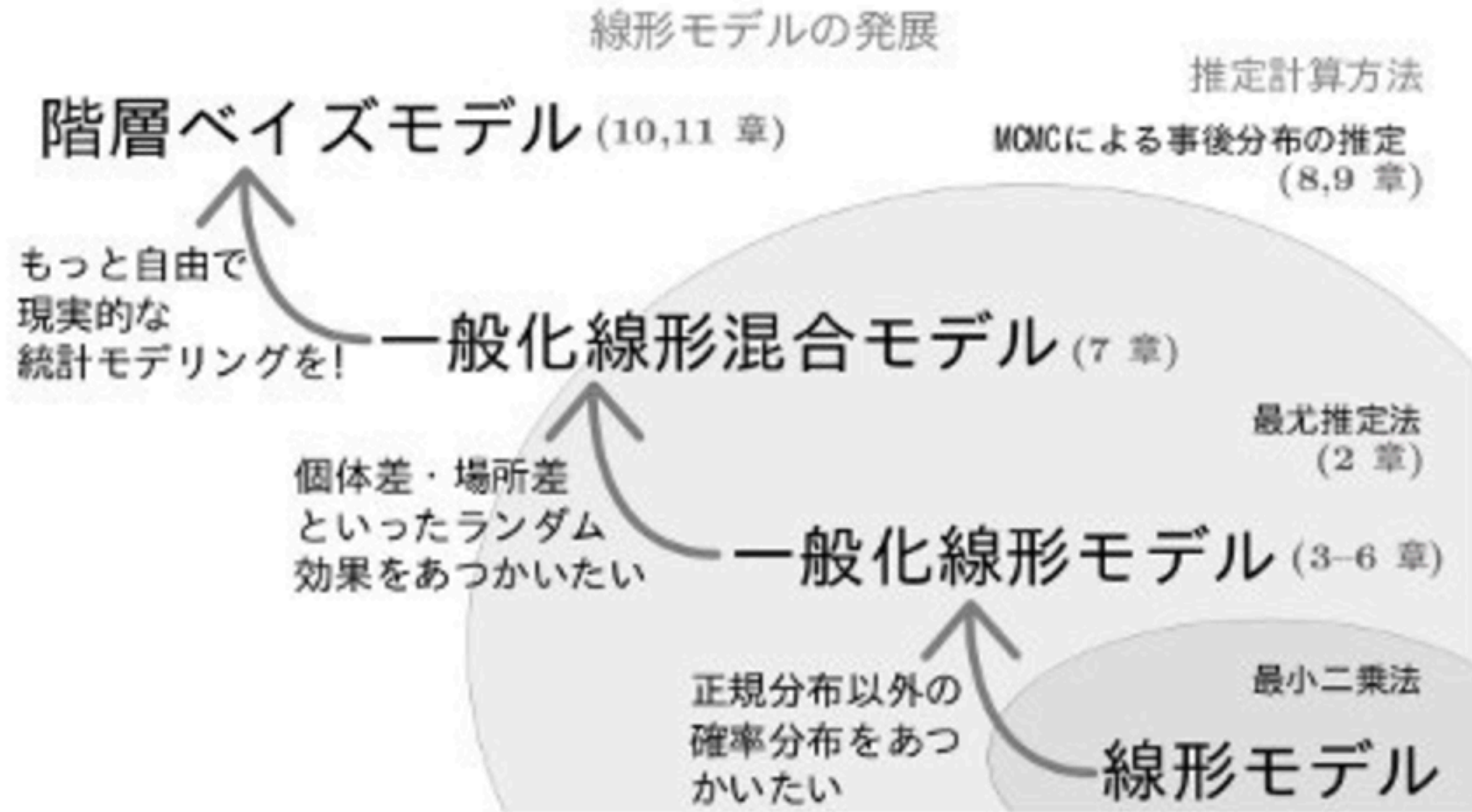


# 統計勉強会

## 7章 一般化線形混合モデル (GLMM) 個体差のモデリング

清水 翔太郎

これまでとこれから



# GLMからGLMMへ

- GLMは現実のデータ解析には対応しきれていない
- 理由：GLMの「説明変数以外は全部均質」という仮定はたいていの現実のデータ（例えば自然界のデータ）には合致しないから
- GLMM（generalized linear mixed model）は、「人間が測定できない・測定しなかった個体差」を組み込んだGLM
- 個体差・場所差：データ化されていない原因不明の差異
  - 説明変数として扱えるような数量はここでは個体差とは呼ばない

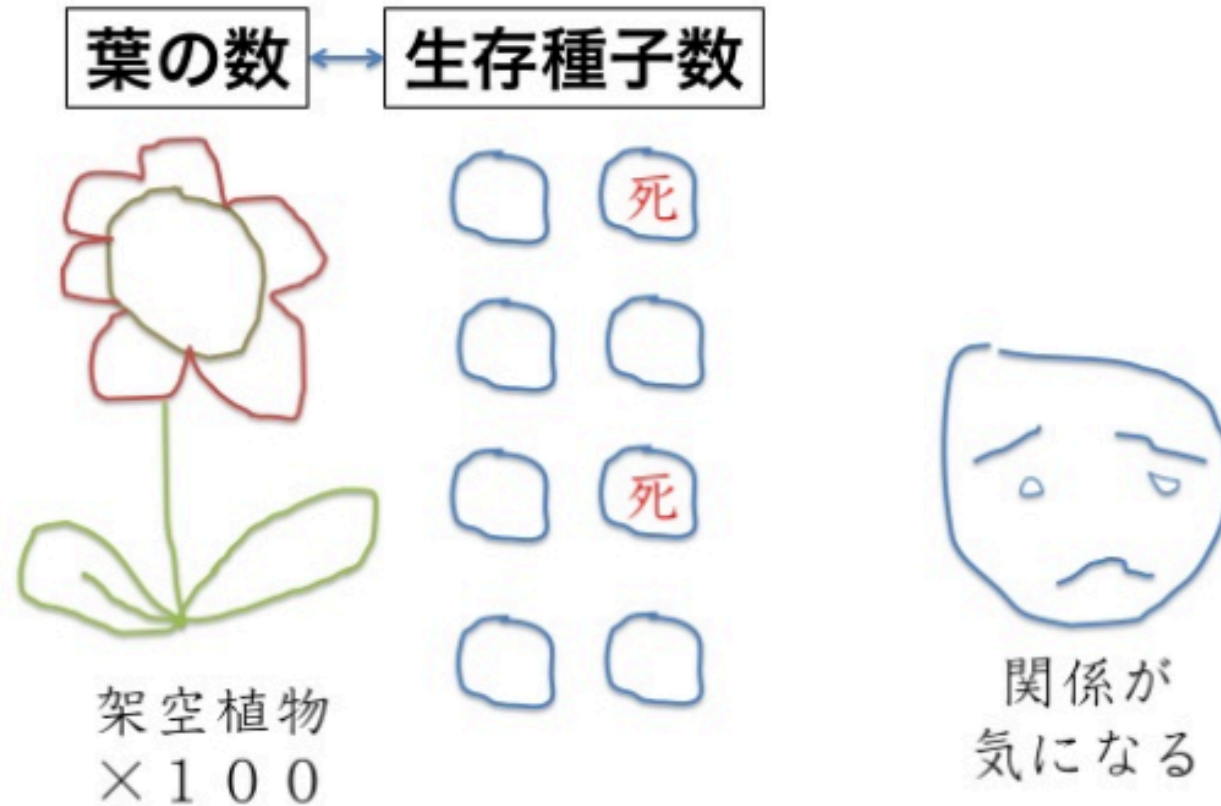
## 7.1 例題：GLMでは説明できないカウントデータ

## 例題：種子の生存確率

- 架空植物の各個体から8個の種子を取ってくる
- 生存種子数が葉数とともにどのように増大するか
- 100個体（葉数2枚が20個体、葉数3枚が20個体、…、計100個体）
- 種子の生存数 $y$ 個（ $0 \leq y \leq 8$ 、整数）
- 葉数 $x$ 枚（ $2 \leq x \leq 6$ 、整数）

# 例題：種子の生存確率

## 図にするとこうなる



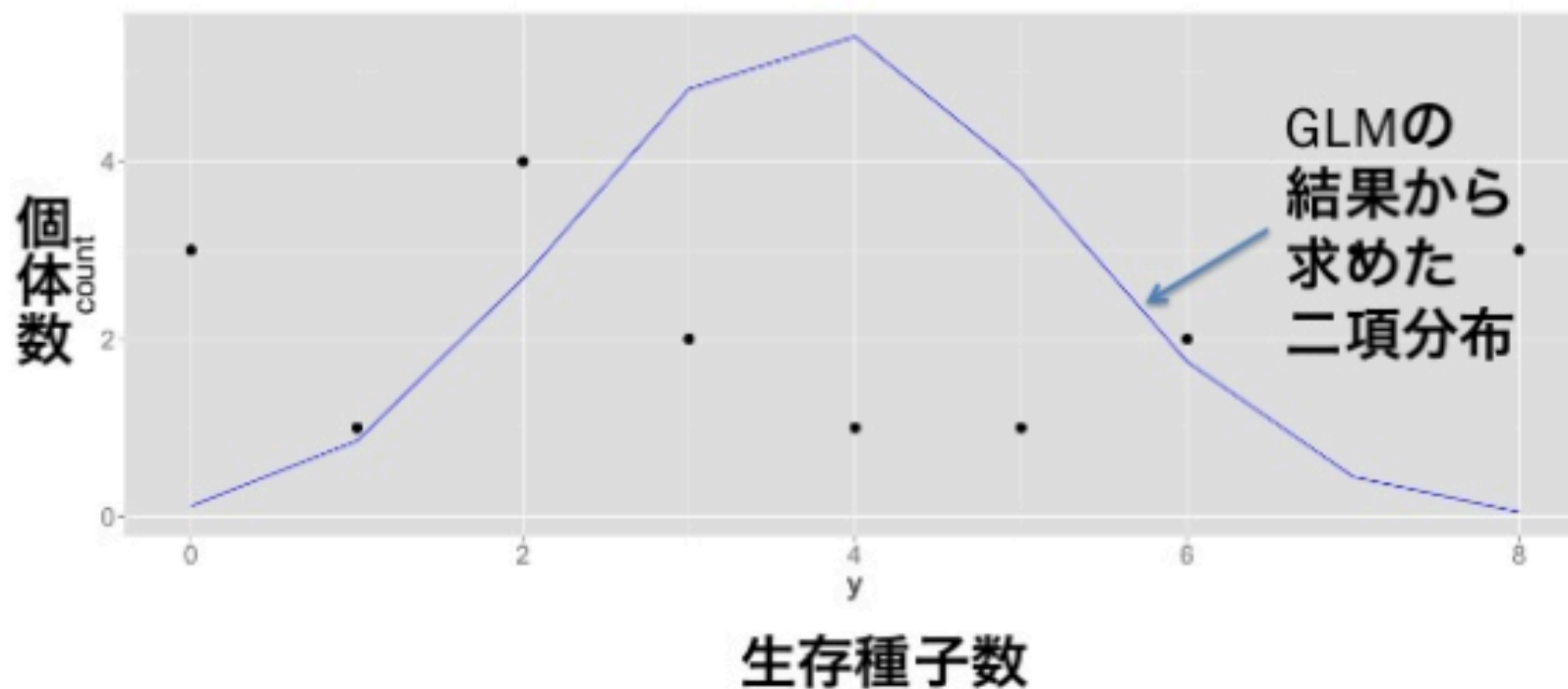
# データの可視化 ～ GLMを用いたモデリング

一旦、Rに移ります

# 実際の分布とGLMから求めた二項分布

## 全然二項分布じゃない

葉の数4枚の場合の生存種子数と個体数の関係





## 7.2 過分散と個体差

# 過分散と個体差

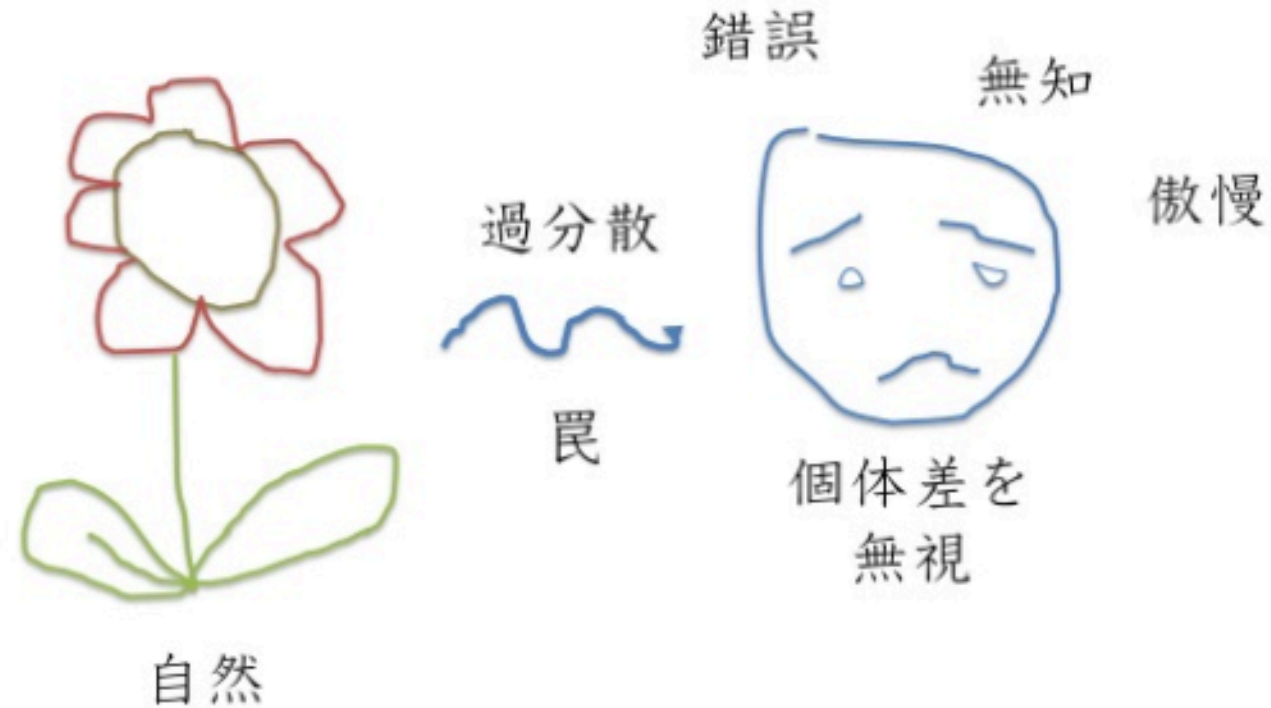
- 過分散：データから得られる分散が平均から推定される分散（二項分布で期待される分散）に比べて大きすぎる状況
- 過分散が起きるのは観測できていない個体差があるから

# 観測されていない個体差がもたらす過分散の極端な例

- 何らかの個体差によって確率分布と実際のデータが乖離する例
  - 平均生存種子数は4個だが、全個体の半数のは生存種子数が0個、残りの半分は生存種子数が8個
    - 二項分布に従っていると仮定すると4個付近が一番多くなるはず
  - 平均点は50点だが、クラスの半分は0点、残りの半分は100点
    - 正規分布に従っていると仮定すると50点付近が一番多くなるはず

# 例題：種子の生存確率

## 図にするとこうなる



# 観測されていない個体差の例

➤ 遺伝子、年齢 -> 個体差

➤ 生育環境 -> 場所差



今後はまとめて個体差と呼ぶ

※今回の例の場合、葉の数は個体によって異なるが個体差とは呼ばない

➤ 個体差の条件：個体によって異なること & 観測されていないデータであること

• 原因を全て明らかにすることは不可能だが、影響は取り込みたい

## 7.3 一般化線形混合モデル

# GLMMの概要

固定効果

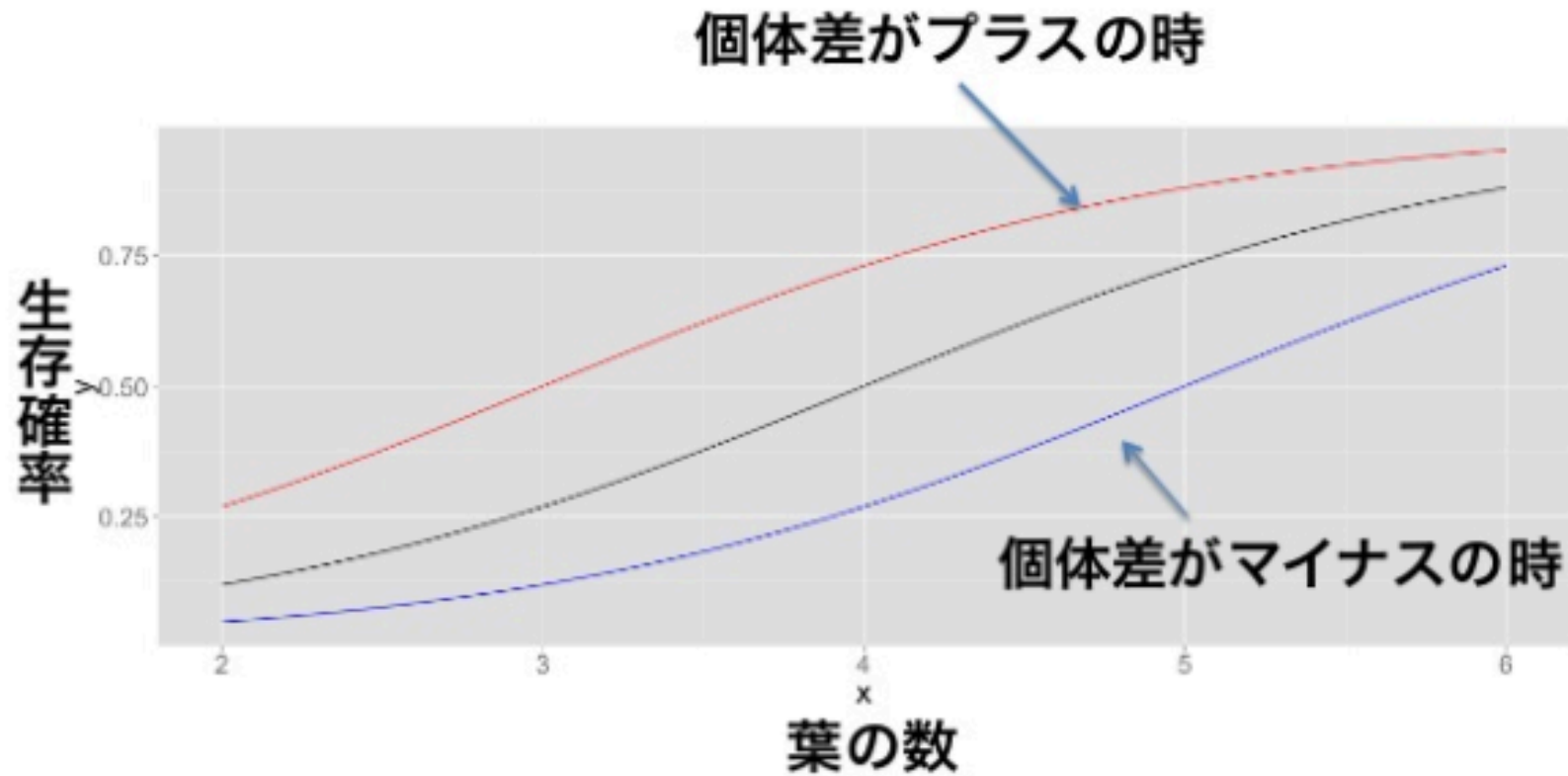
ランダム効果

$$\text{logit}(q_i) = \beta_1 + \beta_2 x_i + r_i$$

個体差として追加  
個体間で独立した正規分布  
平均は0  
標準偏差はsとして任意に設定

# GLMMの概要

## 個体差で生存確率は変わる





# ランダム切片モデルとランダム傾きモデル

- ランダム切片モデル

- 個体差によって切片が変化する -> グラフが縦に平行移動するような感覚

- ランダム傾きモデル

- 個体差によって傾きが変化する -> グラフの勾配が回転するような感覚

# GLMM豆知識

- 様々な呼び名が存在する（らしい）
  - 階層線形モデル（HLM）
  - マルチレベルモデル
  - ランダム効果モデル
  - 成長曲線モデル

次回：ランダム効果の確率分布・最尤推定 etc.

- 力尽きました