統計モデリング入門 2013（1）

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

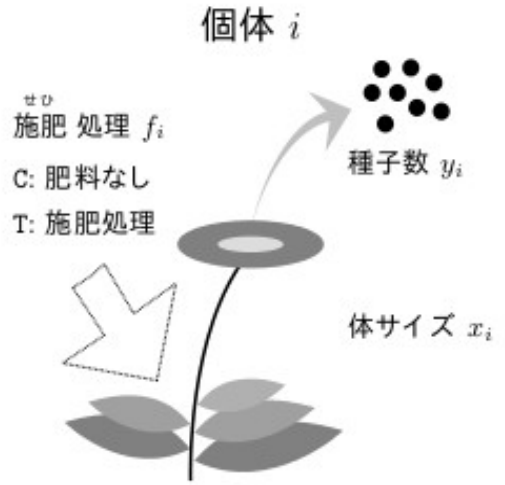


図 3.1 この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

統計モデリング授業の web page

http://goo.gl/82dgC

更新: 2013-06-27 15:24:25

生態学のデータ解析 - 統計学授業 2013

- 統計学の授業やります (2013 年度前期後半, 2013 年 7 月)
  - 「植物生態学特論 I」の一部, 2013 年の 7/1 (月) から開始
    - 誰でも参加できます
    - 事前のもうしこみなどは何も必要ありません (聴講のみで，単位を必要としない場合)
    - 全部ではなく部分的に参加してもらってもかまいません
  - 教科書「統計モデリング入門」
    - 参考: 2008 年の講義のーと
  - 月曜日・水曜日の 3 講目 (13:00-14:30)
  - 教室: 地球環境科学研究院 A 棟 A809 (エレヴェイターでてすぐ前の部屋)
  - 短縮 URL: http://goo.gl/82dgC

[おもな内容]

- 第 1 回: 7/01 (月) 観測されたパターンを説明する統計モデル
- 第 2 回: 7/03 (水) 確率分布と最尤推定
- 第 3 回: 7/08 (月) 一般化線形モデル: ポアソン回帰
- 第 4 回: 7/10 (水) モデル選択と検定
- 第 5 回: 7/17 (水) 一般化線形モデル: ロジスティック回帰

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

- 授業登録している人たちは自動的に ML に登録します
  - I will subscribe all registrants of the class to the kubostat ML.
- ML を使って各回の「課題」を出します
  - 回答もメイ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

The main language of this class is Japanese ⋯ Sorry

- I didn't know foreign students are in the class.
- Why in Japanese? ⋯ because even in Japanese, statistics is difficult for Japanese students to understand.
- I compensate for language disadvantages in foreign students when I give grades.
- Questions in English is welcomed.

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

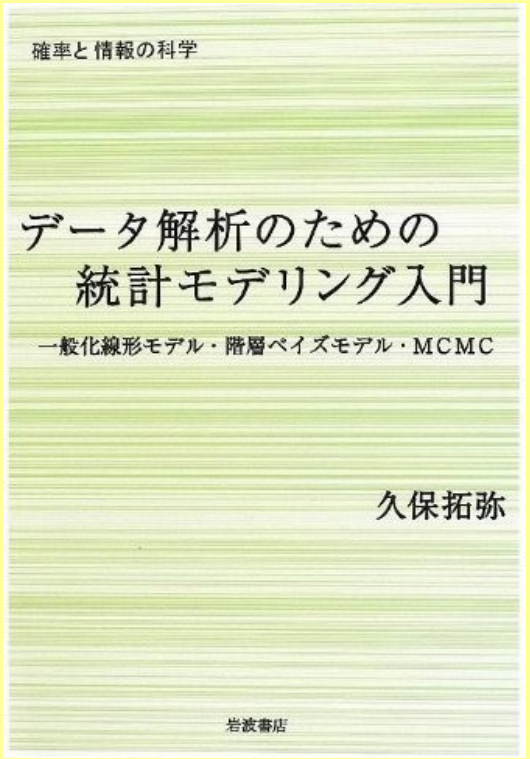
- ブラックボックス統計解析
- とにかく「ゆーい差」さえ出せばよいという発想になっている
- 大学・大学院でもあまりちゃんと教えられていない，教えられるヒトが少ない……とくに近年発達している統計モデリングについて

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!

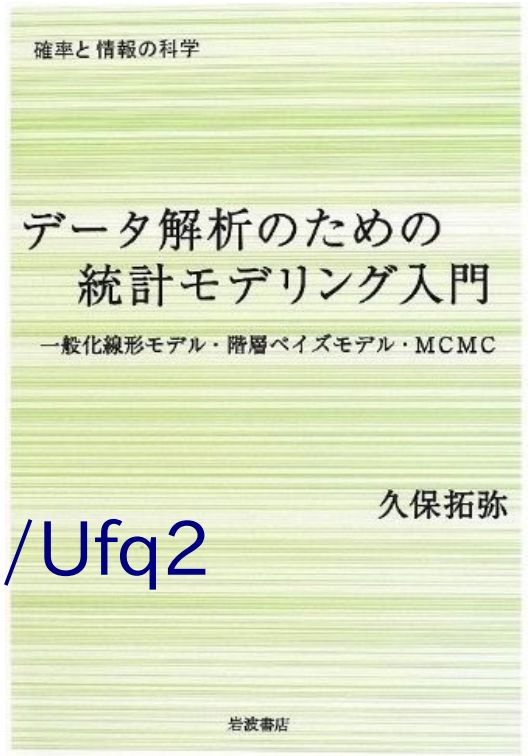
- データ解析で使われるの中でも比較的簡単な統計モデルを理解しよう
- 「ゆーい差」さえ出せばよいという発想をやめて，データと統計モデルの対応関係をよく見よう（作図重要）
- 統計ソフトウェア R を使い始めよう

教科書とソフトウェア

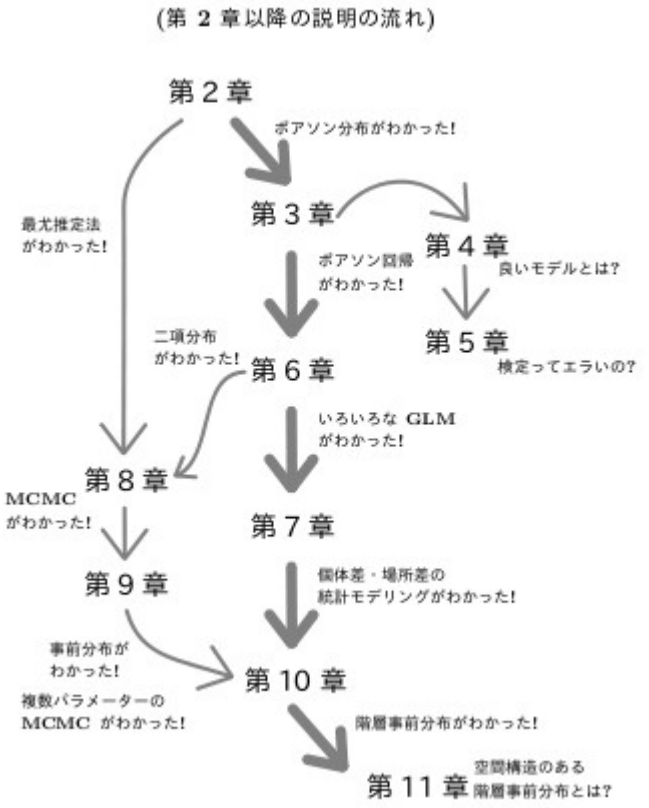

確率と情報の科学
データ解析のための
統計モデリング入門
一般化線形モデル・階層ベイズモデル・MCMC
久保拓弥
岩波書店

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

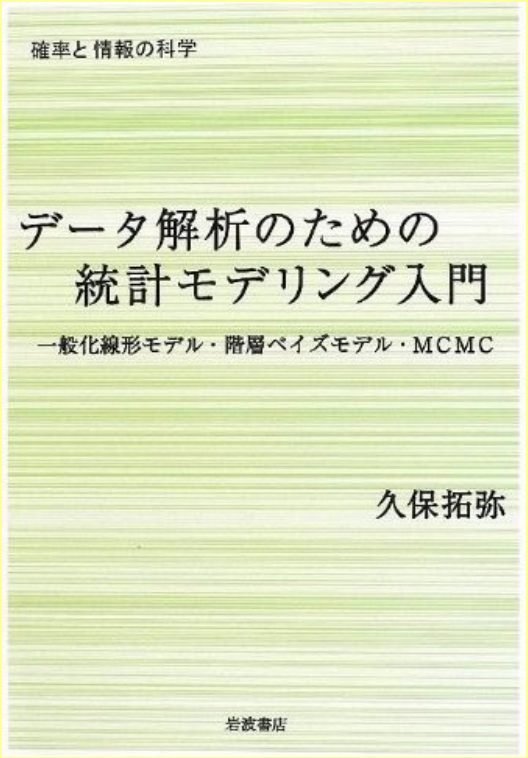
統計ソフトウェア R

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

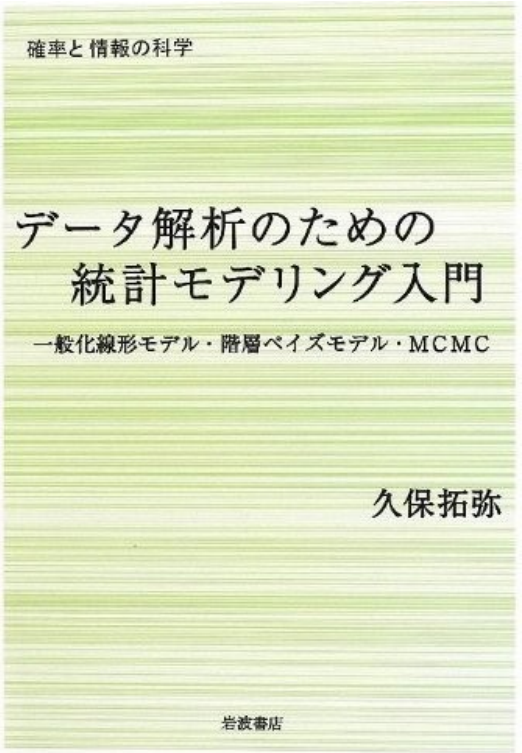
この教科書でも R を
使って問題を解決する
方法を説明しています

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

「統計モデル」とは何か?

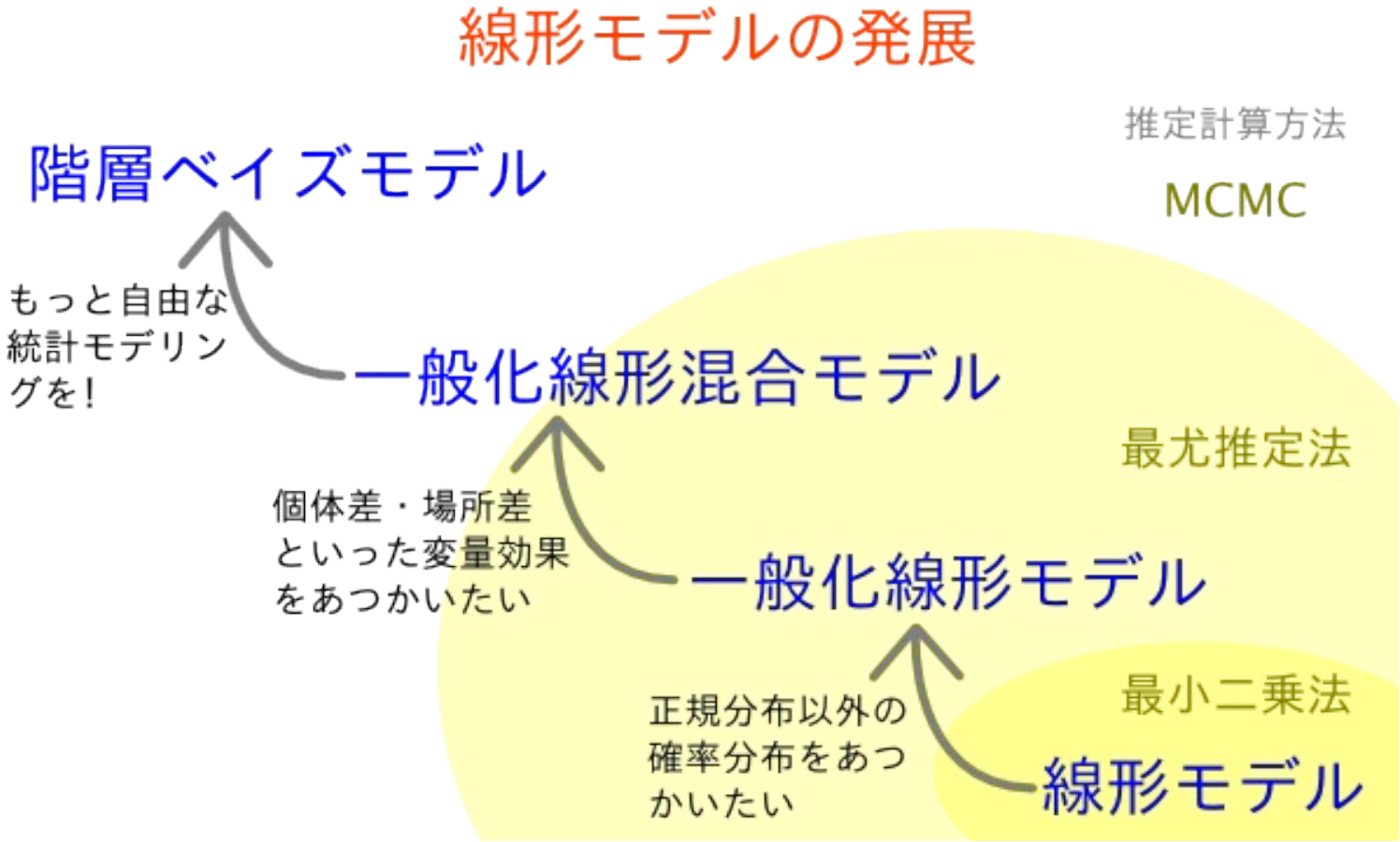
どんな統計解析においても
統計モデルが使用されている



- 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

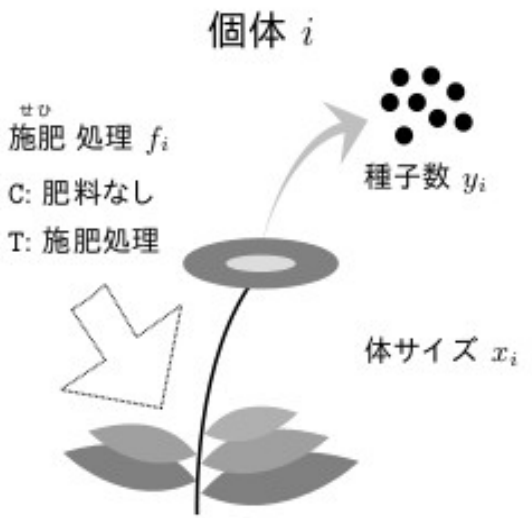
「統計モデリング入門」の主張

「何でも正規分布」じゃないだろ!



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）



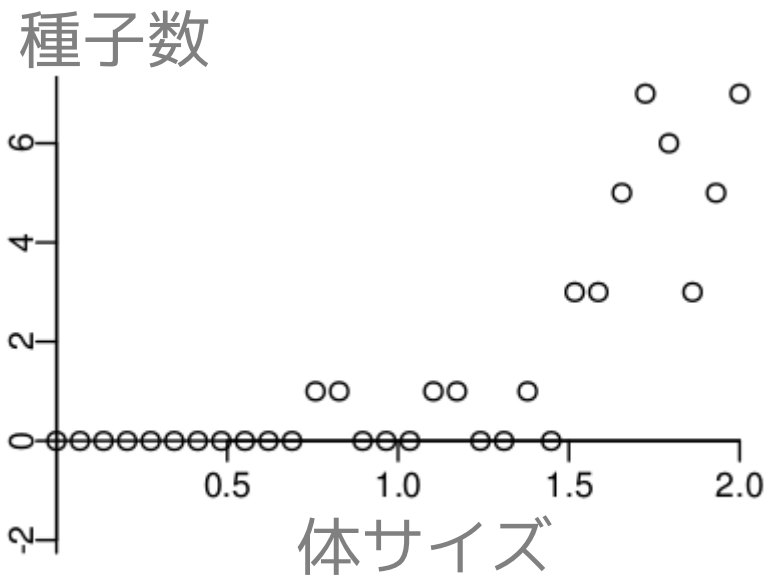


図 3.1　この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（ 個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

一般化線形モデル - ばらつきをよく見る

Generalized Linear Model, GLM

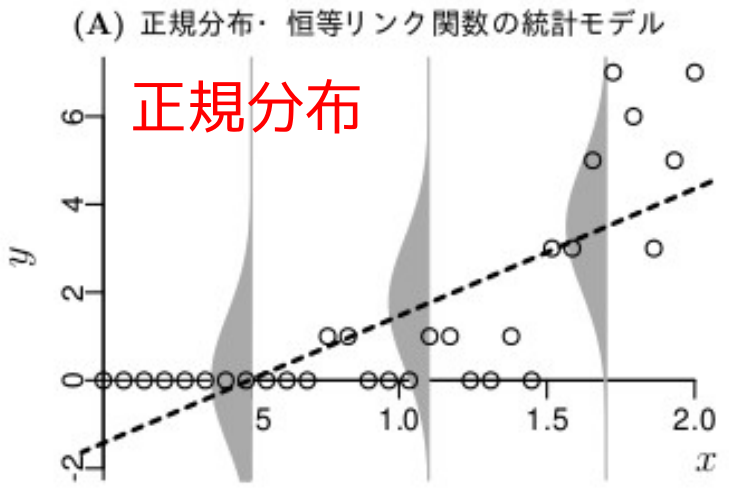


図 3.9 回帰モデルと確率分布の関係．また別の架空データに対して GLM をあてはめた例．破線は x とともに変化する平均値．グレイで

線形モデルの発展

推定計算方法

階層ベイズモデル

MCMC

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

最小二乗法

正規分布以外の確率分布をあつかいたい

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

第 1 回: 7/01 (月) 観測されたパターンを説明する統計モデル

第 2 回: 7/03 (水) 確率分布と最尤推定

第 3 回: 7/08 (月) 一般化線形モデル: ポアソン回帰

第 4 回: 7/10 (水) モデル選択と検定

第 5 回: 7/17 (水) 一般化線形モデル: ロジスティック回帰

第 6 回: 7/22 (月) 一般化線形混合モデル

第 7 回: 7/24 (水) 階層ベイズモデル

7/3（水）

統計モデリング入門 2013 (2)

確率分布と最尤推定

久保拓弥 kubo@ees.hokudai.ac.jp

北大環境科学院の講義 http://goo.gl/82dgC

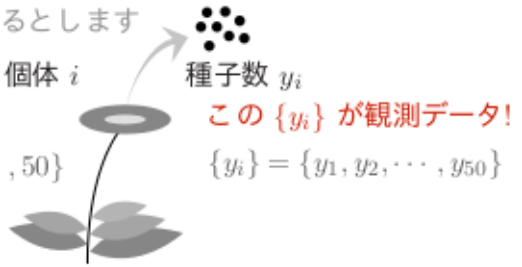
2013–07–03

ファイル更新時刻: 2013–06–25 15:00

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

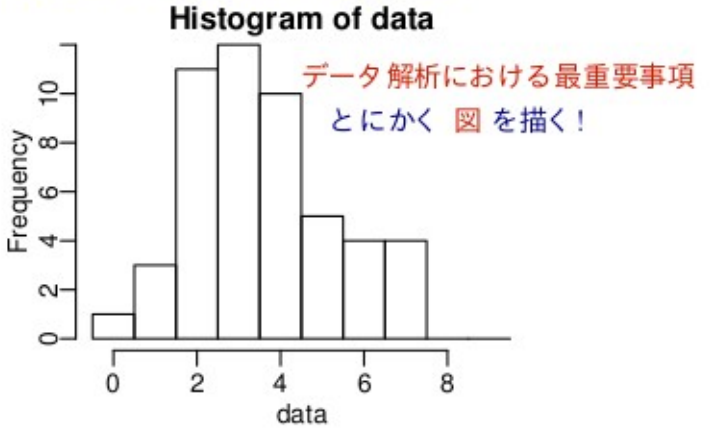


このデータ $\{y_i\}$ がすでに R に格納されていた，としましょう

```
> 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

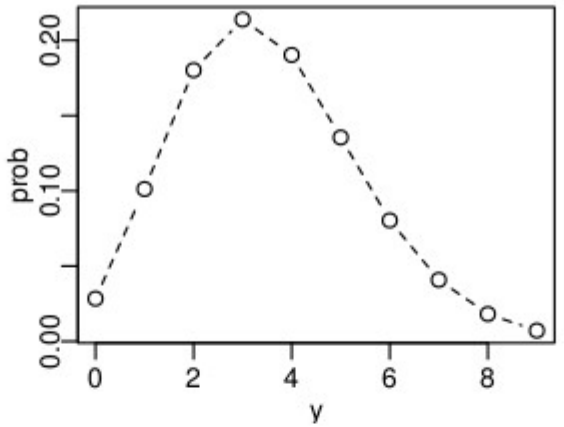


図 4　平均 $\lambda = 3.56$ のポアソン分布．種子数 y とその確率 prob の関係が示されている．図 3 の表を図にしたもの．R の plot() 関数の引数，type = "b" によって「丸と折れ線による図示」，lty = 2 によって「折れ線は破線で」と指示している．

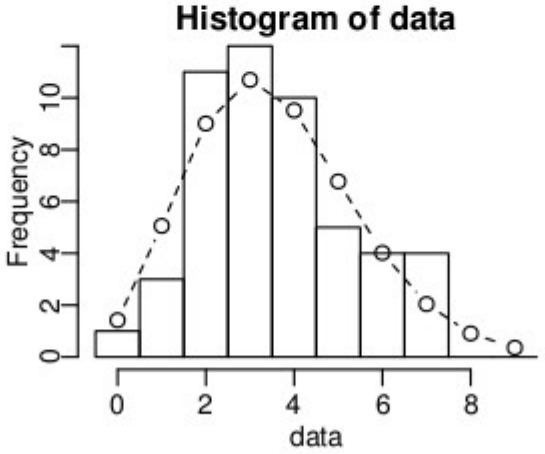


図 5　観測データと確率分布の対応をながめる．ヒストグラムは図 2 と同じ．それに重ねられている丸と破線は y 個の種子をもつ個体数の予測．平均 3.56 の図 4 のポアソン分布の確率分布に全個体数 50 をかけて得られる．

さいゆう最尤推定という考えかたを説明します

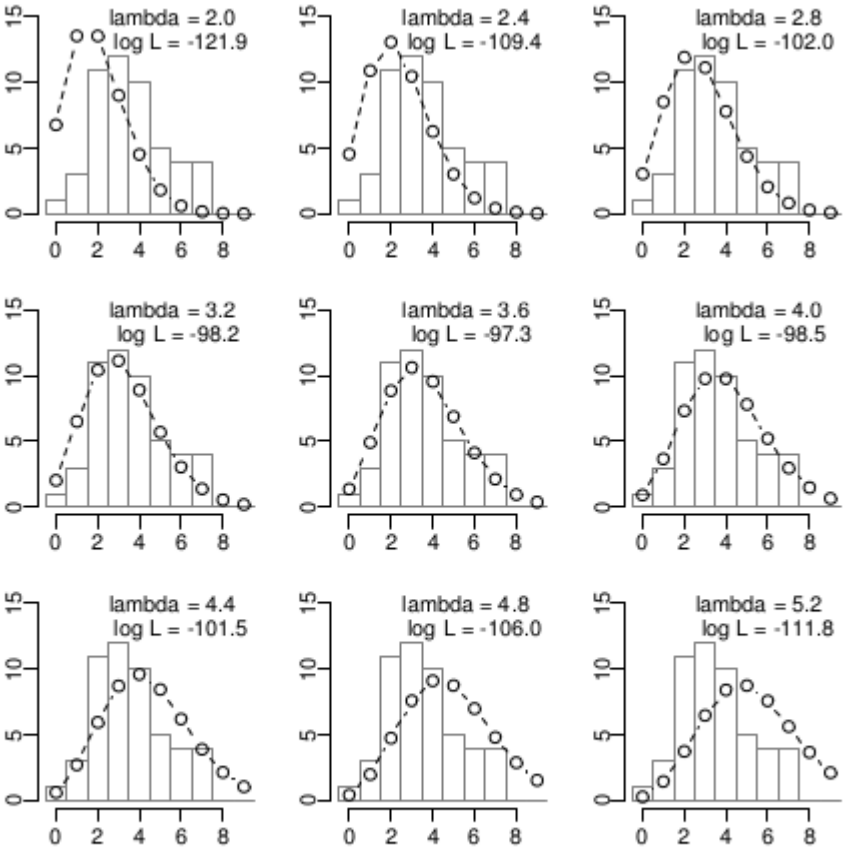
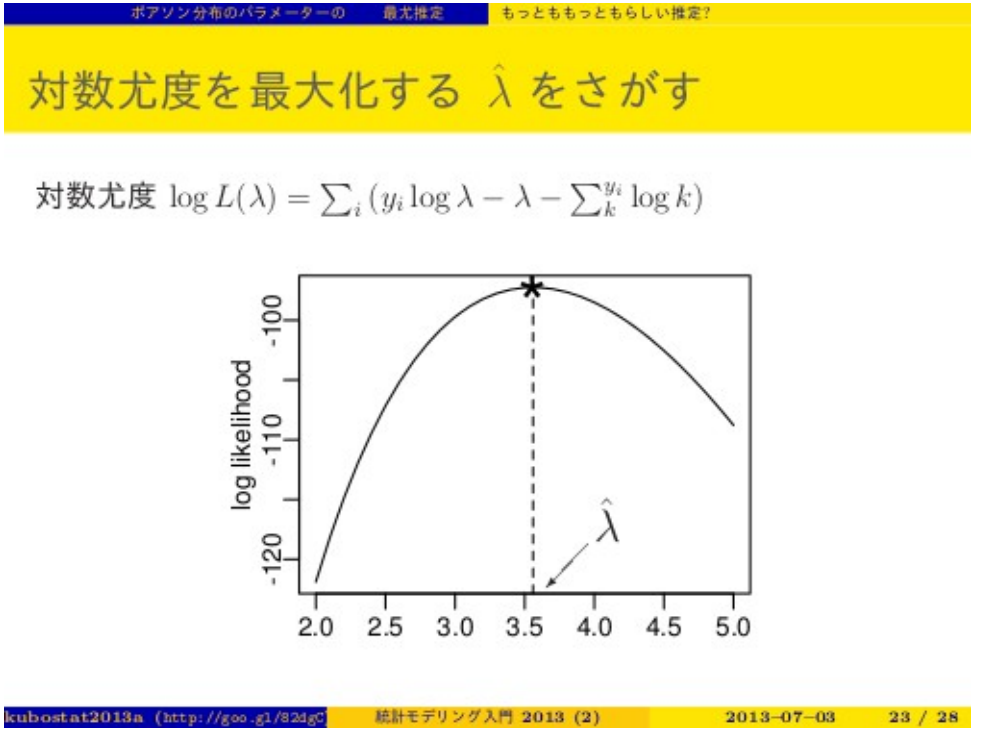


図 7　平均 λ（lambda）を変化させていったポアソン分布と，観測データへのあてはまりの良さ（対数尤度 logL）．すべてのヒストグラムは図 2 と同じ．



7/8（月）

統計モデリング入門 2013 (3)

一般化線形モデル: ポアソン回帰

久保拓弥 kubo@ees.hokudai.ac.jp

北大環境科学院の講義 http://goo.gl/82dgC

2013–07–08

ファイル更新時刻: 2013–06–25 15:24

ここで登場する ---
「何でも正規分布」ではダメ！という発想

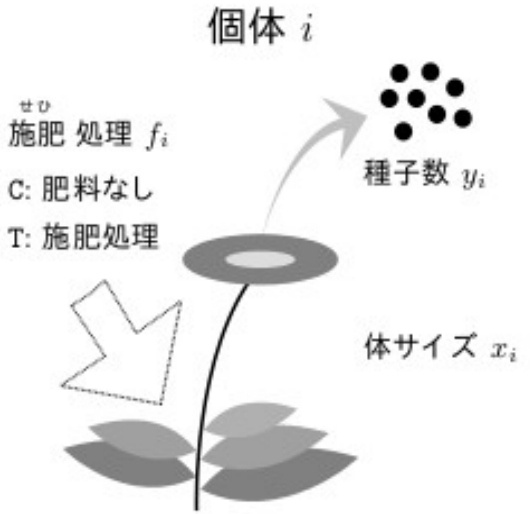


図 3.1　この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

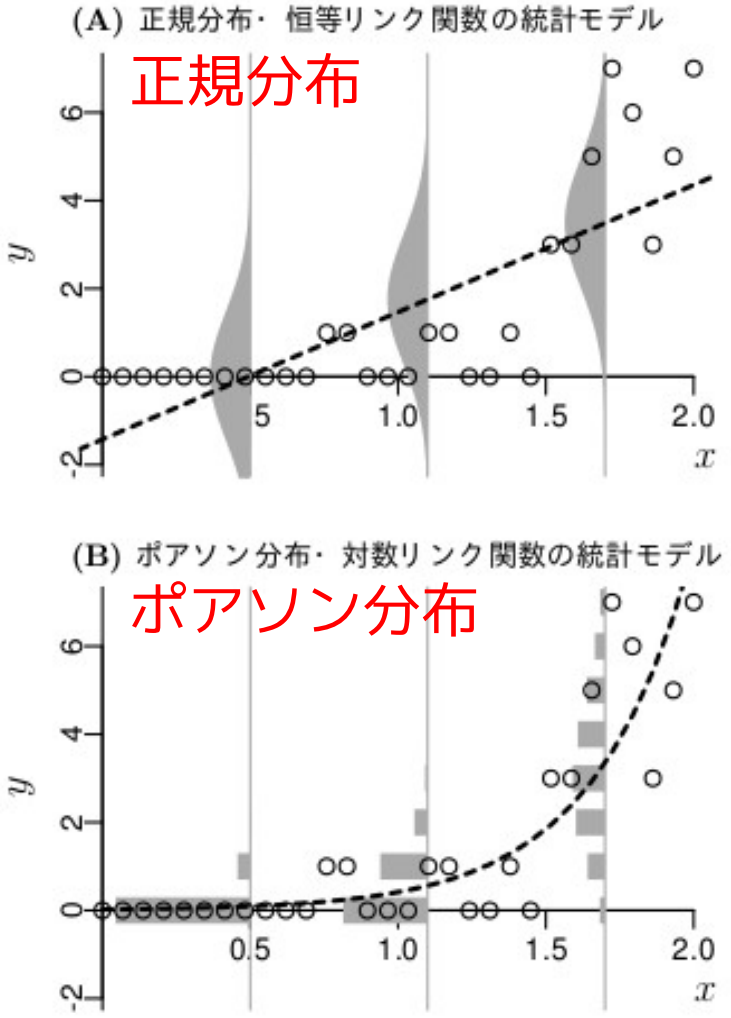


図 3.9　回帰モデルと確率分布の関係．また別の架空データに対して GLM をあてはめた例．破線は x とともに変化する平均値．グレイで

Free の統計
ソフトウェア **R** で統計モデリング

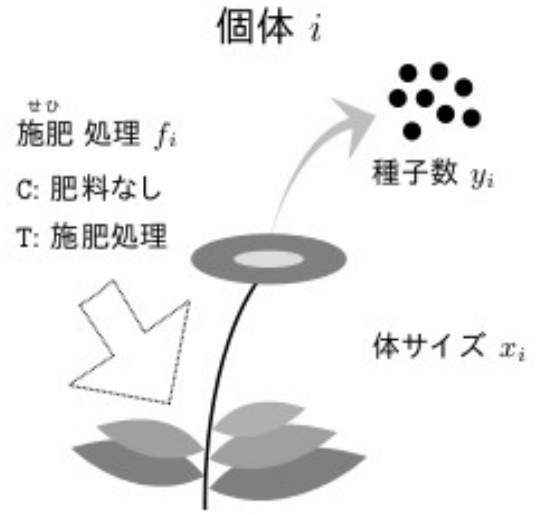
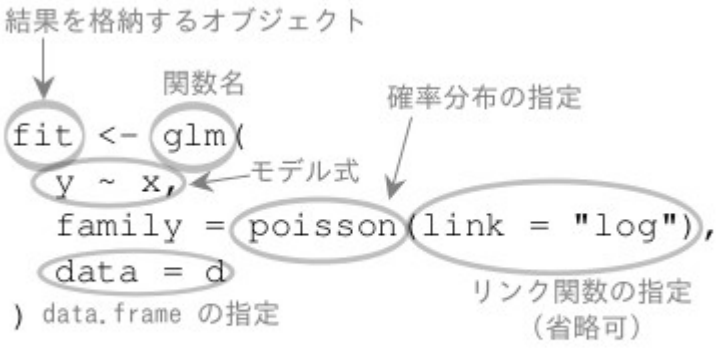


図 3.1　この例題に登場する架空植物の第 i 番目の個体
体サイズ（ 個体の大きさ）x_i と肥料をやる施肥処理
にどう影響しているのかを知りたい.



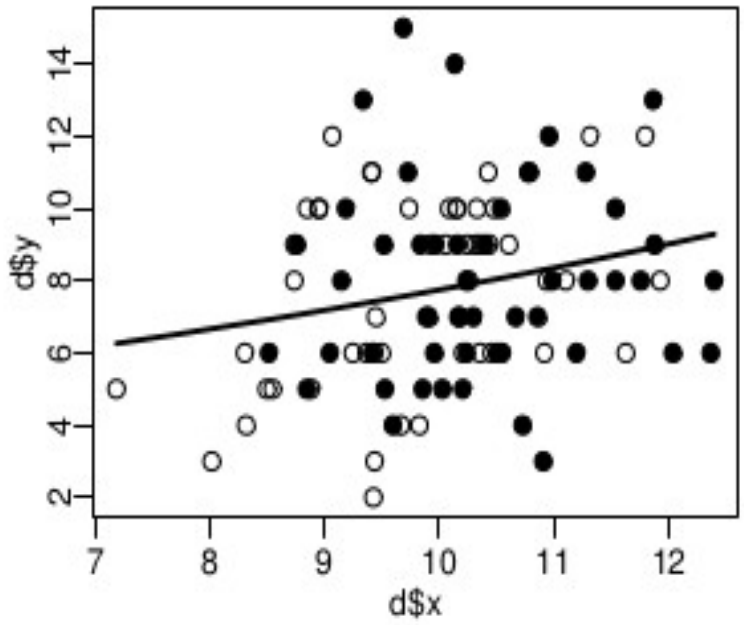


図 17　平均種子数 λ の予測. 図 12 に λ の予測値（実線）を上がきしたもの.

7/10（水）

統計モデリング入門 2013 (4)

モデル選択と検定

久保拓弥 kubo@ees.hokudai.ac.jp

北大環境科学院の講義 http://goo.gl/82dgC

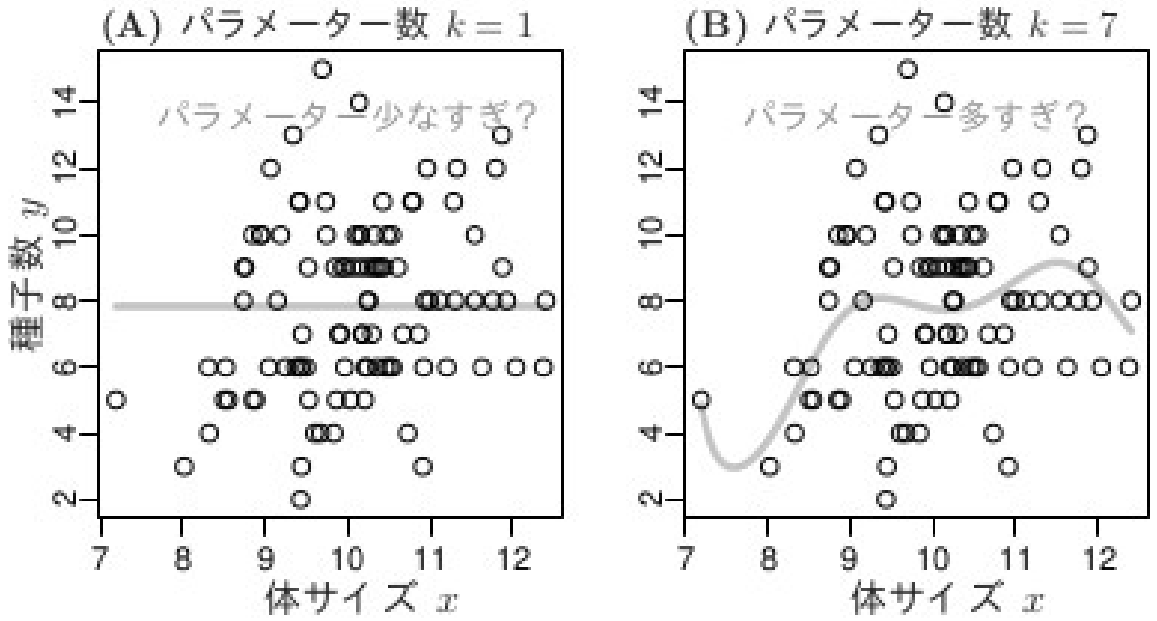
2013–07–10

ファイル更新時刻: 2013–06–25 15:44

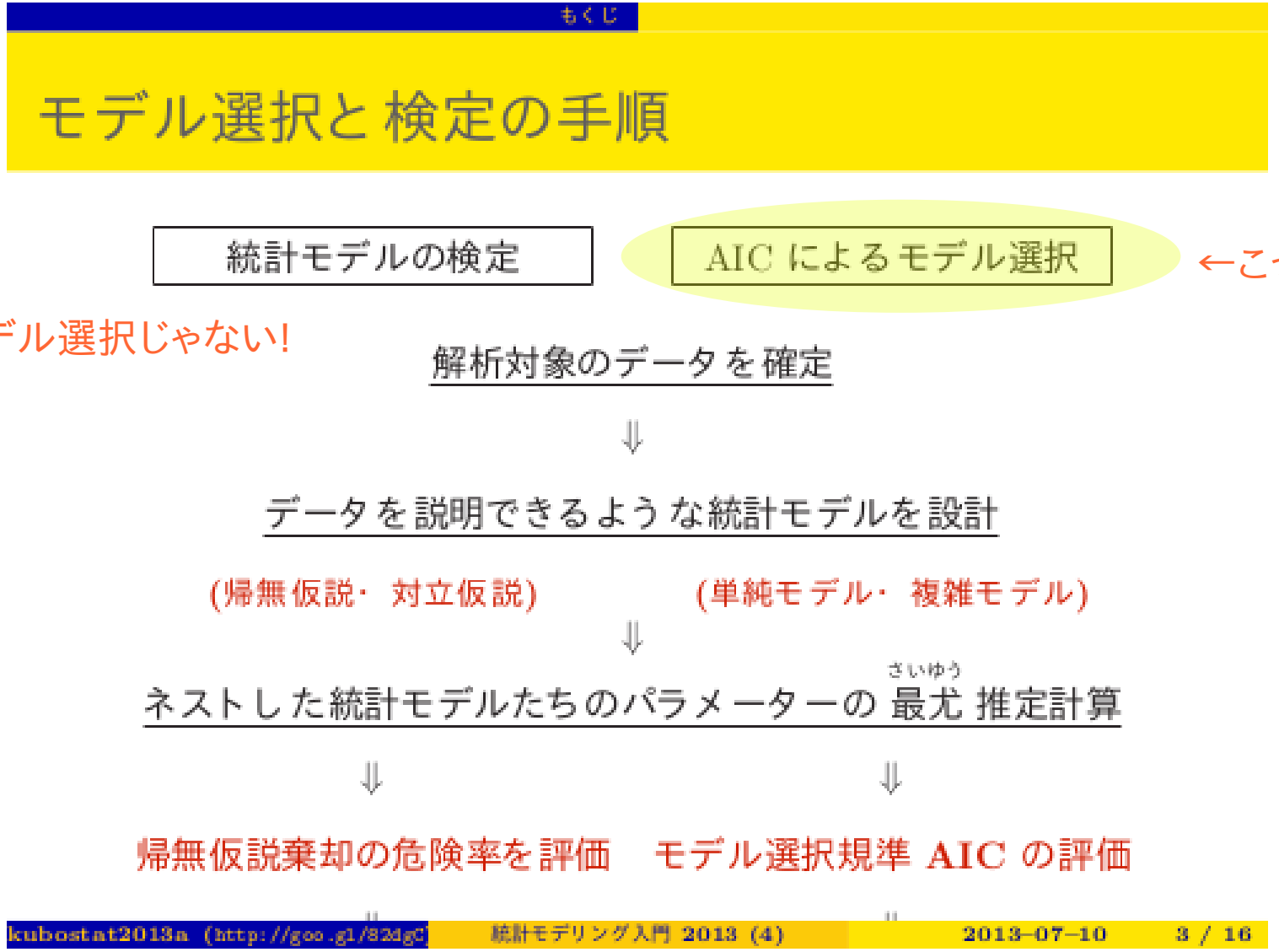
Q. モデル選択とは何か？

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?

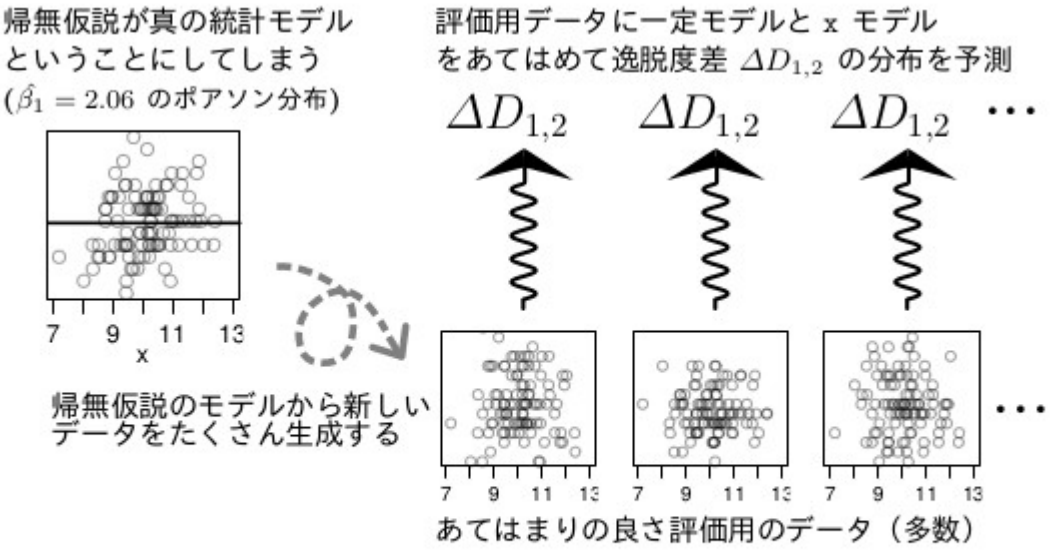


図 6　尤度比検定に必要な $\Delta D_{1,2}$ の分布の生成．まず帰無仮説である一定モデル（$\hat{\beta}_1 = 2.06$, p.6 参照）が真の統計モデルだと仮定し，そこから得られるデータを使って逸脱度差 $\Delta D_{1,2}$ がどのような分布になるかを調べる．

7/17（水）

統計モデリング入門 2013 (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

北大環境科学院の講義 http://goo.gl/82dgC

2013–07–17

ファイル更新時刻: 2013–06–25 16:36

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

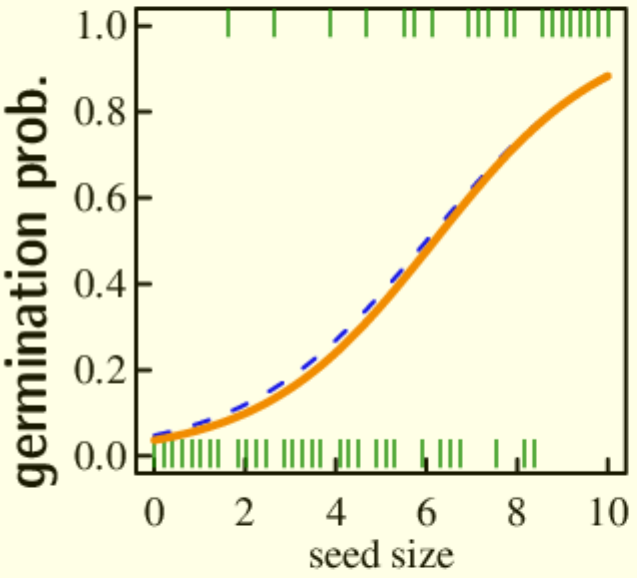
ちなみにこれは x と o/x を比較してるんだから，反比例みたいな偽「負の相関」がでるのはあたりまえ

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

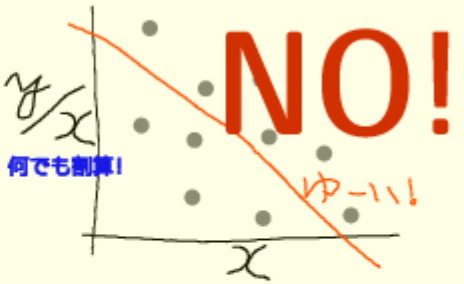
GLM のひとつ，ロジスティック回帰を使おう

データにあわせたより良い統計モデリングを!

おススめ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



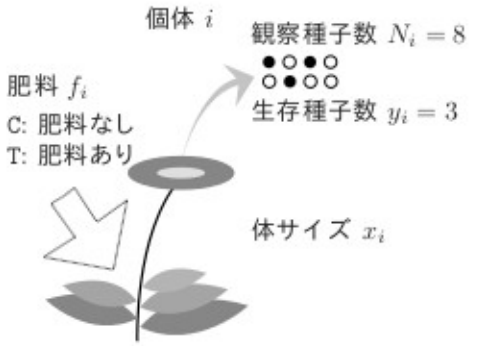
コツ: 不自然にデータをこねくりまわさない
データの性質・構造にあったモデリングを!

2012–11–02 k4 (2012–10–26 17:07 修正版) 43/ 44

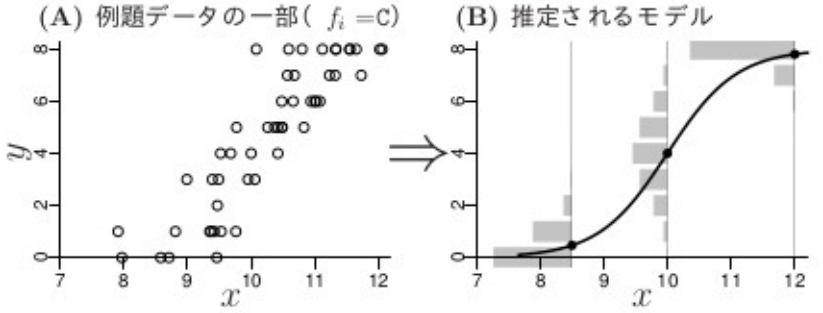
GLM のひとつ，ロジスティック回帰を使おう

またいつもの例題? …… ちょっとちがう

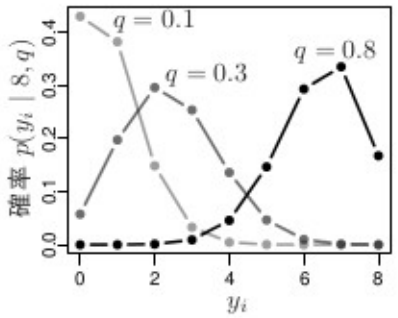
8 個の種子のうち y 個が 発芽可能 だった! …… というデータ



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



7/22（月）

統計モデリング入門 2013 (6)

一般化線形混合モデル

久保拓弥 kubo@ees.hokudai.ac.jp

北大環境科学院の講義 http://goo.gl/82dgC

2013–07–22

ファイル更新時刻: 2013–06–27 12:17

GLM ではうまく対処できない問題

この授業であつかう統計モデルたち

線形モデルの発展

推定計算方法

階層ベイズモデル — MCMC

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

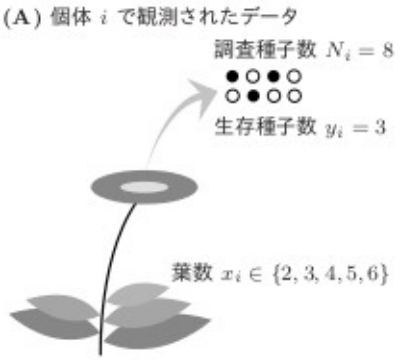
一般化線形モデル — 最小二乗法

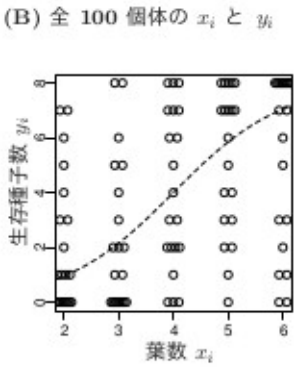
正規分布以外の確率分布をあつかいたい

線形モデル

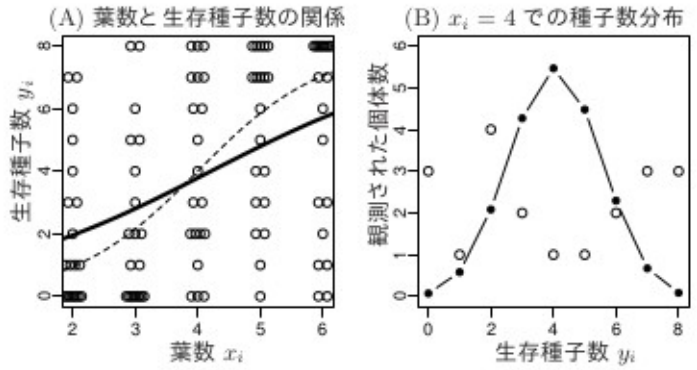
統計モデル勉強のプラン: 線形モデルを発展させる

今日の例題: 種子の生存確率, ただし……





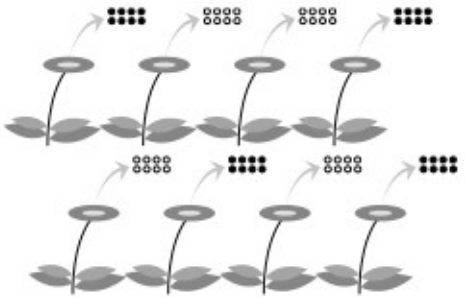
GLM では説明できないばらつき!



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

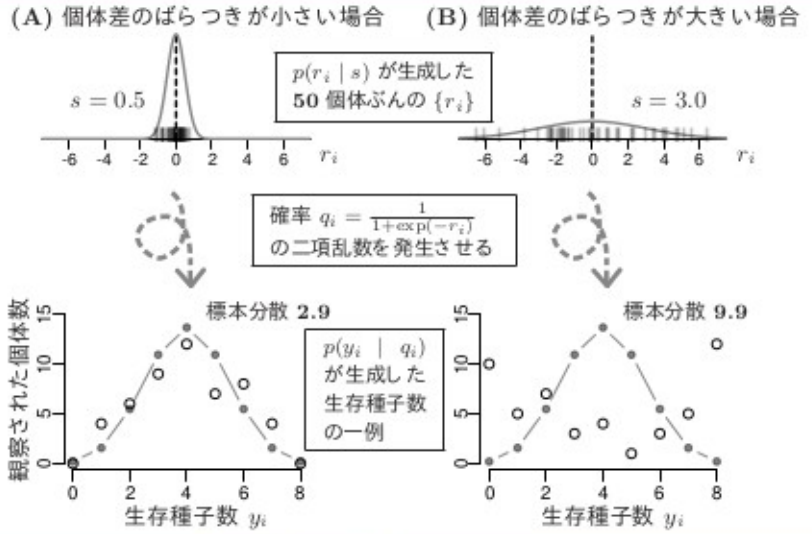
過分散 (overdispersion) とは何か?



kubostat2013a (http://goo.gl/82dgC)　統計モデリング入門 2013 (6)　2013–07–22　9 / 21

個体差のばらつきをあらわす確率分布　平均的な個体や「異端」な個体のばらつき

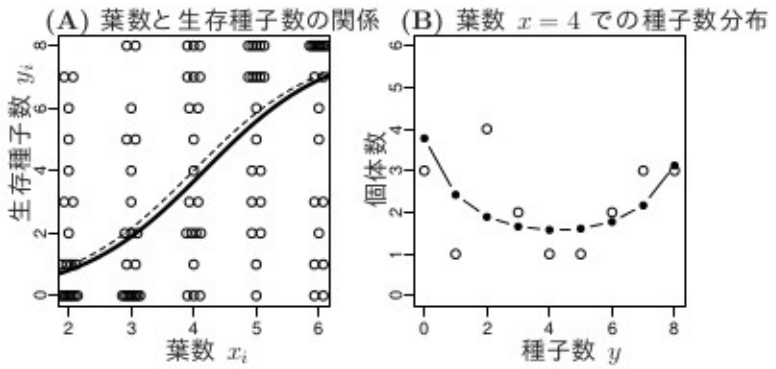
個体差 r_i の分布と過分散の関係



kubostat2013a (http://goo.gl/82dgC)　統計モデリング入門 2013 (6)　2013–07–22　14 / 21

一般化線形混合モデルの最尤推定　「積分する」とは分布を混ぜること

推定された GLMM を使った予測



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

7/24（水）

2013–07–24

全部で 7 回中の 7 回目

統計モデリング入門 2013 (7)

階層ベイズモデル

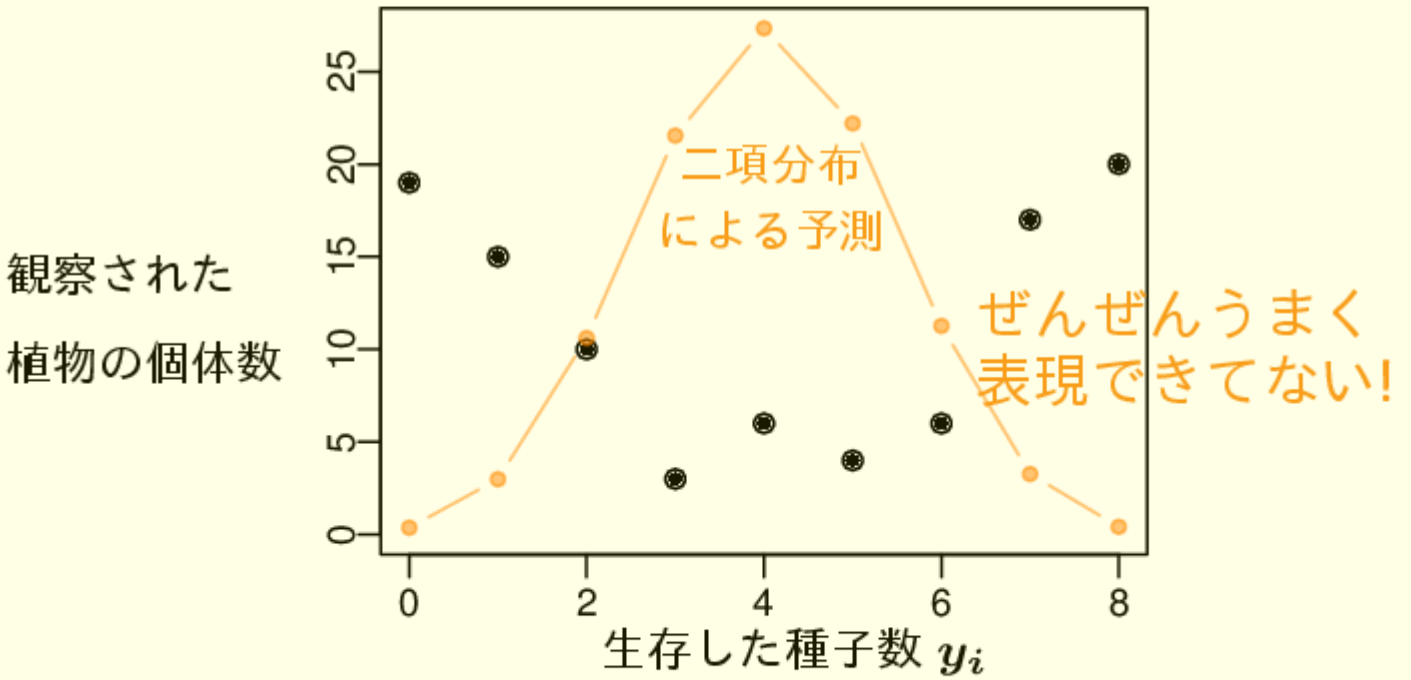
久保拓弥 kubo@ees.hokudai.ac.jp

http://goo.gl/OyB2k

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?!

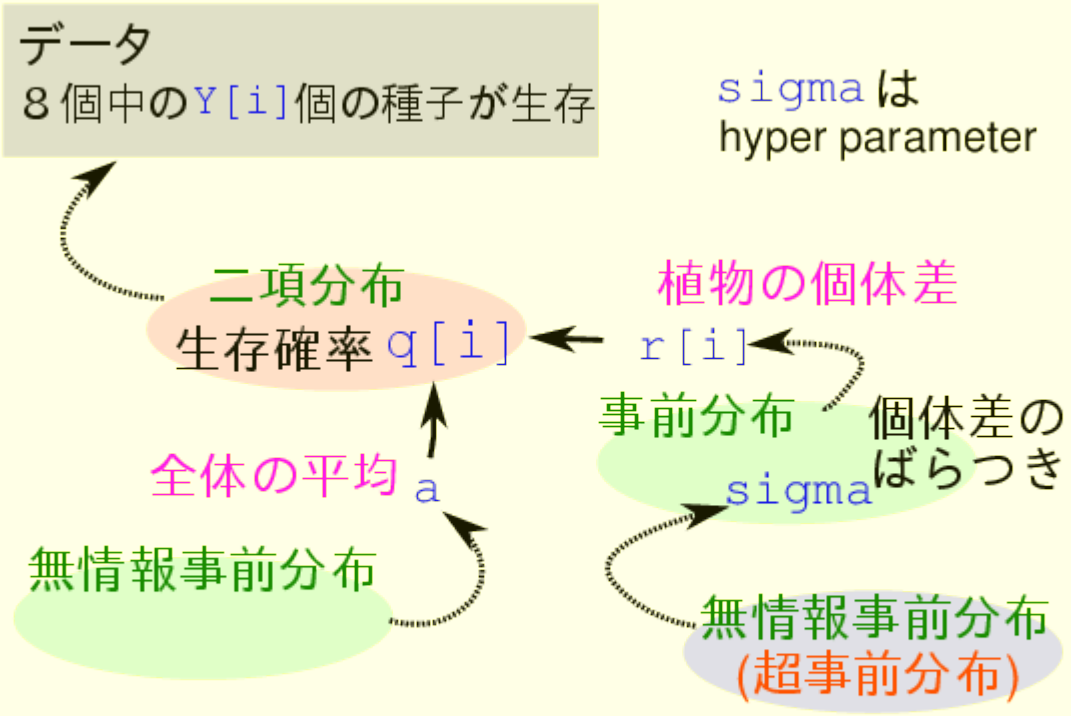
100 個体の植物の合計 800 種子中 403 個の生存が見られたので，平均生存確率は 0.50 と推定されたが……



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

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

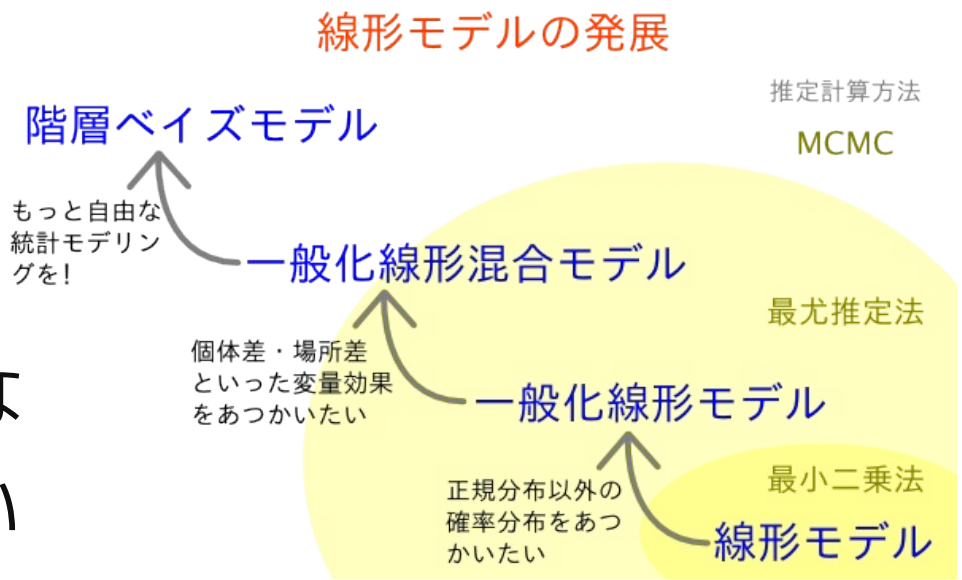
2012–11–02 k5 (2012–10–26 17:26 修正版) 52/ 101

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・種差などめんどう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

今日のハナシはここまで

時間があれば R インストール実演

第 1 回: 7/01 (月) 観測されたパターンを説明する統計モデル

第 2 回: 7/03 (水) 確率分布と最尤推定

第 3 回: 7/08 (月) 一般化線形モデル: ポアソン回帰

第 4 回: 7/10 (水) モデル選択と検定

第 5 回: 7/17 (水) 一般化線形モデル: ロジスティック回帰

第 6 回: 7/22 (月) 一般化線形混合モデル

第 7 回: 7/24 (水) 階層ベイズモデル

次回以降も楽しく勉強をすすめましょう!