

# 統計モデリング入門 2015 (a)

An Introduction to Statistical Modeling

## 観測されたパターンを説明する統計モデル

久保拓弥 (北海道大・環境科学)

[kubo@ees.hokudai.ac.jp](mailto:kubo@ees.hokudai.ac.jp)

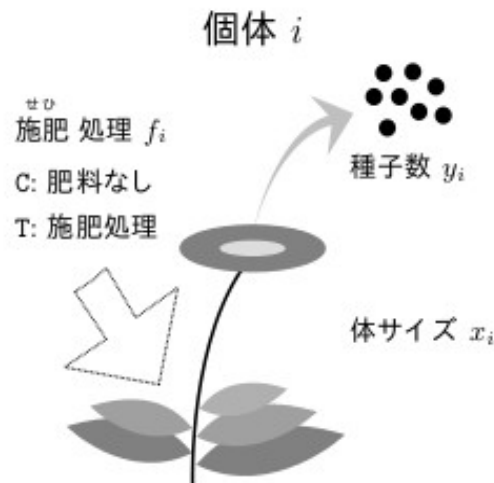


図 3.1 この例題に登場する架空植物の第  $i$  番目の個体. この植物の体サイズ (個体の大きさ)  $x_i$  と肥料をやる施肥処理  $f_i$  が種子数  $y_i$  にどう影響しているのかを知りたい.

The main language of this class is  
Japanese ... Sorry

- Why in Japanese? ... because even in Japanese, **statistics is difficult** for Japanese students to understand.
- I will **compensate for language disadvantages** in foreign students when I give grades.
- **Questions in English are always welcomed!**

# 統計モデリング授業の web page

<http://goo.gl/76c4i>

植物生態学特論 I (Advanced Course of Plant Ecology I)

生態学の統計モデリング 7月6日から

Statistical Modeling for Ecology, commence on July 6

13:00 - 14:30, Monday and Wednesday

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

- 授業メーリングリスト (Course Mailing List)

- <http://goo.gl/f0vCn8> で早めに登録してください。登録のユーザーインターフェイスは日本語に変更できます。授業の資料ダウンロードの連絡などします。
- subscribe at <http://goo.gl/f0vCn8> immediately, or you can not download course handouts.

- 授業 Web Site: <http://goo.gl/76c4i>

# 統計モデリング授業 Mailing List

<http://goo.gl/f0vCn8>

植物生態学特論 I (Advanced Course of Plant Ecology I)

生態学の統計モデリング 7月6日から

Statistical Modeling for Ecology, commence on July 6

13:00 - 14:30, Monday and Wednesday

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

- 授業メーリングリスト (Course Mailing List)

- <http://goo.gl/f0vCn8> で早めに登録してください。登録のユーザーインターフェイスは日本語に変更できます。授業の資料ダウンロードの連絡などします。
- subscribe at <http://goo.gl/f0vCn8> immediately, or you can not download course handouts.

- 授業 Web Site: <http://goo.gl/76c4i>

# この統計モデリング授業の Mailing List (ML) **kubostat**

- ML を使って各回の「課題」を出します
  - 回答もメールで送信してください
- 成績評価は「課題」の回答
  - 出欠関係なし（欠席の連絡いりません）
- 単位とらない人も ML 登録してください
  - 講義資料のダウンロード案内などあります

# Performance Rating

- E-mail assignment (via Mailing List)
  - That's ALL!
- Attendance? NOT care.

What for Statistical  
Modeling?

なぜデータ解析の方法を  
勉強しなければならないのか？

# All you depend on statistics

whenever you conclude something based on your data

- データ解析がおかしいと結論もおかしい
- Crazy data analysis → Crazy results
- 統計解析わからんと批判的に読めない
- A lack of statistical knowledge →  
no critical reading of papers




データ解析はあまり重視されてなかった  
内容がわからなくてもソフトウェアにまるなげ

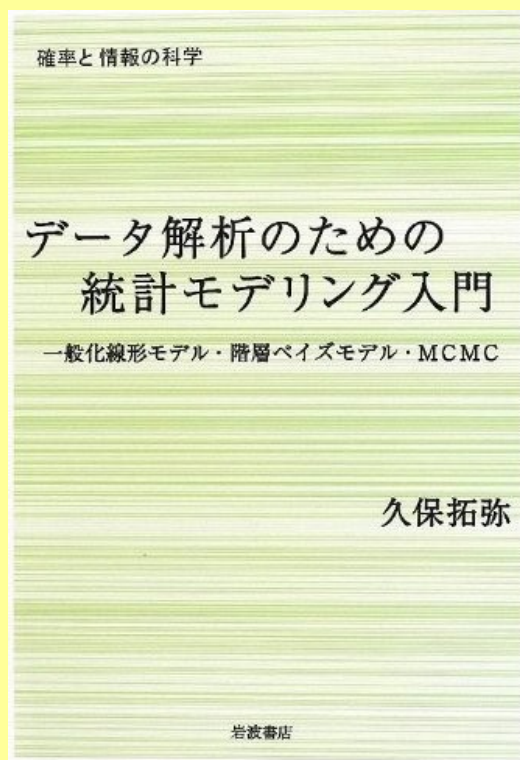
- ブラックボックス統計解析
  - No “Blackbox” statistics!
- とにかく「ゆーい差」さえ出せばよいという発想になっている
  - Don't blindly believe “Significance” !

# この授業のねらい (aim)

できるだけ内容を理解して統計ソフトウェアを使おう!

- Understand how to fit statistical models to your data データにあてはめられる統計モデルを作ろう
- Use the statistical software  to show your data structure

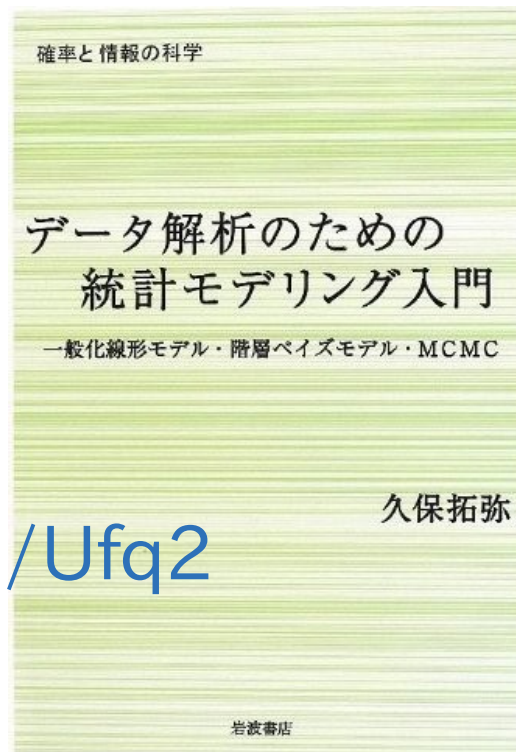
# 教科書とソフトウェア



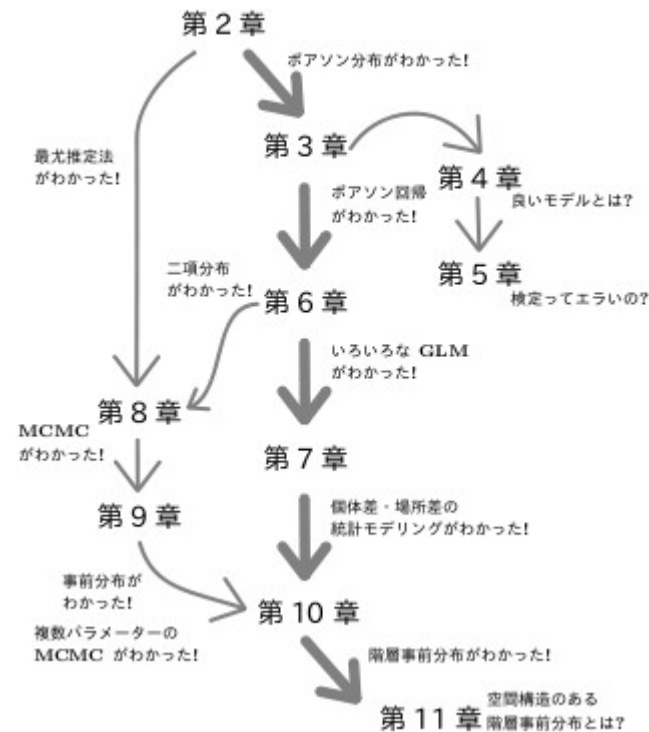
# この授業は「統計モデリング入門」 にそった内容を説明します

my text book (in Japanese)  
(第 2 章以降の説明の流れ)

著者: 久保拓弥  
出版社: 岩波書店  
2012-05-18 刊行  
価格 3990 円



<http://goo.gl/Ufq2>



割引販売 3000 円!!

# 「統計モデリング入門」のもとになった「講義のーと」もあります

北海道大学学術成果コレクション  
HUSCAP  
Hokkaido University collection of Scholarly and Academic Papers  
Copyright(c) 2005 Hokkaido University Library, All Rights Reserved.

北海道大学 | 附属図書館 | HUSCAP

検索 Language: 日本語

Hokkaido University Collection of Scholarly and Academic Papers >  
環境科学院・地球環境科学研究所 >  
雑誌発表論文等 >

コレクション一覧  
研究科等一覧  
紀要・研究報告集  
学位論文  
開催学会会議録  
開催イベント  
歴史的に貴重な文献  
学協会  
よく読まれる文献  
学内の方へ  
HUSCAPで著作を公開するには

フルテキスト

<a href="#">kubostat2008a.pdf</a>	第1回	260.69 kB	PDF	<a href="#">見る/開く</a>
<a href="#">kubostat2008b.pdf</a>	第2回	156.71 kB	PDF	<a href="#">見る/開く</a>
<a href="#">kubostat2008c.pdf</a>	第3回	434.56 kB	PDF	<a href="#">見る/開く</a>
<a href="#">kubostat2008f.pdf</a>	第6回	219.43 kB	PDF	<a href="#">見る/開く</a>
<a href="#">kubostat2008g.pdf</a>	第7回	246.95 kB	PDF	<a href="#">見る/開く</a>
<a href="#">kubostat2008e.pdf</a>	第5回	238.75 kB	PDF	<a href="#">見る/開く</a>
<a href="#">kubostat2008d.pdf</a>	第4回	184.92 kB	PDF	<a href="#">見る/開く</a>

タイトル: 講義のーと : データ解析のための統計モデリング  
著者: 久保, 拓弥  
キーワード: 生態学  
データ解析

授業 web page に「講義のーと」へのリンクがあります! <http://goo.gl/82dgC>

# 統計ソフトウェア R



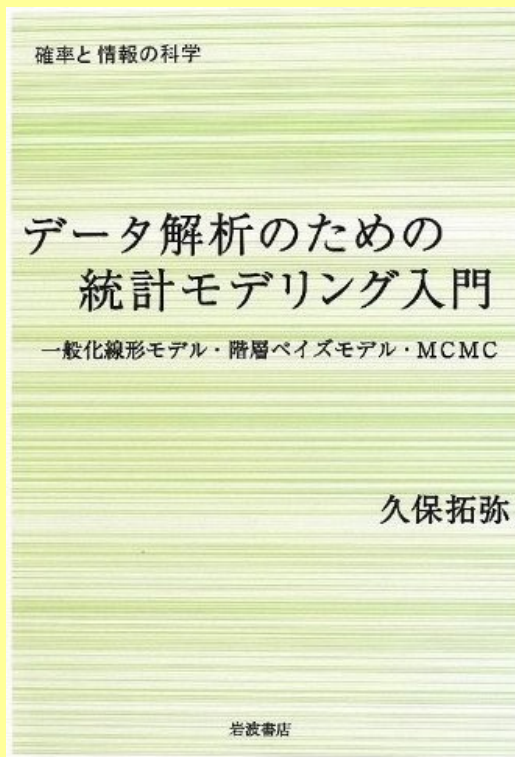
統計学の勉強には良い統計ソフトウェアが必要!

- 無料で入手できる
- 内容が完全に公開されている
- 多くの研究者が使っている
- 作図機能が強力

この教科書でも R を  
使って問題を解決する  
方法を説明しています



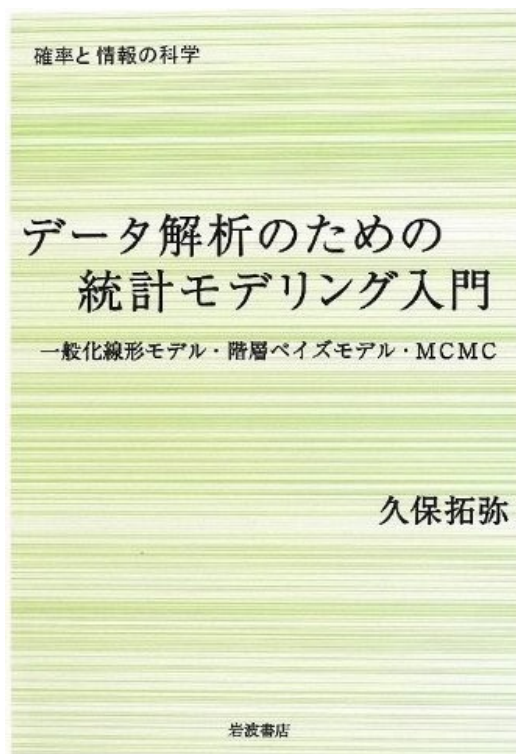
# 統計モデルとは何か？



# 「統計モデル」とは何か？

どんな統計解析においても  
統計モデルが使用されている

- 観察によってデータ化された現象を説明するために作られる
- 確率分布が基本的な部品であり、これはデータにみられるばらつきを表現する手段である
- データとモデルを対応づける手つづきが準備されていて、モデルがデータにどれぐらい良くあてはまっているかを定量的に評価できる





# 「統計モデリング入門」の主張

「何でも正規分布」じゃないだろ!

## 線形モデルの発展

推定計算方法

MCMC

階層ベイズモデル

もっと自由な  
統計モデリン  
グを!

一般化線形混合モデル

最尤推定法

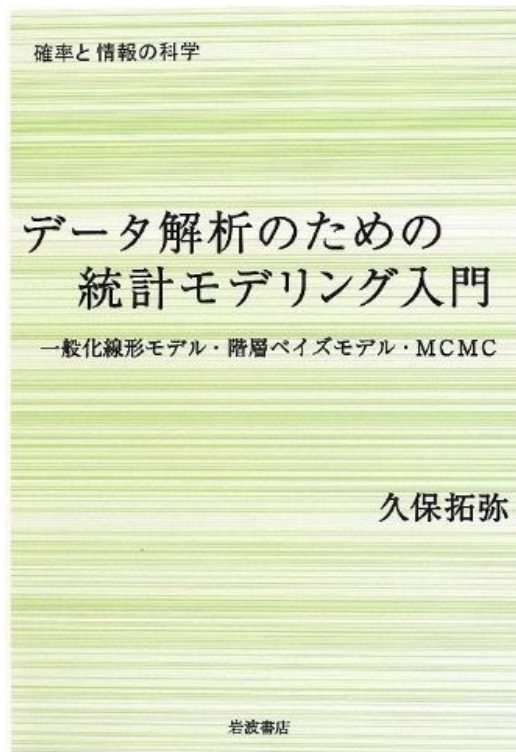
個体差・場所差  
といった変量効果  
をあつかいたい

一般化線形モデル

正規分布以外の  
確率分布をあつ  
かいたい

最小二乗法

線形モデル



# 「統計モデリング入門」の主張

right probability distribution  
for right statistical modeling

## The Evolution of Linear Models

Hierarchical Bayesian Model  
(HBM)

Parameter  
Estimation  
MCMC

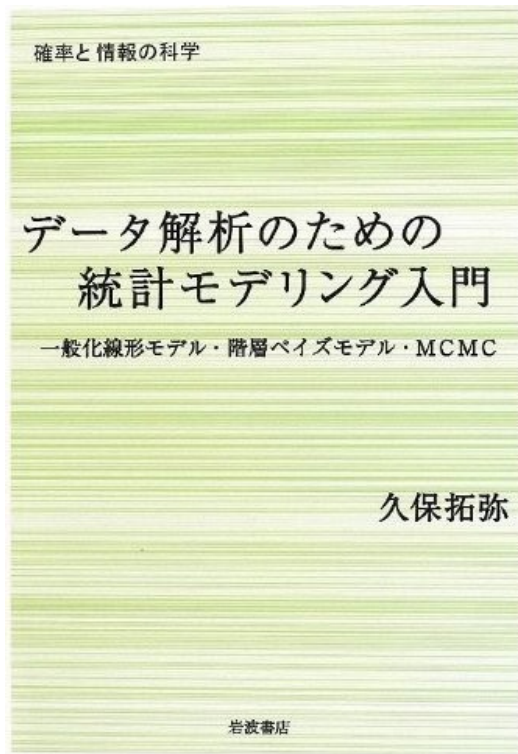
Generalized Linear Mixed Model  
(GLMM)

Generalized  
Linear Model (GLM)

MLE

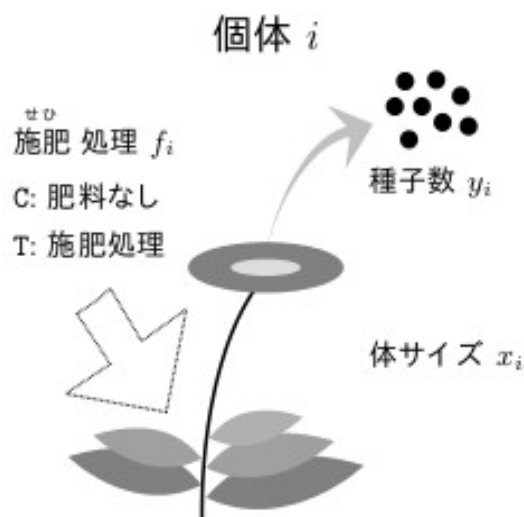
MSE

Linear Model



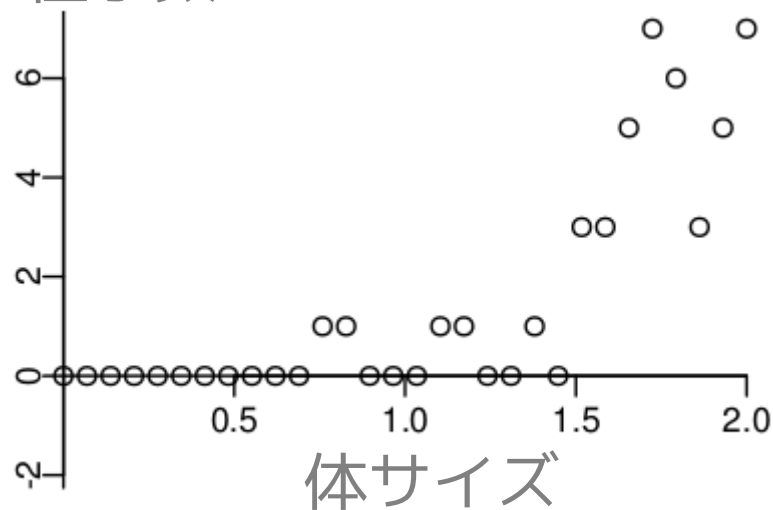
# たとえばこんなデータがあったしましょう

An example  
(次の時間の例題)



number of seeds

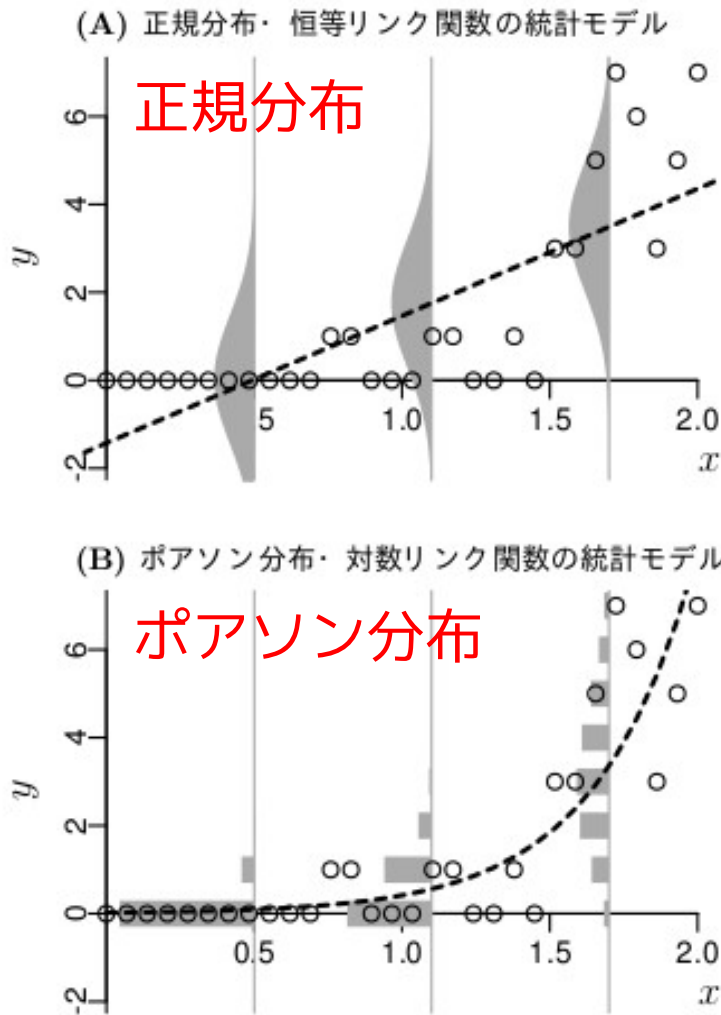
種子数



plant body size

# 一般化線形モデル - ばらつきをよく見る

Don't use the normal distribution  
without seeing data!



線形モデルの発展

階層ベイズモデル

もっと自由な  
統計モデリン  
グを!

一般化線形混合モデル

推定計算方法  
MCMC

個体差・場所差  
といった変量効果  
をあつかいたい

一般化線形モデル

最尤推定法

正規分布以外の  
確率分布をあつ  
かいたい

最小二乗法  
線形モデル

0 個, 1 個, 2 個と数えられる種子数が  
「正規分布」なわけないだろ!!

3.9 回帰モデルと確率分布の関係. また別の架空データに対して GLM をあてはめた例. 破線は  $x$  とともに変化する平均値. グレイで

# 全体の流れ (1/2)

第 1 回: 7/06 (月) 観測されたパターンを説明する統計モデル

Introduction

第 2 回: 7/08 (水) 確率分布と最尤推定

Probability Distributions and

Maximum Likelihood Estimation (MLE)

第 3 回: 7/13 (月) 一般化線形モデル: ポアソン回帰

Generalized Linear Model (GLM):

Poisson Regression

# 全体の流れ (2/2)

第 4 回: 7/15 (水) モデル選択と検定

Model Selection and Statistical Test

第 5 回: 7/22 (水) 一般化線形モデル: ロジスティック回帰

GLM: Logistic Regression

第 6 回: 7/27 (月) 一般化線形混合モデル

Generalized Linear Mixed Model (GLMM)

第 7 回: 7/29 (水) 階層ベイズモデル

Bayesian GLMM and Markov Chain Monte Carlo

# 統計モデリング入門 2015 (b)

probability distribution and maximum likelihood estimation  
確率分布と最尤推定

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

北大環境科学院の講義 <http://goo.gl/76c4i>

2015-07-08

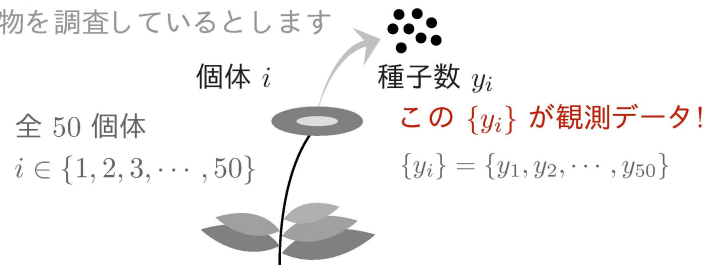
ファイル更新時刻: 2015-07-02 16:53

# 単純化した例題

例題: 種子数の統計モデリング まあ、かなり単純な例から始めましょう

number of seeds per plant individual  
こんなデータ (架空) があったとしましょう

まあ、なんだかこういうヘんな植物を調査しているとします



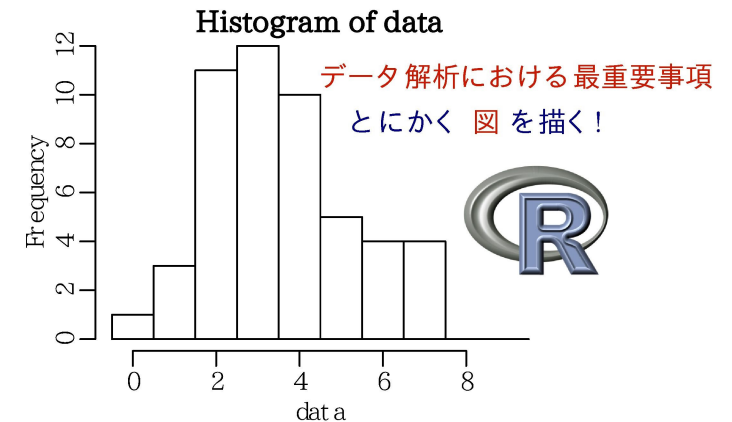
このデータ  $\{y_i\}$  がすでに R という統計ソフトウェアに格納されていた、としましょう

```
> data
[1] 2 2 4 6 4 5 2 3 1 2 0 4 3 3 3 3 4 2 7 2 4 3 3 3 4
[26] 3 7 5 3 1 7 6 4 6 5 2 4 7 2 2 6 2 4 5 4 5 1 3 2 3
```

例題: 種子数の統計モデリング まあ、かなり単純な例から始めましょう

start with data plotting, always  
とりあえずヒストグラムを描いてみる

```
> hist(data, breaks = seq(-0.5, 9.5, 1))
```





# カウントデータはポアソン分布を使って説明できないかを調べる

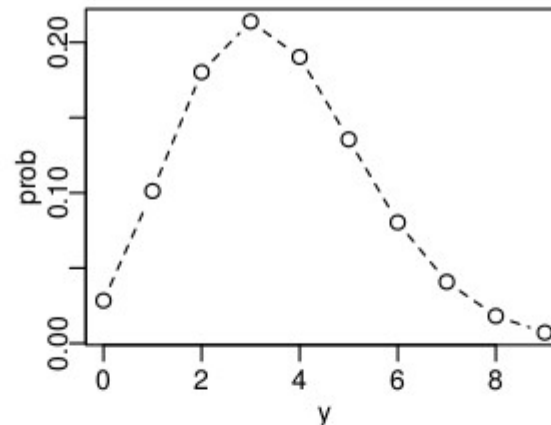


図 4 平均  $\lambda = 3.56$  のポアソン分布. 種子数  $y$  とその確率  $\text{prob}$  の関係が示されている. 図 3 の表を図にしたもの. R の `plot()` 関数の引数, `type = "b"` によって「丸と折れ線による図示」, `lty = 2` によって「折れ線は破線で」と指示している.

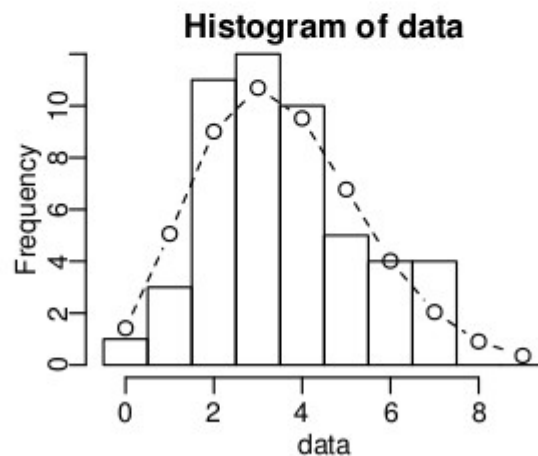
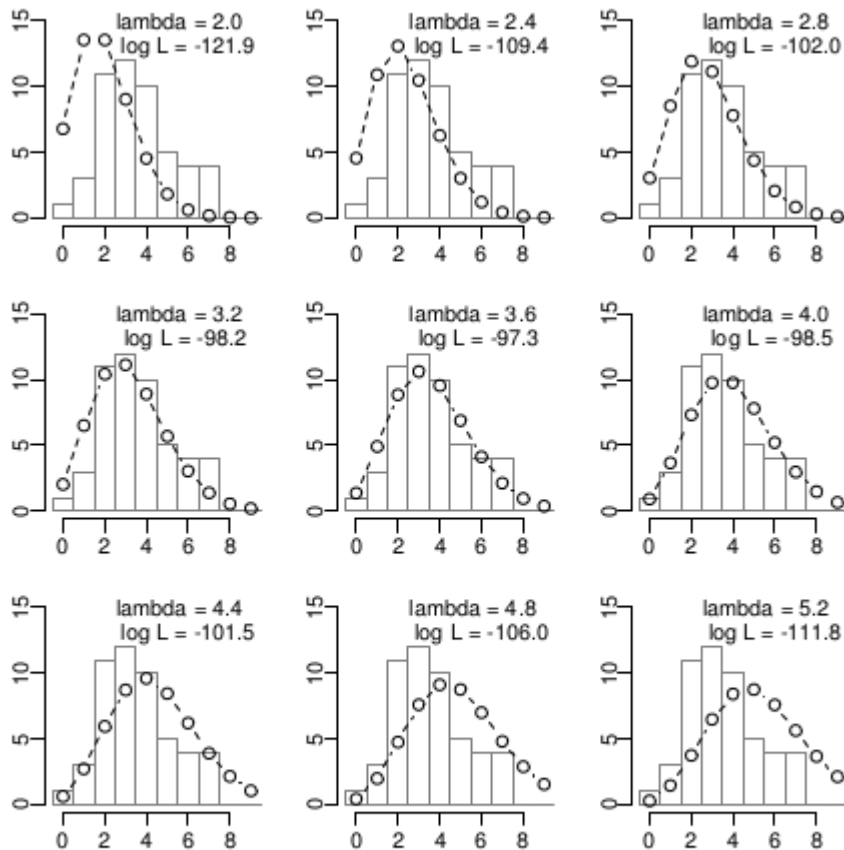


図 5 観測データと確率分布の対応をながめる. ヒストグラムは図 2 と同じ. それに重ねられている丸と破線は  $y$  個の種子をもつ個体数の予測. 平均 3.56 の図 3 のポアソン分布の確率分布に全個体数 50 をかけて得られる.

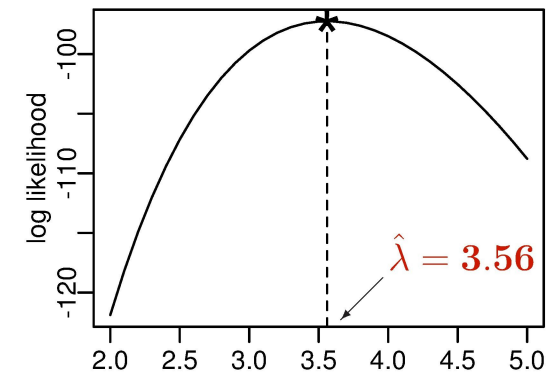
さいゆう

# 最尤推定という考えかたを説明します



ポアソン分布のパラメータの 最尤推定 もっとももらしい推定?  
seek the maximum likelihood estimate,  $\hat{\lambda}$   
対数尤度を最大化する  $\hat{\lambda}$  をさがす

$$\text{対数尤度 } \log L(\lambda) = \sum_i (y_i \log \lambda - \lambda - \sum_k y_i \log k)$$



$$\frac{d \log L}{d \lambda} = 0 \text{ より}$$

図 7 平均  $\lambda$  (lambda) を変化させていったポアソン分布と、観測データへのあてはまりの良さ (対数尤度  $\log L$ )。すべてのヒストグラムは図 2 と同じ。

7/13 (月)

# 統計モデリング入門 2015 (c)

Poisson regression, a generalized linear model (GLM)  
一般化線形モデル: ポアソン回帰

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

北大環境科学院の講義 <http://goo.gl/76c4i>

2015-07-13

ファイル更新時刻: 2015-07-02 16:24

# ここで登場する ---

## 「何でも正規分布」ではダメ! という発想

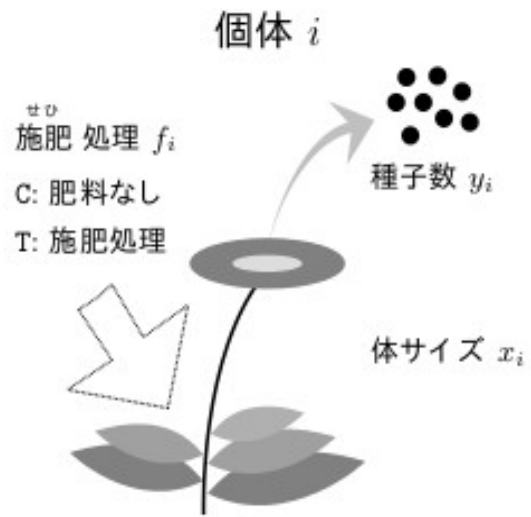


図 3.1 この例題に登場する架空植物の第  $i$  番目の個体. この植物の体サイズ (個体の大きさ)  $x_i$  と肥料をやる施肥処理  $f_i$  が種子数  $y_i$  にどう影響しているのかを知りたい.

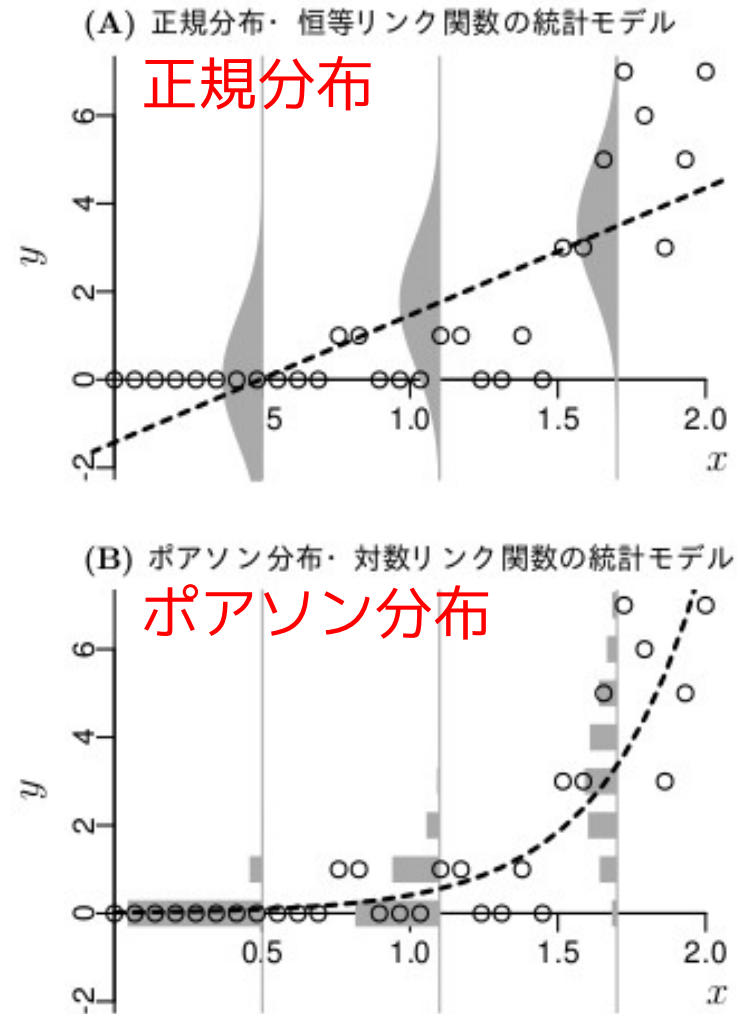


図 3.9 回帰モデルと確率分布の関係. また別の架空データに対して GLM をあてはめた例. 破線は  $x$  とともに変化する平均値. グレイで

# R で統計モデリング

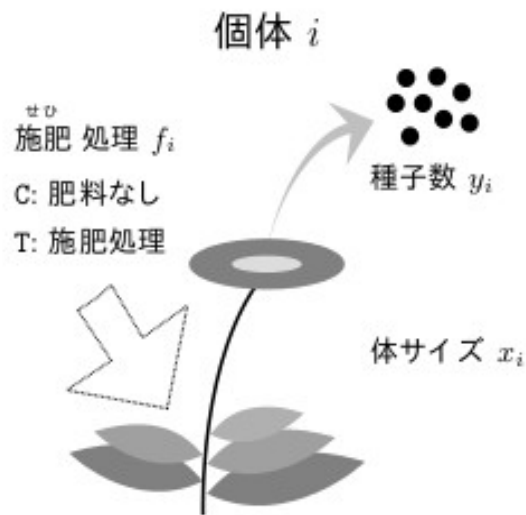


図 3.1 この例題に登場する架空植物の第  $i$  番目の個体  
体サイズ(個体の大きさ)  $x_i$  と肥料をやる施肥処理、  
にどう影響しているのかを知りたい。

結果を格納するオブジェクト

```
fit <- glm(
  y ~ x,
  family = poisson(link = "log"),
  data = d
)
```

関数名: glm  
モデル式: y ~ x  
確率分布の指定: poisson  
リンク関数の指定 (省略可): link = "log"  
data.frame の指定: data = d

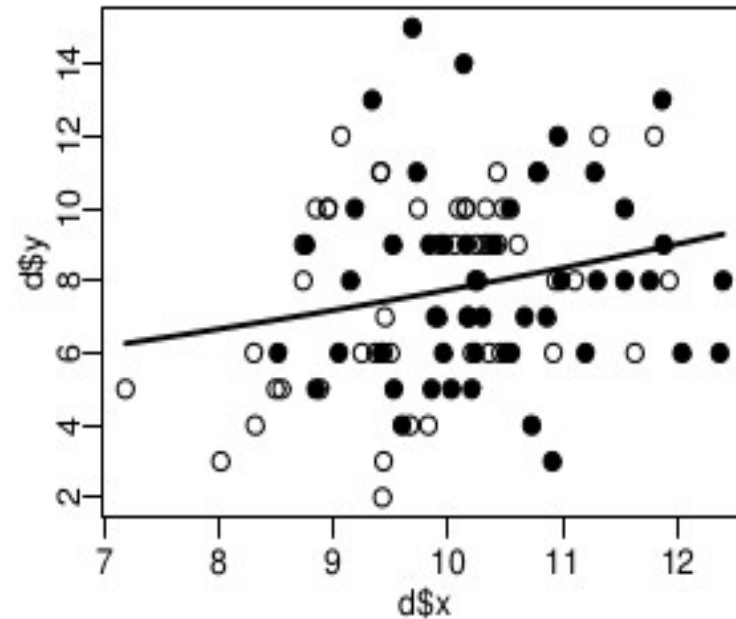


図 17 平均種子数  $\lambda$  の予測. 図 12 に  $\lambda$  の予測値 (実線) を上かきしたもの。

7/15 (水)

# 統計モデリング入門 2015 (d)

モデル選択と検定

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

北大環境科学の講義 <http://goo.gl/76c4i>

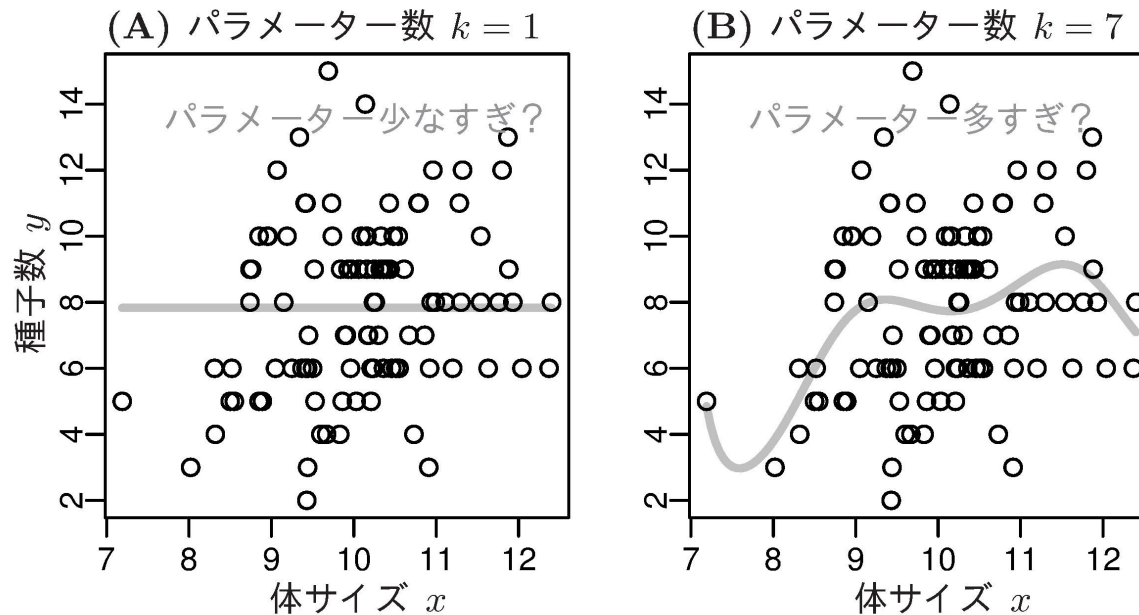
2015-07-15

ファイル更新時刻: 2015-07-02 16:24

# Q. モデル選択とは何か？

もくじ

パラメータ数は多くても少なくてもヘン？



What is the “best?” parameter number  $k$ ?

# A. より良い予測をする統計モデルを探すこと

統計学的な検定 そして、その非対称性  
But their procedures are similar  
しかしモデル選択と検定の手順は途中まで同じ

統計モデルの検定

AICによるモデル選択

←こっちだ!

検定はモデル選択じゃない!

解析対象のデータを確定



データを説明できるような統計モデルを設計

(帰無仮説・対立仮説)

(単純モデル・複雑モデル)



ネストした統計モデルたちのパラメータの さいゆう 最尤推定計算



帰無仮説棄却の危険率を評価



モデル選択規準 AIC の評価



# 統計学って「検定」のこと?

「検定」って何なの?

「検定」ってエラいの?

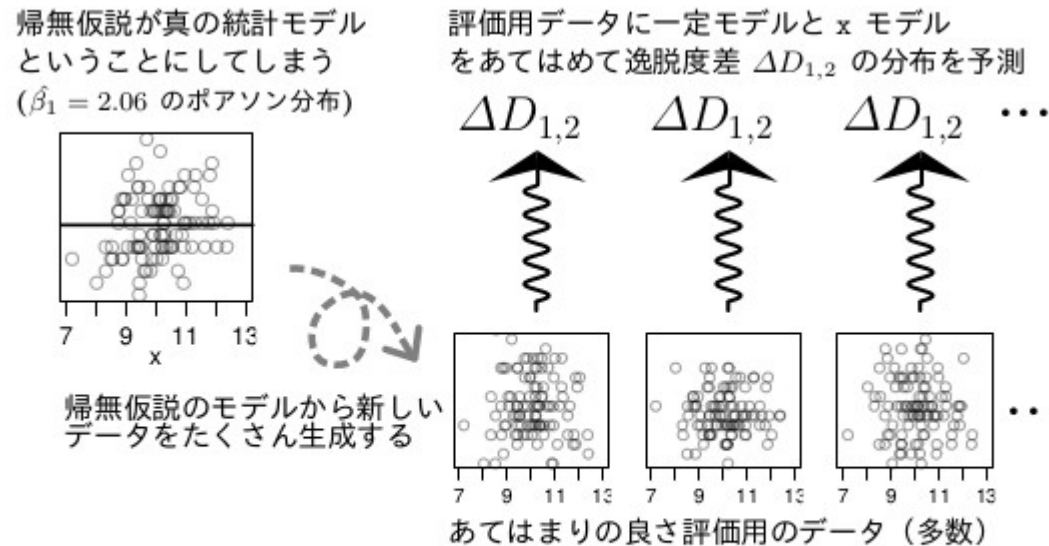


図 6 尤度比検定に必要な  $\Delta D_{1,2}$  の分布の生成。まず帰無仮説である一定モデル ( $\hat{\beta}_1 = 2.06$ , p. 参照) が真の統計モデルだと仮定し、そこから得られるデータを使って逸脱度差  $\Delta D_{1,2}$  がどのような分布になるかを調べる。

7/22 (水)

# 統計モデリング入門 2015 (e)

GLM                      logistic regression  
一般化線形モデル: ロジスティック回帰

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

北大環境科学院の講義 <http://goo.gl/76c4i>

2015-07-22

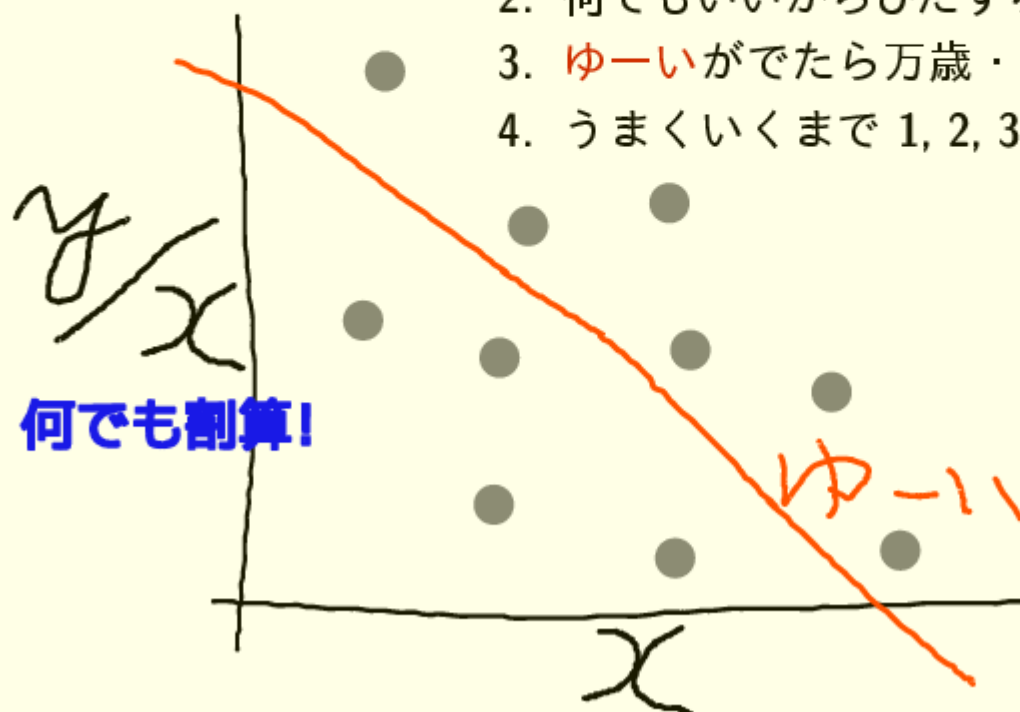
ファイル更新時刻: 2015-07-02 16:24

# 生物学のデータ解析は「割算」しまくり!!

## この悪しき「割算」な統計学

世間でよくみかけるおススメできない作法の例

1. データどんどん割算・割算
2. 何でもいからひたすらセンをひく
3. ゆーいがでたら万歳・万歳
4. うまくいくまで 1, 2, 3 ぐるぐる

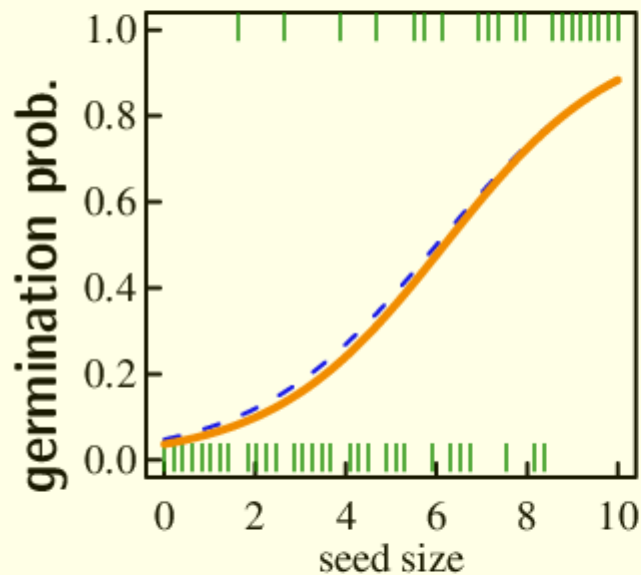


ちなみにこれは  $w$  と  $0/w$  を比較してるんだから、反比例みたいな偽「負の相関」ができるのはあたりまえ

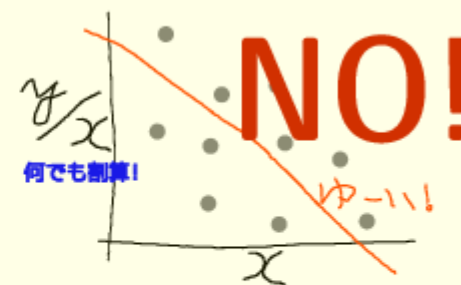
# GLM のひとつ, ロジスティック回帰を使おう

データにあわせたより良い統計モデリングを!

おススメできないデータ解析を回避するための注意点



- むやみに 区画わけしない!
- 何でも 割り算するな!
- たくさん 図を描く
- 「観測データを説明する 確率分布は何か?」を考える



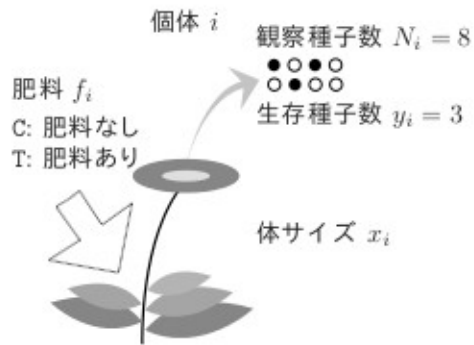
コツ: 不自然にデータをこねくりまわさない  
データの性質・構造にあったモデリングを!

# GLM のひとつ, ロジスティック回帰を使おう

データと確率分布の対応   どういう関係なのか表示してながめる

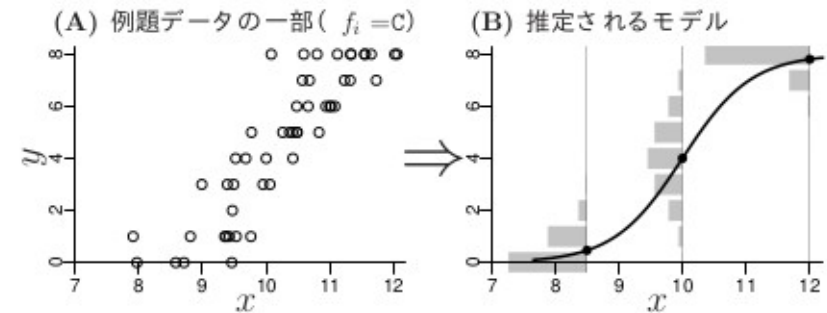
またいつもの例題? …… ちょっとちがう

8 個の種子のうち  $y$  個が **発芽可能** だった! …… というデータ



データと確率分布の対応   どういう関係なのか表示してながめる

ロジスティック回帰とは何なのか?



kubostat2013a (<http://goo.gl/82dgC>)

統計モデリング入門 2013 (5)

2013-07-17  4 / 16

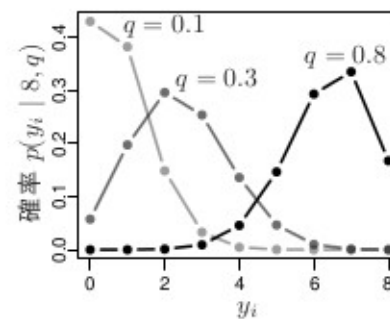
kubostat2013a (<http://goo.gl/82dgC>)

統計モデリング入門 2013 (5)

2013-07-17  9 / 16

データと確率分布の対応   どういう関係なのか表示してながめる

二項分布:  $N$  回のうち  $y$  回, となる確率



7/27 (月)

# 統計モデリング入門 2015 (f)

Generalized Linear Mixed Model (GLMM)  
一般化線形混合モデル

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

北大環境科学院の講義 <http://goo.gl/76c4i>

2015-07-27

ファイル更新時刻: 2015-07-02 16:24

# GLM ではうまく対処できない問題

statistical models appeared in the class  
この授業であつかう統計モデルたち

## The development of linear models

Hierarchical Bayesian Model

Be more flexible

Generalized Linear Mixed Model (GLMM)

Incorporating random effects such as individuality

Generalized Linear Model (GLM)

Always normal distribution? That's non-sense!

Linear model

parameter estimation  
MCMC

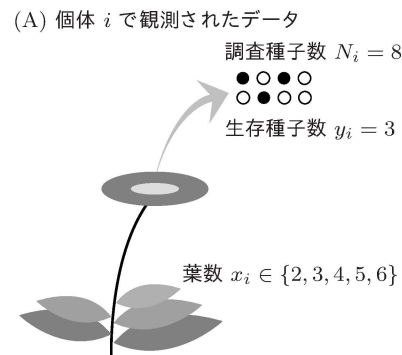
MLE

MSE

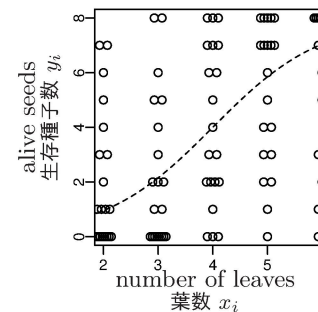
Kubo Doctrine: "Learn the evolution of linear-model family, firstly!"

GLM では説明できない種子データ 「ばらつき」が大きすぎる!

example seed survivorship again, but ...  
今日の例題: 種子の生存確率……前回と同じ?!

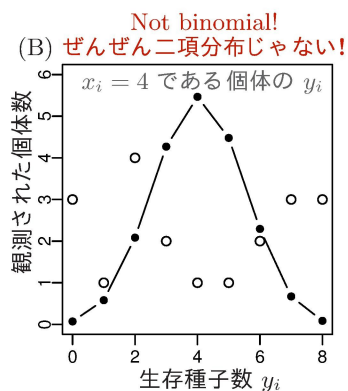
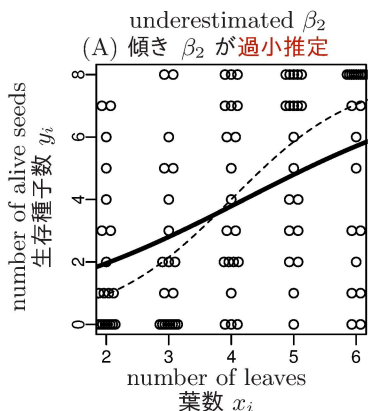


(B) 全 100 個体の  $x_i$  と  $y_i$



GLM では説明できない種子データ 「ばらつき」が大きすぎる!

GLM doesn't work!  
GLM では説明できないばらつき!



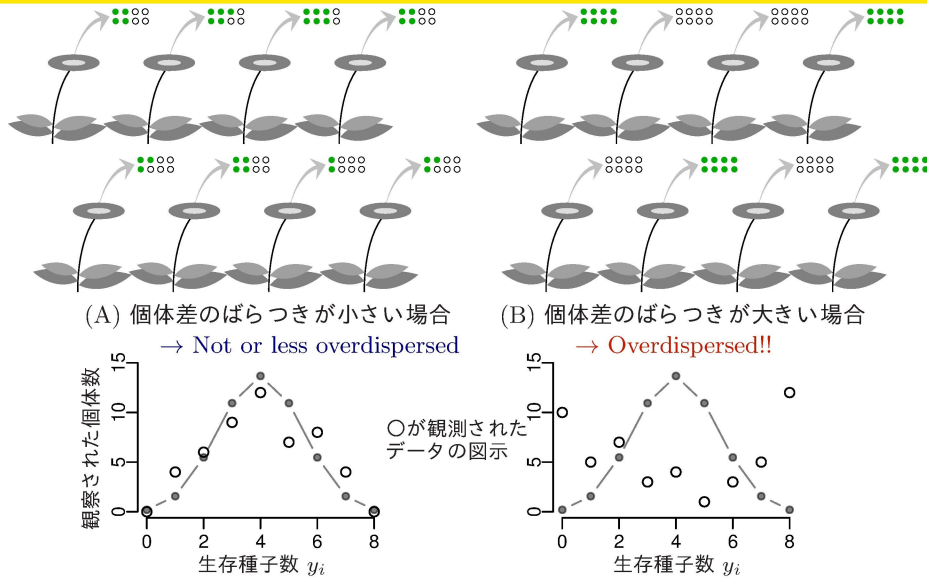
○が観測されたデータの図示

# 今まで「個体差」無視した統計モデル (GLM) を使っていた!

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

## 過分散 (overdispersion) とは何か?



kubostat2015f (<http://goo.gl/76c4i>)

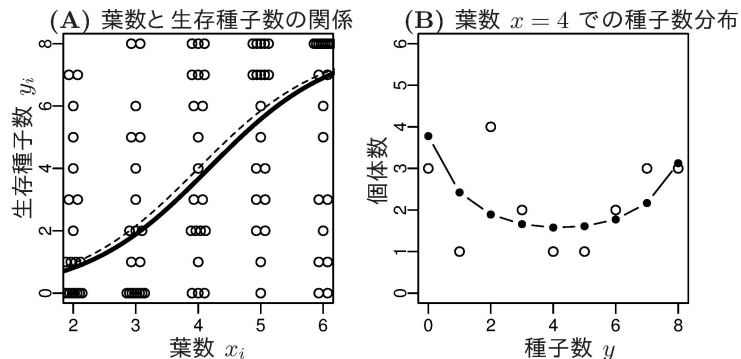
統計モデリング入門 2015 (f)

2015-07-27 10 / 35

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

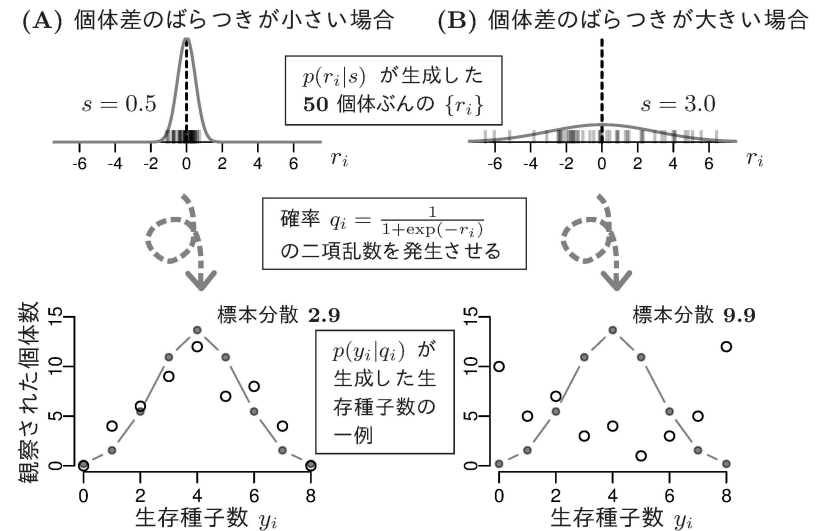
## 推定された GLMM を使った 予測



一般化線形混合モデル

個体差をあらゆるパラメーターを追加

## 個体差 $r_i$ の分布と過分散の関係



kubostat2015f (<http://goo.gl/76c4i>)

統計モデリング入門 2015 (f)

2015-07-27 17 / 35

## 一般化線形混合モデル (Generalized Linear Model, GLMM) を使って問題解決



7/29 (水)

# 統計モデリング入門 2015 (e)

階層ベイズモデル

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

北大環境科学の講義 <http://goo.gl/76c4i>

2015-07-29

ファイル更新時刻: 2015-07-02 16:24

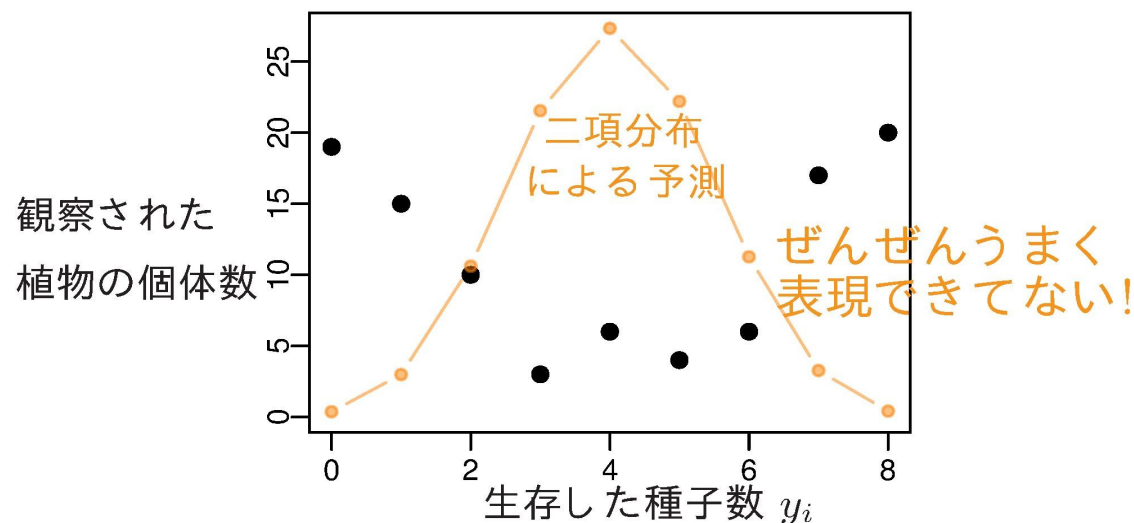
# GLM ではうまく説明できないデータ!?

GLMM は階層ベイズモデルの一種

事前分布をどう選ぶかが重要

また別の観測データ：二項分布だめだめ?!

100 個体の植物の合計 800 種子中 **403 個** の生存が見られたので、平均生存確率は 0.50 と推定されたが……



さっきの例題と同じようなデータなのに?

(「統計モデリング入門」第 10 章の最初の例題)

第 6 回と同じような例題を、こんどはベイズモデルを使ってモデリングします

# GLM を階層ベイズモデル化して対処

GLMM は階層ベイズモデルの一種

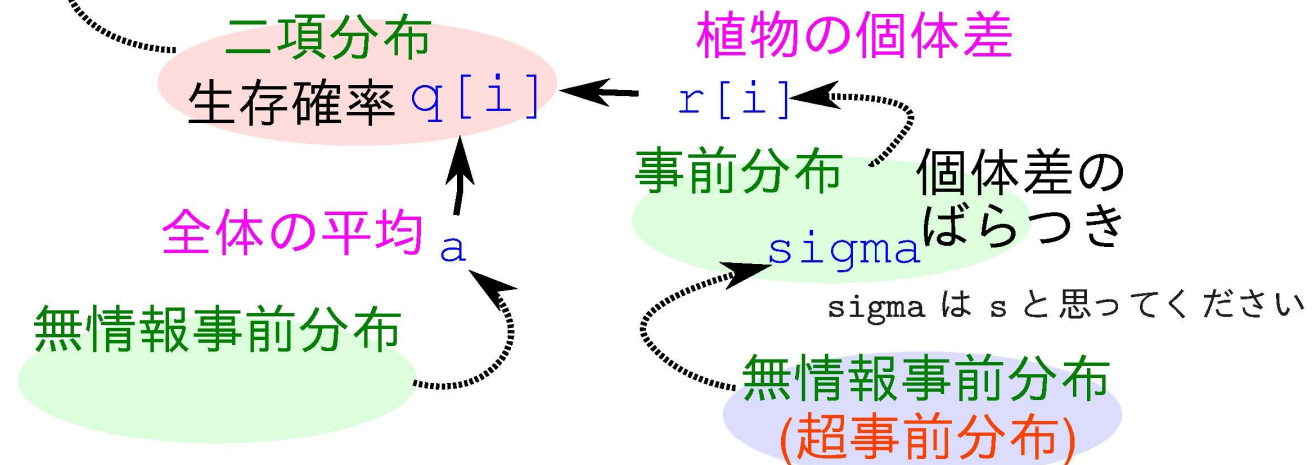
事前分布をどう選ぶかが重要

## なぜ「階層」ベイズモデルと呼ばれるのか？

超事前分布 → 事前分布という階層があるから

データ  
8 個中の  $Y[i]$  個の種子が生存

$\sigma$  は  
hyper parameter



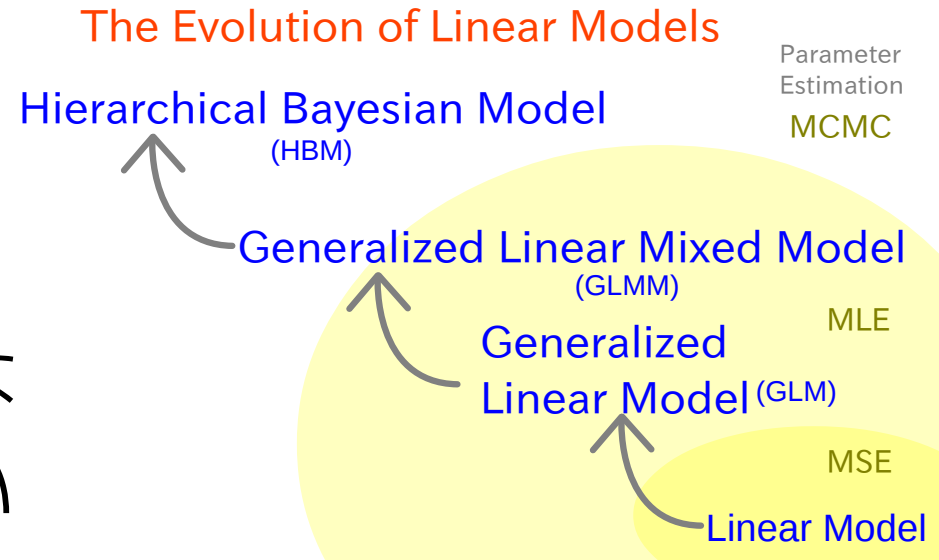
矢印は手順ではなく、依存関係をあらわしている

# なぜ階層ベイズモデルまで勉強するの？

## • 生態学や漁業のデータ

### 解析は難しいから！

- ✓ 個体差・エリア差・空間相関・時間相関・種差などめんどろなことをあつかわないといけない



そういう難しい状況では……

- ベイズモデル化
- そのパラメーターの事後分布を MCMC 法を使って推定するのが無難

# 全体の流れ (1/2)

第 1 回: 7/06 (月) 観測されたパターンを説明する統計モデル

Introduction

第 2 回: 7/08 (水) 確率分布と最尤推定

Probability Distributions and

Maximum Likelihood Estimation (MLE)

第 3 回: 7/13 (月) 一般化線形モデル: ポアソン回帰

Generalized Linear Model (GLM):

Poisson Regression

# 全体の流れ (2/2)

第 4 回: 7/15 (水) モデル選択と検定

Model Selection and Statistical Test

第 5 回: 7/22 (水) 一般化線形モデル: ロジスティック回帰

GLM: Logistic Regression

第 6 回: 7/27 (月) 一般化線形混合モデル

Generalized Linear Mixed Model (GLMM)

第 7 回: 7/29 (水) 階層ベイズモデル

Bayesian GLMM and Markov Chain Monte Carlo