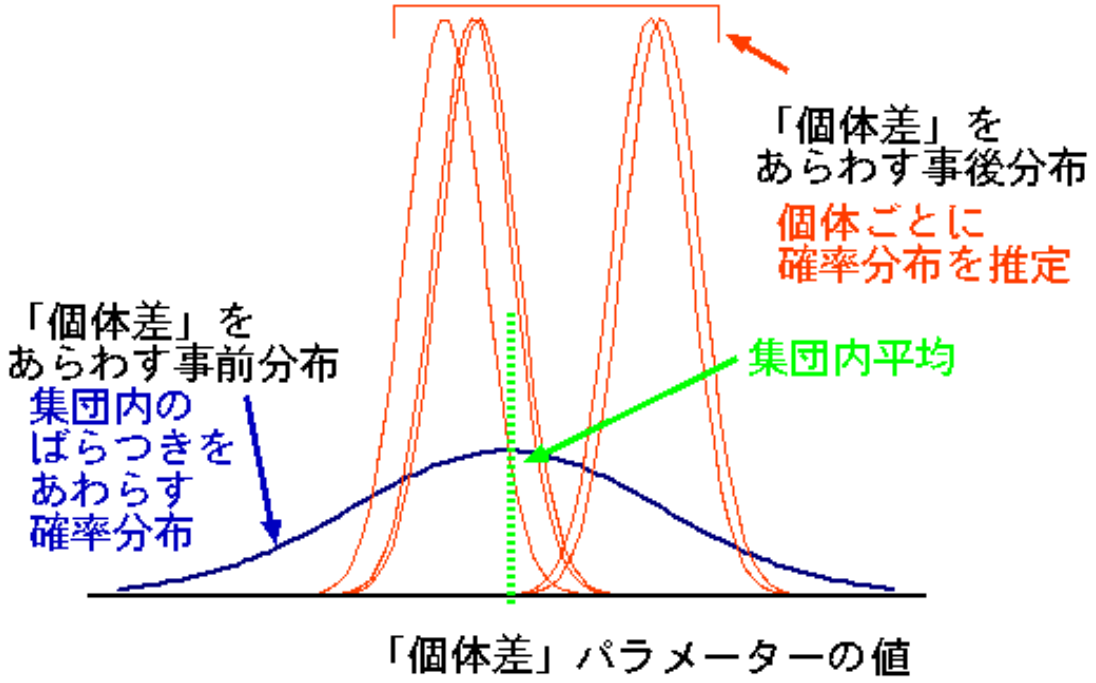
- 樹種差をあらわす層

(Individual)

- 「個体差」をあらわす層
  - 観測のずれなども含む

超樹種・樹種・個体すべてに事前分布が設定される

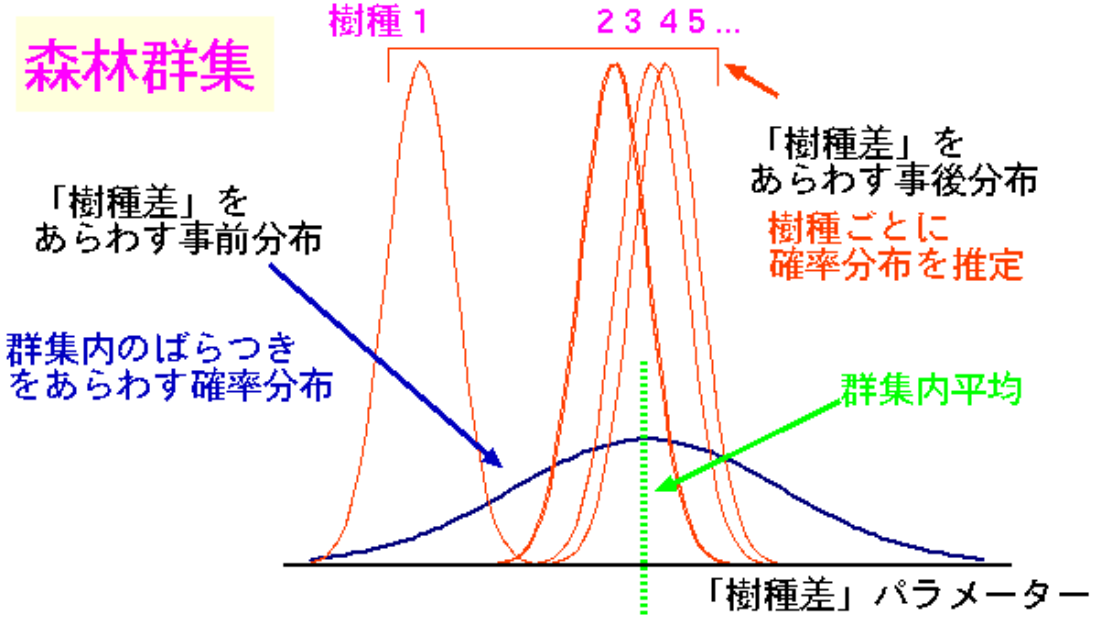
「個体差」: 樹木個体ごとの random effects



階層ベイズモデル

- 事後分布: 個体ごとの「個体差」をあらわす
- 事前分布: 集団内の「個体差」のばらつきをあらわす
- 超事前分布: 事前分布の平均と分散を確率分布としてあらわす

樹種差: あたかも「個体差」のようにあつかう

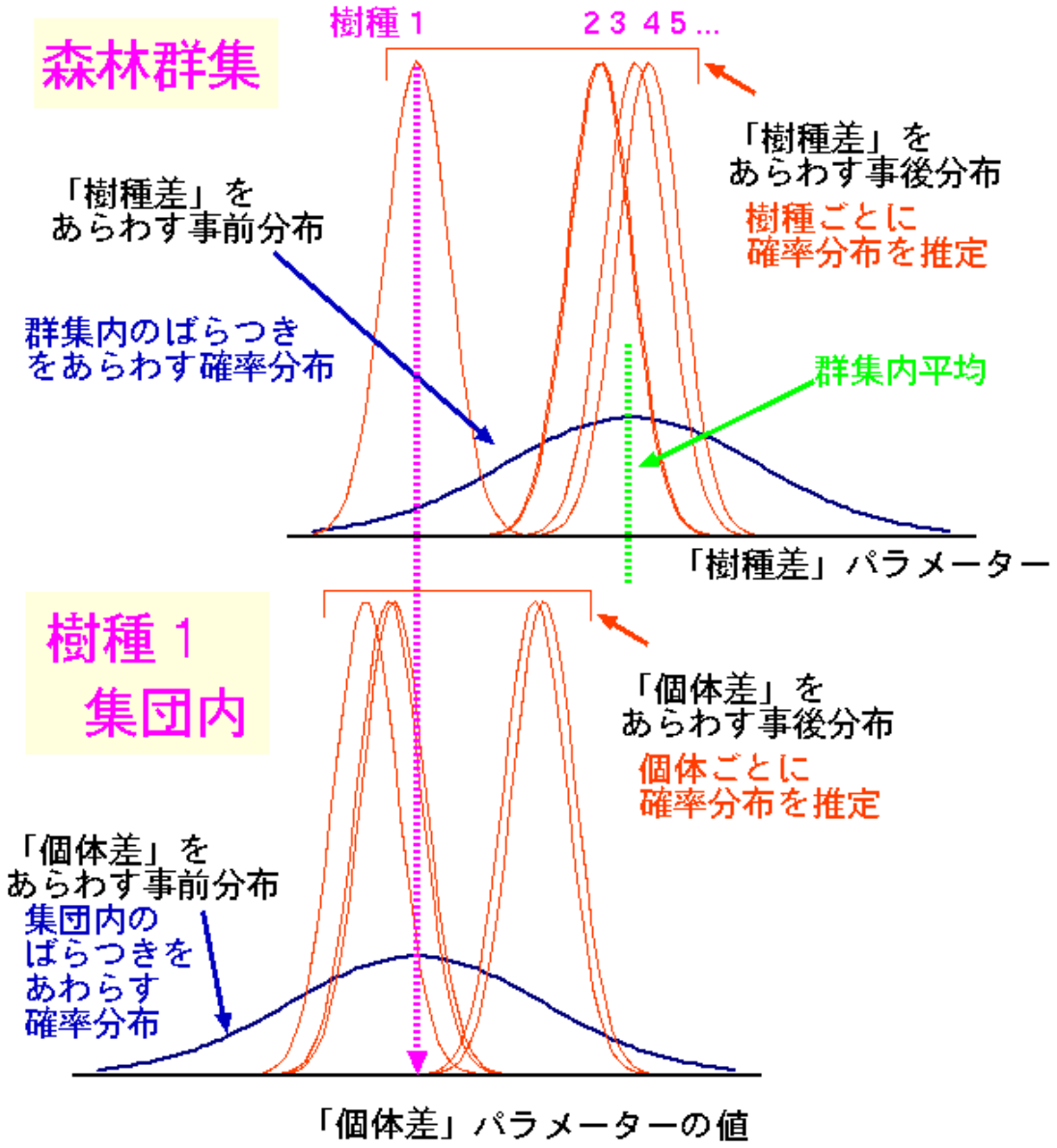


階層ベイズモデル

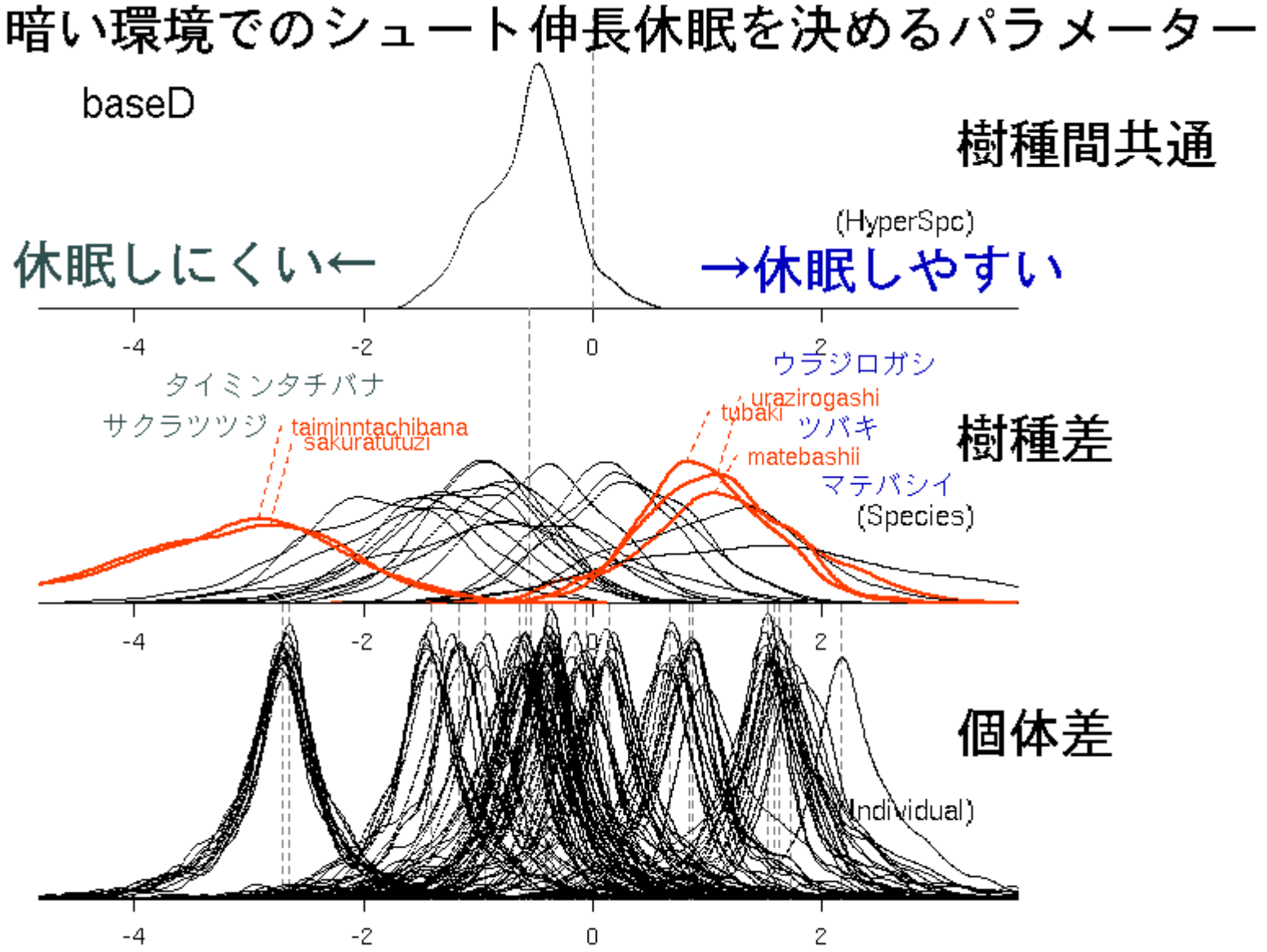
- 事後分布: 樹種との樹種差をあらわす

  → nest して「個体差」の平均
- 事前分布: 森林群集の樹種差のばらつきをあらわす
- 超事前分布: 事前分布の平均と分散を確率分布としてあらわす

「個体差」を樹種差を nest する



推定結果を使ってモデルを説明してみる



屋久島シュート伸長モデルをデータに対応させる

3. 階層ベイズモデルを MCMC 計算であつかう

「種差」「個体差」そして環境の影響を同時に推定

昨年度の
修士論文発表
(牛原阿海さん)

エンドユーザーからみた統計学ツール「含有関係」

(一般化) 線形モデル的に現象を表現する場合

[尤度をあつかうモデル]

「すべてのパラメーターは確率分布」とする Bayes 統計学

階層 Bayes モデルなどなど

[最尤推定法 であつかうモデル]

パラメーター (fixed + random effects) は特定の値

経験 Bayes 法や一般化線形混合モデル (GLMM) などなど

[一般化線形モデル (GLM)]

指数関数族の確率分布 + 線形モデル, fixed effects のみ

[最小二乗法 であつかうモデル]

等分散正規分布 + 線形モデル

直線回帰，いわゆる「分散分析」などなど

今日でてくるベイズ用語の整理

(事後分布) $\propto$ (尤度) $\times$ (事前分布) $\times$ (超事前分布)

- **階層ベイズモデル** $p(\beta, \alpha|y) \propto p(y|\beta)p(\beta|\alpha)p(\alpha)$
  - 推定計算方法: Markov Chain Monte Carlo (MCMC) 法
    - MCMC 計算わざ 1: Metropolis-Hastings 法
    - MCMC 計算わざ 2: Gibbs sampler

(上のふたつは本質的には同じもの)

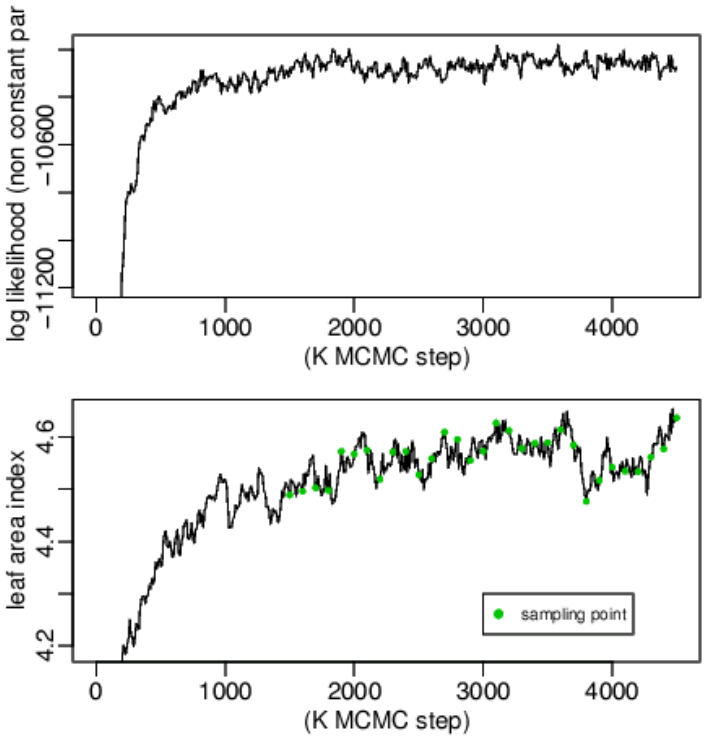
- **経験ベイズ法** $p(\beta, \alpha|y) \propto \int p(y|\beta)p(\beta|\alpha)\mathrm{d}\beta$
  - 推定計算方法: α, β の点推定 (最尤推定)
    - 例: 一般化線形混合モデル (GLMM)
  - 簡単化した階層ベイズモデル，と考えるべきか?

(参照: 石黒ほか. 2004. 階層ベイズモデルとその周辺)

MCMC 計算が何で必要?

Markov Chain Monte Carlo 計算

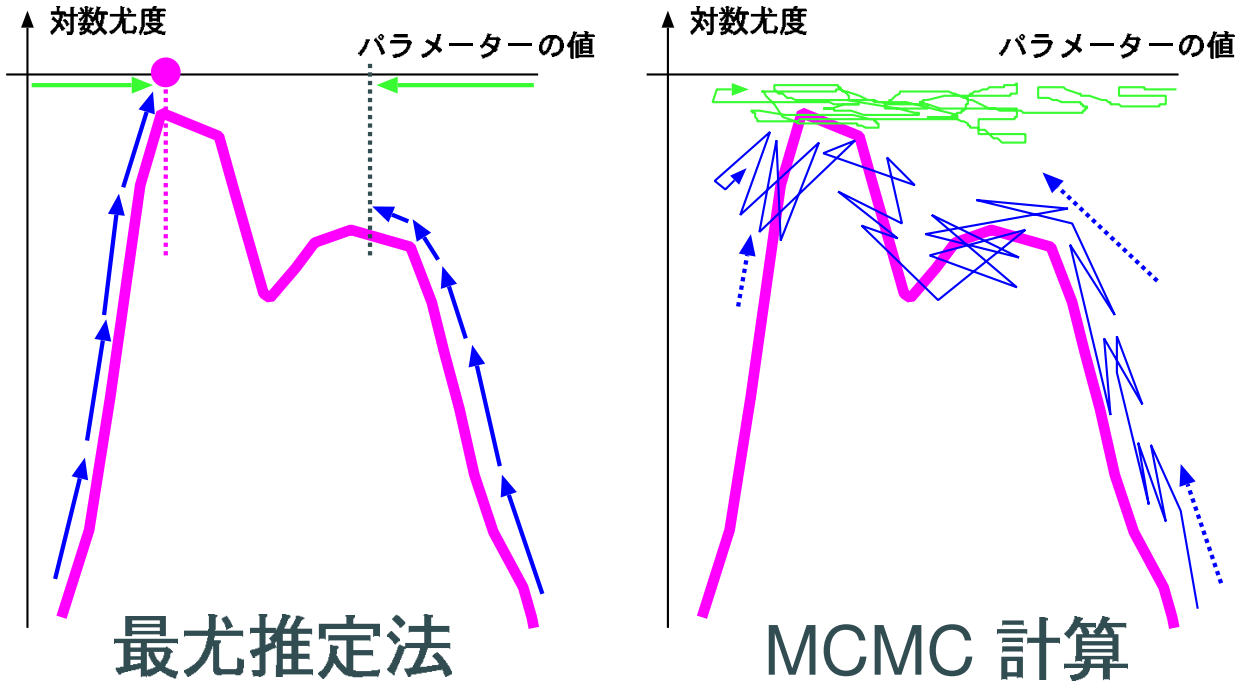
- Gibbs 分布から逐次的に標本抽出 (sampling) する方法
  - 意訳: 「あてはまりの良さそうなところ」を「さまよう」
- どんな初期値から出発しても ……
- 「定常状態」に収束していく，はず?
- 得られた sample は事後分布からの random sample set と考える



- 定常状態になるまでの step は捨てる (burn-in)

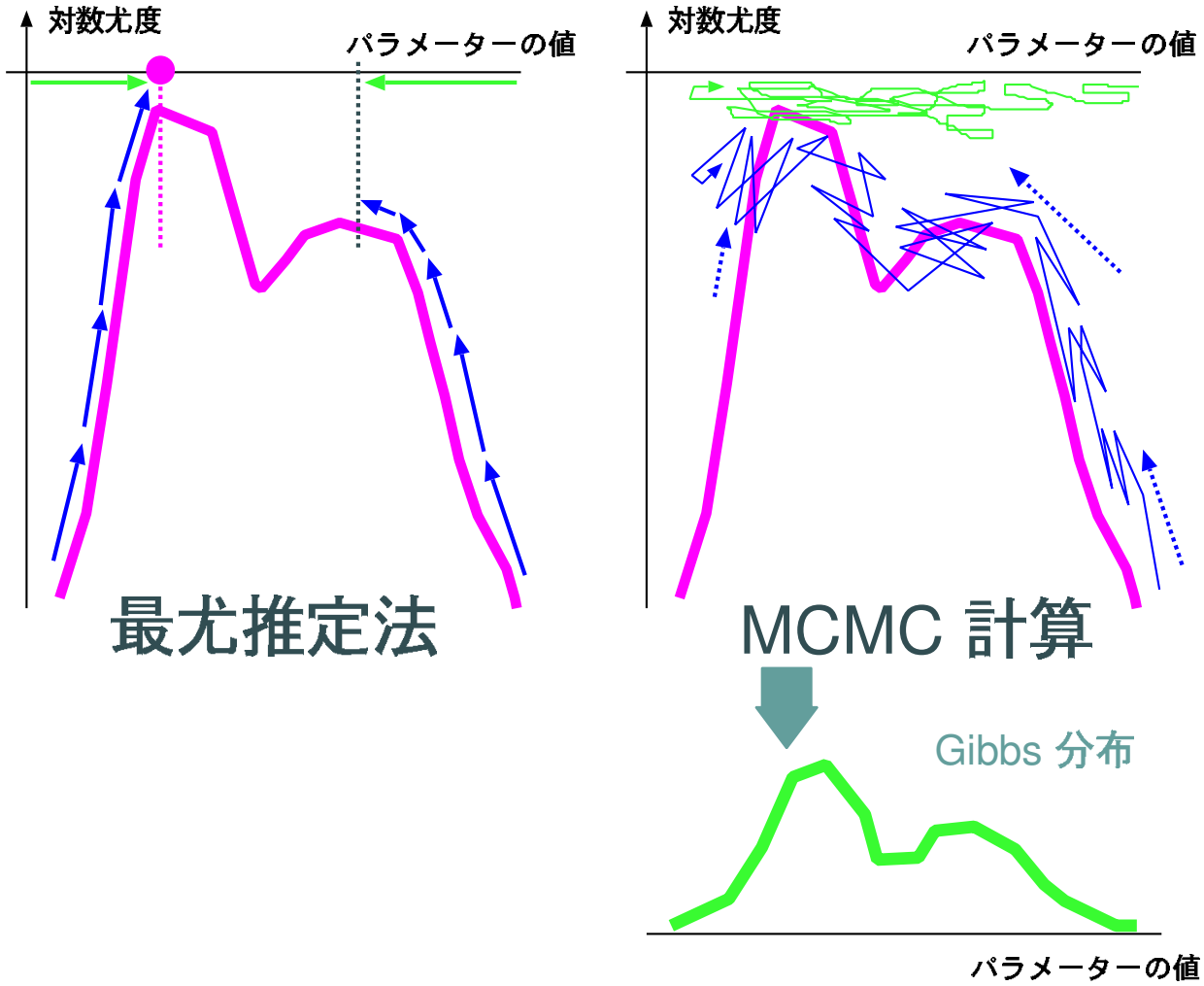
最尤推定法 vs MCMC 計算 (1) ニセ山頂回避

MCMC 計算ではときどき「悪くなる方向」にも動くので



最尤推定法 vs MCMC 計算 (2) 事後分布を得る

MCMC 計算の目的は「最適化」ではなく Gibbs 分布の推定



R まわりの MCMC 計算 / Gibbs sampler

- MCMC 計算はどういうソフトウェアで?
  - 自作する (問題によっては現実的)
  - R package: `library(MCMCpack)` など (いまいち)
  - Gibbs sampler ソフトウェア (R ではない世界)
    - WinBUGS
    - OpenBUGS
    - JAGS
  - WinBUGS と OpenBUGS の関係
    - WinBUGS , 2004 年ごろ開発停止，ソース非公開
    - OpenBUGS は WinBUGS の後継 project, GPL

WinBUGS 1.4.1 とは何か?

- おそらく世界でもっともよく使われている Gibbs sampler
- BUGS 言語の実装
- adaptive rejection sampler
- 2004-09-13 に最新版 (ここで開発停止 → OpenBUGS)
- ソースなど非公開，無料，ユーザー登録必要
- Windows バイナリーとして配布されている
  - Linux 上では WINE 上で動作
  - MacOS X 上でも Darwine など駆使すると動くらしい
- R ユーザーにとっては R2WinBUGS が快適

BUGS 言語で階層ベイズモデルを記述すると……

- Spiegelhalter et al. 1995. BUGS: Bayesian Using Gibbs Sampling version 0.50.

```
model {
    mu ~ dnorm(0, 1.0E-2)
    tau ~ dgamma(1.0E-3, 1.0E-3)
    for (i in 1:n.samples) {
        re[i] ~ dnorm(0.0, tau)
        p[i] <- 1.0 / (1.0 + exp(-(mu + re[i])))
        n.seeds[i] ~ dbin(p[i], n.ovules[i])
    }
}
```

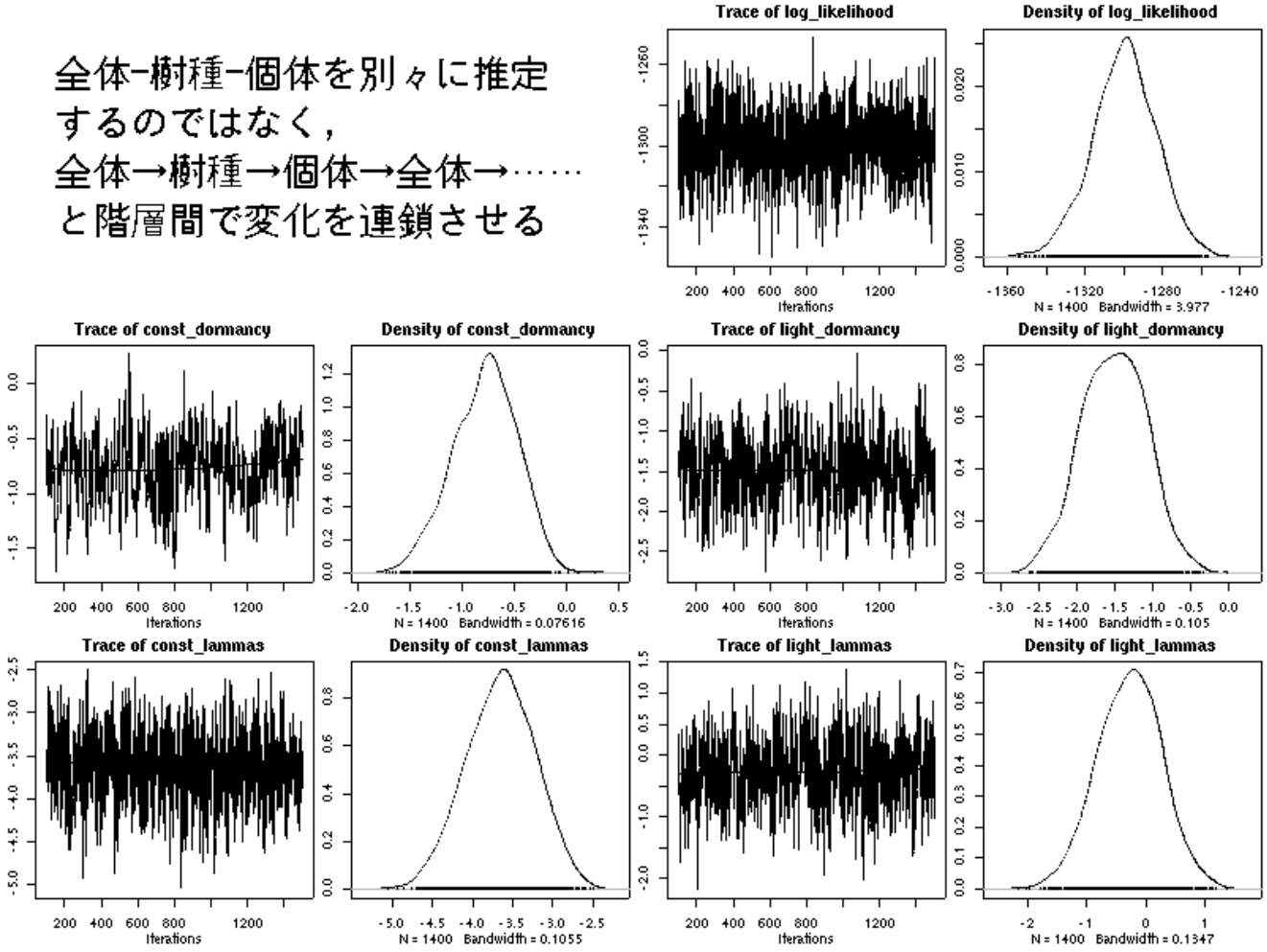
- JAGS だと行末に ; が必要，といった方言がある

推定計算: Markov Chain Monte Carlo (MCMC) 法

逐次的に Gibbs 分布からサンプリングする方法

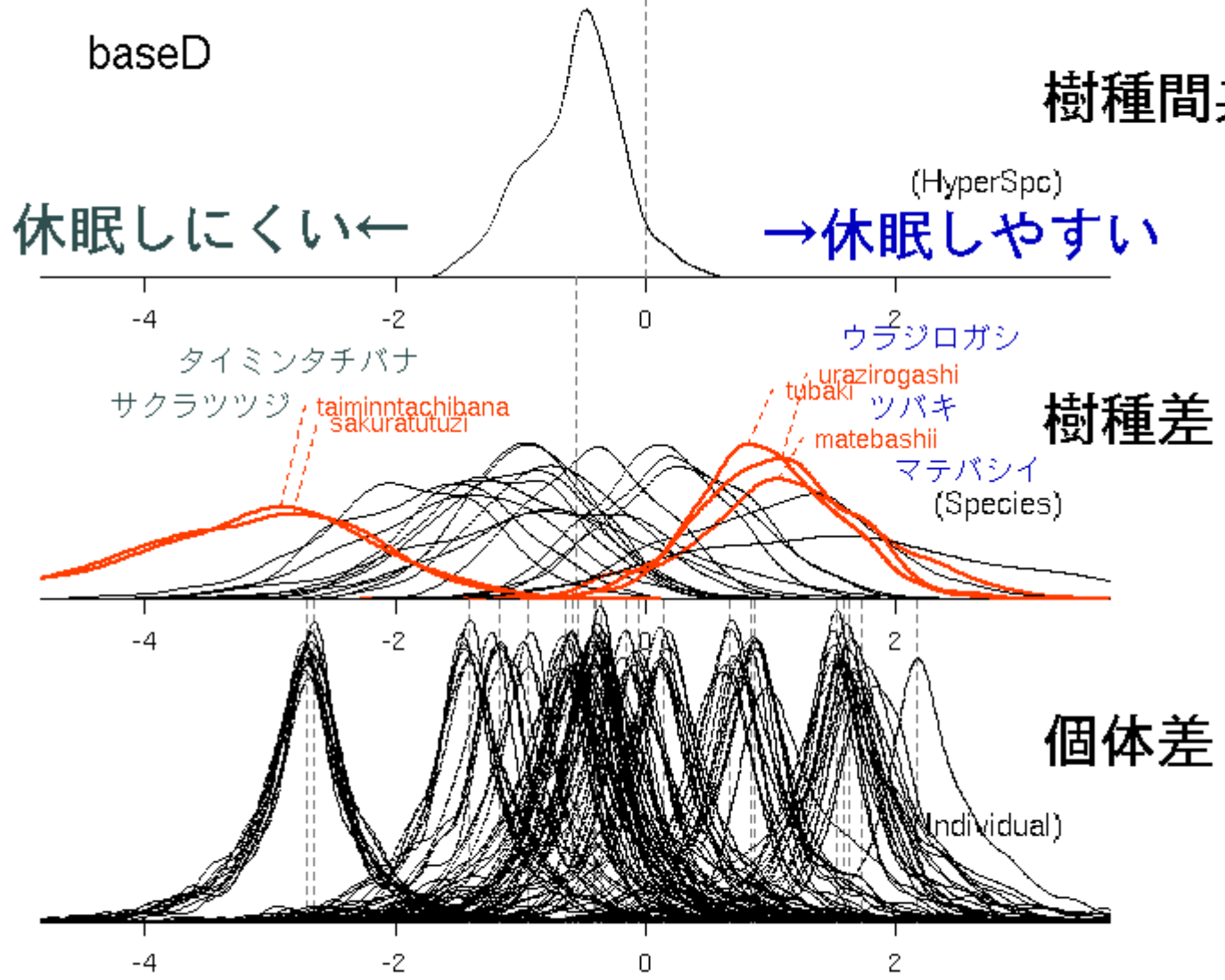
(とりあえず図示)

全体–樹種–個体を別々に推定するのではなく，
全体→樹種→個体→全体→……と階層間で変化を連鎖させる



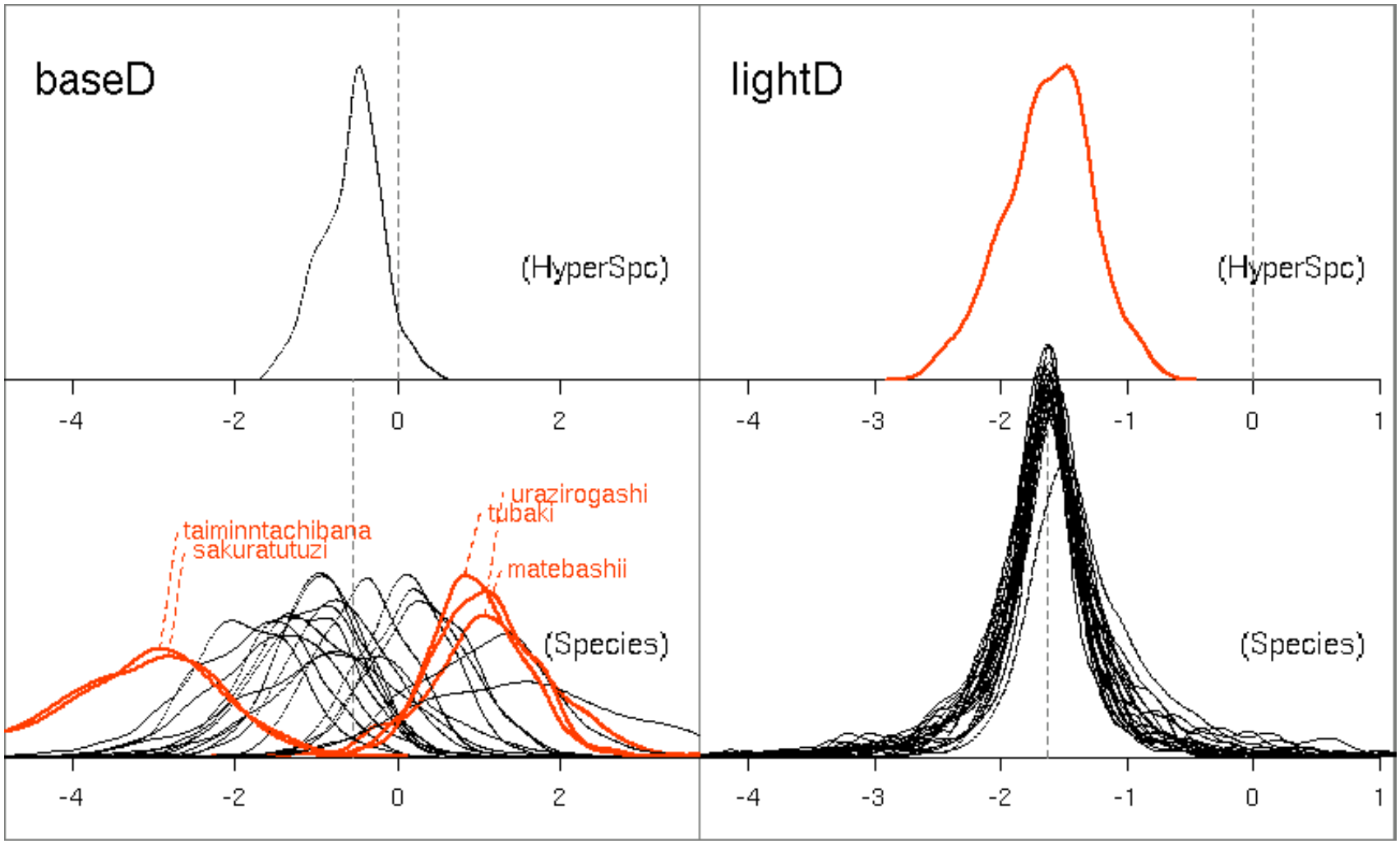
nest したパラメーター事後分布

暗い環境でのシュート伸長休眠を決めるパラメーター



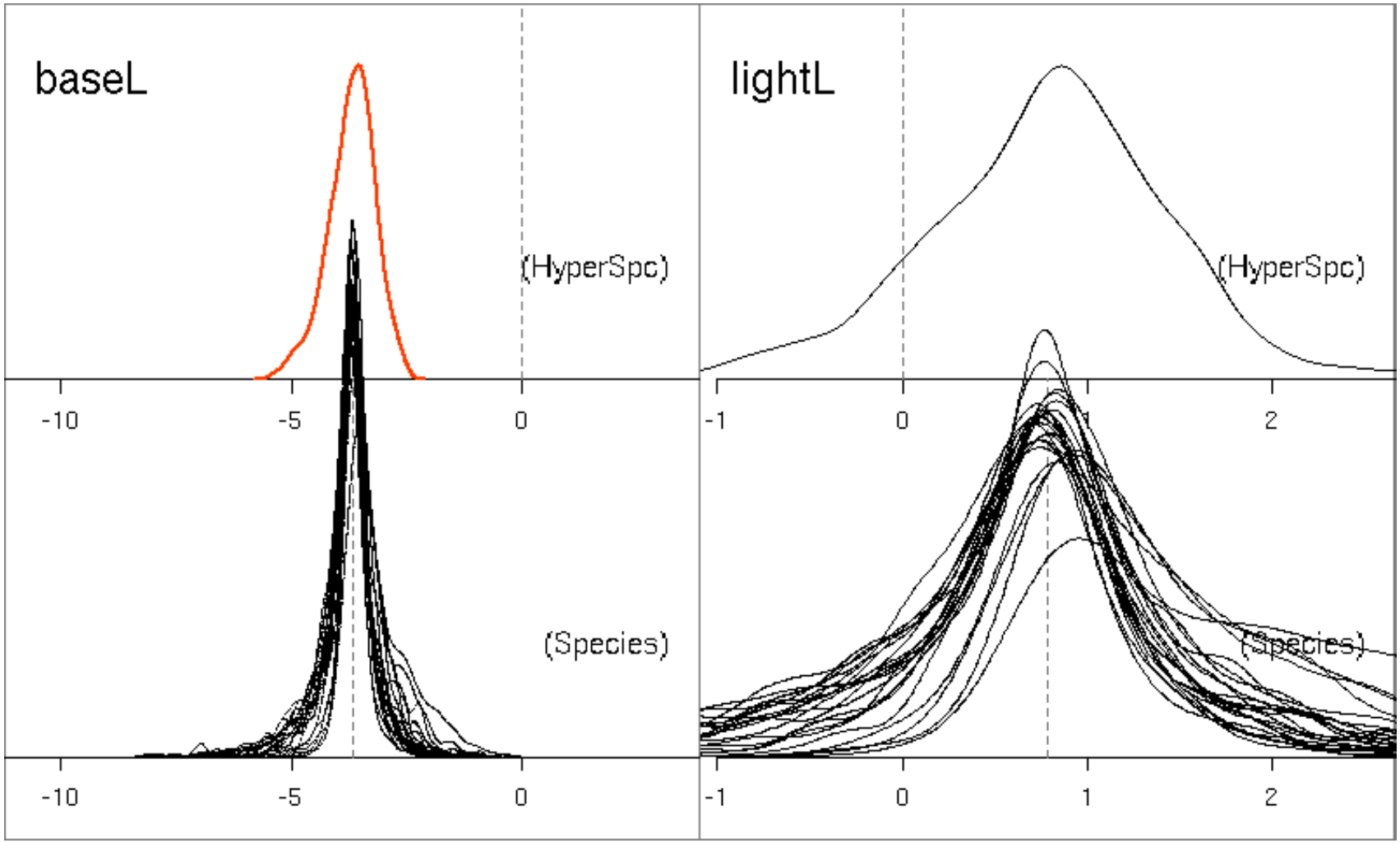
推定結果: nest したパラメーター事後分布

シュート伸長休眠: $\beta_{CD} = \beta_{CD}^{Hyperspecies} + \beta_{CD}^{Species} + \beta_{CD}^{Individual}$



nest したパラメーター事後分布: シュート二度伸び

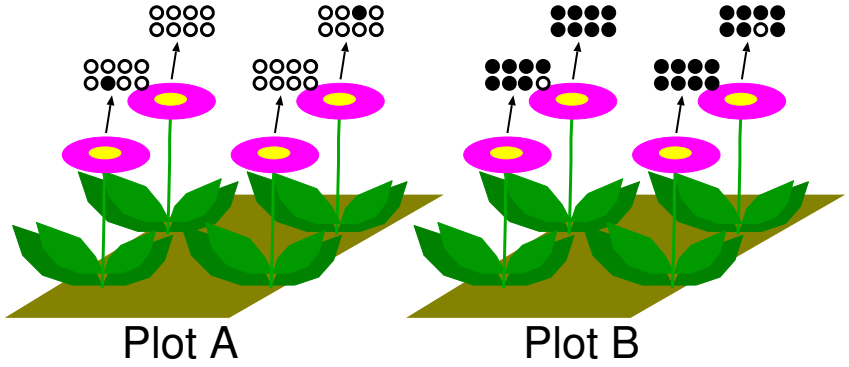
休眠確率は 0.31 (close), 0.09 (open), 二度伸び確率は 0.02



階層ベイズモデルのご利益とは?

階層ベイズモデルでないとうまく表現できない現象がある

- 複数の random effects (個体差・ブロック差・縦断的データ・……)
- 多重 nest した random effects の導入



- 「隠れた」状態をあつかうモデル
  - 例: 「欠側値を補う」処理
- 空間構造ある問題も MCMC 計算で
  - 例: 「隣は似てるよ」効果 – Gaussian Random Field

今日のまとめ: 「個体差」を考慮した解析が必要

1. `glm()` がうまくいかない状況?

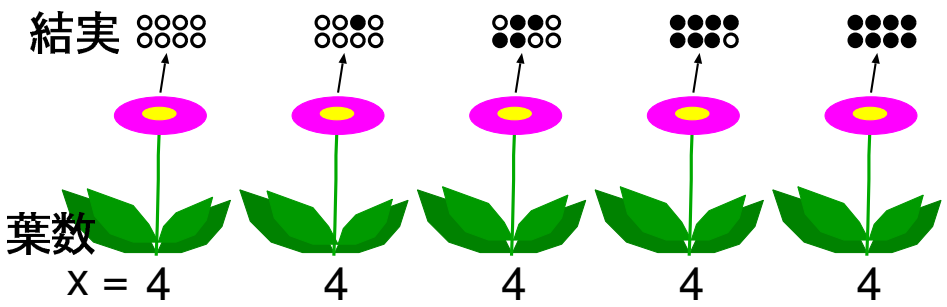
   原因: 「個体差」による過分散 (overdispersion)

2. 屋久島照葉樹 22 樹種のデータ解析

   たくさんの「種差」のモデリング

3. 階層ベイズモデルを MCMC 計算であつかう

   「種差」「個体差」そして環境の影響を同時に推定



「観測できない個体差」などは random effects としてあつかい，調べたい量 (葉数の効果など) に集中するのがうまい統計モデリング.「データとりまくれば解決する」と思いこむのはまちがい.