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

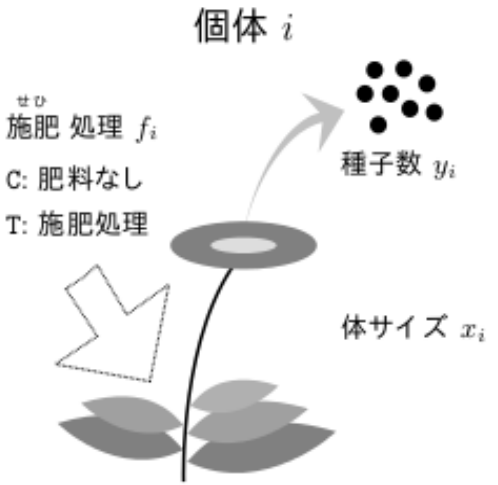


図 3.1　この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

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

「統計モデル」って何?

内容がわからなくてもソフトウェアにまるなげ

“人工知能”?

↑

その実態：機械学習?

↑

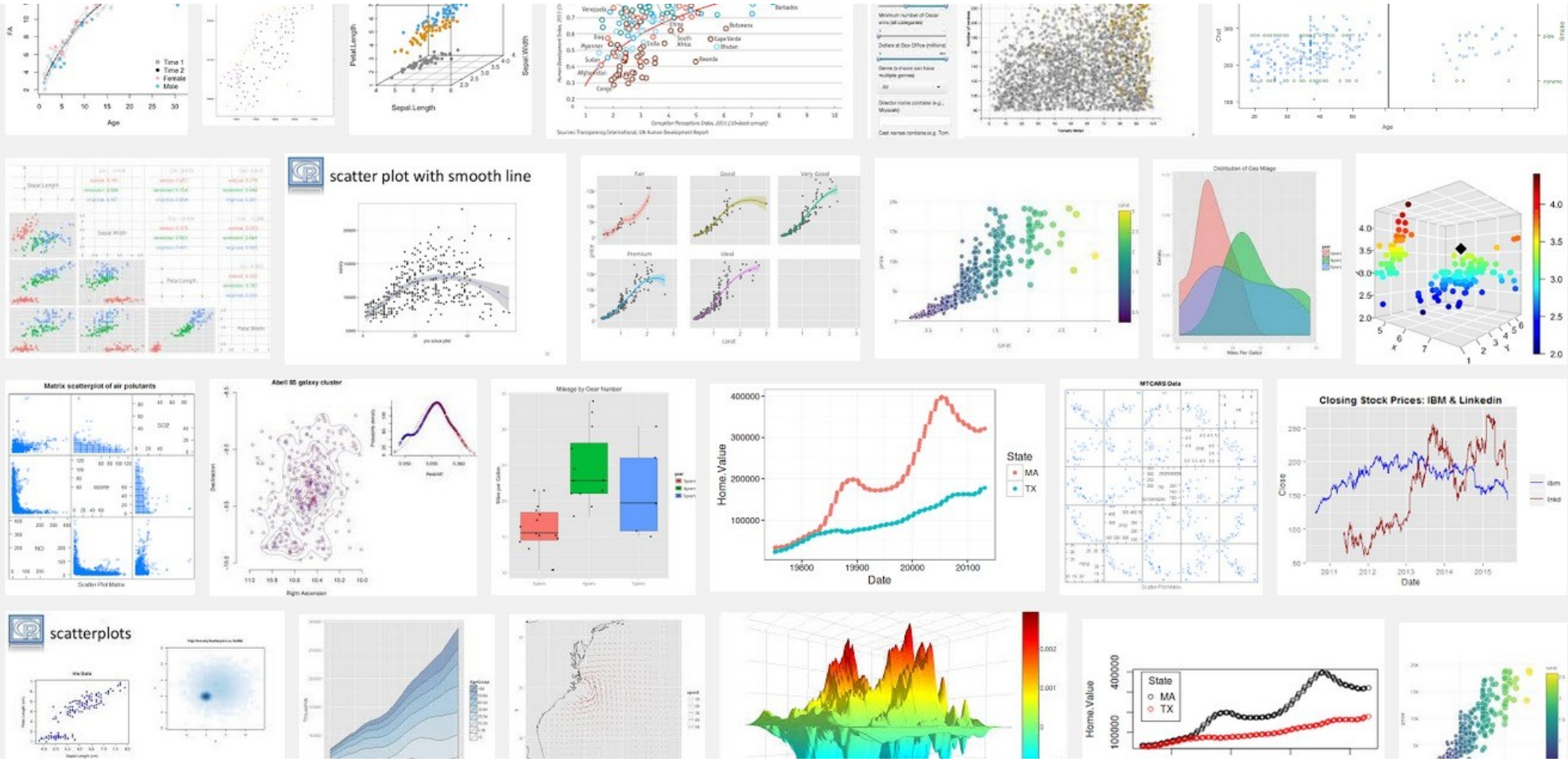
その部品の一部：「統計モデル」

…ぐらいの理解でよいかと…

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

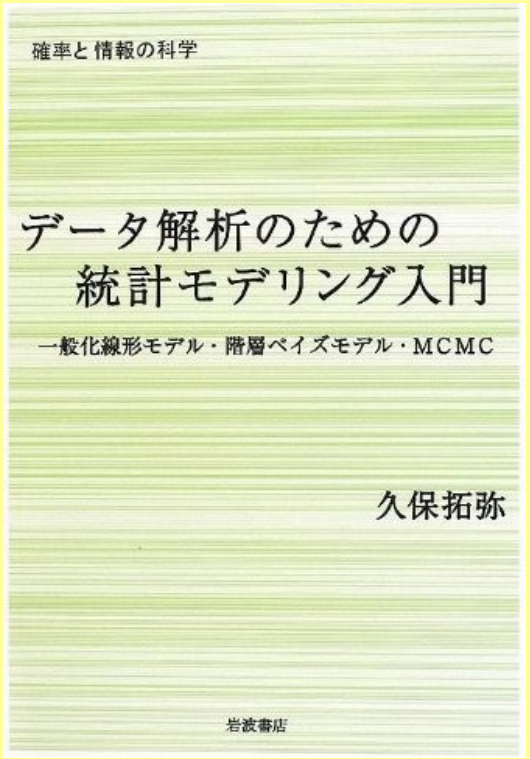
作図のないデータ解析はありえない!

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

教科書とソフトウェア

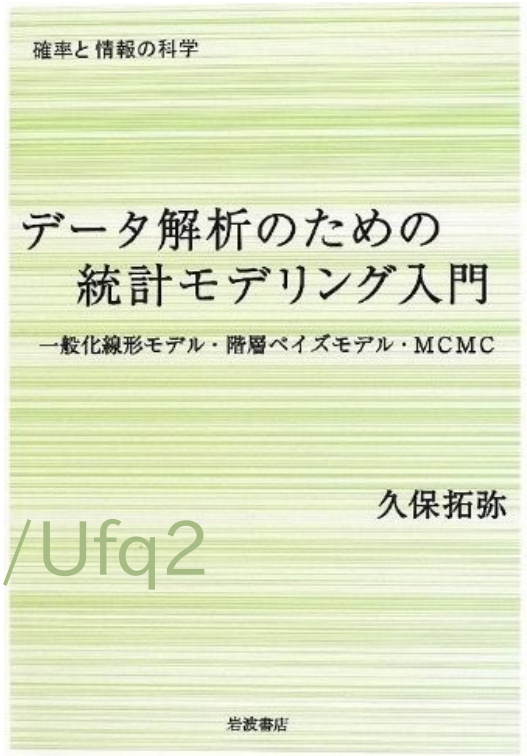

確率と情報の科学
データ解析のための
統計モデリング入門
一般化線形モデル・階層ベイズモデル・MCMC
久保拓弥
岩波書店


R

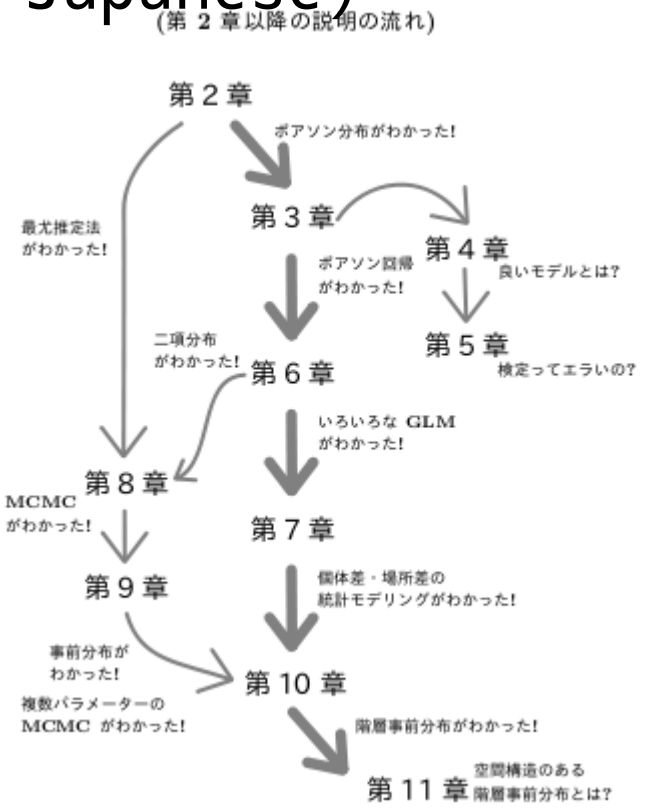
この授業は「統計モデリング入門」にそった内容を説明します

my text book (in Japanese)

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

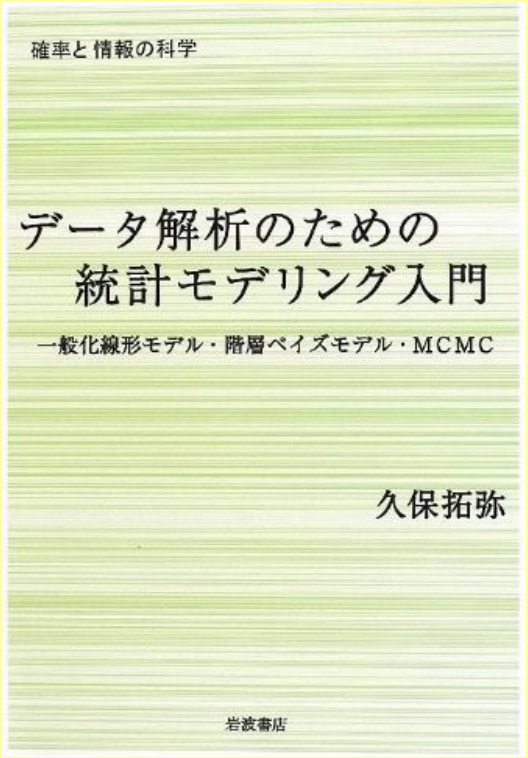
統計ソフトウェア R

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

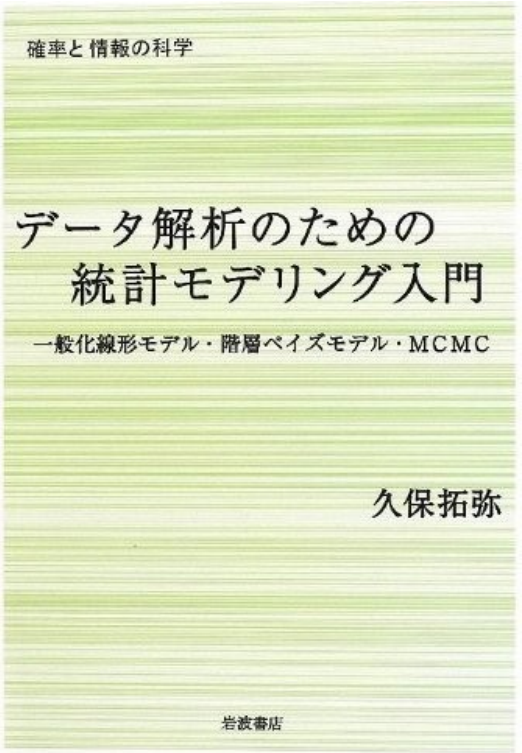
この教科書でも R を
使って問題を解決する
方法を説明しています

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！

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

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

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)

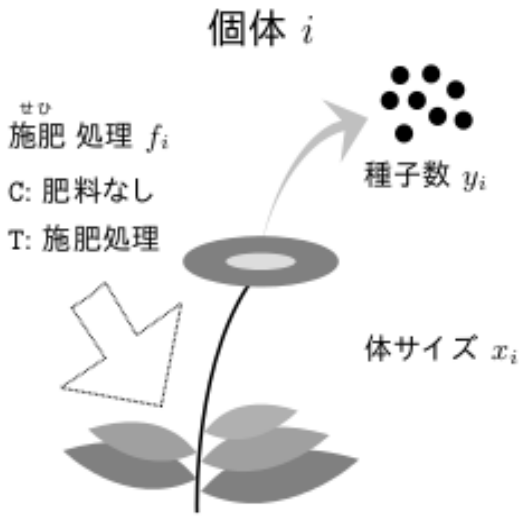
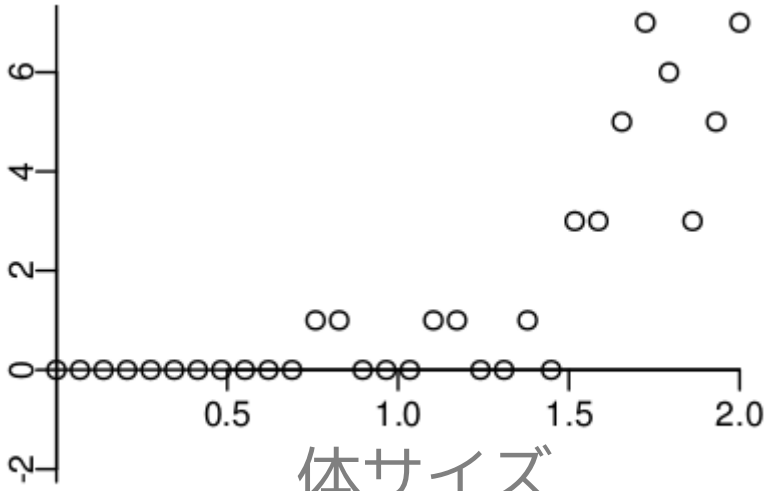


図 3.1　この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（ 個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

number of seeds

種子数



体サイズ

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!

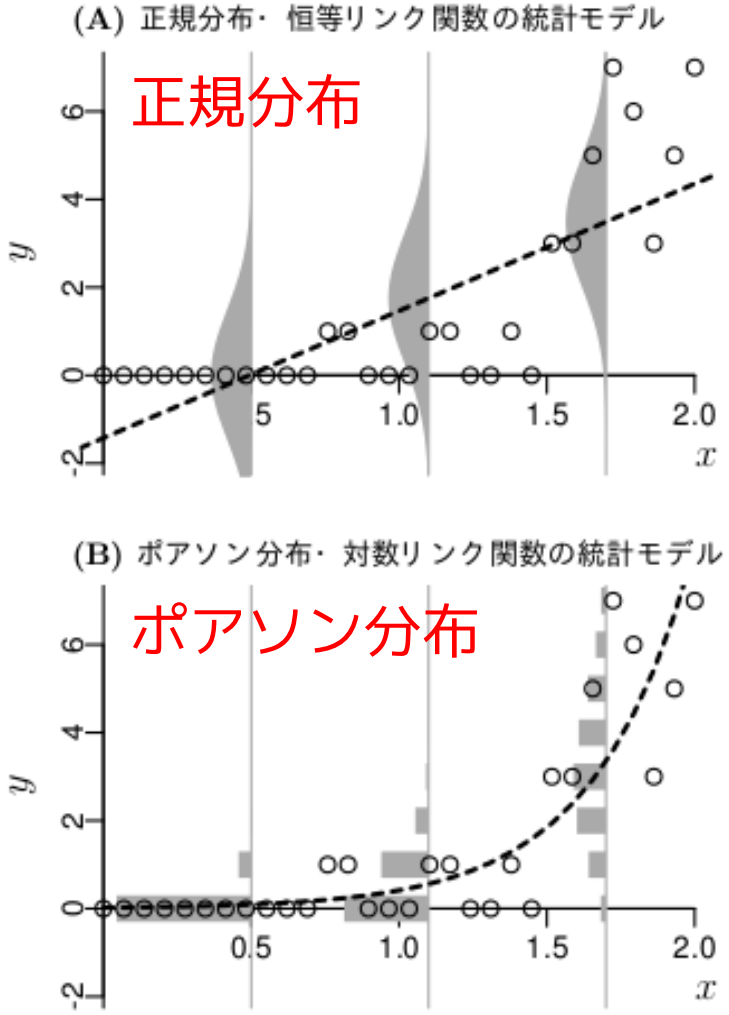


図 3.9 回帰モデルと確率分布の関係．また別の架空データに対して GLM をあてはめた例．破線は x とともに変化する平均値．グレイで

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

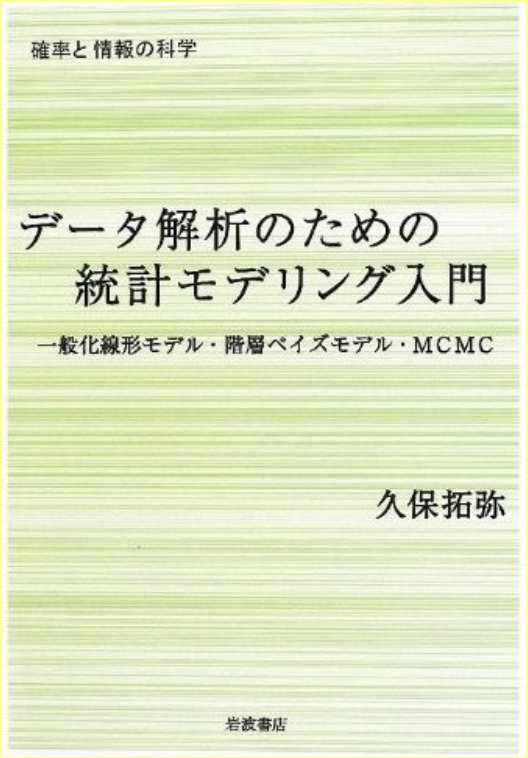
最小二乗法

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!!

もともと誰のための教科書・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
  - No “Blackbox” statistics!
- 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

9/18 の概要

(a) 09:30-10:45 統計モデリング講義の概要

(b) 10:55-12:10 確率分布と最尤推定

(c) 13:20-14:35 R 実習: data.frame 操作と作図

(d) 14:45-16:00 ポアソン回帰の GLM

(e) 16:10-17:25 モデル選択と検定

筑波大 (大塚) 集中講義 2016 (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

筑波大の講義 http://goo.gl/HvRhXn

2016–09–18

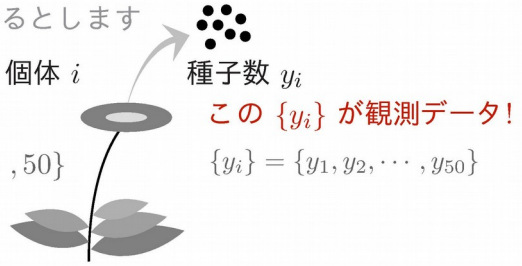
ファイル更新時刻: 2016–09–08 17:19

単純化した例題

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



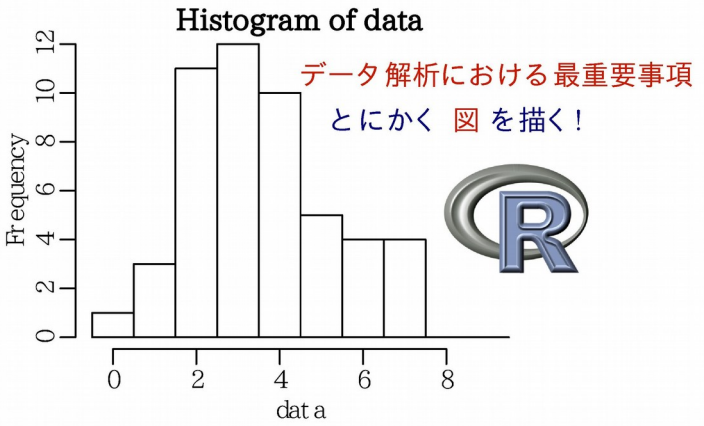
このデータ $\{y_i\}$ がすでに R という統計ソフトウェアに
格納されていた，としましょう

```
> 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
```


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))
```



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

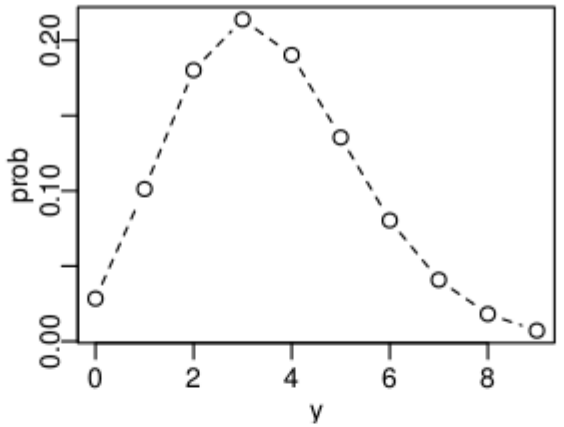


図 4　平均 $\lambda = 3.56$ のポアソン分布．種子数 y とその確率 prob の関係が示されている．図 3 の表を図にしたもの．R の plot() 関数の引数，type = "b" によって「丸と折れ線による図示」，lty = 2 によって「折れ線は破線で」と指示している．

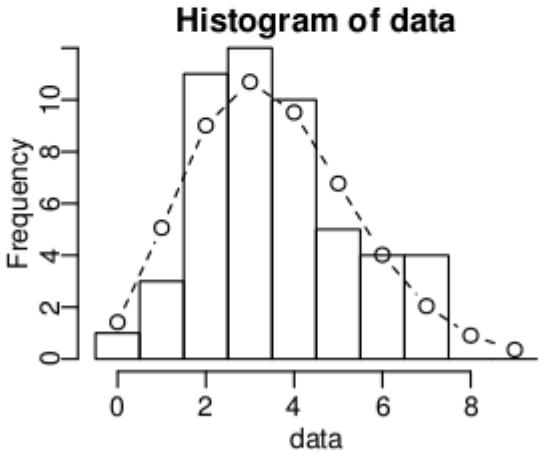


図 5　観測データと確率分布の対応をながめる．ヒストグラムは図 2 と同じ．それに重ねられている丸と破線は y 個の種子をもつ個体数の予測．平均 3.56 の図 4 のポアソン分布の確率分布に全個体数 50 をかけて得られる．

さいゆう最尤推定という考えかたを説明します

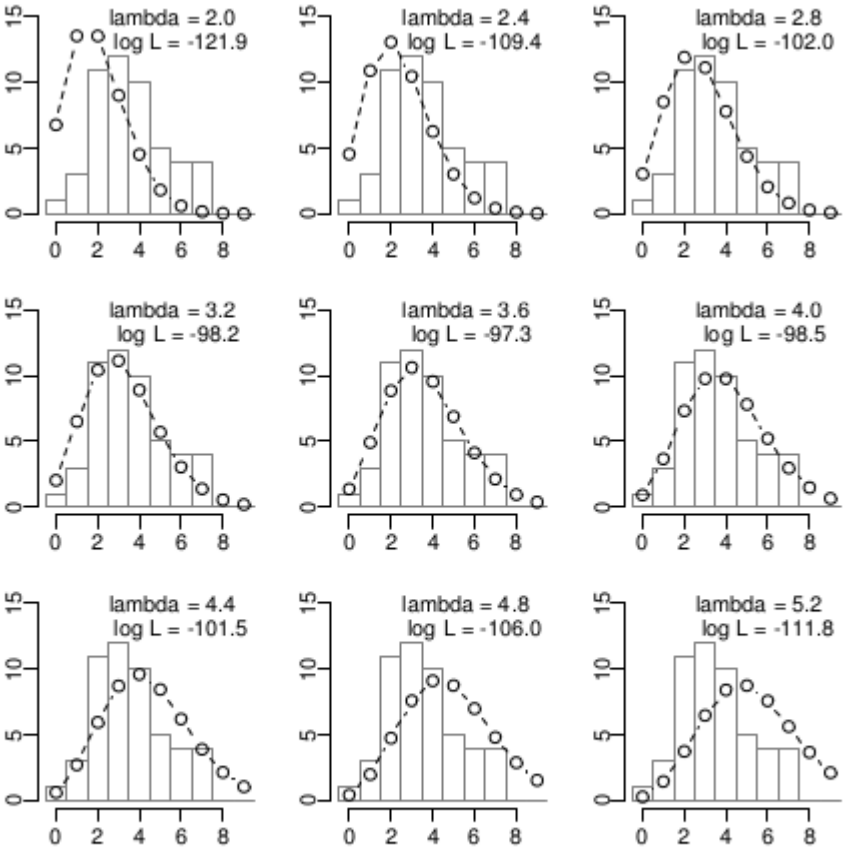
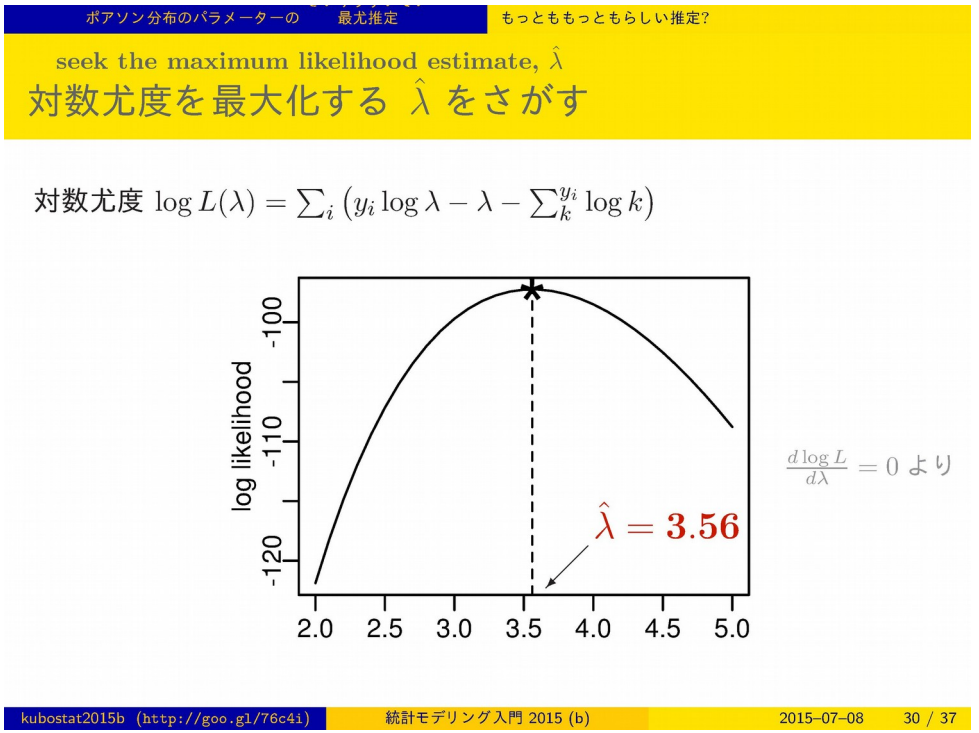


図 7　平均 λ（lambda）を変化させていったポアソン分布と，観測データへのあてはまりの良さ（対数尤度 logL）．すべてのヒストグラムは図 2 と同じ．



筑波大 (大塚) 集中講義 2016 (c)

R の練習: 次の時間の例題データ

久保拓弥 kubo@ees.hokudai.ac.jp

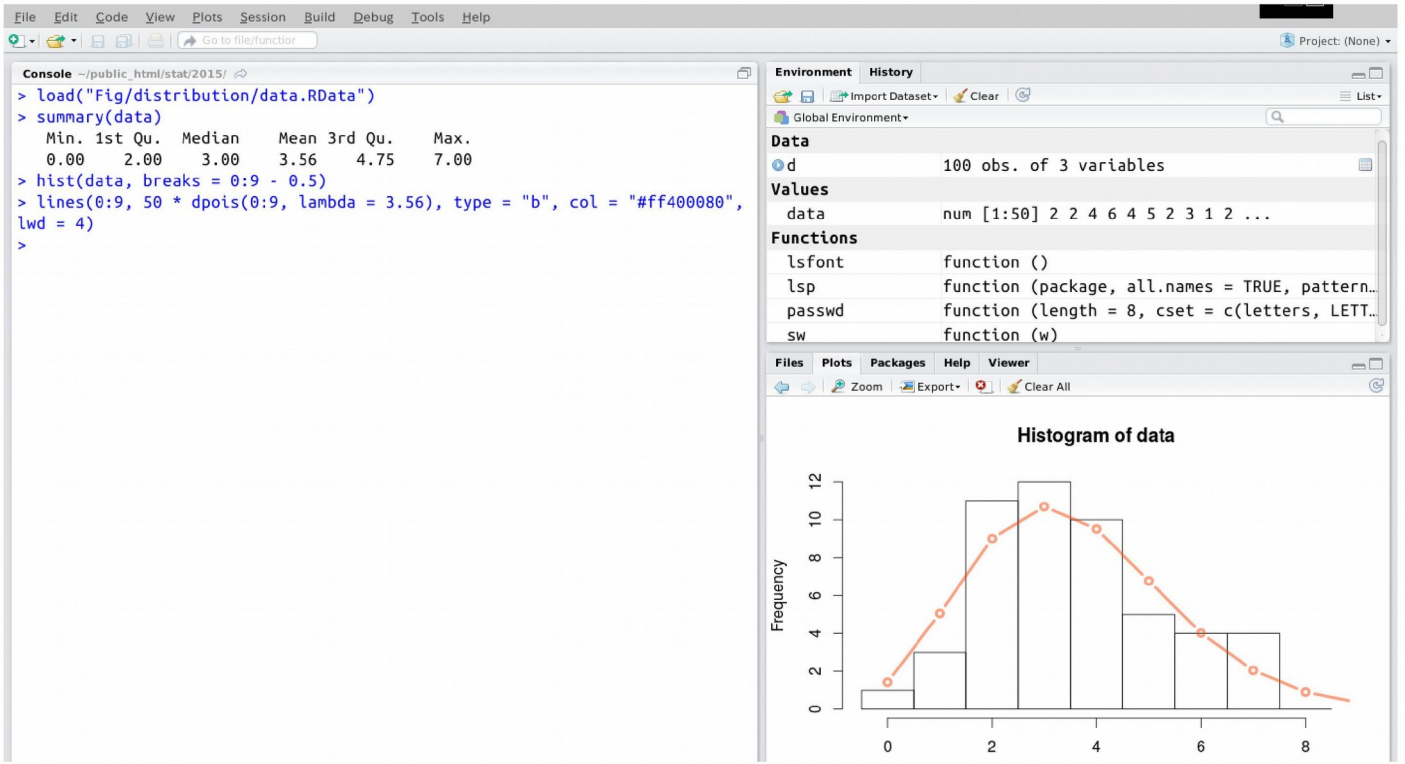
筑波大の講義 http://goo.gl/HvRhXn

2016–09–18

ファイル更新時刻: 2016–09–09 17:27

統計ソフトウェア R の練習

RStudio 使ってみますかね?



筑波大 (大塚) 集中講義 2016 (d)

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

筑波大の講義 http://goo.gl/HvRhXn

2016–09–18

ファイル更新時刻: 2016–09–08 17:40

ここで登場する ---
「何でも正規分布」ではダメ！という発想

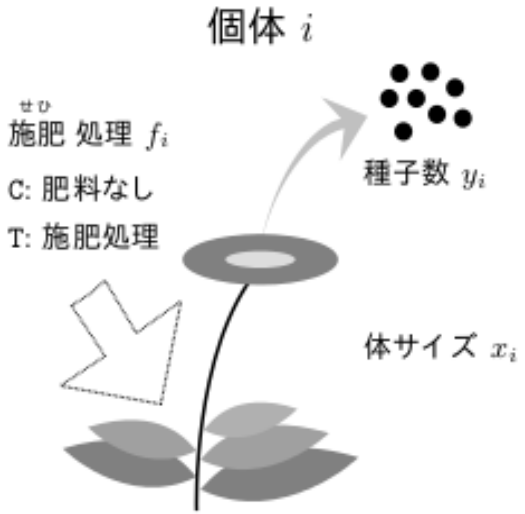


図 3.1　この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

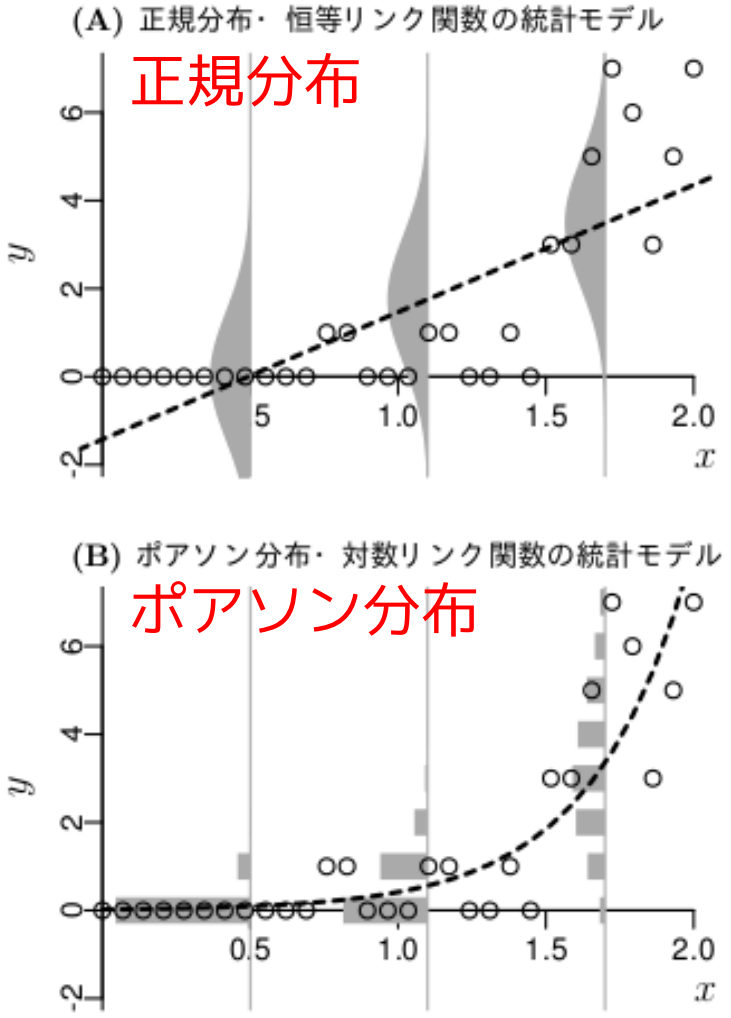


図 3.9　回帰モデルと確率分布の関係．また別の架空データに対して GLM をあてはめた例．破線は x とともに変化する平均値．グレイで

Free の統計ソフトウェア

R で統計モデリング

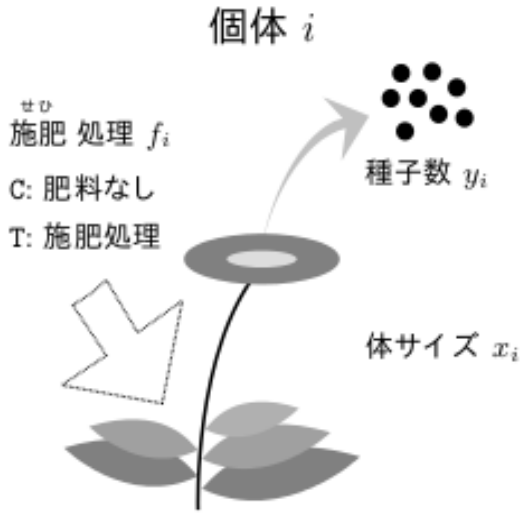
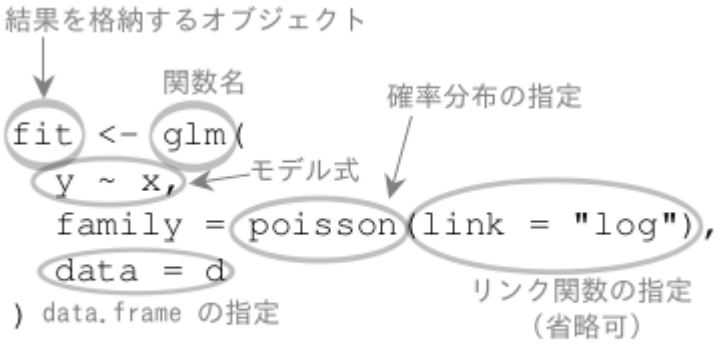


図 3.1　この例題に登場する架空植物の第 i 番目の個体
体サイズ（ 個体の大きさ）x_i と肥料をやる施肥処理
にどう影響しているのかを知りたい．



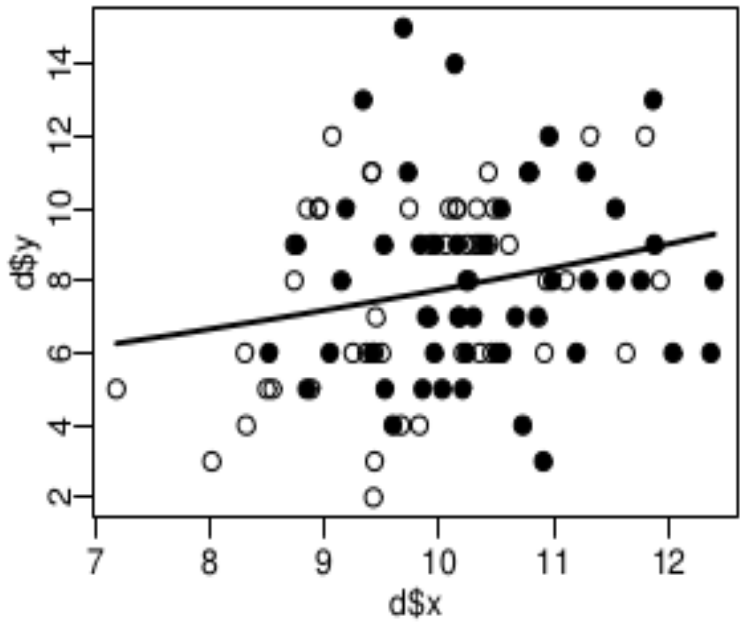


図 17　平均種子数 λ の予測．図 12 に λ の予測値（実線）を上がきしたもの．

筑波大 (大塚) 集中講義 2016 (e)

モデル選択と検定

久保拓弥 kubo@ees.hokudai.ac.jp

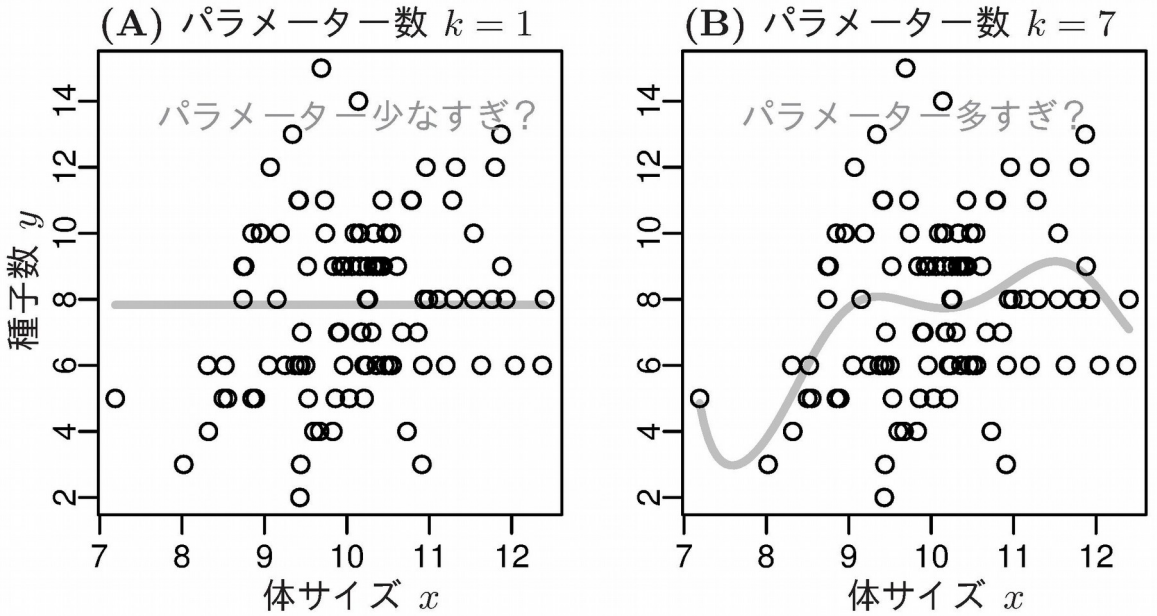
筑波大の講義 http://goo.gl/HvRhXn

2016–09–18

ファイル更新時刻: 2016–09–08 17:48

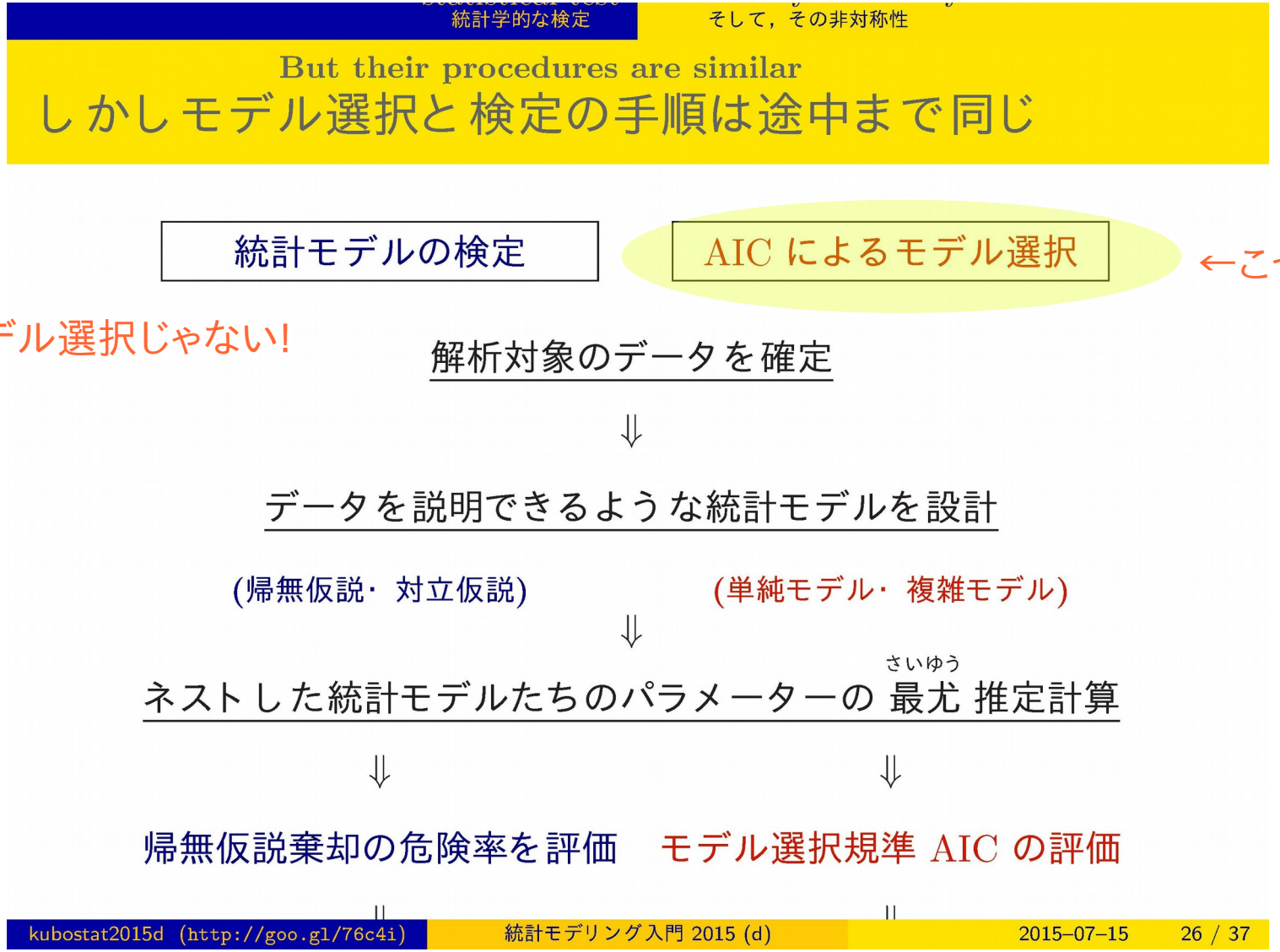
Q. モデル選択とは何か？

パラメーター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?

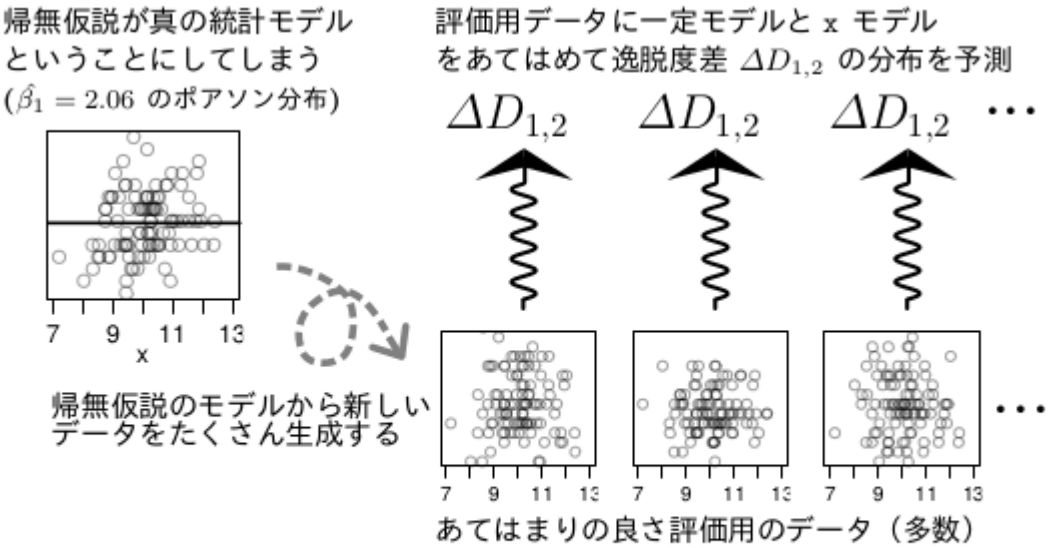


図 6　尤度比検定に必要な $\Delta D_{1,2}$ の分布の生成．まず帰無仮説である一定モデル（$\hat{\beta}_1 = 2.06$, p.6 参照）が真の統計モデルだと仮定し，そこから得られるデータを使って逸脱度差 $\Delta D_{1,2}$ がどのような分布になるかを調べる．

9/19 の概要

(f) 09:30-10:45 ロジスティック回帰の GLM

(g) 10:55-12:10 マルコフ連鎖モンテカルロ法

(h) 13:20-14:35 階層ベイズモデル

(i) 14:45-16:00 階層ベイズモデルの応用:
時系列データの状態空間モデル 1

(j) 16:10-17:25 階層ベイズモデルの応用:
時系列データの状態空間モデル 2

筑波大 (大塚) 集中講義 2016 (f)

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

筑波大の講義 http://goo.gl/HvRhXn

2016–09–19

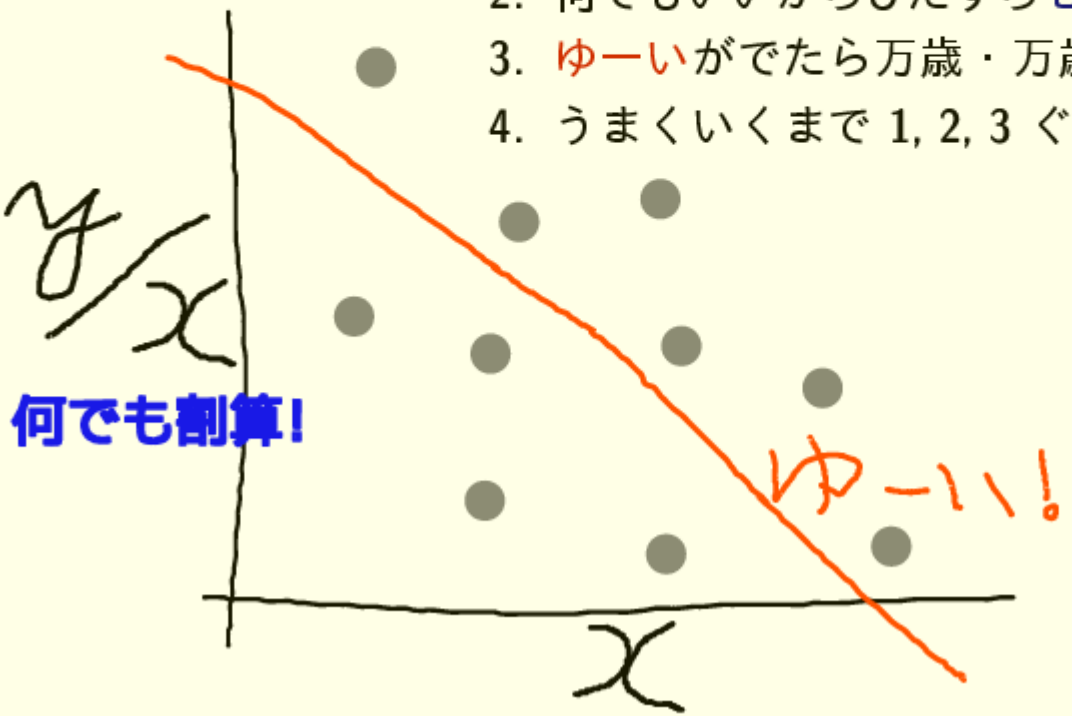
ファイル更新時刻:　2016–09–08　17:58

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



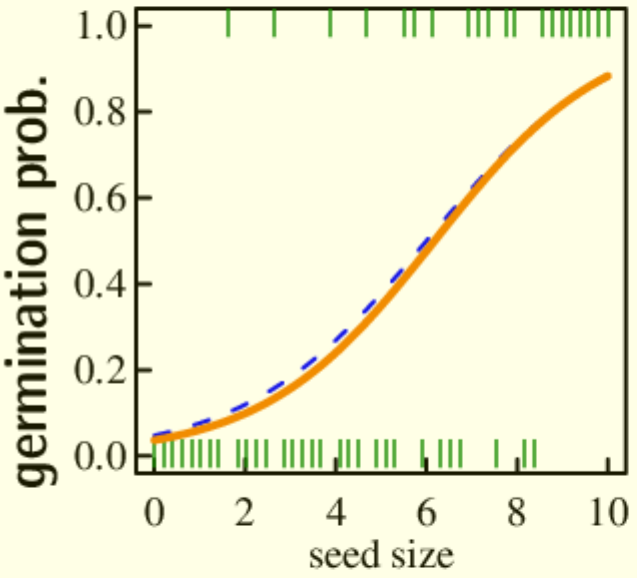
ちなみにこれは x と o/x を比較してるんだから，反比例みたいな偽「負の相関」がでるのはあたりまえ

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

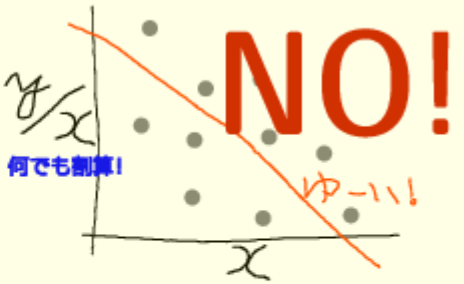
GLM のひとつ，ロジスティック回帰を使おう

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



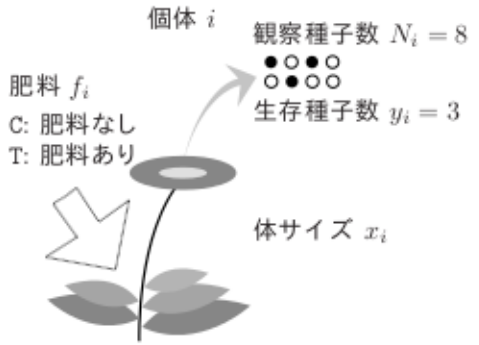
コツ: 不自然にデータをこねくりまわさない
データの性質・構造にあったモデリングを!

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

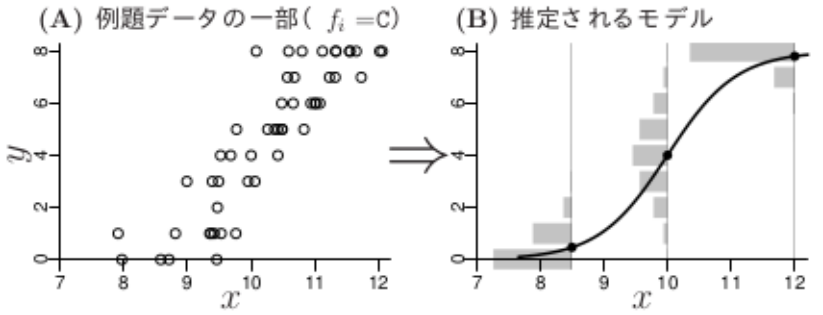
GLM のひとつ，ロジスティック回帰を使おう

またいつもの例題? …… ちょっとちがう

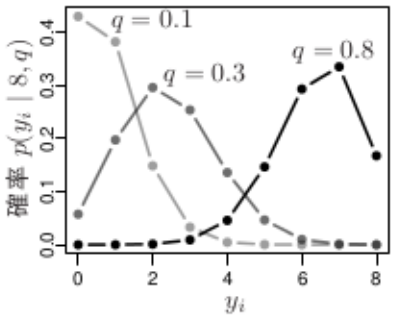
8 個の種子のうち y 個が 発芽可能 だった! …… というデータ



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



筑波大 (大塚) 集中講義 2016 (g)

階層ベイズモデル Hierarchical Bayesian Model

久保拓弥 kubo@ees.hokudai.ac.jp

筑波大の講義 http://goo.gl/HvRhXn

2016–09–19

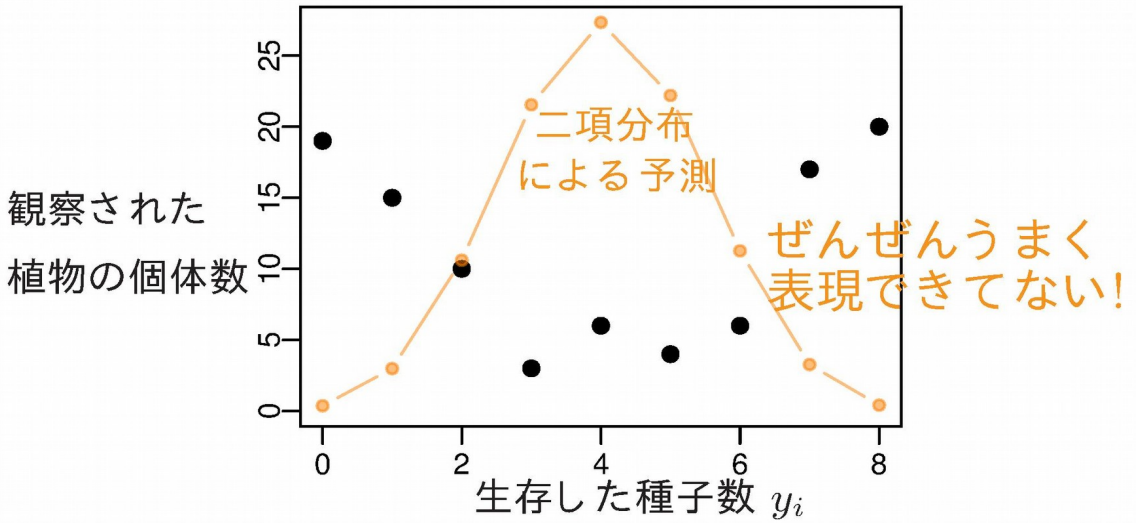
ファイル更新時刻: 2016–09–08 18:02

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?

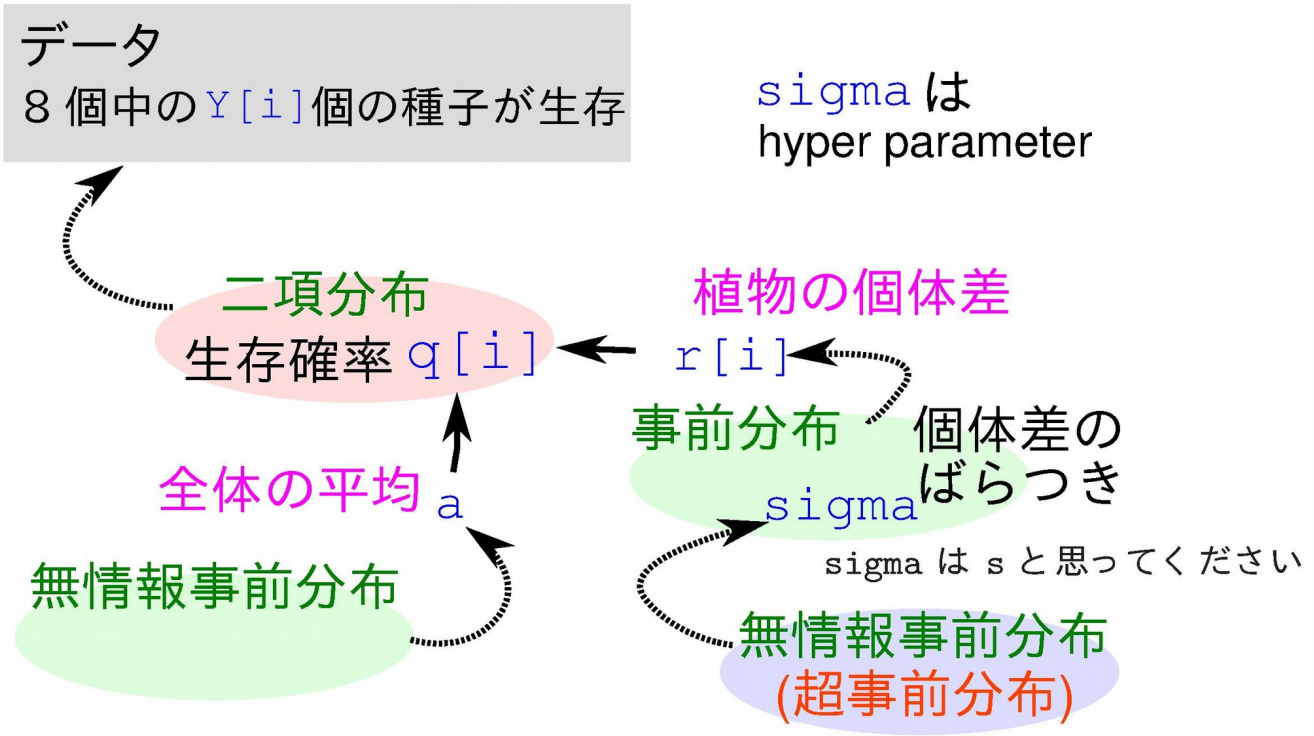
(「統計モデリング入門」第 10 章の最初の例題)

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

なぜ「階層」ベイズモデルと呼ばれるのか?

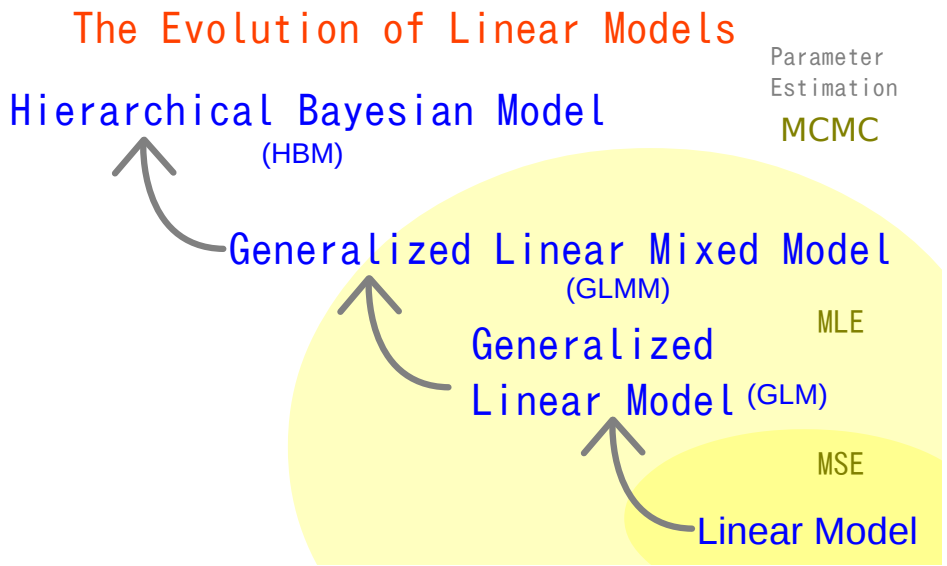
超事前分布 → 事前分布という階層があるから



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うな「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

9/19 (h)

筑波大 (大塚) 集中講義 2016 (h)

階層ベイズモデルと時間変化モデル

久保拓弥 kubo@ees.hokudai.ac.jp

筑波大の講義 http://goo.gl/HvRhXn

2016–09–19

ファイル更新時刻: 2016–09–09 13:29

短い時系列データ

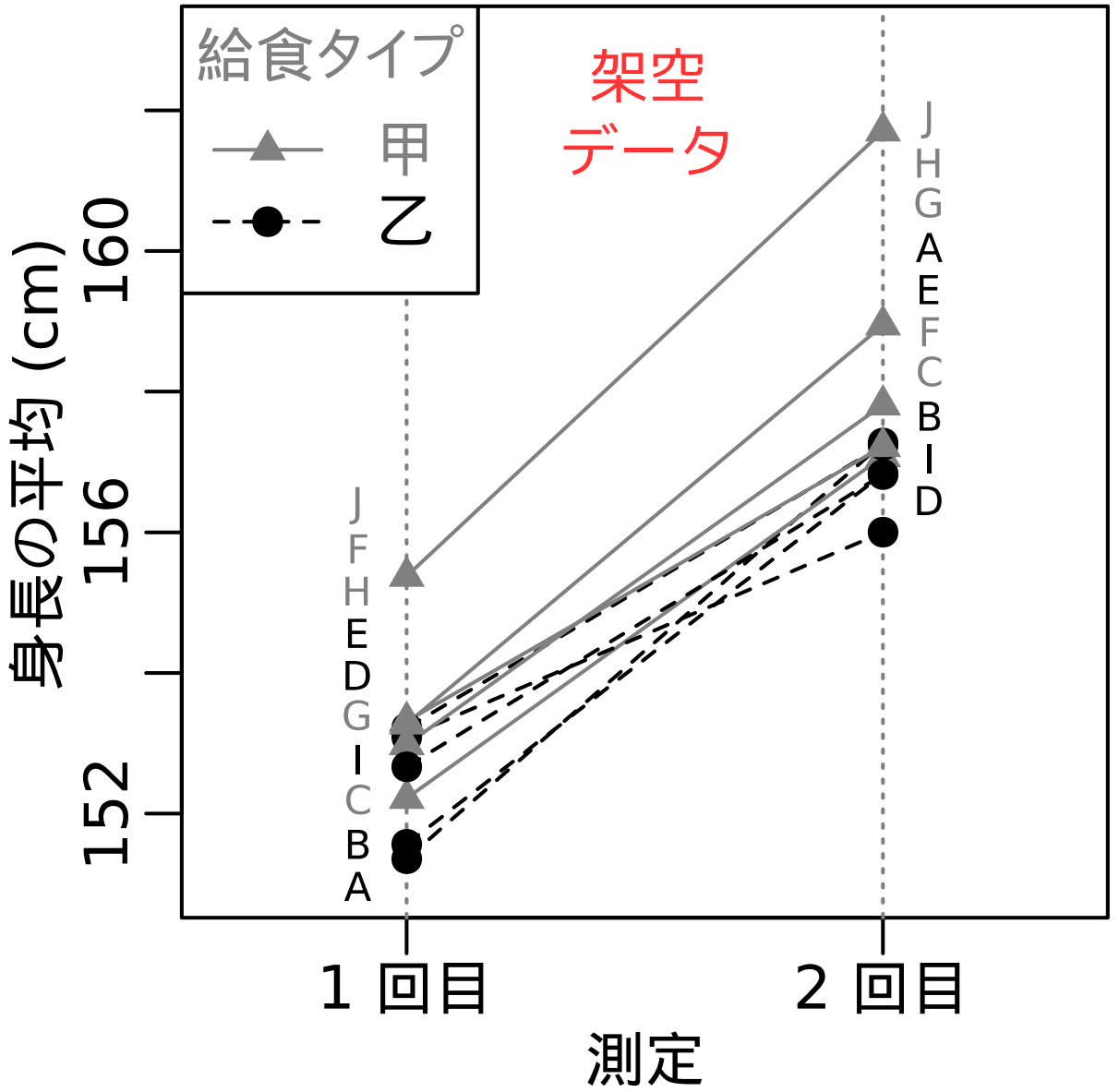
時系列の長短に関係なく

「対応のある」データ点か

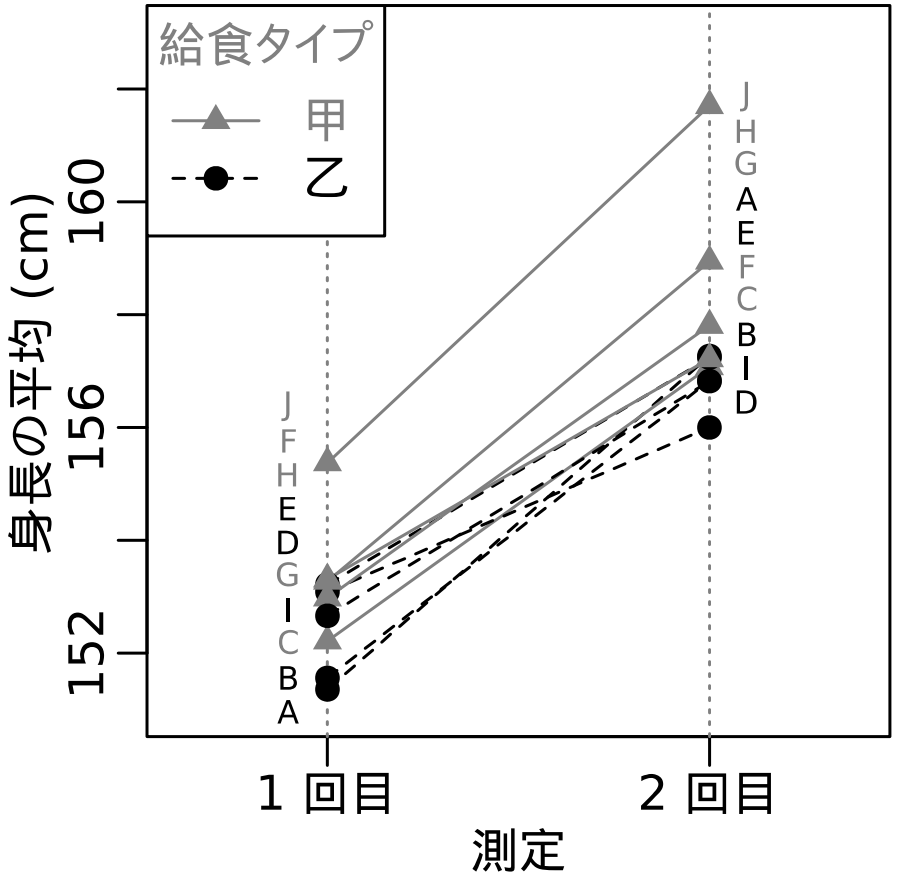
どうかが本質的な問題

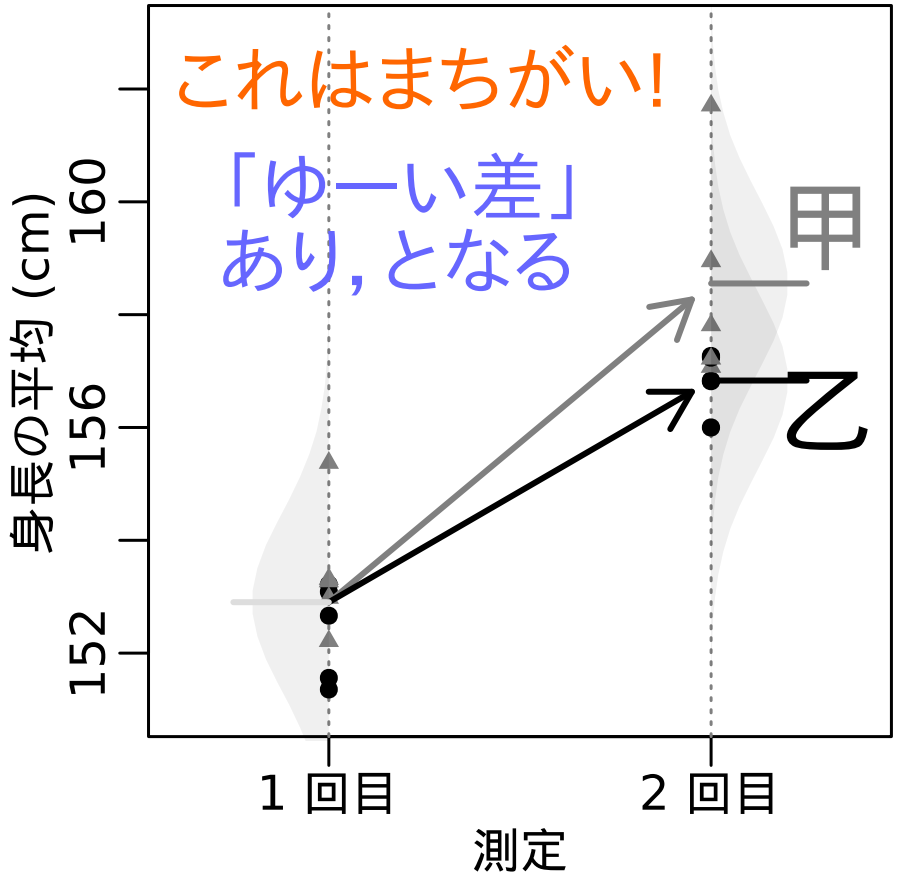
再測定もまた時系列データ

岩波データサイエンス vol.1



対応（paired）を考えてない GLM あてはめ

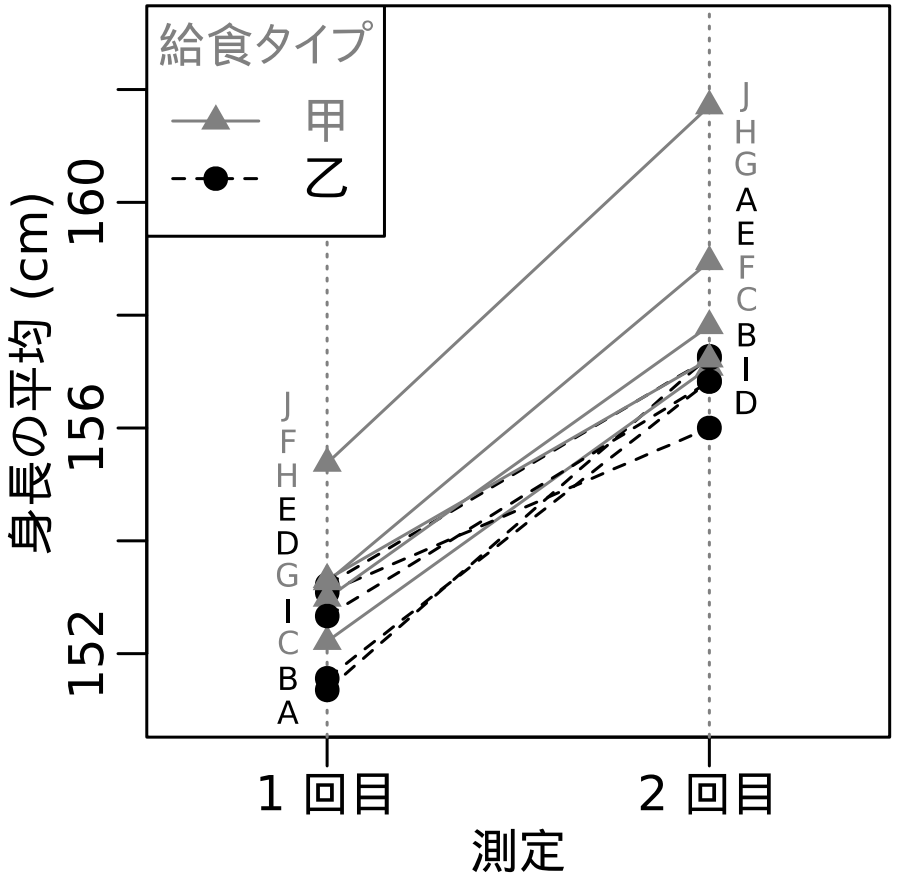


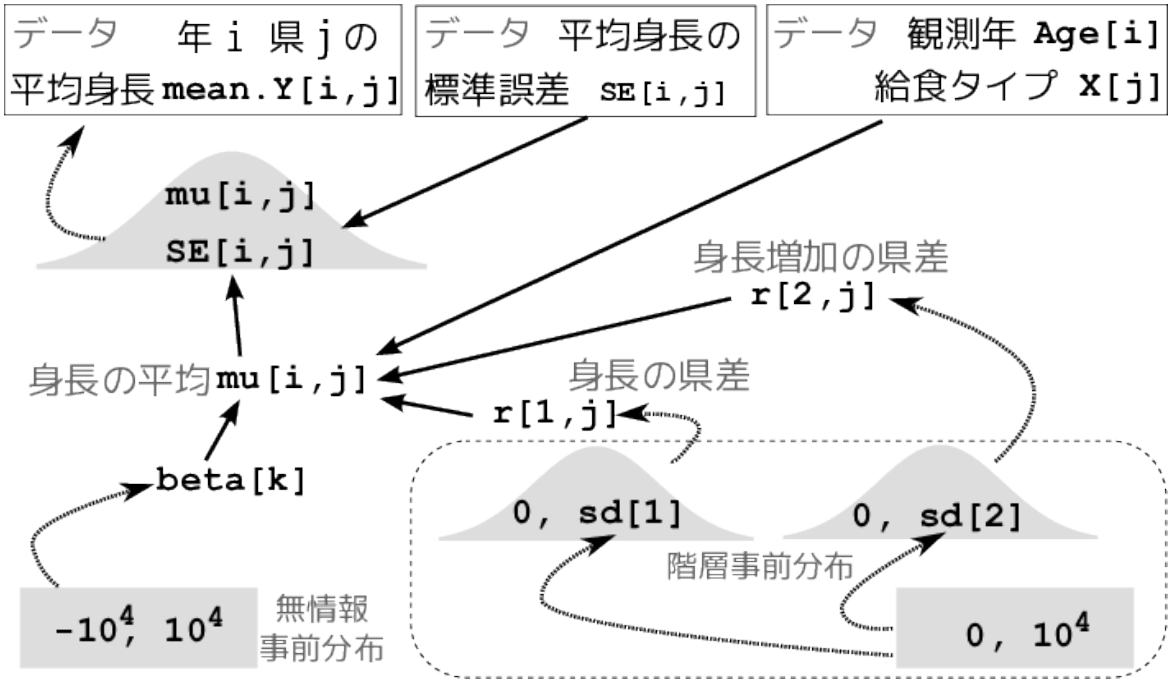


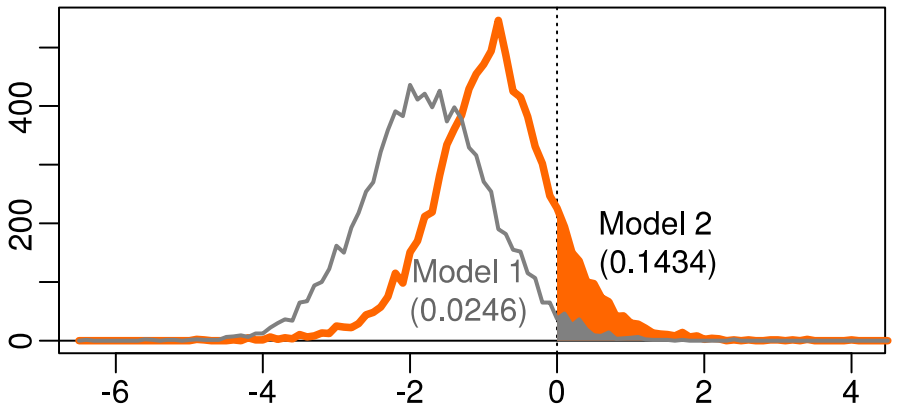
glm(身長 ～ (測定2回目) + (測定2回目):(処理の効果))

同じ対象を二回測定していることを考慮してない

対応（paired）を考慮し，さらに県の差もあるモデル







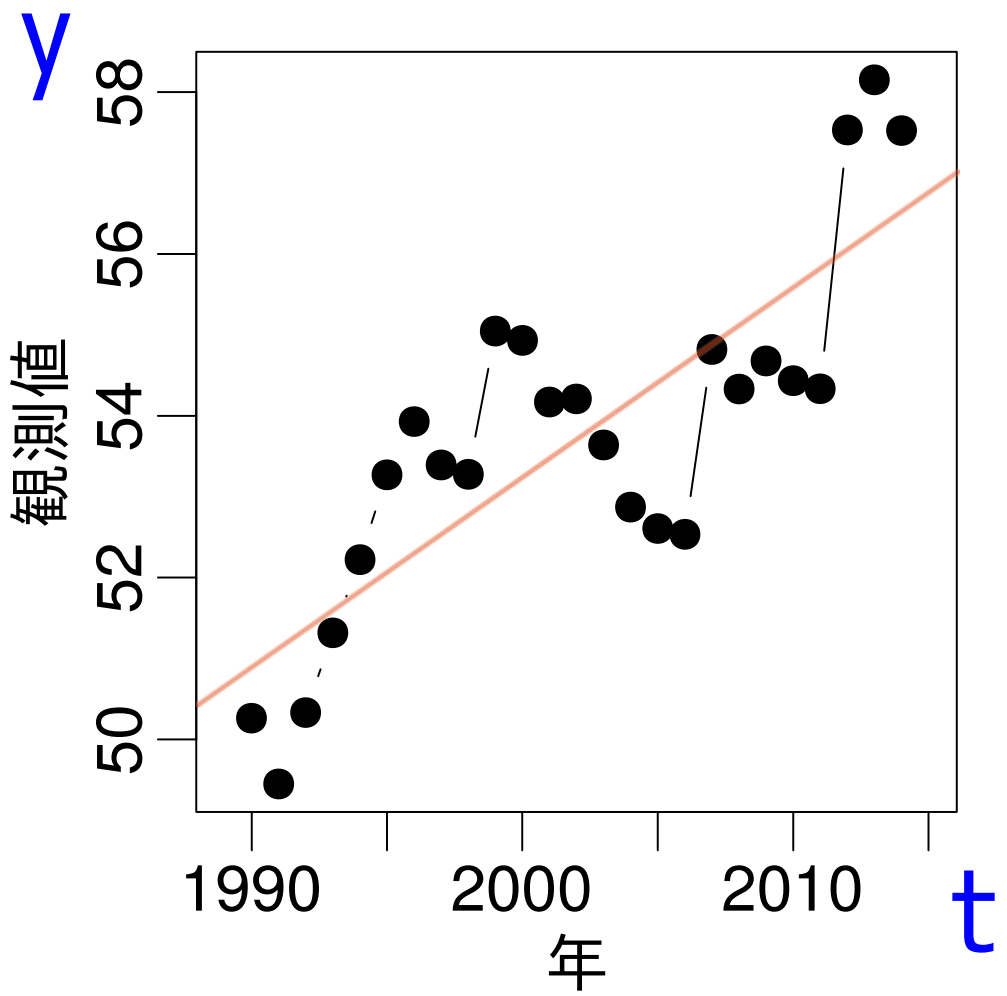
給食効果なし

長い時系列データ

データ上で「時間相関」が見える

「時間相関」のモデリングが必要

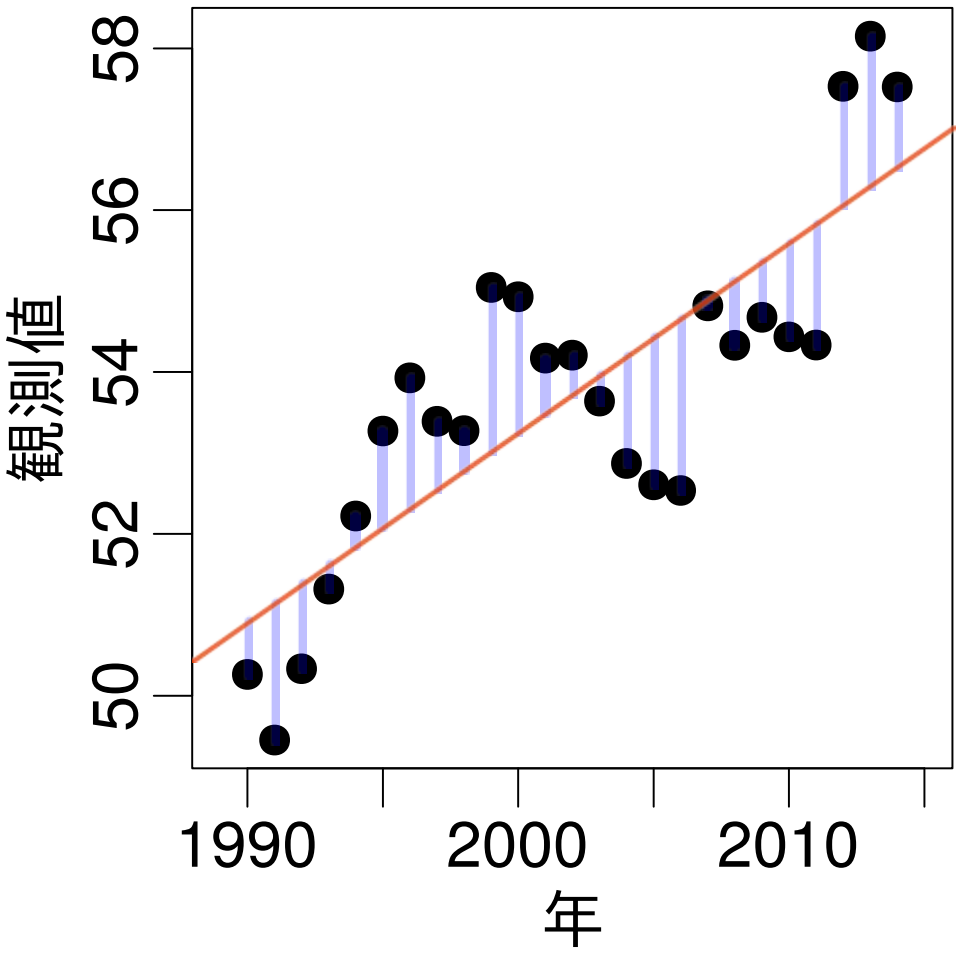
時間相関のある時系列データに…



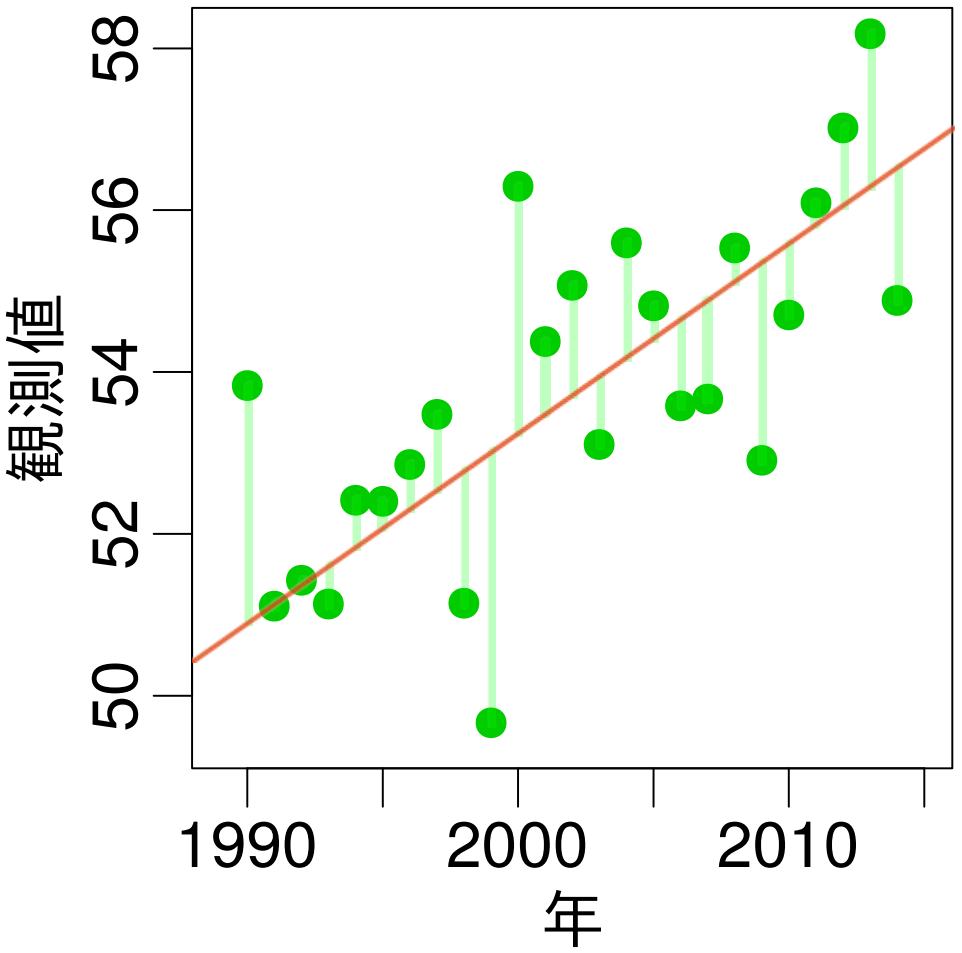
glm(y ~ t)

…と，モデルを
あてはめてみた

時系列の「ずれ」



GLM のずれ

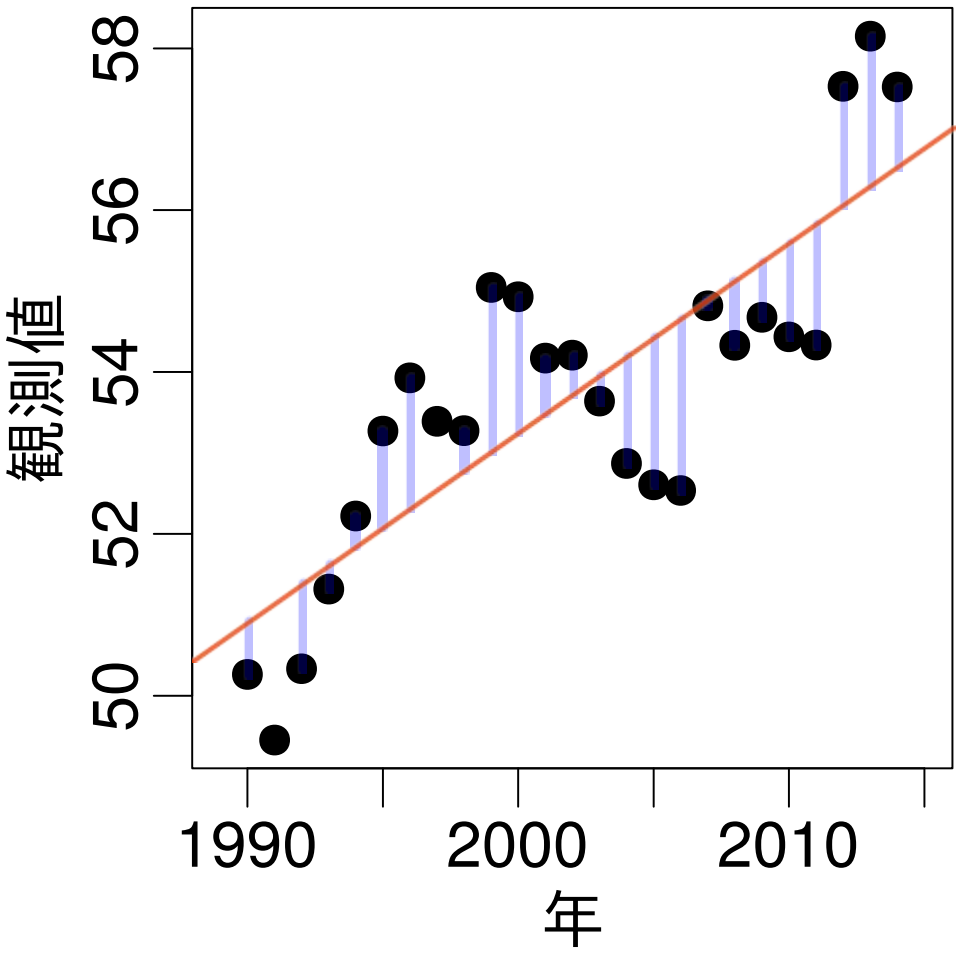


直線からのずれがちがう!

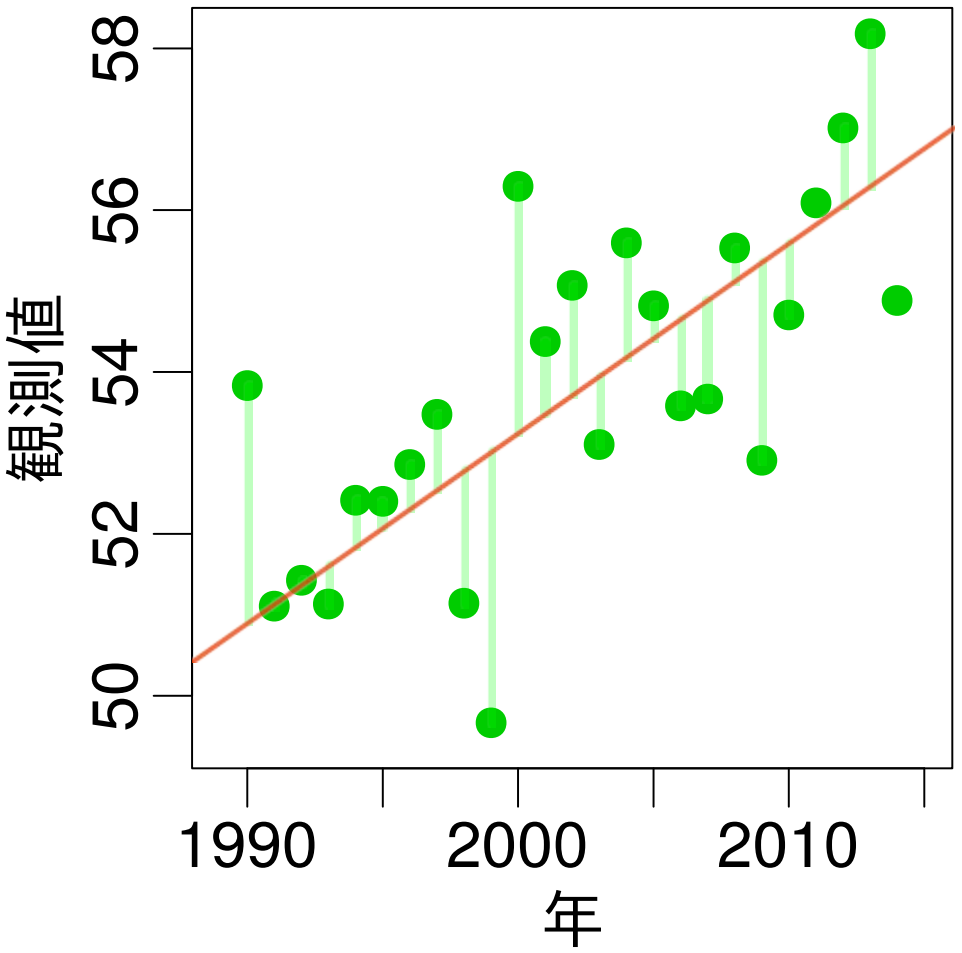
時間的自己相関がある

時間的自己相関がない

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

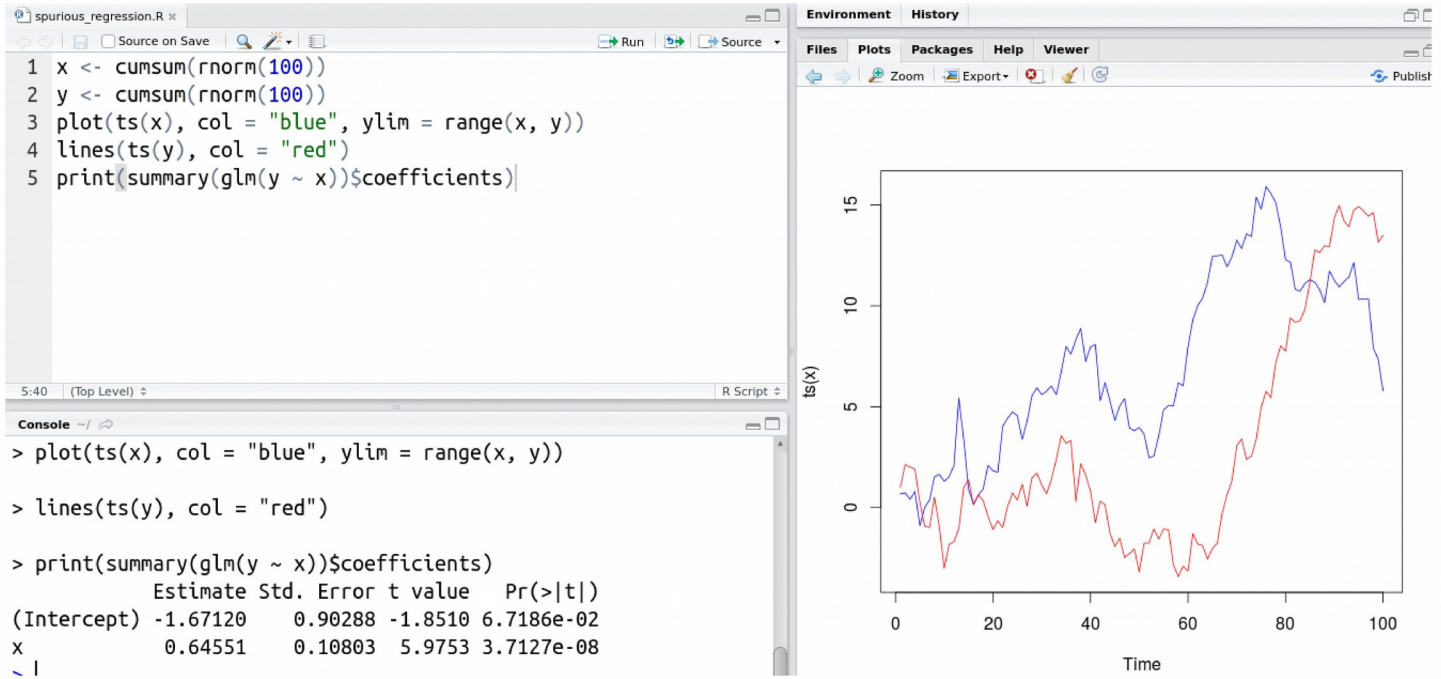
時間的自己相関がない

9/19 (j)

時系列データ解析

状態空間モデル（SSM）の続きと疑わしい回帰（spurious regression）

久保拓弥（北海道大・環境科学）



統計モデルづくりの要点

時系列データの解析は

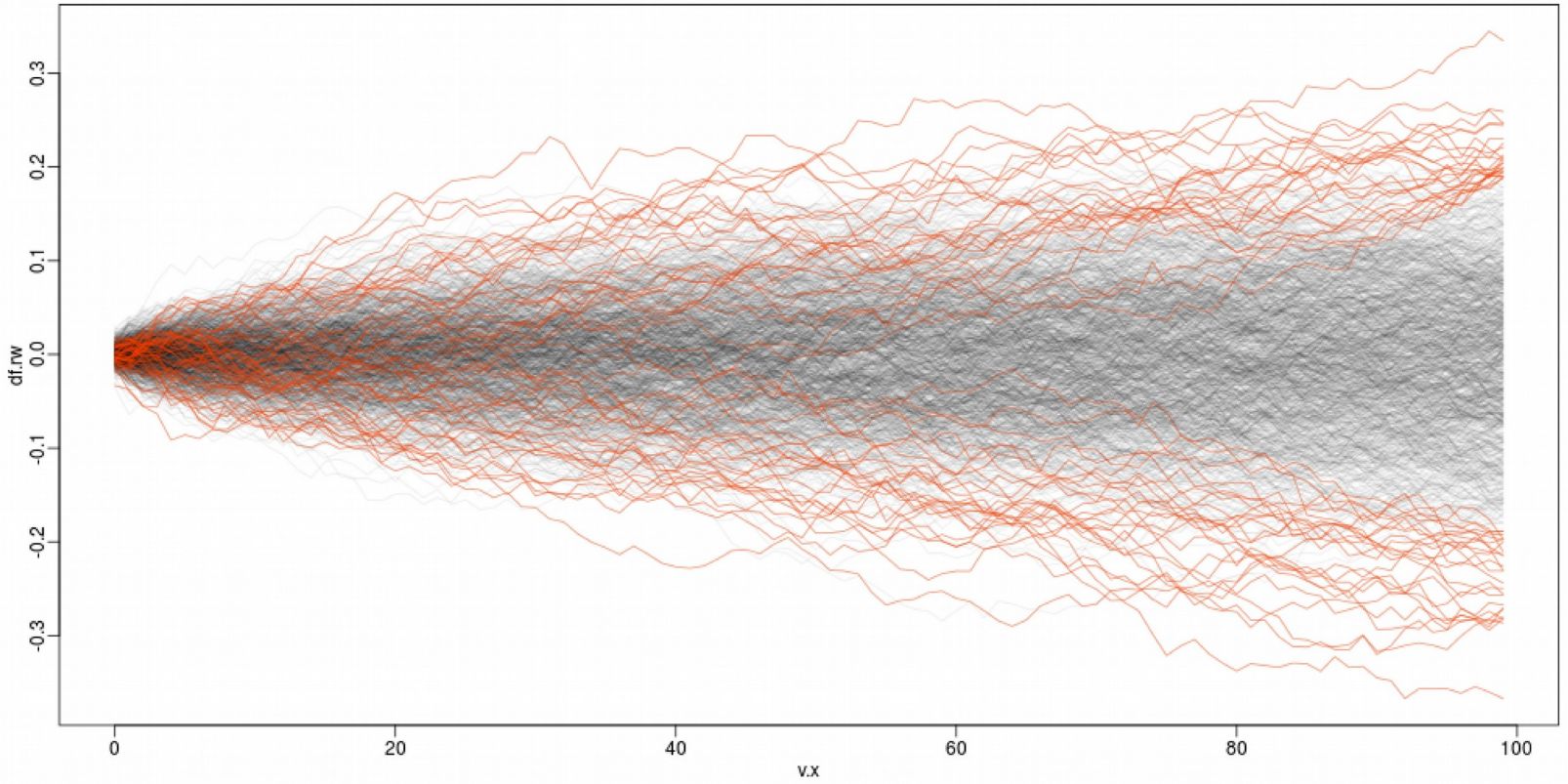
階層ベイズモデル化した

状態空間モデルを使うのが便利

9/19 (i)

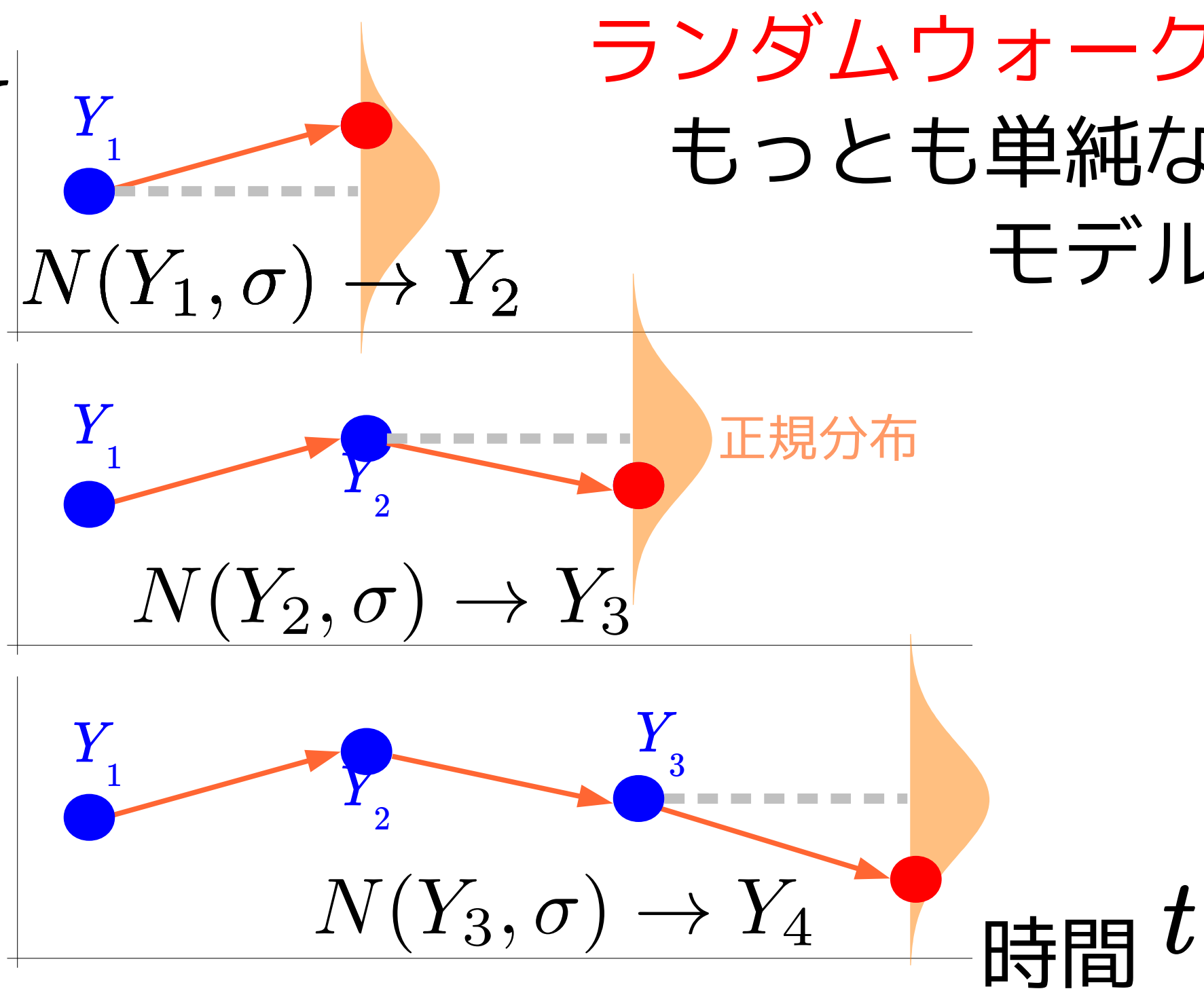
生態学の時系列データ解析でよく見る『あぶない』モデリング

久保拓弥 mailto:kubo@ees.hokudai.ac.jp

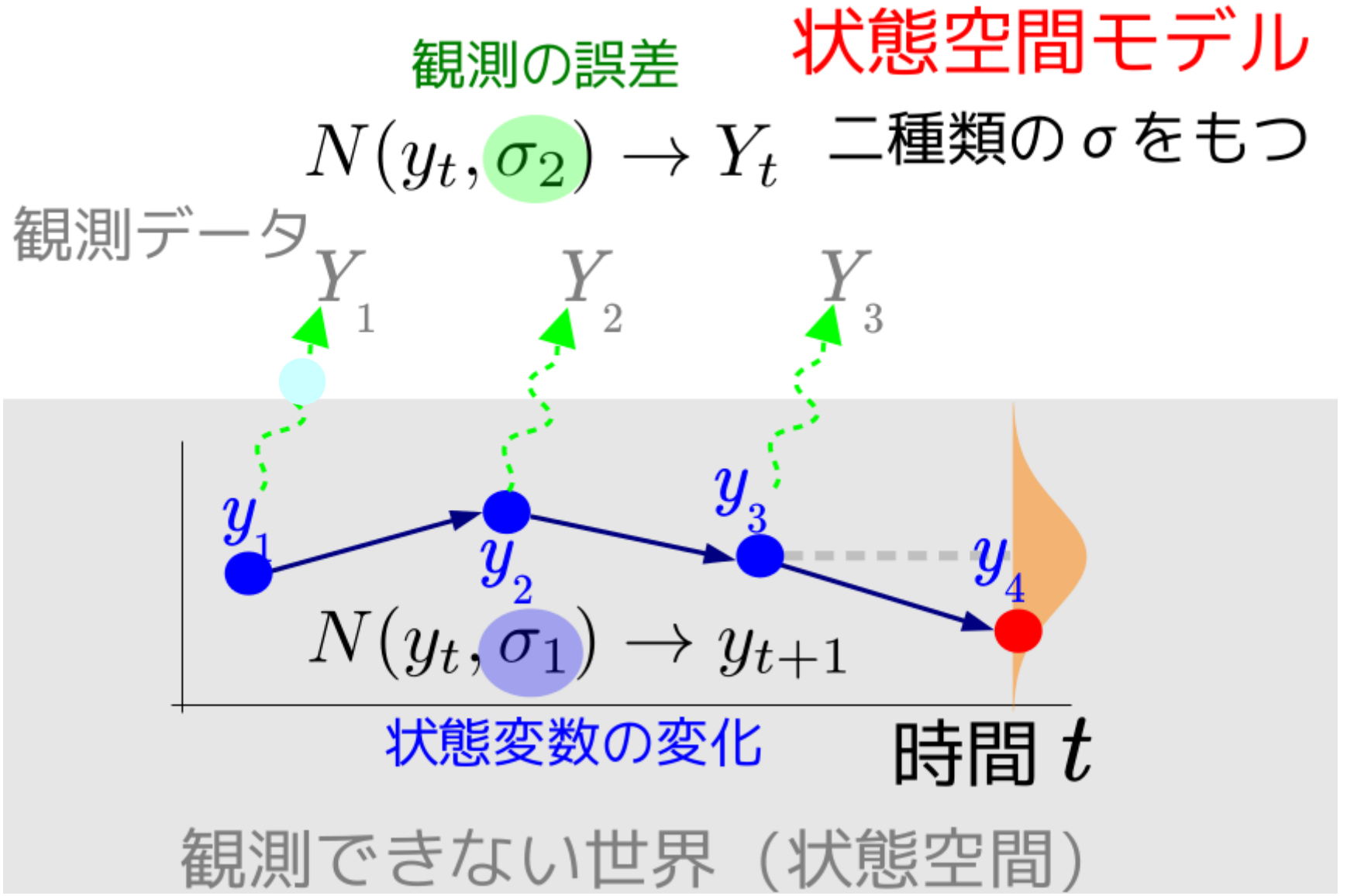


ランダムウォーク

もっとも単純なモデル



状態空間モデル + 観測モデル



全体のながれ…でした

データの性質・構造をよくみて統計モデルを作る