統計モデリング入門 道総研 [07]

一般化線形混合モデル (GLMM)

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

道総研勉強会 http://goo.gl/HQbeoh

2015–11–17

ファイル更新時刻: 2015–11–12 22:11

この時間に説明したいこと

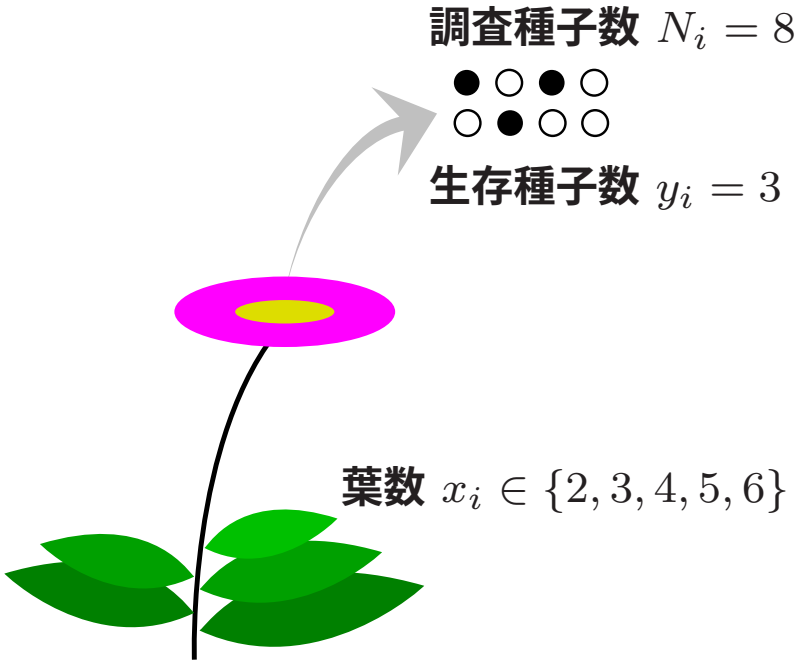
1. **GLM だけでは実際のデータ解析はできない**
   GLM は「個体差」などを無視しているから
2. **一般化線形混合モデル (GLMM) を作って推定**
   個体差 r_i を積分して消す尤度方程式
3. **現実のデータ解析には GLMM が必要**
   個体差・グループ差を考えないといけないから

1. GLM だけでは実際のデータ解析はできない

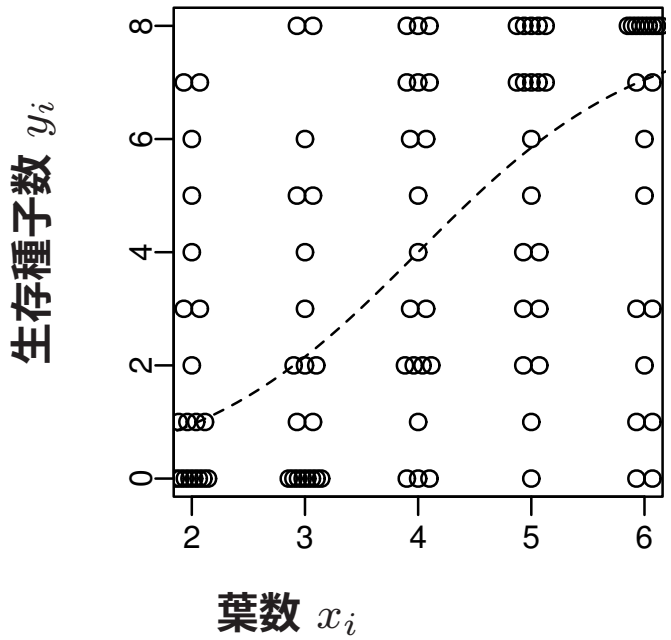
GLM は「個体差」などを無視しているから

種子生存確率の GLMM

(A) 個体 i で観測されたデータ



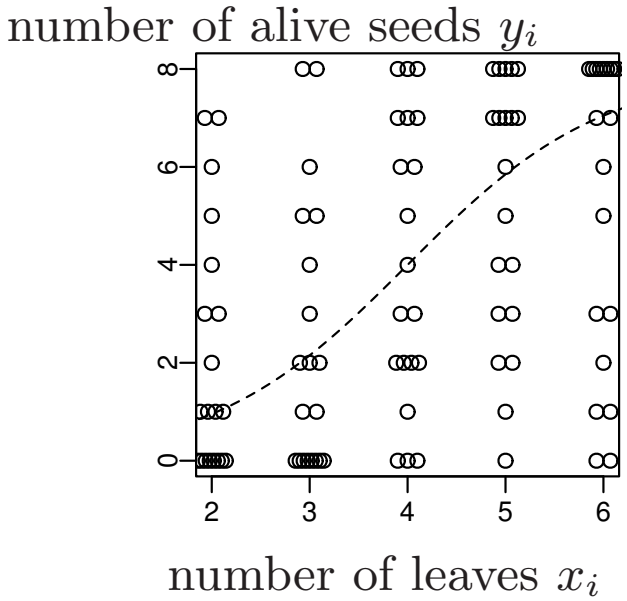
(B) 全 100 個体の x_i と y_i



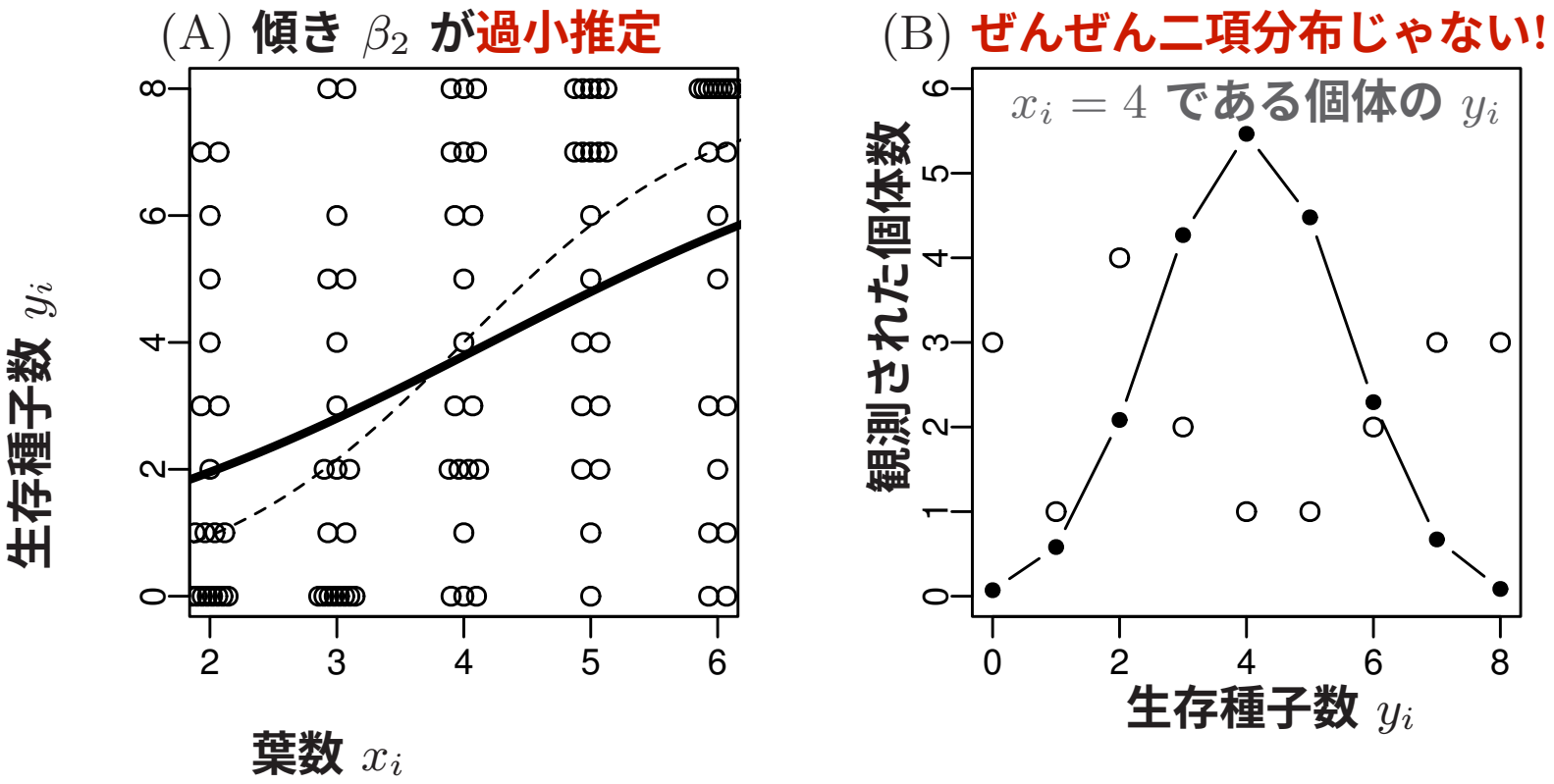
“N 個中の y 個” というデータ $\rightarrow$ ロジスティック回帰?

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**

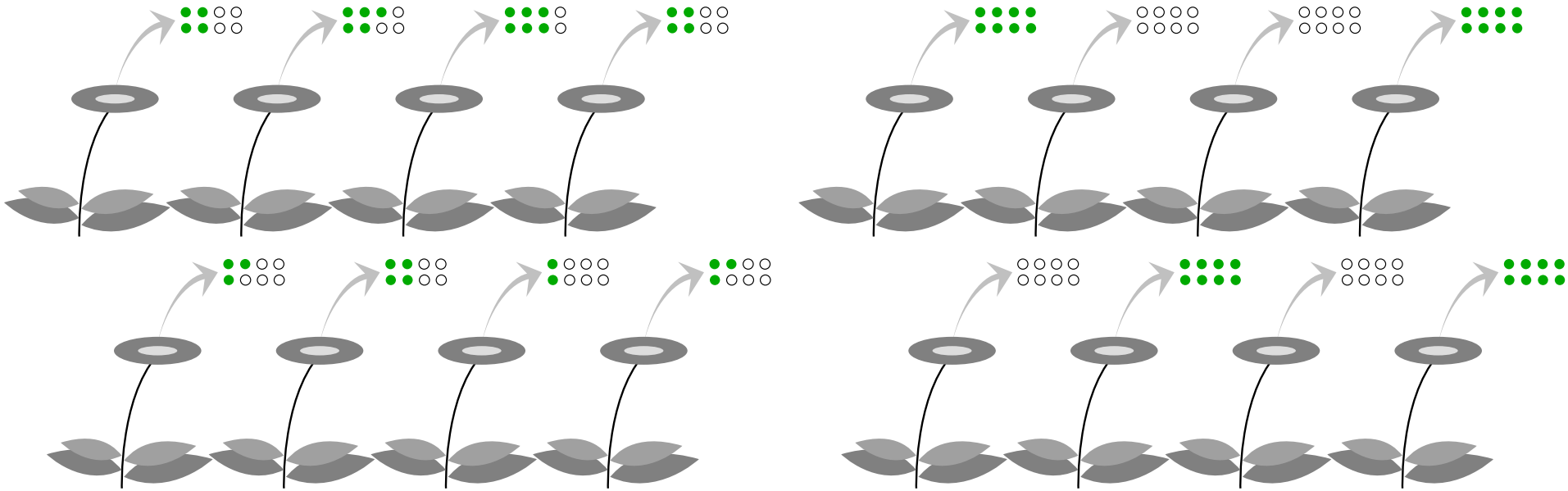


GLM では説明できないばらつき!



○が観測されたデータの図示

過分散 (overdispersion) とは何か?

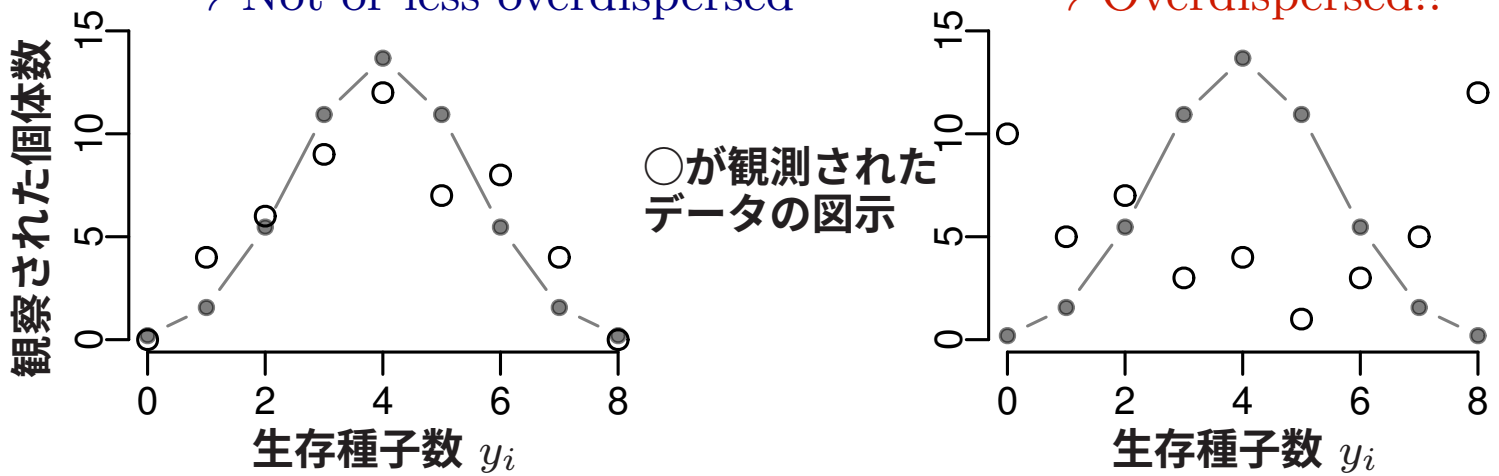


(A) 個体差のばらつきが小さい場合

→ Not or less overdispersed

(B) 個体差のばらつきが大きい場合

→ Overdispersed!!



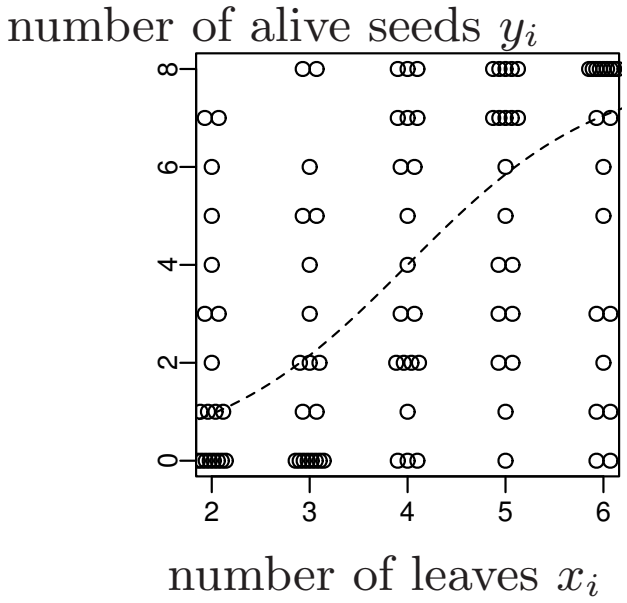
ロジスティック回帰やポアソン回帰といった GLM では全サンプルの均質性を仮定している

現実のカウントデータは，多くの場合「過分散」

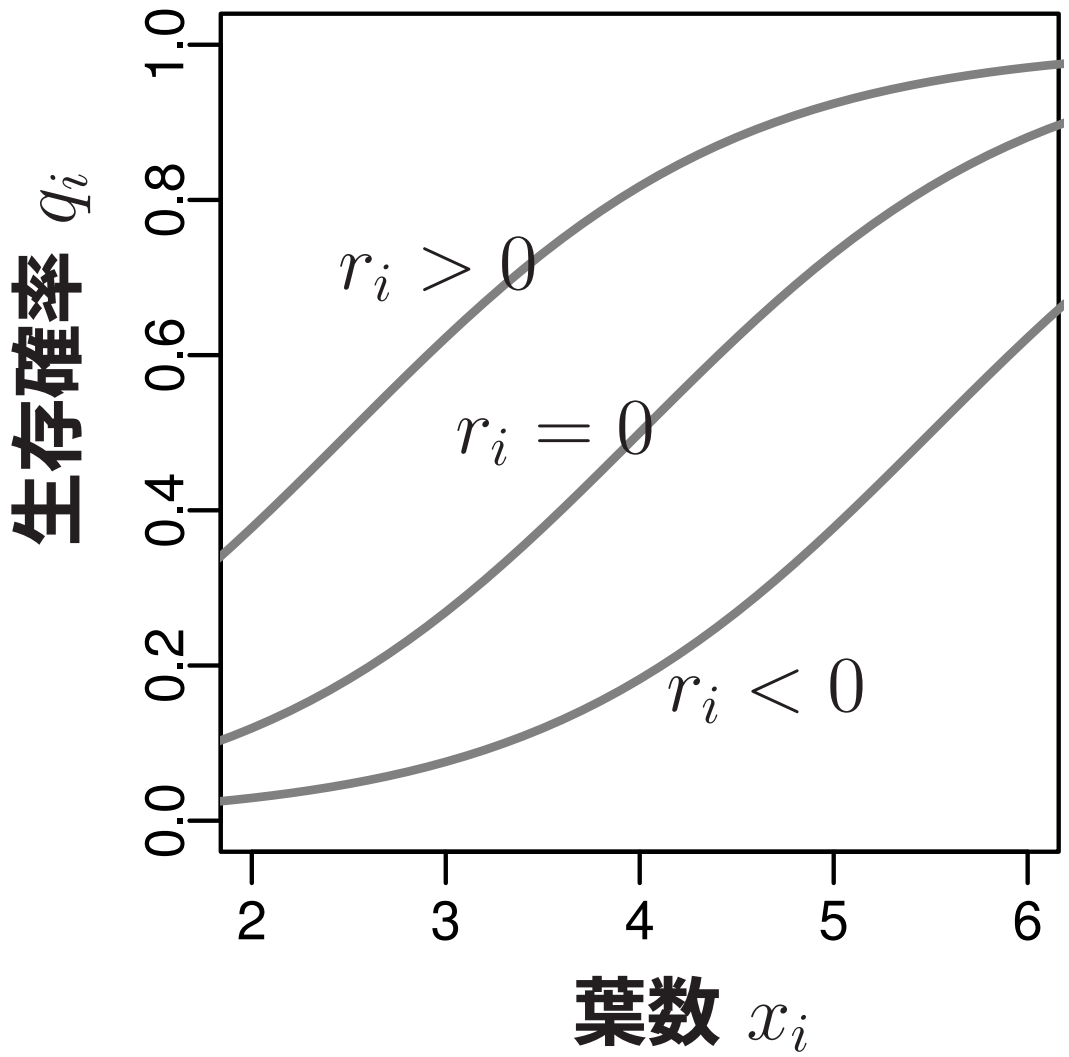
ロジスティック回帰のモデルを改良する

ロジスティック回帰のモデル

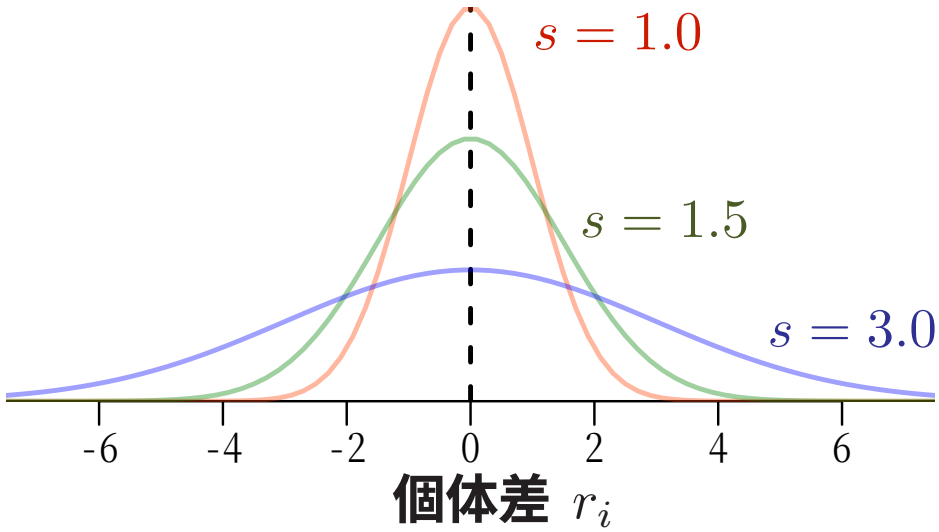
- **確率分布**: **二項分布**
- **線形予測子**: $\beta_1 + \beta_2 x_i + r_i$
- **リンク関数**: logit **リンク関数**



個体 i の個体差を r_i としてみよう



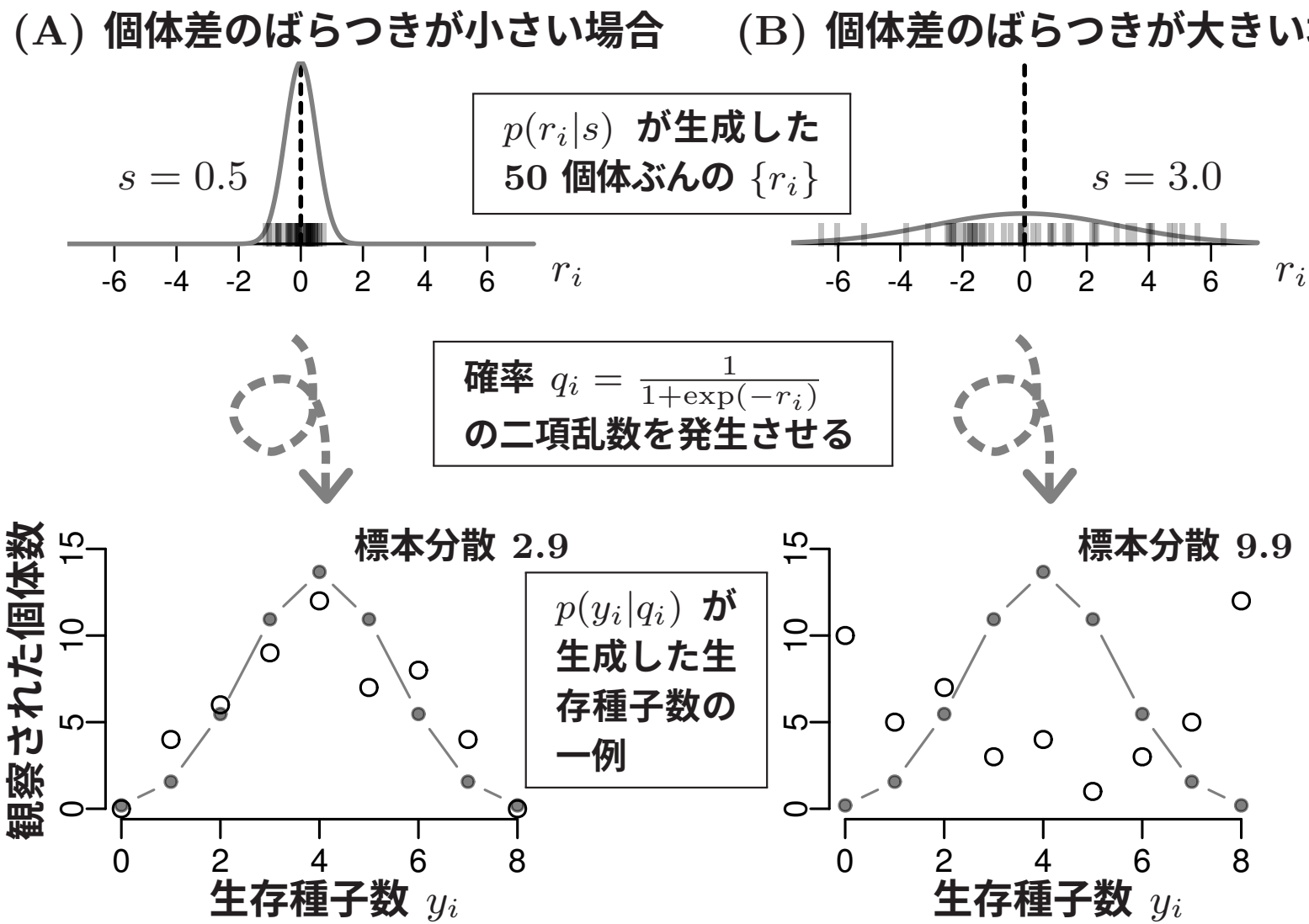
$\{r_i\}$ のばらつきは正規分布だと考えてみる



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

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

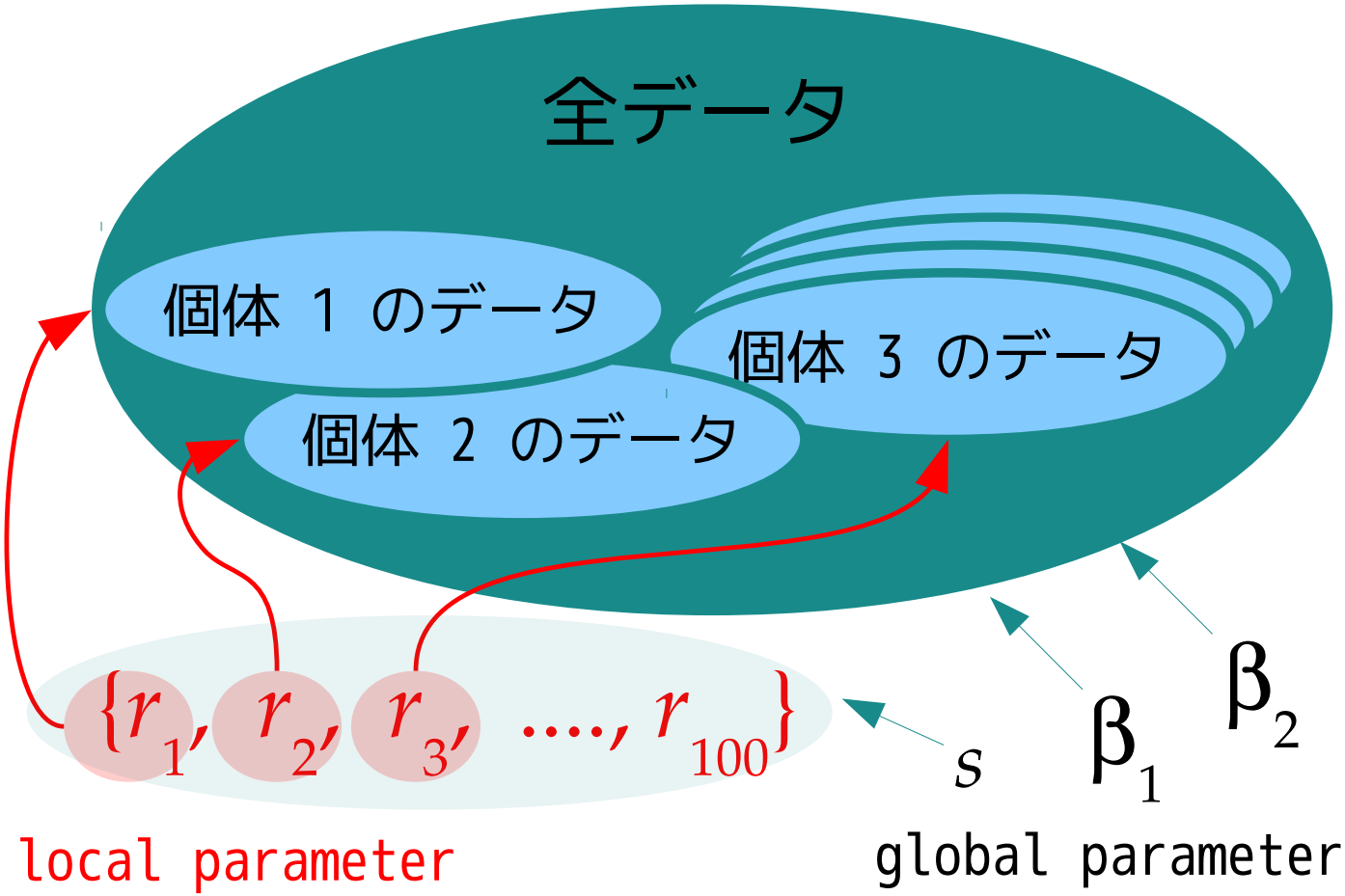
Generalized Linear Mixed Model (GLMM)

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**

2. 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

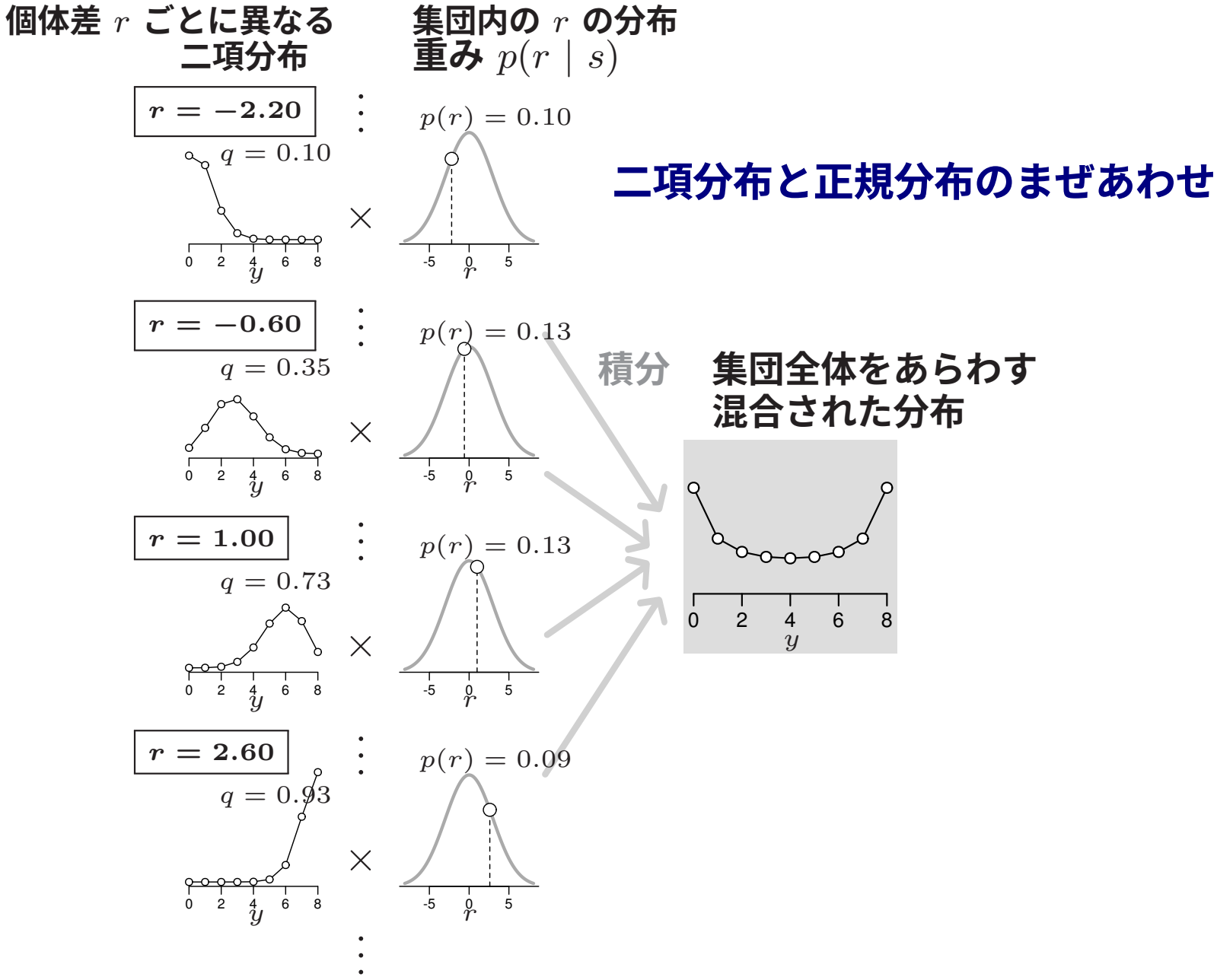
$$L_i = \int_{-\infty}^{\infty} p(y_i \mid \beta_1, \beta_2, r_i)\ p(r_i \mid s) dr_i$$

全データの尤度 — β_1, β_2, s の関数

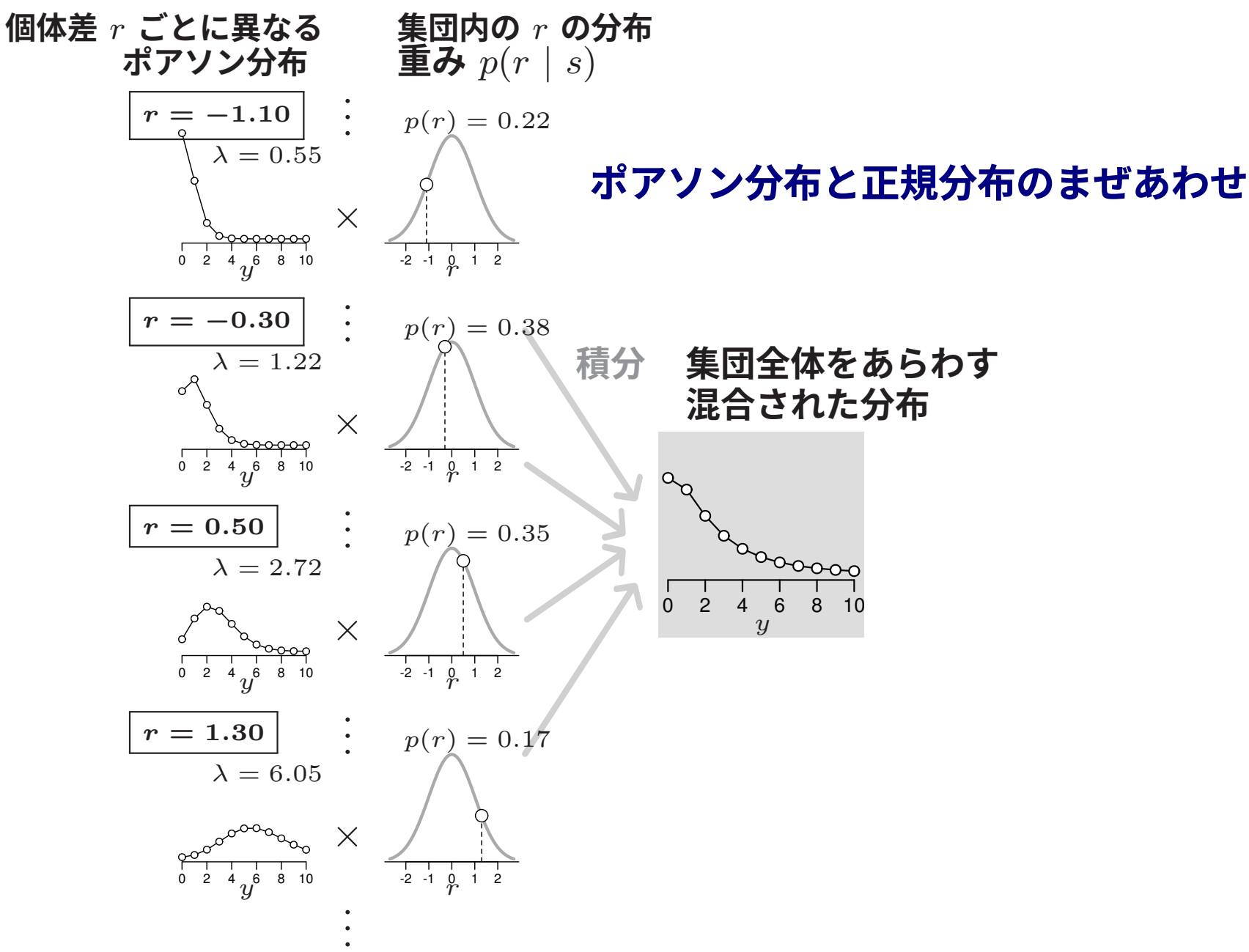
$$L(\beta_1, \beta_2, s) = \prod_i L_i$$

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: $\hat{\beta}_1, \hat{\beta}_2, \hat{s}$

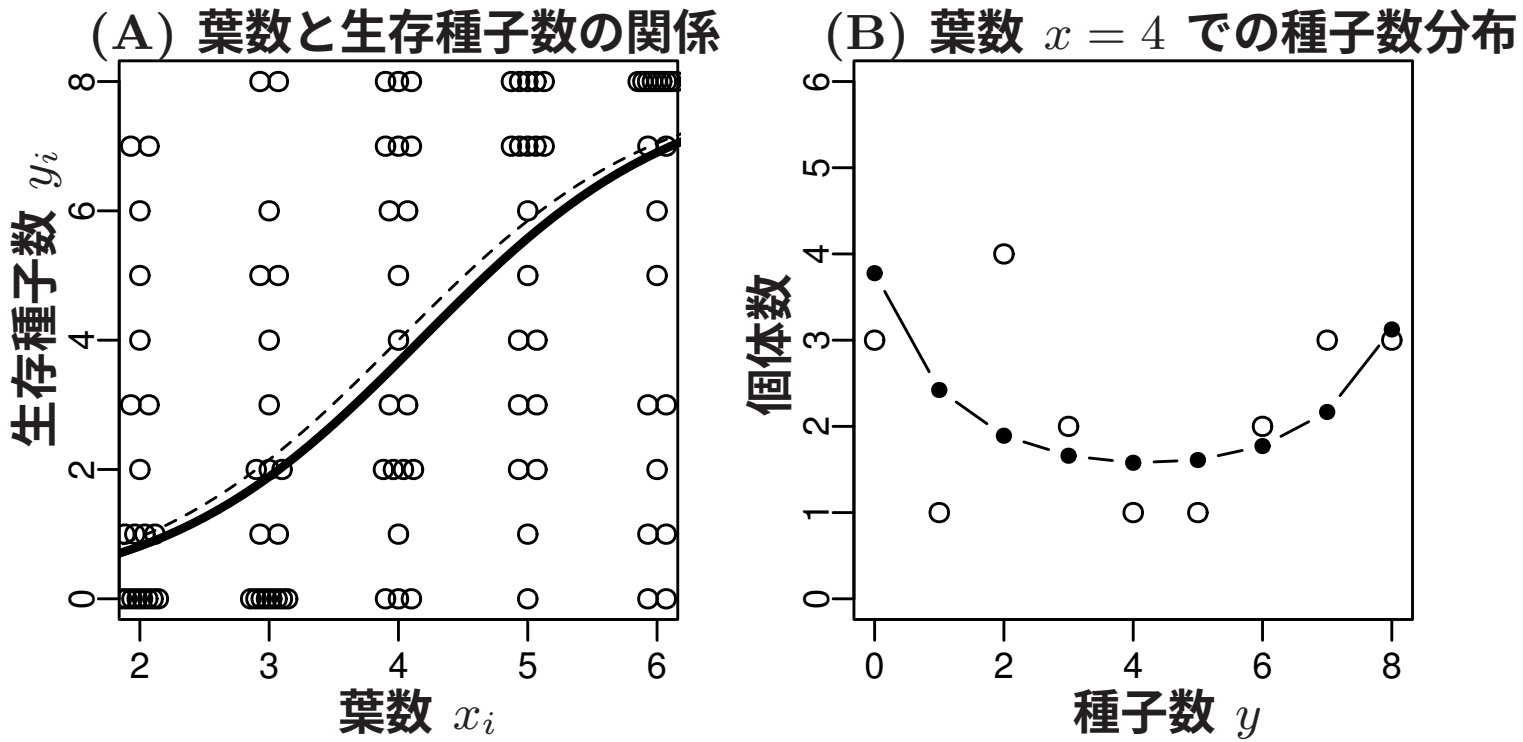
```
> glmmML(cbind(y, N - y) ~ x, data = d, family = binomial,
+ cluster = id)
...(snip)...
             coef se(coef)     z Pr(>|z|)
(Intercept) -4.13    0.906 -4.56  5.1e-06
x            0.99    0.214  4.62  3.8e-06

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

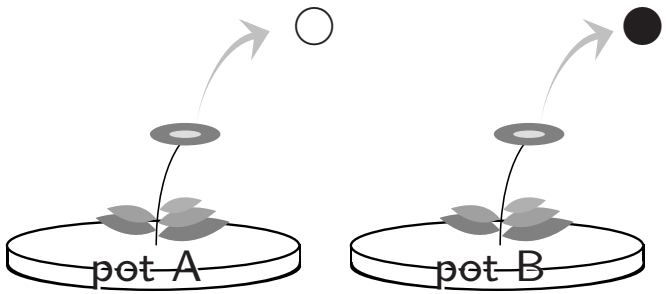


3. 現実のデータ解析には GLMM が必要

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**

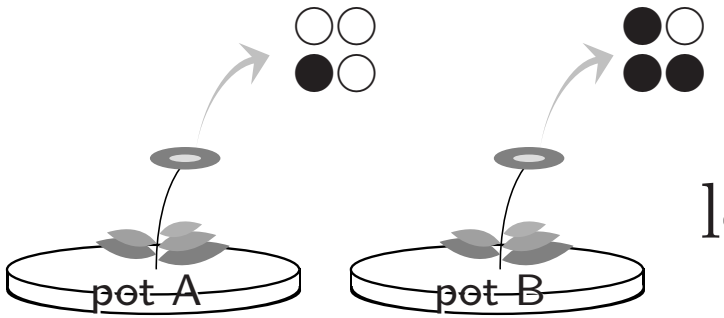


個体差も植木鉢差も推定できない

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

q_i: **種子の生存確率**

(B) **個体は擬似反復，植木鉢は反復**



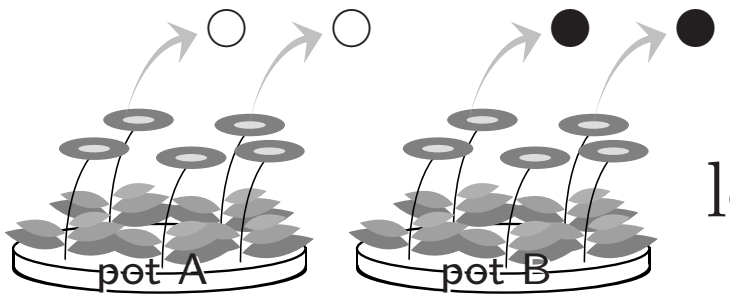
個体差は推定できる
植木鉢差は推定できない

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

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

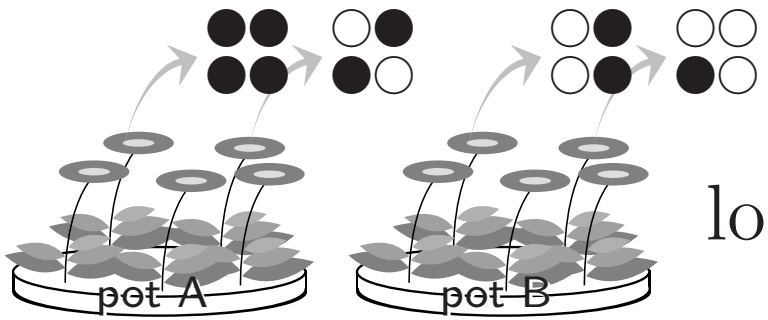
(C) 個体は反復，植木鉢は擬似反復



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 ……