統計モデリング入門 2019（a）

久保の多様性生物学の講義全体の流れの紹介

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

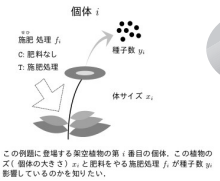




図 3.1 この例題に登場する架空植物の第 $i$ 番目の個体．この植物の体サイズ（個体の大きさ）$x_i$ と肥料をやる施肥処理 $f_i$ が種子数 $y_i$ にどう影響しているのかを知りたい．



## この統計モデリング授業の みなさんの質問に基づく成績評価

- 全 8 回の授業，それぞれについて
  - 「よくわからなかった」箇所などなど…について，久保あてにメイルを送ってください．質問の長さは自由
  - 例「ポアソン分布と二項分布のちがいがわかりません」などなど…
  - 「ここまではわかっているんですよ」といったアピールも可能
  - よく考えられた質問には，よい点をあげたい…



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

- 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



## 統計モデリング授業の web page

http://goo.gl/76c4i

この授業のポータル?



## この授業の主題

よい統計モデリング
とは何だろう?

…あるいは…

悪いモデルとは?

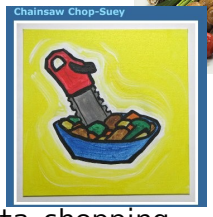
Ecologists' data-cooking
nice data!

http://www.wstephens.net/

## A caricatured example

Ecologists' data-cooking
nice data!

http://www.wstephens.net/




data-chopping
http://alispagnola.blogspot.com/

Data-cooking in Ecology

http://www.wstephens.net/


http://alispagnola.blogspot.com/

data-chopping

http://chan4chan.com/

data-flaming
http://howdoifeelaboutthis.tumblr.com/

Ecologists' data-cooking


your conclusion!

## データ解析は統計モデルを作り！

- データ解析とは，統計モデルをデータにあてはめてみること！
- 統計モデルがおかしいと結論もおかしい
- データの性質・構造などをよくみて，よりよい統計モデルを作ろう（試行錯誤が重要）

まあ，授業しながら，おいおいと…



## 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

- (この授業限定の用語)「ゆーい差」とは?
- 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
- ブラックボックス統計解析



## 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!

- 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 R to show your data structure

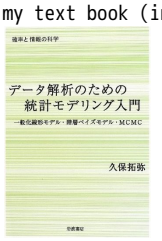


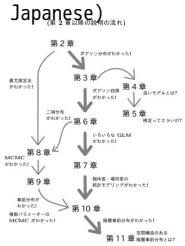
## 教科書とソフトウェア


## 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)

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!!



Statistical software for this course

## 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

追記メモ：RStudio の紹介！



## 統計モデルとは何か?

What? statistical modeling?

# 「統計モデル」とは何か?

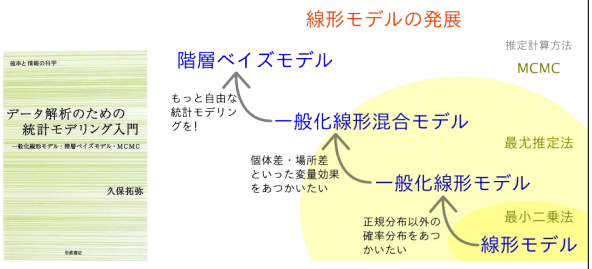
どんな統計解析においても
統計モデルが使用されている


- 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

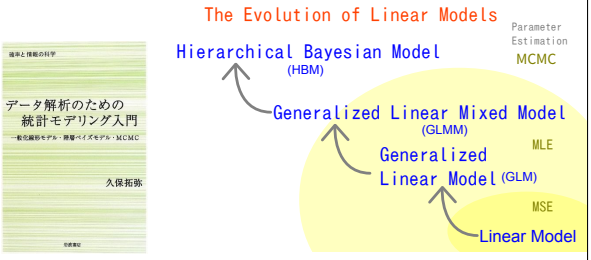
「何でも正規分布」じゃないだろ!






# GLM and extended GLMs!

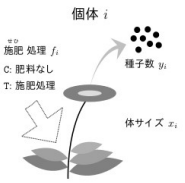
a better statistica model for better data analysis!






# たとえばこんなデータがあったしましょう

An example




number of seeds

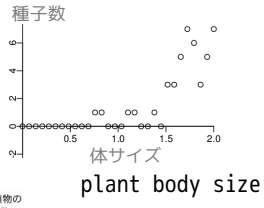
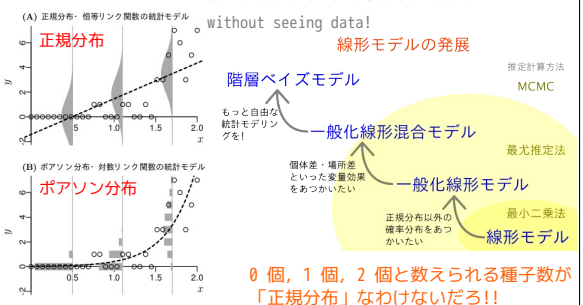



図 3.1 この例題に登場する架空植物の第 $i$ 番目の個体．この植物の体サイズ（個体の大きさ）$x_i$ と肥料をやる施肥処理 $f_i$ が種子数 $y_i$ にどう影響しているのかを知りたい．



# 一般化線形モデル - ばらつきをよく見る

Don't use the normal distribution without seeing data!




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

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（1/3）

第 1 回 観測されたパターンを説明する統計モデル
Introduction

第 2 回 確率分布と最尤推定
Probability Distributions and Maximum Likelihood Estimation (MLE)

第 3 回 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（2/3）

第 4 回 モデル選択と検定
Model Selection and Statistical Test

第 5 回 一般化線形モデル: ロジスティック回帰
GLM: Logistic Regression

第 6 回 階層ベイズモデル 1
Hierarchical Bayesian Models (HBM) 1

# 全体の流れ（3/3）

第 7 回 繰り返し測定の階層ベイズモデル
Bayesian models for repeated measurements

第 8 回 時間変化データのベイズ統計モデル
Bayesian models for time series data

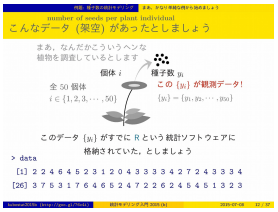
The End!

Overview
Statistical Modeling 2019 (b)

Probability distributions and
maximum likelihood estimation

さまざまな確率分布と最尤推定

## 単純化した例題

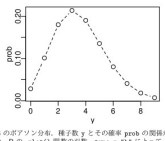


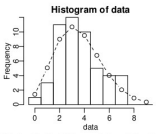

Simplified examples to learn statistical modeling



カウントデータとは，
たとえば植物の種子が
１個，２個，３個，…
と数えられるデータの
ことをさす．

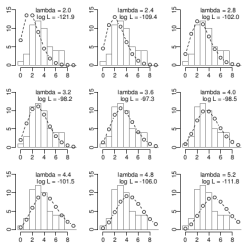
人間の身長などはカウ
ントデータではない

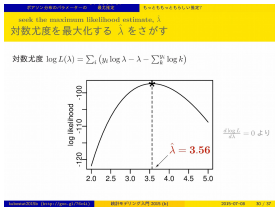





さいゆう
## 最尤推定という考えかたを説明します






How to fit the distribution to the observation?

Maximum likelihood estimation!

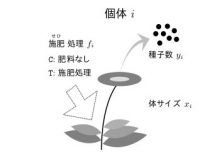


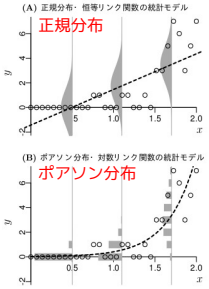
Overview
Statistical Modeling 2019 (c)

Poisson regression
and generalized linear model

ポアソン回帰と GLM

## ここで登場する ---
## 「何でも正規分布」ではダメ！という発想

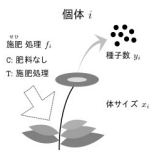


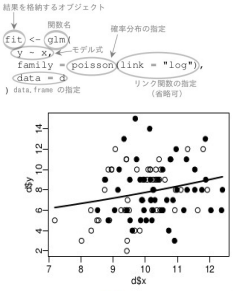




the “normal distirbution is NOT “normal”



Free の統計ソフトウェア R で統計モデリング





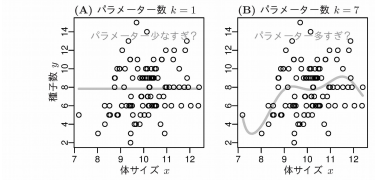

Overview
Statistical Modeling 2019 (d)

Model Selection
and Statistical Test

モデル選択と統計学的検定

## statistical model selection
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 $k$?



## model selection for better predictions
A. より良い予測をする統計モデルを探すこと

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 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?

fallacy of statistical significance?

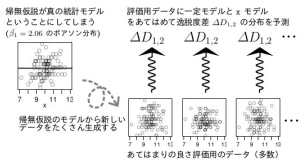



図 6 尤度比検定に必要な $\Delta D_{1,2}$ の分布の生成。まず帰無仮説である一定モデル（$\hat{\beta}_1 = 2.06$, p.[] 参照）が真の統計モデルだと仮定し，そこから得られるデータを使って逸脱度差 $\Delta D_{1,2}$ がどのような分布になるかを調べる。



## Overview
## Statistical Modeling 2019 (e)

Logistic regression,

a generalized linear model

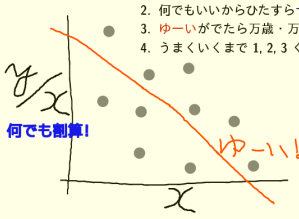
ロジスティック回帰

## mesurement / mesurement?… sounds bad!
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




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



## Important: statistical education for graduate students in ecology

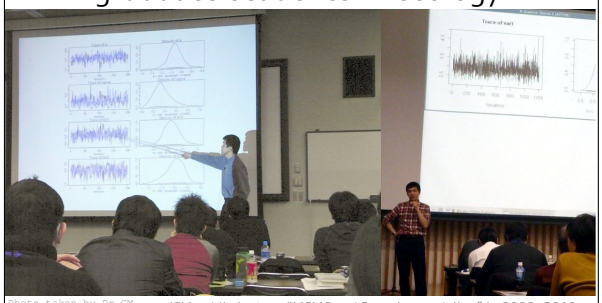


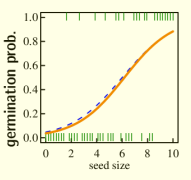
Photo taken by Dr.GM ISM public lecture "MCMC and Bayesian modeling" in 2009, 2010

統計モデリング教育が重要

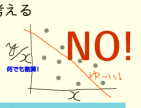
## Use logistic regressions!
GLM のひとつ，ロジスティック回帰を使おう

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

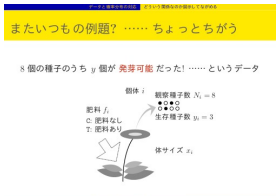


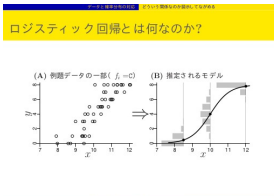

コツ: 不自然にデータをこねくりまわさない
データの性質・構造にあったモデリングを!

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

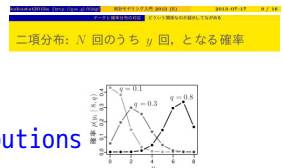


## GLM のひとつ，ロジスティック回帰を使おう




a statistical model
for fractions
using binomial distributions



8/5

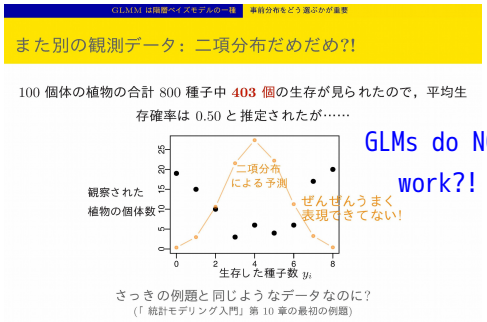
Overview
Statistical Modeling 2019 (f)

Hierarchical Bayesian model
and MCMC sampling

階層ベイズモデルと MCMC

---

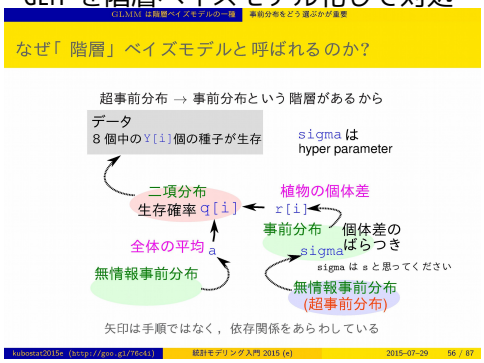
GLM ではうまく説明できないデータ!?




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

---

A solution: Hierarchical Bayesian GLM
GLM を階層ベイズモデル化して対処





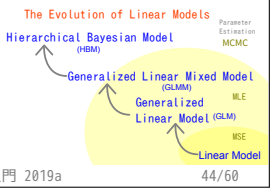

---

なぜ階層ベイズモデルまで勉強するの?

たとえば生態学とかでは…

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

生態学にかぎらず，実験などにおける反復でも，必要…






---

第 7, 8 回は
「時間変化」するデータ
の統計モデリング
(階層ベイズモデルの応用)
Modeling of time-series data as an application of hierarchical Bayesian modeling!

---

Overview
Statistical Modeling 2019 (g)

Modeling time change data
(short term)

短い時系列データの統計モデル

---

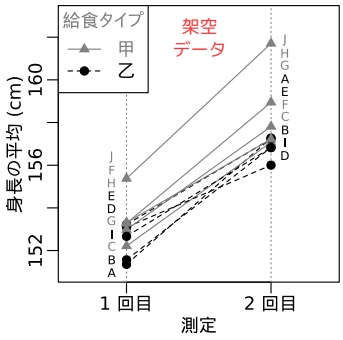
A Time series model
for single step data
短い時系列データ

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

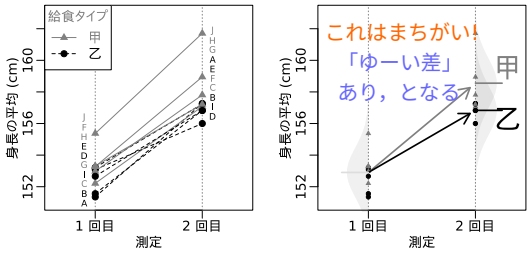
---

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

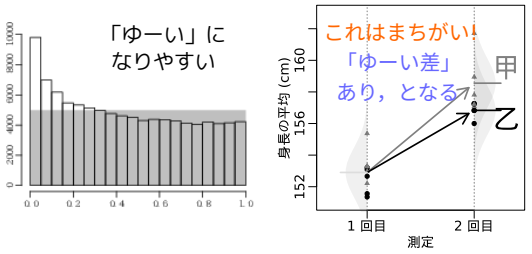



glm(身長 ~ (測定2回目) + (測定2回目):(処理の効果))

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

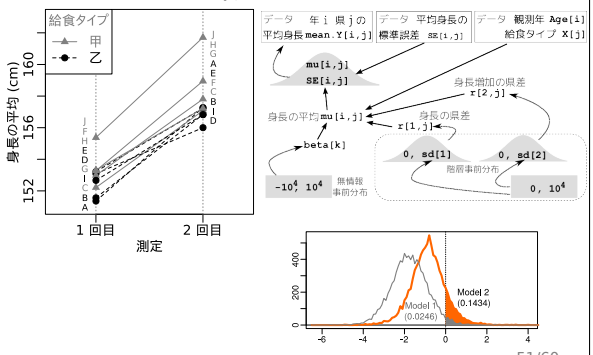



glm(身長 ~ (測定2回目) + (測定2回目):(処理の効果))

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



## Overview
## Statistical Modeling 2019 (h)

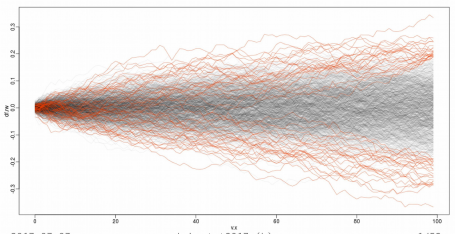
Modeling time series data

(long term)

長い時系列データの統計モデル

## 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

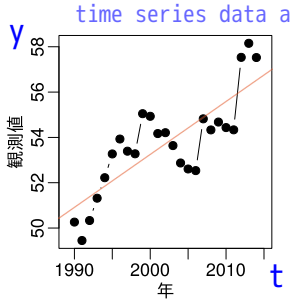
久保拓弥 mailto:kubo@ees.hokudai.ac.jp





## 時間相関のある時系列データに…

time series data and autocorrelation



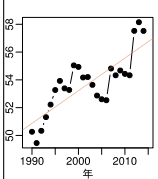

glm(y ~ t)

…と，モデルをあてはめてみた



## 「やったーゆーいだ!!」……??

A fake significance



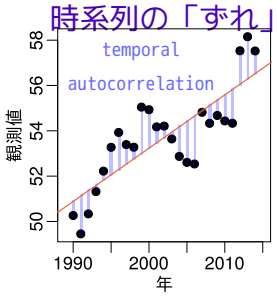
```
> summary(glm(formula = y ~ t))

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.1295  -1.0583  -0.0817   0.9860   2.0188

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -414.5655    71.4761   -5.80  6.6e-06
t              0.2339     0.0357    6.55  1.1e-06
```

これはまちがい→ glm(時系列Y ~ 時間 t)

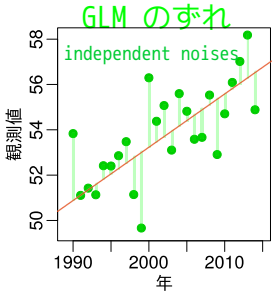
統計モデルがおかしい?



## 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

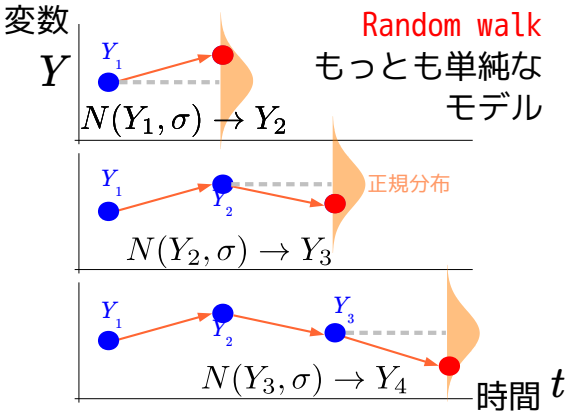
統計モデルづくりの要点

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

Latent state model is a better model to know the characteristics of time-series data





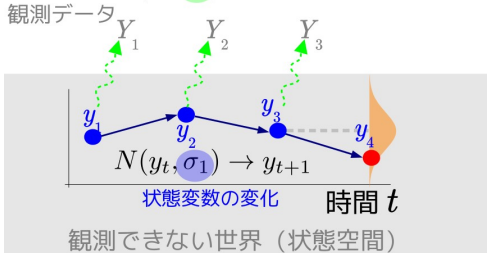

状態空間モデル + 観測モデル

Latent state variables + observation model

観測の誤差 状態空間モデル

$N(y_t, \sigma_2) \to Y_t$ 二種類の $\sigma$ をもつ

観測データ




観測できない世界（状態空間）



この時間はここまで

any questions?