茨城大学集中講義・統計モデリング入門（a）

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

茨城大学理工学研究科

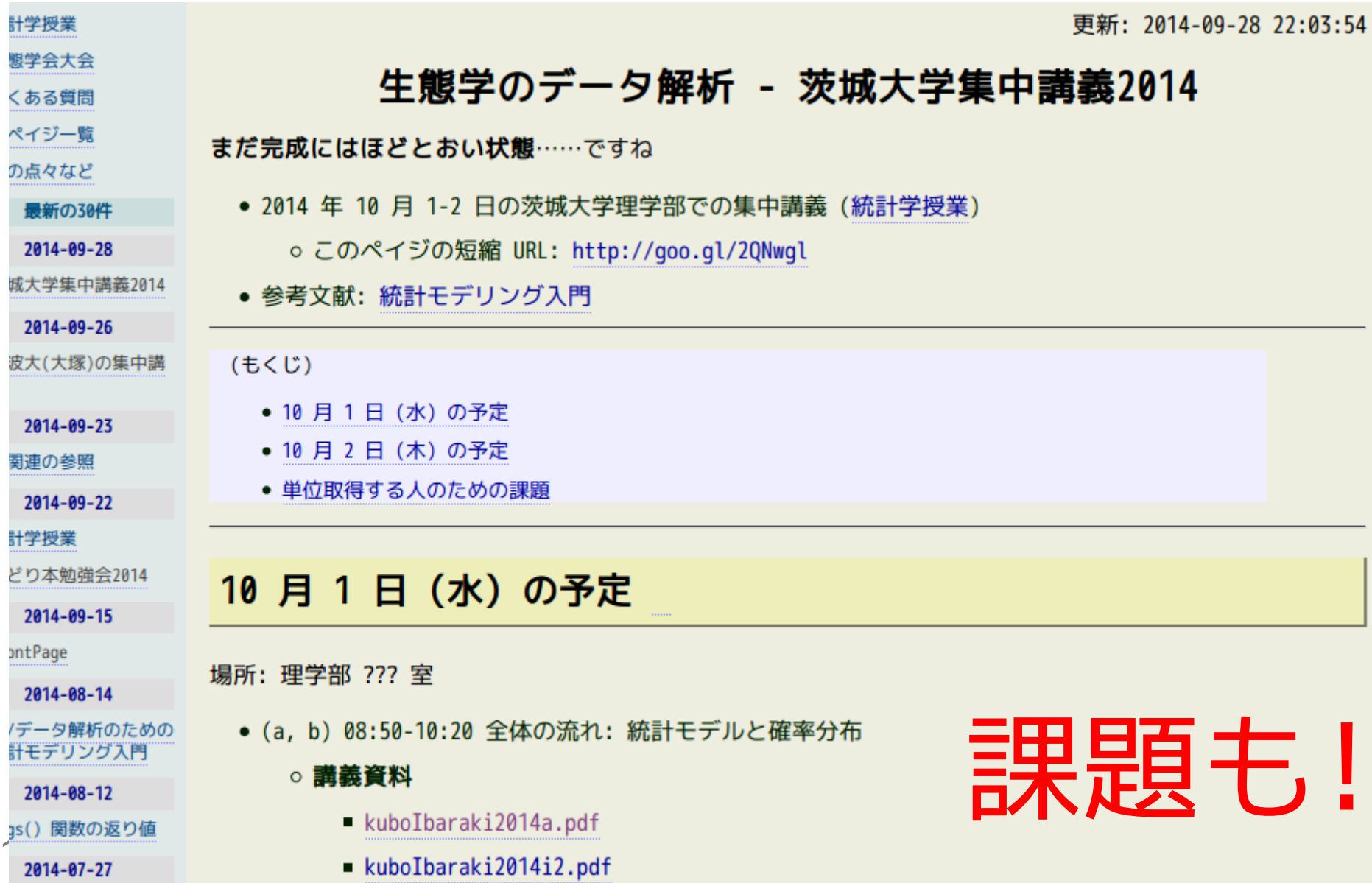
・生物系特別講義 IV

・先端科学トピックス

2014 年 10/1-10/2

統計モデリング授業の web page

http://goo.gl/2QNwgl



計学授業
態学会大会
くある質問
ページ一覧
の点々など
最新の30件
2014-09-28
城大学集中講義2014
2014-09-26
波大(大塚)の集中講
2014-09-23
関連の参照
2014-09-22
計学授業
どり本勉強会2014
2014-09-15
ontPage
2014-08-14
/データ解析のための
計モデリング入門
2014-08-12
gs() 関数の返り値
2014-07-27

更新: 2014-09-28 22:03:54

生態学のデータ解析 - 茨城大学集中講義2014

まだ完成にはほどとおい状態……ですね

- 2014 年 10 月 1-2 日の茨城大学理学部での集中講義（統計学授業）
  - このページの短縮 URL: http://goo.gl/2QNwgl
- 参考文献: 統計モデリング入門

（もくじ）

- 10 月 1 日（水）の予定
- 10 月 2 日（木）の予定
- 単位取得する人のための課題

---

10 月 1 日（水）の予定

場所: 理学部 ??? 室

- (a, b) 08:50-10:20 全体の流れ: 統計モデルと確率分布
  - **講義資料**
    - kuboIbaraki2014a.pdf
    - kuboIbaraki2014i2.pdf

課題も！

自己紹介：久保拓弥

- 北大の環境科学院という学部のない大学院
- 生態学に関係するデータ解析とかやってます
  - 野外調査をしない生態学者
- データは誰か別の人（共同研究者）がとってきてくれます

そもそも生態学って何?

- 生物の数の変化や分布や生活の様子を調べる
- いろいろな動植物が対象

統計モデルは
データ解析の道具

なぜデータ解析の方法を
勉強しなければ
ならないのか?

科学のデータ解釈は統計的手法に依存

「データ→結論」のつなぎめ

- データ解析がおかしいと結論もおかしい
- データ解析を悪用して結論をねつぞうできる
- 論文を読むときにデータ解析の部分がわからないと「どうしてこのデータからこの結論が導かれたのか，妥当といえるのか」などがわからない→論文を批判的に読めない

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

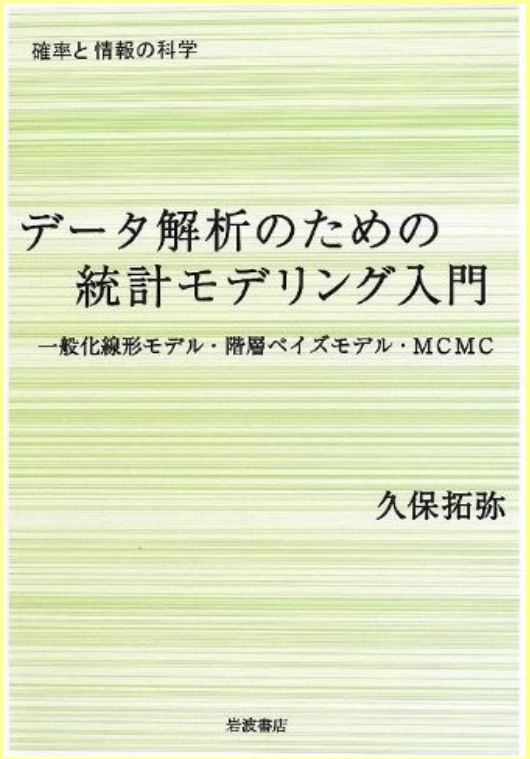
- ブラックボックス統計解析
- とにかく「ゆーい差」さえ出せばよいという発想になっている
- 大学・大学院でもあまりちゃんと教えられていない，教えられるヒトが少ない……とくに近年発達している統計モデリングについて

この授業のねらい

できるだけ内容を理解して統計ソフトウェアを使おう!

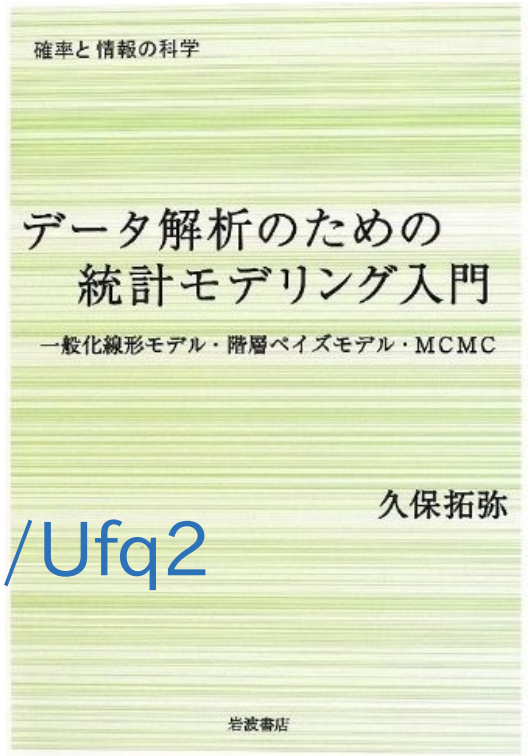
- データ解析で使われるの中でも比較的簡単な統計モデルを理解しよう
- 「ゆーい差」さえ出せばよいという発想をやめて，データと統計モデルの対応関係をよく見よう（作図重要）
- 統計ソフトウェア R を使い始めよう

教科書とソフトウェア

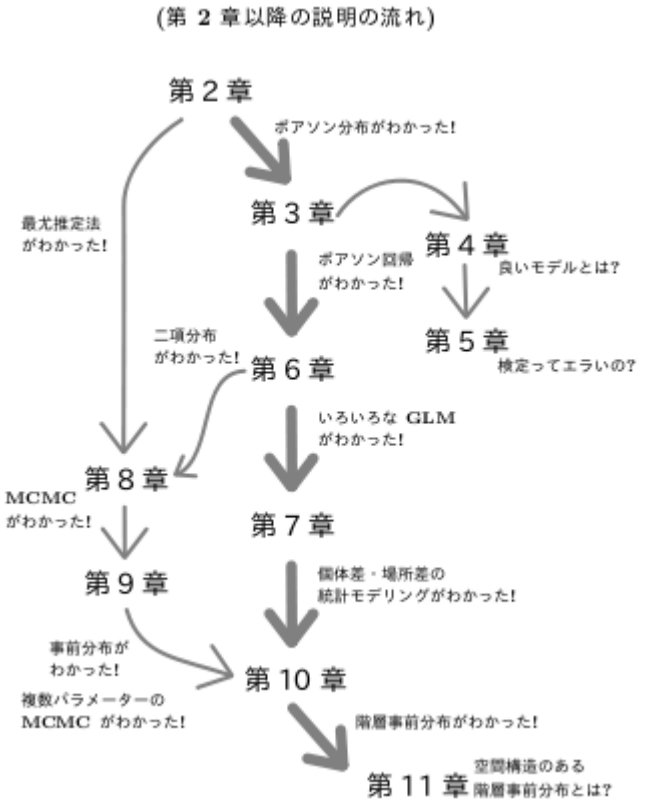

確率と情報の科学
データ解析のための
統計モデリング入門
一般化線形モデル・階層ベイズモデル・MCMC
久保拓弥
岩波書店

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



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

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

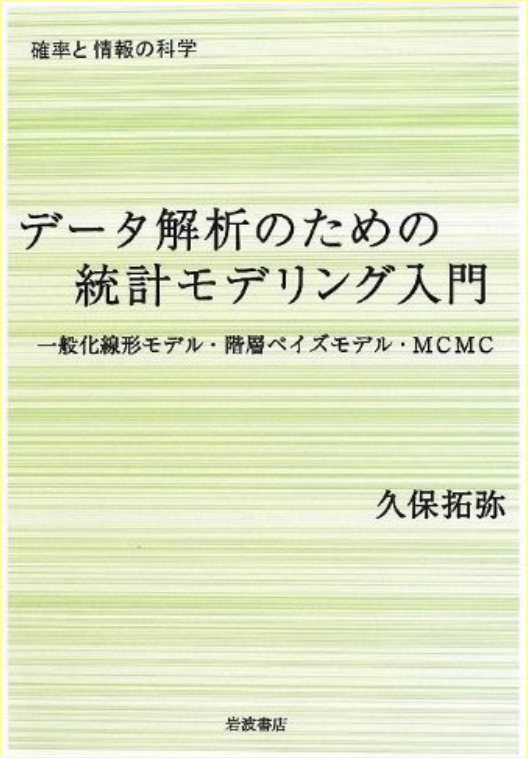
統計ソフトウェア R

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

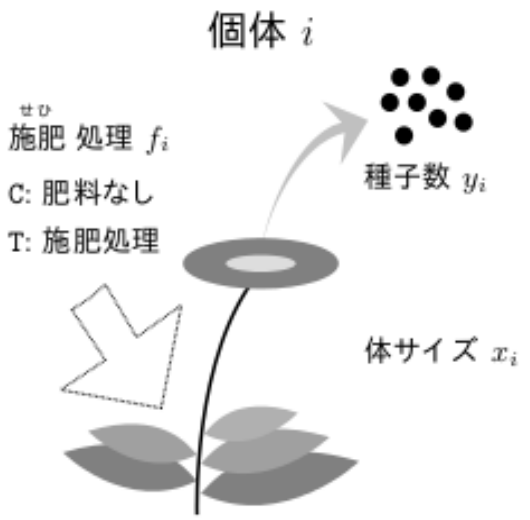
この教科書でも R を
使って問題を解決する
方法を説明しています

統計モデルとは何か?


確率と情報の科学
データ解析のための
統計モデリング入門
一般化線形モデル・階層ベイズモデル・MCMC
久保拓弥
岩波書店

たとえばこんなデータがあったしましょう

（次回の講義の例題）



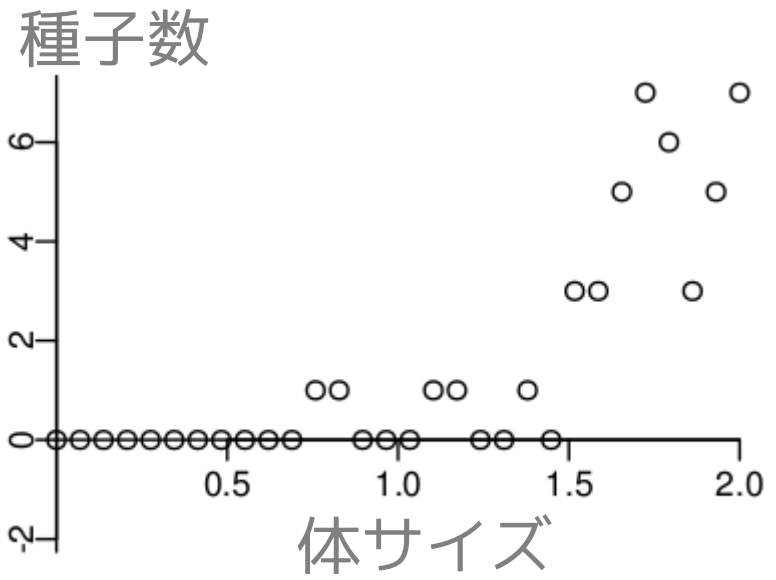
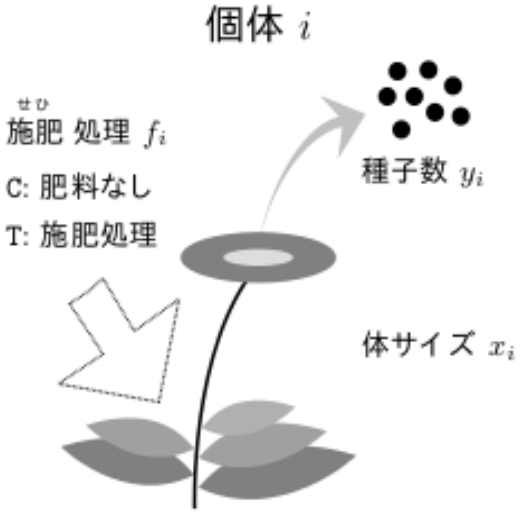
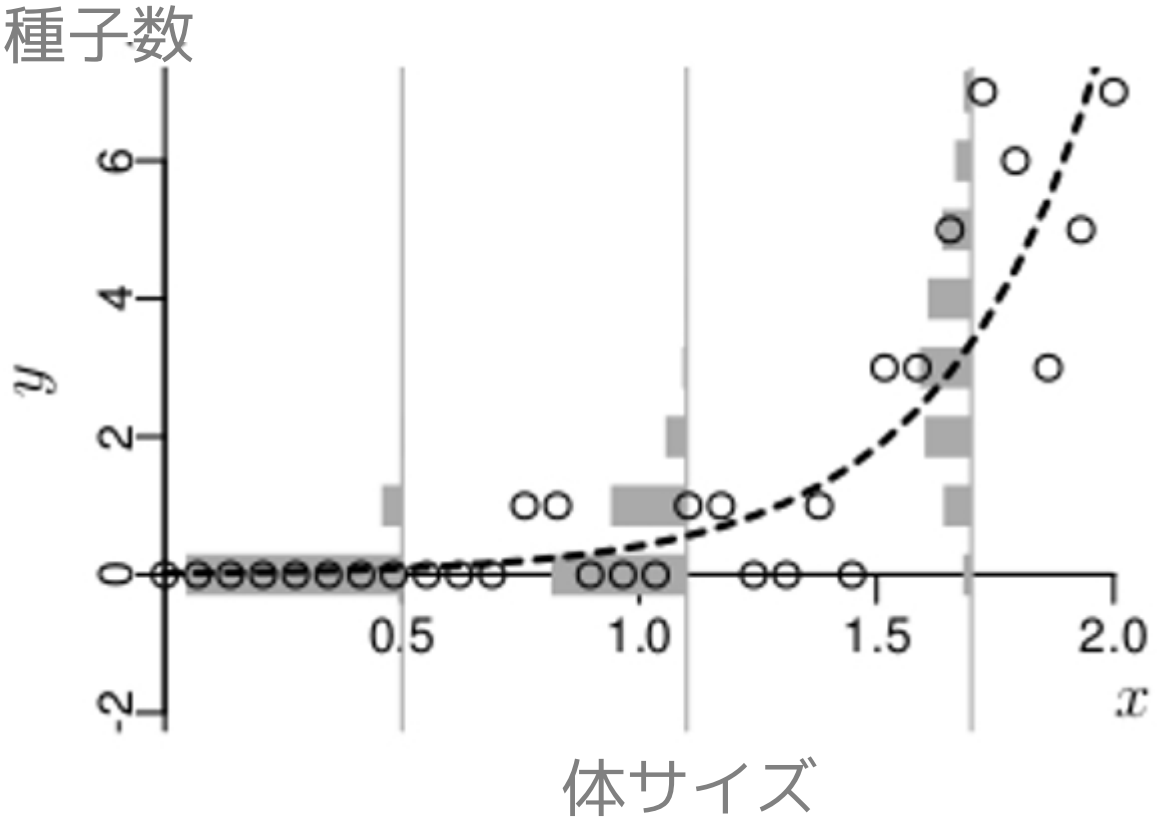


図 3.1　この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（ 個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

こういう「定量的な説明」があったらいいな……

なんとなく「わかった」ような気分?





数式で書かれた
「統計モデル」を
準備する
それをデータに
あてはめる

種子数の平均値はサイズ x とともに増大する

平均値が増大するとばらつきが変化する

……などなど……

「統計モデル」のしくみを理解しよう!

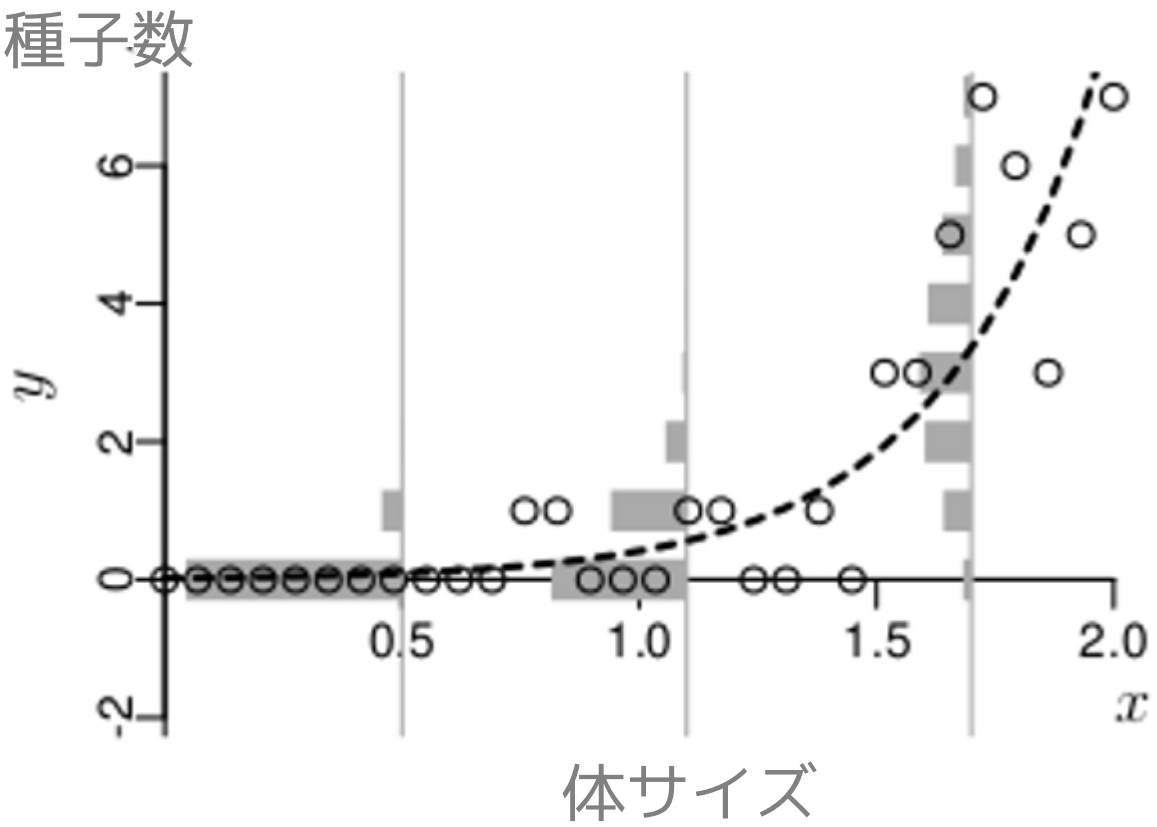
もうすこし「わかった」ような気分?

種子数の平均値はサイズ x とともに増大する

➡ どのように変化するのか?
数式で書くとどうなる?

平均値が増大するとばらつきが変化する

➡ どのようにばらつくのか?
確率分布?



統計モデルをデータにうまくあてはめる

➡ どのようにあてはめるのが妥当なのか? パラメーター推定法?

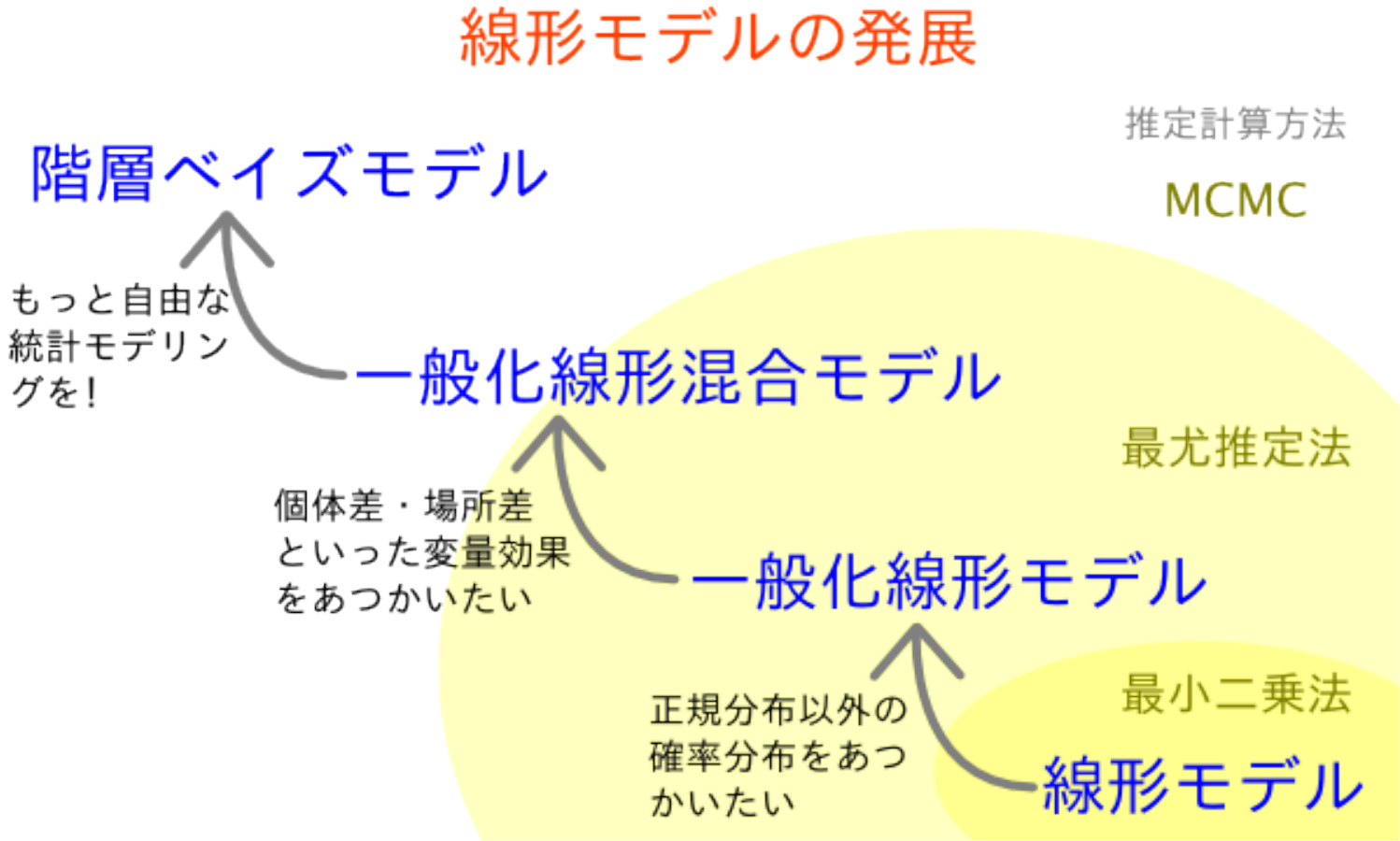
「統計モデル」とは何か?

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!



一般化線形モデル - ばらつきをよく見る

Generalized Linear Model, GLM

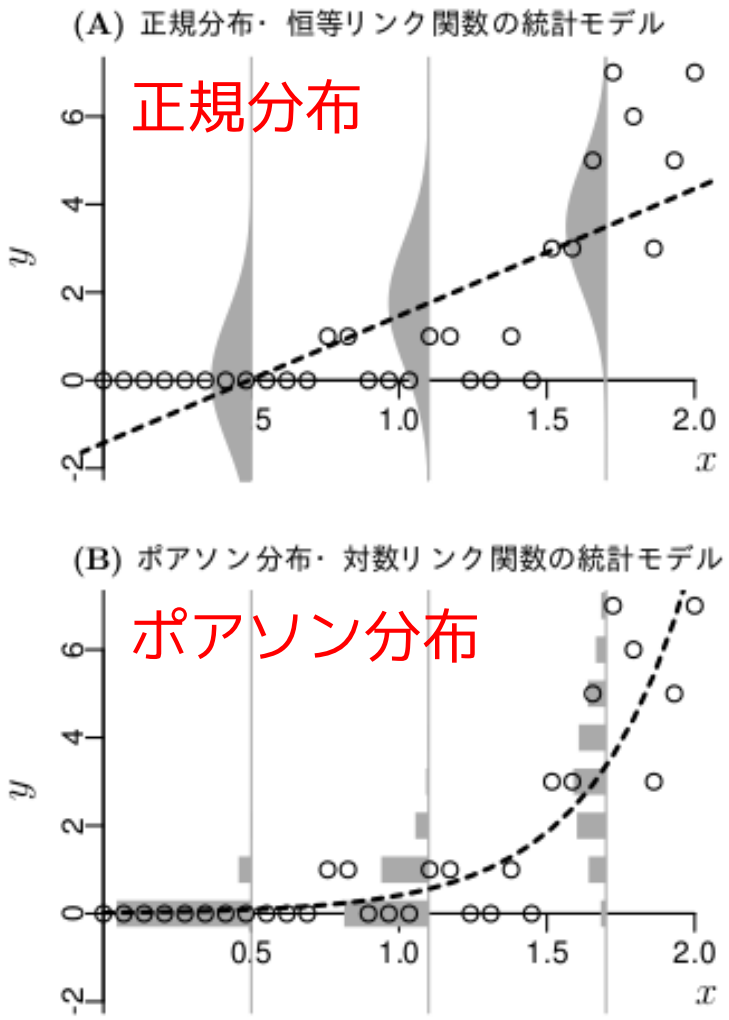


図 3.9 回帰モデルと確率分布の関係．また別の架空データに対して GLM をあてはめた例．破線は x とともに変化する平均値．グレイで

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

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!!

全体の流れ

10/1（水）

第 1 講時（a）観測されたパターンを説明する統計モデル

+（b）確率分布と最尤推定

第 2 講時:（c）一般化線形モデル: ポアソン回帰

第 3 講時:（d）モデル選択と検定

第 4 講時:（e）一般化線形モデル: ロジスティック回帰

第 5 講時:（f）階層ベイズモデル

10/2（木）第 1,2 講時: 情報基盤センターで R 実習（r1, b-f）

10/1（水）第 1 講時

統計モデリング入門 2014 (2)

確率分布と最尤推定

久保拓弥 kubo@ees.hokudai.ac.jp

茨城大集中講義 http://goo.gl/2QNwgl

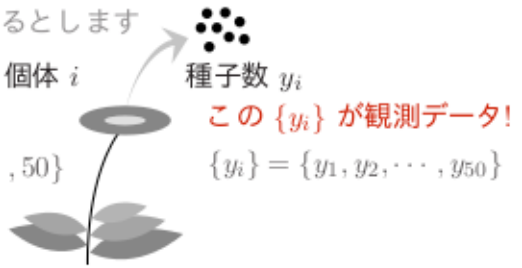
2014–10–01

ファイル更新時刻: 2014–09–25 14:39

単純化した例題

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

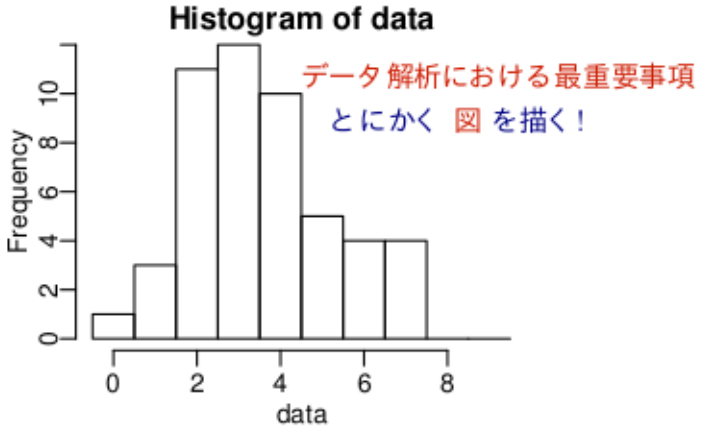


このデータ $\{y_i\}$ がすでに R に格納されていた，としましょう

```
> 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
```


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))
```



データ解析における最重要事項
とにかく 図 を描く!

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

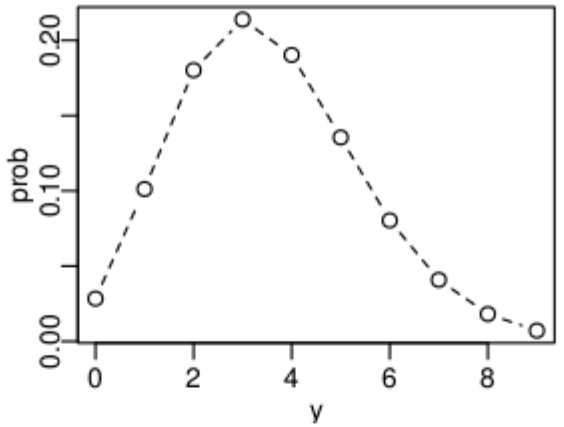


図 4　平均 $\lambda = 3.56$ のポアソン分布．種子数 y とその確率 prob の関係が示されている．図 3 の表を図にしたもの．R の plot() 関数の引数，type = "b" によって「丸と折れ線による図示」，lty = 2 によって「折れ線は破線で」と指示している．

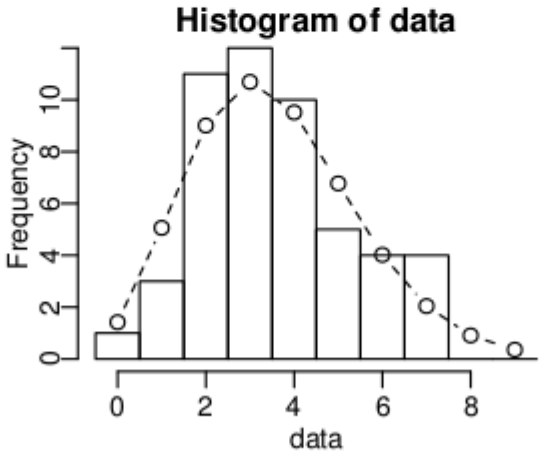


図 5　観測データと確率分布の対応をながめる．ヒストグラムは図 2 と同じ．それに重ねられている丸と破線は y 個の種子をもつ個体数の予測．平均 3.56 の図 4 のポアソン分布の確率分布に全個体数 50 をかけて得られる．

さいゆう最尤推定という考えかたを説明します

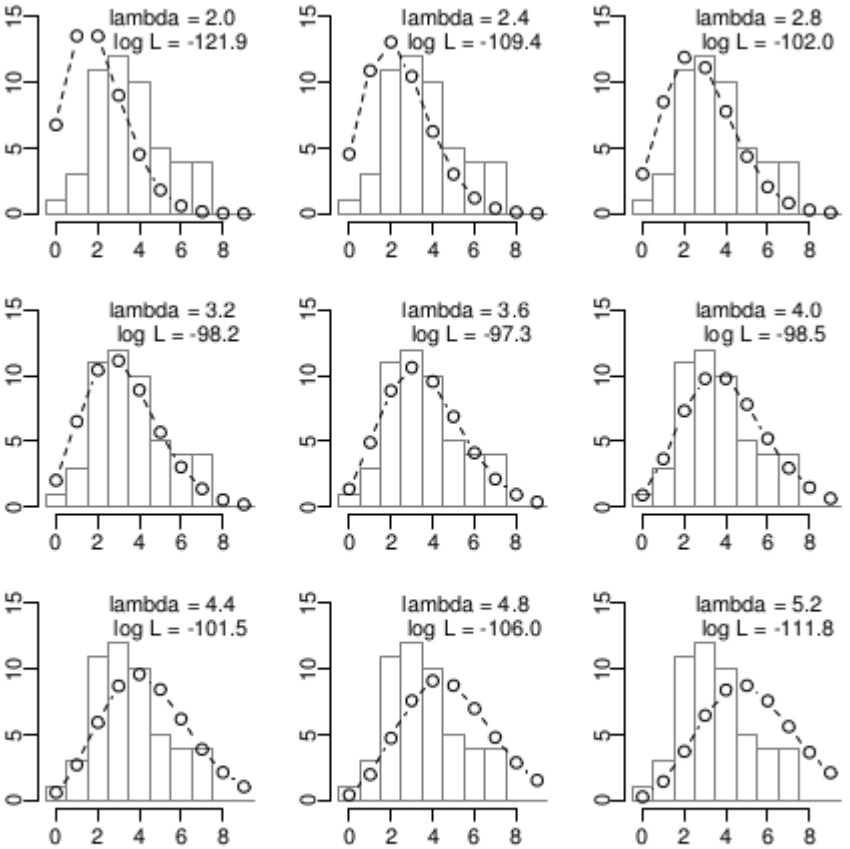
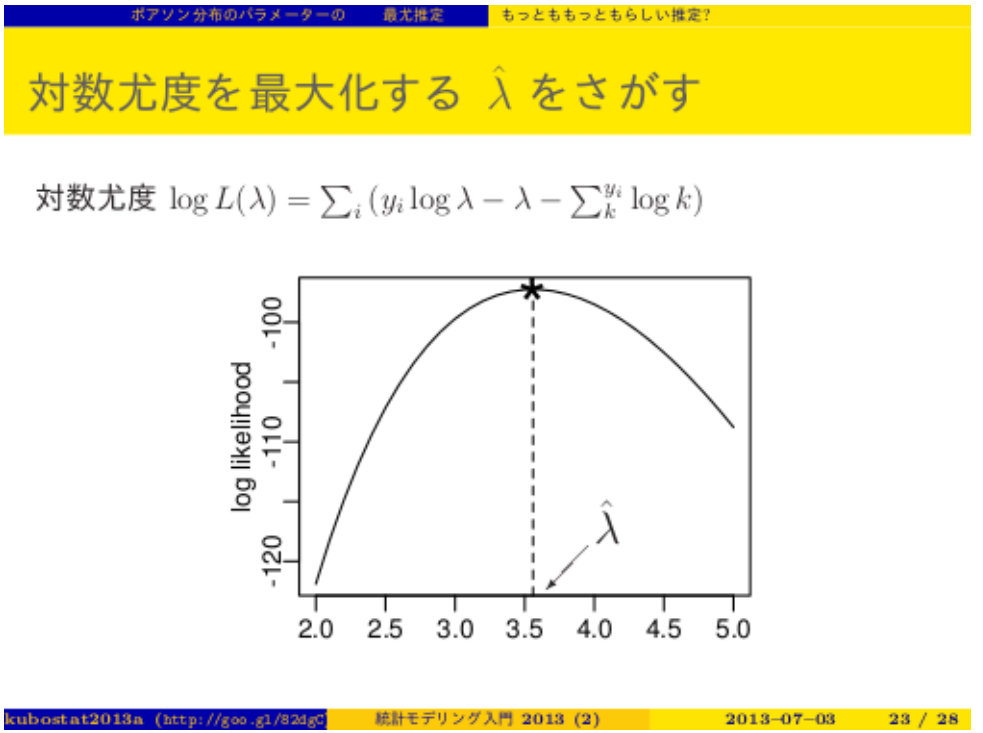


図 7　平均 λ（lambda）を変化させていったポアソン分布と，観測データへのあてはまりの良さ（対数尤度 logL）．すべてのヒストグラムは図 2 と同じ．



10/1（水）第 2 講時

統計モデリング入門 2014 (3)

一般化線形モデル: ポアソン回帰

久保拓弥 kubo@ees.hokudai.ac.jp

茨城大集中講義 http://goo.gl/2QNwgl

2014–10–01

ファイル更新時刻: 2014–09–29 12:44

ここで登場する ---
「何でも正規分布」ではダメ！という発想

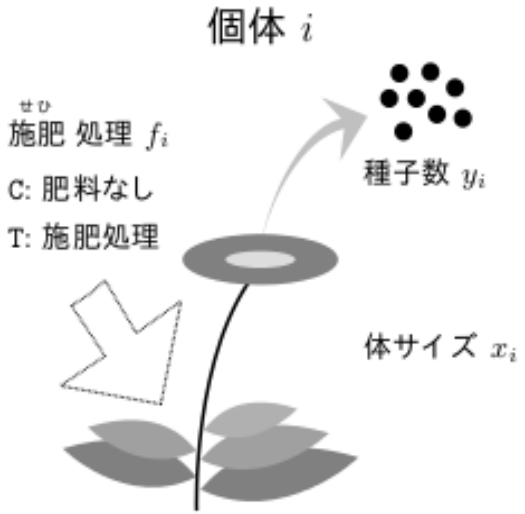


図 3.1　この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

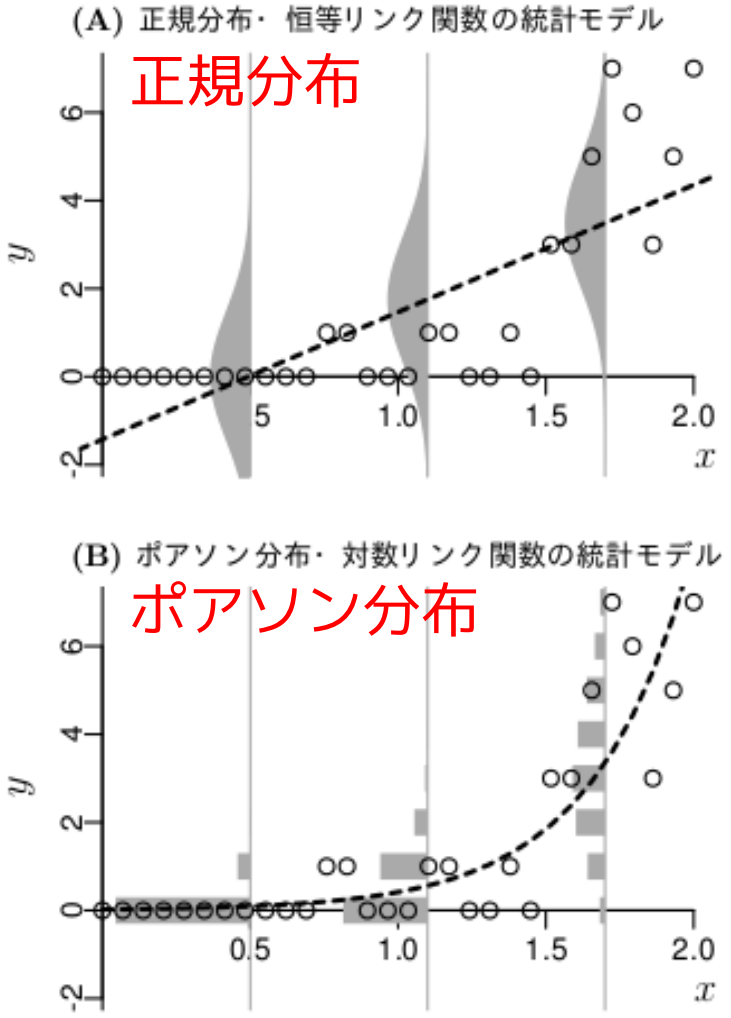


図 3.9　回帰モデルと確率分布の関係．また別の架空データに対して GLM をあてはめた例．破線は x とともに変化する平均値．グレイで

Free の統計ソフトウェア

R で統計モデリング

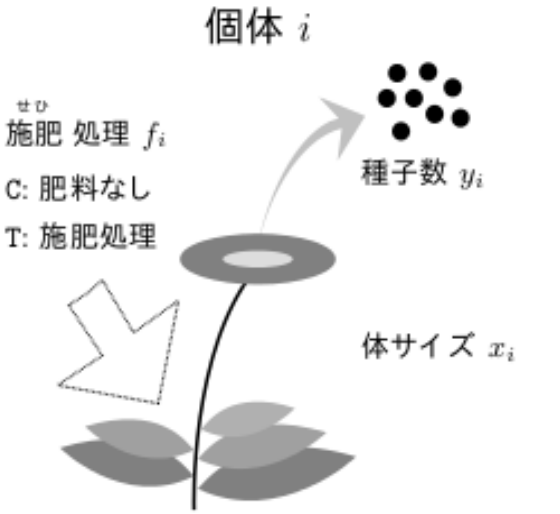
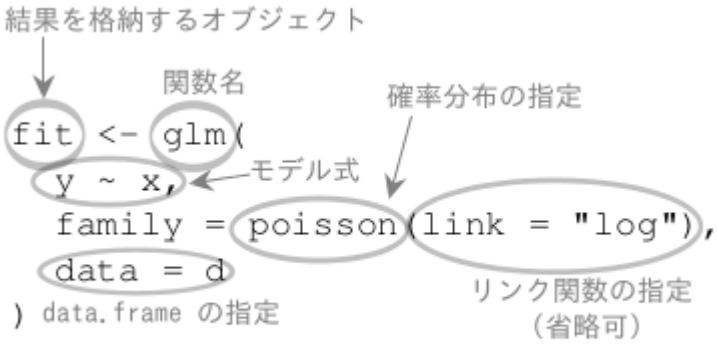


図 3.1　この例題に登場する架空植物の第 i 番目の個体
体サイズ（ 個体の大きさ）x_i と肥料をやる施肥処理
にどう影響しているのかを知りたい.



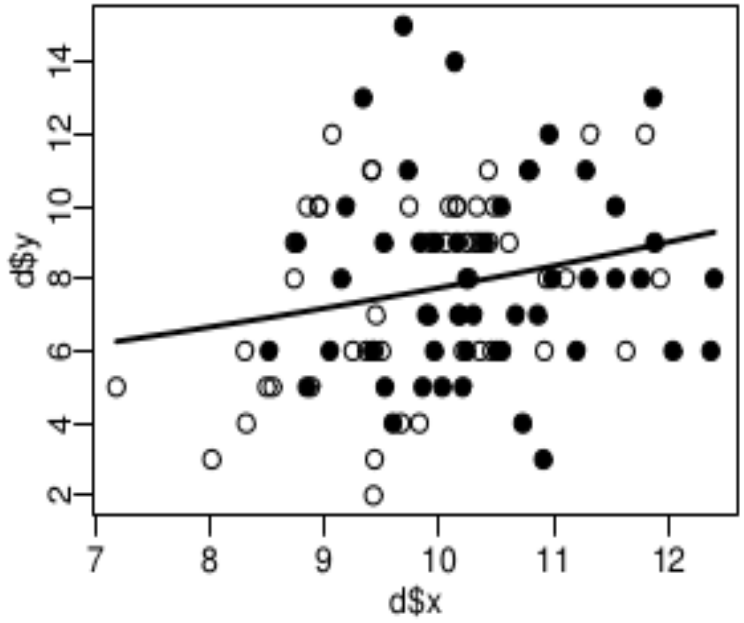


図 17　平均種子数 λ の予測. 図 12 に λ の予測値（実線）を上がきしたもの.

10/1（水）第 3 講時

統計モデリング入門 2014 (4)

モデル選択と検定

久保拓弥 kubo@ees.hokudai.ac.jp

茨城大集中講義 http://goo.gl/2QNwgl

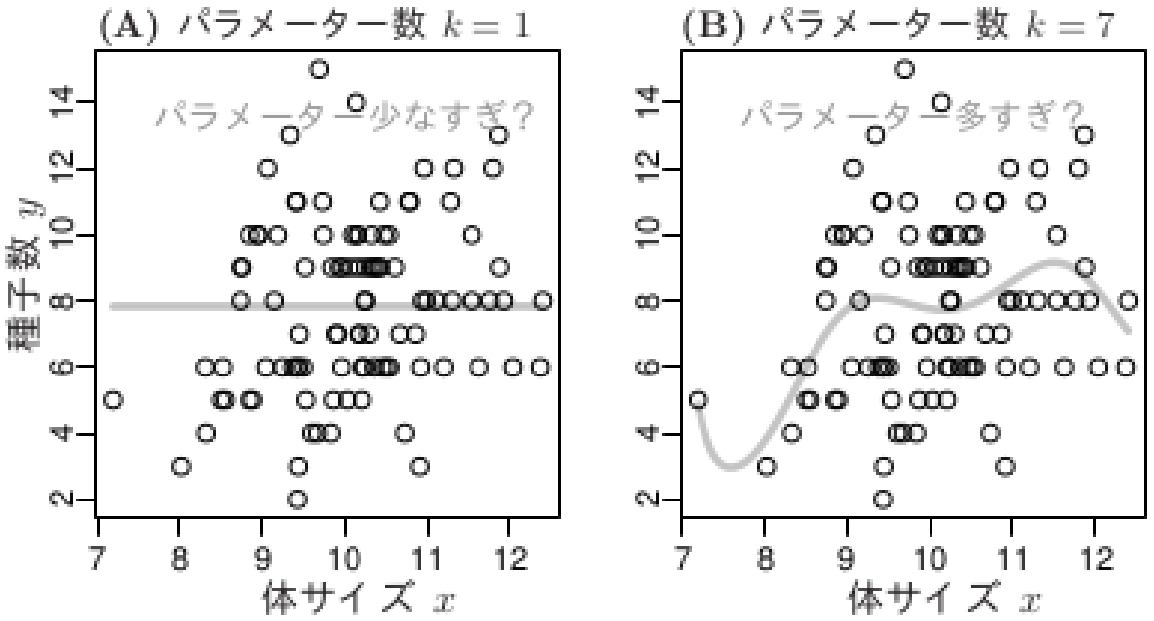
2014–10–01

ファイル更新時刻: 2014–09–29 12:44

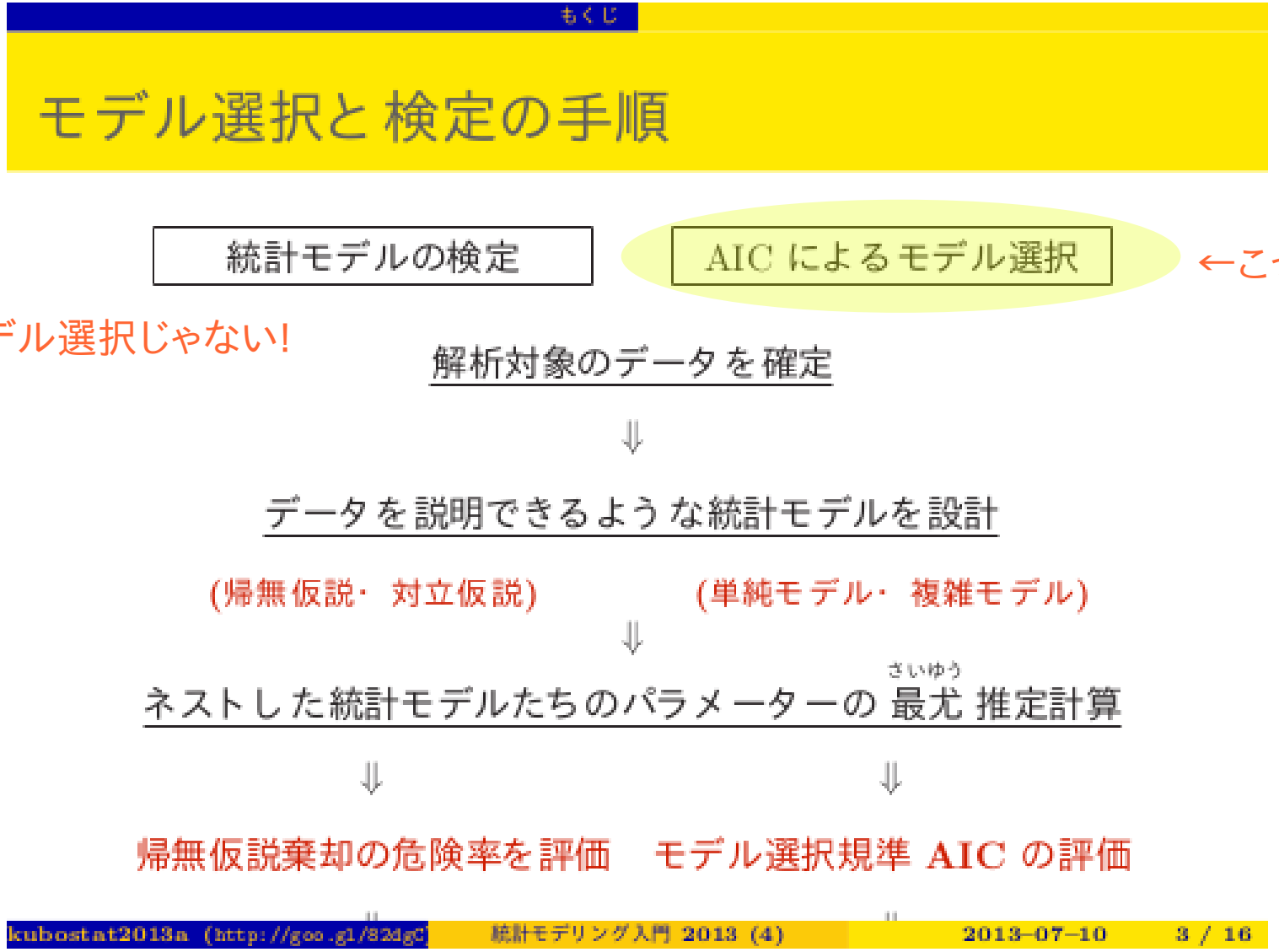
Q. モデル選択とは何か？

データと確率分布の対応　どういう関係なのか図示してながめる

パラメーター数は多くても少なくてもヘン？



A. より良い予測をする統計モデルを探すこと



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?

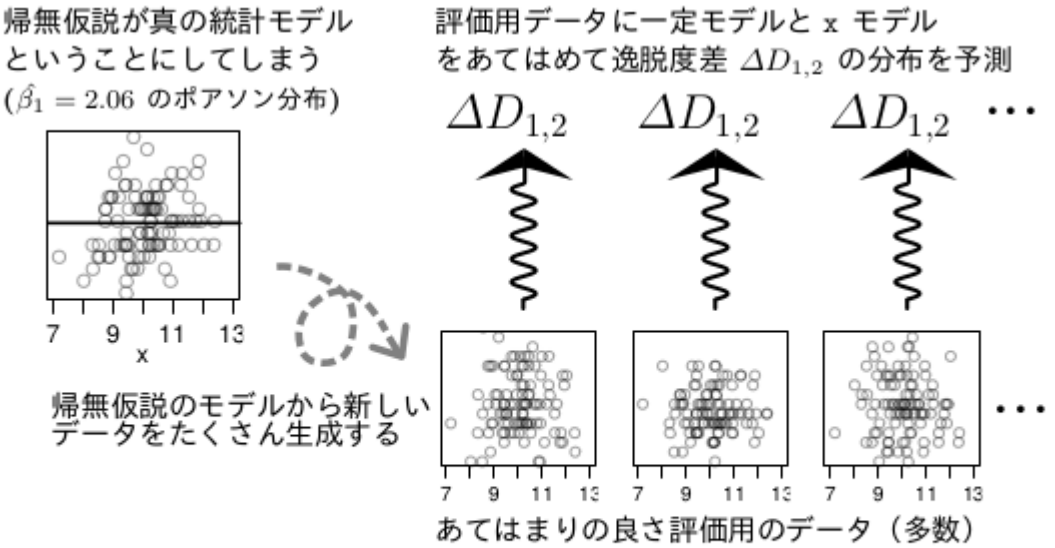


図 6　尤度比検定に必要な $\Delta D_{1,2}$ の分布の生成．まず帰無仮説である一定モデル（$\hat{\beta}_1 = 2.06$, p.6 参照）が真の統計モデルだと仮定し，そこから得られるデータを使って逸脱度差 $\Delta D_{1,2}$ がどのような分布になるかを調べる．

10/1（水）第 4 講時

統計モデリング入門 2014 (5)

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

茨城大集中講義 http://goo.gl/2QNwgl

2014–10–01

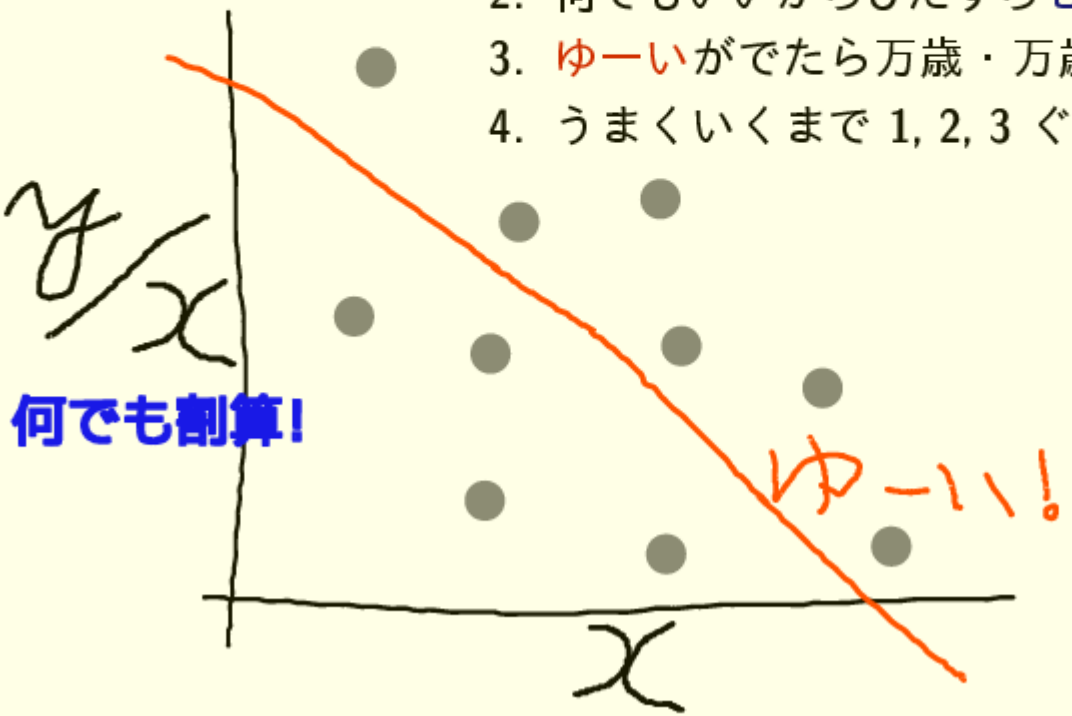
ファイル更新時刻: 2014–09–29 12:44

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めできない作法の例

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



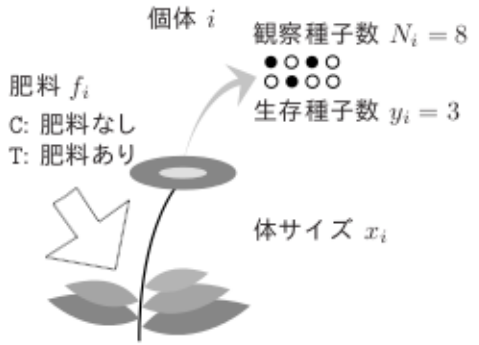
ちなみにこれは x と o/x を比較してるんだから，反比例みたいな偽「負の相関」がでるのはあたりまえ

2012–11–02 k4 (2012–10–26 17:07 修正版) 14/ 44

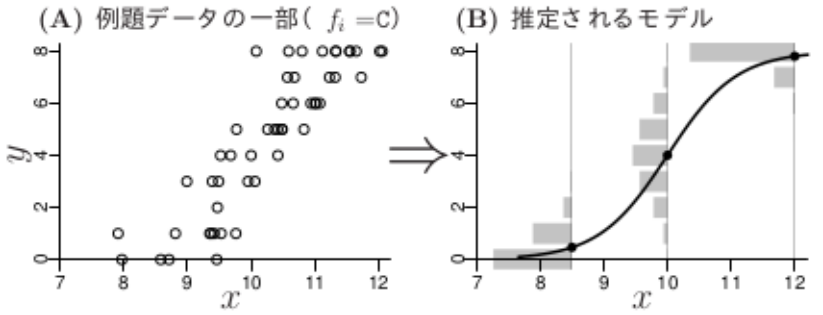
GLM のひとつ，ロジスティック回帰を使おう

またいつもの例題? …… ちょっとちがう

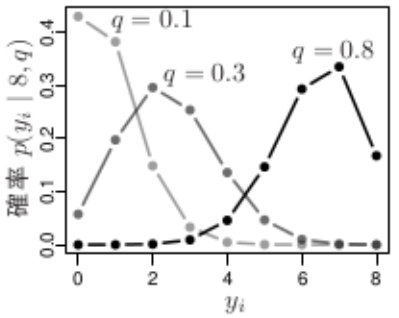
8 個の種子のうち y 個が 発芽可能 だった! …… というデータ



ロジスティック回帰とは何なのか?



二項分布: N 回のうち y 回，となる確率



10/1（水）第 5 講時

統計モデリング入門 2014 (6)

階層ベイズモデル (先端科学トピックス)

久保拓弥 kubo@ees.hokudai.ac.jp

茨城大集中講義 http://goo.gl/2QNwgl

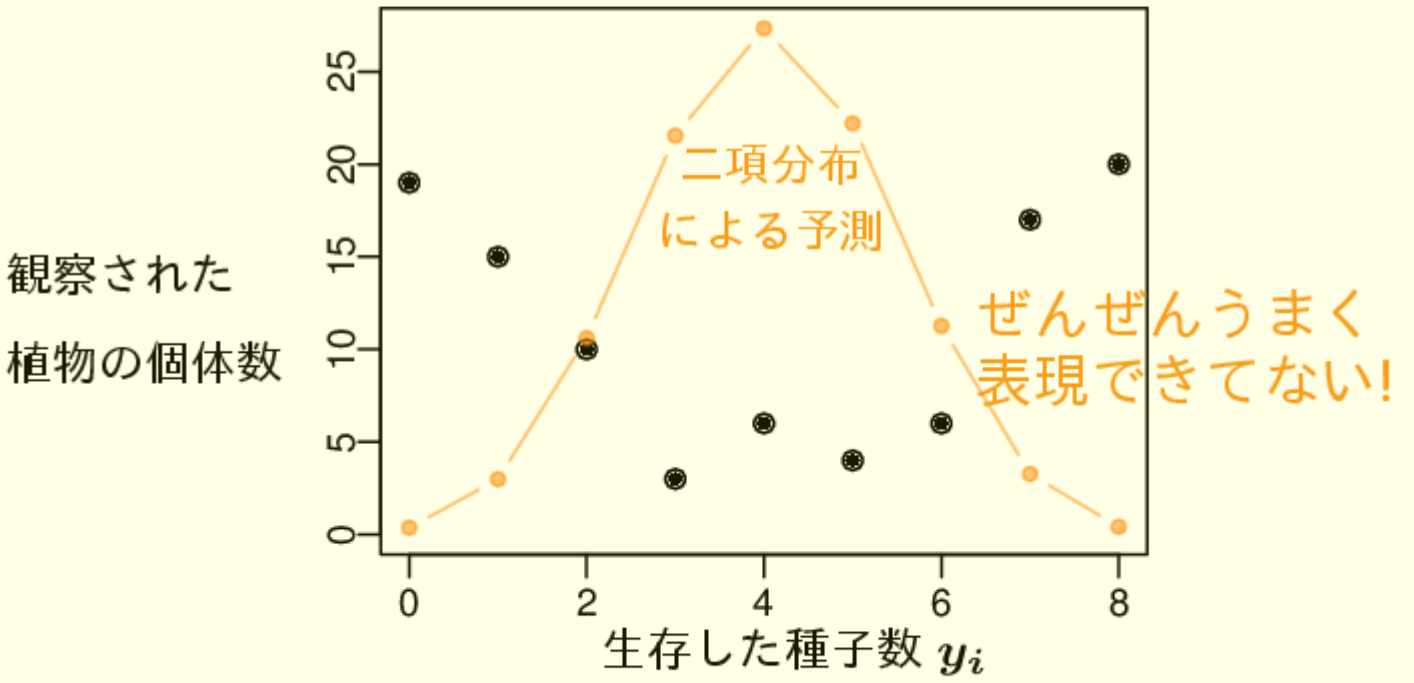
2014–10–01

ファイル更新時刻: 2014–09–29 12:44

GLM ではうまく説明できないデータ!?

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……



要点: 現実のデータを解析するためには，さらに工夫が必要!

現実のデータの複雑さを表現できる統計モデルをつくる!

個体差・地域差・生物種差・
空間相関・時間相関など
めんどうなことをあつかわない
といけない

GLM にそういう要因を組みこむ

データに複雑なモデルをあてはめる工夫をする (パラメーター推定法の改善)