統計モデリング入門 BSJ2015 (3)

二項分布の GLM と GLMM

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

日本行動計量学会 春セミナー http://goo.gl/vIdtcv

2015–03–07

ファイル更新時刻: 2015–03–07 00:40

この時間に説明したいこと

1. “N 個のうち k 個が生きてる” タイプのデータ
   - 上限のあるカウントデータ
2. ロジスティック回帰のモデル
   - もっともよく使われる GLM
3. ロジスティック回帰の部品
   - 二項分布 binomial distribution と logit link function
4. GLM だけでは実際のデータ解析はできない
   - 一般化線形混合モデル (GLMM) 登場!
5. 一般化線形混合モデル (GLMM) を作って推定
   - 個体差 r_i を積分して消す尤度方程式
6. 現実のデータ解析には GLMM が必要
   - 個体差・グループ差を考えないといけないから

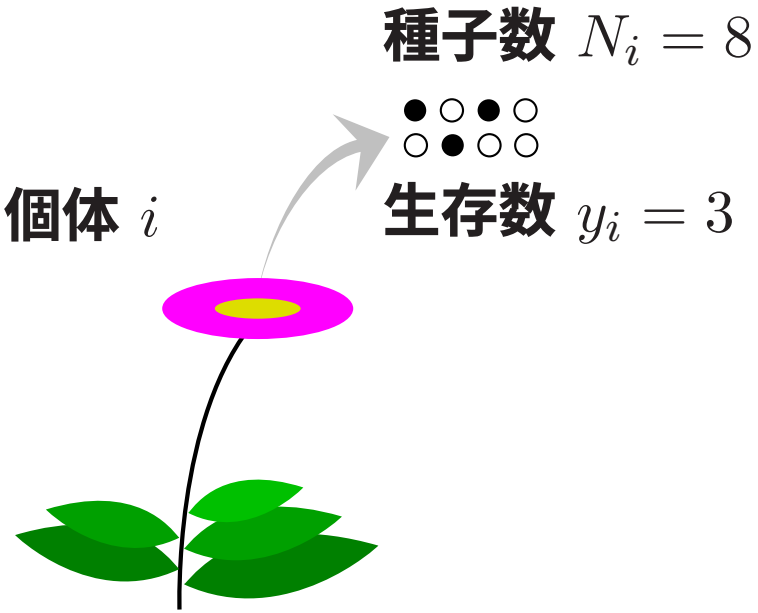
1. “N 個のうち k 個が生きてる”タイプのデータ

上限のあるカウントデータ

ポアソン分布ではなく二項分布で

例題: 植物の種子の生存確率

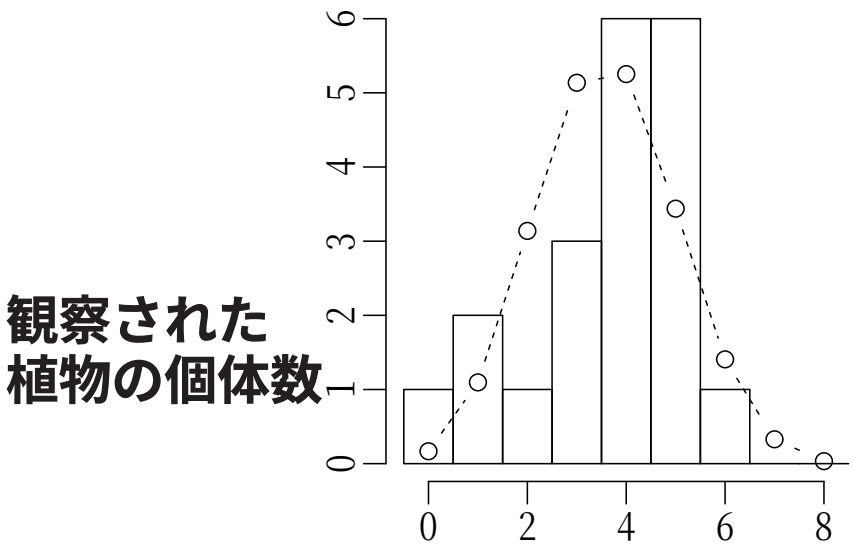
- 架空植物の種子の生存を調べた
- **種子**: 生きていれば発芽する
  - どの個体でも 8 個 の種子を調べた
- **生存確率**: ある種子が生きている確率



- データ: 植物 20 個体，合計 160 種子の生存の有無を調べた
- 73 種子が生きていた — このデータを統計モデル化したい

たとえばこんなデータが得られたとしましょう

個体ごとの生存数	0	1	2	3	4	5	6	7	8
観察された個体数	1	2	1	3	6	6	1	0	0



これは個体差なしの均質な集団

生存確率 q と二項分布の関係

- 生存確率を推定するために**二項分布** という確率分布を使う
- 個体 i の N_i 種子中 y_i 個が生存する確率

$$p(y_i \mid q) = \binom{N_i}{y_i} q^{y_i} (1-q)^{N_i - y_i},$$

- ここで仮定していること
  - **個体差はない**
  - つまり **すべての個体で同じ生存確率** q

尤度（ゆうど）：20 個体ぶんのデータが観察される確率

- 観察データ $\{y_i\}$ が確定しているときに
- パラメータ q は値が自由にとりうると考える
- 尤度 は 20 個体ぶんのデータが得られる確率の積，パラメータ q の関数として定義される

$$L(q|\{y_i\}) = \prod_{i=1}^{20} p(y_i \mid q)$$

個体ごとの生存数	0	1	2	3	4	5	6	7	8
観察された個体数	1	2	1	3	6	6	1	0	0

対数尤度方程式と最尤推定

- この尤度 $L(q \mid \text{データ})$ を最大化するパラメータの推定量 $\hat{q}$ を計算したい
- 尤度を対数尤度になおすと

$$\log L(q \mid \text{データ}) = \sum_{i=1}^{20} \log \binom{N_i}{y_i} + \sum_{i=1}^{20} \{y_i \log(q) + (N_i - y_i) \log(1-q)\}$$

- この対数尤度を最大化するように未知パラメーター q の値を決めてやるのが**最尤推定**

最尤推定 (MLE) とは何か

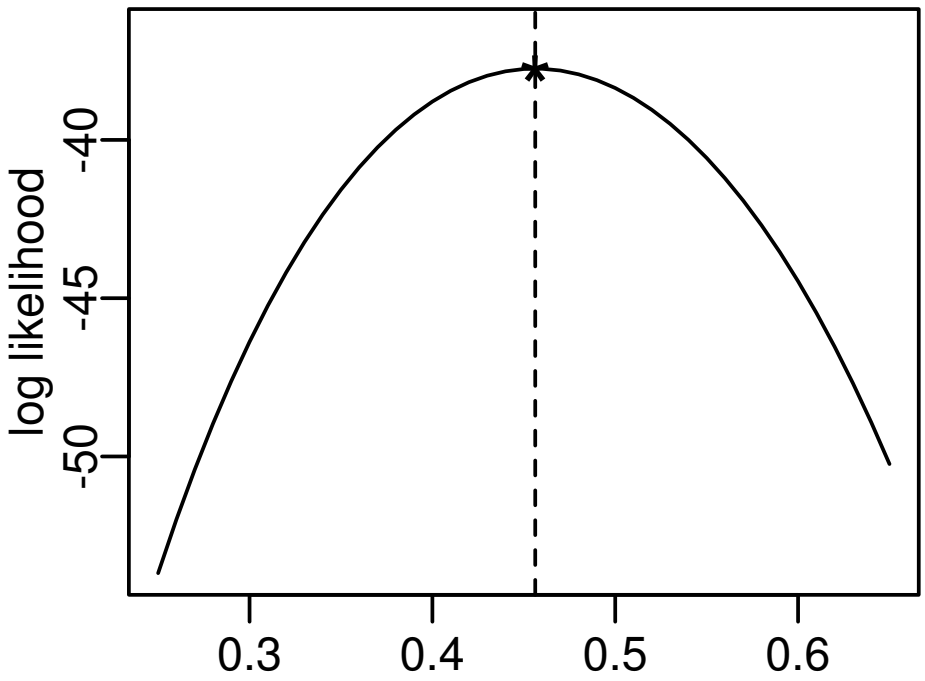
- **対数尤度 $L(q \mid$ データ$)$ が最大になるパラメーター q の値をさがしだすこと**

- **対数尤度 $\log L(q \mid$ データ$)$ を q で偏微分して 0 となる $\hat{q}$ が対数尤度最大**

$$\partial \log L(q \mid \text{データ}) / \partial q = 0$$

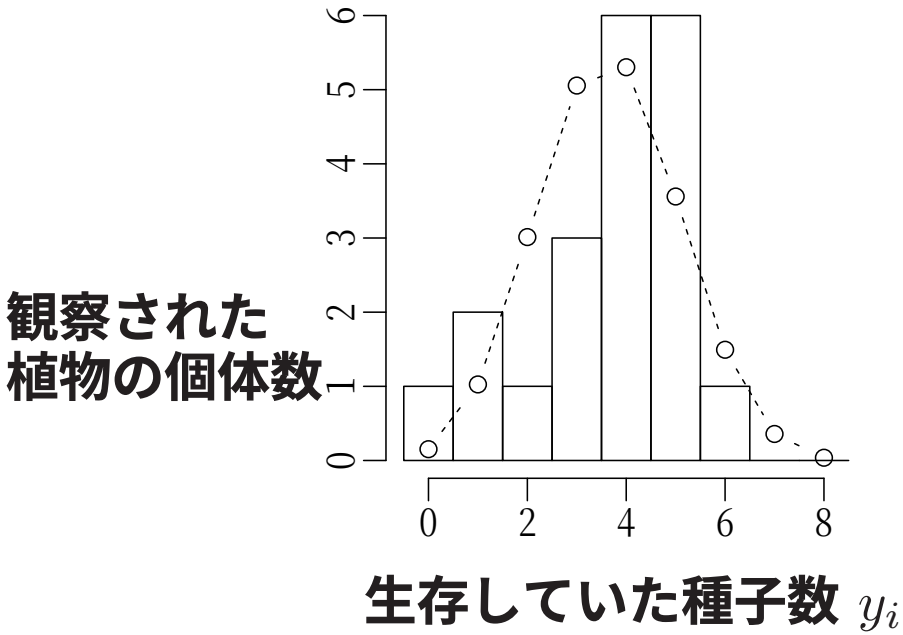
- **生存確率 q が全個体共通の場合の最尤推定量・最尤推定値は**

$$\hat{q} = \frac{\text{生存種子数}}{\text{調査種子数}} = \frac{73}{160} = 0.456 \text{ ぐらい}$$



二項分布で説明できる 8 種子中 y_i 個の生存

$\hat{q} = 0.46$ **なので** $\binom{8}{y} 0.46^y \ 0.54^{8-y}$



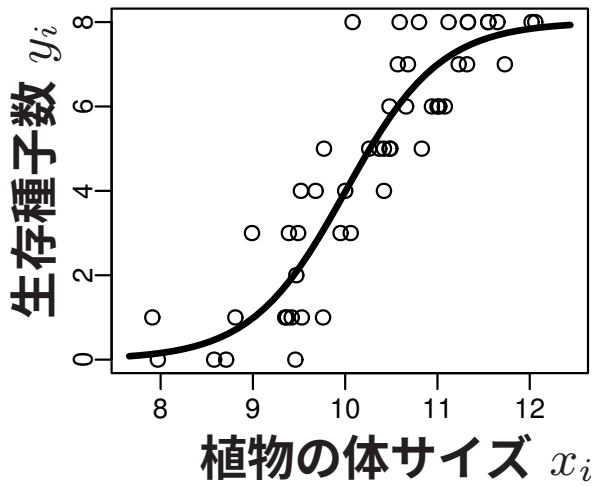
2. ロジスティック回帰のモデル

もっともよく使われる GLM

GLM のひとつである logistic 回帰モデルを指定する

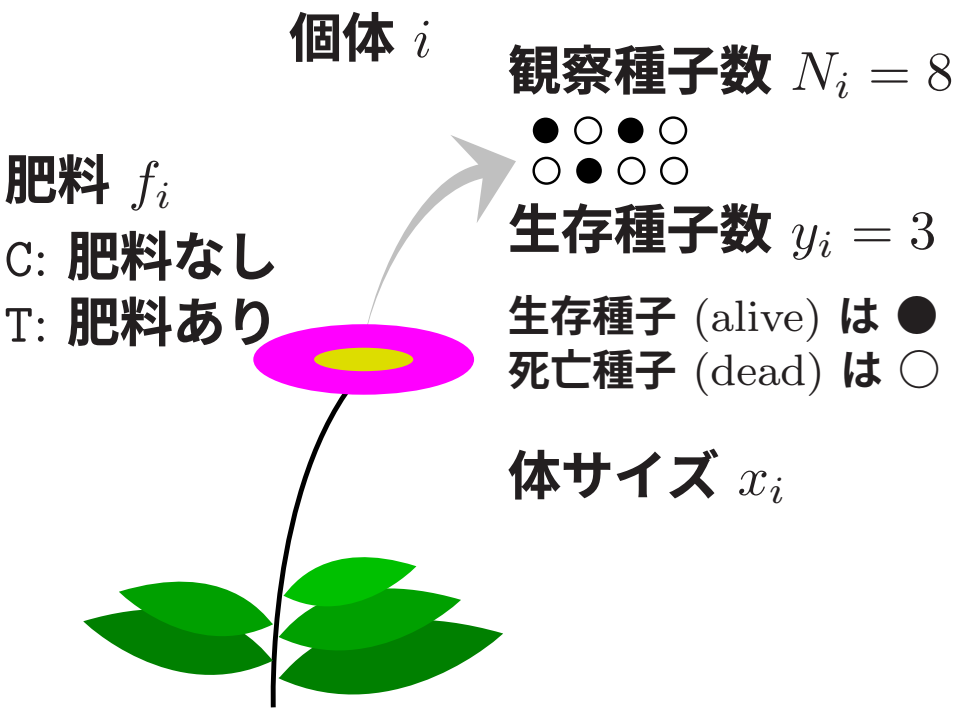
ロジスティック回帰のモデル

- **確率分布**: **二項分布**
- **線形予測子**: e.g., $\beta_1 + \beta_2 x_i$
- **リンク関数**: logit **リンク関数**



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



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/2014/Fig/binomial/data4a.csv")
```

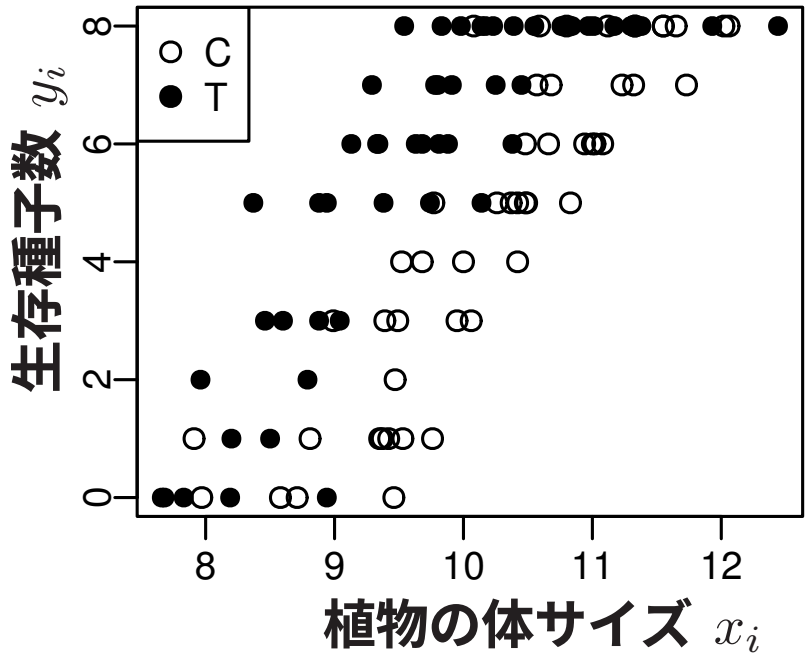
データは d と名付けられた data frame (表みたいなもの) に格納される

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 がきいている?

3. ロジスティック回帰の部品

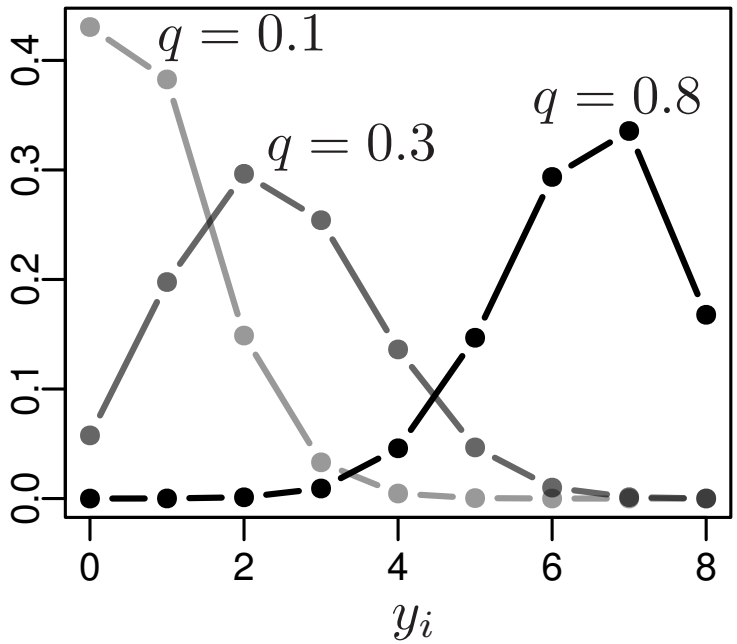
二項分布 binomial distribution と logit link function

二項分布: N 回のうち y 回，となる確率

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

$\binom{N}{y}$ 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 $p(y_i \mid 8, q)$

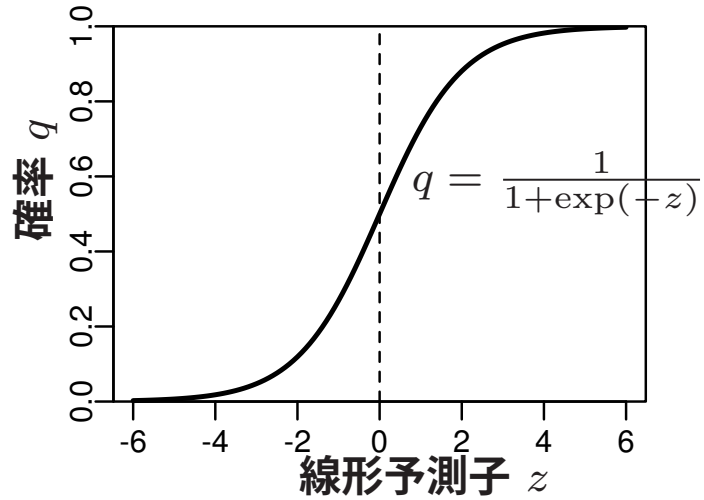


ロジスティック曲線とはこういうもの

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

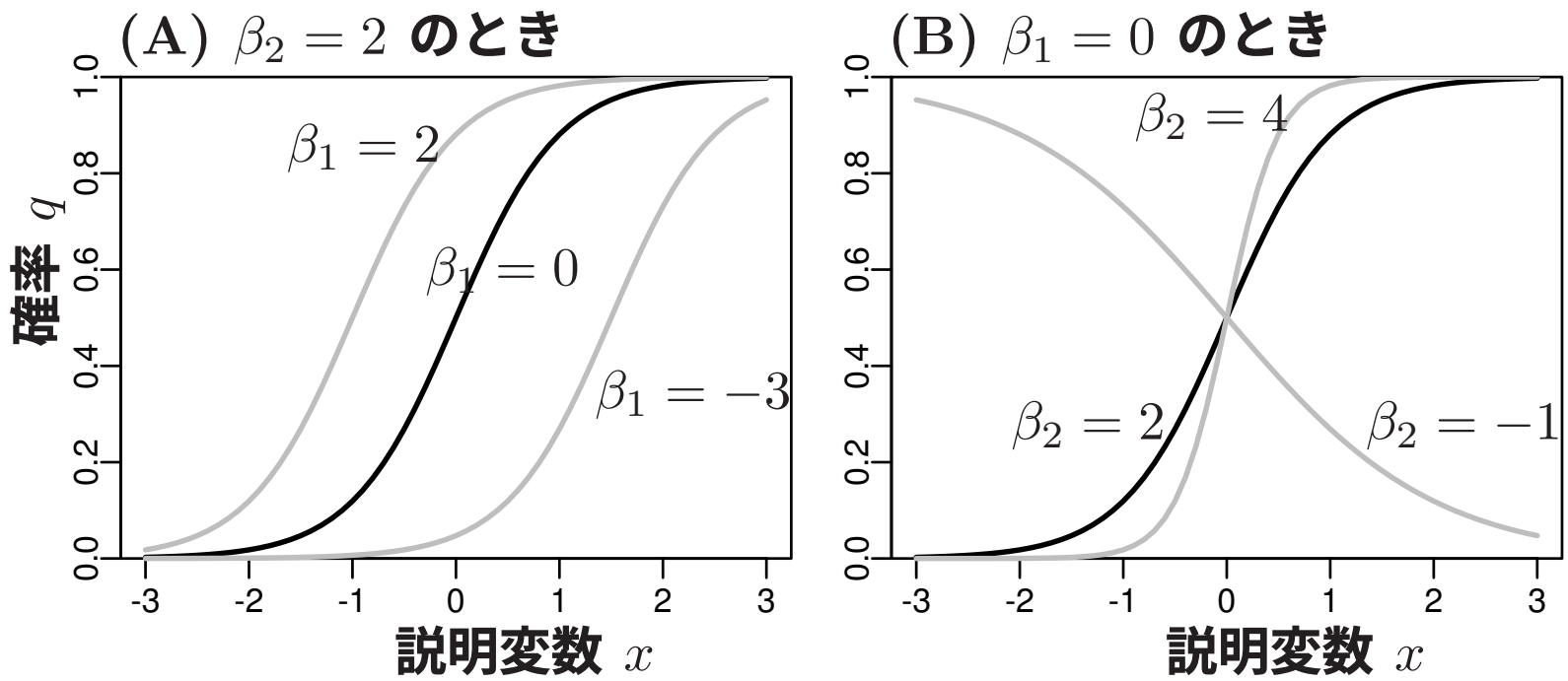
$$q_i = \mathrm{logistic}(z_i) = \frac{1}{1 + \exp(-z_i)}$$

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



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**

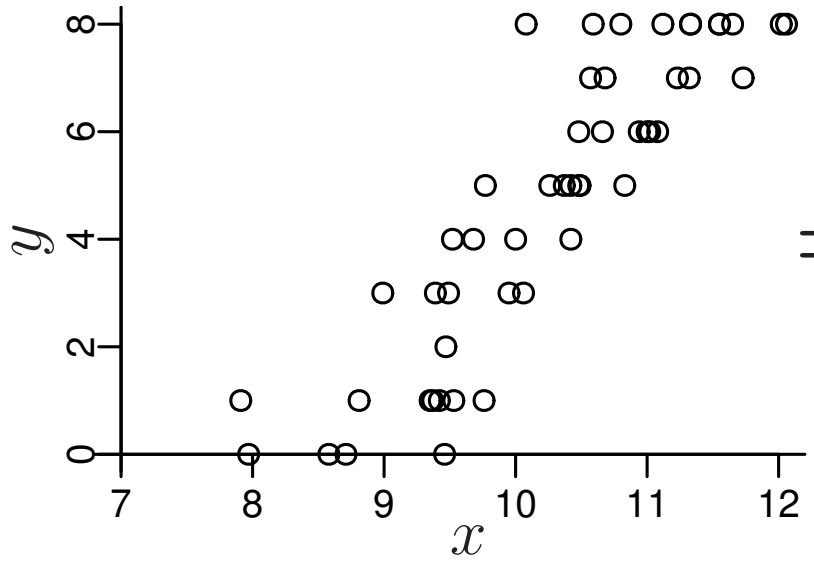
$$\text{logit}(q) = \log \frac{q}{1-q} = \beta_1 + \beta_2 x$$

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

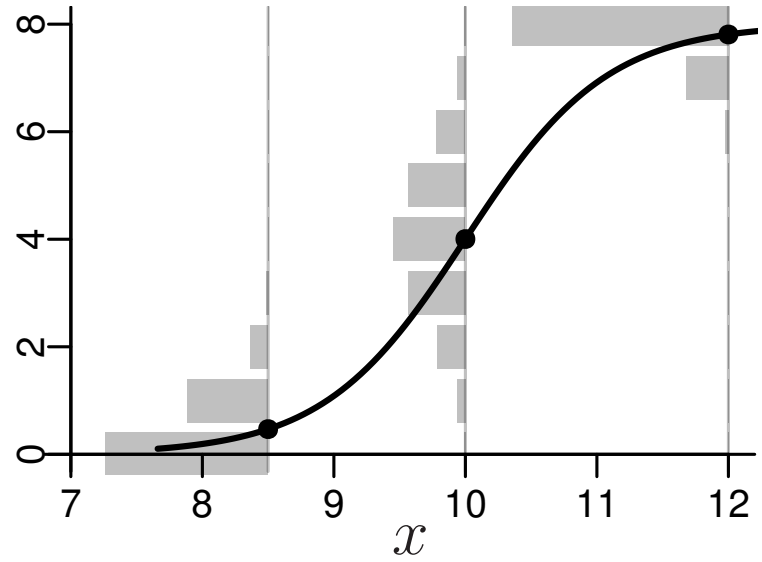
logit is the inverse function of logistic function, vice versa

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）

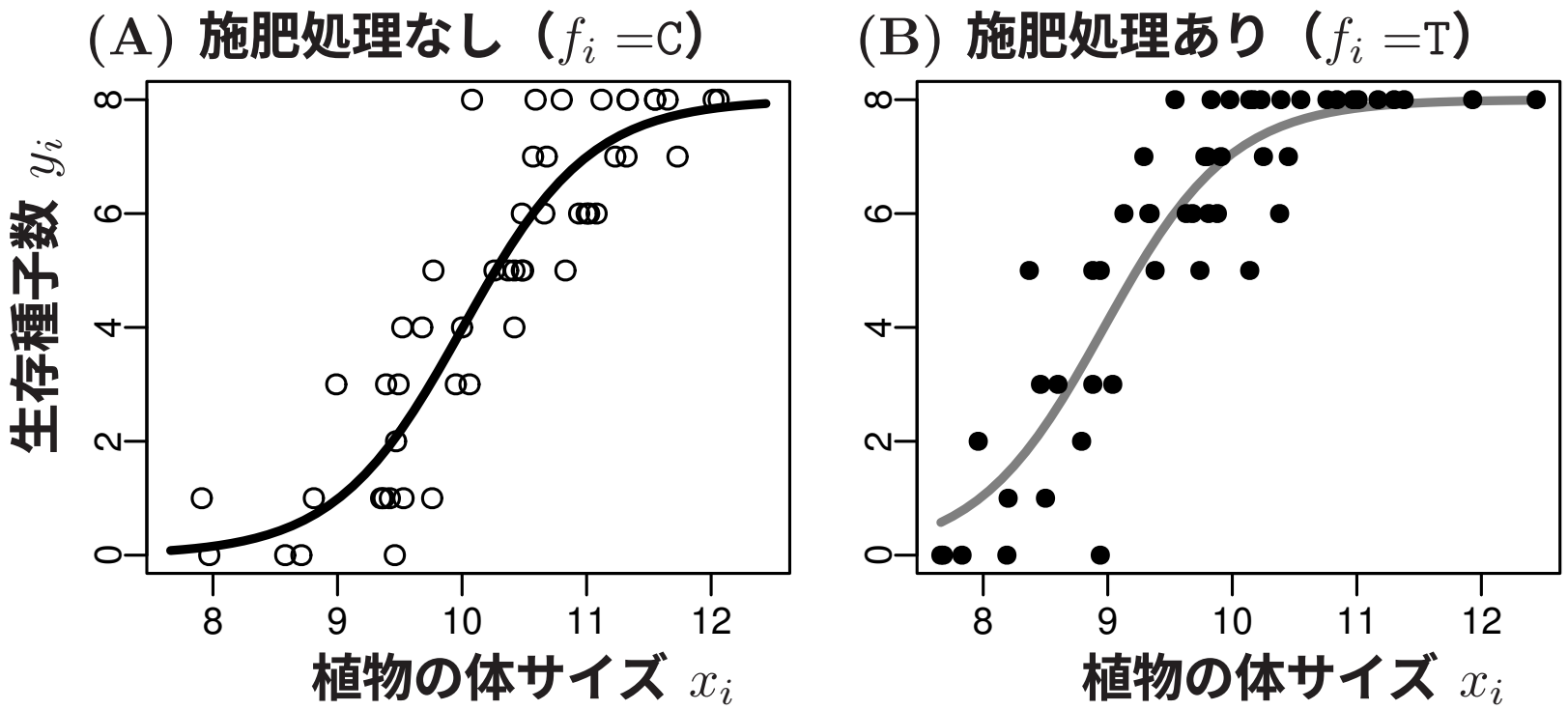


(B) 推定されるモデル




```
> 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

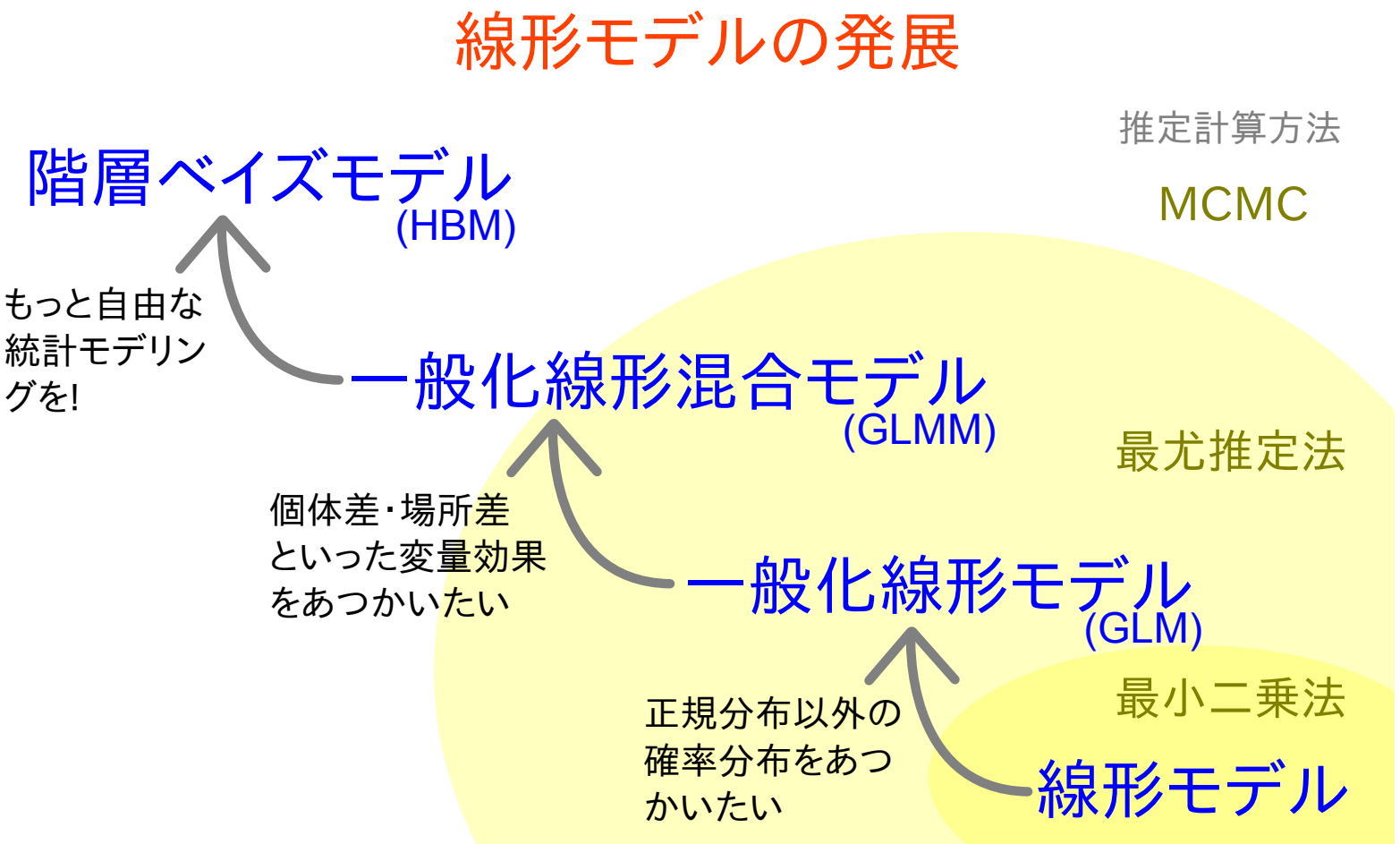


4. GLM だけでは実際のデータ解析はできない

一般化線形混合モデル (GLMM) 登場!

GLM は「個体差」などを無視しているところが問題

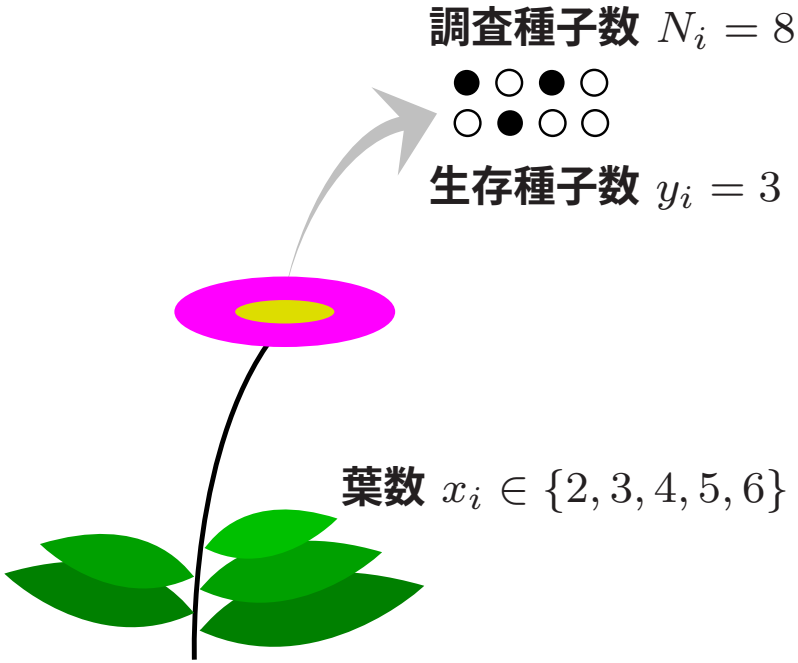
“統計モデリング入門” に登場する統計モデル



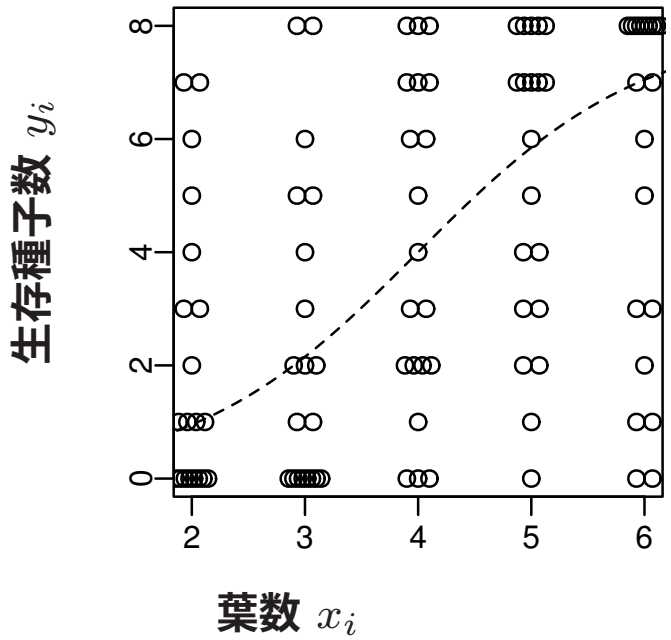
データの特徴にあわせて線形モデルを改良・発展させる

種子生存確率の GLMM

(A) 個体 i で観測されたデータ



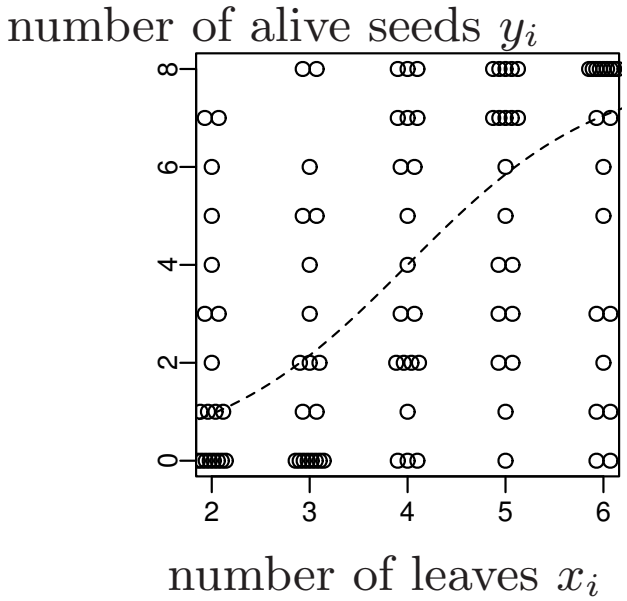
(B) 全 100 個体の x_i と y_i



“N 個中の y 個” というデータ $\rightarrow$ ロジスティック回帰?

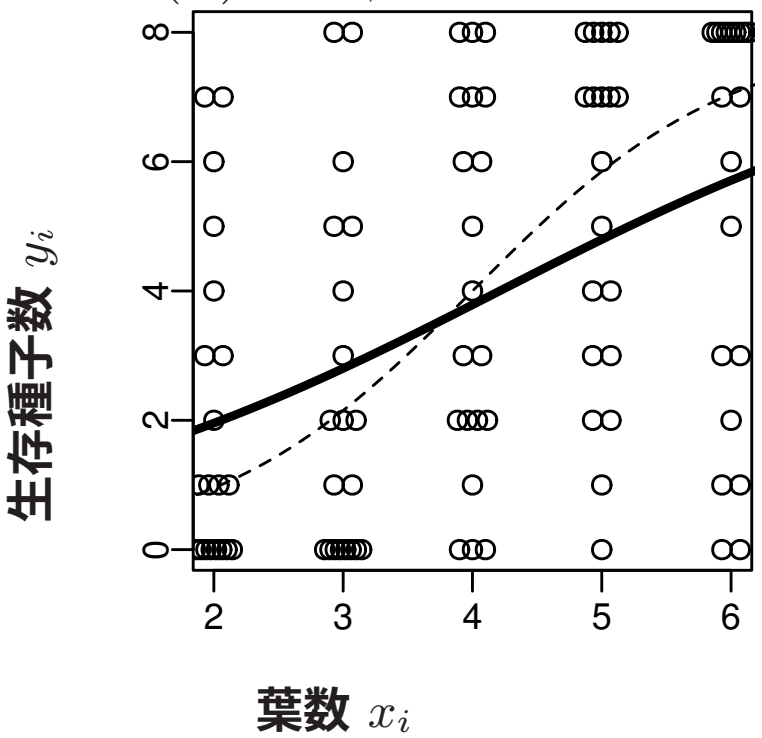
ロジスティック回帰のモデル

- **確率分布**: **二項分布**
- **線形予測子**: $\beta_1 + \beta_2 x_i$
- **リンク関数**: logit **リンク関数**

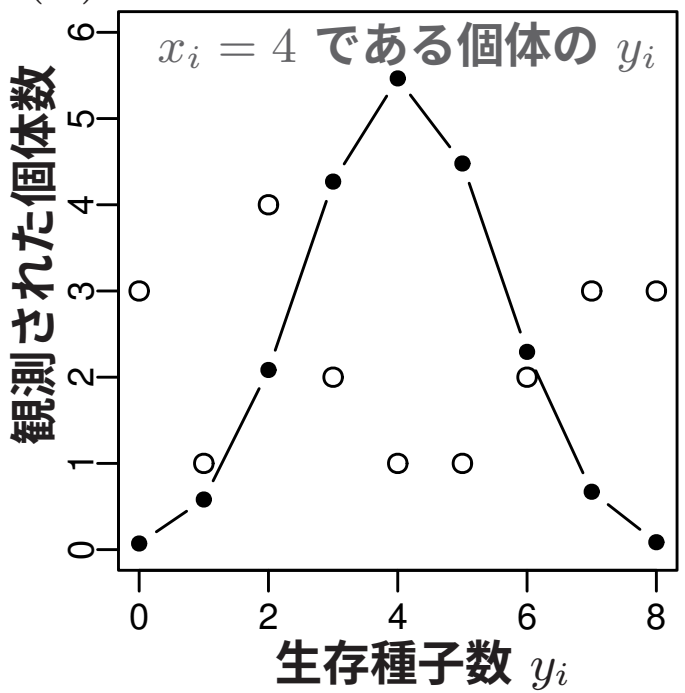


GLM では説明できないばらつき!

(A) 傾き β_2 が過小推定

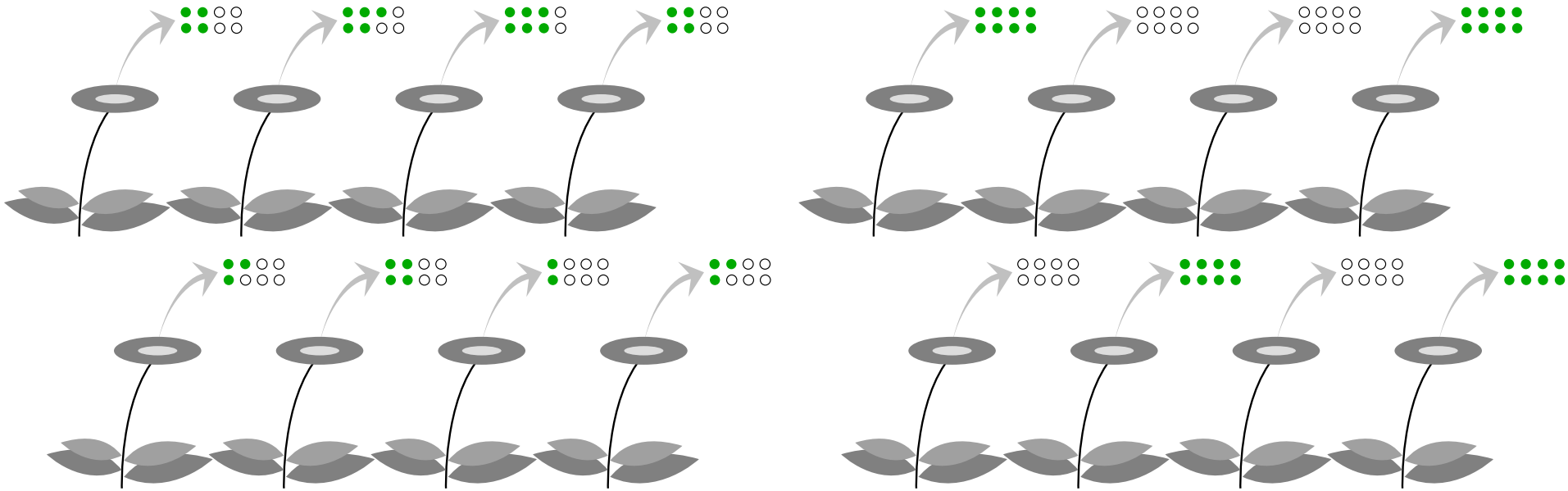


(B) ぜんぜん二項分布じゃない!



○が観測されたデータの図示

過分散 (overdispersion) とは何か?

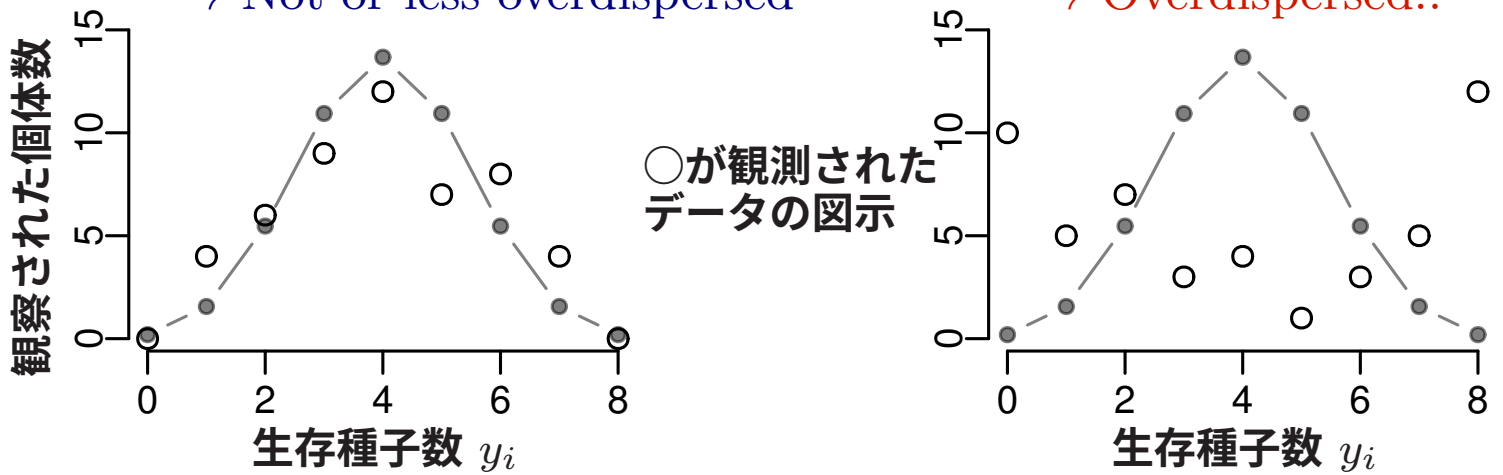


(A) 個体差のばらつきが小さい場合

→ Not or less overdispersed

(B) 個体差のばらつきが大きい場合

→ Overdispersed!!



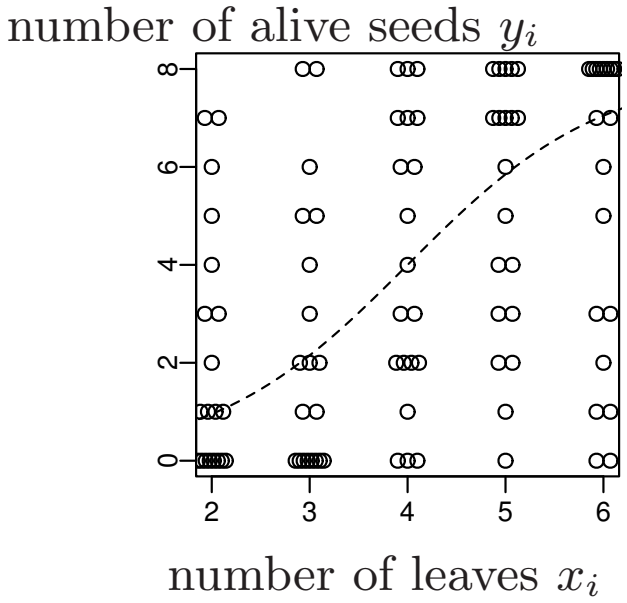
ロジスティック回帰やポアソン回帰といった GLM では全サンプルの均質性を仮定している

現実のカウントデータは，多くの場合「過分散」

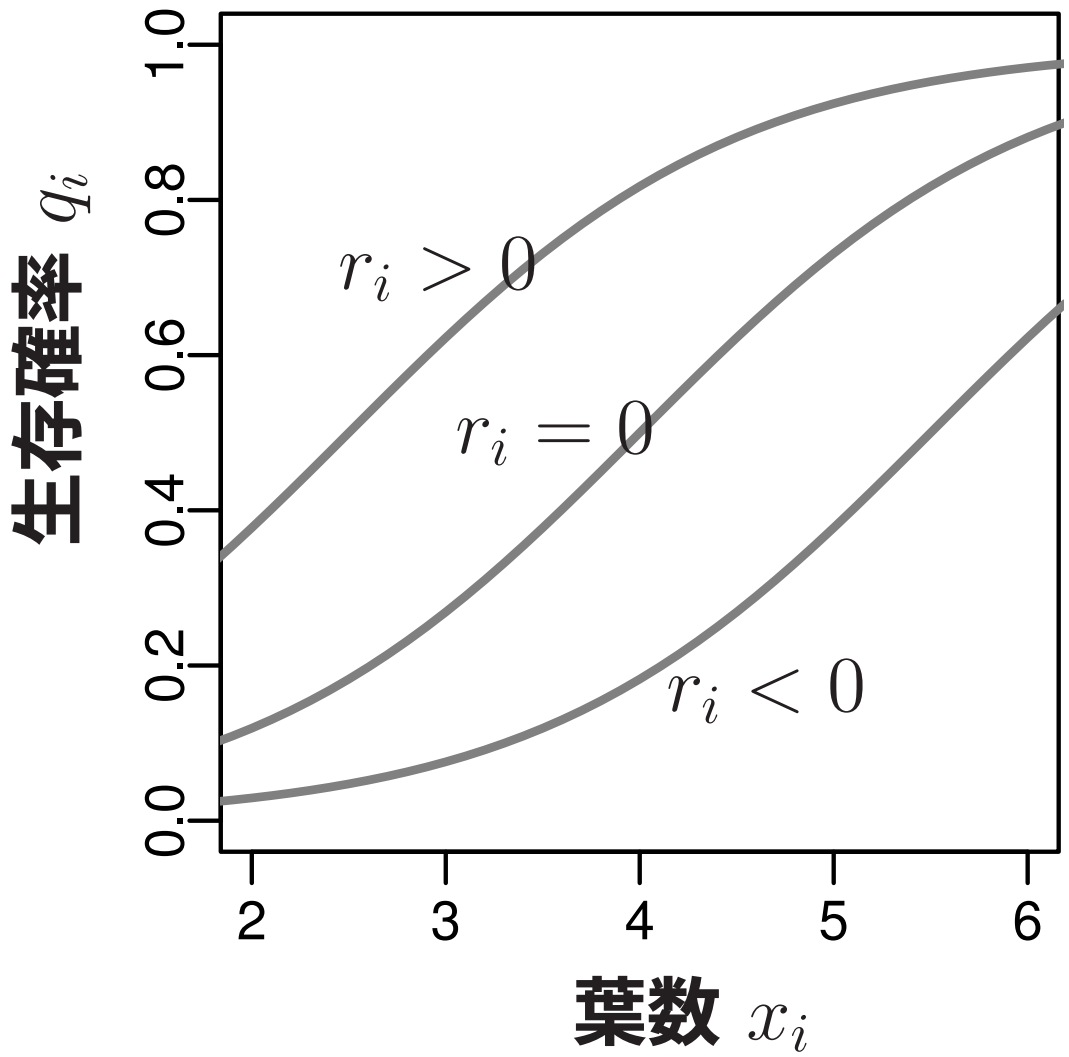
ロジスティック回帰のモデルを改良する

ロジスティック回帰のモデル

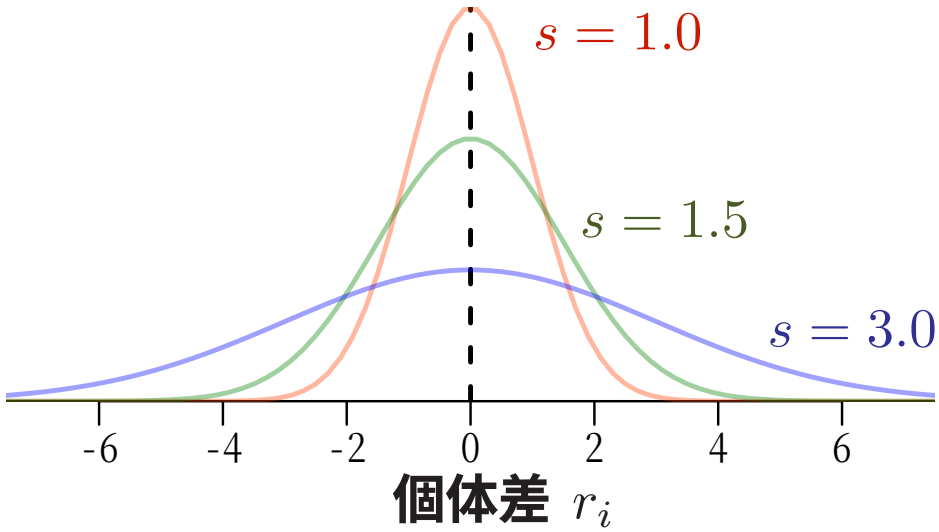
- **確率分布**: **二項分布**
- **線形予測子**: $\beta_1 + \beta_2 x_i + r_i$
- **リンク関数**: logit **リンク関数**



個体 i の個体差を r_i としてみよう



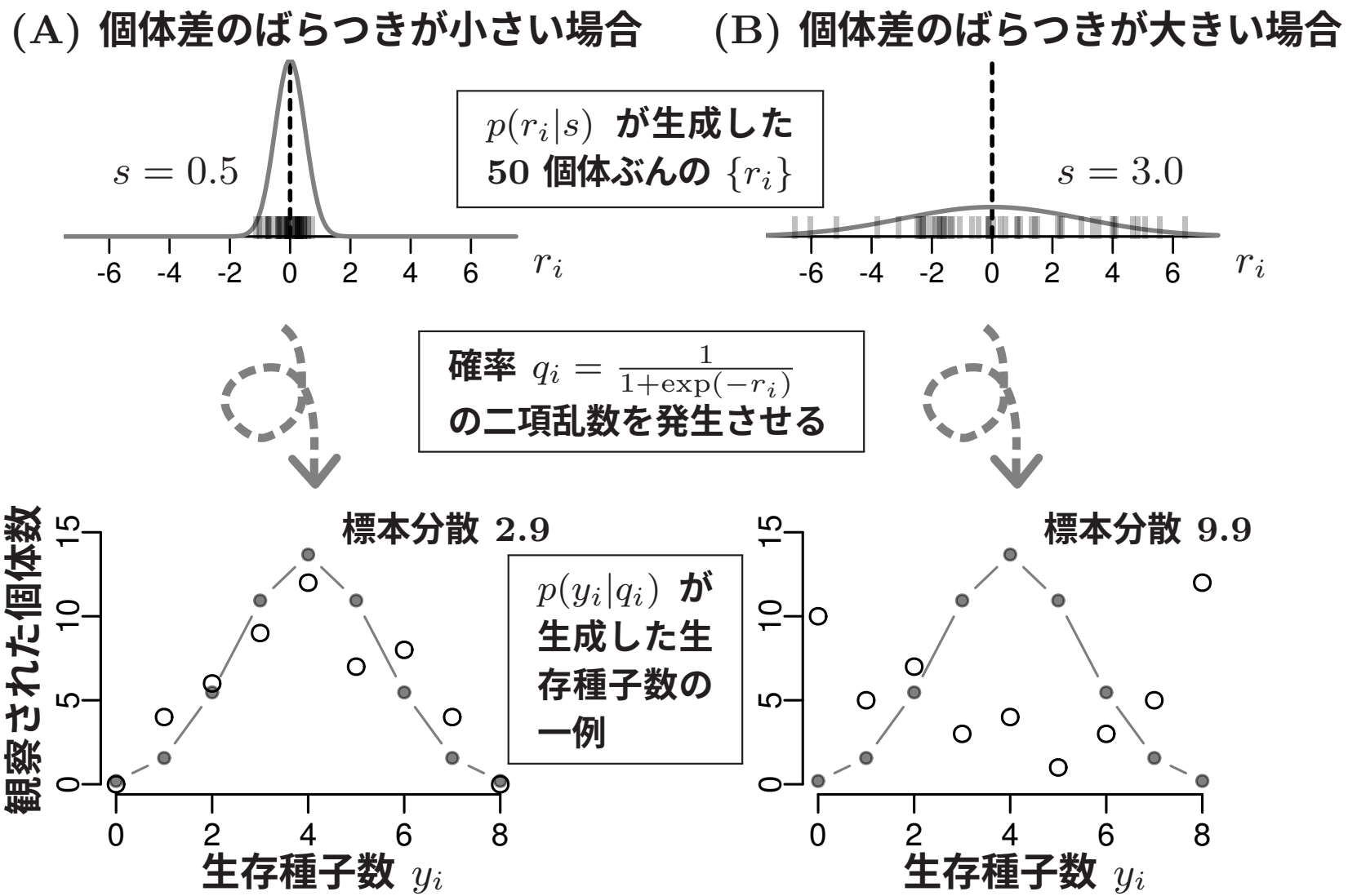
$\{r_i\}$ のばらつきは正規分布だと考えてみる



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

この確率密度 $p(r_i \mid s)$ 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



固定効果 と ランダム効果

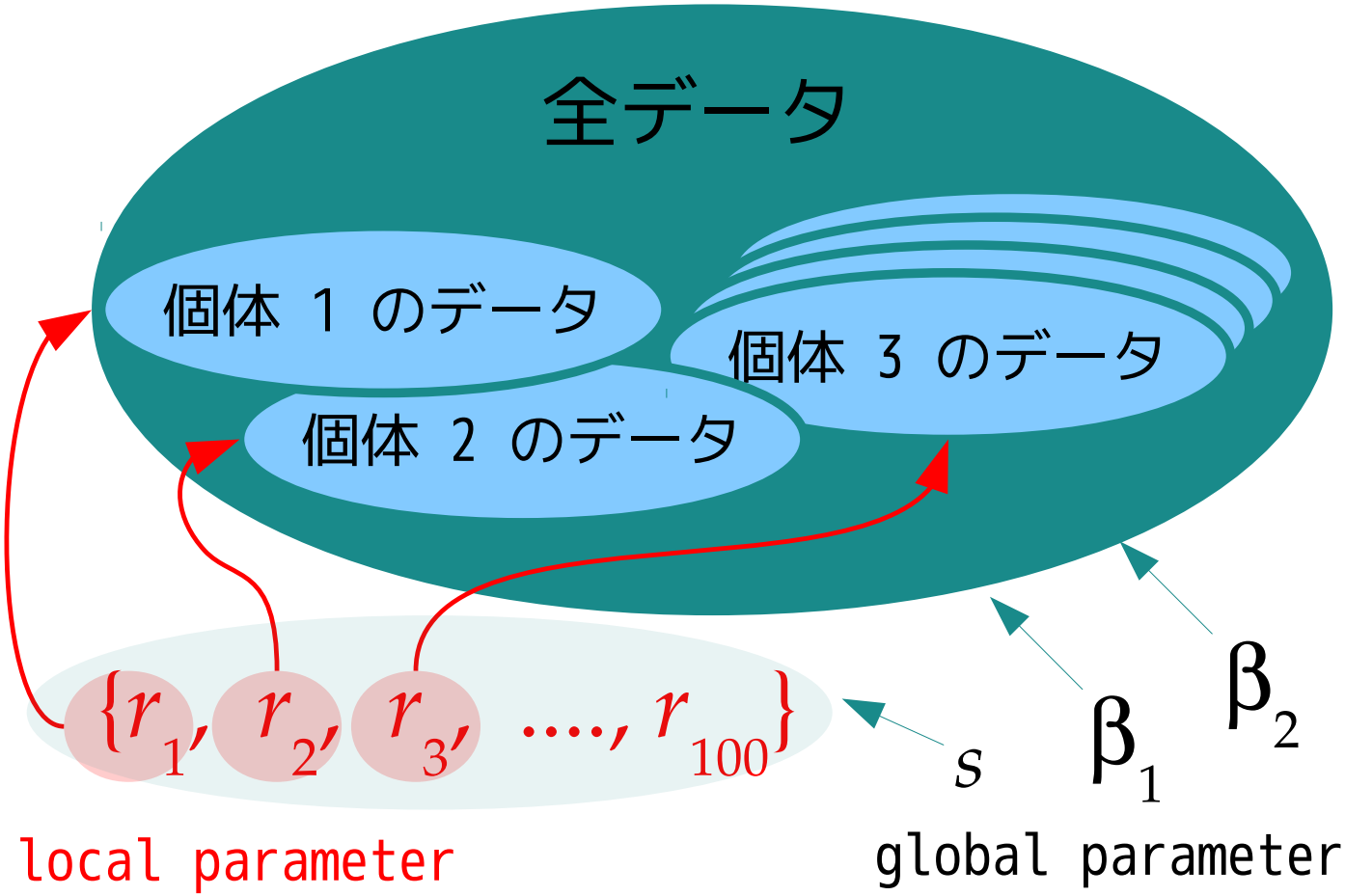
Generalized Linear Mixed Model (GLMM)

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……?**

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?

global parameter と local parameter

Generalized Linear Mixed Model (GLMM)

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

- fixed effects: $\beta_1 + \beta_2 x_i$
  - global parameter — **全個体に共通**
- **全個体のばらつき** s **も** global parameter
- random effects: $+r_i$
  - local parameter — **個体** i **だけを説明**

5. 一般化線形混合モデル (GLMM) を作って推定

個体差 r_i を積分して消す尤度方程式

global parameter と local parameter

Generalized Linear Mixed Model (GLMM)

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 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 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
```

個々の r_i を推定するには データが少なすぎるから

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 — 二項分布

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

個体差 r_i のばらつき — 正規分布

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

個体 i の 尤度 — r_i を消す

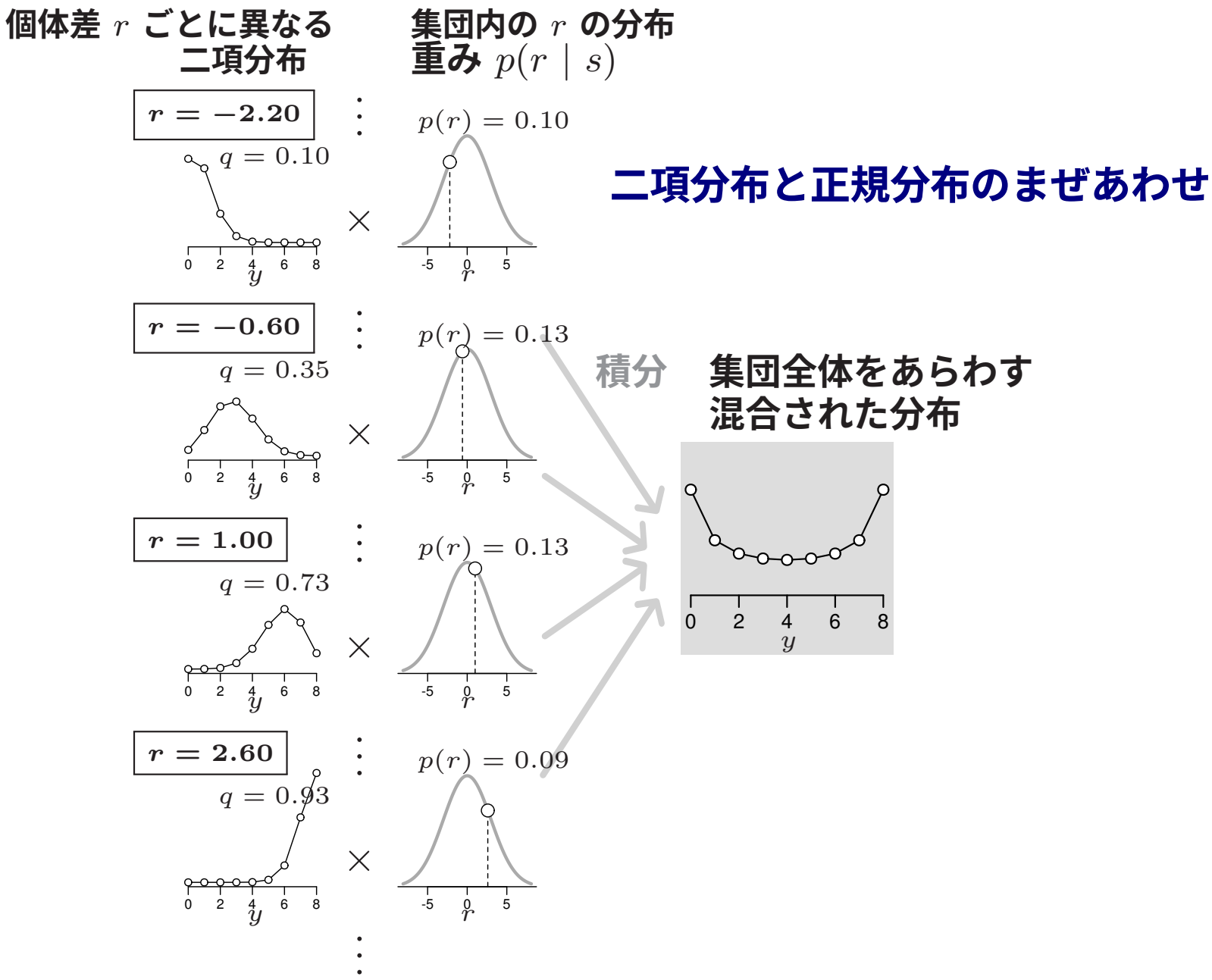
$$L_i = \int_{-\infty}^{\infty} p(y_i \mid \beta_1, \beta_2, r_i)\ p(r_i \mid s) dr_i$$

全データの尤度 — β_1, β_2, s の関数

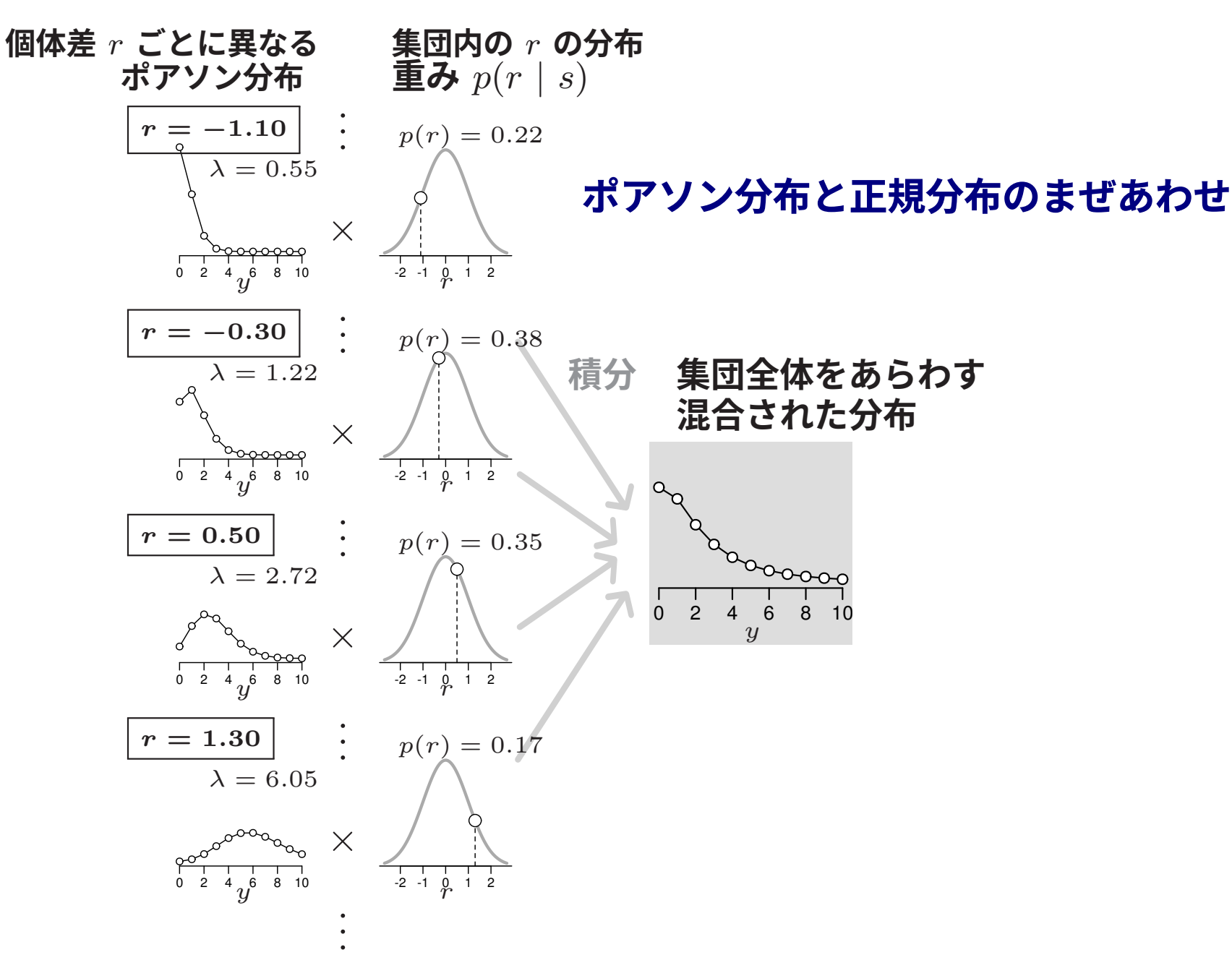
$$L(\beta_1, \beta_2, s) = \prod_i L_i$$

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

Integral of $r_i \rightarrow$ mixture distribution of the binomial and Gaussian distributions



二項分布と正規分布のまぜあわせ


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 $p(r \mid s)$
$r = -1.10$
$\lambda = 0.55$
$p(r) = 0.22$
$r = -0.30$
$\lambda = 1.22$
$p(r) = 0.38$
$r = 0.50$
$\lambda = 2.72$
$p(r) = 0.35$
$r = 1.30$
$\lambda = 6.05$
$p(r) = 0.17$
y
r
積分
集団全体をあらわす
混合された分布
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: $\hat{\beta}_1, \hat{\beta}_2, \hat{s}$

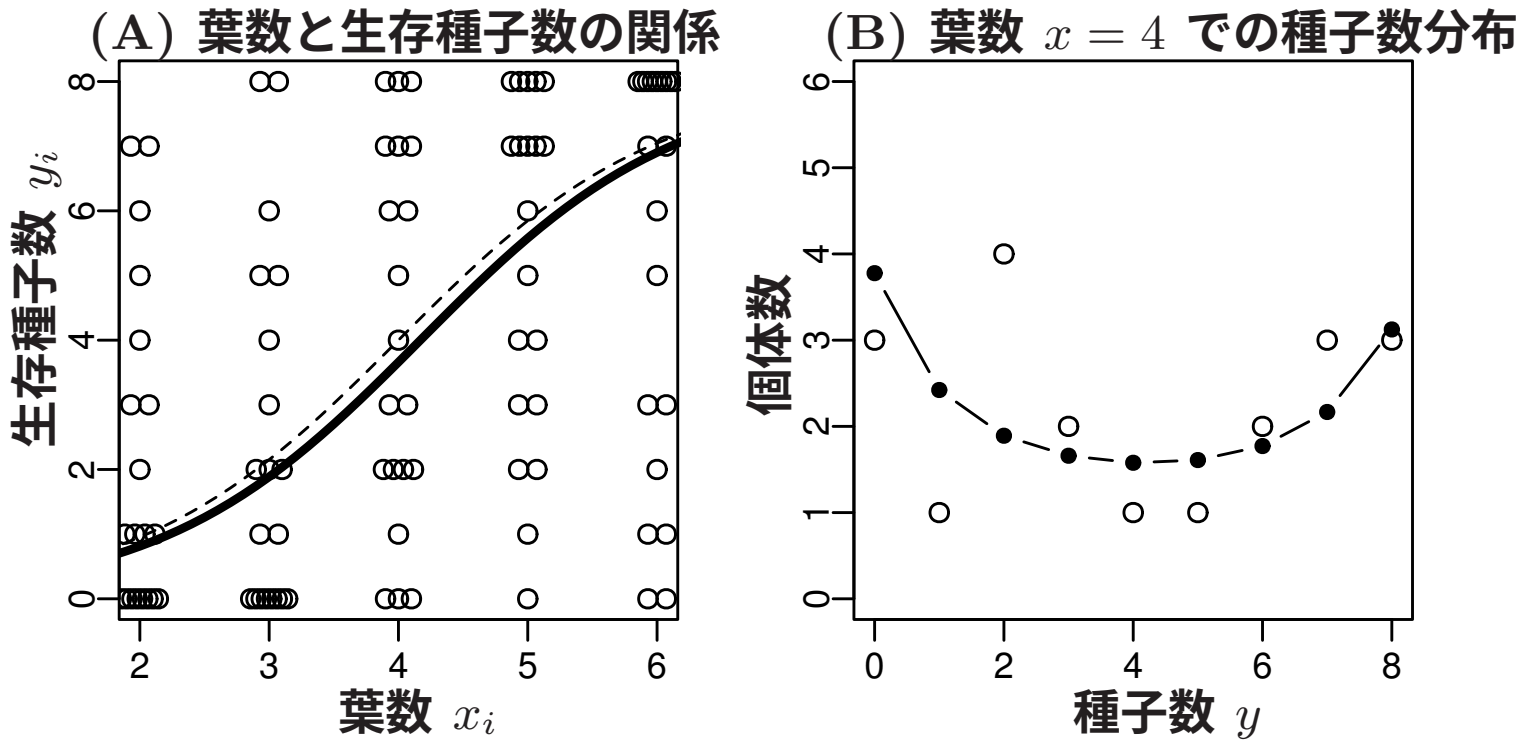
```
> glmmML(cbind(y, N - y) ~ x, data = d, family = binomial,
+ cluster = id)
...(snip)...
             coef se(coef)     z Pr(>|z|)
(Intercept) -4.13    0.906 -4.56  5.1e-06
x            0.99    0.214  4.62  3.8e-06

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 を使った予測

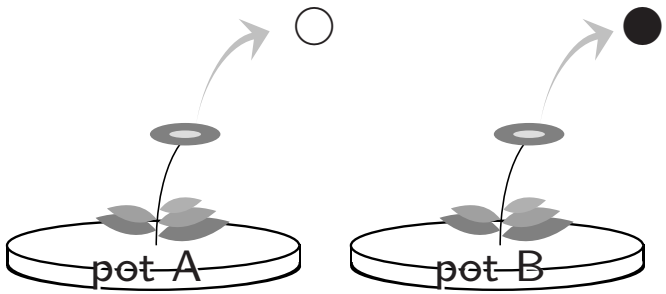


6. 現実のデータ解析には GLMM が必要

個体差・グループ差を考えないといけないから

個体差 + 場所差の GLMM I

(A) **個体・植木鉢が反復**

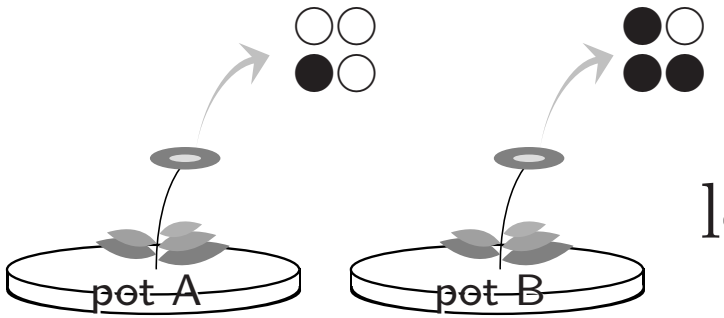


個体差も植木鉢差も推定できない

$$\mathrm{logit} q_i = \beta_1 + \beta_2 x_i \ (\mathrm{GLM})$$

q_i: **種子の生存確率**

(B) **個体は擬似反復，植木鉢は反復**



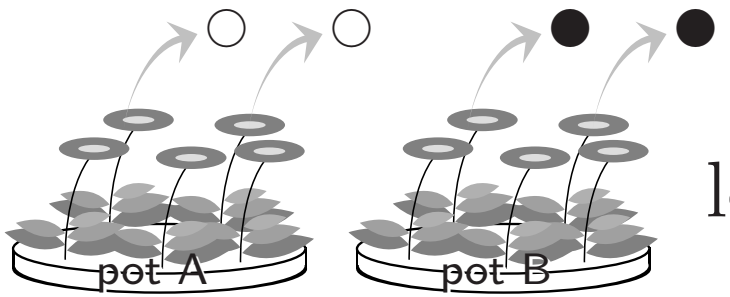
個体差は推定できる
植木鉢差は推定できない

$$\mathrm{logit} q_i = \beta_1 + \beta_2 x_i + r_i$$

より正確にいうと (A) (B) **は個体差と植木鉢差の区別がつかない**

個体差 + 場所差の GLMM II

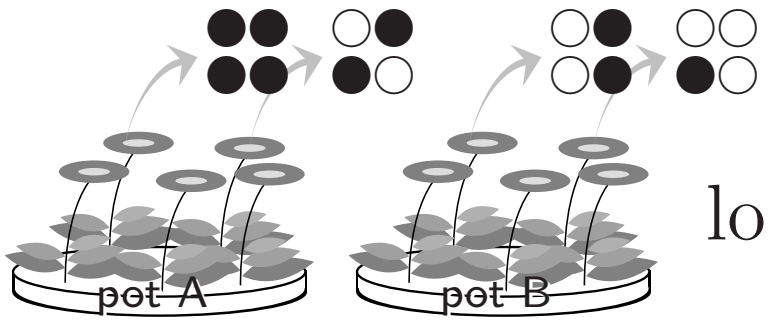
(C) 個体は反復，植木鉢は擬似反復



個体差は推定できない
植木鉢差は推定できる

$$\mathrm{logit} q_i = \beta_1 + \beta_2 x_i + r_j$$

(D) 個体・植木鉢が擬似反復



個体差も植木鉢差も
推定できる

$$\mathrm{logit} q_i = \beta_1 + \beta_2 x_i + r_i + r_j$$

複雑なモデルほど最尤推定は困難，しかも多くのデータが必要

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 ……