

# 統計モデリング入門 2016 (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)

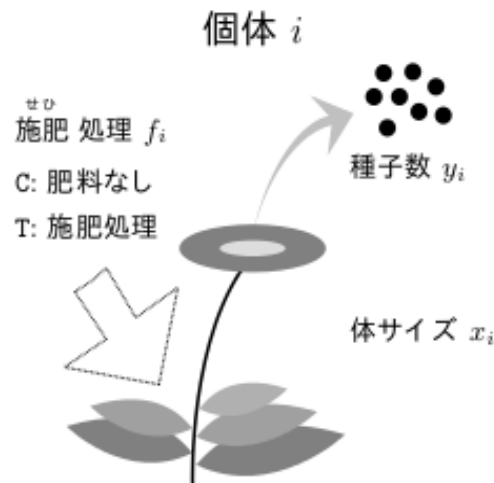
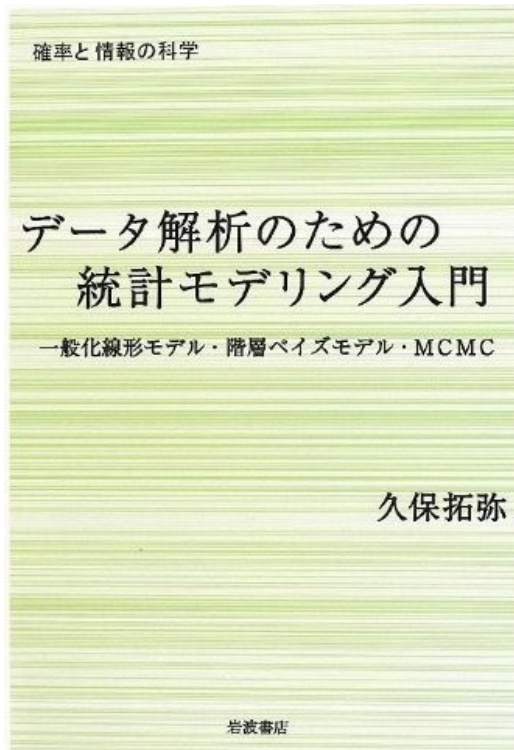


図 3.1 この例題に登場する架空植物の第  $i$  番目の個体. この植物の体サイズ (個体の大きさ)  $x_i$  と肥料をやる施肥処理  $f_i$  が種子数  $y_i$  にどう影響しているのかを知りたい.

この授業の目的:

「データにあわせて

統計モデルを作る」

…という考えかたに慣れる

「脱」ぶらっくぼっくす統計!

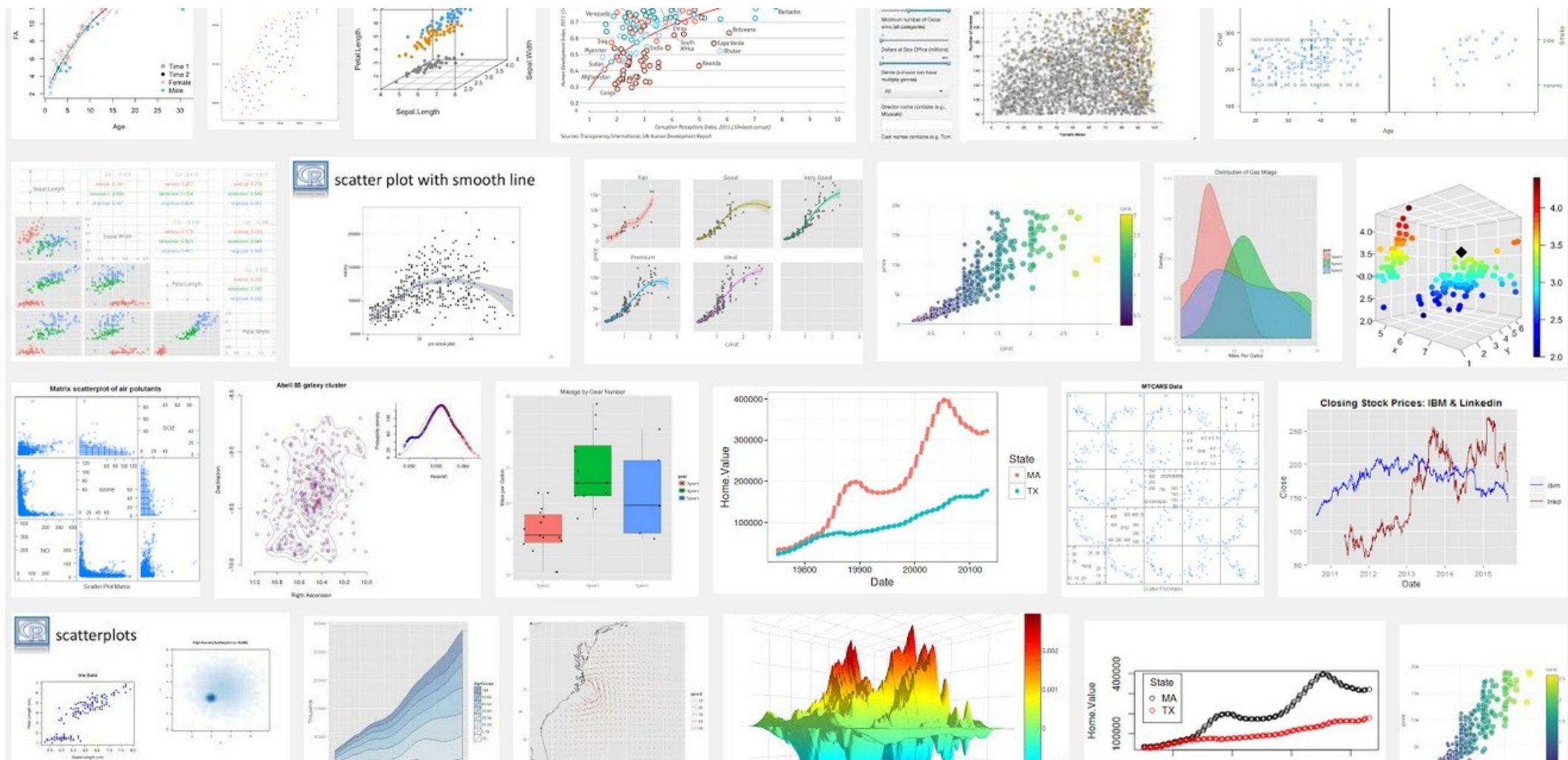
# ぶらっくぼっくす統計！？

- 何をやっているのかよくわからないまま統計ソフトウェアにまるなげ
- とにかく  $p < 0.05$  …とか “\*\*\*” なんかが表示されれば「自分は正しい」「好きなように結論してよい」
- なんでも正規分布だ！

データ解析で  
もっとも重要なことは?

統計モデル?…必ずしもそうではない

# データを図示すること!!



google 画像検索の結果の一例

作図のないデータ解析はありえない!

# じゃ、データ図示の授業やったら？

- ・うーむ…作図は art?

自分の中では体系化されていない

ダメな作図は指摘できる

よい作図の方針はよくわからない

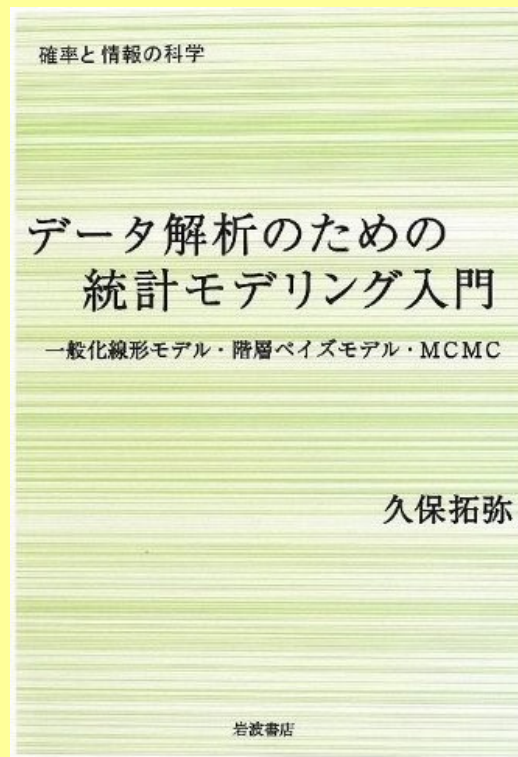
- ・統計モデリングは science

簡単なものから高度なものへステップアップ

何がダメか、比較的明瞭

ただし、明日は作図の練習ですよ！

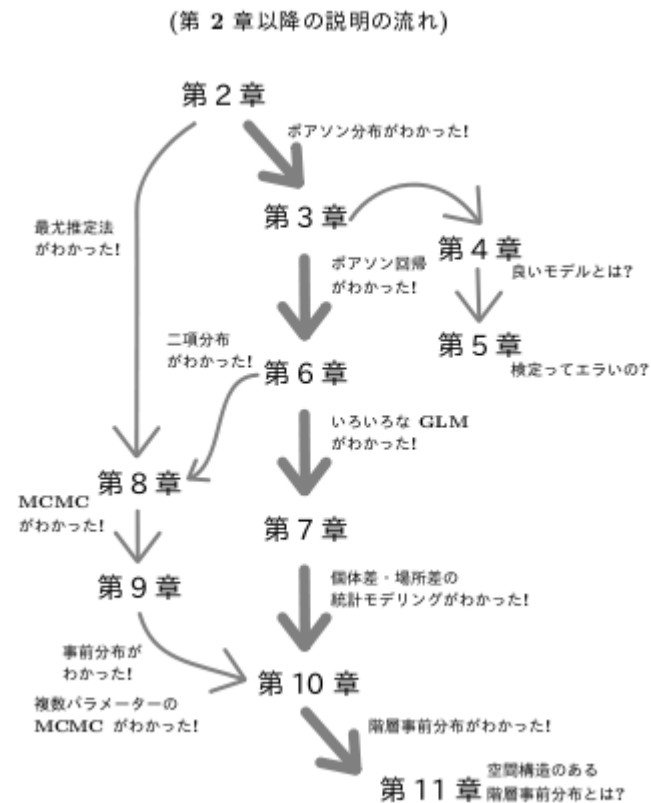
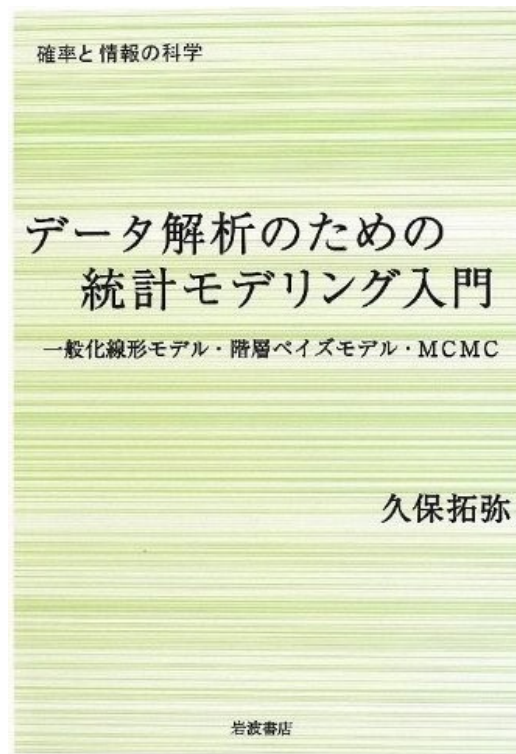
# 教科書とソフトウェア





# 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

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>



# 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



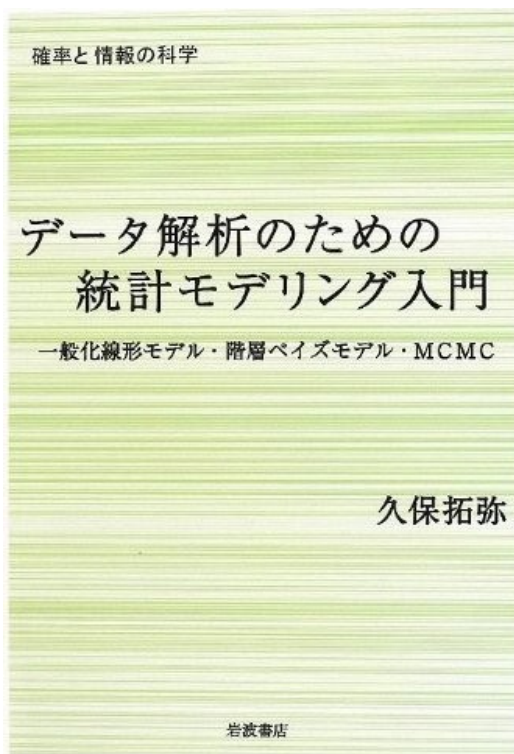
# 統計モデルとは何か？



# 「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

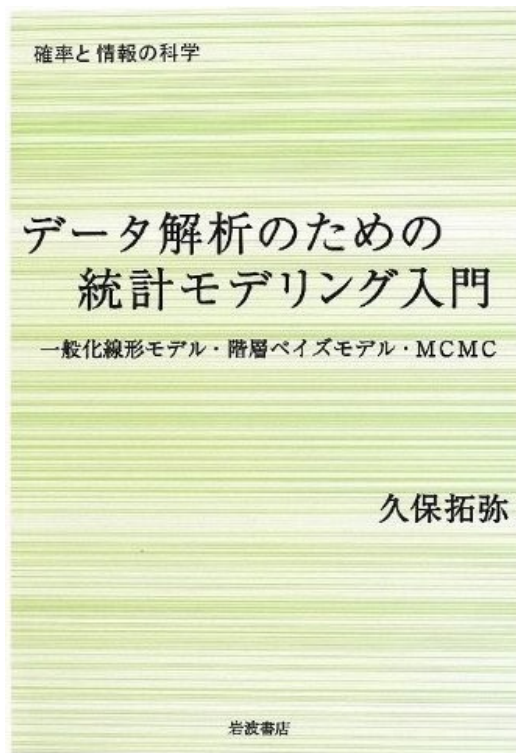
- 観察によってデータ化された現象を説明するために作られる
- 確率分布が基本的な部品であり、これはデータにみられるばらつきを表現する手段である
- データとモデルを対応づける手つづきが準備されていて、モデルがデータにどれぐらい良くあてはまっているかを定量的に評価できる



# 「統計モデリング入門」の主張

「何でも正規分布」じゃないでしょ！

## 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

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

線形モデル

推定計算方法

MCMC

最尤推定法

最小二乗法

# たとえばこんなデータがあったしましょう

(次の時間の例題)

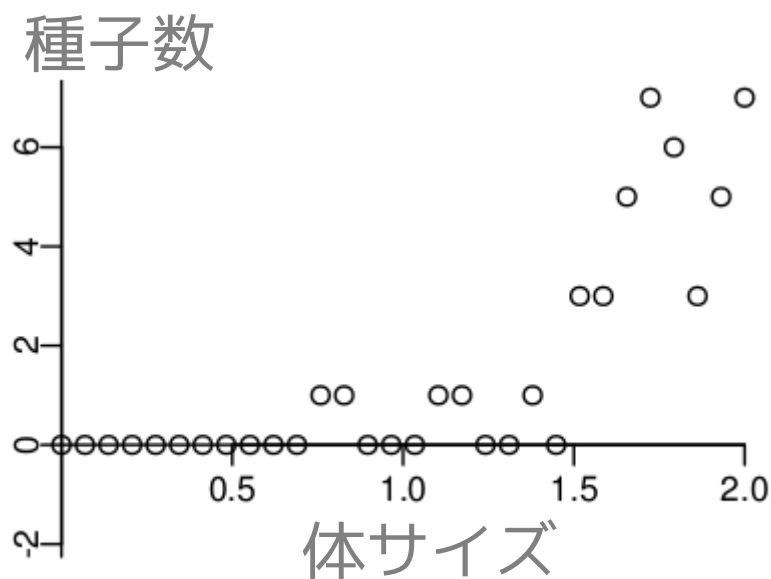
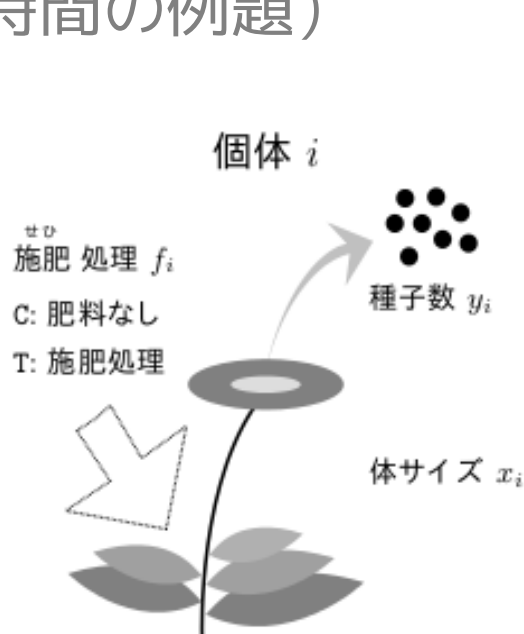
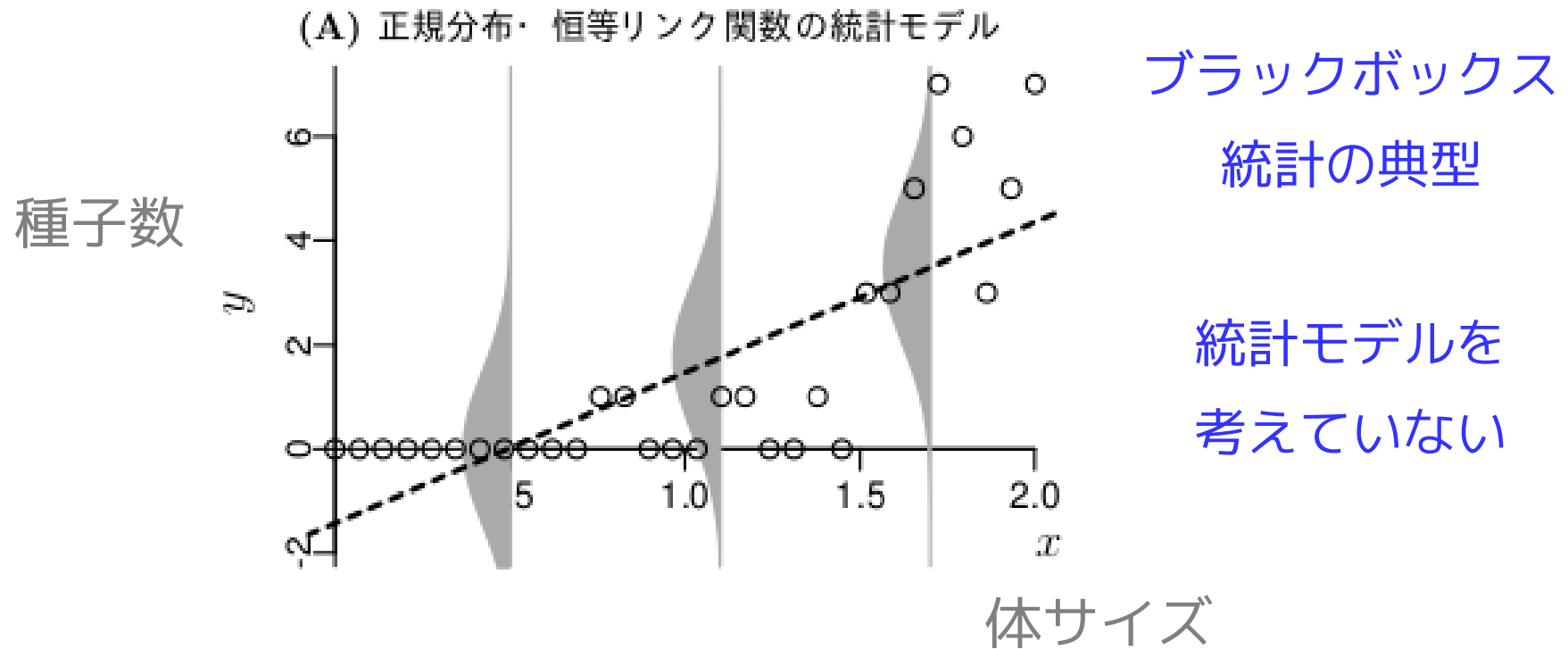


図 3.1 この例題に登場する架空植物の第  $i$  番目の個体. この植物の体サイズ (個体の大きさ)  $x_i$  と肥料をやる施肥処理  $f_i$  が種子数  $y_i$  にどう影響しているのかを知りたい.

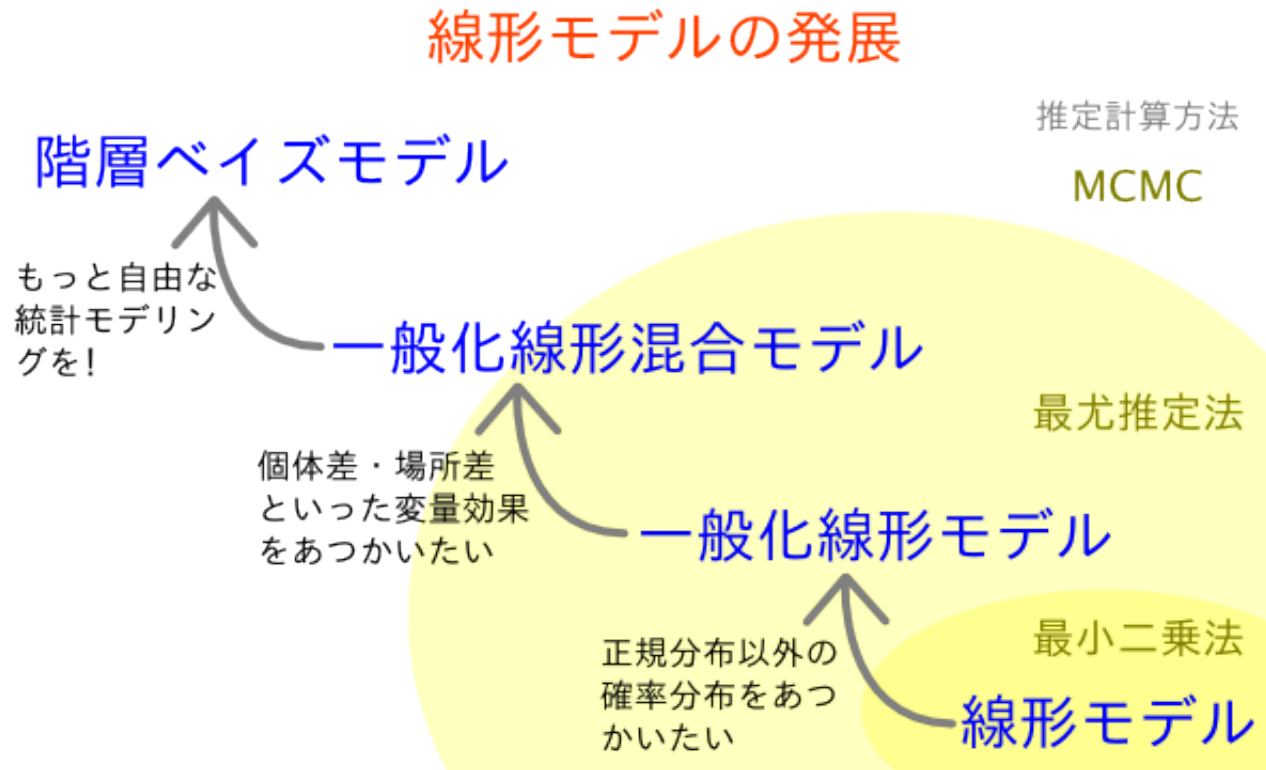
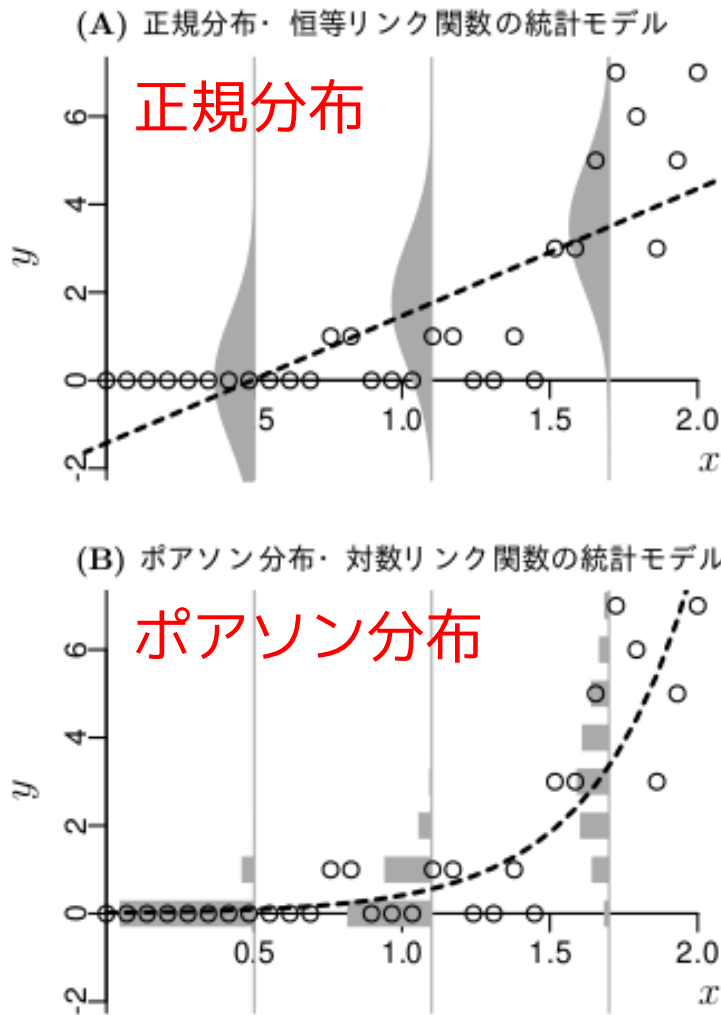
# 残念なデータ解析



0 個, 1 個, 2 個と数える

種子数が「正規分布」……??

# データの値のばらつきをよく見る

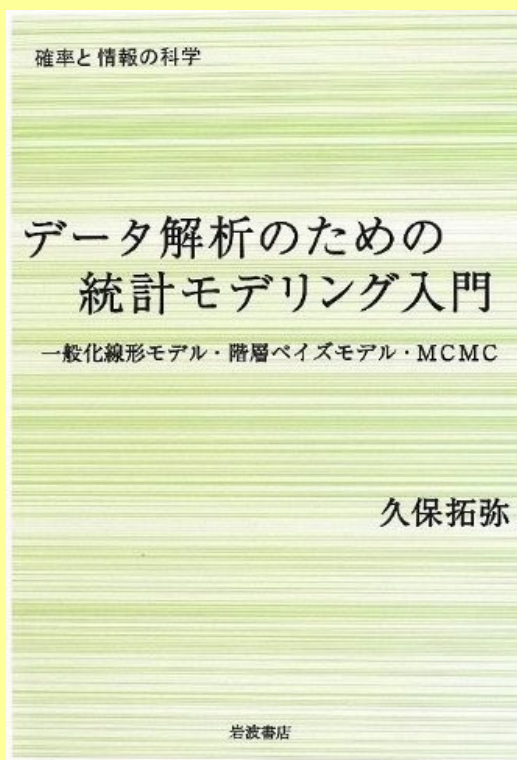


0 個, 1 個, 2 個と数えられる種子数が「正規分布」なわけないだろ!!

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



# 誰のための教科書・講義？



あまり勉強しない「理系」院生のための…

「理系」院生の秘密：あまりよく考えない

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
- とにかく「ゆーい差」さえ出せばよいという  
発想(「理系」だけにありがちな思考?)

統計モデルを理解する

→ 「脱」ブラックボックス

# 11/15 の概要

(a, b) 08:30-10:00 概要, 確率分布と最尤推定

(c) 10:15-11:45 一般化線形モデルとモデル選択

(R1) 13:10-14:40 R 実習 : データフレーム操作と作図

(R2) 14:55-16:25 R 実習 : GLM あてはめと作図

(d) 16:40-18:10 検定とロジスティック回帰



# 11/16 の概要

(e) 8:30-10:00 階層ベイズモデル

(f) 10:15-11:45 時間変化データの階層ベイズモデル

(午後) データ解析相談 (希望する人がいれば)

場所は?

# 新潟大集中講義 2016 (b)

確率分布と最尤推定

久保拓弥 `kubo@ees.hokudai.ac.jp`

新潟大の講義 <http://goo.gl/aFLLHZ>

2016-11-15

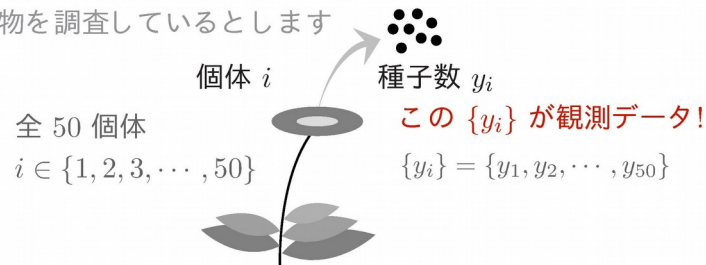
ファイル更新時刻: 2016-11-10 15:29

# 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ンな  
植物を調査しているとします



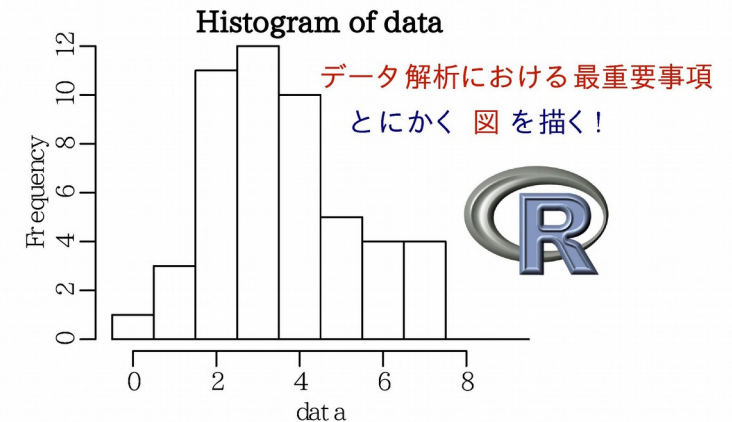
このデータ  $\{y_i\}$  がすでに R という統計ソフトウェアに  
格納されていた, としましょう

```
> 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

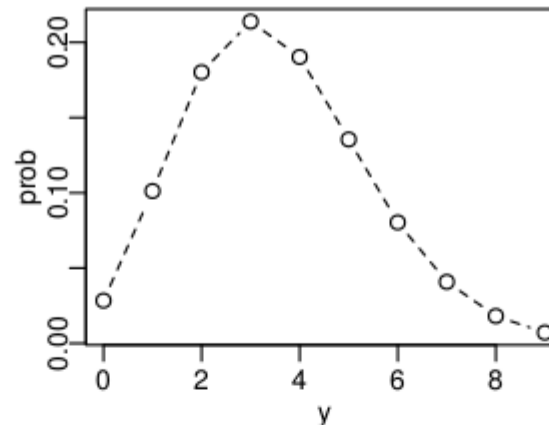


図 4 平均  $\lambda = 3.56$  のポアソン分布. 種子数  $y$  とその確率  $\text{prob}$  の関係が示されている. 図 3 の表を図にしたもの. R の `plot()` 関数の引数, `type = "b"` によって「丸と折れ線による図示」, `lty = 2` によって「折れ線は破線で」と指示している.

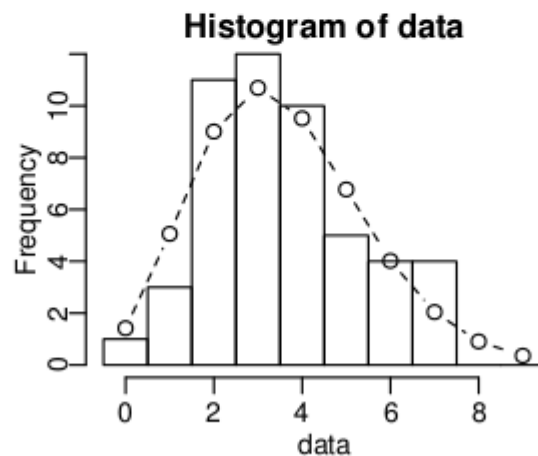
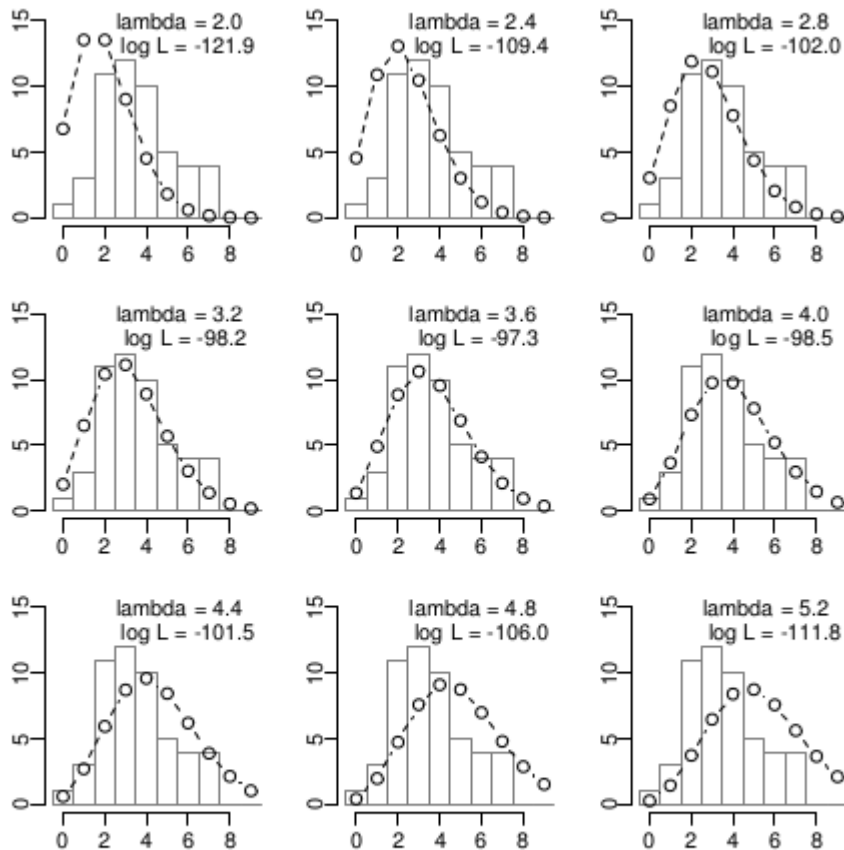


図 5 観測データと確率分布の対応をながめる. ヒストグラムは図 3 と同じ. それに重ねられている丸と破線は  $y$  個の種子をもつ個体数の予測. 平均 3.56 の図 3 のポアソン分布の確率分布に全個体数 50 をかけて得られる.



さいゆう

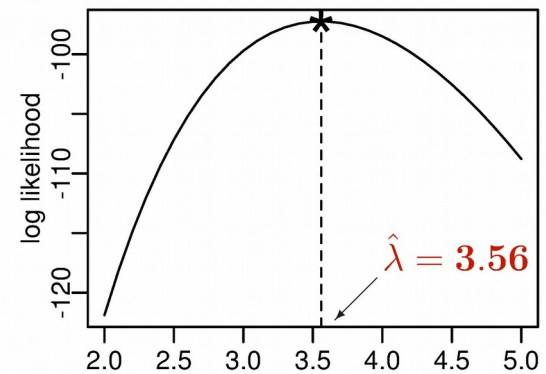
# 最尤推定という考えかたを説明します



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

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



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

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。

# 新潟大集中講義 2016 (c)

一般化線形モデル (ポアソン回帰) とモデル選択

久保拓弥 `kubo@ees.hokudai.ac.jp`

新潟大の講義 <http://goo.gl/aFLLHZ>

2016-11-15

ファイル更新時刻: 2016-11-10 15:29

# ここで登場する ---

## 「何でも正規分布」ではダメ! という発想

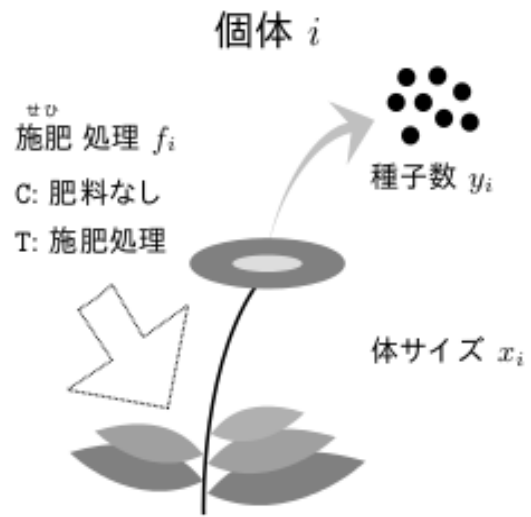


図 3.1 この例題に登場する架空植物の第  $i$  番目の個体。この植物の体サイズ（個体の大きさ） $x_i$  と肥料をやる施肥処理  $f_i$  が種子数  $y_i$  にどう影響しているのかを知りたい。

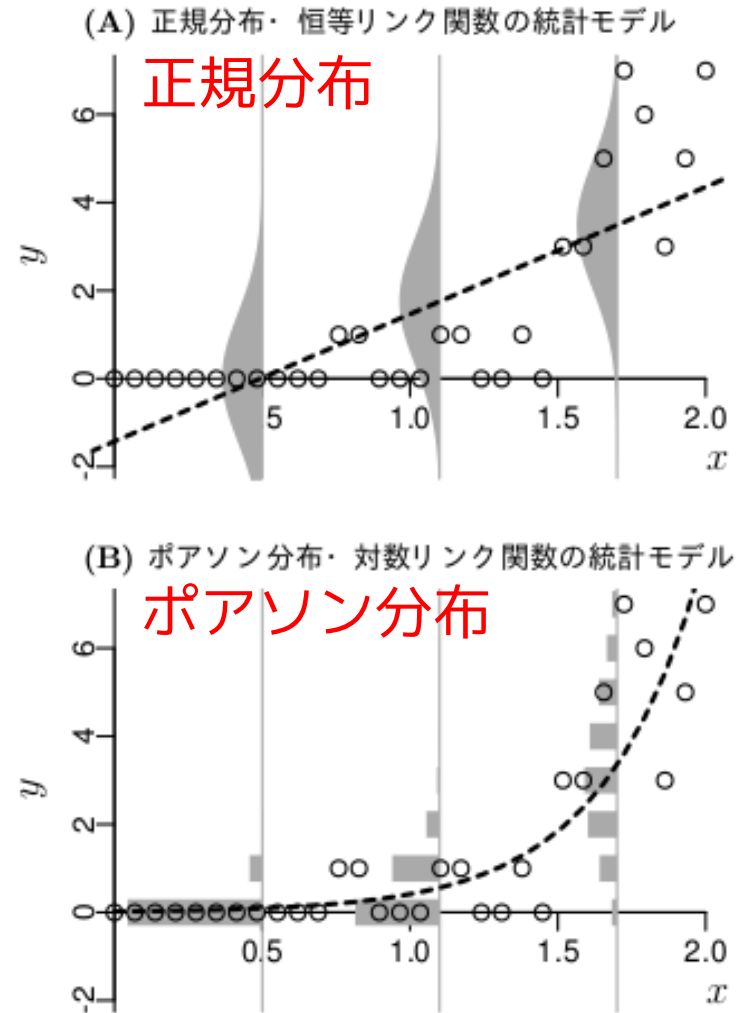


図 3.9 回帰モデルと確率分布の関係。また別の架空データに対して GLM をあてはめた例。破線は  $x$  とともに変化する平均値、グレイで

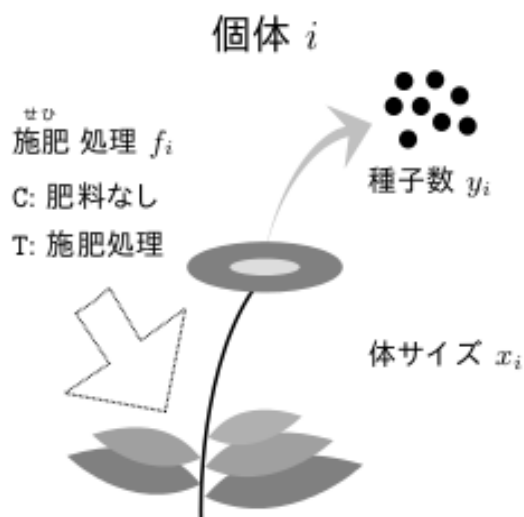


図 3.1 この例題に登場する架空植物の第  $i$  番目の個体  
体サイズ(個体の大きさ)  $x_i$  と肥料をやる施肥処理、  
にどう影響しているのかを知りたい。

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  
確率分布の指定  
モデル式  
リンク関数の指定 (省略可)  
) data.frame の指定

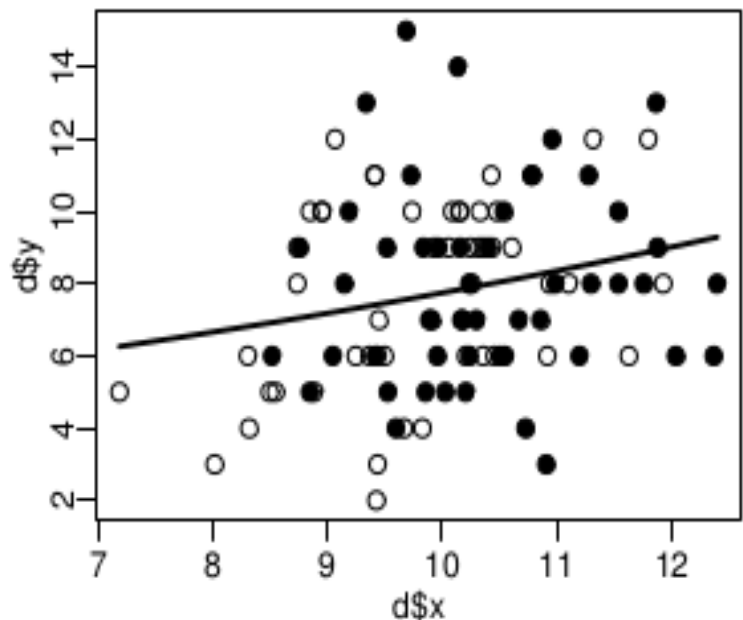
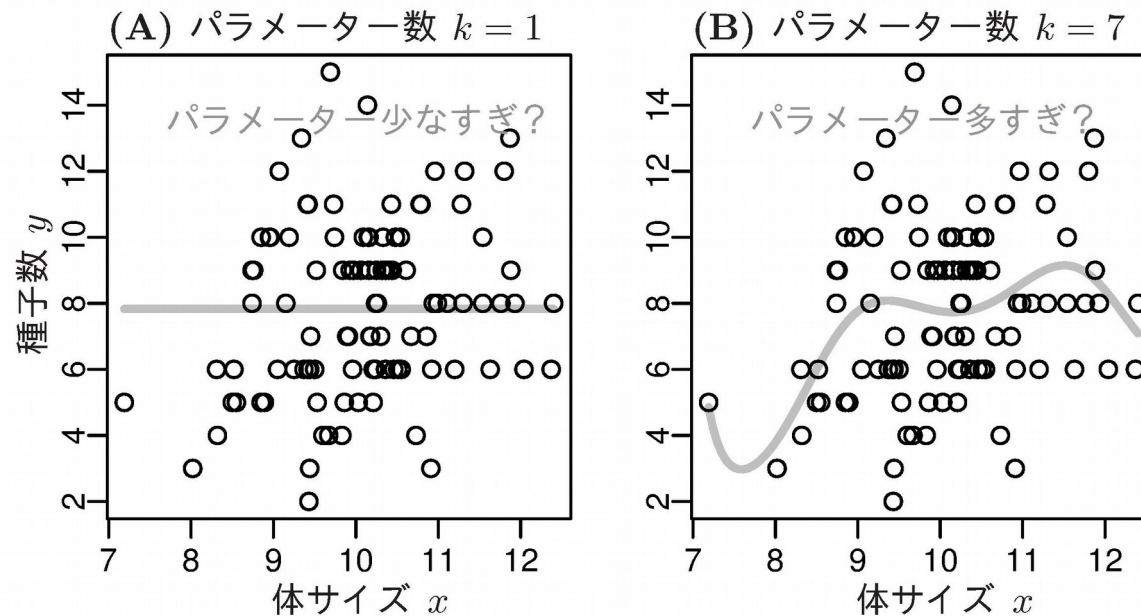


図 17 平均種子数  $\lambda$  の予測. 図 12 に  $\lambda$  の予測値 (実線) を上げきしたものの。

# 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  $k$ ?

# 新潟大集中講義 2016 (d)

統計学的検定 と ロジスティック回帰

久保拓弥 kubo@ees.hokudai.ac.jp

新潟大の講義 <http://goo.gl/aFLLHZ>

2016-11-15

ファイル更新時刻: 2016-11-10 15:29



# 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ーの 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?

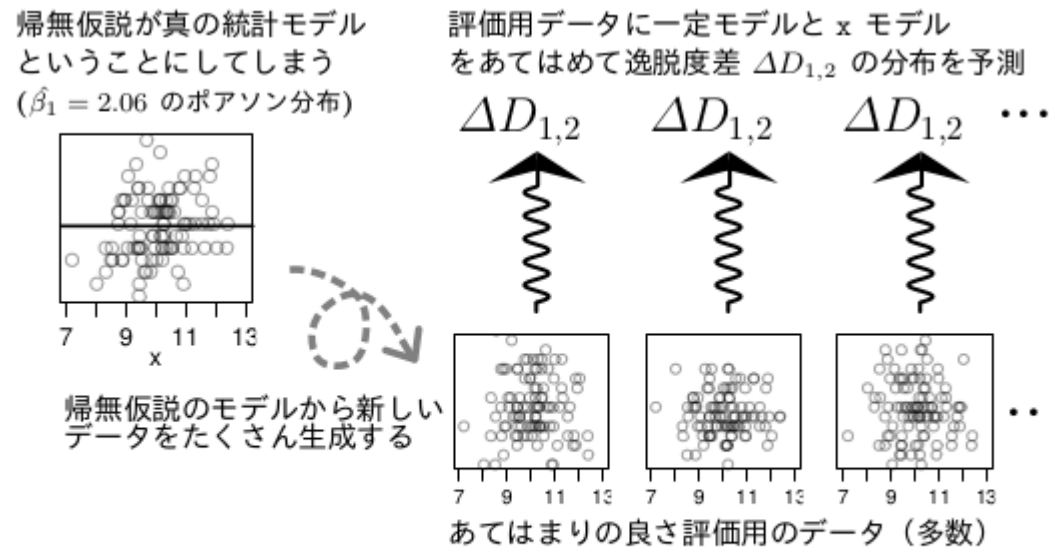
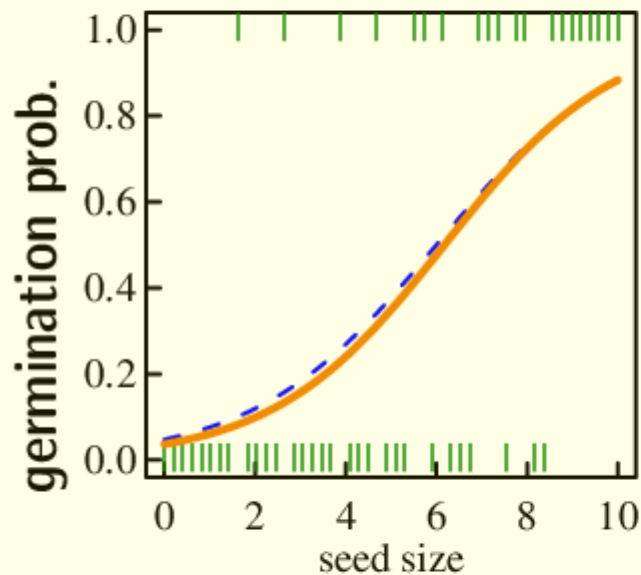


図 6 尤度比検定に必要な  $\Delta D_{1,2}$  の分布の生成。まず帰無仮説である一定モデル ( $\hat{\beta}_1 = 2.06$ , p. 参照) が真の統計モデルだと仮定し、そこから得られるデータを使って逸脱度差  $\Delta D_{1,2}$  がどのような分布になるかを調べる。

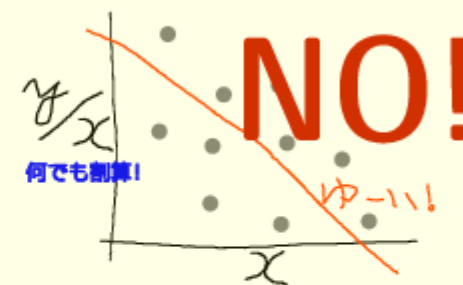
# GLM のひとつ, ロジスティック回帰を使おう

データにあわせて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!

# 新潟大集中講義 2016 (e)

階層ベイズモデル – 個体差・場所差のモデリング

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

新潟大の講義 <http://goo.gl/aFLLHZ>

2016-11-15

ファイル更新時刻: 2016-11-10 15:29

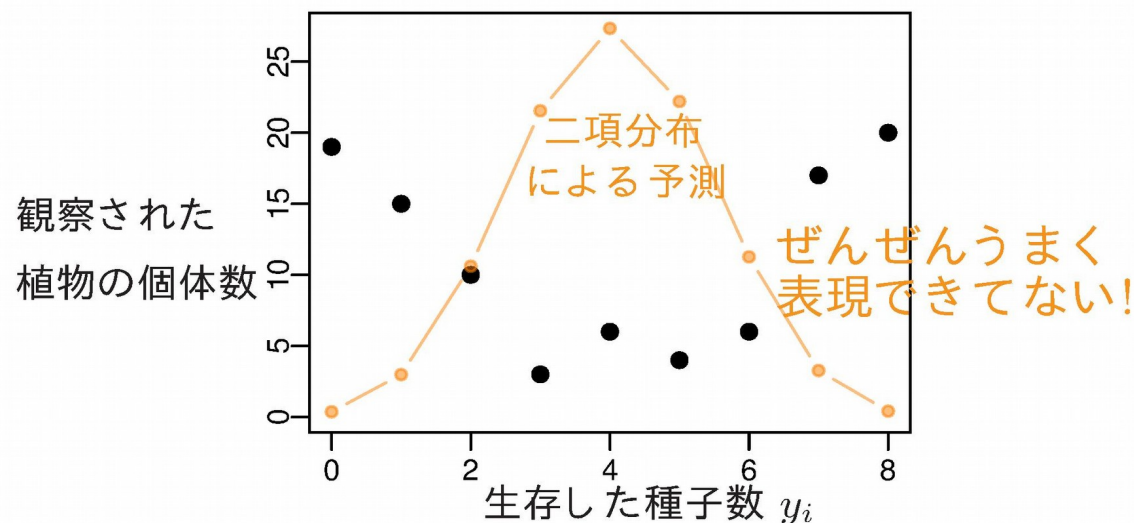
# GLM ではうまく説明できないデータ!?

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

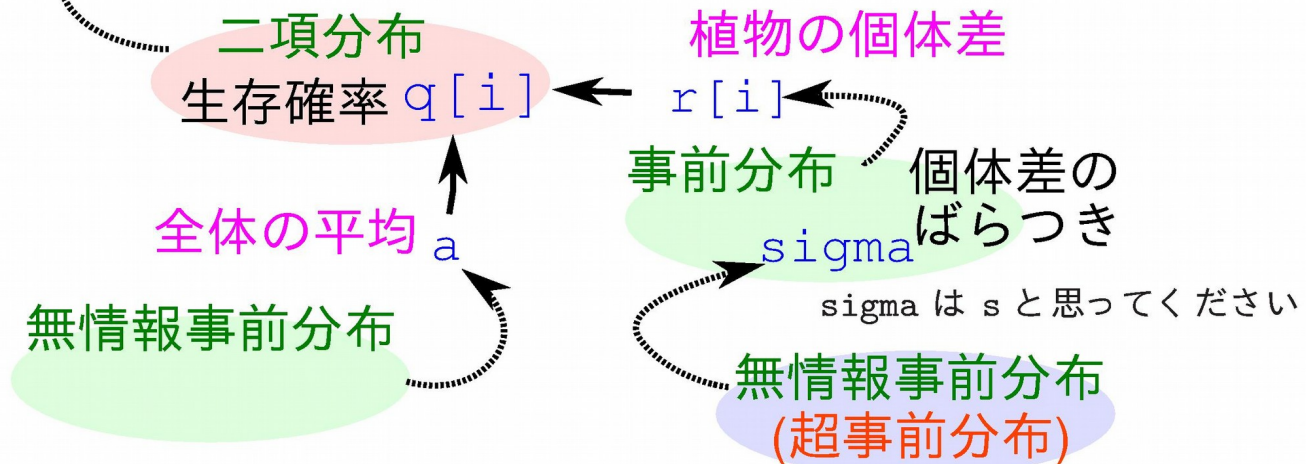
事前分布をどう選ぶかが重要

## 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

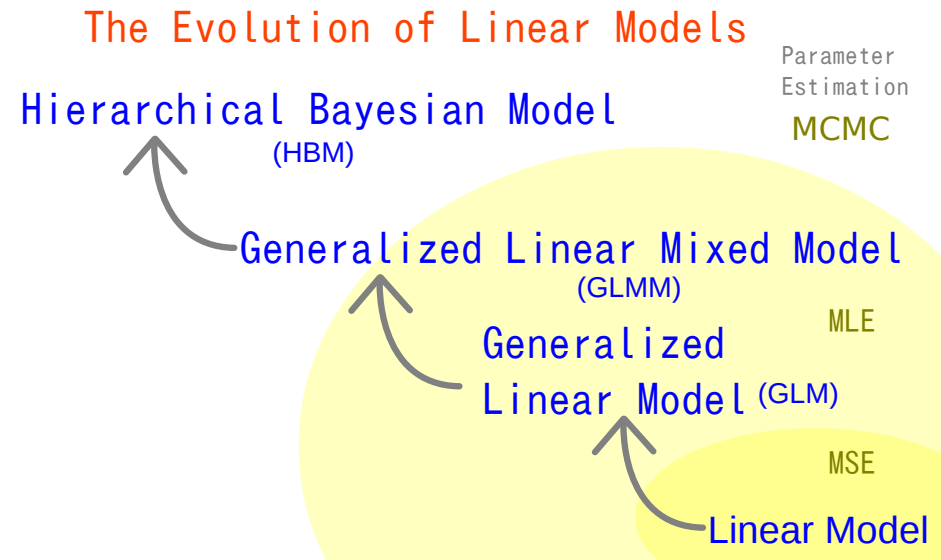


矢印は手順ではなく、依存関係をあらわしている



# なぜ階層ベイズモデルまで勉強するの？

- ✓ 個体差・店舗差・地区差・空間相関・時間相関などめんどろな「細かい差異」をあつかわないといけない

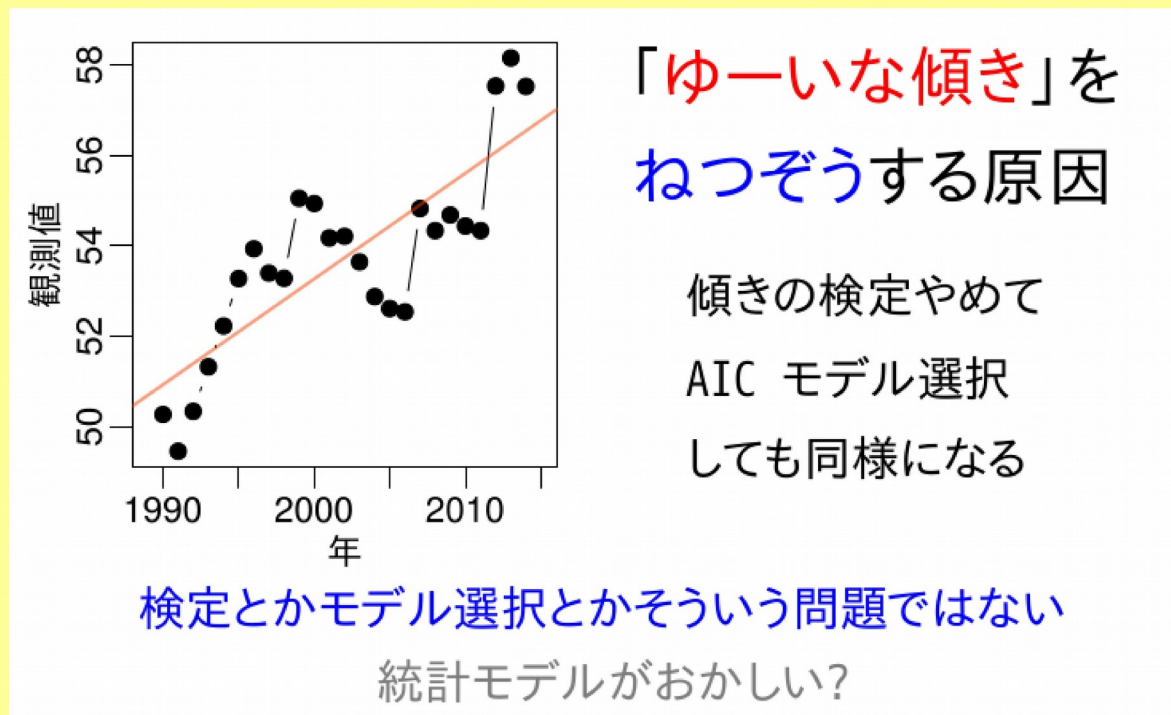


そういう難しい状況では……

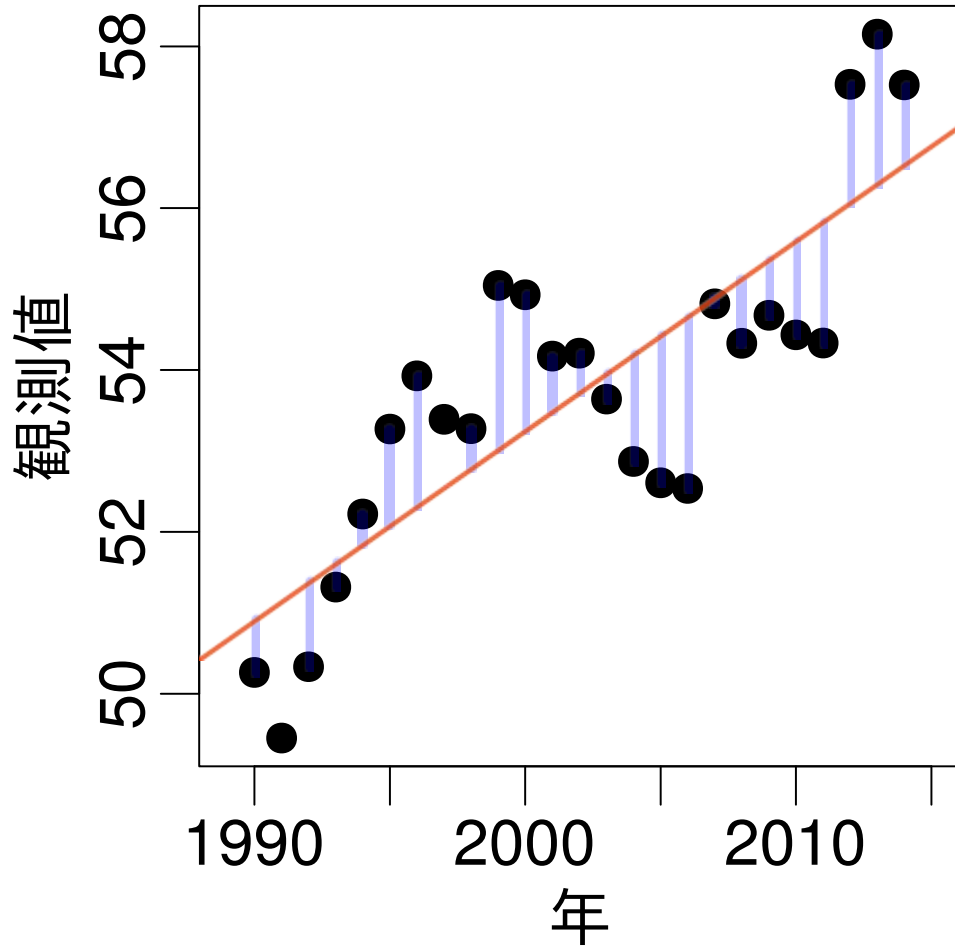
- 「差異」の階層ベイズモデル化
- そのパラメーターの事後分布を MCMC 法を使って推定するのが無難

## (f) 階層ベイズモデル - 時間変化の統計モデリング

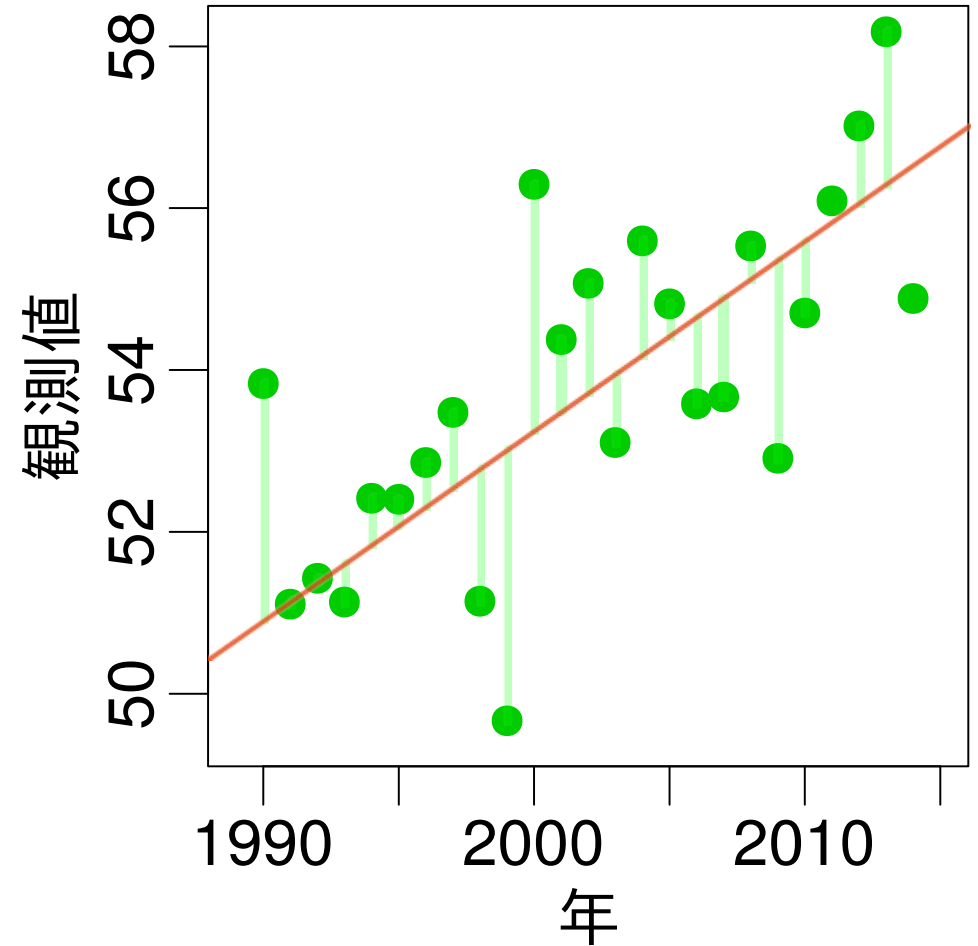
### (危1) 時系列データを GLM で



# 時系列の「ずれ」



# GLM のずれ



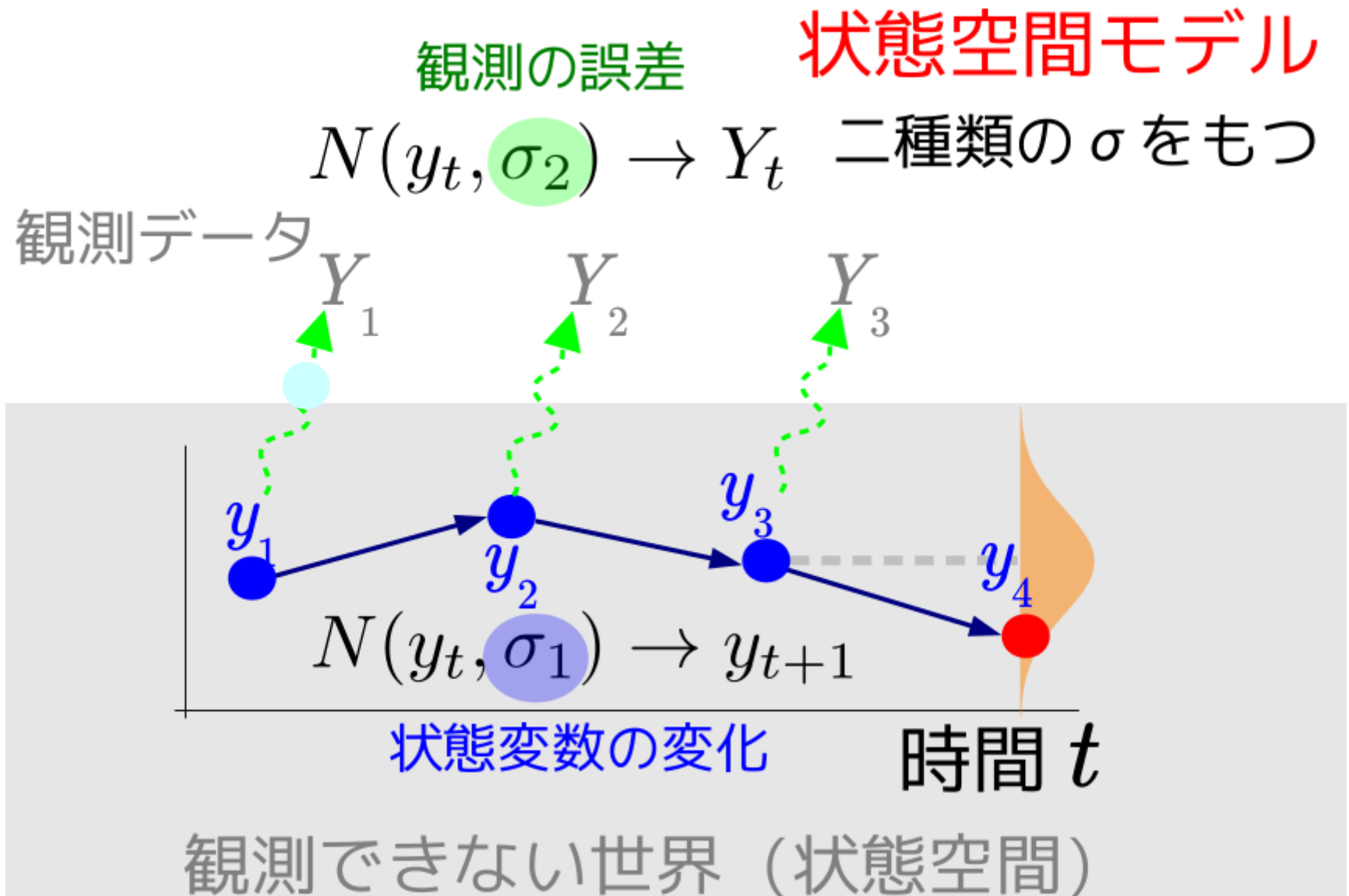
直線からのずれがちがう!

時間的自己相関がある

時間的自己相関がない



# 状態空間モデル + 観測モデル



# 統計モデリング入門, 全体のながれ

データの性質・構造をよくみて統計モデルを作る

