

生態学会新潟大会 (JES53) 自由集会 JY8

データ解析で出会う統計的問題— 「個体差」のモデリング

R でやってみる「<sup>random effects</sup>個体差」  
の統計モデリング

一般化線形混合モデル (GLMM) 入門

<http://hosho.ees.hokudai.ac.jp/~kubo/ce/2006/>

講釈: 久保拓弥 [kubo@ees.hokudai.ac.jp](mailto:kubo@ees.hokudai.ac.jp)

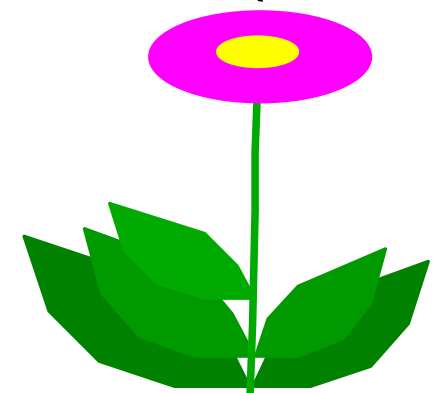
(いきなり!) 今日の架空植物: 胚珠 → 種子となる確率

[架空植物の性質あれこれ]

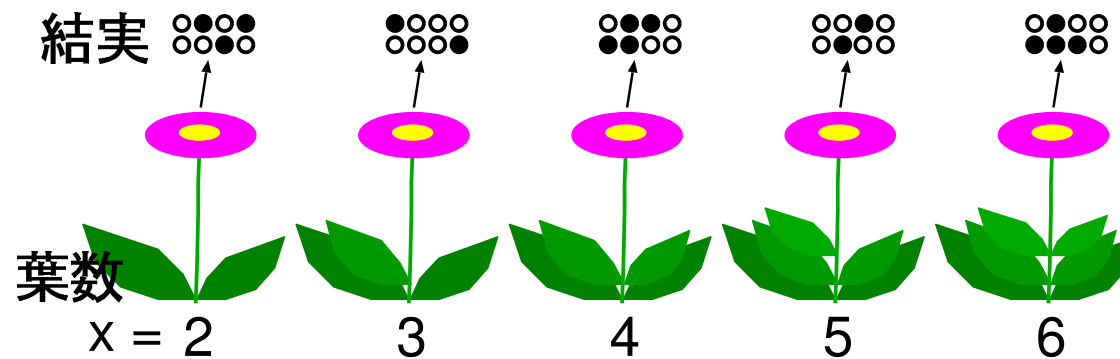
- 花の胚珠数は 8
- 結実率  $p$ : ある胚珠が種子になる確率

: 結実成功胚珠, 結実失敗胚珠

結実 



葉数  $x = 5$



葉数  $x$  と結実確率の関係に興味ある

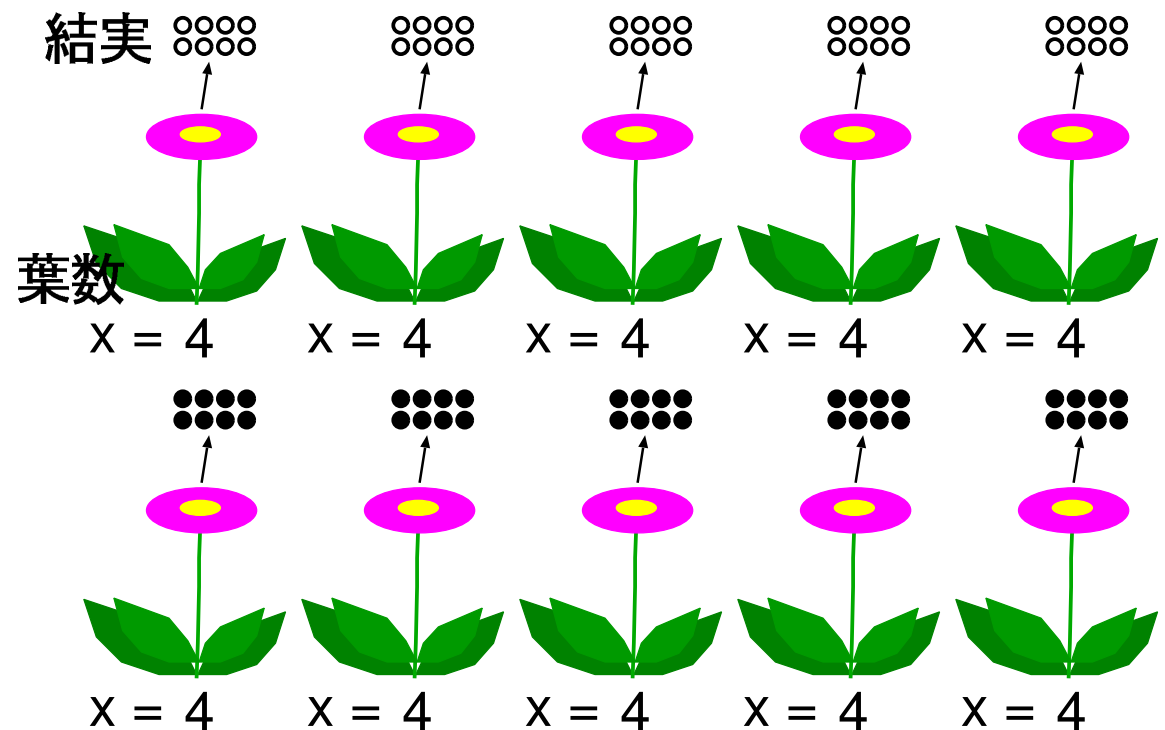
# 今日の重要用語: overdispersion と「個体差」

観測データ例:

- 完全結実失敗個体 5
- 完全結実成功個体 5

.....で「平均結実率 **0.5** でした」だって???

## Overdispersion!



- (この例の場合) 二項分布から逸脱してる状況を overdispersion という
- $p = 0.5$  の二項分布で期待される SD は  $\sqrt{8 \times 0.5^2} = 1.41$
- しかしこの観測データ例の SD は **4.28**
- ばらつきがでかすぎるので overdispersion とよぶ

# 今日の重要用語: overdispersion と「個体差」

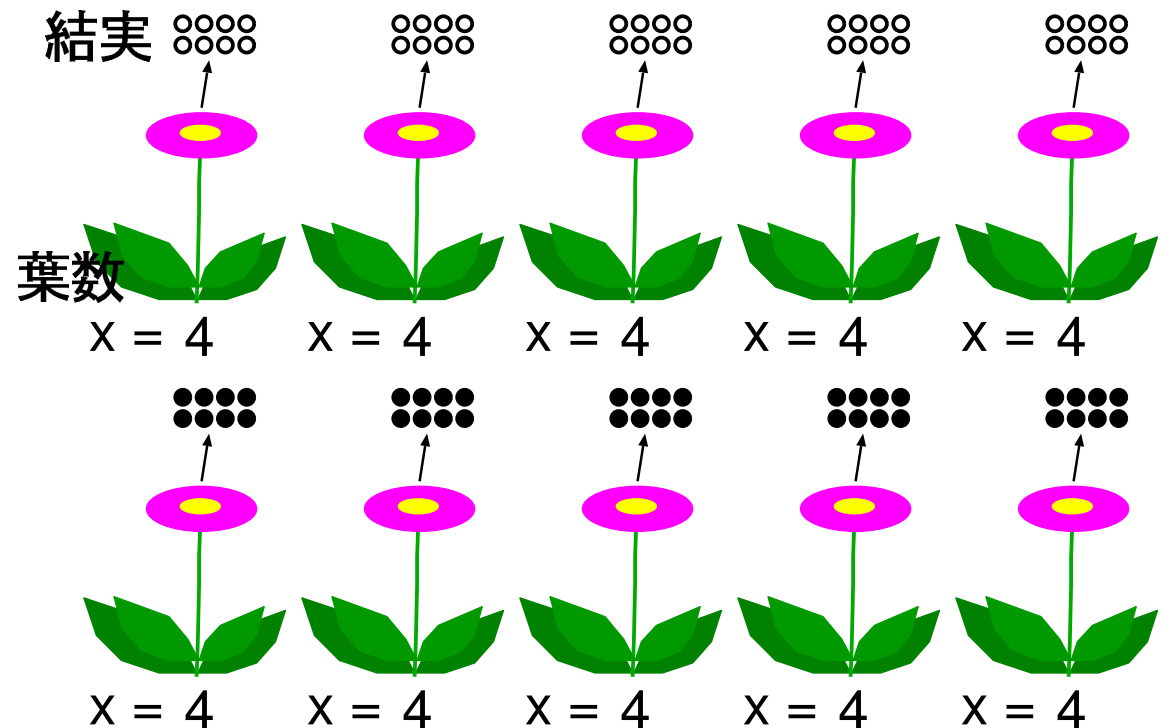
観測データ例:

- 完全結実失敗個体 5
- 完全結実成功個体 5

.....で「平均結実率 0.5 でした」だって???

## Overdispersion!

- 「説明変数である葉数  $x$ 」はそろえているのにばらつく.....
- ばらつきを増大させるわけわからん効果が **random effects**
- ここでは random effects を「個体差」と仮称する
- 葉数  $x$  など **fixed effects** は「個体差」とは呼ばない



# 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

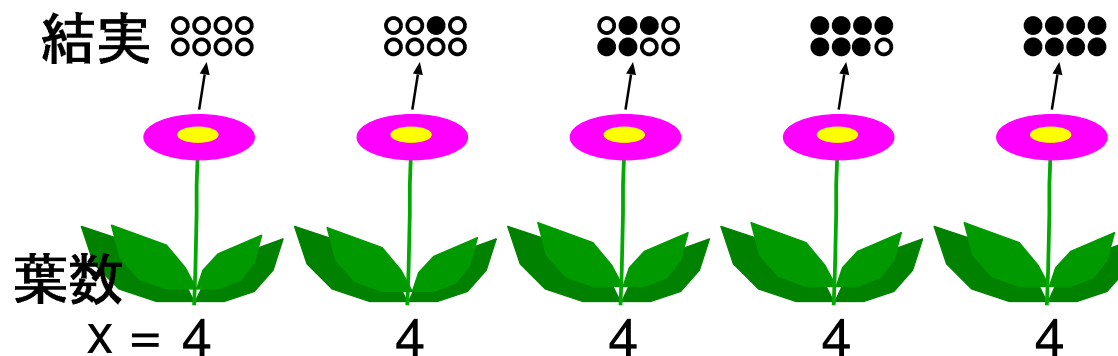
データを何回とりなおしてもダメ?

## 2. 究明篇: 「個体差」 見つけかた作りかた

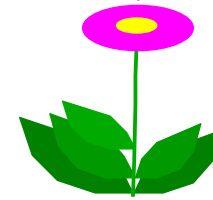
過分散 (overdispersion) を表現するモデル

## 3. 解決篇: 一般化線形混合モデル

Fixed & Random effects 同時に



結実  $\circ\circ\circ\circ$



葉数  $x = 5$

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

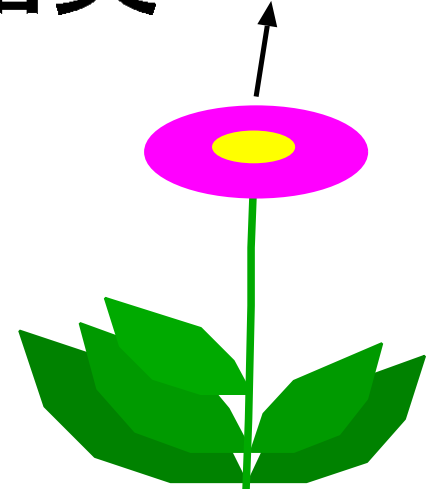
データを何回とりなおしてもダメ?

# 今日の架空植物 (続) : Q. 胚珠が種子になる確率?

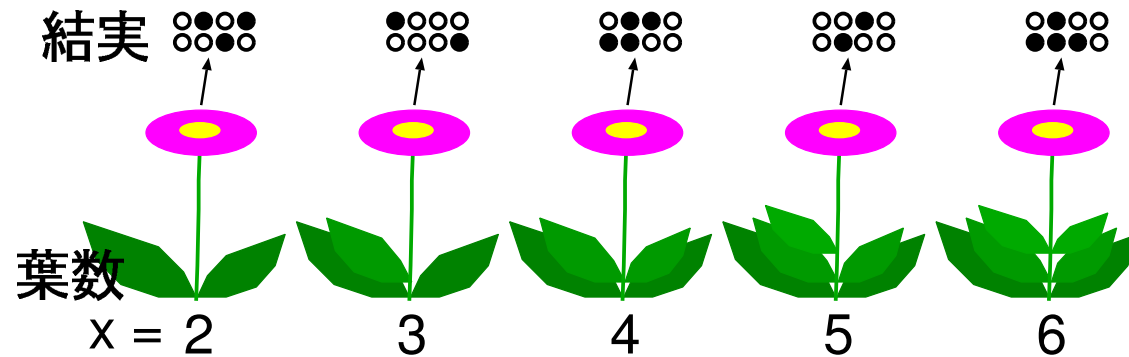
[架空植物の性質あれこれ]

- 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)

結実 



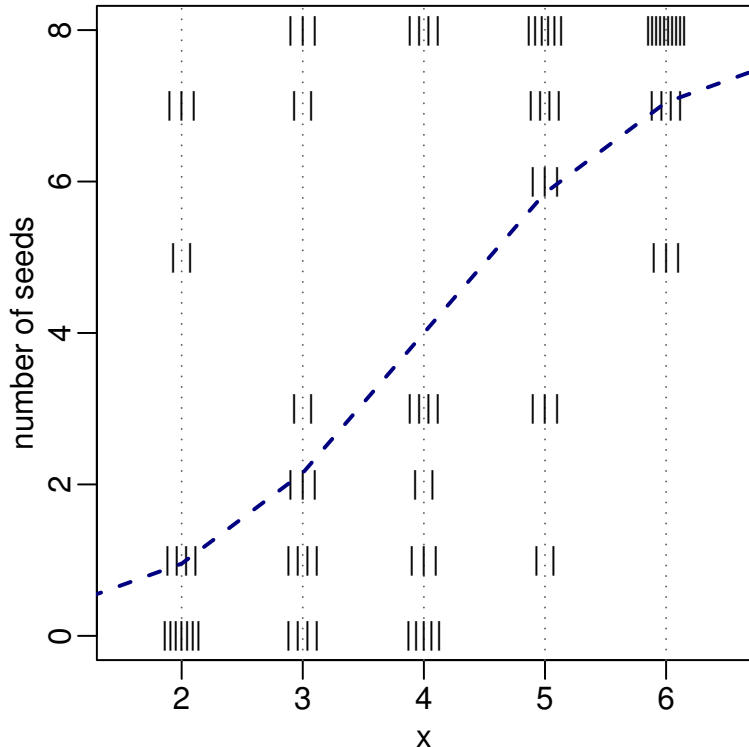
葉数  $x = 5$



- 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]

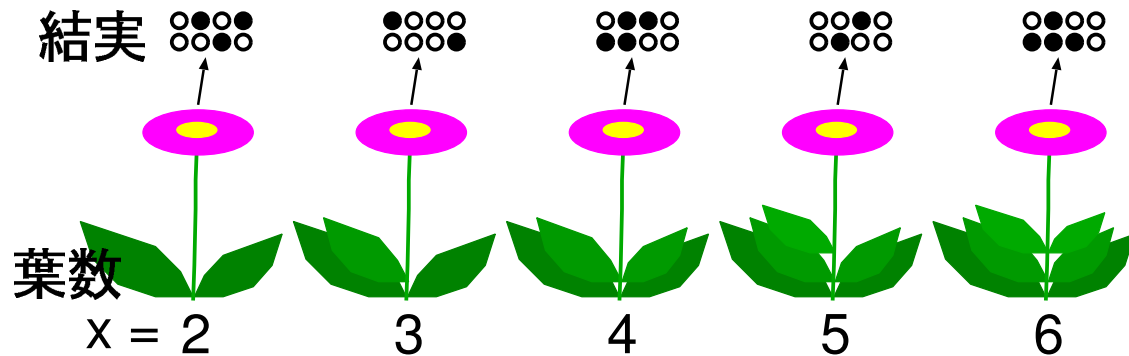
- 葉数  $x$  大きいほど結実率が高い

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

$a = -4$  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)

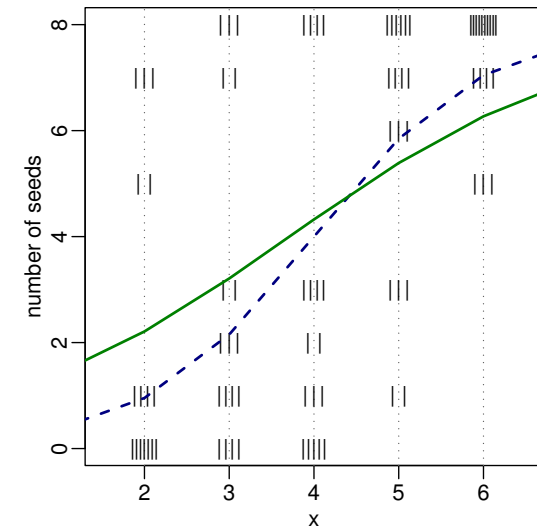


# こんなのロジスティック回帰で簡単に.....?

```
> 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)...
```

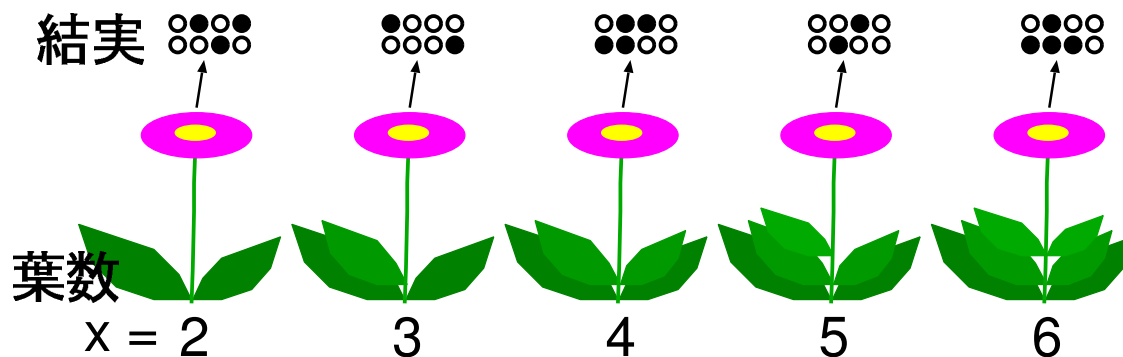
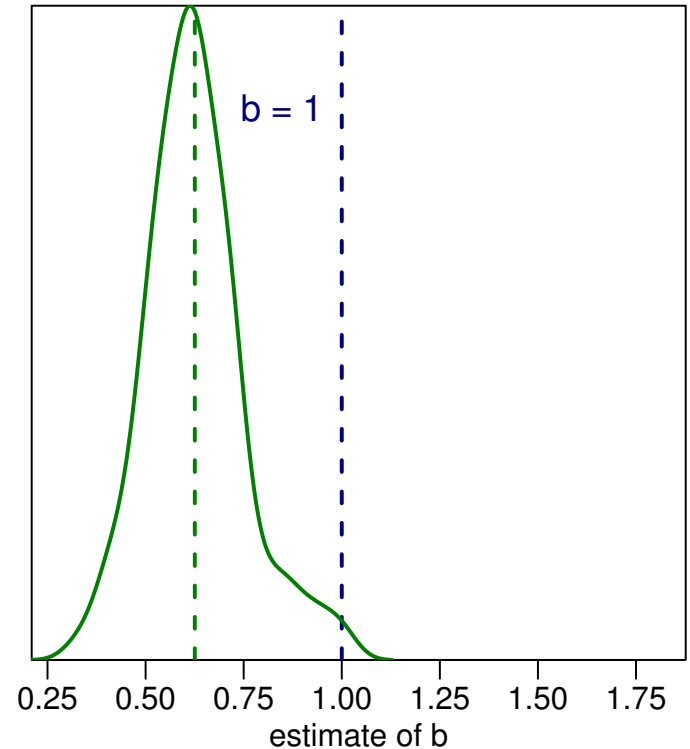


なんか葉数パラメーター  
の推定値がずれてるんで  
すけど.....  $\hat{b} = 0.5627$   
(真の値は  $b = 1$ )



# 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 回ほど繰り返してしまっただ
- `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?

- 気を取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

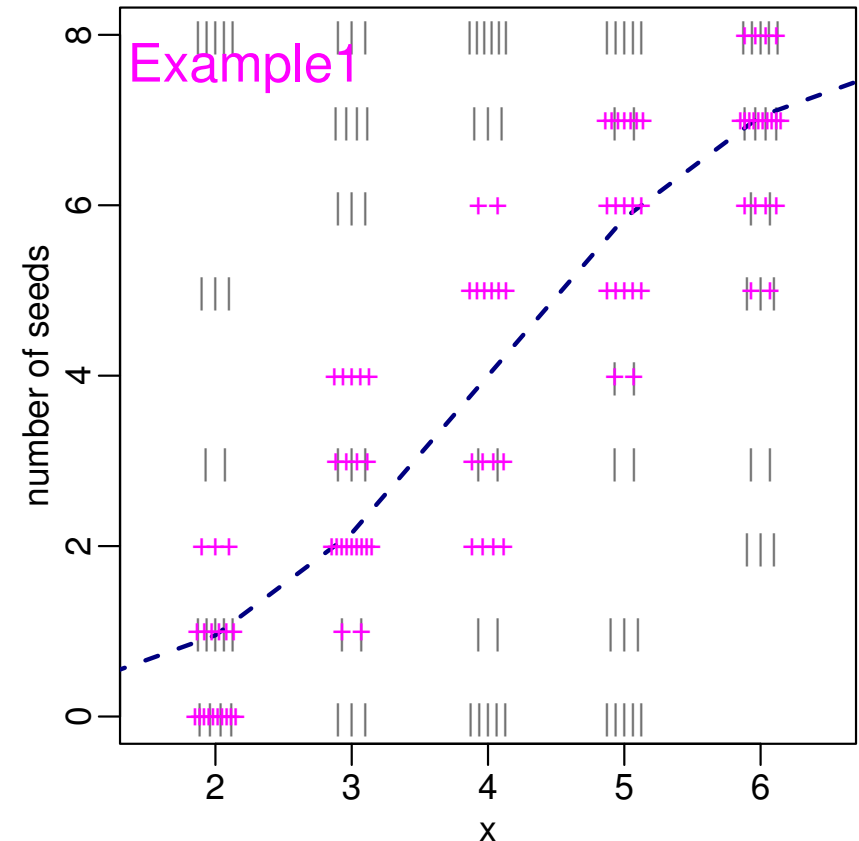
$$\frac{8!}{y!(8-y)!} p^y (1-p)^{8-y}$$

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

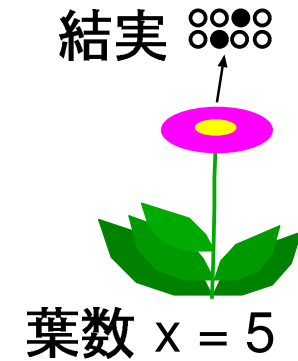
- **[R による実験で調べる]** 結実率を .....

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

- .....とにおいて 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ートした**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.....?



人間が測定してない・できないばらつき



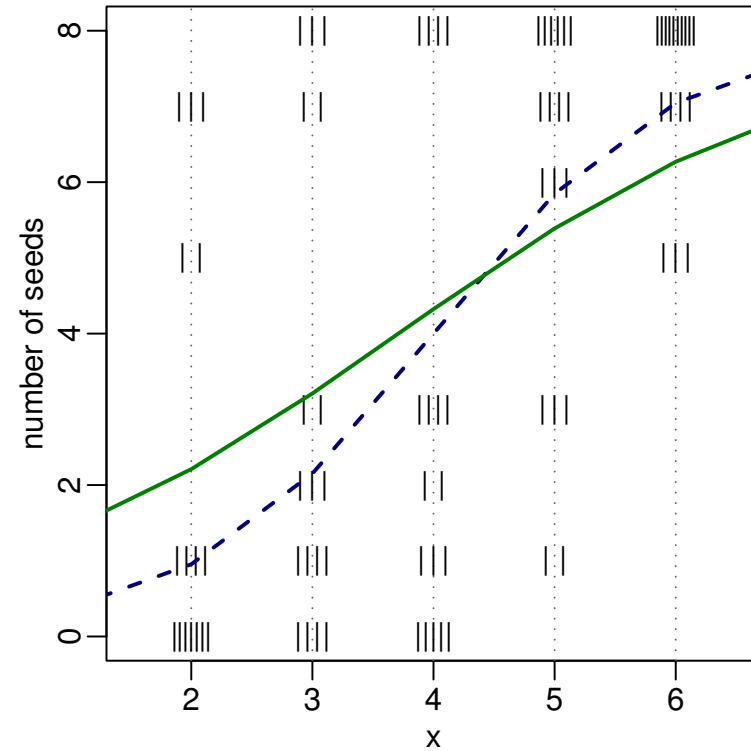
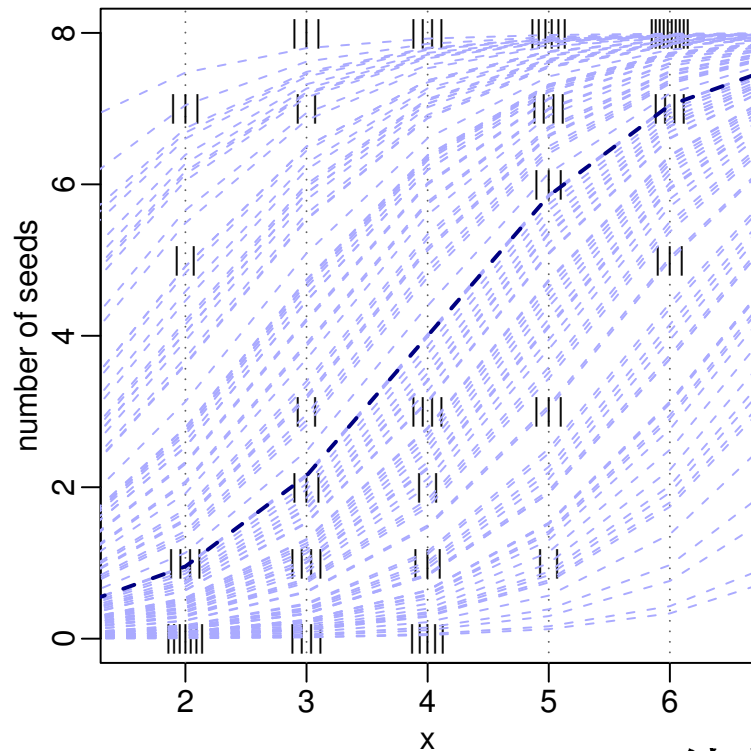
## 2. 究明篇: 「個体差」見つけかた作りかた

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

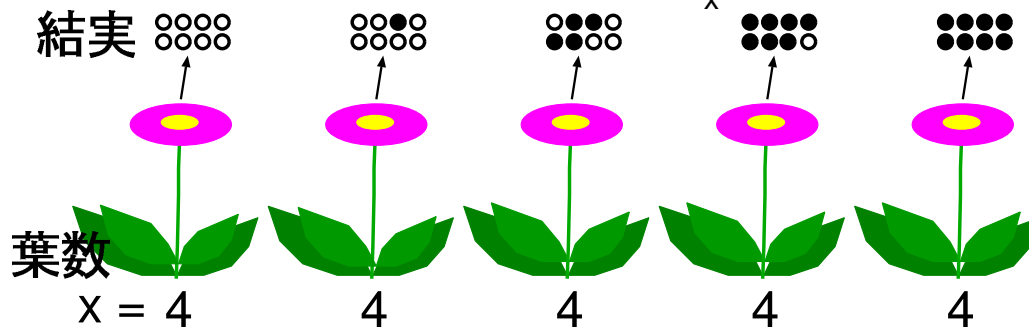
「傾き」 ( $b$ ) には「個体差」ない  
 「切片」 ( $a$ ) には「個体差」ある

「個体差」を考慮していない `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.....?

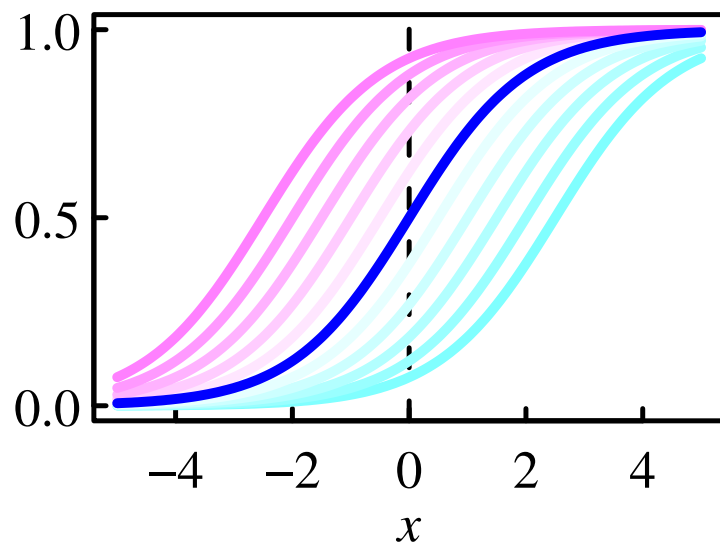


# (ちょっと復習) 「ロジスティック関数」って何だっけ?

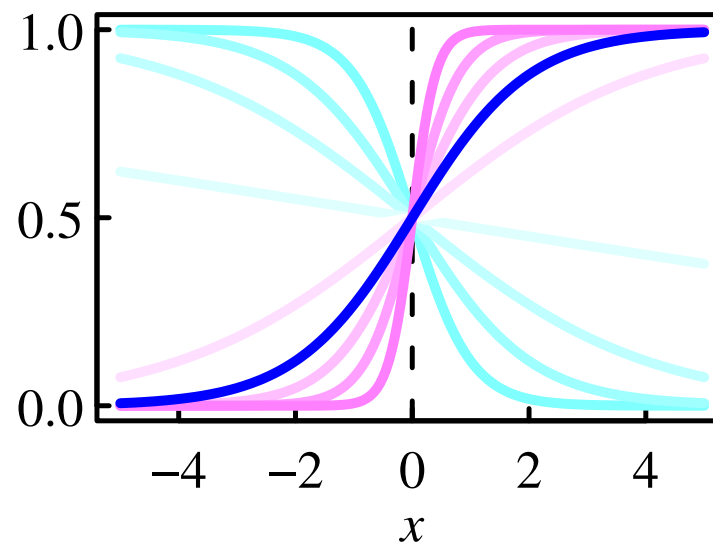
$$p = \frac{1}{1 + \exp(-(a + bx))}$$

( $\exp(Z) = e^Z$  のこと)

$a$  だけ変化させる



$b$  だけ変化させる

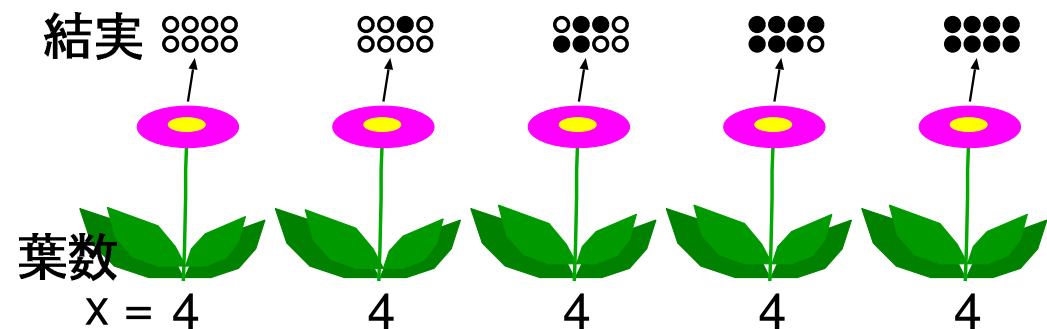
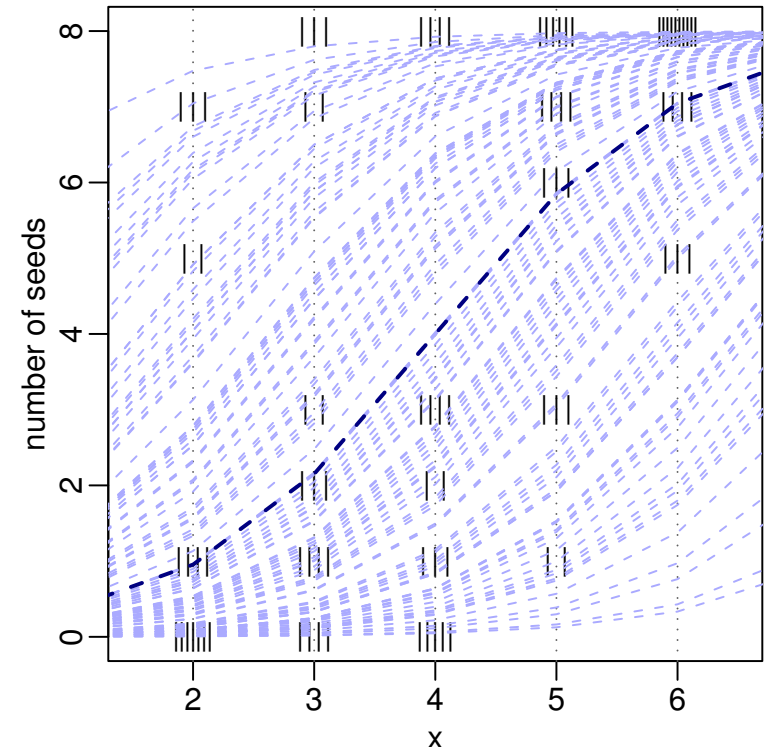


つまりパラメーター  $\{a, b\}$  や説明変数  $x$  がどんな値をとっても確率  $p$  は  $0 \leq p \leq 1$  となる便利な関数

# 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  $x$  などに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，土壤中の栄養塩類，  
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!)



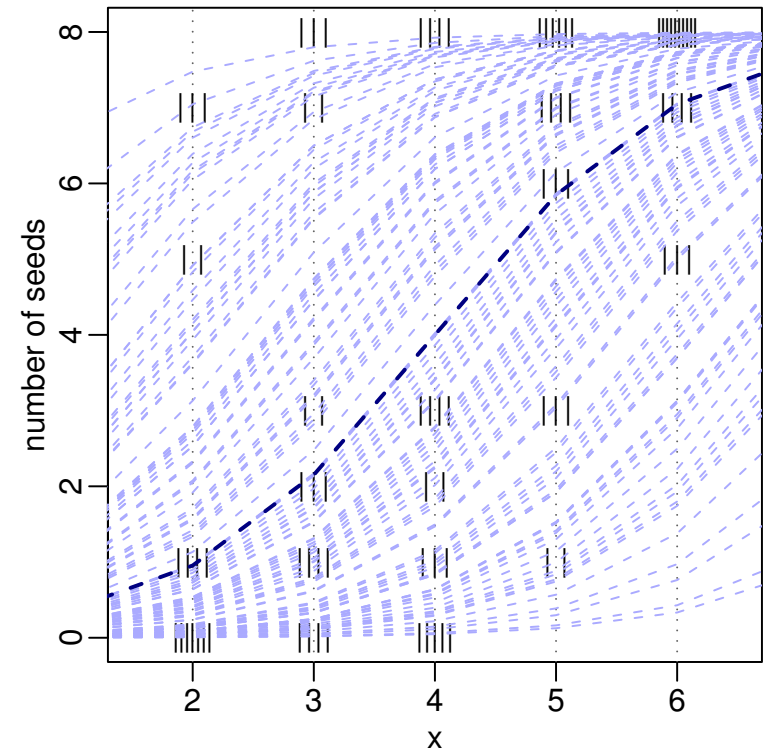
# 「個体差」とは何か? (統計モデリング的側面)

- 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





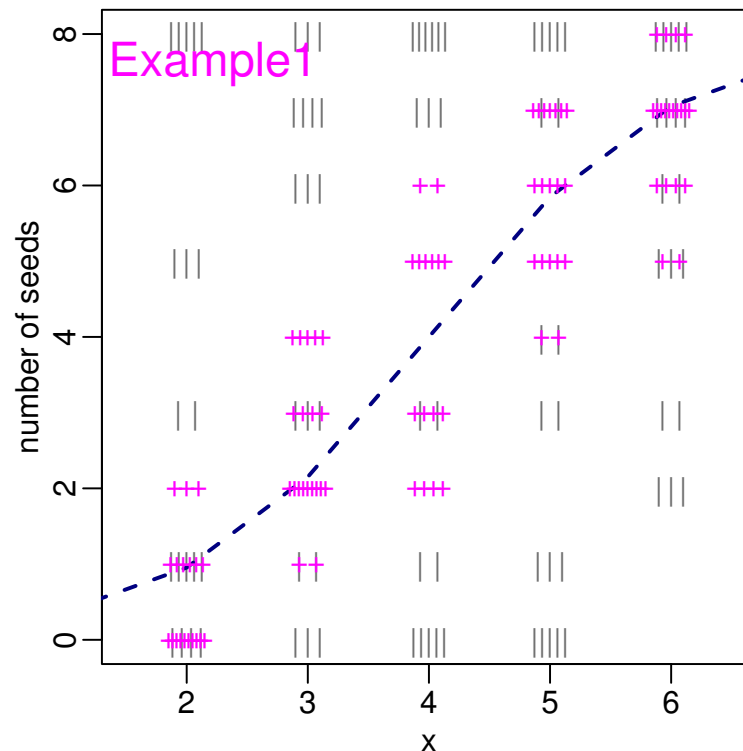
# Random effects がもたらす過分散 (overdispersion)

- Random effects なしの結実率モデル

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

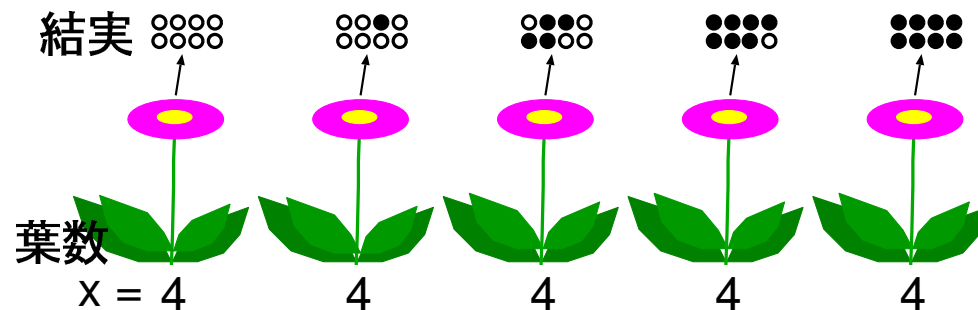
を仮定して二項乱数を生成させると + のようなデータが得られる

- 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)

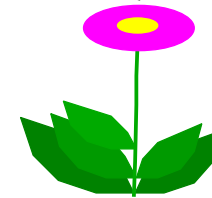


- これを過分散 (overdispersion) という

- 過分散を発見したら「個体差」無視できないと考える



結実 



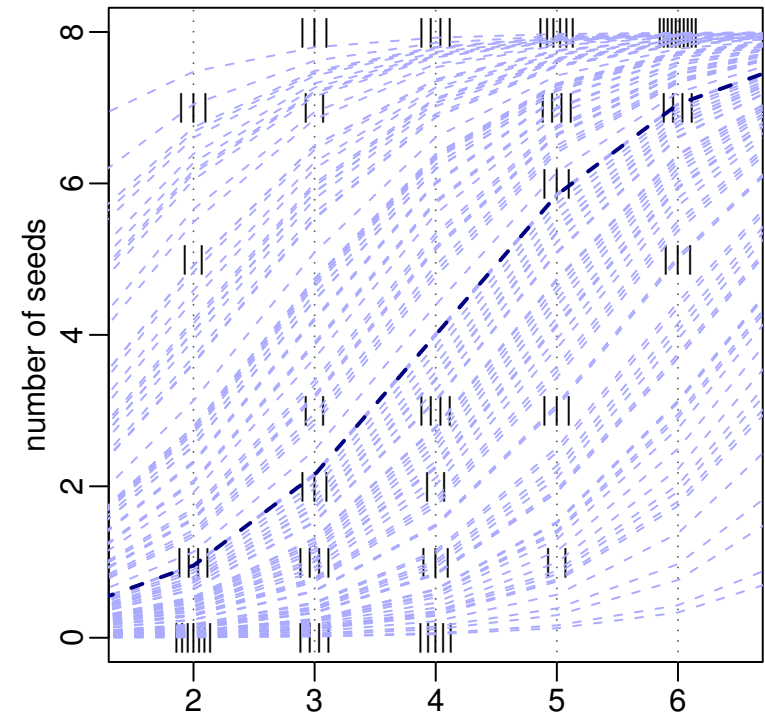
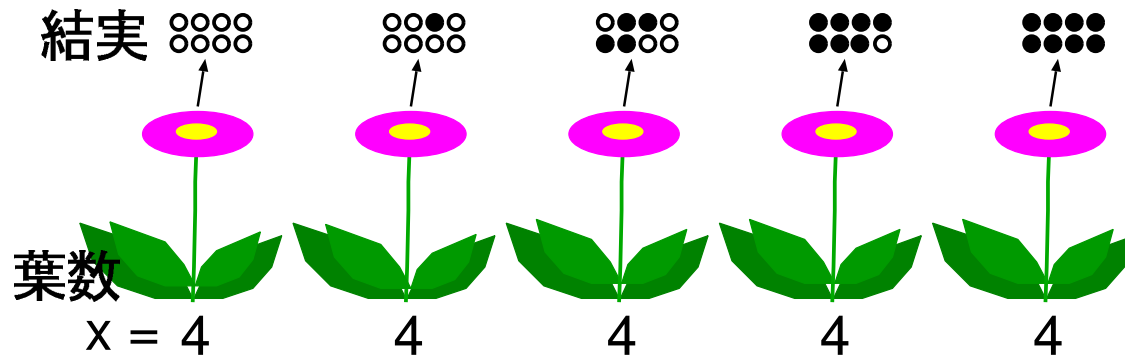
葉数  $x = 5$

「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

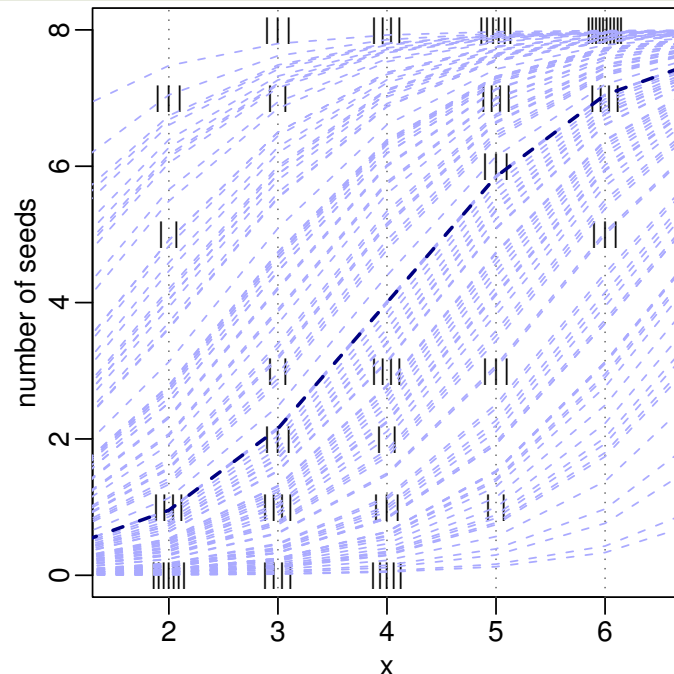
# 「個体差」 (random effects) を表現する統計モデル

- 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

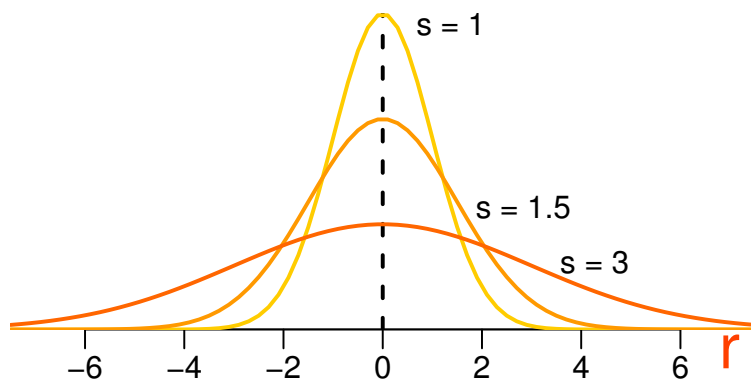
$$p(a, b, r) = \frac{1}{1 + \exp(-(a + bx + r))}$$



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

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

- 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

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

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

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

## 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

$$f(y_i | a, b, r_i) g(r_i, s)$$

- 「個体差」  $r_i$  はわからないので全ての可能性の期待値をとる

$$\int_{-\infty}^{\infty} f(y_i | a, b, r_i) g(r_i, s) dr_i \quad (\text{これはある個体 } i \text{ の尤度である})$$

- 全個体の尤度は

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

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 .....

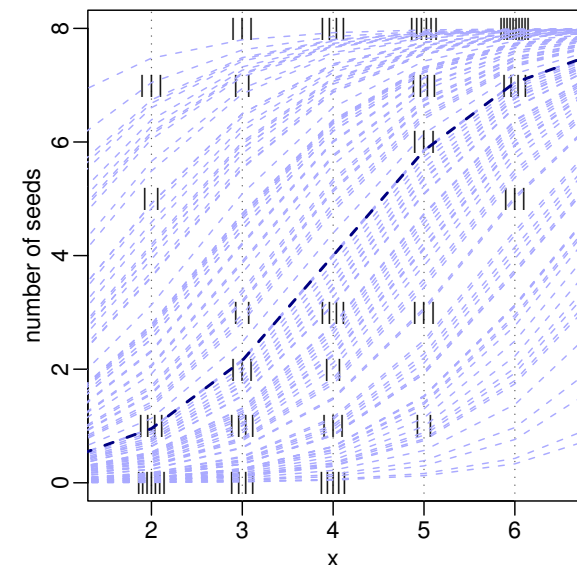
ロジスティック回帰・ポアソン回帰など一般化線形モデル (GLM) に random effects の項を加えて混合モデル化したものの総称



GLMM: generalized linear mixed model

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

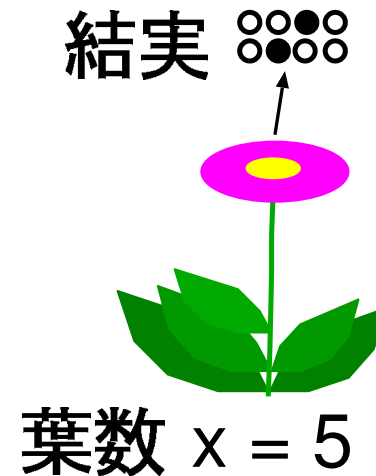
- 混合モデルの **尤度を数値積分** で計算する
- 「**定数項**」のみが random effects になりうる (つまり「横方向のずれ」だけ)
- family = binomial or poisson だけ
- Göran Broström さんが開発・発展させている



# 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 (結実していない) 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, family = 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)
- `family = 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

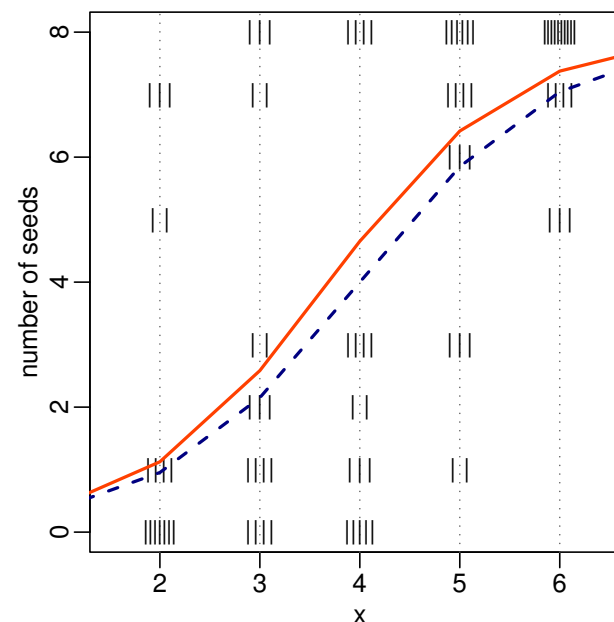
# 混合 logistic モデルの最尤推定 , `glmmML()` で (3)

## `glmmML()` の推定結果を読む

```
> glmmML(n.seed ~ 1 + x, family = binomial, data = d, cluster = d$id)
...(略)...
```

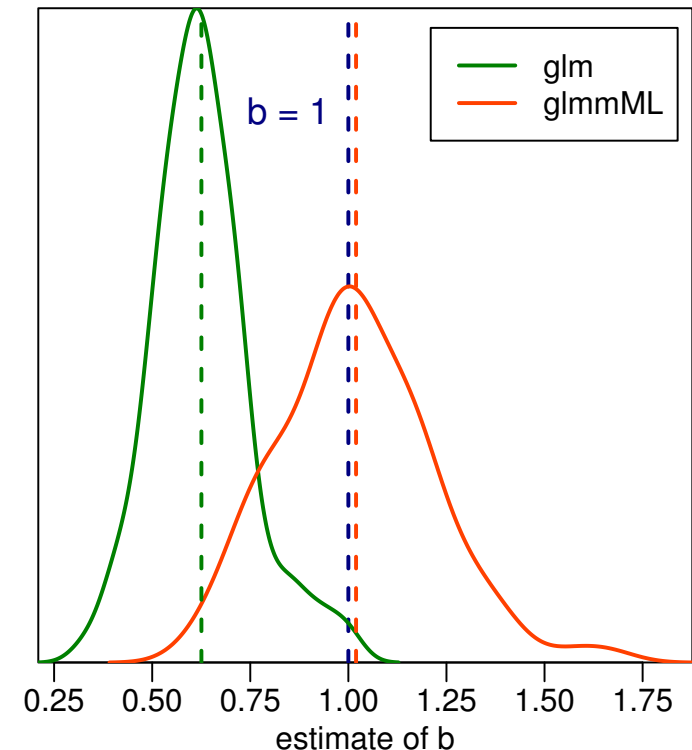
```
      coef se(coef)      z Pr(>|z|)
(Intercept) -3.95   1.222 -3.23  0.00120
x             1.07   0.311  3.44  0.00057
Standard deviation in mixing distribution:  2.34
...(略)...
```

- 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

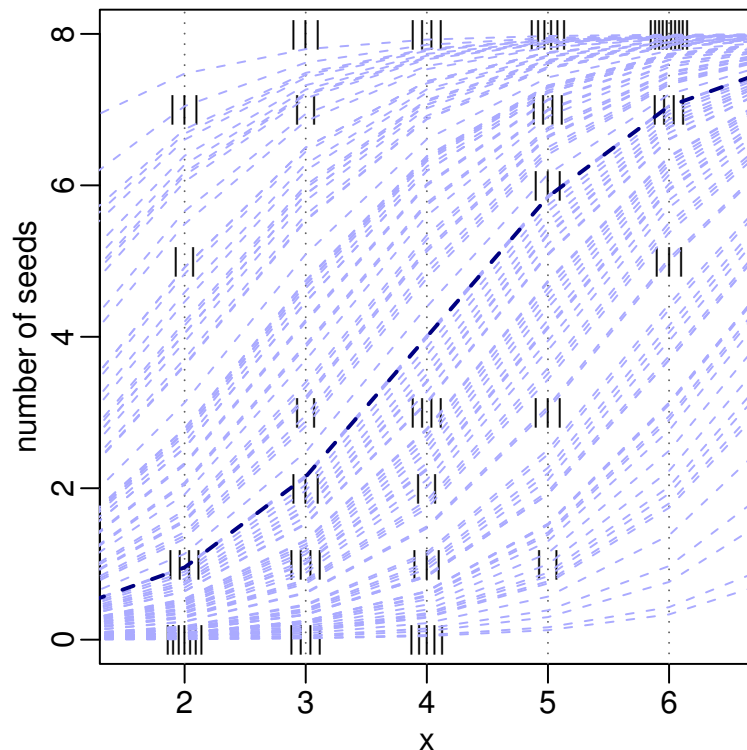
- 「glmmML() でホントにだいじょうぶか」と心配になってまたべつの集団 (100 個体) から種子データとりなおしてみた
- `glmmML(n.seed = 1 + x, ...)` やりなおしてみた
- これを何度も何度も.....**100** 回ほど繰り返してしまった
- 推定値は偏らない
- しかし推定値のばらつきは大きい
  - 改善するには ..... 標本数 (100 個体) をさらに増やすしかないだろう
  - 注: `glm()` 使った推定では, いくら標本数ふやしても偏ったまま



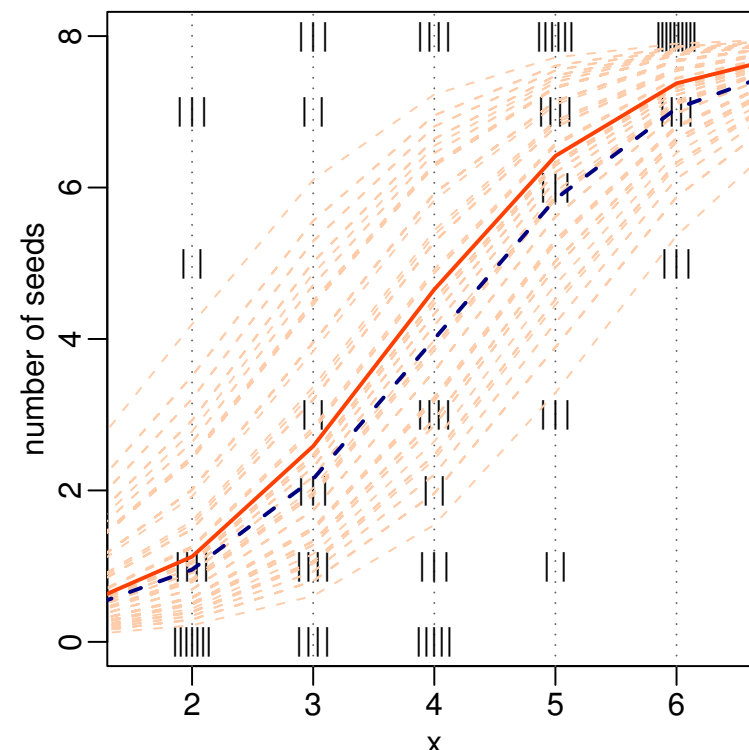
# glmmML() 推定結果: 「個体差」は過小推定される

- 「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



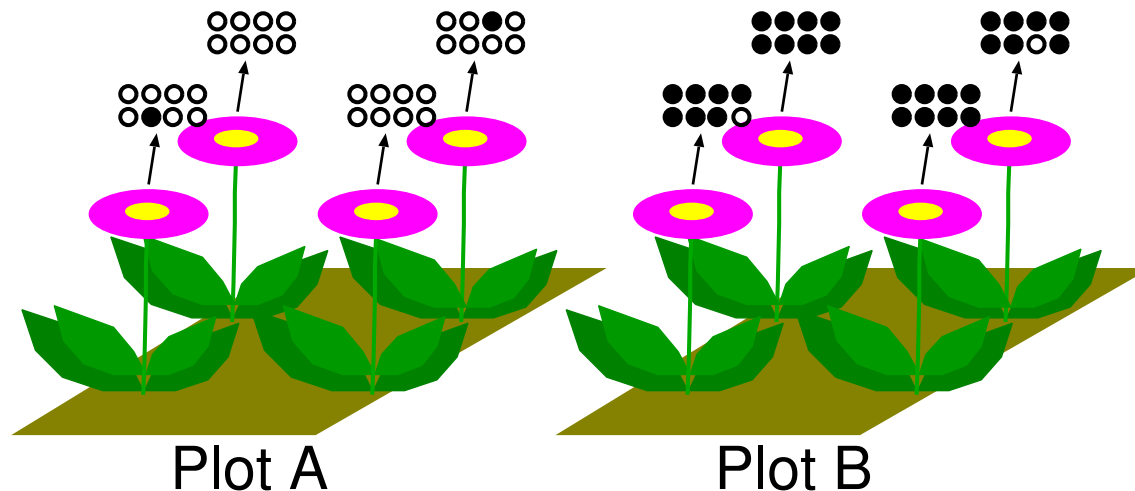
glmmML() の推定結果



- `fit <- glmmML(...)` と推定結果を格納して, `fit$frail` とすると「個体ごとの random effects の期待値」が得られる

## 「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

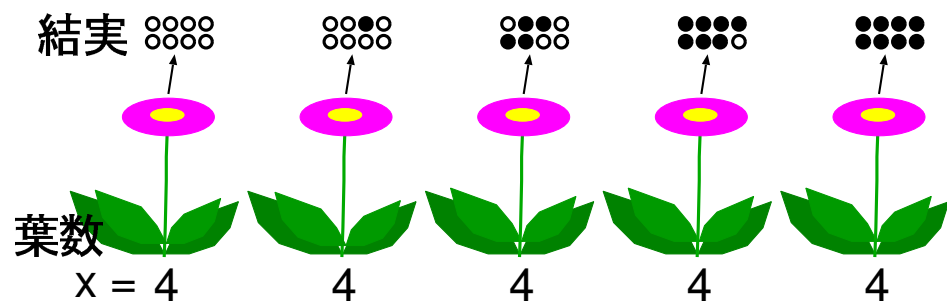
「個体差」があるときに「偏った推定結果」などが得られる

## 2. 究明篇: 「個体差」見つけかた作りかた

Overdispersion は random effects は説明できるかも

## 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.

## 今日は説明できなかったコトども

- ネストされた (階層性ある) 状況での複数の random effects 入れるモデリング
- R でできるいろいろな GLMM
- 縦断的 (longitudinal) データの解析と混合モデル
- 空間構造のある random effects, ランダムフィールド
- 経験ベイズモデルと混合モデルの関係
- めんどろな混合モデルと Markov Chain Monte Carlo (MCMC) 計算の関係

..... などなどなど

またいつかべつの機会に.....