統計モデルの基礎: 何でも「割算」するな!

ロジスティック回帰と GLMM

久保拓弥 kubo@ees.hokudai.ac.jp

生態学基礎論 (**生物多様性論** II) http://goo.gl/mDYZzP

2015–01–21

ファイル更新時刻: 2015–01–21 16:23

agenda

この時間のハナシ I

1. “N 個のうち k 個が生きてる” タイプのデータ
   count data or categorical data with upper bound
   上限のあるカウントデータ
2. ロジスティック回帰の部品
   logistic regression
   二項分布 binomial distribution と logit link function
3. ちょっとだけ交互作用項 について
   interaction term
   complicate terms in linear predictor
   線形予測子の中の複雑な項
4. 何でも「割算」するな!
   NO $\frac{\text{data}}{\text{data}}$ statistics!
   「脱」割算の offset 項わざ
   use GLM with offset term
5. GLM では説明できない種子データ
   GLM is not enough!
   「ばらつき」が大きすぎる!
   overdispersion data
6. 過分散と個体差
   観測されていない個体差がもたらす過分散
   overdispersion caused by individual differences

agenda

この時間のハナシ II

Genralized Liear Mixed Model

❼ **一般化線形混合モデル**

parameters for indviduals

個体差をあらわすパラメーターを追加

Maximum likelihood estimation for GLMM

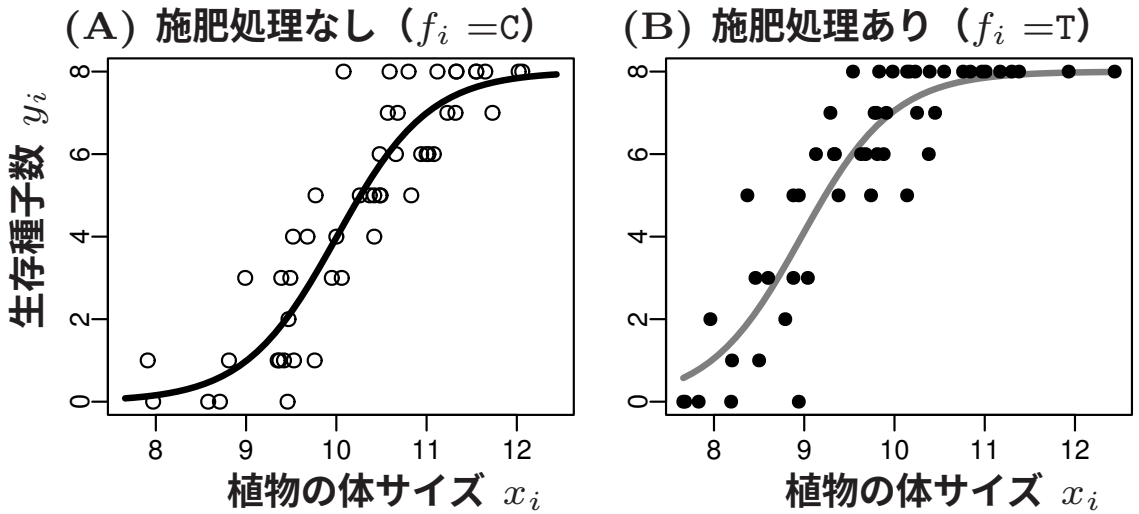
❽ **一般化線形混合モデルの最尤推定**

個体差 r_i を積分して消す尤度方程式

Real data are ... harder!!

❾ **実際のデータ構造はもっと複雑**

複数の階層，時間や空間の構造などなど

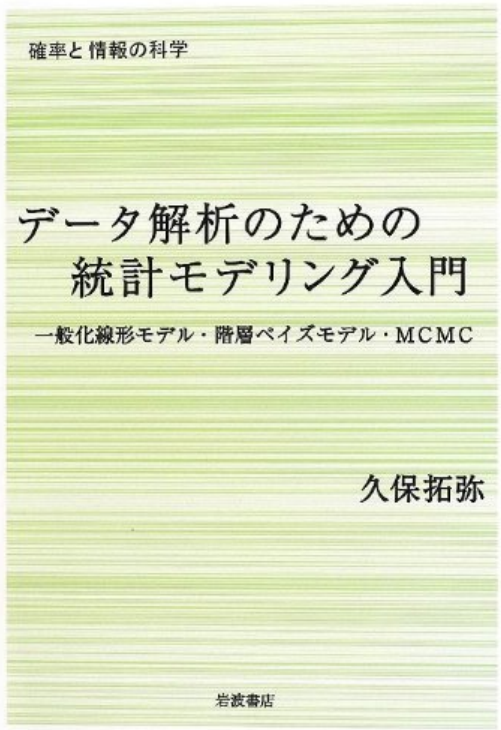


今日の内容と「統計モデリング入門」との対応

http://goo.gl/Ufq2

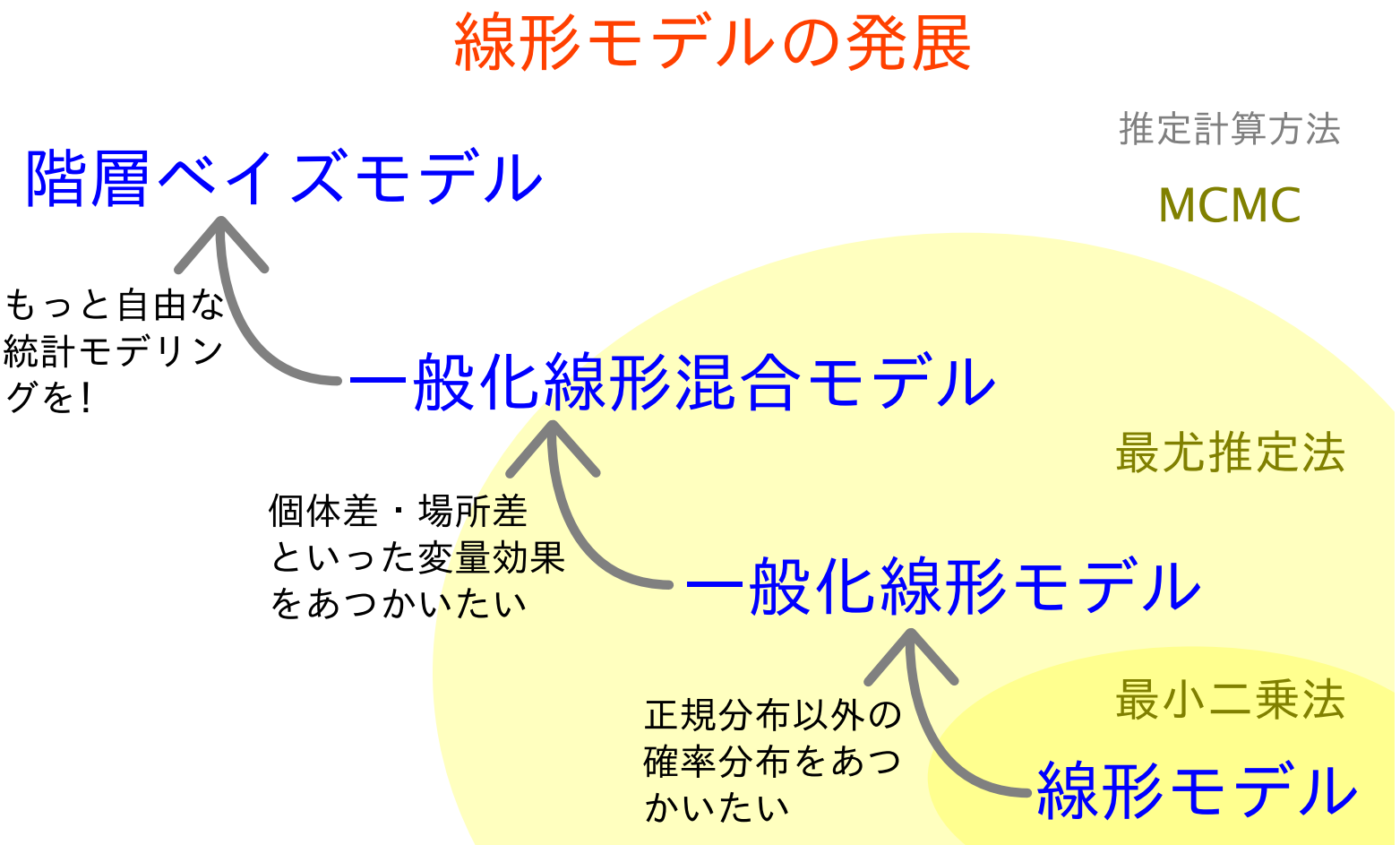
今日はおもに「第 6–7 章」の内容を説明します．

- 著者: 久保拓弥
- 出版社: 岩波書店
- 2012–05–18 刊行



statistaical models appeared in the class

この授業であつかう統計モデルたち



データの特徴にあわせて線形モデルを改良・発展させる

一般化線形モデルって何?

Generalized Linear Model

一般化線形モデル (GLM)

- ポアソン回帰 (Poisson regression)
- **ロジスティック回帰 (logistic regression)**
- 直線回帰 (linear regression)
- ……

how to specify GLM

一般化線形モデルを作る

Generalized Linear Model

一般化線形モデル (GLM)

probability distribution

- **確 率 分 布 は**?

linear predictor

- **線形予測子は**?

link function

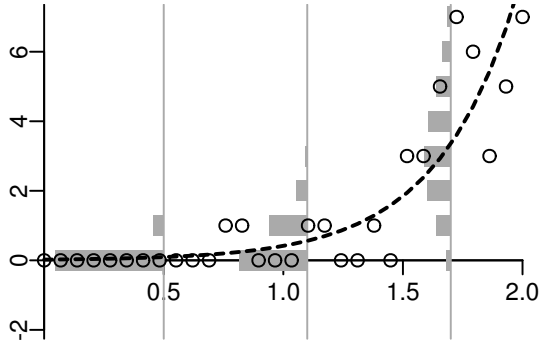
- **リンク関数は**?

how to specify Poisson regression model, a GLM

GLM のひとつであるポアソン回帰モデルを指定する

ポアソン回帰のモデル

- 確率分布 (probability distribution): ポアソン分布 (Poisson distribution)
- 線形予測子 (linear predictor): e.g., $\beta_1 + \beta_2 x_i$
- リンク関数 (link function): 対数リンク関数 (log link function)



how to specify logistic regression model, a GLM

GLM のひとつであるlogistic 回帰モデルを指定する

ロジスティック回帰のモデル

probability distribution　binomial distribution

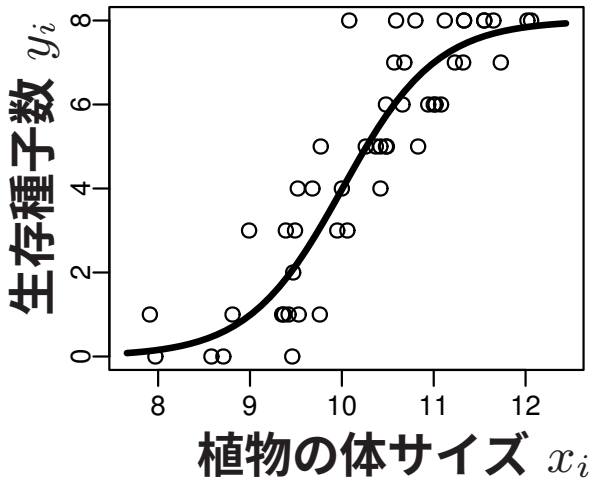
- **確率分布**: **二項分布**

linear predictor

- **線形予測子**: e.g., $\beta_1 + \beta_2 x_i$

link function

- **リンク関数**: logit **リンク関数**

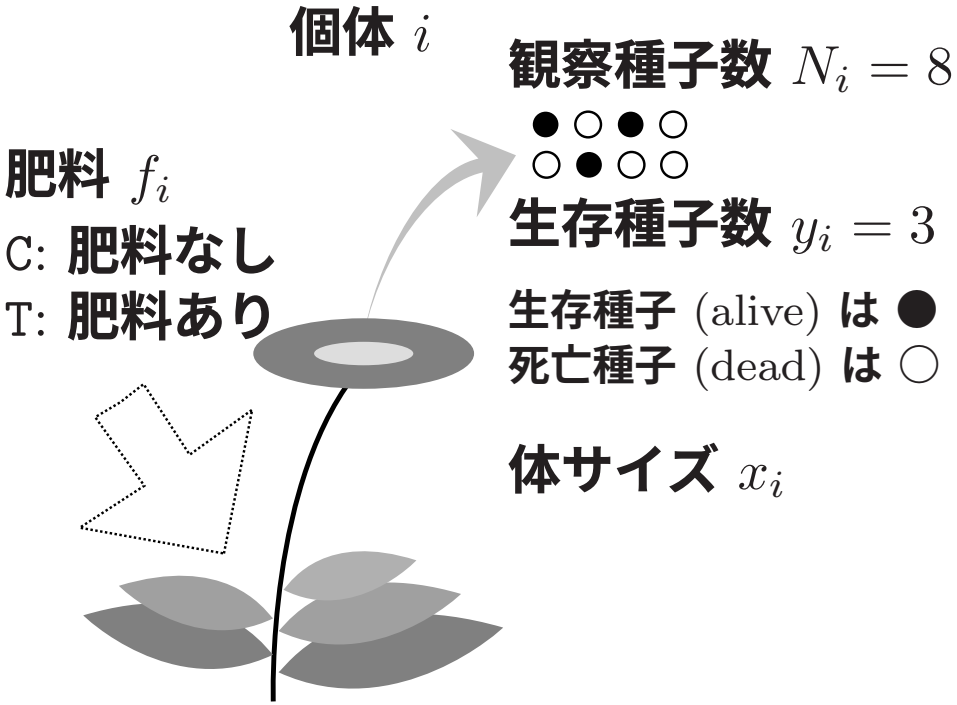


1. “N 個のうち k 個が生きてる” タイプのデータ

count data or categorical data with upper bound
上限のあるカウントデータ

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



Reading data file

データファイルを読みこむ 

data4a.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("data4a.csv")
```

or

```
> d <- read.csv(
+ "http://hosho.ees.hokudai.ac.jp/~kubo/stat/2013/fig/binomial/data4a.csv")
```

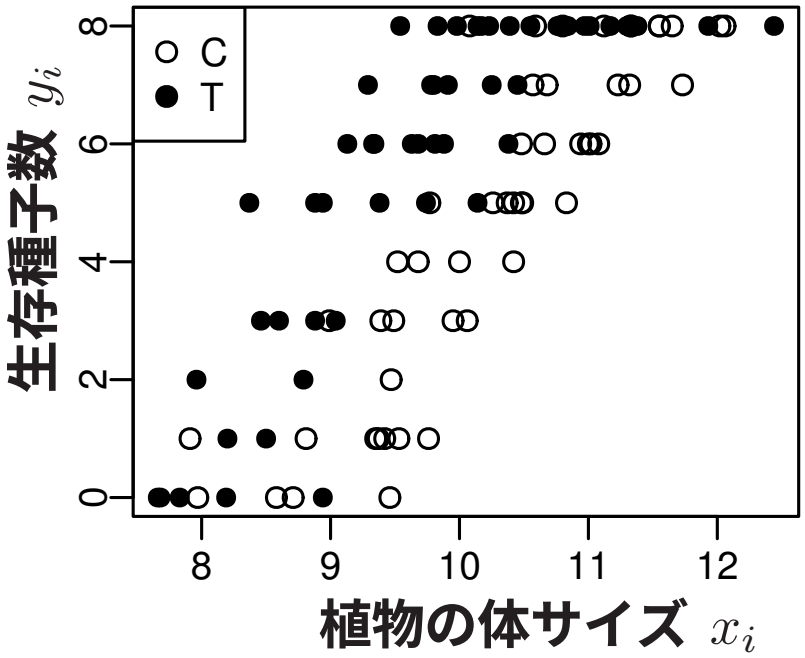
データは d と名付けられた data frame (「表」みたいなもの) に格納される

data frame d を調べる

```
> summary(d)
       N              y                x           f
 Min.   :8   Min.   :0.00   Min.   : 7.660   C:50
 1st Qu.:8   1st Qu.:3.00   1st Qu.: 9.338   T:50
 Median :8   Median :6.00   Median : 9.965
 Mean   :8   Mean   :5.08   Mean   : 9.967
 3rd Qu.:8   3rd Qu.:8.00   3rd Qu.:10.770
 Max.   :8   Max.   :8.00   Max.   :12.440
```

まずはデータを図にしてみる

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



今回は施肥処理 (fertilization) がきいている (effective)?

logistic regression

2. ロジスティック回帰の部品

二項分布 binomial distribution と logit link function

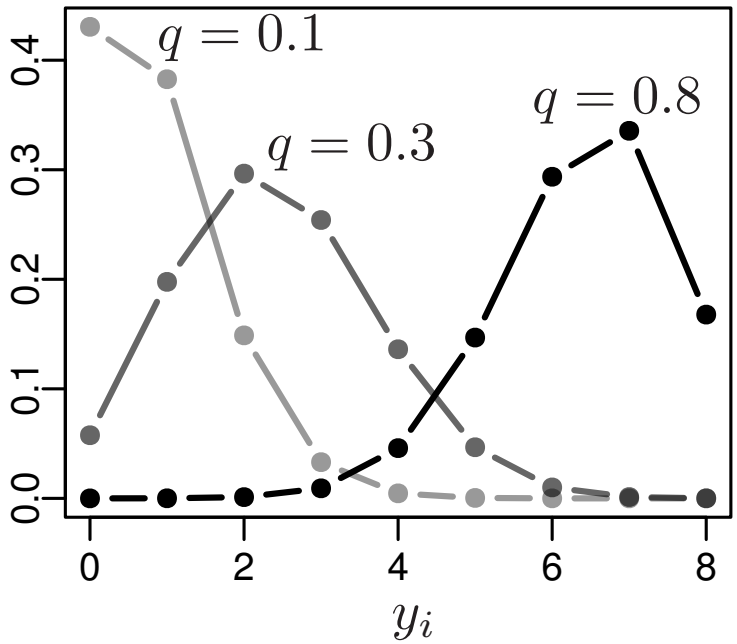
binomial distribution

二 項 分 布：N 回のうち y 回，となる確率

$$p(y|N,q) = \binom{N}{y} q^y (1-q)^{N-y}$$

$\binom{N}{y}$ は「N 個の観察種子の中から y 個の生存種子を選びだす場合の数」

確率 $p(y_i|8,q)$

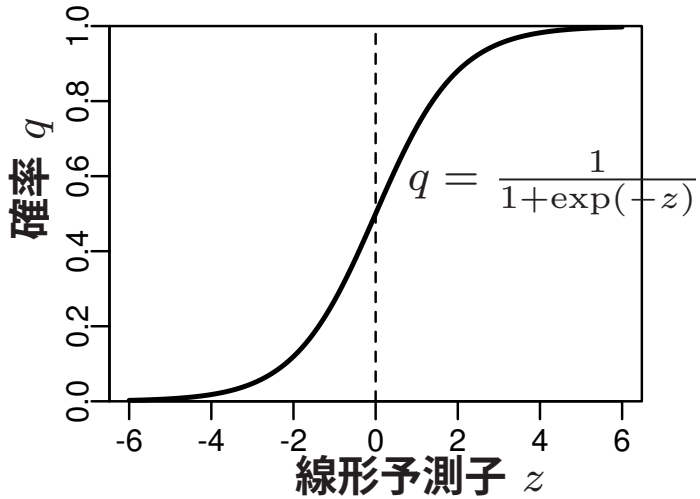


logistic curve
ロジスティック曲線とはこういうもの

ロジスティック関数の関数形 (z_i: linear predictor **線形予測子**, e.g. $z_i = \beta_1 + \beta_2 x_i$)

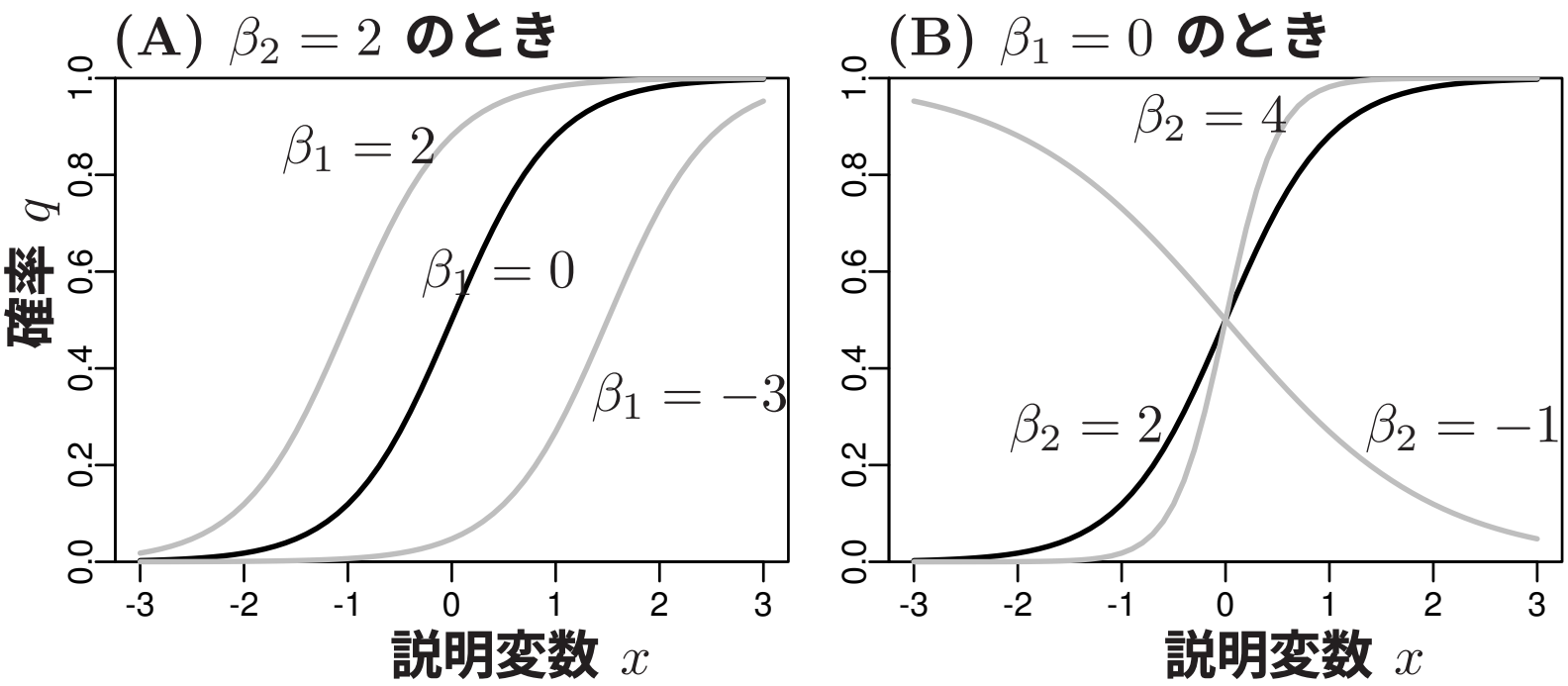
$$q_i = \mathrm{logistic}(z_i) = \frac{1}{1 + \exp(-z_i)}$$

```
> logistic <- function(z) 1 / (1 + exp(-z)) # 関数の定義
> z <- seq(-6, 6, 0.1)
> plot(z, logistic(z), type = "l")
```



β_1 and β_2 change logistic curve
パラメーターが変化すると……

黒い曲線は $\{\beta_1, \beta_2\} = \{0, 2\}$. (A) $\beta_2 = 2$ と固定して β_1 を変化させた場合.
(B) $\beta_1 = 0$ と固定して β_2 を変化させた場合.



パラメーター $\{\beta_1, \beta_2\}$ や説明変数 x がどんな値をとっても確率 q は $0 \leq q \leq 1$ となる便利な関数

logit link function

- logistic **関数**

$$q = \frac{1}{1 + \exp(-(\beta_1 + \beta_2 x))} = \text{logistic}(\beta_1 + \beta_2 x)$$

- logit **変換**

$$\text{logit}(q) = \log \frac{q}{1 - q} = \beta_1 + \beta_2 x$$

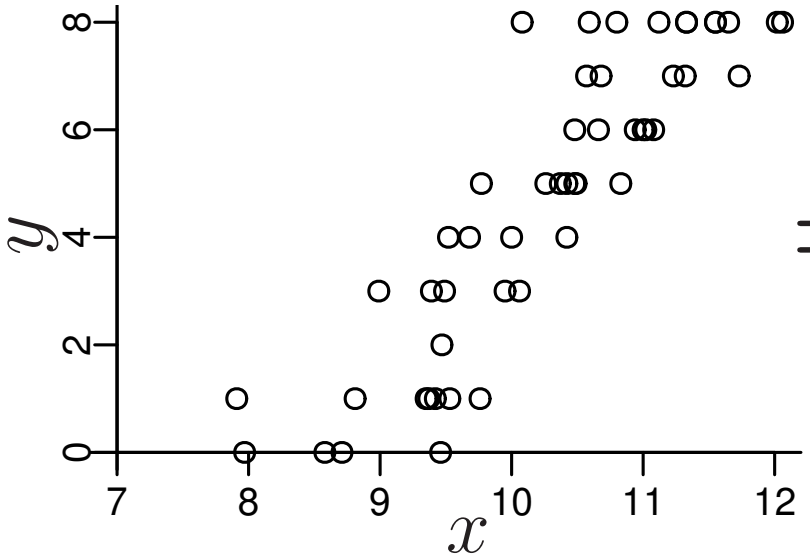
logit **は** logistic **の逆関数，** logistic **は** logit **の逆関数**

logit is the inverse function of logistic function, vice versa

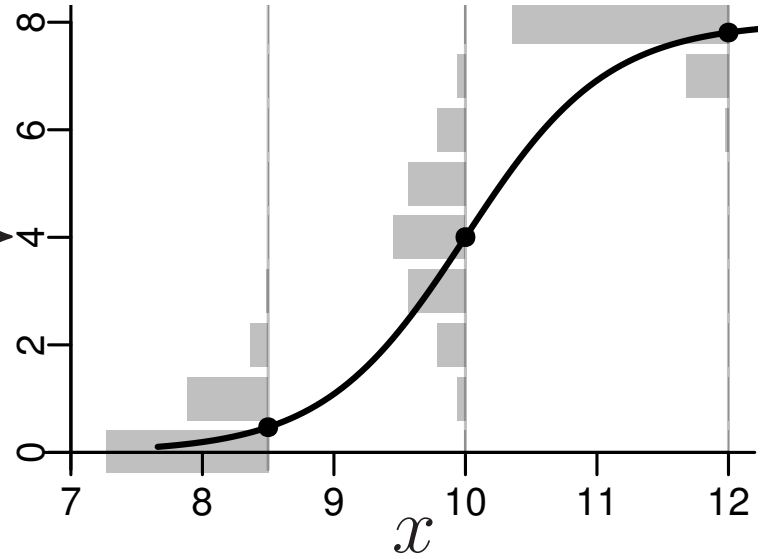
logistic regression MLE for β_1 and β_2

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）

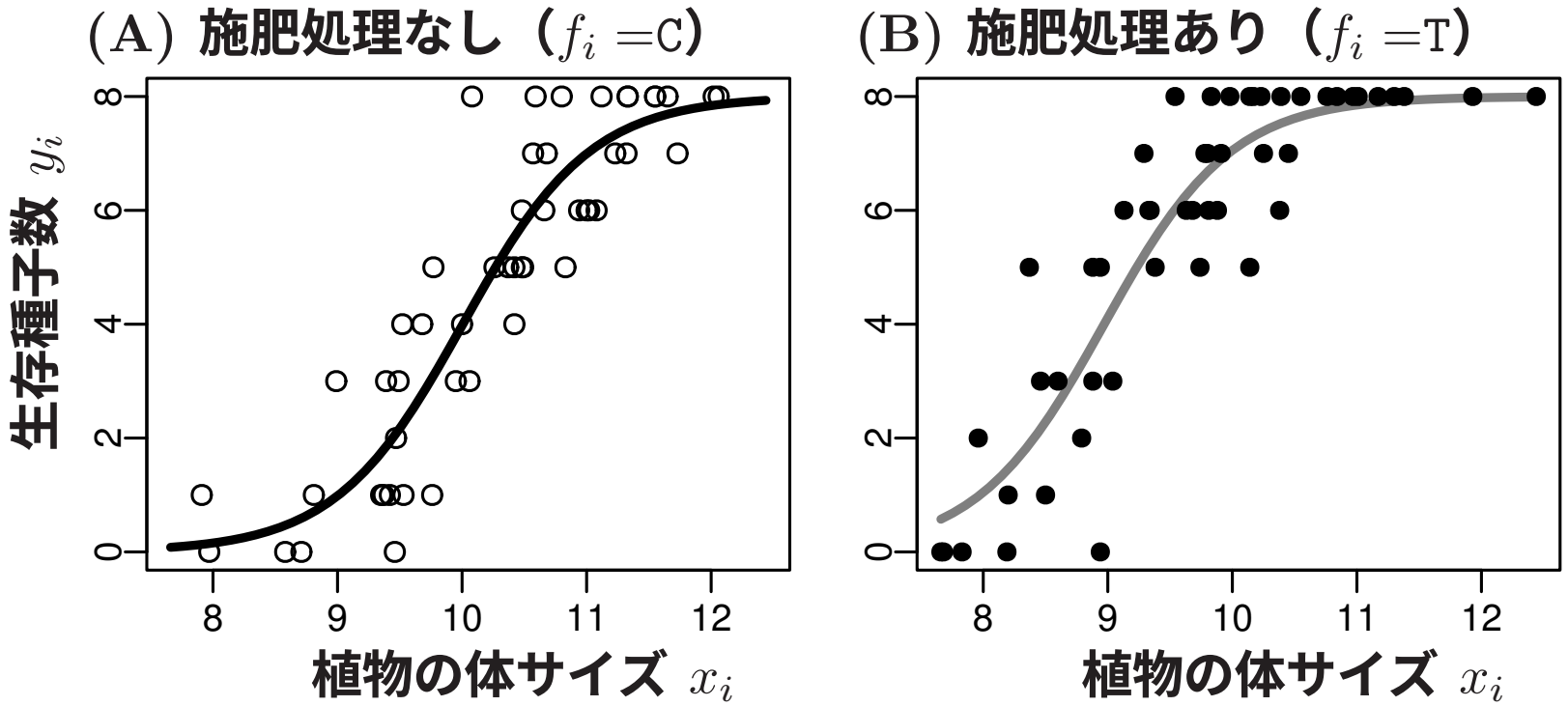


(B) 推定されるモデル




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

統計モデルの予測: 施肥処理によって応答が違う



interaction term

3. ちょっとだけ交互作用項 について

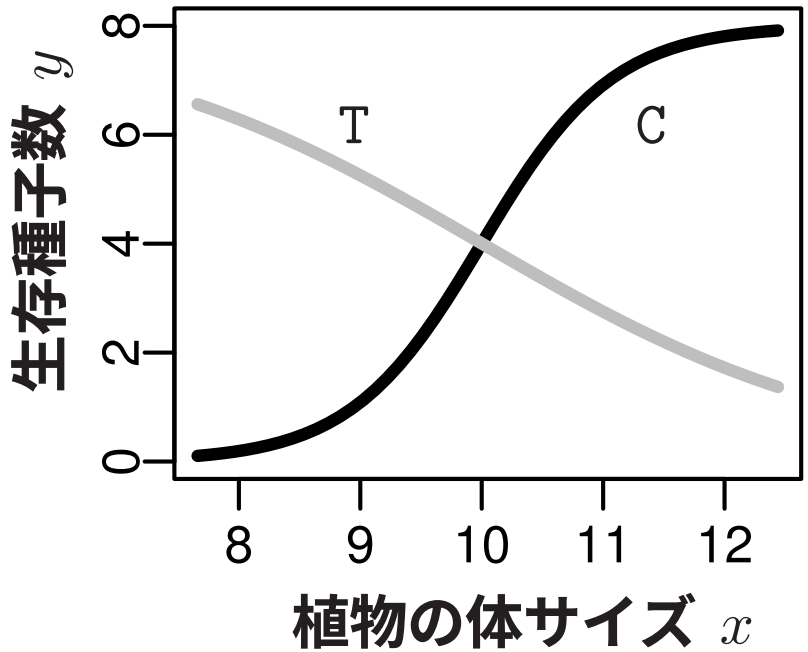
complicate terms in linear predictor

線形予測子の中の複雑な項

交互作用項とは何か?

$$\mathrm{logit}(q) = \log \frac{q}{1-q} = \beta_1 + \beta_2 x + \beta_3 f + \beta_4 x f$$

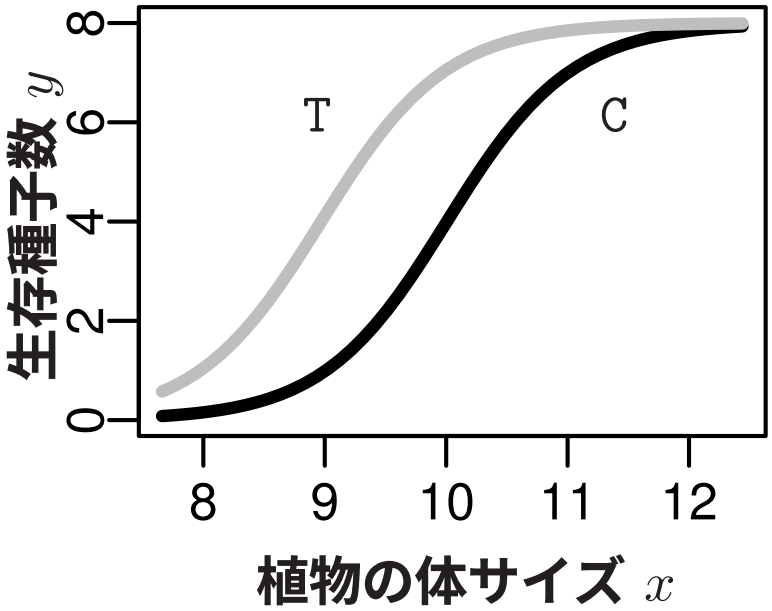
... in case that $\beta_4 < 0$, sometimes it predicts ...



in today's example
この例題データの場合，交互作用はない
no interaction effect

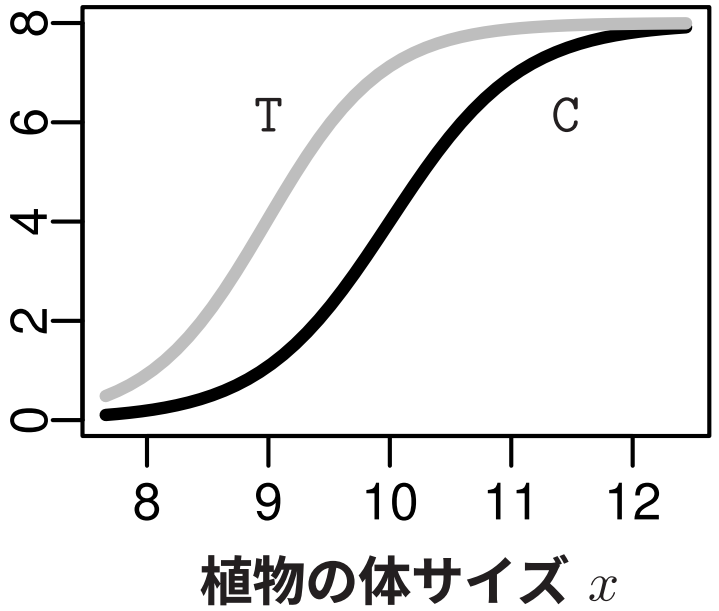
`glm(y ~ x + f, ...)`

(A) 交互作用のないモデル



`glm(y ~ x + f + x:f, ...)`

(B) 交互作用のあるモデル



little difference
差がほとんどない

NO $\frac{\text{data}}{\text{data}}$ statistics!

4. 何でも「割算」するな!

use GLM with offset term

「脱」割算の offset 項わざ

an enhanced Poisson GLM

ポアソン回帰を強めてみる

割算値ひねくるデータ解析はなぜよくないのか?

- **観測値 / 観測値** がどんな確率分布にしたがうのか見とおしが悪く， さらに説明要因との対応づけが難しくなる
- **情報が失われる**: 「10 打数 3 安打」 と「200 打数 60 安打」，「どちらも 3 割バッター」と言ってよいのか?
- 割算値を使わないほうが見とおしのよい， **合理的なデータ解析**ができる (今回の授業の主題)
- したがって割算値を使ったデータ解析は不利な点ばかり， そんなことをする必要性はどこにもない

How to avoid data/data?
避けられるわりざん

avoidable data/data values
- **避けられる割算値**
  - probability
    確　率

    例: N **個のうち** k **個にある事象が発生する確率**

    use statistical model with binomial distribution
    対策: **ロジスティック回帰など二項分布モデルで**

  - indices such as densities
    密度などの指数

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

    use offset term!　　described later
    対策: offset **項わざ** — **このあと解説!**

unfortunately, sometimes fractions appear ...

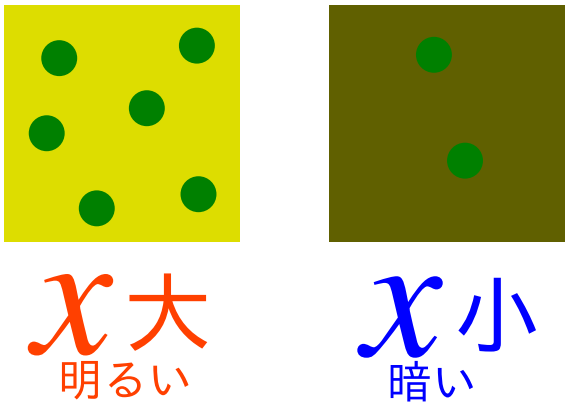
避 け に く い わ り ざ ん

hard to avoid ...

- **避けにくい割算値**
  - outputs from some measuring machines
    測定機器が内部で割算した値を出力する場合
  - sometimes we have no choice but plot data/data values ...
    割算値で作図せざるをえない場合があるかも

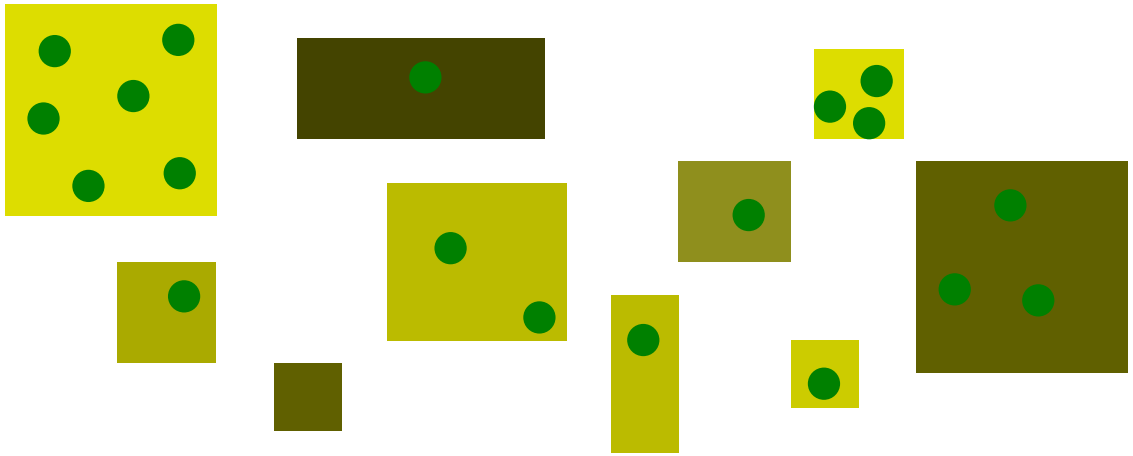
example population densities in research plots
offset 項の例題: 調査区画内の個体密度

- 何か架空の植物個体の密度が「明るさ」(x: light intensity index) に応じて どう変わるかを知りたい
- 明るさ (light index) は $\{0.1, 0.2, \cdots, 1.0\}$ の 10 段階で観測した



これだけなら単純に `glm(..., family = poisson)` とすればよいのだが ……

What? Differences in plot size?!
「場所によって調査区の面積を変えました」?!



- 明るさ x と面積 A を同時に考慮する必要あり
- ただし「密度 $=$ 個体数 $/$ 面積」といった割算値解析はやらない!
- `glm()` の `offset` 項わざでうまく対処できる
- ともあれその前に観測データを図にしてみる

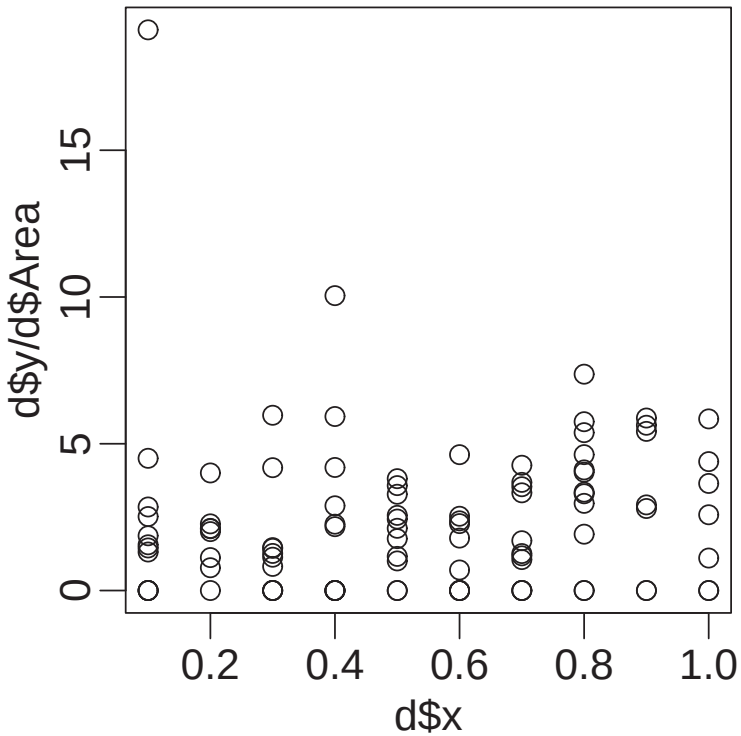
R の data.frame: 面積 Area, 明るさ x, 個体数 y

light index　number of plants

```
> load("d2.RData")
> head(d, 8) # 先頭 8 行の表示
      Area   x y
1 0.017249 0.5 0
2 1.217732 0.3 1
3 0.208422 0.4 0
4 2.256265 0.1 0
5 0.794061 0.7 1
6 0.396763 0.1 1
7 1.428059 0.6 1
8 0.791420 0.3 1
```

明るさ vs 割算値図の図

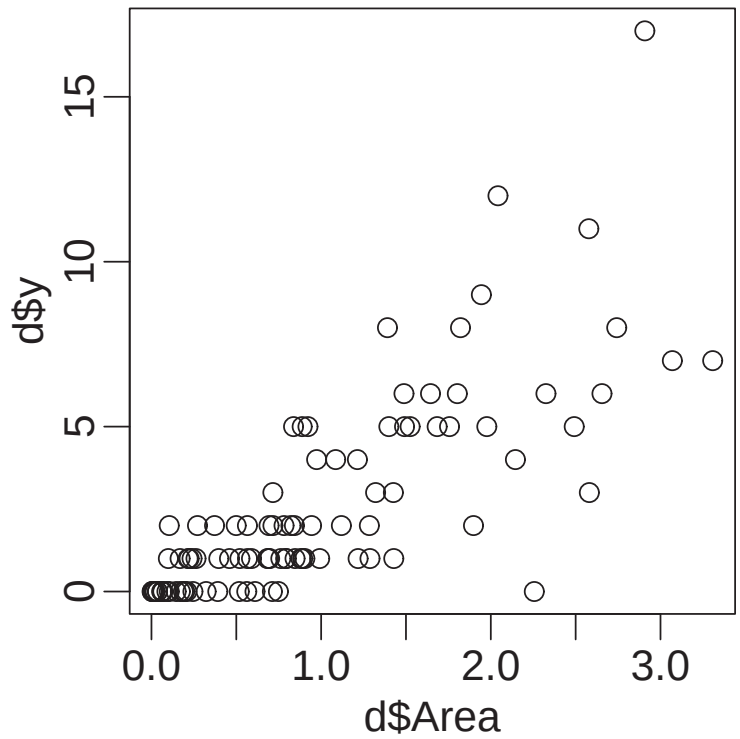
```
> plot(d$x, d$y / d$Area)
```



いまいちよくわからない ……

面積 A vs 個体数 y の図

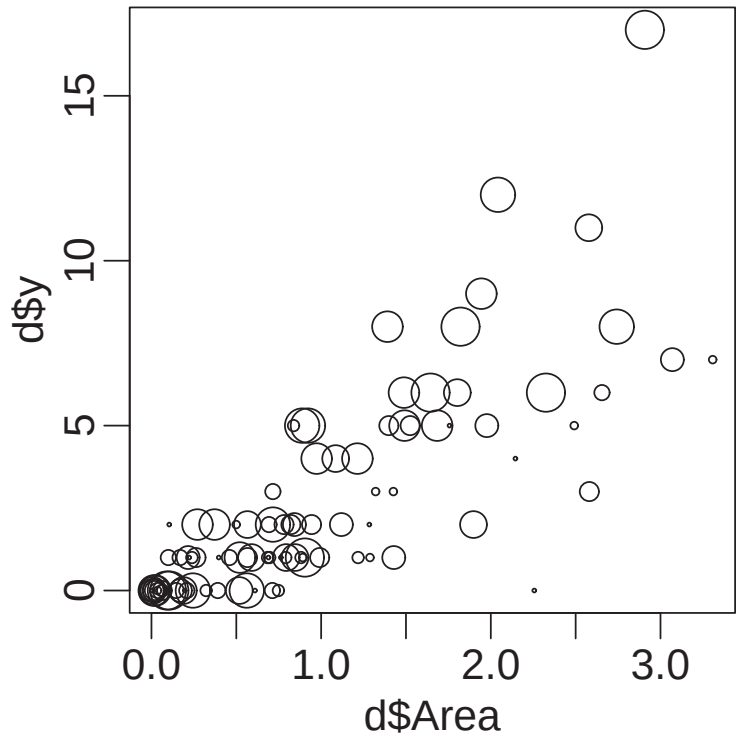
```
> plot(d$Area, d$y)
```



面積 A とともに区画内の個体数 y が増大するようだ

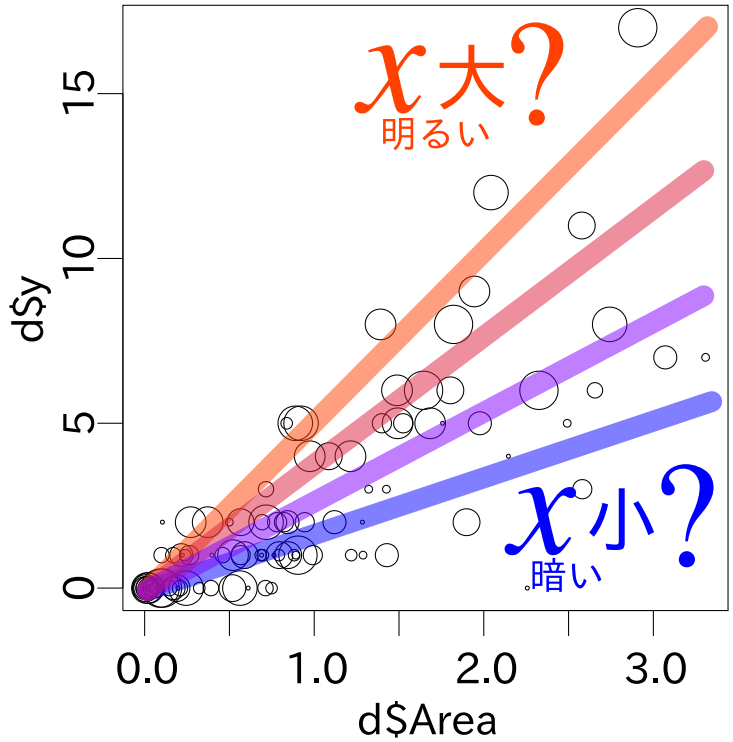
明るさ x の情報（マルの大きさ）も図に追加

```
> plot(d$Area, d$y, cex = d$x * 2)
```



同じ面積でも明るいほど個体数が多い?

密度が明るさ x に依存する統計モデル



- 区画内の個体数 y の**平均**は面積 $\times$ 密度
- 密度は明るさ x で変化する

「平均個体数 = 面積 × 密度」モデル

1. **ある区画 i の応答変数 y_i は平均 λ_i のポアソン分布にしたがうと仮定:**

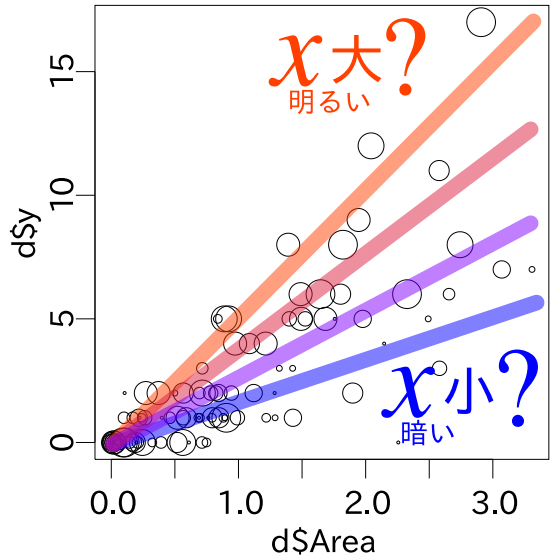
$$y_i \sim \mathrm{Pois}(\lambda_i)$$

2. **平均値 λ_i は面積 A_i に比例し，密度は明るさ x_i に依存する**

$$\lambda_i = A_i \exp(\beta_1 + \beta_2 x_i)$$

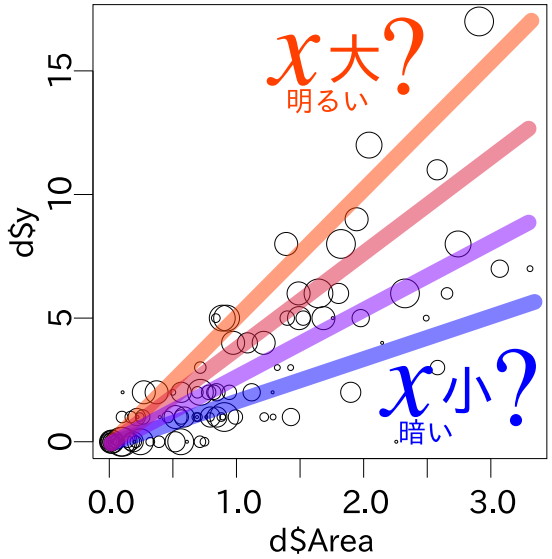
つまり $\lambda_i = \exp(\beta_1 + \beta_2 x_i + \log(A_i))$ **となるので**
$\log(\lambda_i) = \beta_1 + \beta_2 x_i + \log(A_i)$ **線形予測子は右辺のようになる**

このとき $\log(A_i)$ **を** offset **項とよぶ (係数** β **がない)**



この問題は GLM であつかえる!

- family: poisson, **ポアソン分布**
- link **関数**: "log"
- **モデル式** : y ~ x
- offset **項の指定** : log(Area)

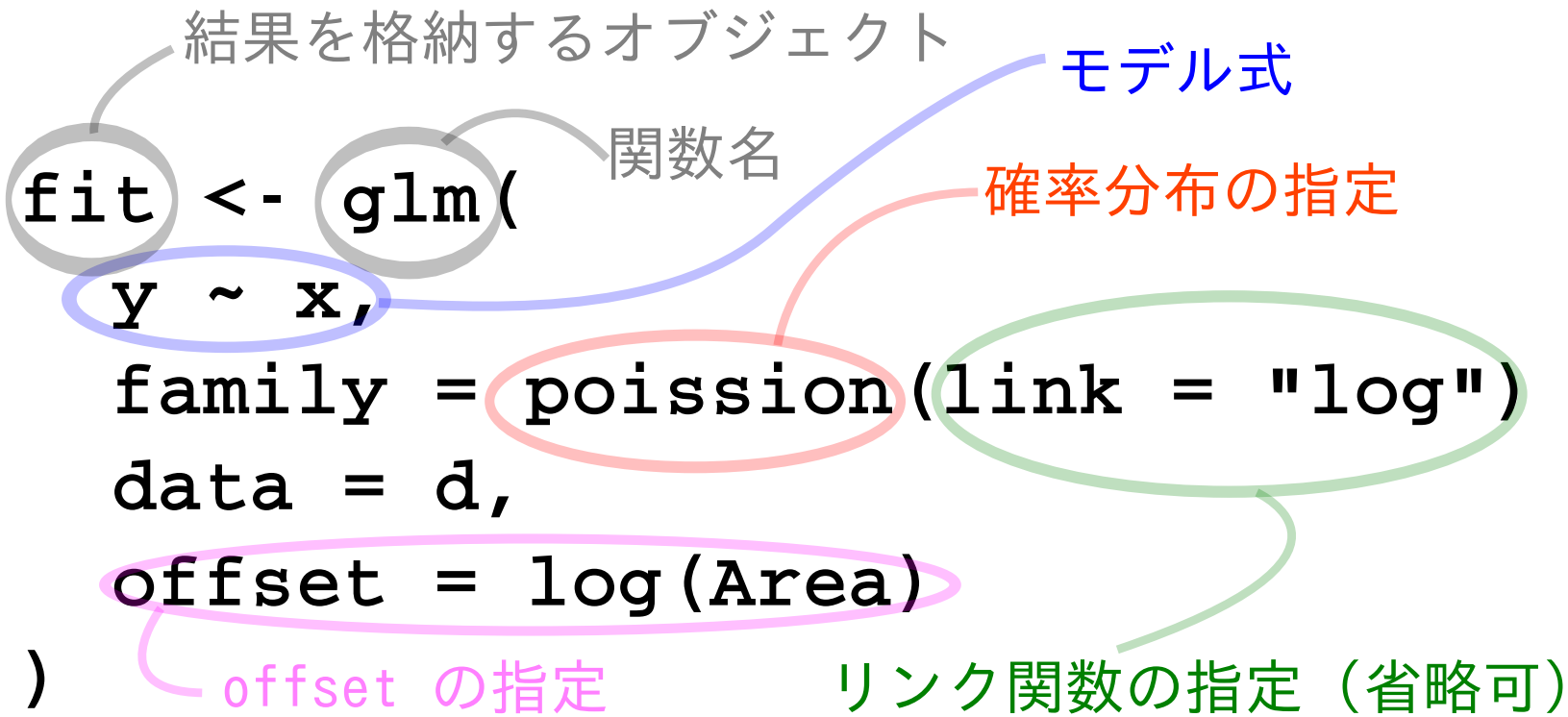


- **線形予測子** $z = \beta_1 + \beta_2\ \mathrm{x} +\ \mathrm{log(Area)}$

  a, b **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 $\lambda = \exp(z) = \exp(\beta_1 + \beta_2\ \mathrm{x} +\ \mathrm{log(Area)})$
- **応答変数 は平均** λ **のポアソン分布に従う**:

glm() 関数の指定


結果を格納するオブジェクト
関数名
モデル式
確率分布の指定
fit <- glm(
y ~ x,
family = poission(link = "log")
data = d,
offset = log(Area)
)
offset の指定
リンク関数の指定（省略可）

Rの glm() 関数による推定結果

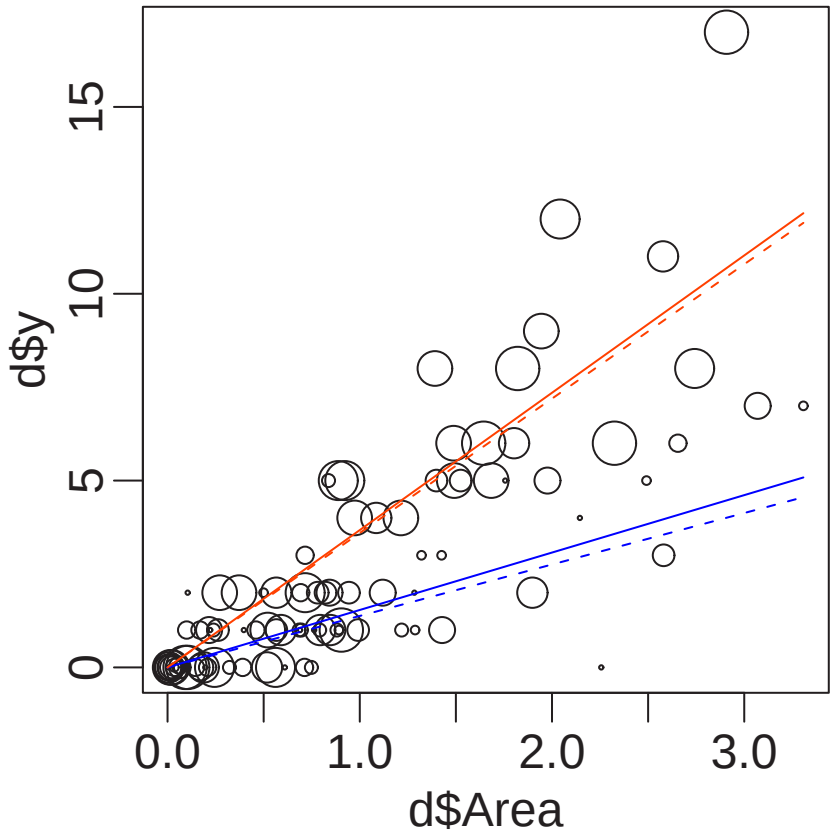
```
> fit <- glm(y ~ x, family = poisson(link = "log"), data = d,
  offset = log(Area))
> print(summary(fit))

Call:
glm(formula = y ~ x, family = poisson(link = "log"), data = d,
    offset = log(Area))

(... 略...)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.321      0.160    2.01    0.044
x              1.090      0.227    4.80  1.6e-06
```

Plotting the model prediction based on estimation
推定結果にもとづく予測を図にしてみる



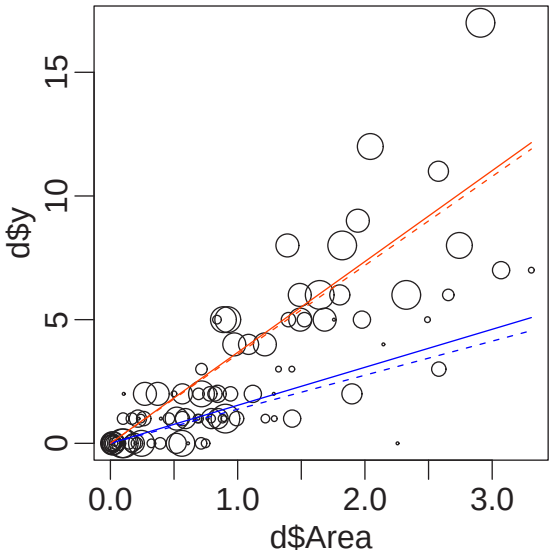
$x = 0.9$
light environment

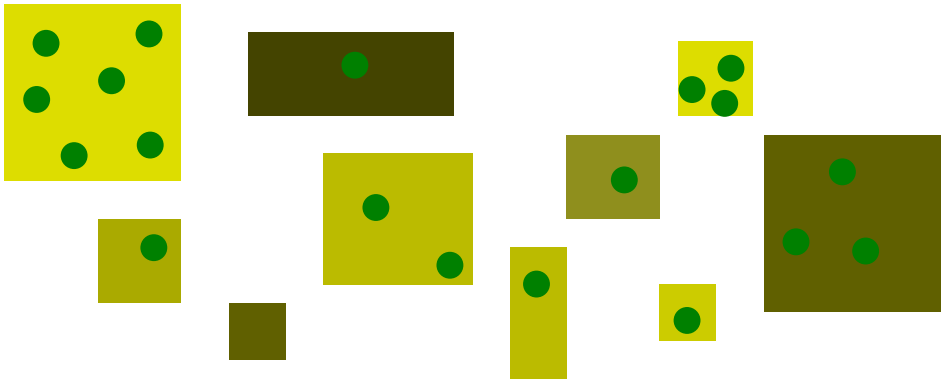
$x = 0.1$
dark environment

- solid lines / 実線は glm() の推定結果にもとづく prediction / 予測
- dotted lines / 破線はデータ生成時に指定した関係 “true” model

まとめ: glm() の offset 項わざで「脱」割算

- 平均値が面積などに比例する場合は，この面積などを offset 項 として指定する
- 平均 = 面積 × 密度，というモデルの密度 を exp(線形予測子) として定式化する





Improve your statisitcal model and remove data/data values!
統計モデルを工夫してわりざんやめよう

avoidable data/data values
- **避けられる割算値**
  probability
  - **確　率**

    例: N **個のうち** k **個にある事象が発生する確率**

    use statistical model with binomial distribution
    対策: **ロジスティック回帰など二項分布モデルで**

  indices such as densities
  - **密度などの指数**

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

    use offset term!　　Improve your statistical model!
    対策: offset **項わざ** — **統計モデリングの工夫**!

GLM is not enough!

5. GLM では説明できない種子データ

overdispersion data

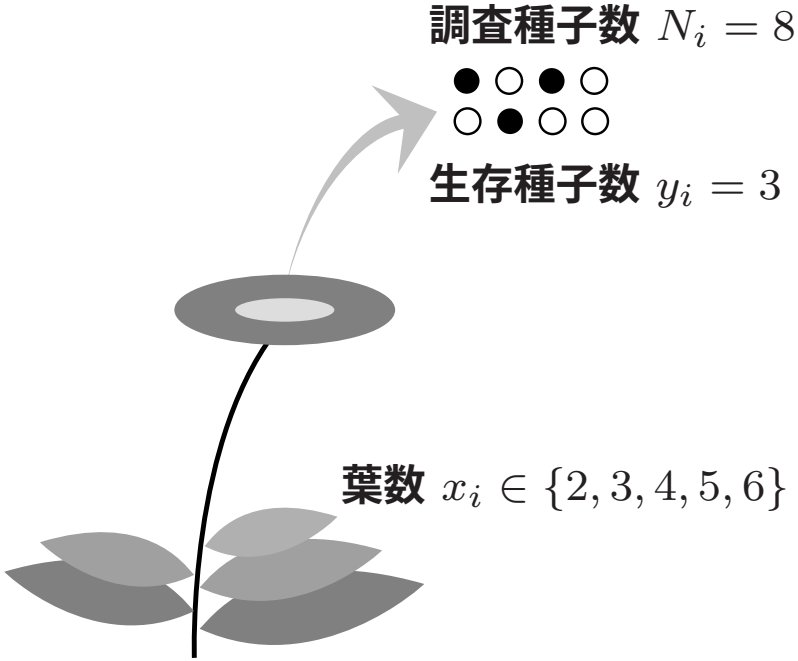
「ばらつき」が大きすぎる!

データにひそむ過分散

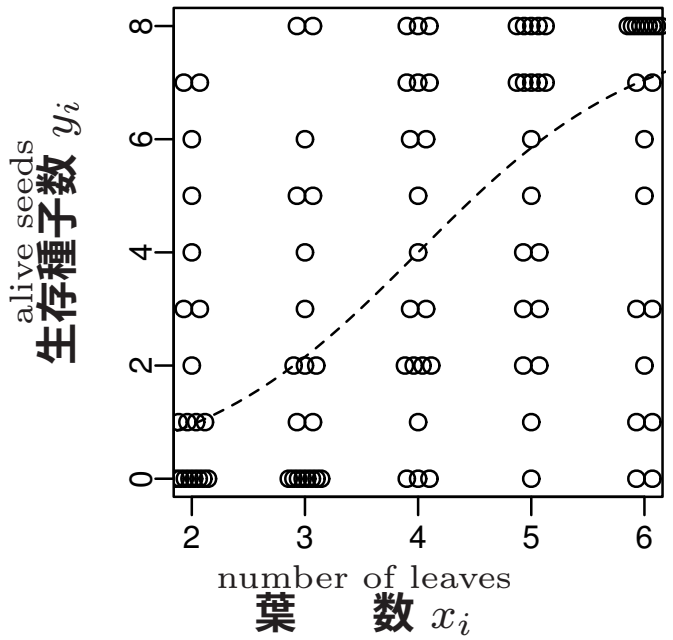
another example　　seed survivorship again, but ...

GLMM の例題: 種子の生存確率……また同じ?!

(A) 個体 i で観測されたデータ



(B) 全 100 個体の x_i と y_i

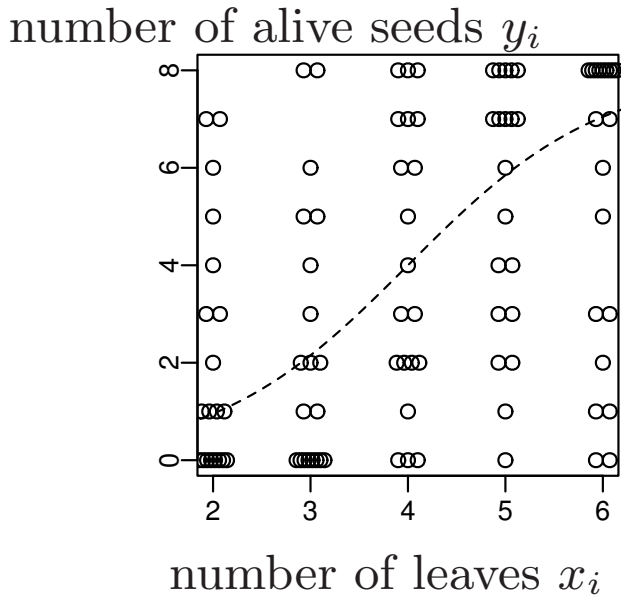


logistic regression as usual?

“N 個中の y 個” 型のデータ→ロジスティック回帰?

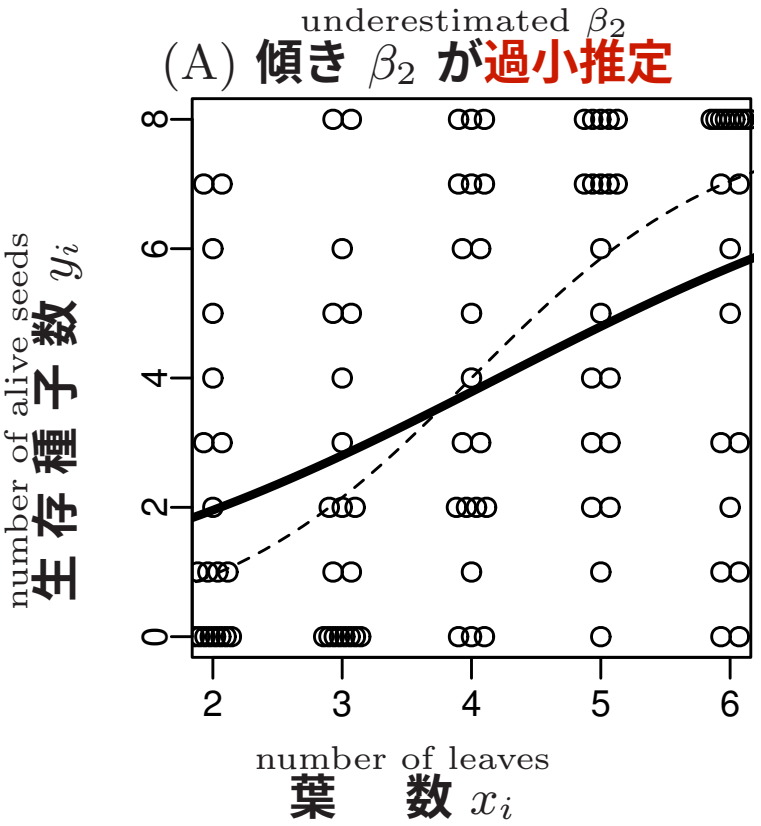
ロジスティック回帰のモデル

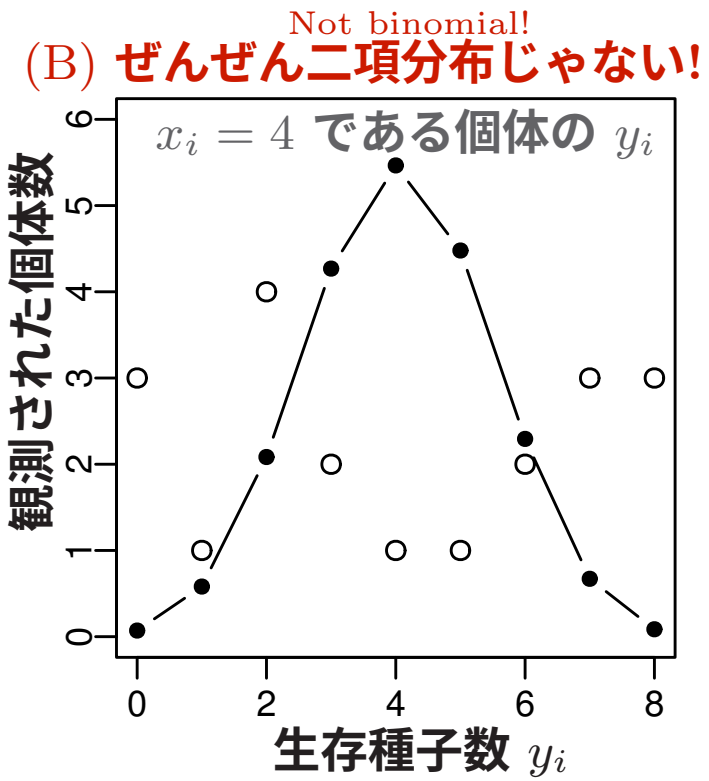
- probability distribution 確率分布: binomial distribution 二項分布
- linear predictor 線形予測子: $\beta_1 + \beta_2 x_i$
- link function リンク関数: logit リンク関数



GLM doesn't work!

GLM では説明できないばらつき!





○が観測されたデータの図示

overdispersion caused by individual differences

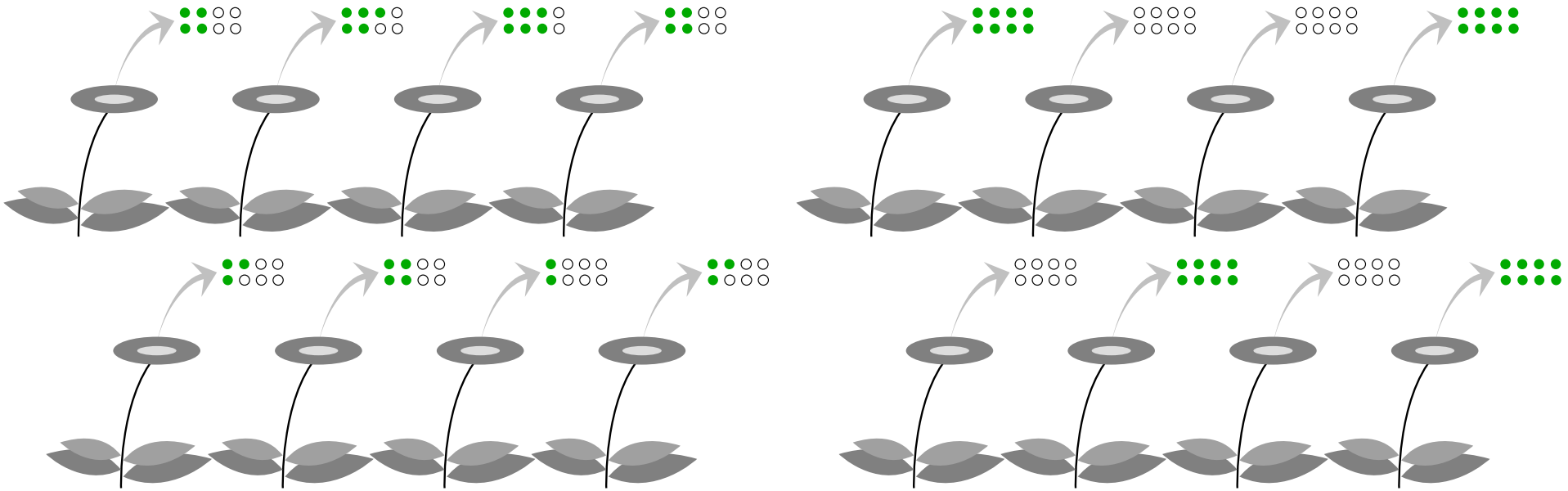
6. 過 分 散 と 個 体 差

観測されていない個体差がもたらす過分散

unobservable differences

観測されてない個体差って?

過分散 (overdispersion) とは何か?

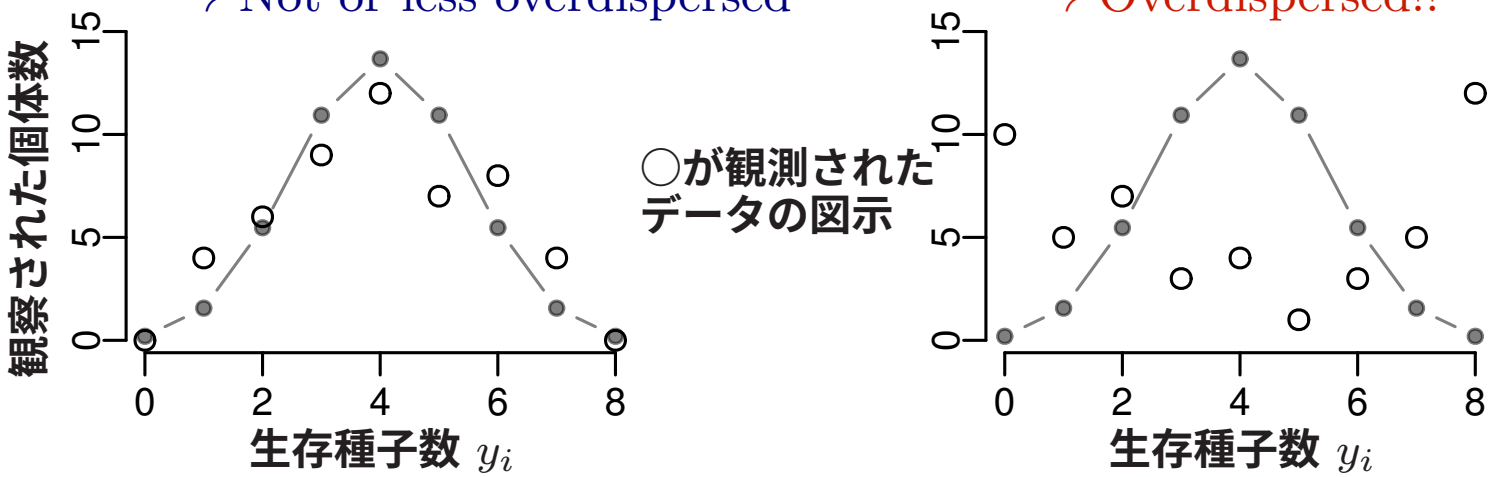


(A) 個体差のばらつきが小さい場合

→ Not or less overdispersed

(B) 個体差のばらつきが大きい場合

→ Overdispersed!!



ロジスティック回帰やポアソン回帰といったGLMでは全サンプルの均質性を仮定している

GLM does not take into account individual differences

現実のカウントデータはほとんど過分散

Almost all "real" data are overdispersed!

Genralized Liear Mixed Model

7. 一般化線形混合モデル

parameters for indviduals

個体差をあらわすパラメーターを追加

fixed effects　　random effects

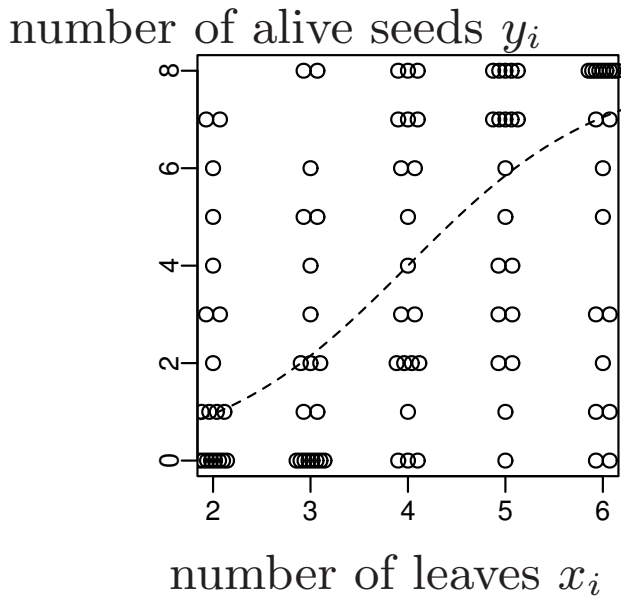
固定効果 と ランダム効果

an improvement of logistic regression model

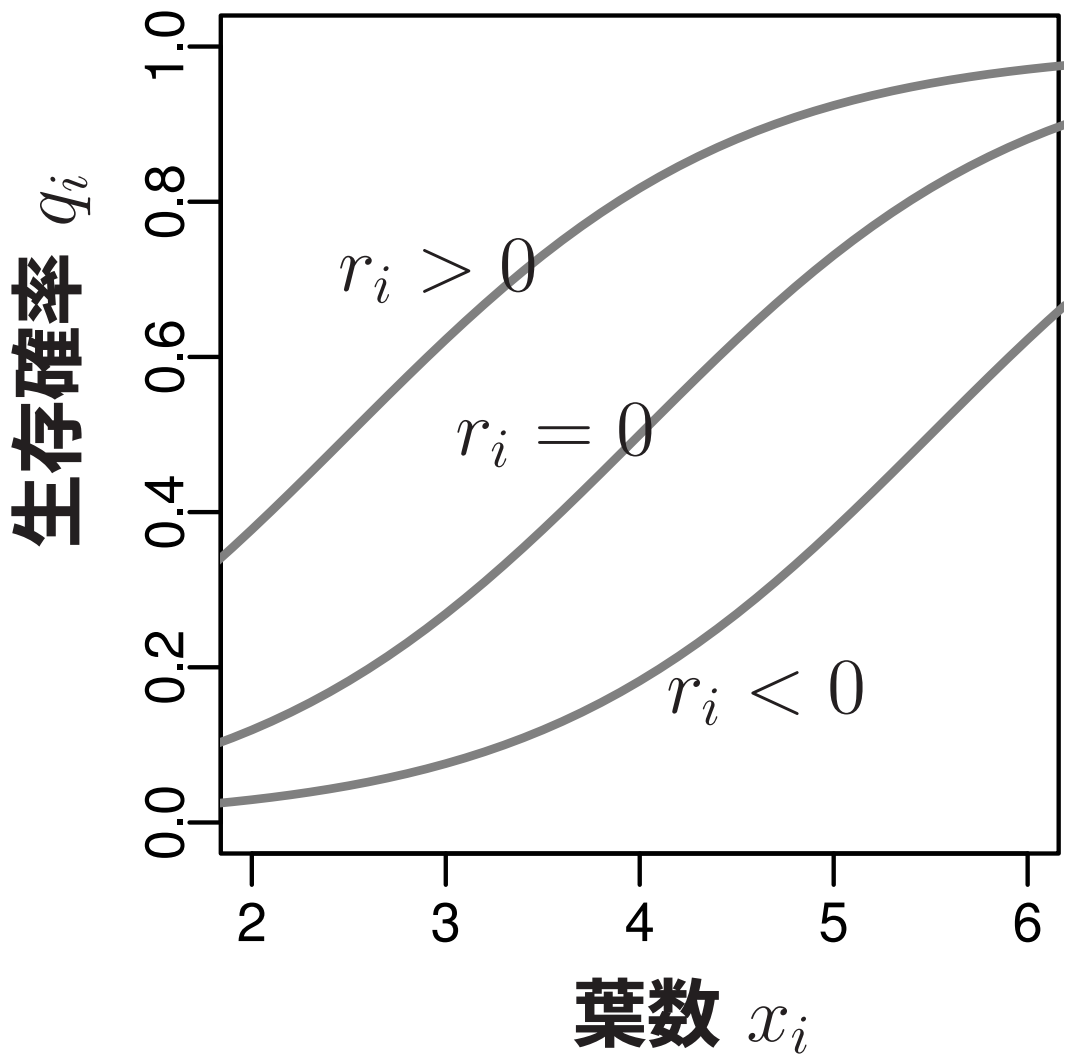
ロジスティック回帰のモデルを改良する

ロジスティック回帰のモデル

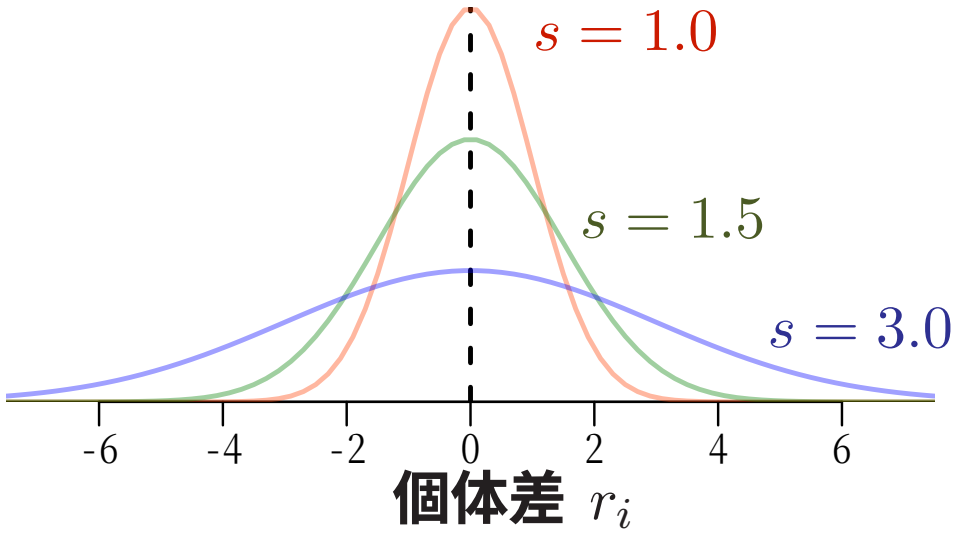
- 確率分布 (probability distribution): 二項分布 (binomial distribution)
- 線形予測子 (linear predictor): $\beta_1 + \beta_2 x_i + r_i$
- リンク関数 (link function): logit リンク関数



個体 i の個体差を r_i としてみよう



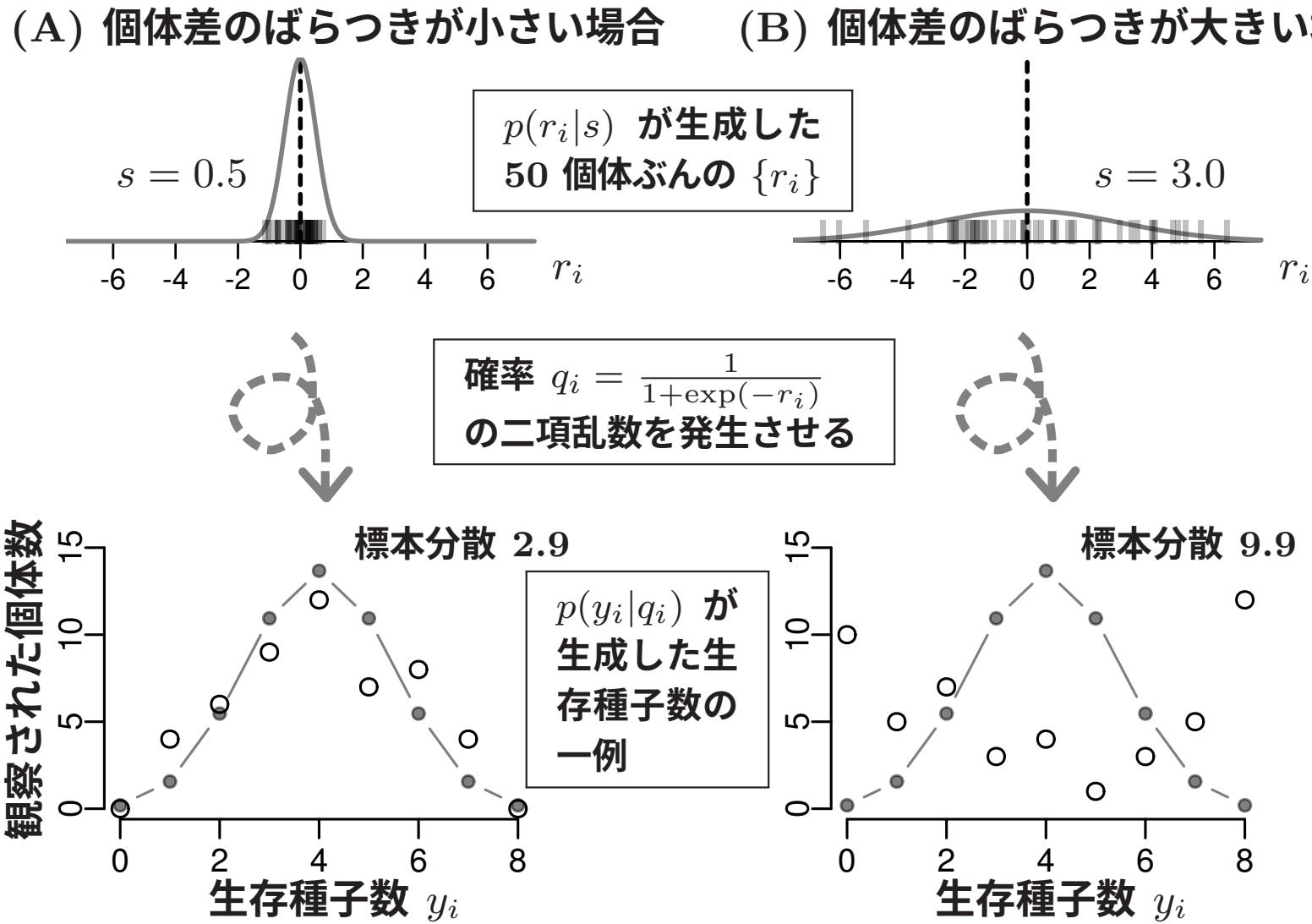
suppose $\{r_i\}$ follow the Gausssian distribution
$\{r_i\}$ のばらつきは正規分布だと考えてみる



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

この確率密度 $p(r_i|s)$ は r_i の「出現しやすさ」をあらわしていると解釈すればよいでしょう．r_i がゼロにちかい個体はわりと「ありがち」で，r_i の絶対値が大きな個体は相対的に「あまりいない」．

個体差 r_i の分布と過分散の関係



a numerical experiment using random numbers

ちょっと乱数を使った数値実験をしてみましょう

```
> # defining logistic function
> logistic <- function(z) { 1 / (1 + exp(-z)) }
> # random numbers following binomial distribution
> rbinom(100, 8, prob = logistic(0))
> # random numbers following Gausssian distribution
> rnorm(100, mu = 0, sd = 0.5)
> r <- rnorm(100, mu = 0, sd = 0.5)
> # random numbers following ... ?
> rbinom(100, 8, prob = logistic(0 + r))
```

fixed effects random effects
固定効果 と ランダム効果

Generalized Linear Mixed Model (GLMM)

linear predictor

で使う Mixed **な 線形予測子**: $\beta_1 + \beta_2 x_i + r_i$

- fixed effects: $\beta_1 + \beta_2 x_i$
- random effects: $+r_i$

fixed? random? **よくわからん……?**

どうでもいい用語説明

伝統的な訳語としては ……

fixed effects — **母数効果**

random effects — **変量効果**

なんだかよくわかりませんね

混合モデル補足

データのばらつきが正規分布である混合モデルは

線形混合モデル (linear mixed model, LMM)

random effects **は「独立とみなせないデータ」**

の「ずれ」をあらわす —

- **個体差の例**: **同じ個体から複数のデータをとっている，など**
- **グループ差の例**: **市内の小学校で共通テストをやったときの「学校差」**

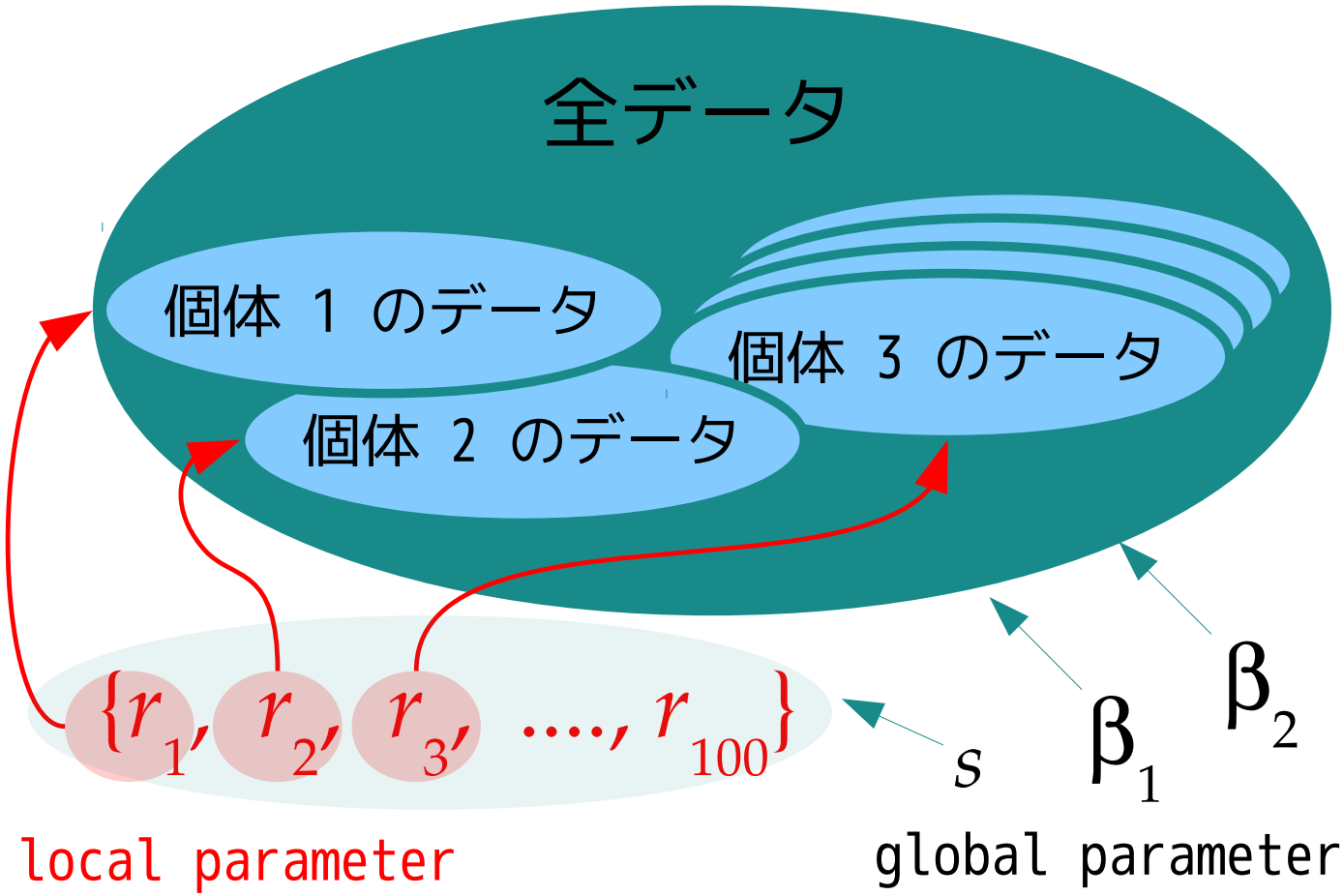
global parameter, local parameter と分類してみる?

Generalized Linear Mixed Model (GLMM)
で使う線形予測子: $\beta_1 + \beta_2 x_i + r_i$

- fixed effects: $\beta_1 + \beta_2 x_i$
  - global parameter — **全個体を説明**
- **全個体のばらつき s も** global parameter
- random effects: $+r_i$
  - local parameter — **個体 i に関する説明**

(**注**) global/local parameter **は久保の造語**

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?

Maximum likelihood estimation for GLMM

8. 一般化線形混合モデルの最尤推定

個体差 r_i を積分して消す尤度方程式

「積分する」とは分布を混ぜること

個体差 r_i は最尤推定できない

$$\text{local parameters: } \{r_1, r_2, \cdots, r_{100}\}$$

全 100 個体に対して，個体ごとにいちいち r_i の値を最尤推定する
と飽和モデル (saturation model) の推定になってしまう

```
> d <- read.csv("data.csv")
> head(d)
  N y x id
1 8 0 2  1
2 8 1 2  2
3 8 2 2  3
4 8 4 2  4
5 8 1 2  5
6 8 0 2  6
```

尤度関数の中で r_i を積分してしまえばよい

データ y_i のばらつき — binomial distribution **二 項 分 布**

$$p(y_i|\beta_1,\beta_2) = \binom{8}{y_i} q_i^{y_i}(1-q_i)^{8-y_i}$$

個体差 r_i のばらつき — Gaussian distribution **正 規 分 布**

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

個体 i の likelihood **尤 度** — to remove r_i **r_i を消す**

$$L_i = \int_{-\infty}^{\infty} p(y_i|\beta_1,\beta_2,r_i)\ p(r_i|s)dr_i$$

likedhood for all data
全データの尤度 — **β_1, β_2, s の関数**

$$L(\beta_1,\beta_2,s) = \prod_i L_i$$

global parameter と local parameter

Generalized Linear Mixed Model (GLMM)

linear predictor

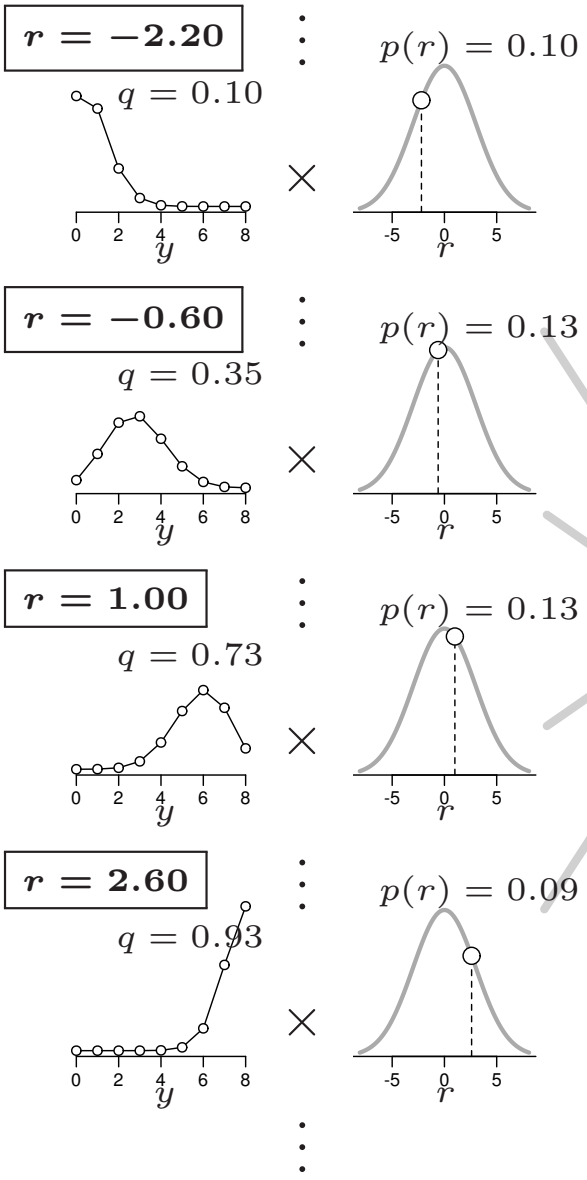
で使う Mixed な 線形予測子: $\beta_1 + \beta_2 x_i + r_i$

- global parameter は最尤推定できる
  - fixed effects: β_1, β_2
  - 全個体のばらつき: s
- local parameter は最尤推定できない
  - random effects: $\{r_1, r_2, \cdots, r_{100}\}$

個体差 r_i について積分するということは二項分布と正規分布をまぜあわせること

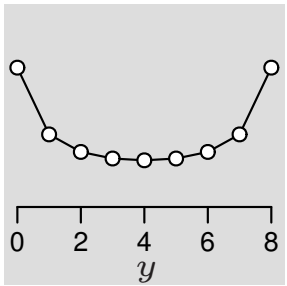
個体差 r ごとに異なる二項分布

集団内の r の分布
重み $p(r|s)$



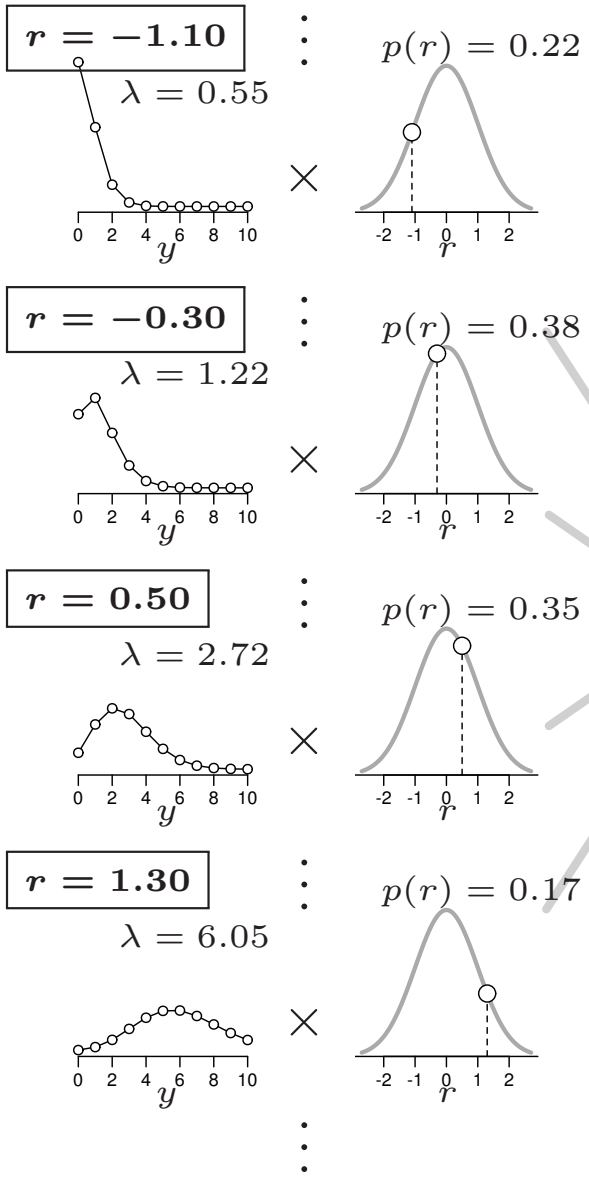
binomial and Gaussian distributions
二項分布と正規分布のまぜあわせ

集団全体をあらわす混合された分布



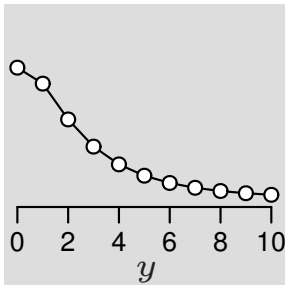
個体差 r ごとに異なる
ポアソン分布

集団内の r の分布
重み $p(r|s)$



積分

集団全体をあらわす
混合された分布



Poisson and Gaussian distributions
ポアソン分布と正規分布のまぜあわせ

glmmML package を使って GLMM の推定

```
> install.packages("glmmML") # if you don't have glmmML
> library(glmmML)
> glmmML(cbind(y, N - y) ~ x, data = d, family = binomial
+ cluster = id)

> d <- read.csv("data.csv")
> head(d)
  N y x id
1 8 0 2  1
2 8 1 2  2
3 8 2 2  3
4 8 4 2  4
5 8 1 2  5
6 8 0 2  6
```

GLMM の推定値: $\hat{\beta}_1, \hat{\beta}_2, \hat{s}$ estimates

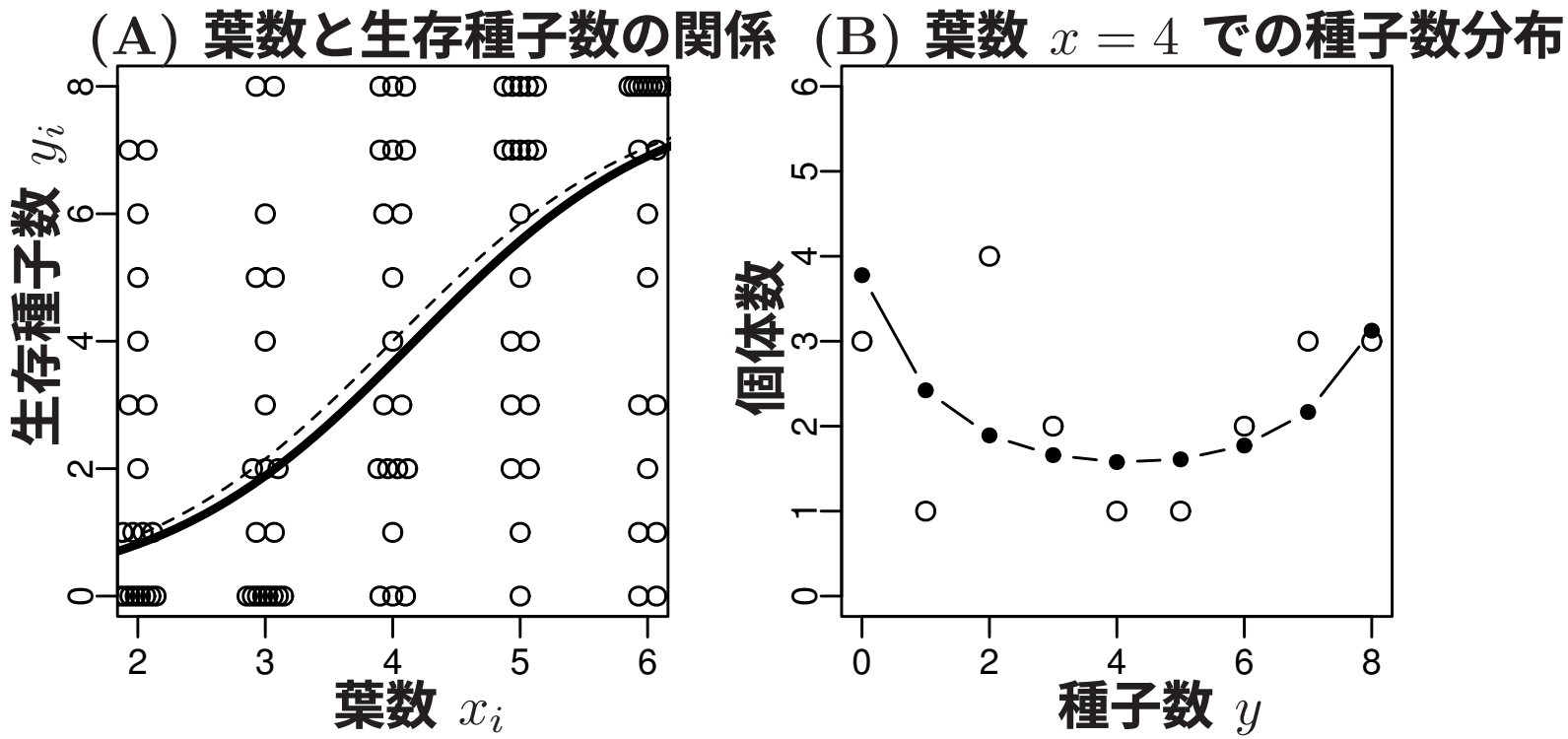
```
> glmmML(cbind(y, N - y) ~ x, data = d, family = binomial,
+ cluster = id)
...(snip)...
             coef se(coef)     z Pr(>|z|)
(Intercept) -4.13    0.906 -4.56  5.1e-06
x            0.99    0.214  4.62  3.8e-06

Scale parameter in mixing distribution:  2.49 gaussian
Std. Error:                              0.309

Residual deviance: 264 on 97 degrees of freedom  AIC: 270
```

$$\hat{\beta}_1 = -4.13, \hat{\beta}_2 = 0.99, \hat{s} = 2.49$$

推定された GLMM を使った 予 prediction 測



summary

GLMM まとめ

- 現実のデータ解析では個体差・場所差の効果を統計モデルに組みこまなければならない
- これらは歴史的には random effects とよばれてきた
- 実際のところは — 統計モデルには global parameter と local parameter があると考えればよい
- GLMM では global parameter を最尤推定する— local parameter は積分して消す
- local parameter が増えると (e.g. 個体差 + 場所差) パラメーター推定がたいへんになる — ということで ……

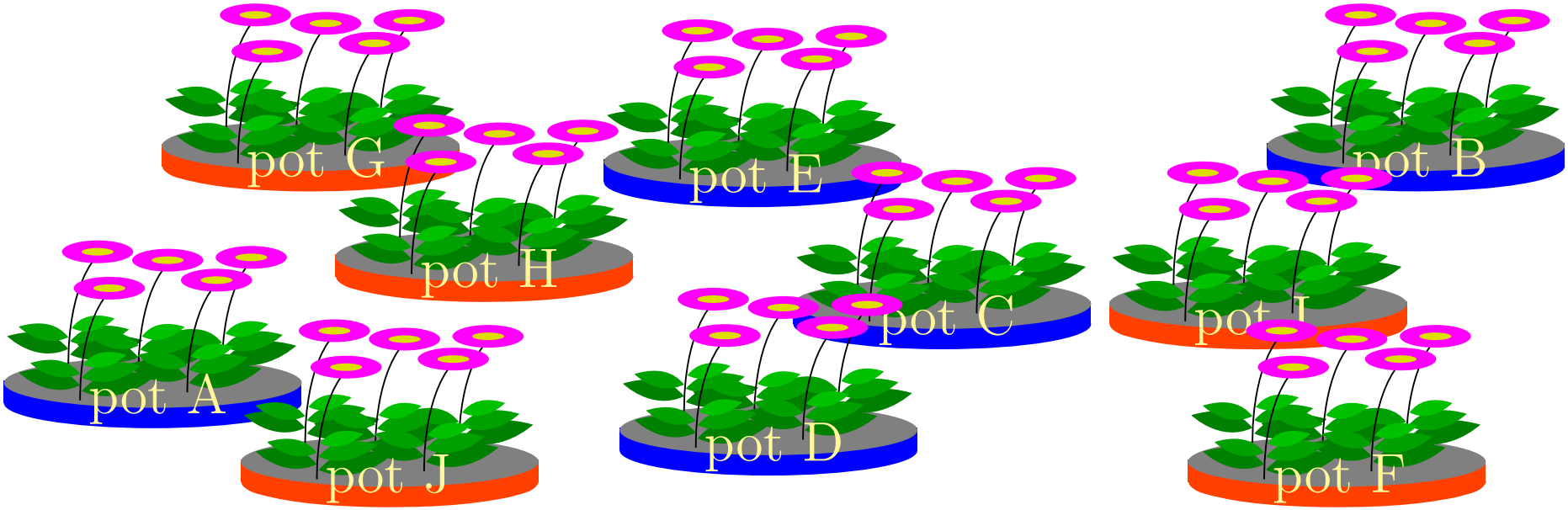
Real data are ... harder!!

9. 実際のデータ構造はもっと複雑

複数の階層，時間や空間の構造などなど

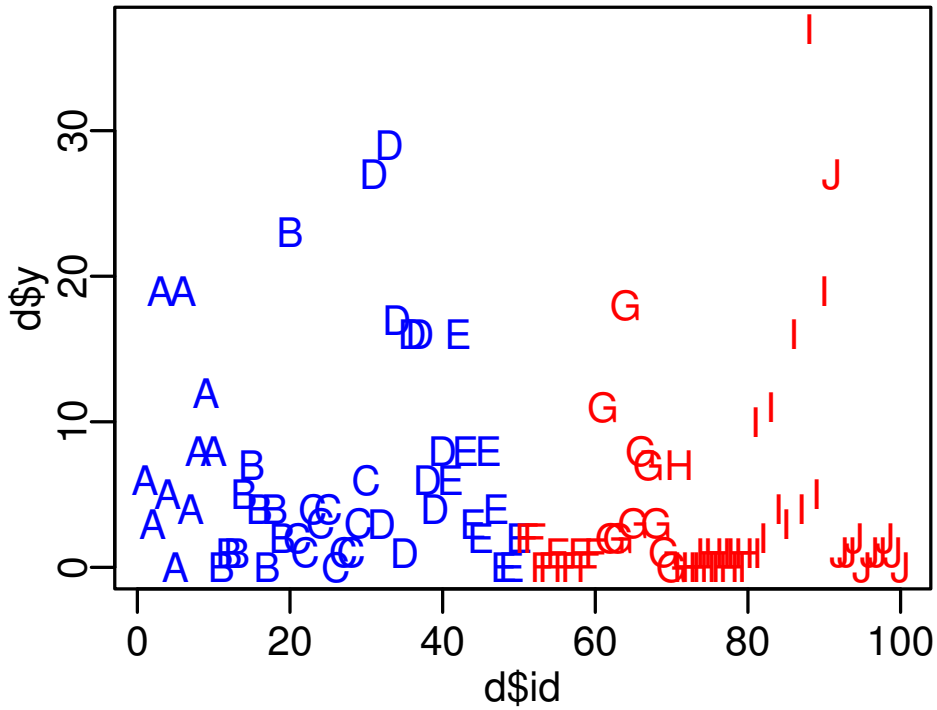
How to model the nested structure of data

架空植物の例題: またまた種子数データ



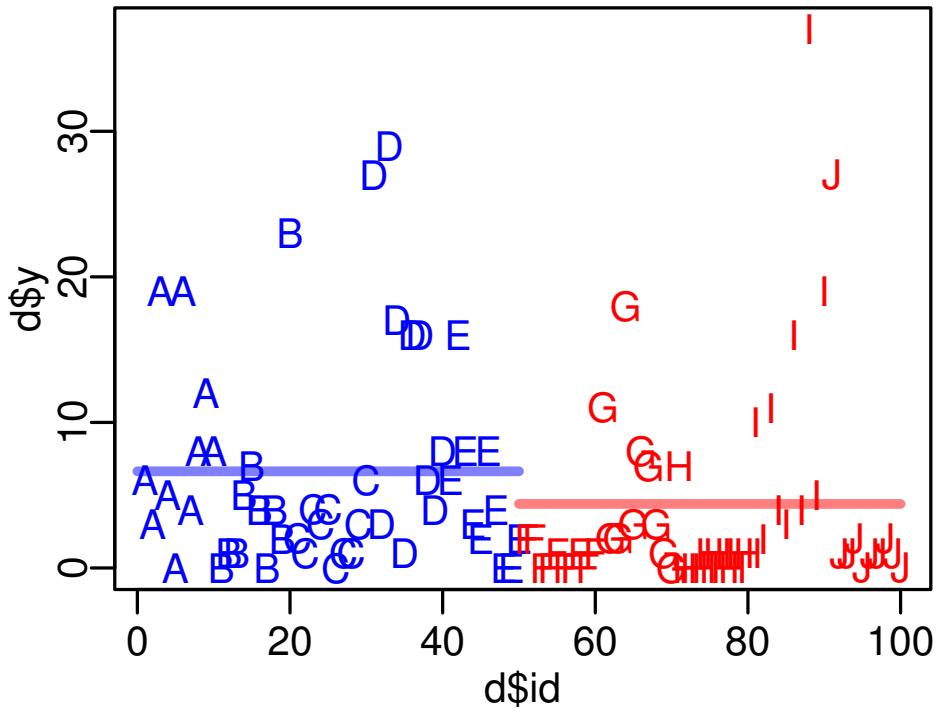
- 肥料をやったら個体ごとの種子数 y_i が増えるかどうかを調べたい
- 植木鉢 10 個，各鉢に 10 個体の架空植物 (合計 100 個体)
  - コントロール $(f_j = \mathbf{C})$ 5 鉢 (合計 50 個体)
  - 肥料をやる処理 $(f_j = \mathbf{T})$ 5 鉢 (合計 50 個体)

データはとにかく図示する!!



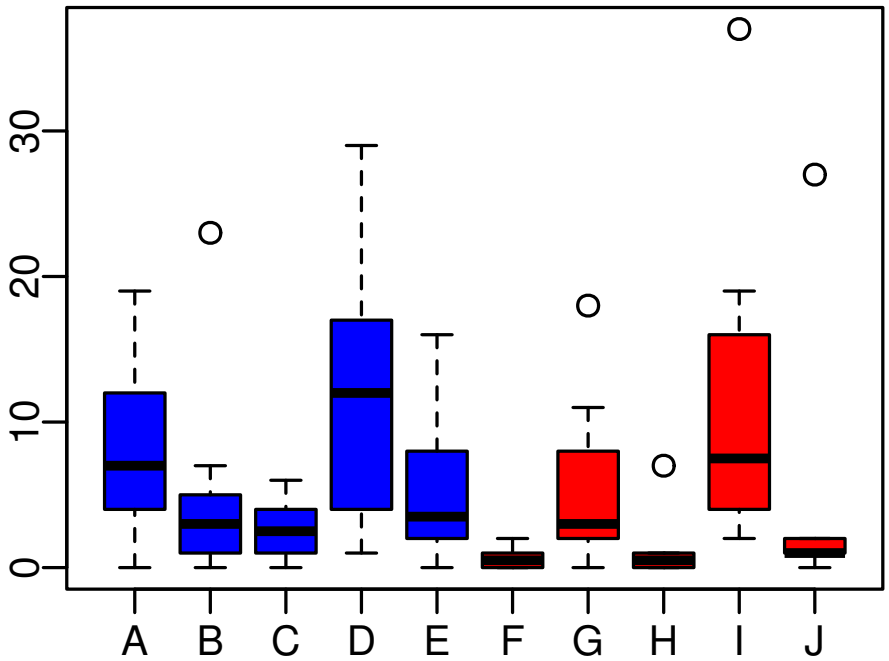
- `plot(d$id, d$y, pch = as.character(d$pot), ...)`
- **コントロール・処理 でそんなに差がない**?

処理ごとの平均も図に追加してみる



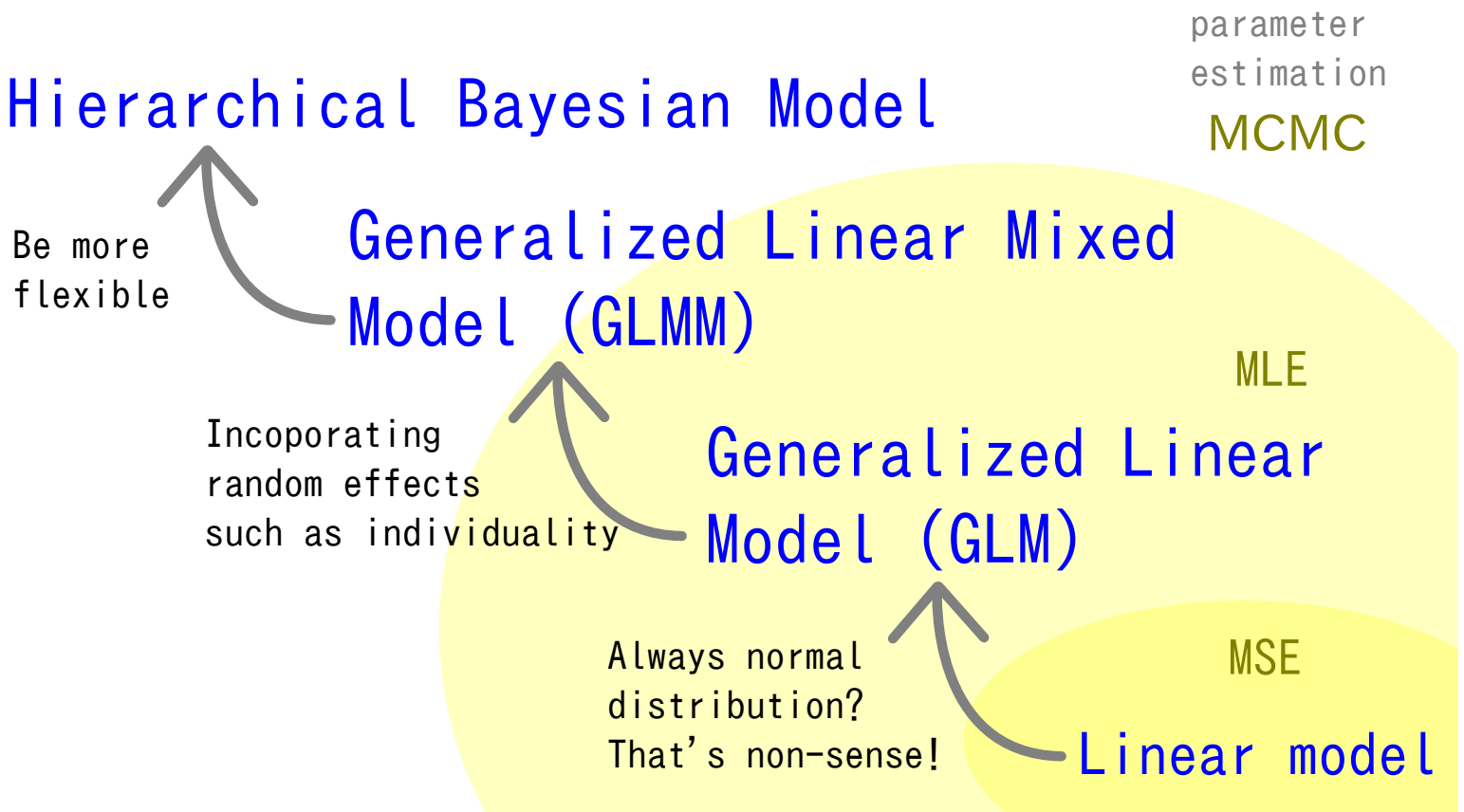
- むしろ **処理** のほうが平均種子数が低い?

個体差だけでなく植木鉢差もありそう?



- `plot(d$pot, d$y, col = rep(c("blue", "red"), each = 5))`
- **植木鉢由来の** random effects **みたいなものはブロック差と呼ばれる**

どうすればよいか? 階層ベイズモデル化!



このあとごく簡単に紹介するので，

くわしくは教科書・ネット上の資料を参照してください

GLM: 個体差もブロック差も無視

```
> summary(glm(y ~ f, data = d, family = poisson))
...(略)...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.8931     0.0549   34.49  < 2e-16
fT           -0.4115     0.0869   -4.73  2.2e-06
...(略)...
```

- 肥料をやる処理 (f) をすると，平均種子数が下がる?
- AIC でモデル選択しても同じような結果に

GLMM: 個体差だけ考慮，ブロック差は無視

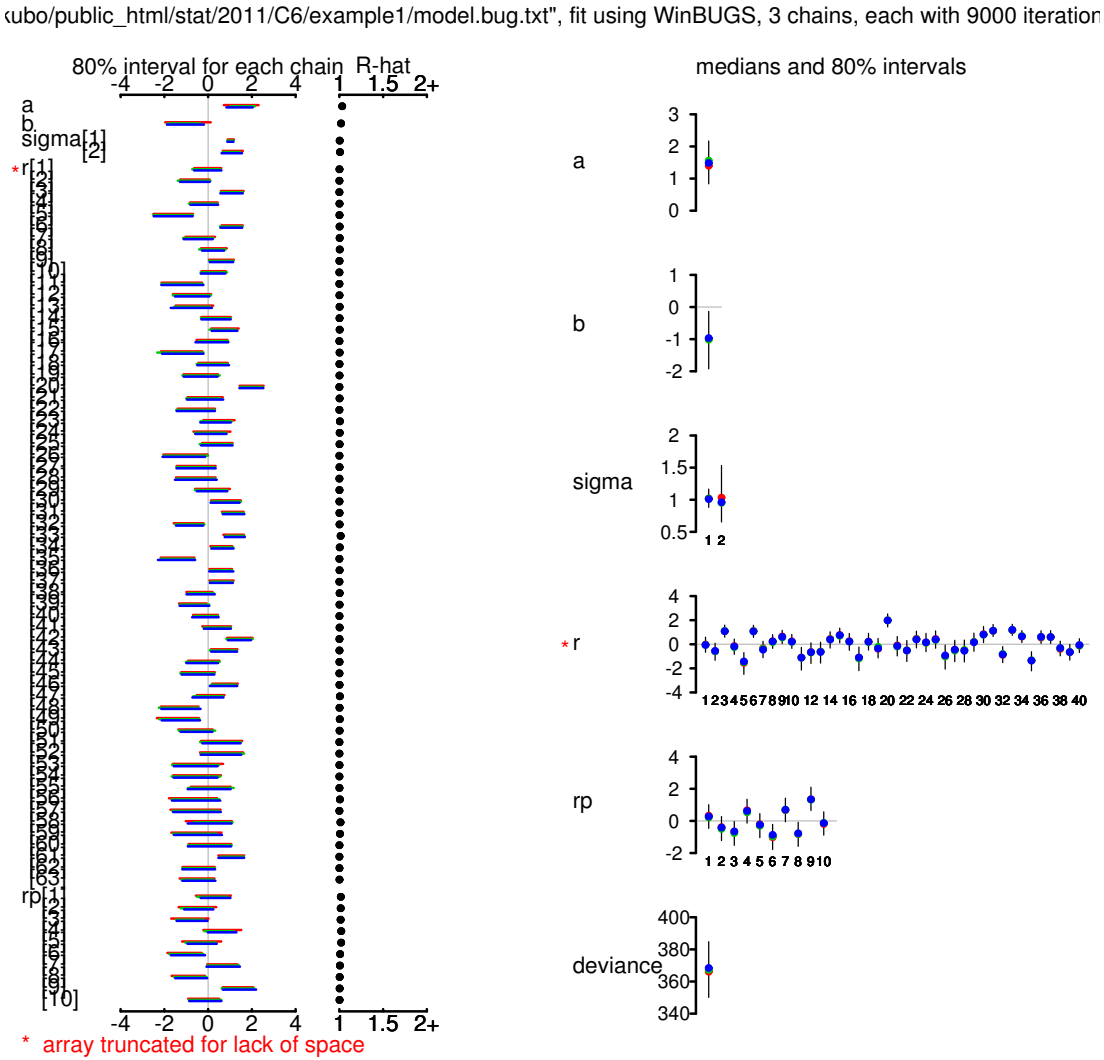
```
> library(glmmML)
> summary(glmmML(y ~ f, data = d, family = poisson,
+ cluster = id))
...(略)...
              coef se(coef)     z Pr(>|z|)
(Intercept)  1.351    0.192  7.05  1.8e-12
fT          -0.737    0.280 -2.63  8.4e-03
...(略)...
```

- やっぱり同じ?
- むしろ肥料処理の悪影響が強い?

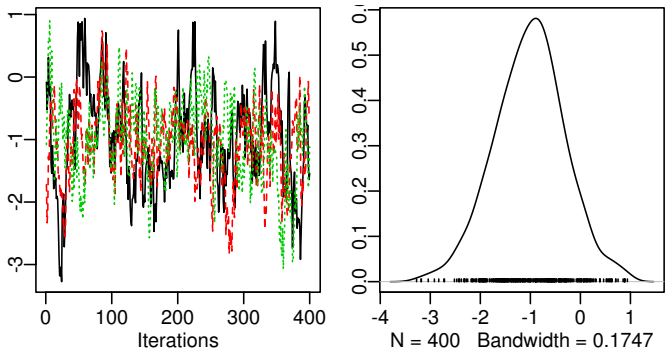
個体差 + ブロック差を考える階層ベイズモデル

- ここでは log リンク関数を使う
- 平均の対数 $\log(\lambda_i) = a + bf_i +$ (個体差) + (ブロック差)
- 事前分布の設定
  - 切片 a と f_i の係数 b は無情報事前分布 (すごく平らな正規分布)
  - 個体差とブロック差は階層的な事前分布 (それぞれ標準偏差 σ_1, σ_2 の正規分布，平均はゼロ)
  - 標準偏差 σ_* は無情報事前分布 ($[0, 10^4]$ の一様分布)

WinBUGS による事後分布の推定，R で収束判定



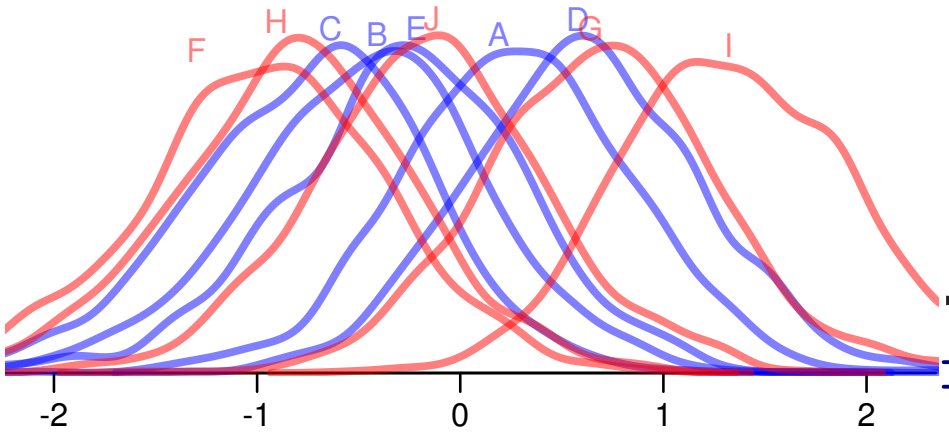
肥料の効果 (パラメーター b) はなさそう?



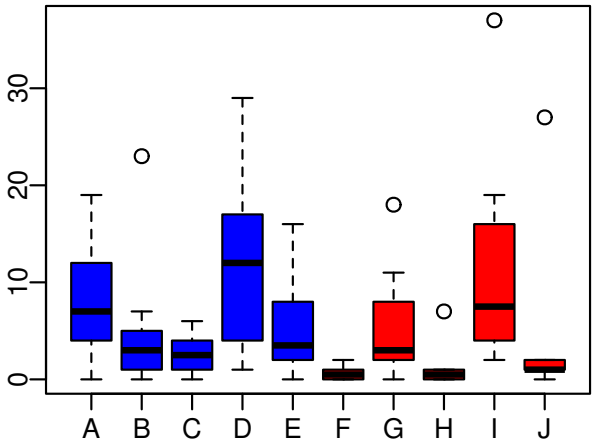

```
> print(post.bugs, digits.summary = 3)
...(略)...
             mean     sd    2.5%     25%     50%     75%   97.5%  Rhat n.eff
a           1.501  0.529   0.482   1.157   1.493   1.852   2.565 1.032   240
b          -1.016  0.706  -2.436  -1.476  -0.993  -0.565   0.395 1.019   450
sigma[1]    1.020  0.114   0.822   0.939   1.014   1.089   1.265 1.004   510
...(略)...
```

この架空データを生成した種子数シミュレイションでは，肥料の効果は<u>まったく無い</u>と設定していた

推定された植木鉢の差（ブロック差）



ブロック差
rp[j]



統計モデリングの手ぬきは危険!

- random effects つまり 個体差・ブロック差が大きい
- random effects の影響が大きいときには，fixed effects の大きさが見えにくくなる— ニセの「効果」が見えることもあれば，見えるはずの傾向が隠されることも
  - **個体差・ブロック差の階層ベイズモデルが必要!**
- もしブロック差を人為的に小さくできないなら，ブロック数をもっと増やして，より正確な植木鉢の効果のばらつきを正確に推定するしかない