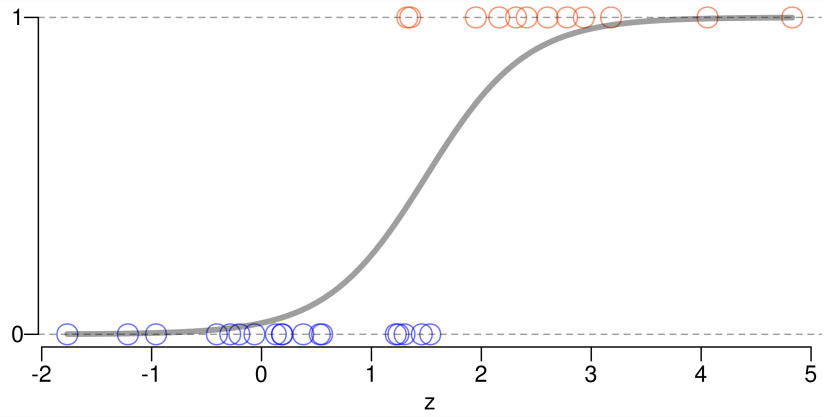
特別講義「生物統計学」

概要紹介・全体の流れについて

久保拓弥（北海道大・環境科学）

kubo@ees.hokudai.ac.jp

- この講義についての連絡
- 統計学の役割
- 統計モデルとは何か
- この講義全体を一覧



講義についての連絡事項

- この講義には web page があります
  - 単位取得のためのレポート問題・提出についての情報
  - 講義資料や例題のデータなどがダウンロード可能

http://goo.gl/wijx2

レポート提出
しめきりは 11/16

提出方法は
web page を参照

全体の流れ

「生物統計学」の時間わり

11/1 (木) 午後

- **(k1)** 統計モデルとは何か? なぜ必要か?
- **(k2)** 確率分布と一般化線形モデル (GLM)
- **(k3)** モデル選択と統計学的検定

11/2 (金) 午前 — R の練習

- **(r1)** R をちょっと使ってみる，図を作る
- **(r2)** R で `glm()` 使って，また図を作る

11/2 (金) 午後

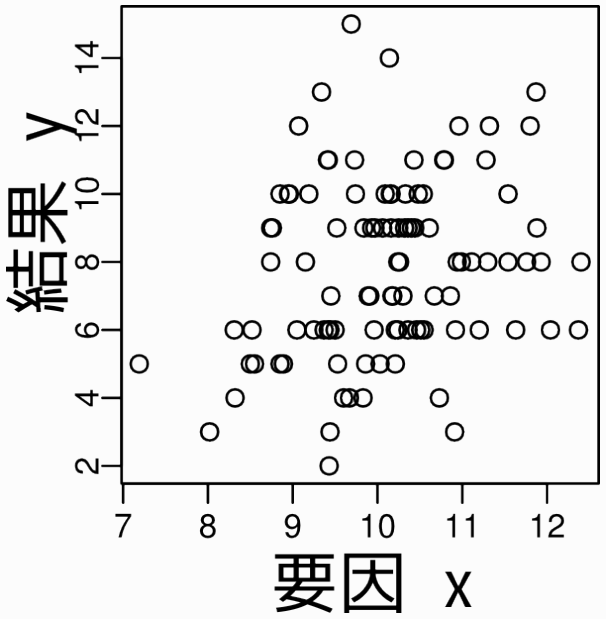
- **(k4)** 何でも割算するな! — GLM で解決できる
- **(k5)** 個体差をいれて階層ベイズモデルを作ろう
- **(k6)** いろいろな階層ベイズモデル

自己紹介

- 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
  - 安楽椅子生態学者
  - 統計モデルを使ったデータ解析
- データは誰か別の人（共同研究者）がとってきてくれます

統計学的手法の役割

科学的な手法の役割（のひとつ）

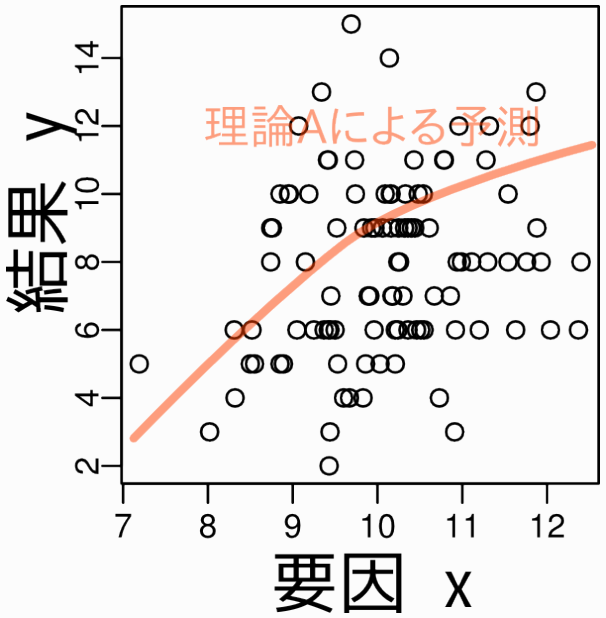


観測された現象の背後にひそむ，何か単純な原理を知りたい!

たとえば何かこのような観測データが得られたとする

- 生物個体ごとに異なる要因 X によって興味のある結果 Y が変化するらしい
- データはばらついているが，この関係を要約できないだろうか？

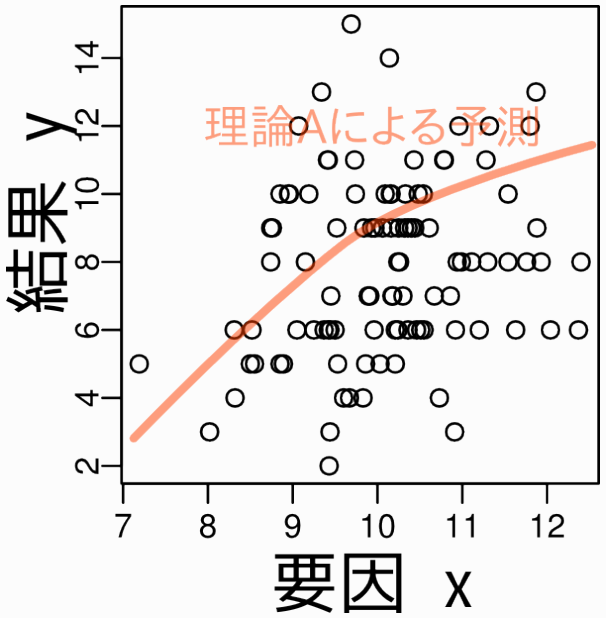
なにか理論にもとづく予測

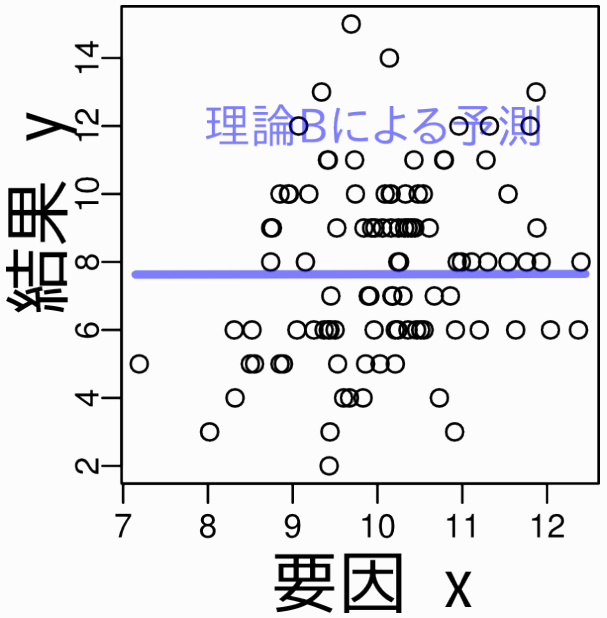


理論Aとやらを使うと要因Xと結果Yの関係が説明された?!

「良い」予測をするのが良い科学理論である!

しかし理論ってのはいくつでも作れるわけで……

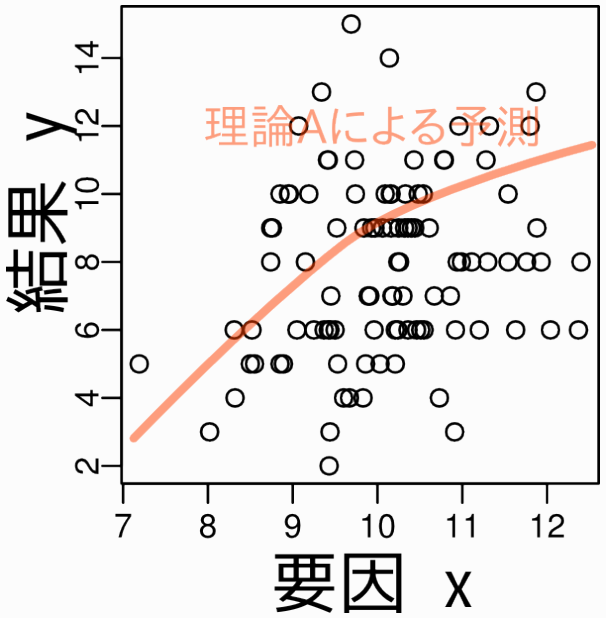


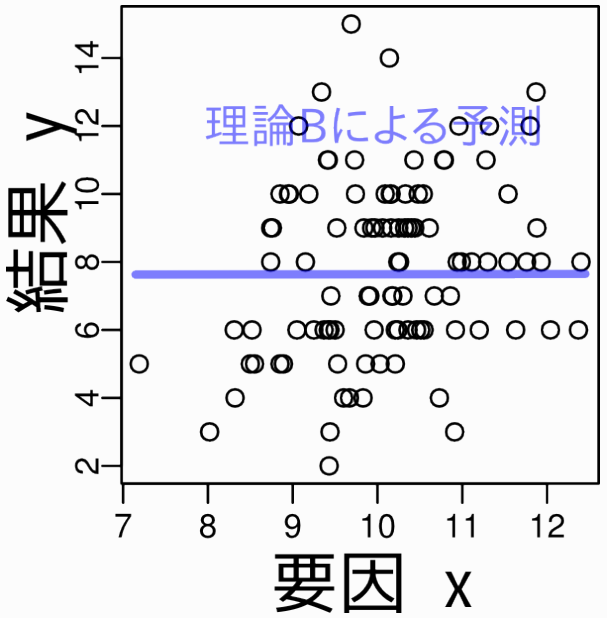


どっちが良い理論?

「良い」予測をするのが
良い科学理論である!

統計学的な手法の役割（のひとつ）

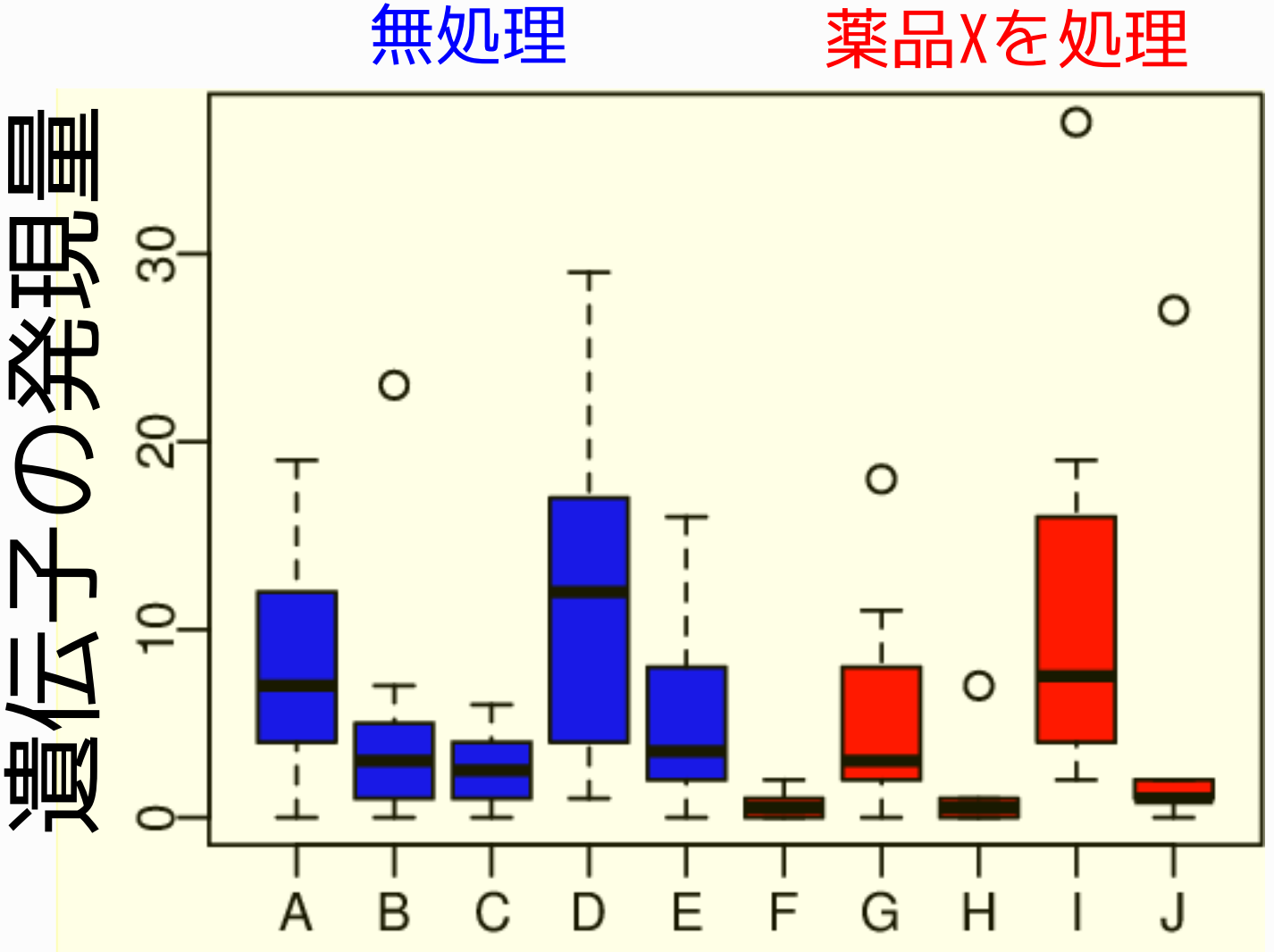




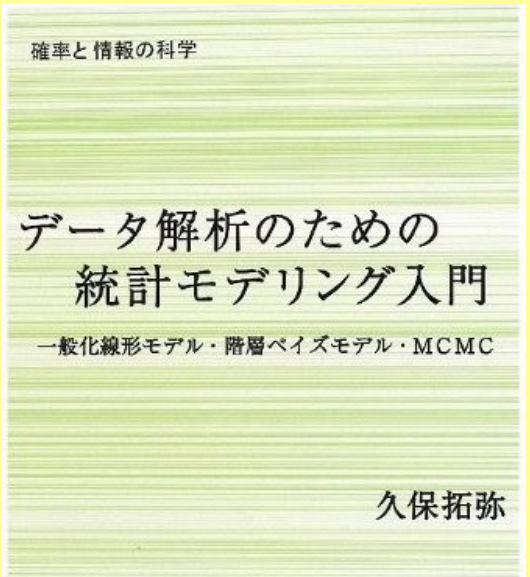
統計学的な手法でできること

- どちらがよりよい予測なのかを評価する
- 各理論それぞれであてはまりのよい線を探す

どんな生物学の分野でも統計学的な手法は必要



統計モデルとは何か?
&
「統計モデリング入門」宣伝



「統計モデル」とは何か?

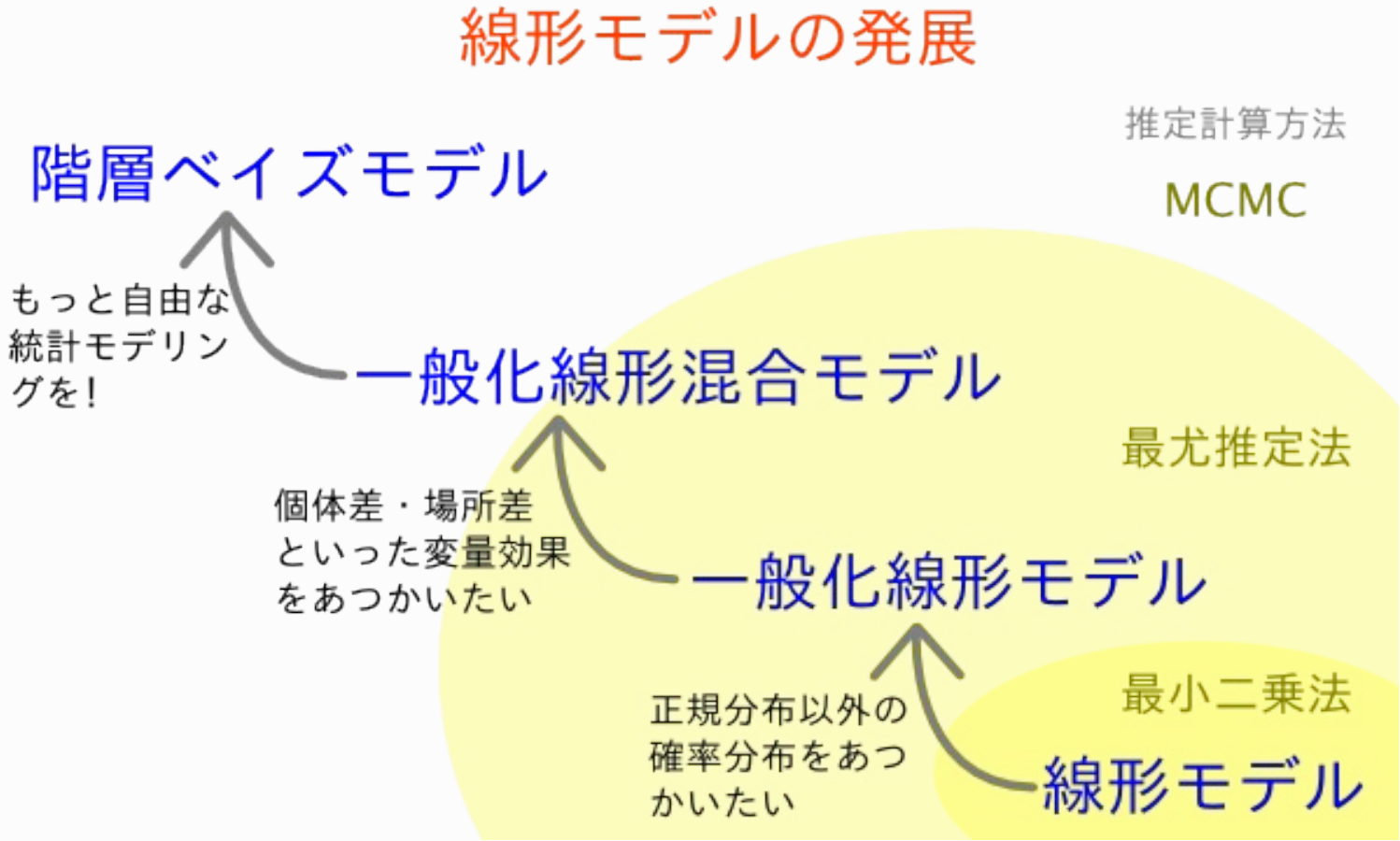
どんな統計解析においても
統計モデルが使用されている

- 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!

2012-05-18 刊行
岩波書店



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）

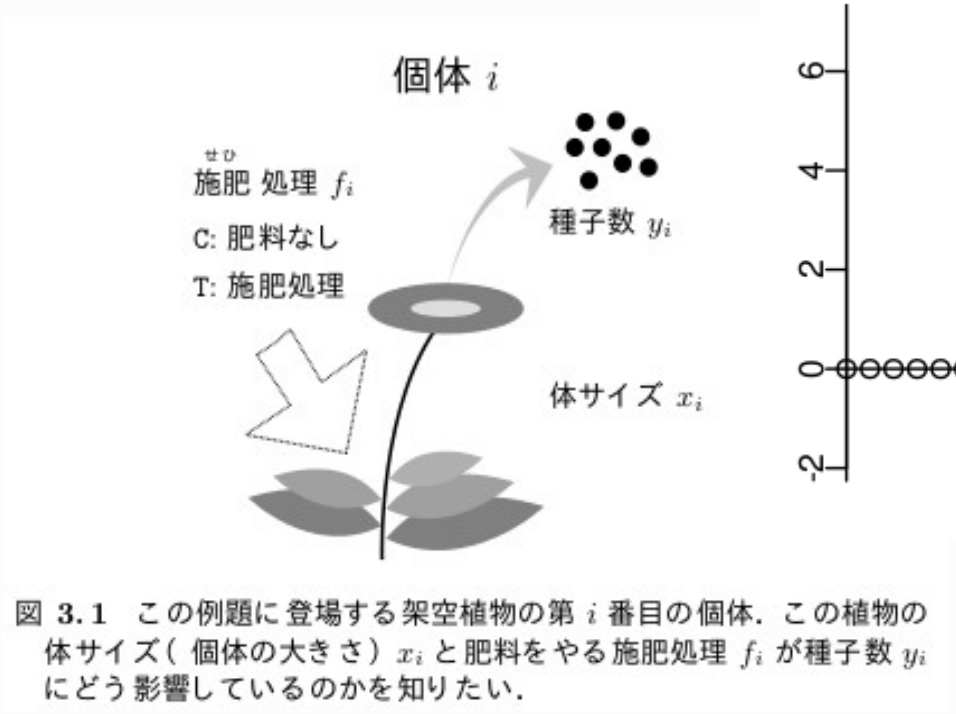
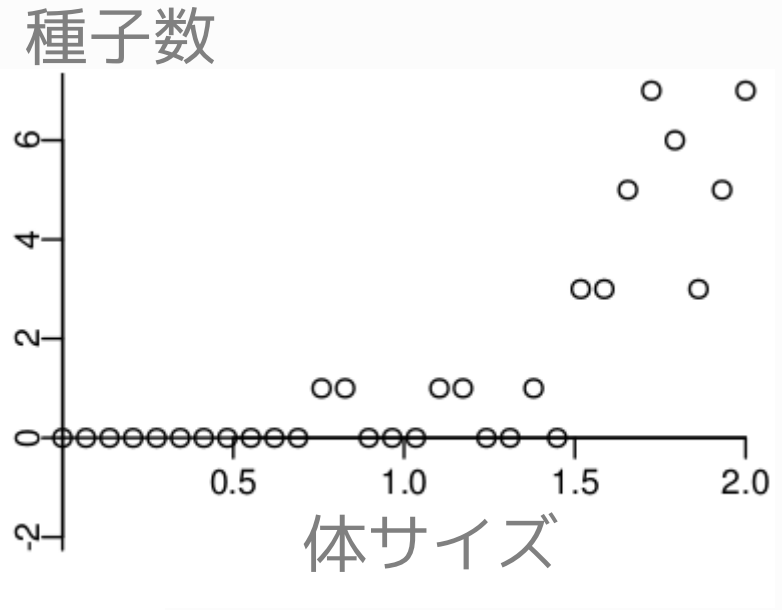


図 3.1　この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（ 個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．



GLM - ばらつきをよく見る

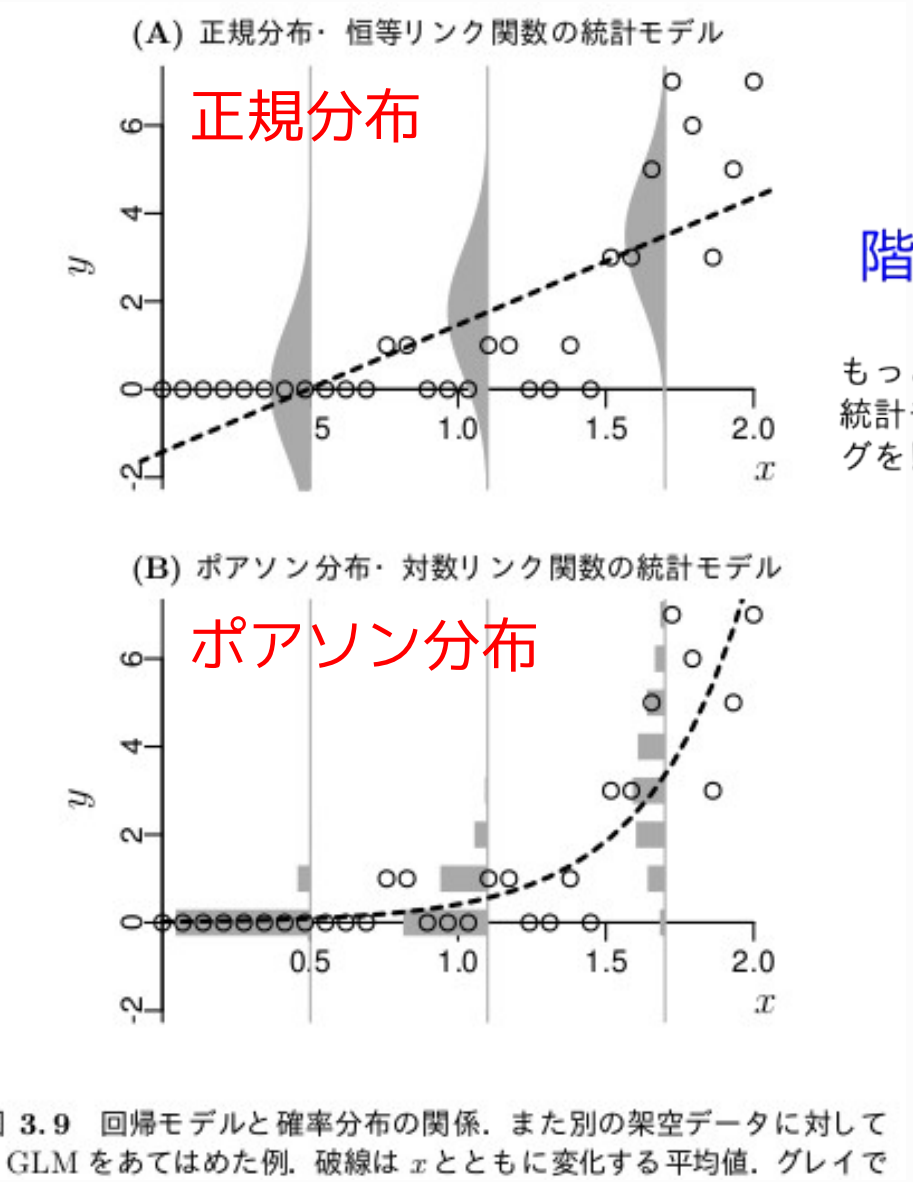


図 3.9 回帰モデルと確率分布の関係．また別の架空データに対して GLM をあてはめた例．破線は x とともに変化する平均値．グレイで

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

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!!

この講義全体を一覧

「生物統計学」の時間わり

11/1 (木) 午後

- **(k1)** 統計モデルとは何か? なぜ必要か?
- **(k2)** 確率分布と一般化線形モデル (GLM)
- **(k3)** モデル選択と統計学的検定

11/2 (金) 午前 — R の練習

- **(r1)** R をちょっと使ってみる，図を作る
- **(r2)** R で `glm()` 使って，また図を作る

11/2 (金) 午後

- **(k4)** 何でも割算するな! — GLM で解決できる
- **(k5)** 個体差をいれて階層ベイズモデルを作ろう
- **(k6)** いろいろな階層ベイズモデル

11/1 午後 k2

データ解析のための統計モデリング

全 6 回中の第 2 回 (2012–11–01 k2)

確率分布と一般化線形モデル (GLM)

久保拓弥 kubo@ees.hokudai.ac.jp

神戸大の集中講義 web http://goo.gl/wijx2

この講義のーとは「データ解析のための統計モデリング入門」を再編したものです

「統計モデリング入門」 web http://goo.gl/Ufq2

もっと勉強したい人はこの教科書を読んでね

もくじ

単純化した例題

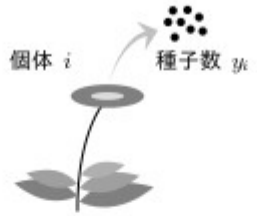


図 1　この章の例題の架空データ．架空植物の第 i 番目の個体．この植物の種子数 y_i．植物個体ごとの葉数やサイズなどについては，何のデータもない．「個体のもつ種子数をどう表現すればよいか」という単純な問題だけを検討する．

ここでは，以下のようなことを説明してみたいと思います:

- 統計学で使う**確率分布** は「正規分布」だけではない!
  - データをよくみて，それをうまく説明できるような確率分布をさがそう
- データの散布図に「セン」をひけばよし，といった考えかたはやめよう!
  - データをよくみて，よりよい「モデル」をつくろう

さてさて……いきなりハナシを始めてしまいますが……**確率分布**（probability distribution）は統計モデルの本質的な部品であり，データにみられるさまざまな「ばらつき」を表現します．この章では，このような「表現の部品としての確率分布」という考えかたを説明するために，簡単な例題データと確率分布の対応づけについて考えます．

1　ひとつめの例題: 種子数の統計モデリング

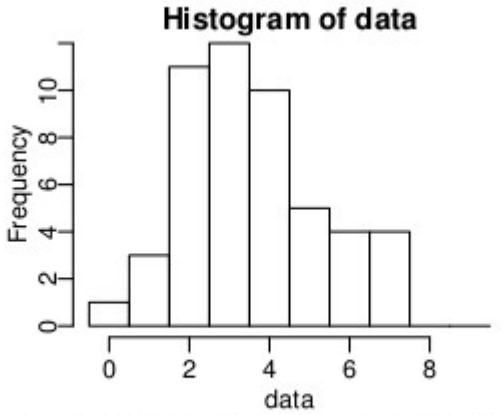


図 2　例題の種子数データのヒストグラム（度数分布図）．横軸は種子数，縦軸は架空
数．全個体数は 50．R の `hist()` 関数による図示．

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

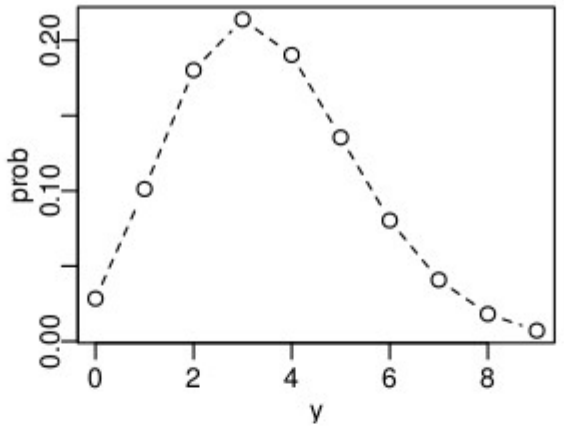


図 4　平均 $\lambda = 3.56$ のポアソン分布．種子数 y とその確率 prob の関係が示されている．図 3 の表を図にしたもの．R の plot() 関数の引数，type = "b" によって「丸と折れ線による図示」，lty = 2 によって「折れ線は破線で」と指示している．

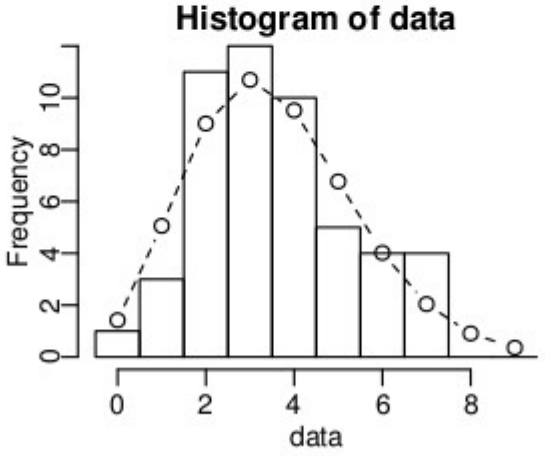


図 5　観測データと確率分布の対応をながめる．ヒストグラムは図 2 と同じ．それに重ねられている丸と破線は y 個の種子をもつ個体数の予測．平均 3.56 の図 4 のポアソン分布の確率分布に全個体数 50 をかけて得られる．

さいゆう
最尤推定という考えかたを説明します

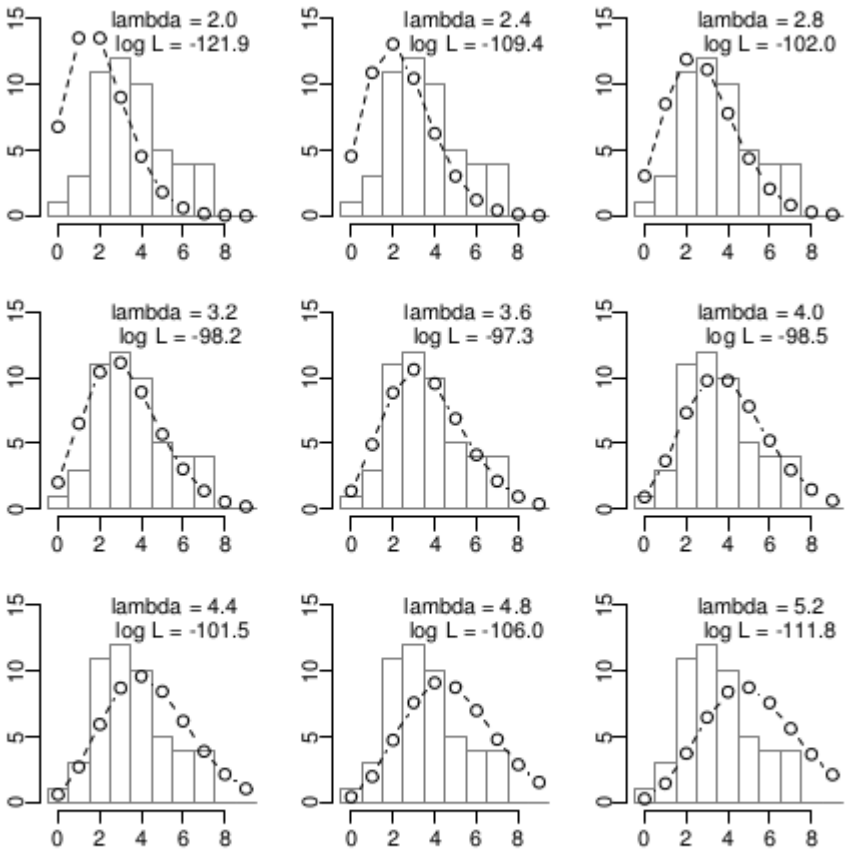


図 7　平均 λ（lambda）を変化させていったポアソン分布と，観測データへのあてはまりの良さ（対数尤度 logL）．すべてのヒストグラムは図 2 と同じ．

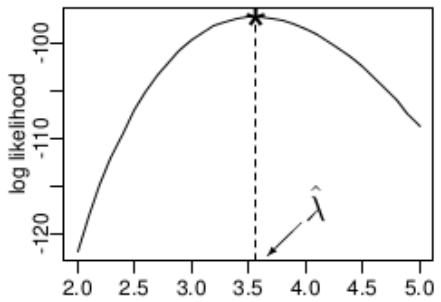


図 8　この章の例題の観測データ（植物 50 個体の種子数）のもとでの λ（横軸）と対数尤度の関係．λ = 3.56 で対数尤度が最大になるので，これが最尤推定値 λ̂ となる．図 7 の λ を連続的に変化させた場合に対応している．

ここで登場する ---
「何でも正規分布」ではダメ！という発想

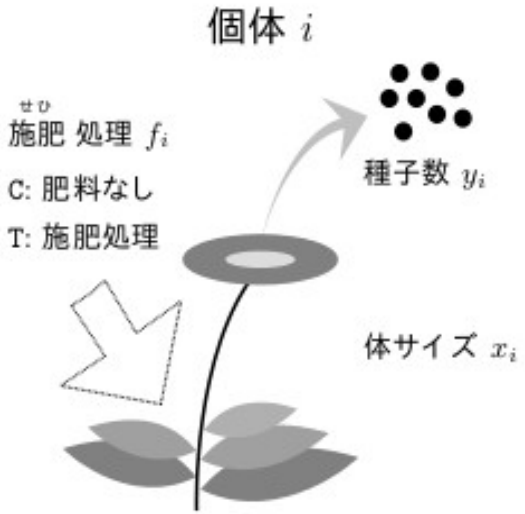


図 3.1　この例題に登場する架空植物の第 i 番目の個体．この植物の体サイズ（個体の大きさ）x_i と肥料をやる施肥処理 f_i が種子数 y_i にどう影響しているのかを知りたい．

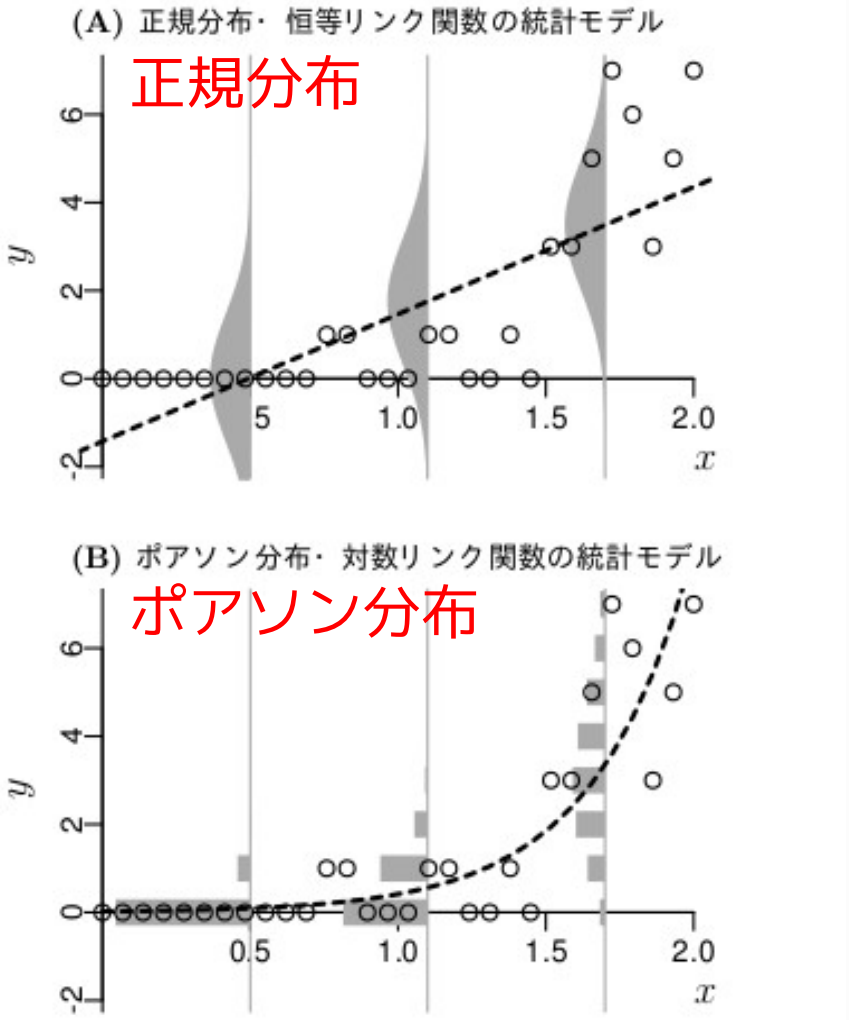


図 3.9　回帰モデルと確率分布の関係．また別の架空データに対して GLM をあてはめた例．破線は x とともに変化する平均値．グレイで

Free の統計ソフトウェア

R で統計モデリング

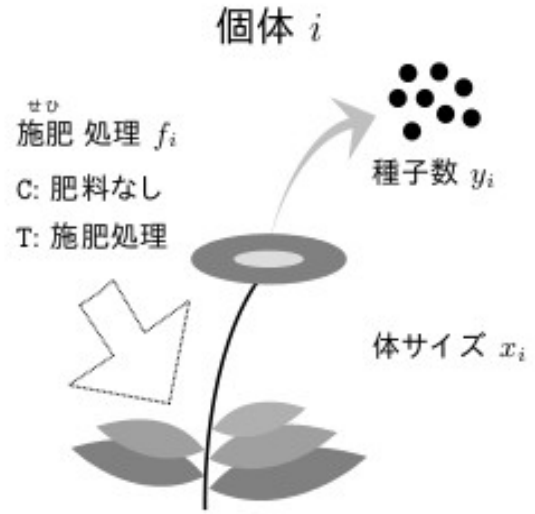
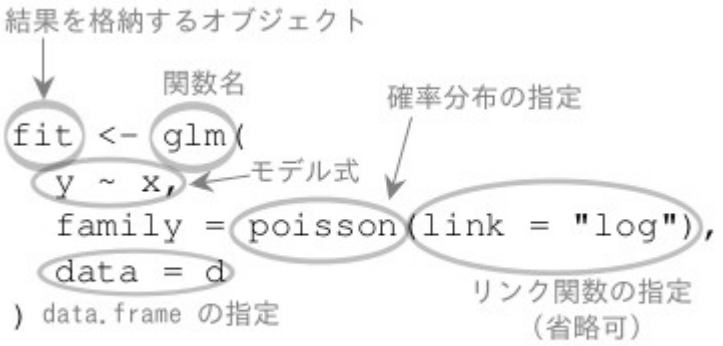


図 3.1　この例題に登場する架空植物の第 i 番目の個体
体サイズ（ 個体の大きさ）x_i と肥料をやる施肥処理
にどう影響しているのかを知りたい.



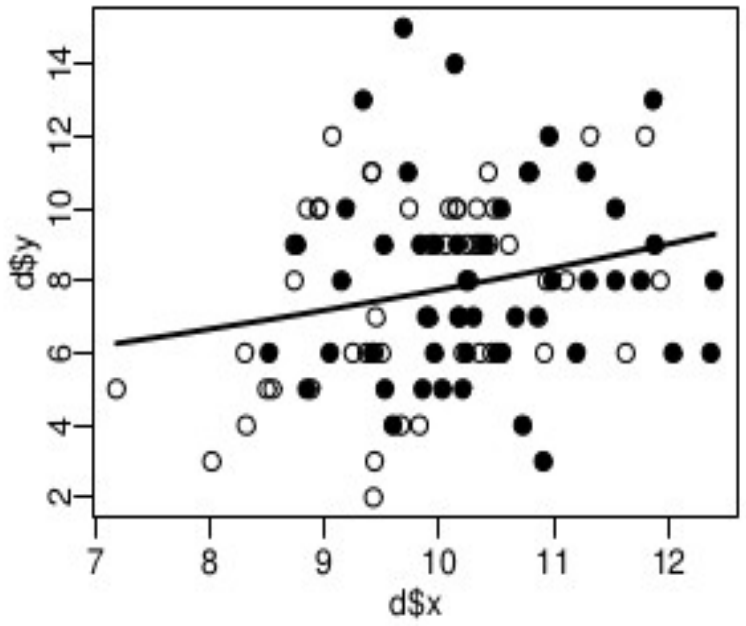


図 17　平均種子数 λ の予測. 図 12 に λ の予測値（実線）を上がきしたもの.

11/1 午後 k3

データ解析のための統計モデリング

全 6 回中の第 3 回 (2012–11–01 k3)

モデル選択と統計学的検定

久保拓弥 kubo@ees.hokudai.ac.jp

神戸大の集中講義 web http://goo.gl/wijx2

この講義のーとは「データ解析のための統計モデリング入門」を再編したものです

統計モデリング本 web http://goo.gl/Ufq2

もっと勉強したい人は「統計モデリング入門」を読んでね

もくじ

Q. より良い予測をする統計モデルは？

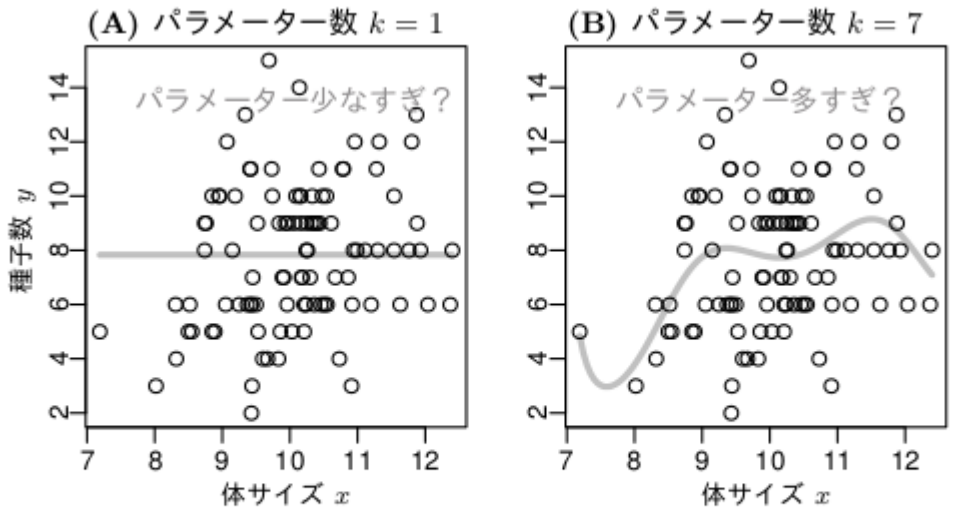


表 2　種子数モデルの最大対数尤度と逸脱度．前の時間のポアソン回帰モデルの種類，最尤推定したパラメーター数 k，最大対数尤度 $\log L^*$，Deviance，Residual deviance の表．各モデルについては図 2 も参照．

モデル	k	$\log L^*$	Deviance $-2\log L^*$	Residual deviance
定数	1	-237.6	475.3	89.5
f	2	-237.6	475.3	89.5
x	2	-235.4	470.8	85.0
x + f	3	-235.3	470.6	84.8
フル	100	-192.9	385.8	0.0

表 3　表 2 に AIC の列を追加した．

モデル	k	$\log L^*$	Deviance $-2\log L^*$	Residual deviance	AIC
定数	1	-237.6	475.3	89.5	477.3
f	2	-237.6	475.3	89.5	479.3
x	2	-235.4	470.8	85.0	474.8
x + f	3	-235.3	470.6	84.8	476.6
フル	100	-192.9	385.8	0.0	585.8

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?

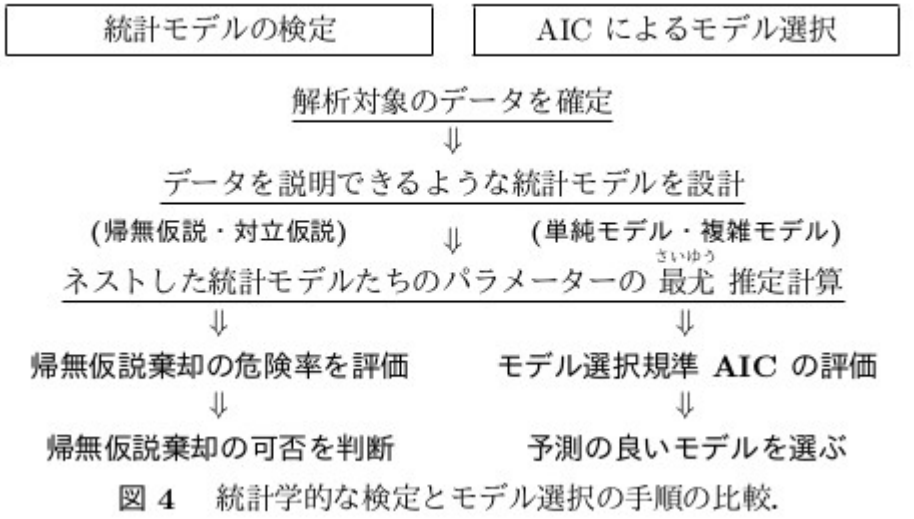


図 4　統計学的な検定とモデル選択の手順の比較.

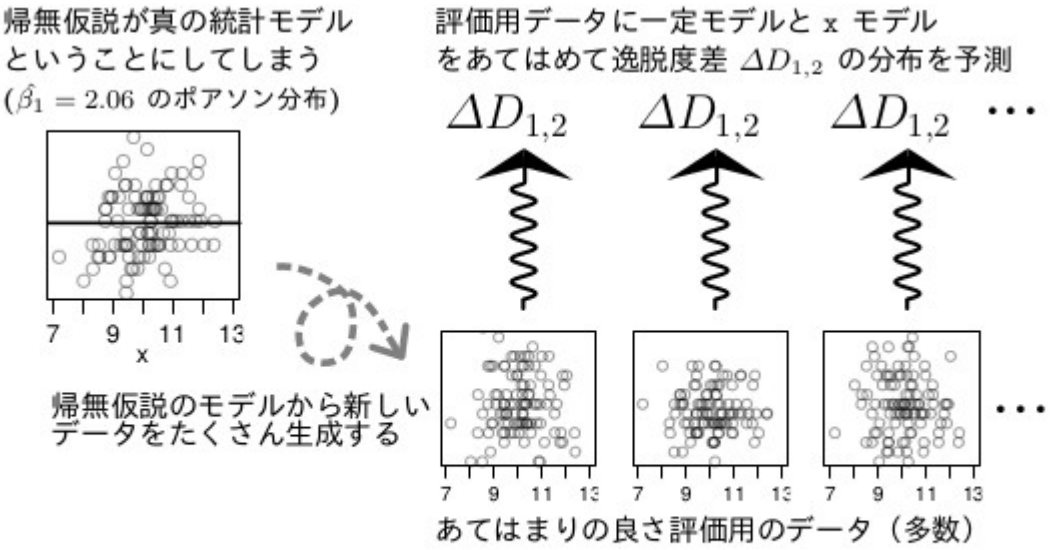


図 6　尤度比検定に必要な $\Delta D_{1,2}$ の分布の生成．まず帰無仮説である一定モデル（$\hat{\beta}_1 = 2.06$, p.6 参照）が真の統計モデルだと仮定し，そこから得られるデータを使って逸脱度差 $\Delta D_{1,2}$ がどのような分布になるかを調べる.

11/2 午前 r1, r2

R の練習 (r1) 2012–11–02

久保拓弥 kubo@ees.hokudai.ac.jp

この授業の web page: http://goo.gl/wijx2

統計ソフトウェア R は研究にたいへん役にたつ free software (無料で入手でき，しかも内部を自由に調べられる) です．今回は R のデータ操作・作図の基本わざを説明します．
R を使ったデータ解析の基本的な流れは次のようになります:

1. データを読みこむ (データフレイム data.frame を作る)
2. 読みこんだデータをいろいろ整理する (データフレイムの操作)
3. データをさまざまな方法で図示する
4. 統計モデリングの設計・あてはめを行う
5. あてはめの結果やモデルの予測を図示する
6. 解析結果をさまざまな方法で出力し，保存する

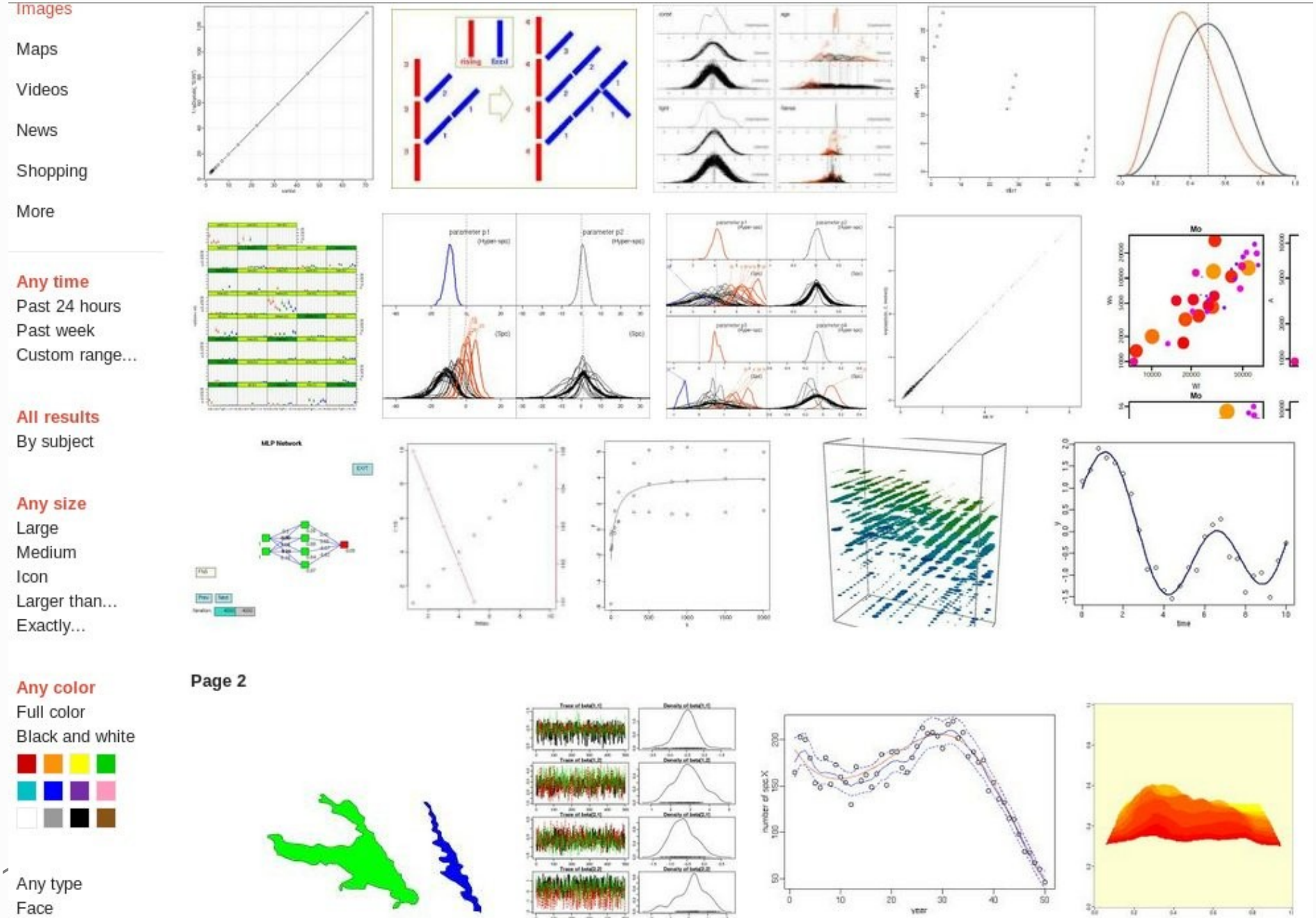
今日は時間も限られているので，データの読みこみ，基本的ななデータフレイム操作，簡単な図示について説明します．上述の授業 web site のあちこちを見て，さらに発展したわざも勉強してください．

1 R でデータフレイムの操作

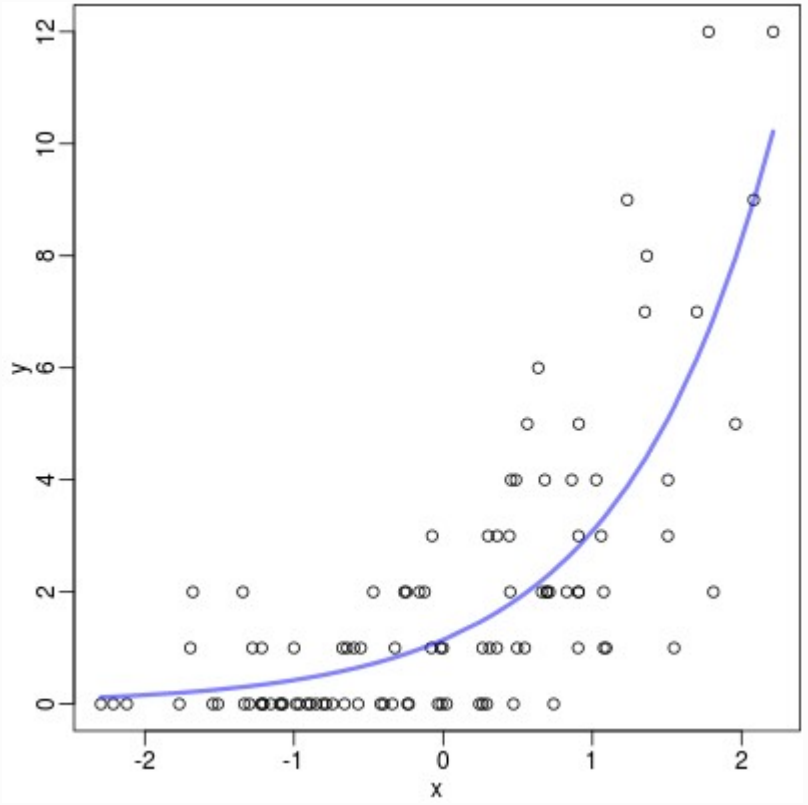
R で作図

久保のサイトを

Google 画像検索

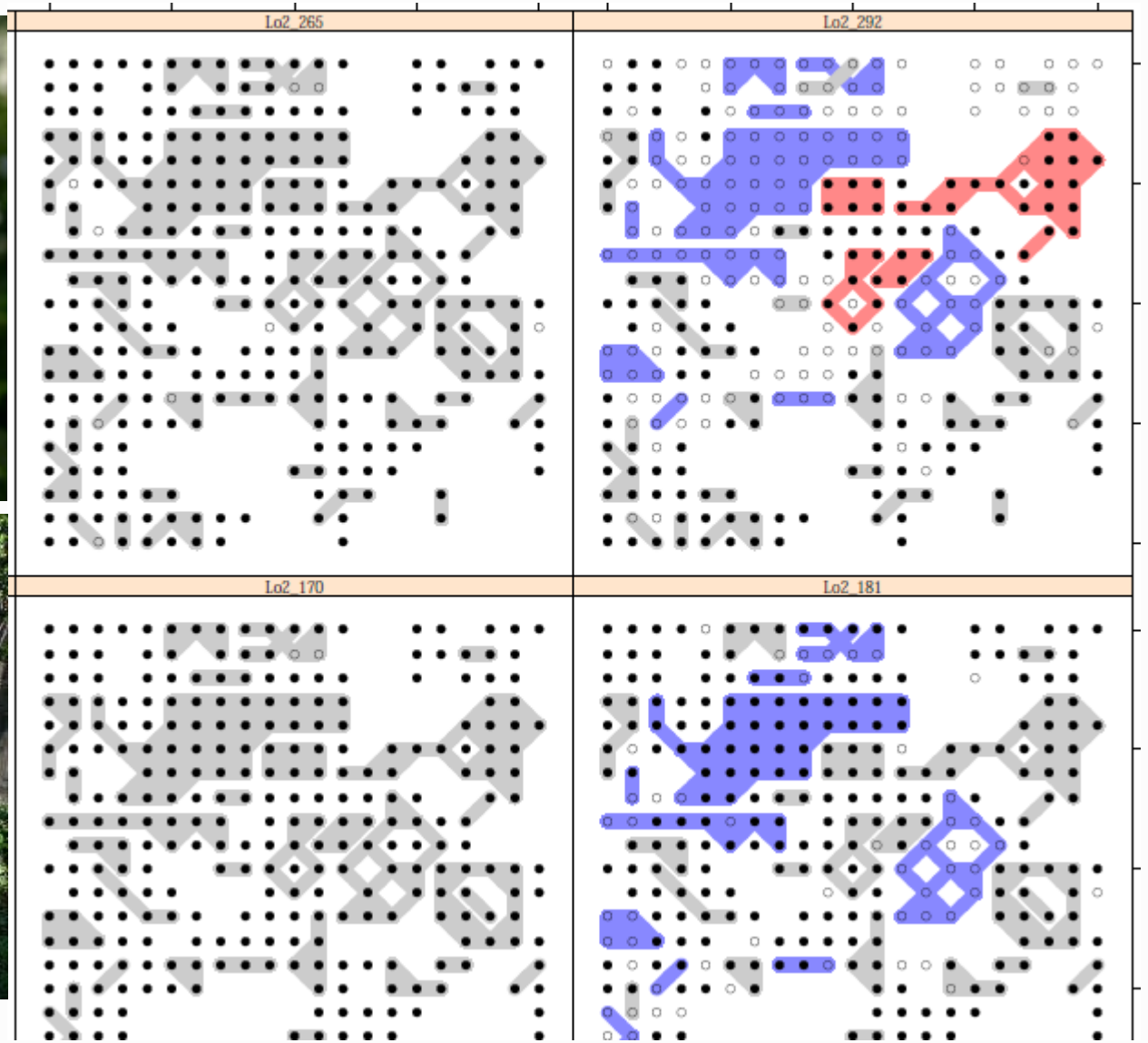


R 作図の基本わざ「重ね描き」




```
plot(x, y) # まずはワクとデータをプロットしてから……
lines(     # 予測線を「追加」する!
    x, exp(predict(fit, newdata = data.frame(x))),
    col = "#0000ff80", lwd = 3
)
```

コンロンソウの「超」個体と遺伝子の特徴をあらわす地図



写真引用
http://blog.goo.ne.jp/taronpe-1944/e/062474b7d1401c3b745f5fcfe8d0d58c

11/2 午後 k4

2012–11–02 k4

神戸大学理学部特別講義

「生物統計学」 (2012 年 11 月) 投影資料

全部で 6 回中の 4 回目

何でも割算するな! — GLM で解決できる

久保拓弥 kubo@ees.hokudai.ac.jp

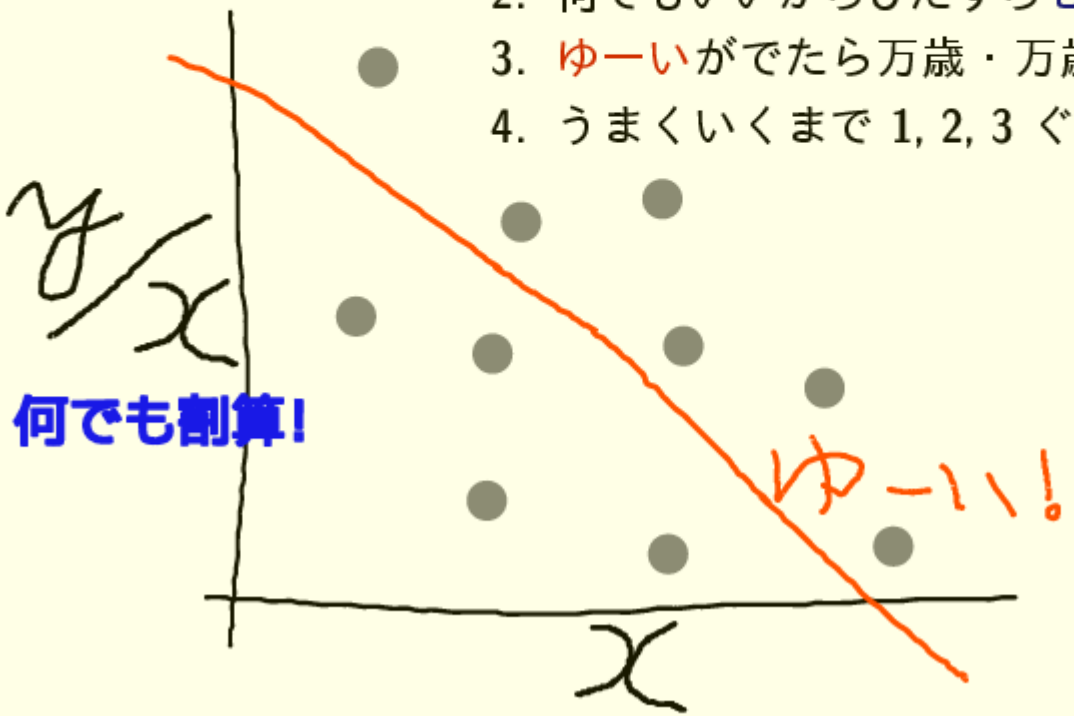
http://goo.gl/wijx2

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



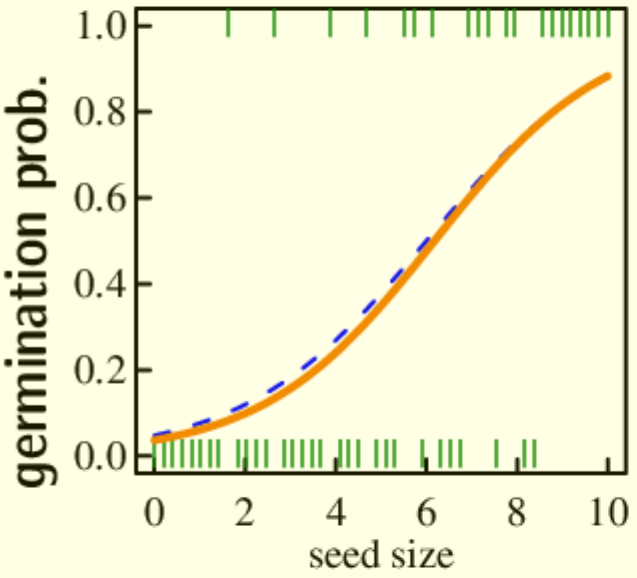
ちなみにこれは x と $\circ/x$ を比較してるんだから，反比例みたいな偽「負の相関」がでるのはあたりまえ

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

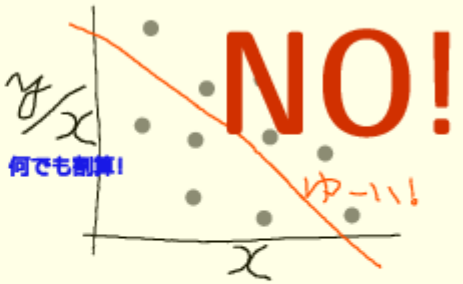
GLM のひとつ，ロジスティック回帰を使おう

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

11/2 午後 k5

2012–11–02 k5

神戸大学理学部特別講義

「生物統計学」 (2012 年 11 月) 投影資料

全部で 6 回中の 5 回目

階層ベイズモデルの基礎

個体差のモデリング

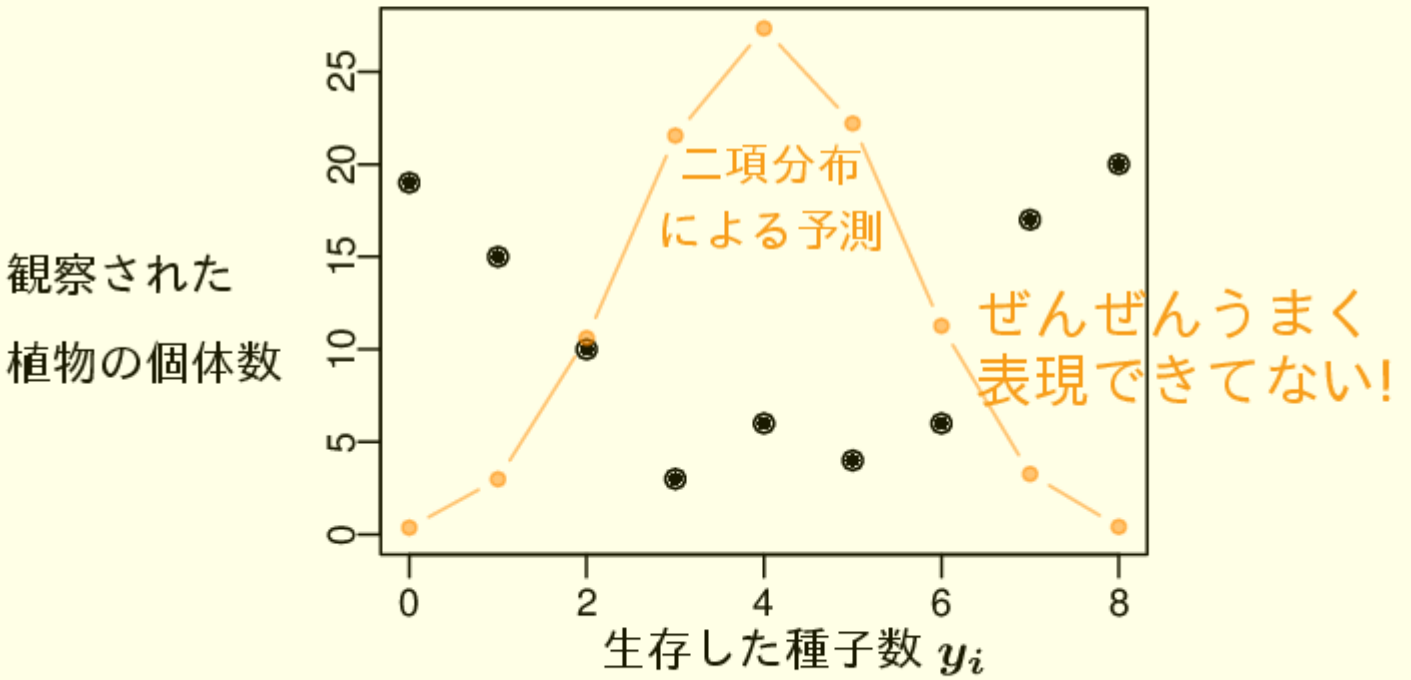
久保拓弥 kubo@ees.hokudai.ac.jp

http://goo.gl/wijx2

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?!

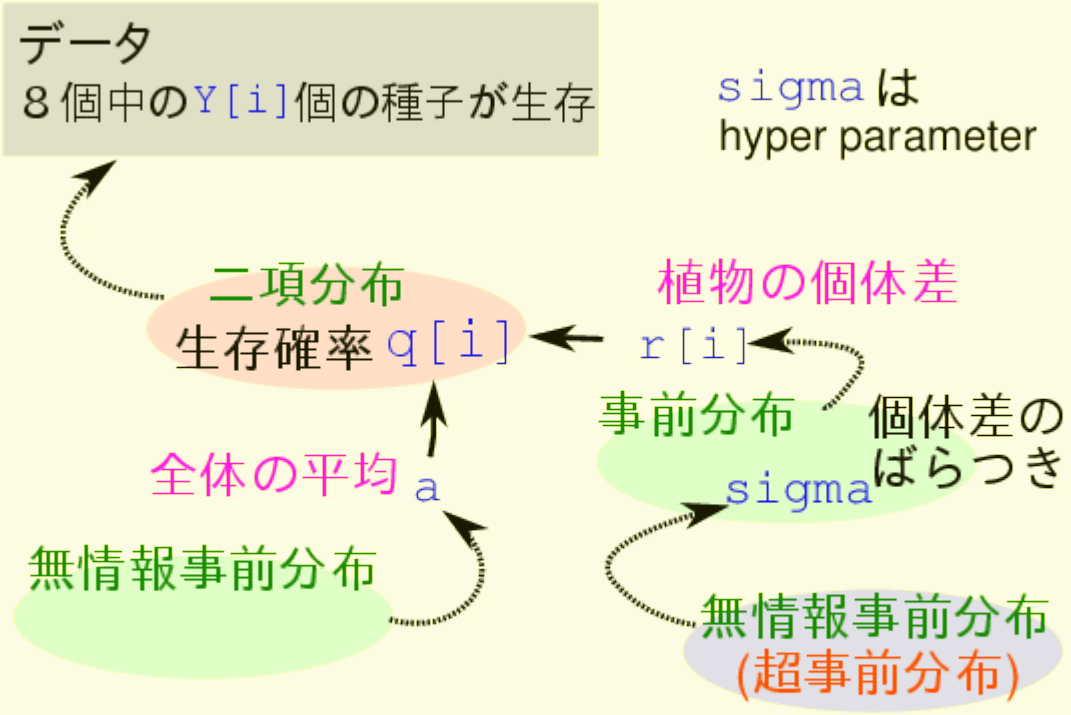
100 個体の植物の合計 800 種子中 403 個の生存が見られたので，平均生存確率は 0.50 と推定されたが……



さっきの例題と同じようなデータなのに?

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

11/2 午後 k6

2012–11–02 k6

神戸大学理学部特別講義

「生物統計学」 (2012 年 11 月) 投影資料

全部で 6 回中の 6 回目

階層ベイズモデルの応用

空間・時間構造などをあつかう

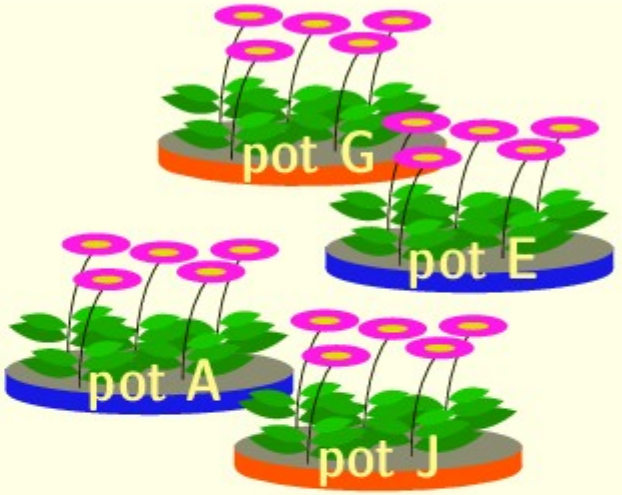
久保拓弥 kubo@ees.hokudai.ac.jp

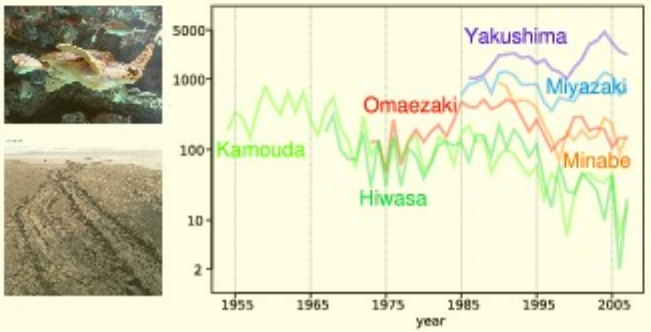
http://goo.gl/wijx2

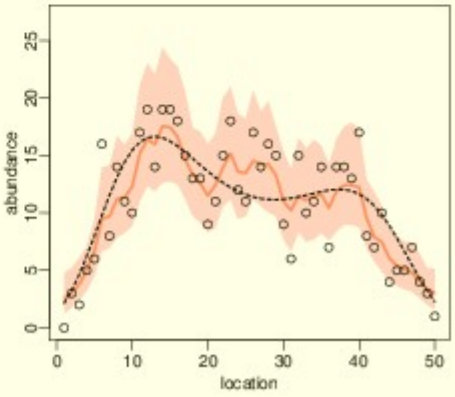
いろいろな例を紹介

今回のハナシ: いろいろな階層ベイズモデル

1. 個体差 + ブロック差というネストしたランダム効果
2. 「隣と似ている」空間相関のあるランダム効果
3. 時間変化する潜在変数: ウミガメ上陸数の統計モデル





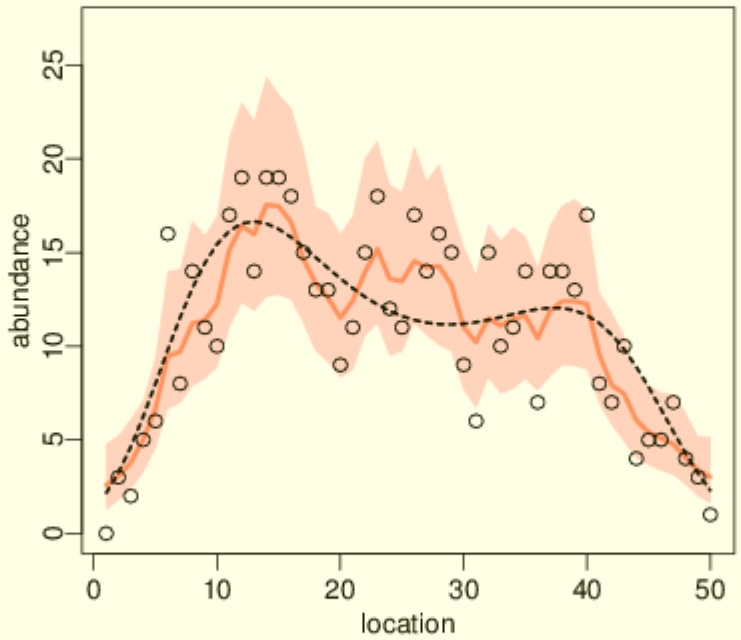


2012–11–02 k6 (2012–10–26 17:28 修正版) 2/ 56

空間構造のあるデータ

解析の目的: まずはこんな推定をしてみたい

空間相関を考慮するモデル
欠測データなし

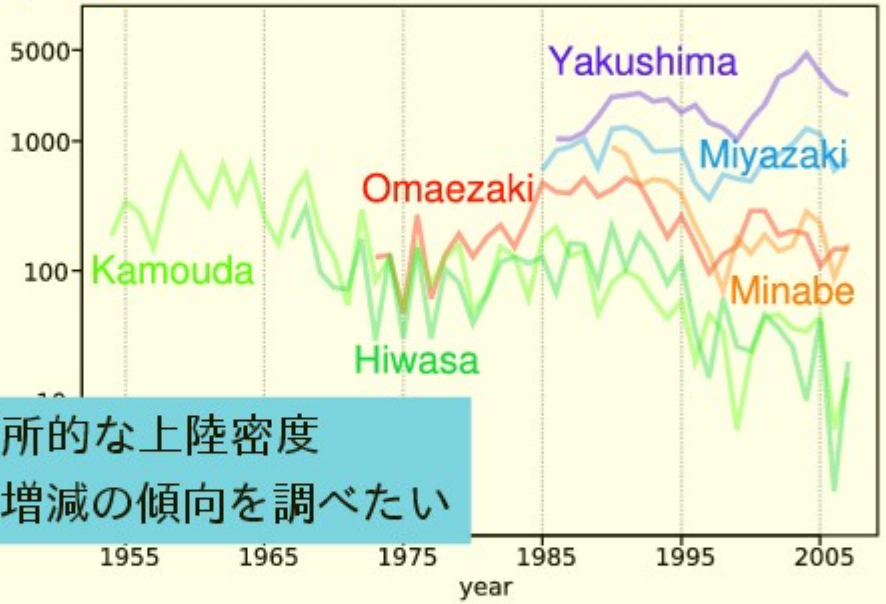


(彩色された領域は平均値の事後分布の 95% 区間，曲線は中央値)

長期ウミガメデータ

問: 産卵場所としての利用が減少している海岸は?

産卵場所として不適 → 上陸数の減少となっている?



局所的な上陸密度
の増減の傾向を調べたい

- 上陸密度 ⟺ 産卵地としての海岸の良さ?
- 上陸数変動の状態空間モデル (階層ベイズモデルの一種)

2012–11–02 k6 (2012–10–26 17:28 修正版) 41/ 56

今日・明日，いっしょに勉強しましょう

「生物統計学」の時間わり

11/1 (木) 午後

- **(k1)** 統計モデルとは何か? なぜ必要か?
- **(k2)** 確率分布と一般化線形モデル (GLM)
- **(k3)** モデル選択と統計学的検定

11/2 (金) 午前 — R の練習

- **(r1)** R をちょっと使ってみる，図を作る
- **(r2)** R で `glm()` 使って，また図を作る

11/2 (金) 午後

- **(k4)** 何でも割算するな! — GLM で解決できる
- **(k5)** 個体差をいれて階層ベイズモデルを作ろう
- **(k6)** いろいろな階層ベイズモデル