多様性生物学基礎論 (生物多様性論 I): 生物多様性の現在

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」を考慮した統計解析・統計モデリング

一般化線形混合モデル (GLMM) 入門

http://hosho.ees.hokudai.ac.jp/~kubo/stat/2005/

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**: 2006.01.23 (月)
  「数えられる」データの統計解析・統計モデリング

  一般化線形モデル (GLM) 入門
- **第 2 回**: 2006.01.25 (水)
  「個体差」を考慮した統計解析・統計モデリング

  一般化線形混合モデル (GLMM) 入門

http://www.r-project.org/

今日のハナシ: データ中の「個体差」にいどむ

1. **疑惑篇: `glm()` がうまくいかない状況**

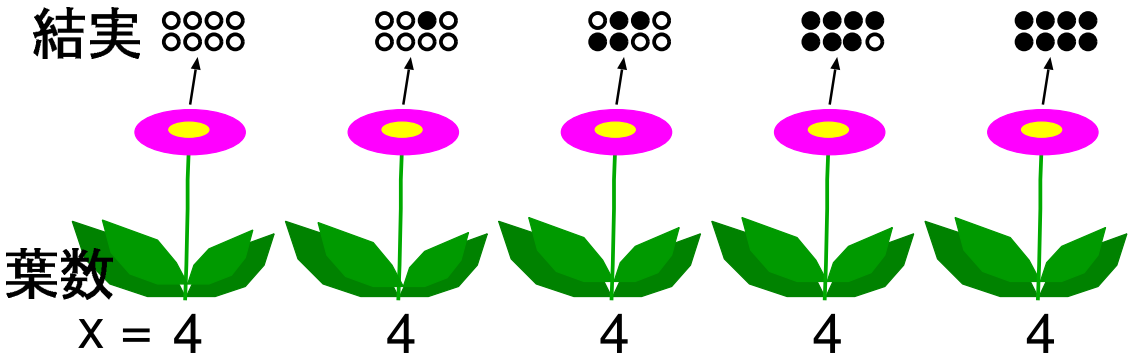
   データを何回とりなおしてもダメ?

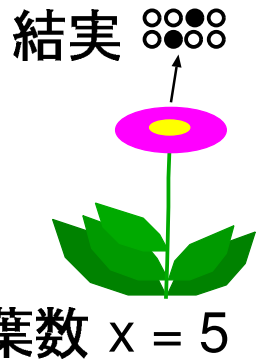
2. **究明篇:「個体差」見つけかた作りかた**

   過分散 (overdispersion) を表現するモデル

3. **解決篇: 一般化線形混合モデル**

   Random effects のばらつきも同時に推定





すごく単純化した状況なのに

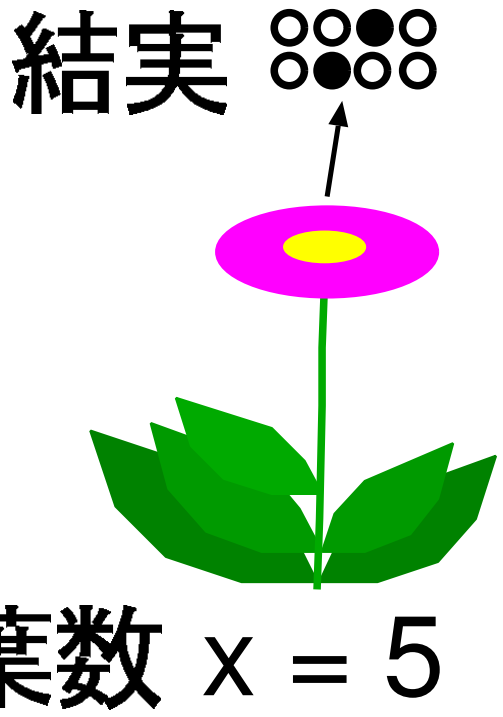
1. 疑惑篇: glm() がうまくいかない状況

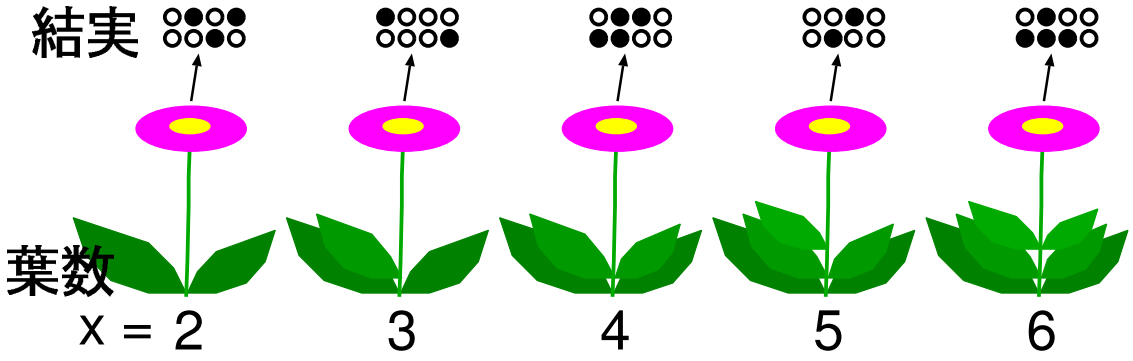
データを何回とりなおしてもダメ?

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)

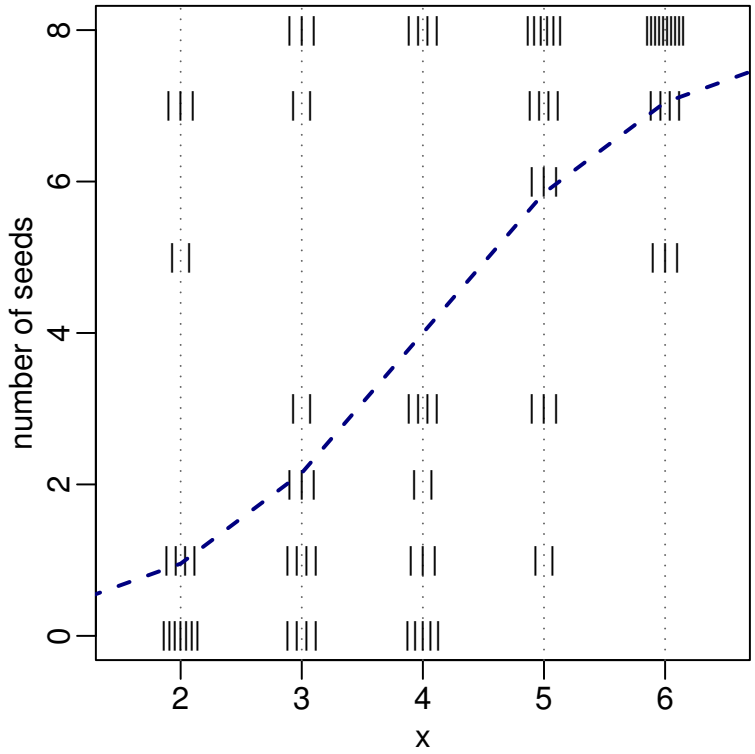


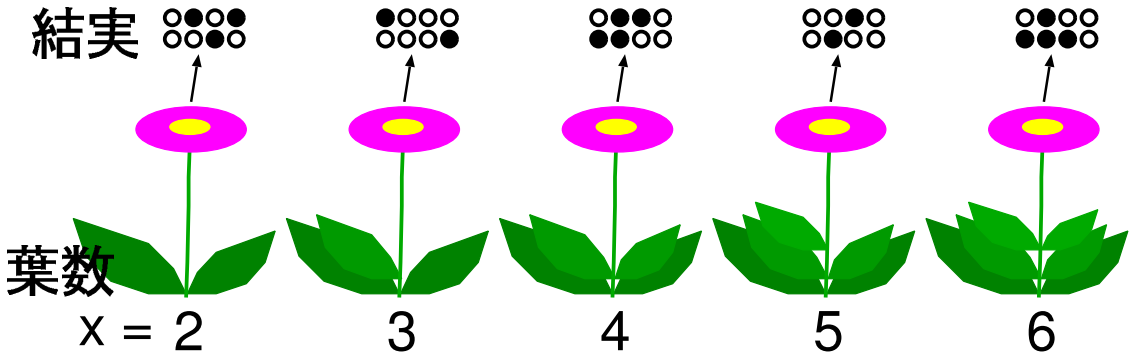


- 葉数 x は個体ごとに 2-6 枚
- 葉数が結実率を決めるらしい

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

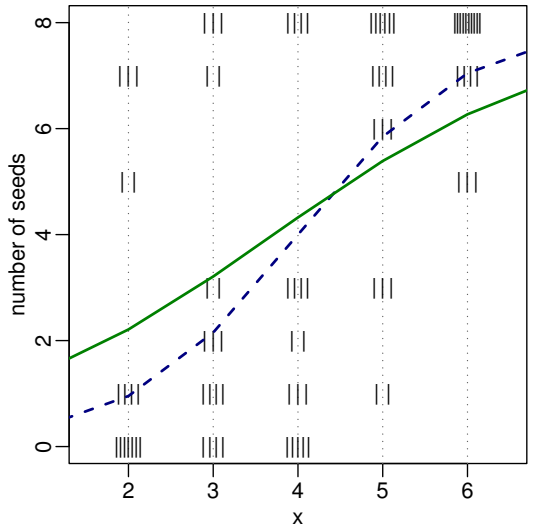
(あとで解説する)

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)...
```

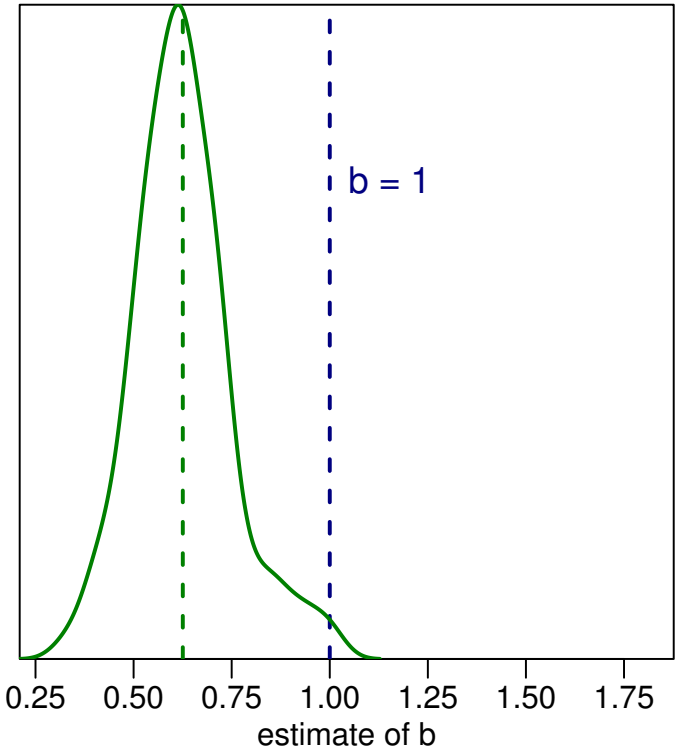
なンか葉数パラメーターの推定値がずれてるんですけど…… $\hat{b} = 0.5627$

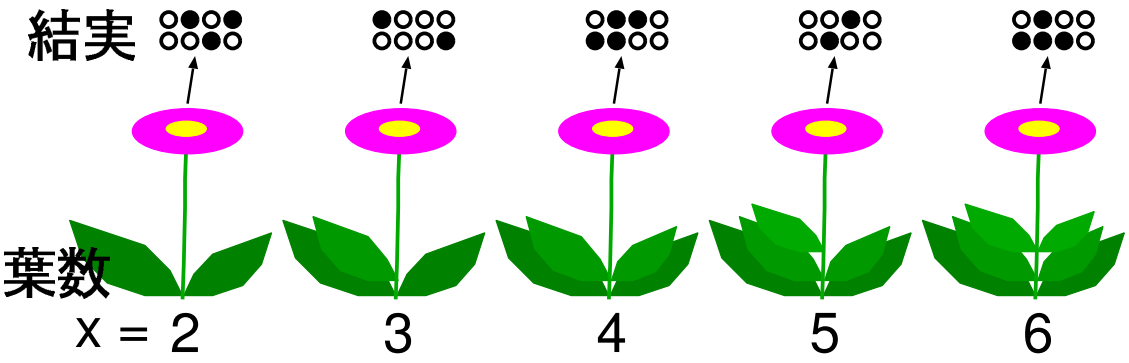
(真の値は $b = 1$)



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?

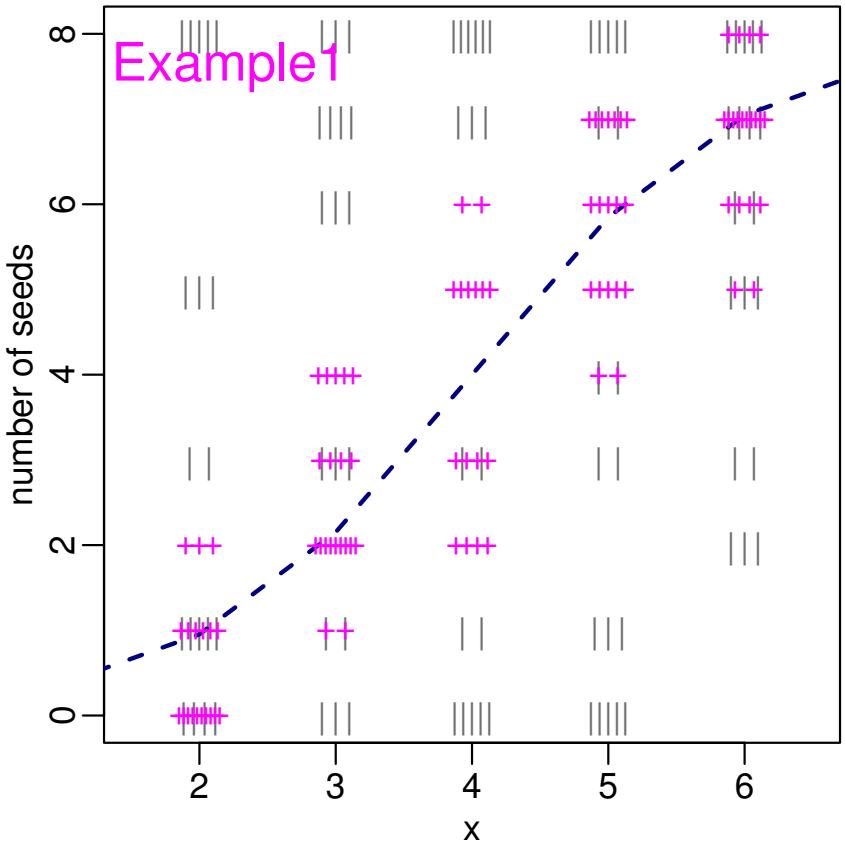
- 気をとりなおして原点から考えなおす
- そもそも logistic 回帰ではデータが**二項分布** (binomial distribution) になることが前提
- 8 個の種子のうち y 個が結実する確率は

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

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

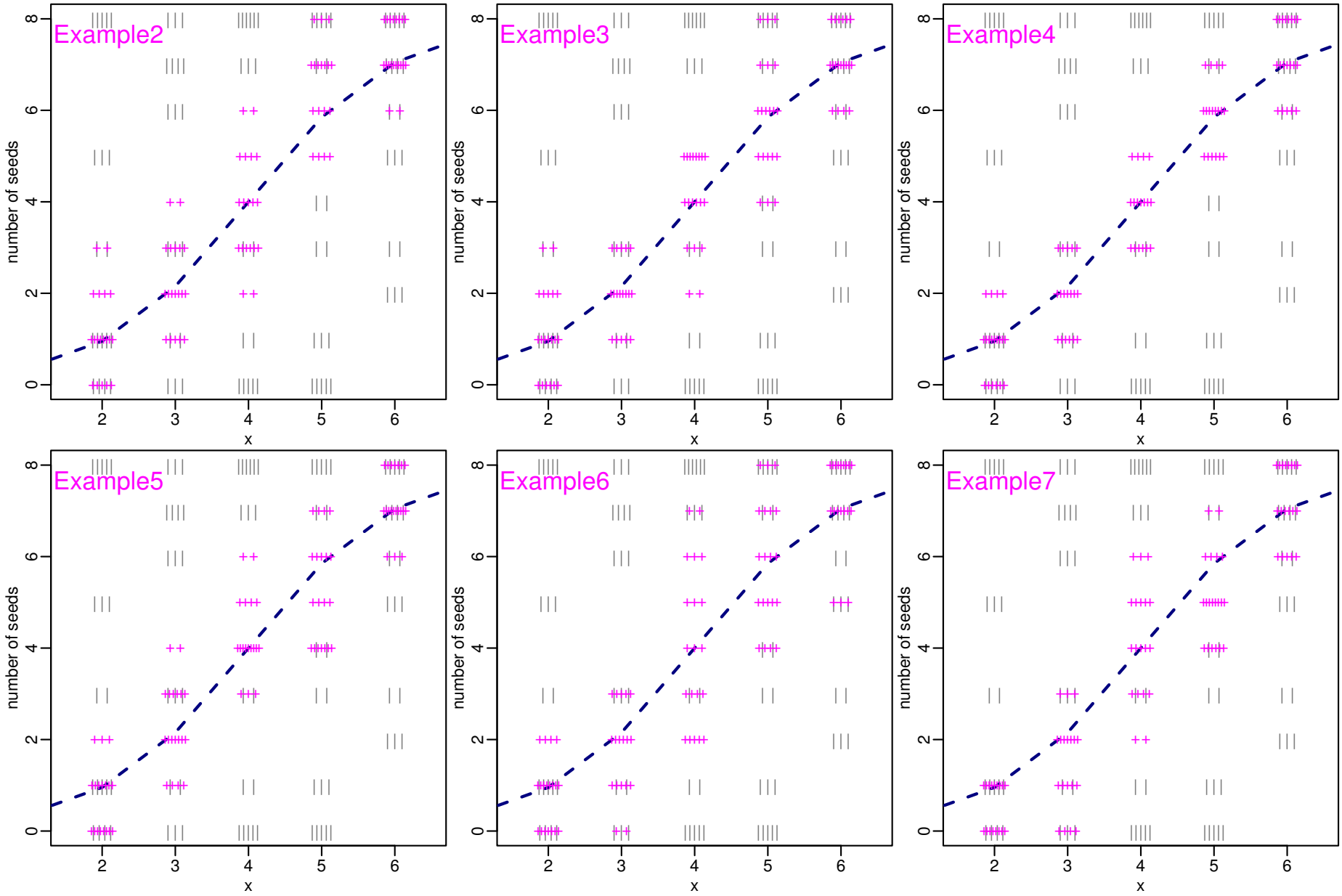
- **[R による実験で調べる]** 結実率を ……

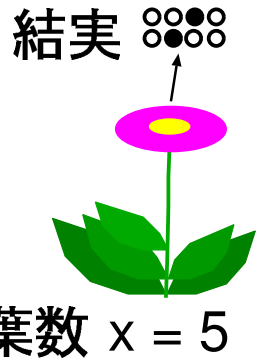
$$p = \frac{1}{1+\exp(-(-4+1\times x)}$$



- ……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……?

そもそもこの観測データ，ばらつきすぎでは? (しつこく)





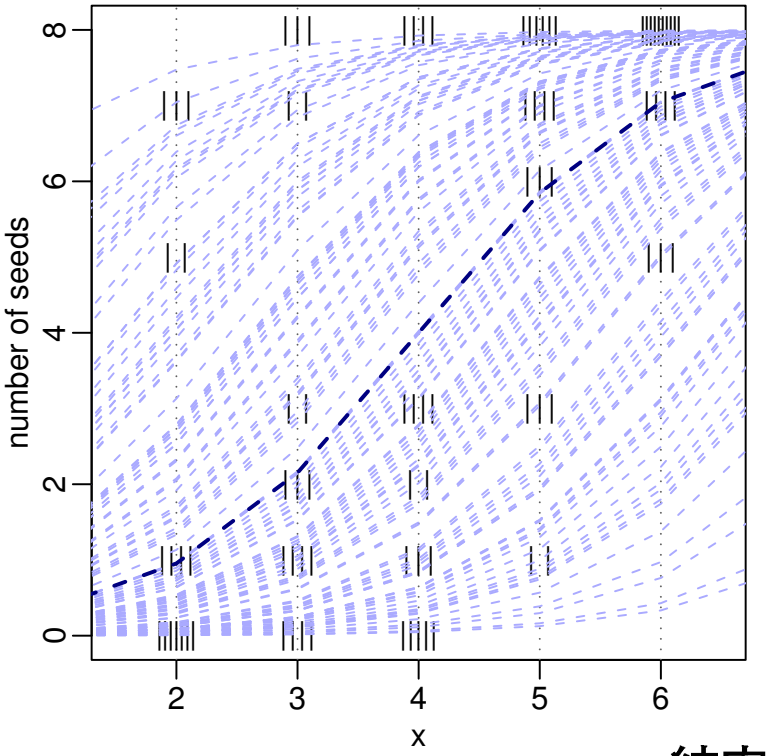
人間が測定してない・できないばらつき

2. 究明篇:「個体差」見つけかた作りかた

過分散 (overdispersion) を表現するモデル

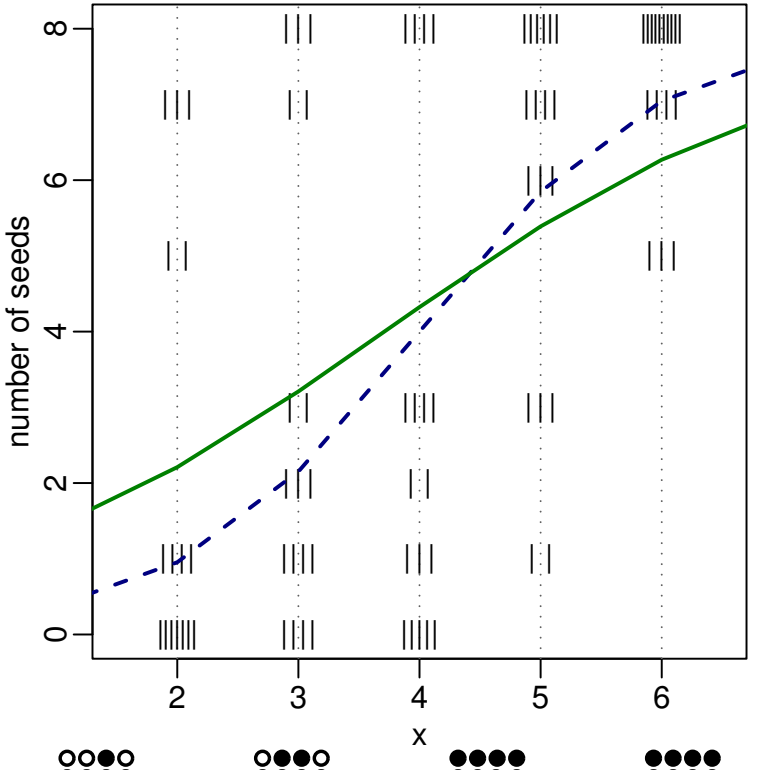
個体ごとの「ずれ」が b の推定を偏らせた

「傾き」 (b) には「個体差」ない
「切片」 (a) には「個体差」ある



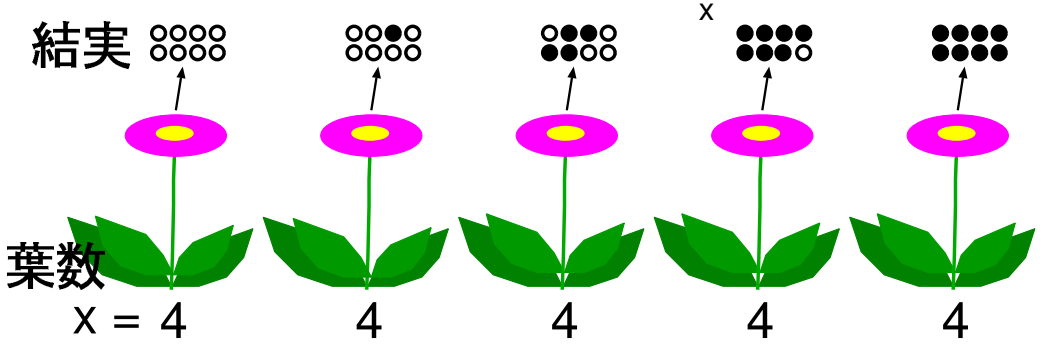
→

「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……?

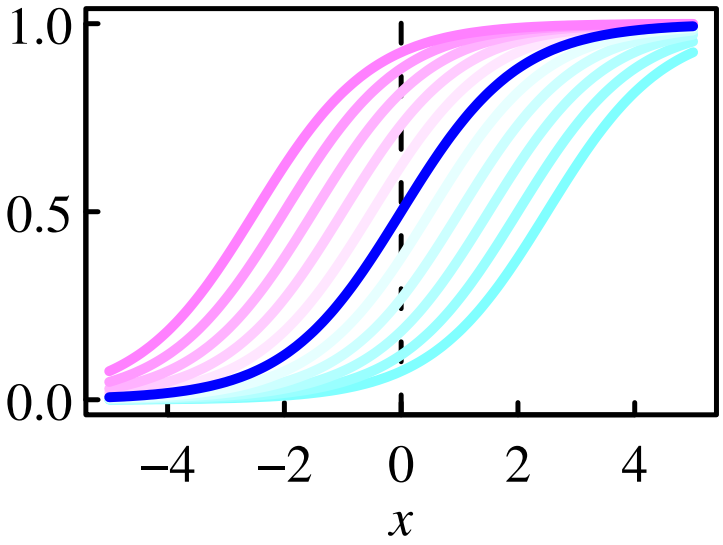


(前回の復習)「ロジスティック関数」って何?

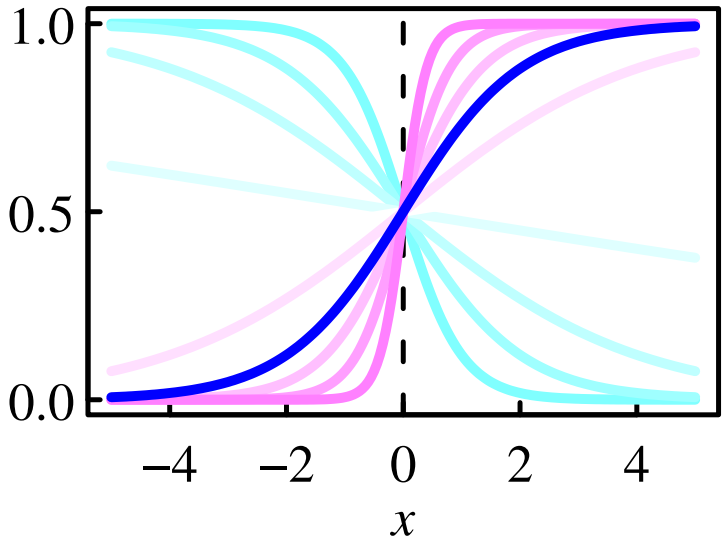
$$p = \frac{1}{1 + \exp(-(a + bx))}$$

($\exp(Z) = \mathrm{e}^Z$ のこと)

a だけ変化させる



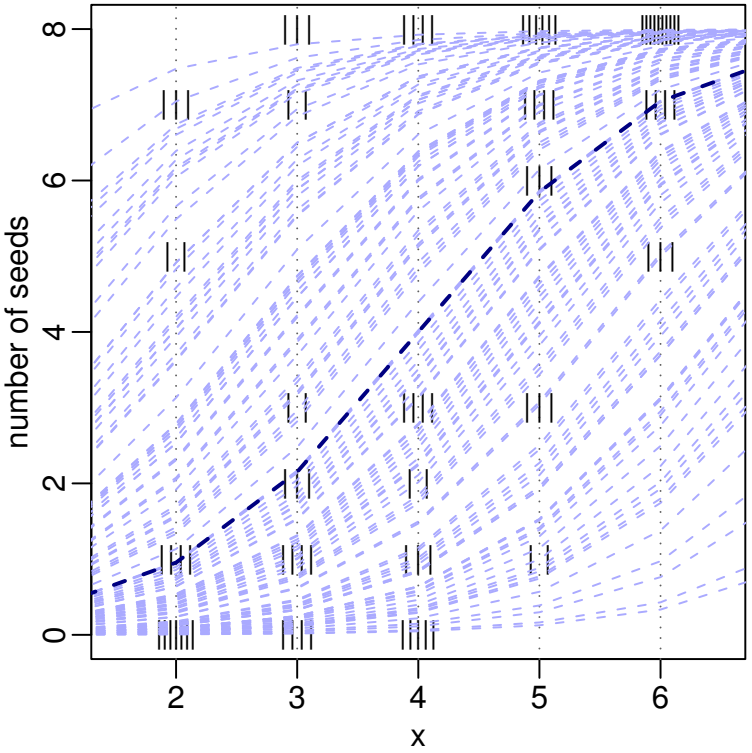
b だけ変化させる

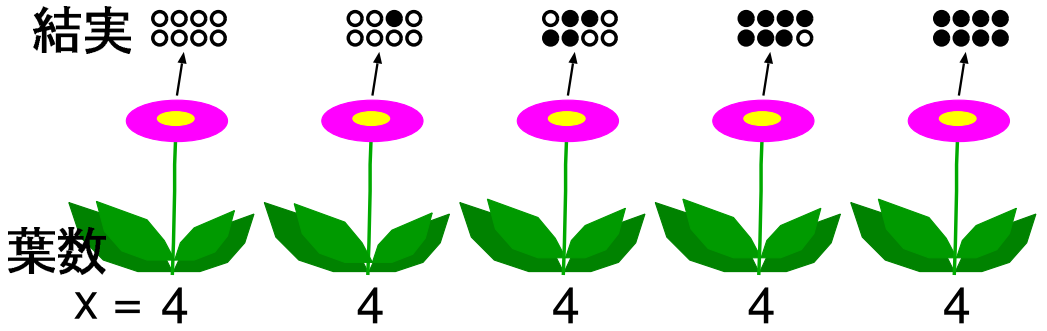


つまりパラメーター $\{a, b\}$ や説明変数 x がどんな値をとっても確率 p は $0 \leq p \leq 1$ となる便利な関数

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!)

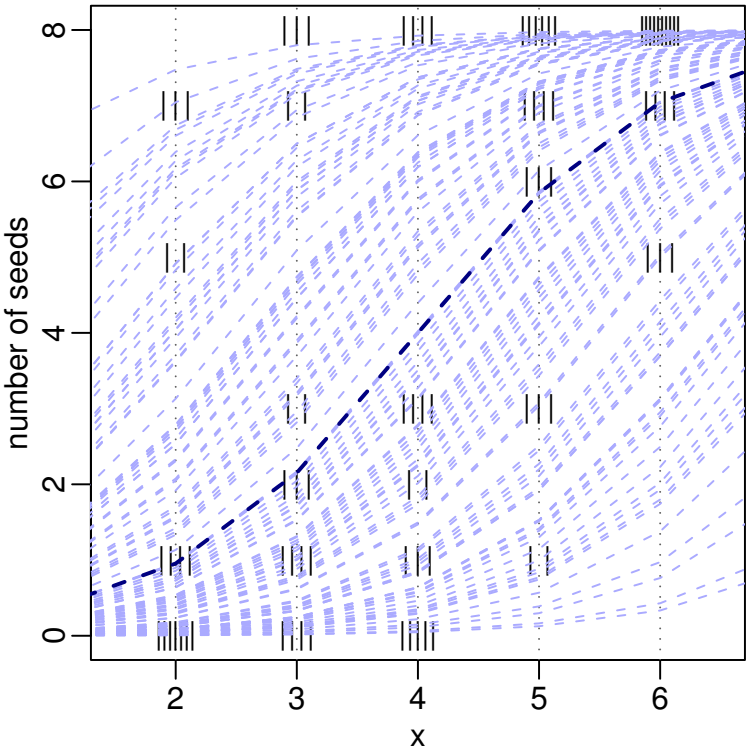
「個体差」とは何か? (統計モデリング的側面)

- 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 $\longrightarrow$ “fixed effects”



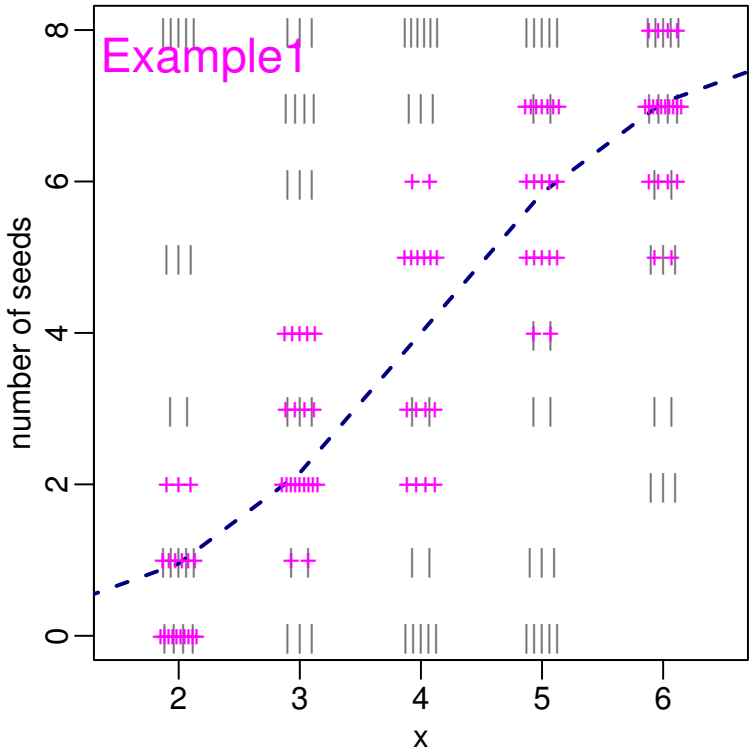
- $?$ は p の平均値を変化させず，ばらつきだけを変えている
  $\longrightarrow$ “random effects”
- fixed effects と random effects を両方ふくむ統計モデルを混合モデル (mixed model) とよぶ

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

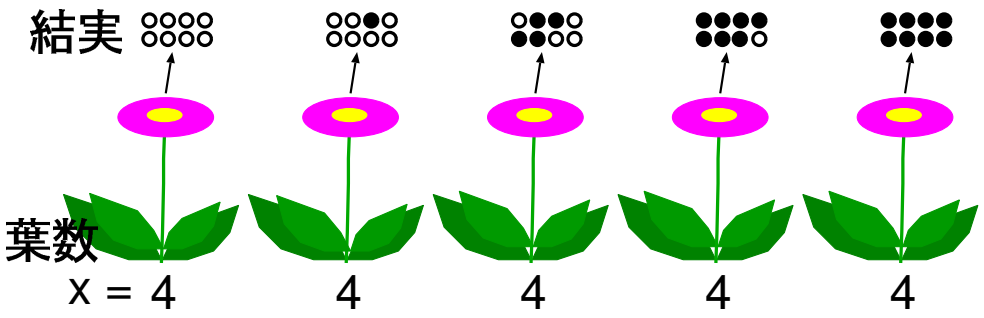
- Random effects なしの結実率モデル

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

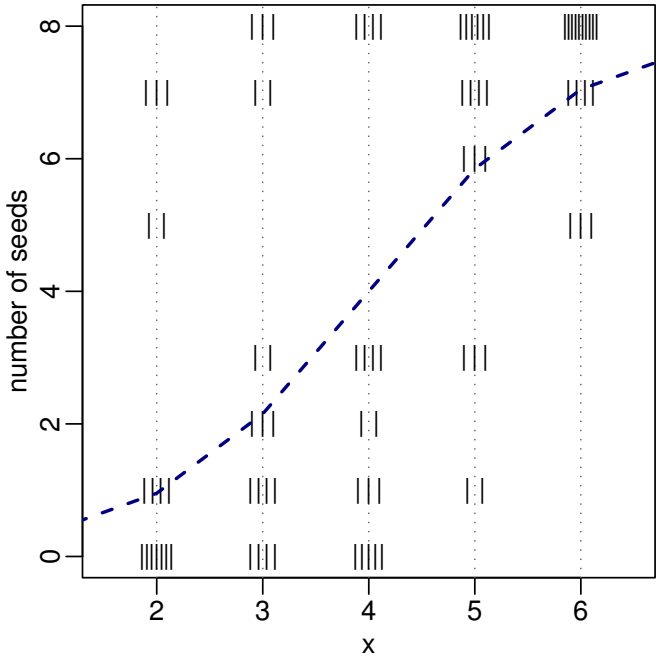
を仮定して二項乱数を生成させると + のようなデータが得られる



- 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



Random effects のある種子結実シミュレイション



- 模倣することは考えること

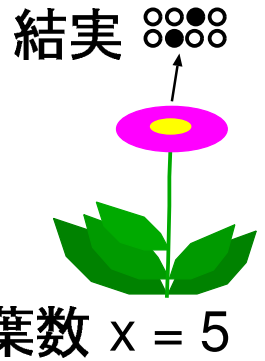
  作ることは理解すること
- 結実率を表現する logistic 曲線

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

- Random effects である ? の部分は平均ゼロの正規乱数を植物個体ごとに与えてやる

```
> re <- rnorm(100, 0, 3) # 平均ゼロで標準偏差 3 の正規乱数 --- 「個体差」
> x <- c(sapply(2:6, function(n) rep(n, 20))) # 葉数 x
> rbinom(100, 8, 1 / (1 + exp(-(-4 + x + re)))) # 「個体差あり」二項乱数

 [1] 6 2 0 0 0 1 8 0 0 1 8 4 4 1 1 4 6 8 1 5 0 0 3 1 0 1 8 8 5 7 6 2 1 7 0 5 6 0 7 8 0
[43] 1 0 2 0 3 3 6 0 0 3 6 0 3 8 2 6 1 8 5 8 7 0 7 0 8 7 3 5 1 8 1 6 0 8 4 3 8 8 1 8 8
[85] 6 6 8 6 1 8 8 7 5 6 8 4 0 5 3 8
```

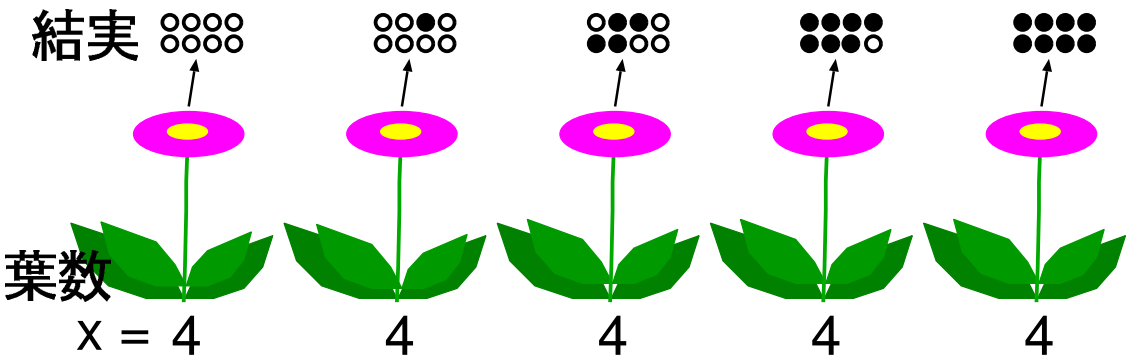


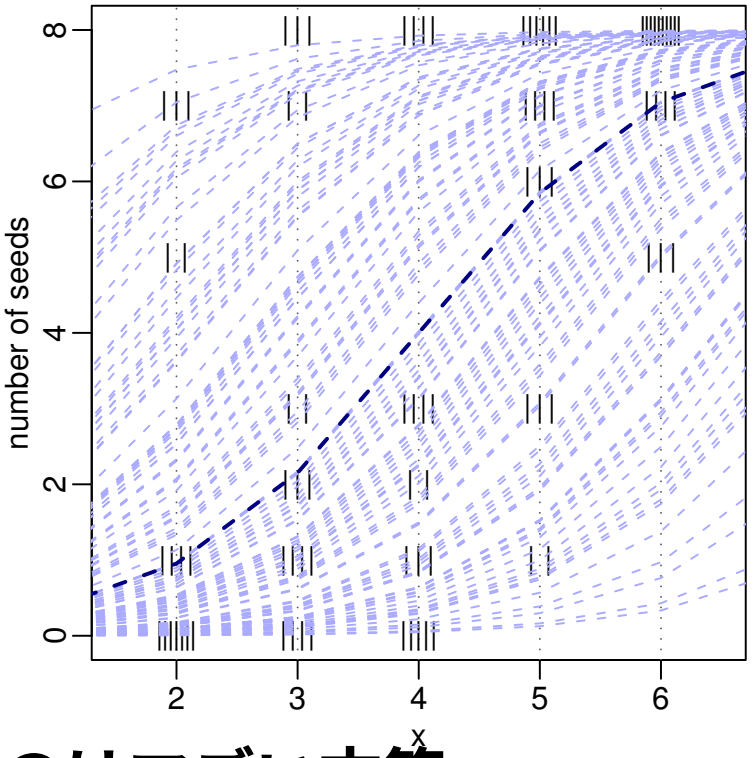
「個体差」くみこむ GLMM

3. 解決篇: 一般化線形混合モデル

Random effects のばらつきも同時に推定

「個体差」 こと random effects をどうあつかうか





- 結実率を表現する logistic 曲線

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

- 個体ごとに「?」の値を直接推定するのはマズい方策
  - パラメーター数むちゃくちゃ増える，自由度が減る
  - 標本個体数増やしても「?」の推定の「良さ」が改善されない

そこで標本個体数を増やしても「個体差」である「?」まわりが面倒にならぬ方法を考える

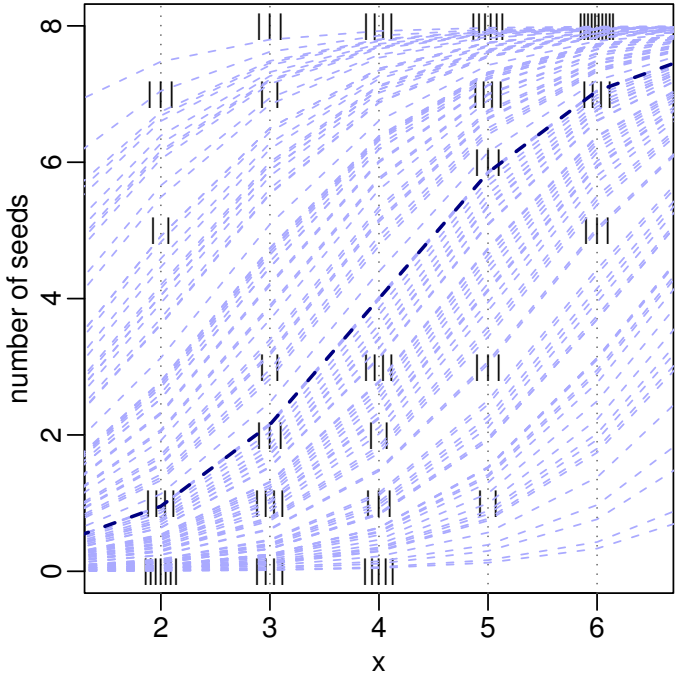
「個体差」 random effects を表現する統計モデル

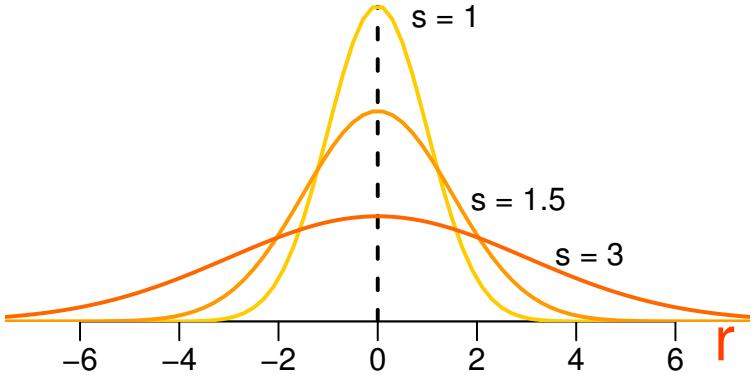
- 結実率を表現する logistic 曲線

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

- random effects をあらわす確率変数を r とする

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





- random effects r は平均ゼロの正規分布 g にしたがうとする

$$g(r, s) = \frac{1}{\sqrt{2\pi s^2}} \exp(-\frac{r^2}{2s^2})$$

- 標準偏差 s は「個体差」のばらつきの大きさ

混合 logistic モデル = fixed + random effects

- ある個体 i の葉数は x_i,「個体差」は r_i とする
- fixed effects $(a + bx)$ と random effects (r) からなる結実率の混合 logistic モデル

$$p_i(a, b, r) = \frac{1}{1 + \exp(-(a + bx_i + r_i))}$$

- ある個体 i で y_i 個の種子が得られる確率は

$$f(y_i \mid a, b, r_i) = \binom{8}{y_i} p_i(a, b, r_i)^{y_i} (1 - p_i(a, b, r_i))^{8-y_i}$$

- 「個体差」こと random effects が r_i である確率 (密度) は

$$g(r_i, s) = \frac{1}{\sqrt{2\pi s^2}} \exp(-\frac{{r_i}^2}{2s^2})$$

混合 logistic モデル の尤度方程式

- ある個体 i で y_i 個の種子が得られて，同時に「個体差」こと random effects が r_i である確率 (密度) は

$$f(y_i \mid a, b, r_i)\ g(r_i, s)$$

- 「個体差」 r_i はわからんので全ての可能性の期待値をとる

$$\int_{-\infty}^{\infty} f(y_i \mid a, b, r_i)\ g(r_i, s) dr_i$$

(これはある個体 i の尤度である)

- 全個体の尤度は

$$L(a, b, s \mid \{y_i\}) = \prod_{i \in \{\text{全個体}\}} \int_{-\infty}^{\infty} f(y_i \mid a, b, r_i)\ g(r_i, s) dr_i$$

この尤度を最大化する $\{a, b, s\}$ を探しあてるのが最尤推定

R で混合モデルの推定計算はどうやればいいのか?

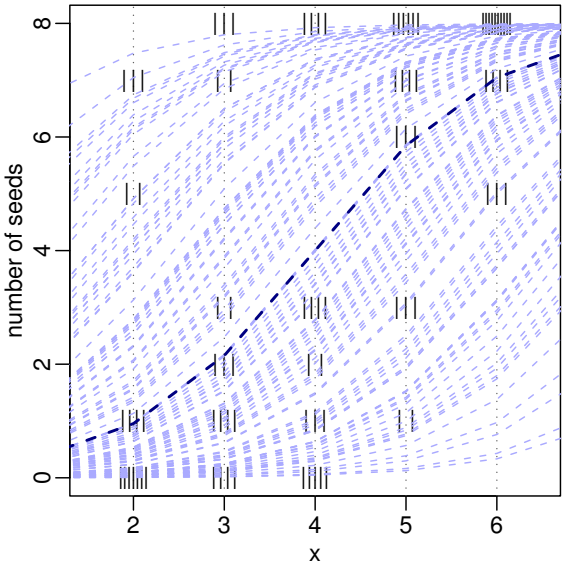
一般化線形混合モデル (GLMM) とは ……

ロジスティック回帰・ポアソン回帰など一般化線形モデル (GLM) に random effects の項を加えて混合モデル化したものの総称

GLMM: generalized linear mixed model

推定計算はとりあえず `glmmML` package の `glmmML()` 関数で

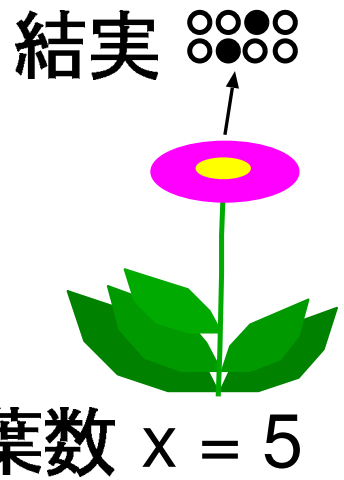
- 混合モデルの **尤度を数値積分** で計算する
- **「定数項」** のみが random effects になりうる (つまり「横方向のずれ」だけ)
- `family = binomial or poisson` だけ
- Göran Broström さんが開発・発展させている



混合 logistic モデルの最尤推定，glmmML() で (1)

まずはデータの準備 (CSV 形式のテキストファイルをよみこむ)

```
> d <- read.csv("d.csv")
> head(d, 10) # d の先頭をみる
    id x n.seed ...
1  f001 2      0 # ここから個体 f001
2  f001 2      0
3  f001 2      0
4  f001 2      0
5  f001 2      0
6  f001 2      1
7  f001 2      0
8  f001 2      0
9  f002 2      1 # ここから個体 f002
10 f002 2      1
```



- n.seed は 0 (結実してない) と 1 (結実している) の値をとる
- 一個体は 8 行で表現される (胚珠数 8 だから)

混合 logistic モデルの最尤推定，glmmML() で (2)

cluster を指定する: 「個体差」を共有する範囲 (個体)

```
> library(glmmML) # glmmML package 読みこみ
> glmmML(n.seed ~ 1 + x, familiy = binomial, data = d, cluster = d$id)
```

glmmML() の指定のいくつかを簡単に説明すると

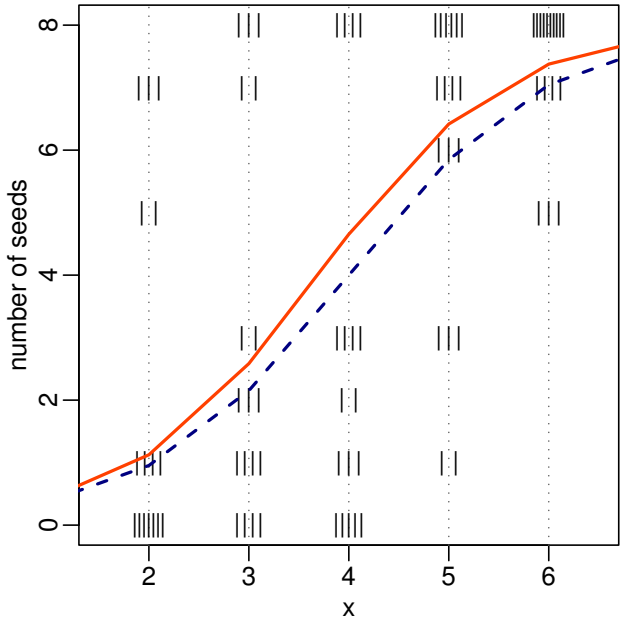
- n.seed ~ 1 + x ……応答変数 n.seed を
  $a + bx$ で説明しろ (fixed effects)
- familiy = binomial ……ばらつきを説明する確率分布は
  二項分布で
- data = d ……データは d オブジェクト
- cluster = d$id ……「個体差」こと random effects r_i は
  個体 i の id もつ 8 個の胚珠で共有される

混合 logistic モデルの最尤推定，glmmML() で (3)

glmmML() の推定結果を読む

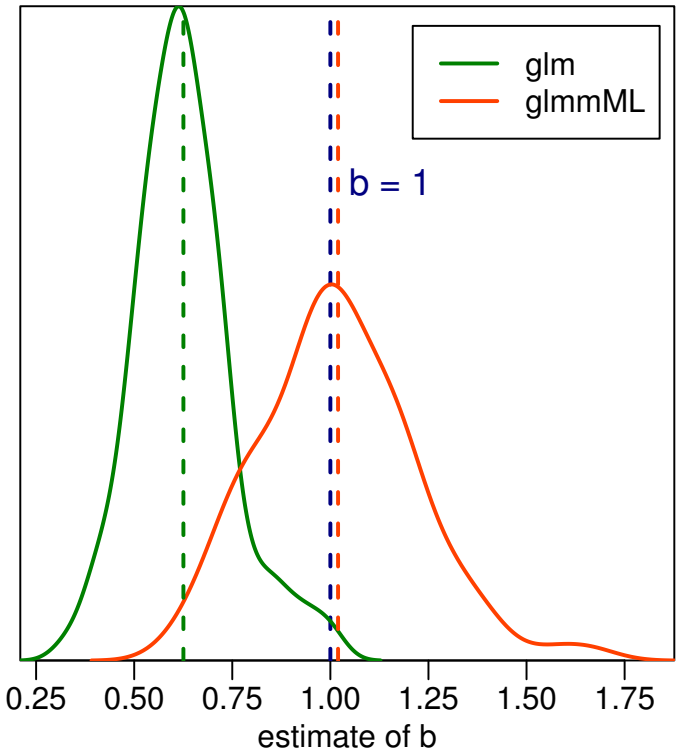
```
> glmmML(n.seed ~ 1 + x, familiy = binomial, data = d, cluster = d$id)
...(略)...
             coef se(coef)     z Pr(>|z|)
(Intercept) -3.95    1.222 -3.23  0.00120
x            1.07    0.311  3.44  0.00057
Standard deviation in mixing distribution:  2.34
...(略)...
```

- 葉数 x の係数 (coef) b の最尤推定値は $\hat{b} = 1.07$
- 「個体差」のばらつき (標準偏差) の推定値は $\hat{s} = 2.34$
  (ホントの値は $s = 3$)



glmmML() 推定結果: 偏りはないけど，ばらつく

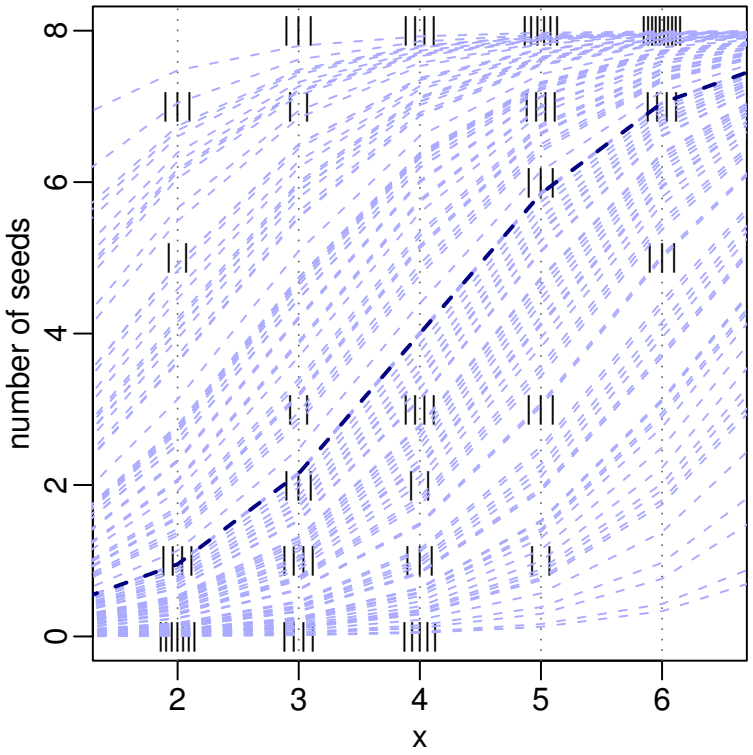
- 「glmmML() でホントにだいじょうぶか」と心配になってまたべつの集団 (100 個体) から種子データとりなおしてみた
- glmmML(n.seed　1 + x, ...) やりなおしてみた
- これを何度も何度も……100 回ほど繰りかえしてしまった
- 推定値は偏らない
- しかし推定値のばらつきは大きい
  - 改善するには …… 標本数 (100 個体) をさらに増やすしかないだろう
  - 注: glm() 使った推定では，いくら標本数ふやしても偏ったまま



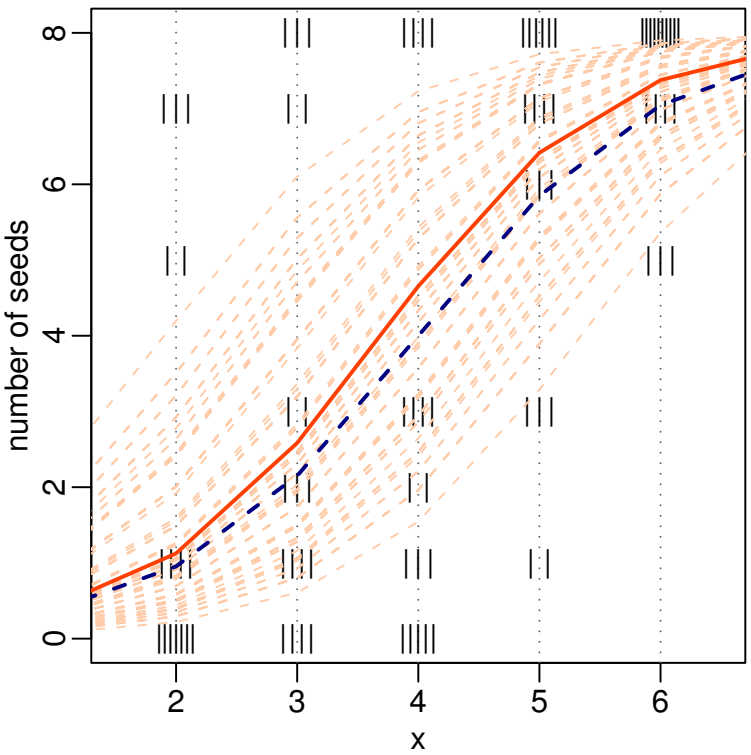
glmmML() 推定結果: 「個体差」は過小推定される

- 「個体差」標準偏差の推定値は $\hat{s} = 2.34$ (ホントの値は $s = 3$)

架空植物のデータ



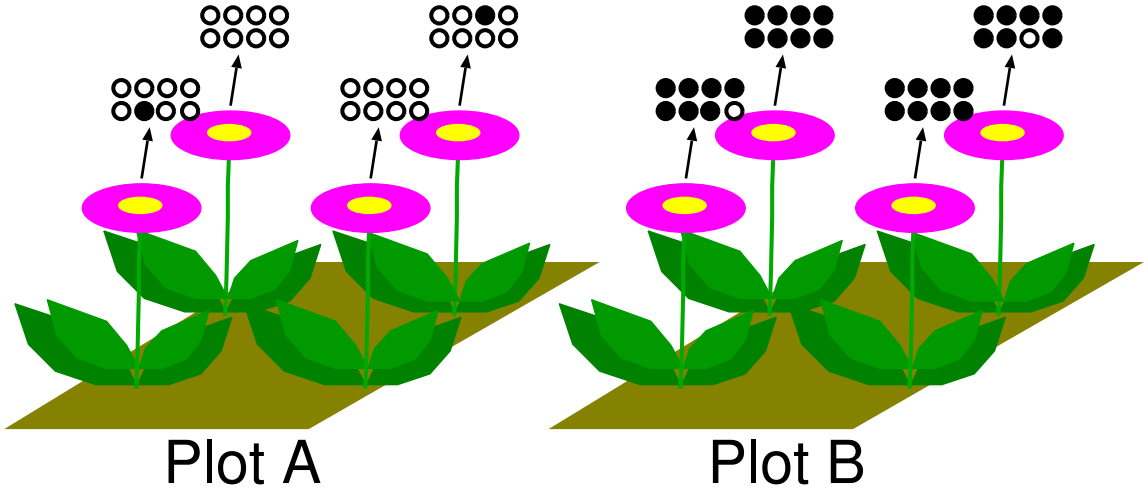
glmmML() の推定結果



- まあ (推定したかった) 葉数パラメーター b の推定結果は良さそうだし,「個体差」とか「関心のない，観測できない」量だから，どうでもいいか……と妥協してみる

「個体差」について，ちょっとつけたし

- 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

今日のまとめ: 「個体差」を考慮した解析が必要

1. **疑惑篇: `glm()` がうまくいかない状況**

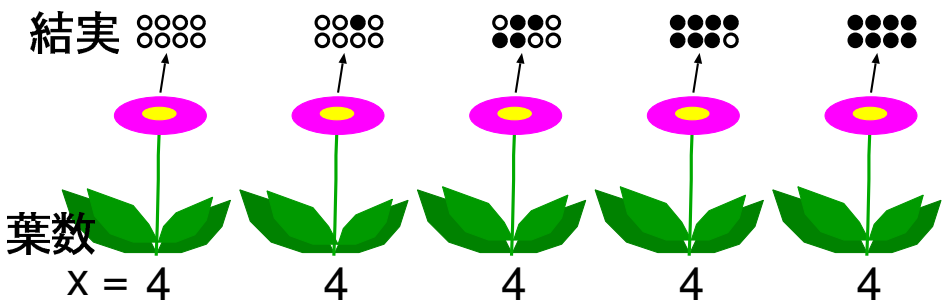
   「個体差」があるときに「偏った推定結果」などが得られる

2. **究明篇:「個体差」見つけかた作りかた**

   過分散 (overdispersion) があったら random effects を導入

3. **解決篇: 一般化線形混合モデル**

   Random effects を平均ゼロの確率変数にしてくみこむ



「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.