2004.11.24

生物多様性論 I: 「統計モデルの基礎」

全部で 2 回講義の 1

統計モデリングと推定を重視してみる

— 理解できる統計学めざして —

http://hosho.ees.hokudai.ac.jp/~kubo/stat/2004/

講釈: 久保拓弥　kubo@ees.hokudai.ac.jp

この 2 回だけの統計学授業でやること

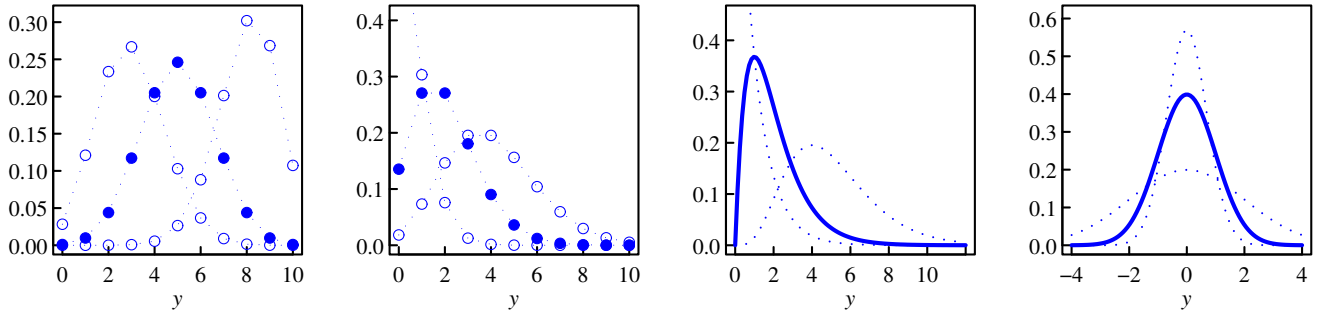
- 自然科学の データ解析 に統計学は必要不可欠
- しかし大半のユーザーは 何もわからん 状態で使ってる
- この授業の目的はその「わからん度」を少しだけ下げる こと

- **第 1 回**: 2004.11.24 (**水**)
  統計モデリングと推定を重視してみる
  — 理解できる統計学めざして
- **第 2 回**: 2004.11.29 (**月**)
  「検定」の使われかたを観察しよう
  — 「検定」ってそんなにエラいのか?

個別的なワザより全体に共通する考えかたを— ただし内容は偏ってるよ

今日のハナシ: 統計学とは何で何が重要か?

1. **とりあえずの「統計学って何 ?」**
   どういうふうに使えて，どう勉強すればいいか
2. **乱数 (標本) と推定**
   今日はこれさえわかれば OK
3. **ダメ推定と良い推定**
   架空だけど具体的な例をながめつつ



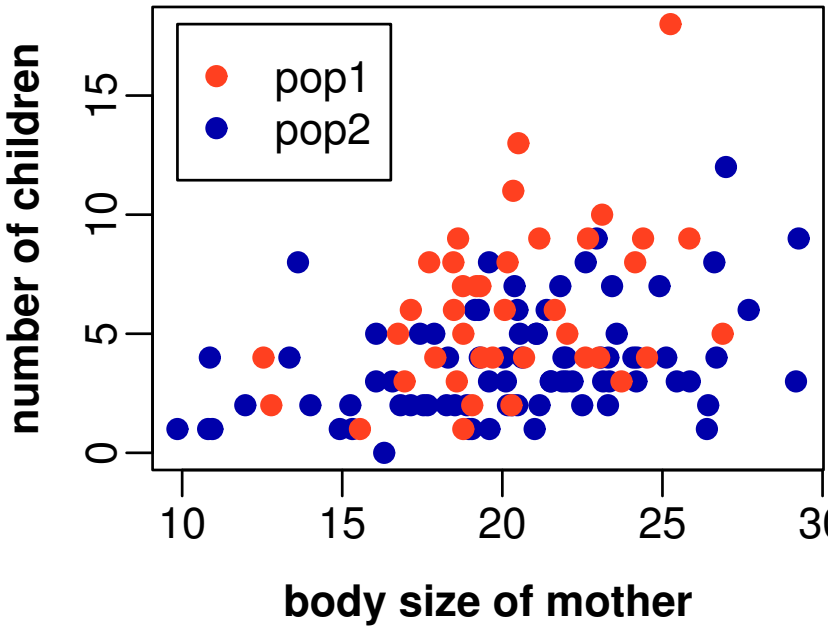
まずは簡単に

1. とりあえずの「統計学って何？」

どういうふうに使えて，どう勉強すればいいか

統計学「こんなことができたらなぁ」例 1

なにか動物の母親サイズ vs 産んだ子の数のデータ (架空) あったときに……

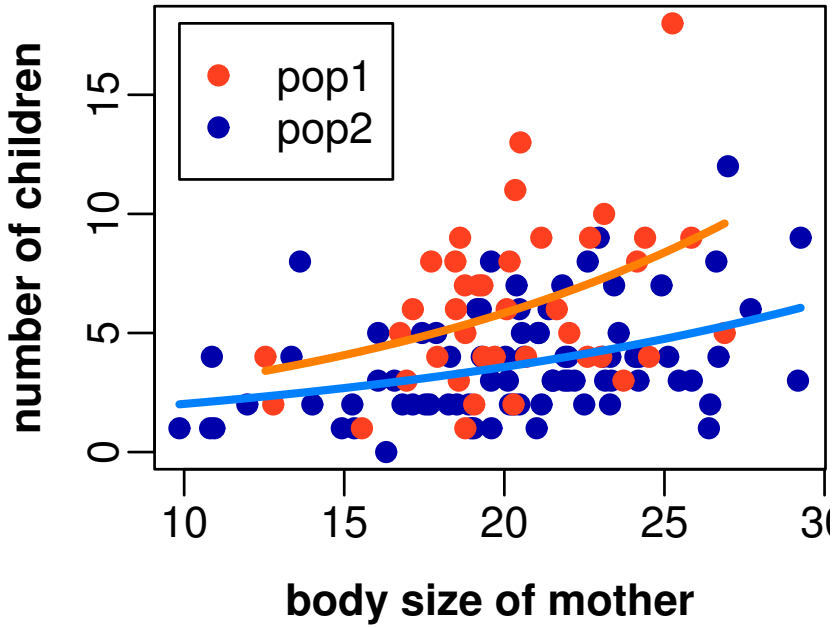


この観測データみてると疑問がいろいろと出てくる:

- 母親のサイズと産仔数はどういう関係なんだろうか?
- 集団 1 と 2 では同じように子供が産まれているんだろうか?

統計学「こんなことができたらなぁ」例 1 続

統計モデル (ポアソン分布モデル) の推定やることで……



- **(平均産仔数)** $= \exp(\alpha + \beta($**母サイズ**$))$
- **集団** 1 **と** 2 **の産仔速度は別のモデルで説明される**

…… てなことを結論できたらなぁ ……

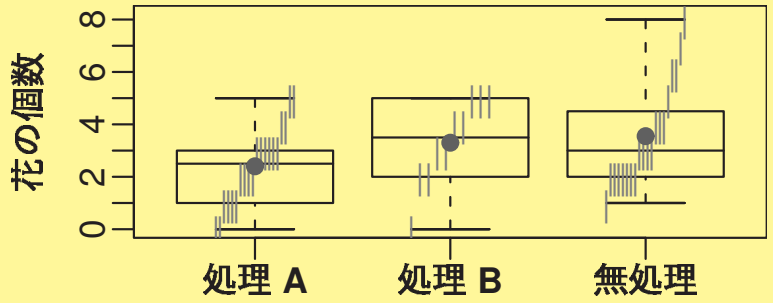
(この 2 回だけの授業ではこの問題の解きかたは説明しない)

2004.11.24

統計学「こんなことができたらなぁ」例 2

なんだかわからない植物の実験データファイル (架空)

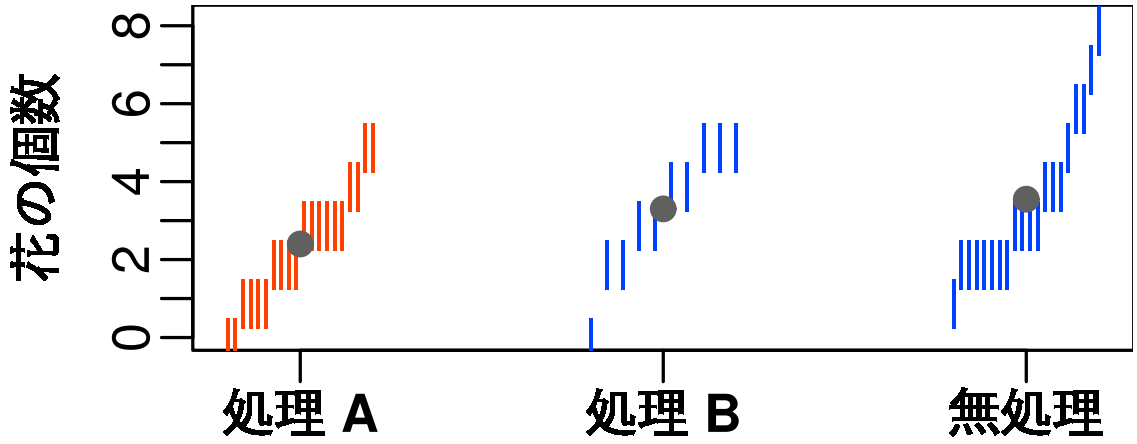
level	n.flower
A	3
A	0
A	1
…	…
B	4
B	1
B	2
…	…
C	5
C	4
C	2
…	…



もしこういうデータあったとして……

- `level`: **水準**
  - `A` — **処理** A (20 **個体**)
  - `B` — **処理** B (10 **個体**)
  - `C` — **無処理** (20 **個体**)
- `n.flower`: **花の個数**
  - `n.flower` $\in \{0, 1, 2, \cdots\}$

統計学「こんなことができたらなぁ」例 2 続



- 無処理と処理 B の間には差が無かった
- しかし処理 A は無処理・処理 B のどちらとも違っていた

…… てなことを結論できたらなぁ ……

(この 2 回だけの授業ではこの問題の解きかたは説明しない)

統計学とおつきあいするためには

あなた自身のココロがまえとして

- われわれが必要とするような統計学は**難しいものではない**，しかしながら勉強するのに**時間はかかる**かもしれない— 確率，という非日常的なヘンてこな抽象概念などとおつきあいするから
- 疑いぶかく— あなたも私もいつでもどこかで何かを誤解している— そして世の中は統計学の**誤用だらけ**

とりあえず

- よい教科書が必要 (インターネット上にもいろいろある)
- よい統計ソフトウェアが必要 (R…… 後述)
- できれば相談できる相手も

必読! 粕谷英一「統計のはなし」

生物学を学ぶ人のための統計のはなし

— きみにも出せる有意差 —

- **著者**: **粕谷英一** (**九州大・理・生物**)
- **出版**: **文一総合出版**
- **サイズ**: A5 **判** / 199 **ページ**
- ISBN: 4-8299-2123-4
- **発行年月**: 1998.3

「この【ピンク本】を読まずにすますことはできない」
(**三中信宏・書評** on "BK1" http://www.bk1.co.jp/)

これ使いましょう: 統計ソフトウェア R

http://www.r-project.org/

- いろいろな OS で使える freeware
- 使いたい機能が充実している
- 作図機能も強力
- S 言語によるプログラミング可能
- よい教科書が出版されつつある
  - 「R による統計解析の基礎」 中澤港 (2003)
  - "Introductory Statistics with R" P. Dalgaard (2002)
  - "Computational Statistics" M. Crawley (2002)
  - ネット上のあちこち

R が変えつつある生態学のデータ解析

- 使いたい手法はたいていそろってる
- 無ければ自分で何でも簡単に作れる
- 統計学的 simulation も簡単にできる

…… となると ……

- データを無理やりある手法にこじつける，ということが不要になる— データの構造にあわせた統計モデリングを行えばよい
- 手法の前提となる統計学の基本(統計モデル) の理解がむしろ重要
- 単純な検定ではなく,「こういう標本のばらつきを生成したメカニズム」の推定のよしあしが問われる

ここが核心部

2. 乱数 (標本) と推定

今日はこれさえわかれば OK

乱数とは何か?

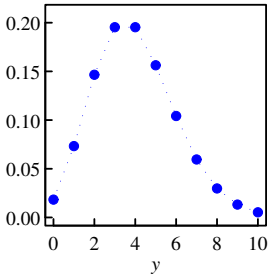
統計学の中核概念

ある 確率分布 (母集団・モデル) から
無作為に得られた値 (標本・データ)

ポアソン分布

R の関数:

dpois(y, $\lambda = 3$)



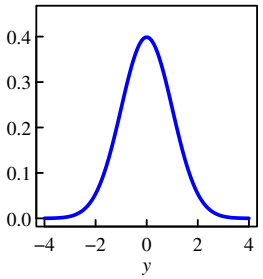
→

```
> rpois(10, lambda = 3)
5 4 3 2 4 2 4 1 7 1
```

正規分布

R の関数:

dnorm(y, $\mu = 0$, $\sigma = 1$)



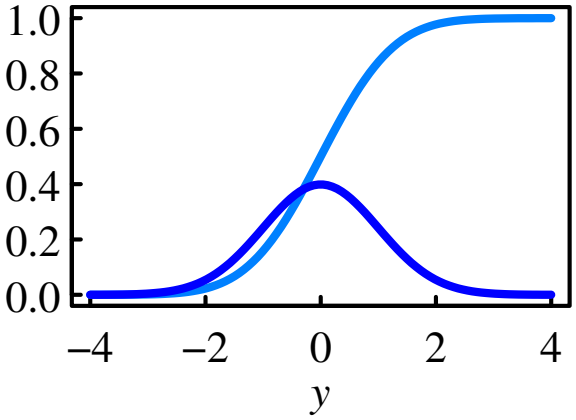
→

```
> rnorm(9, mean = 0, sd = 1)
 1.4851004 -0.9912880 -0.1092131
-2.1752314 -0.3779424  1.1360432
 1.2493592 -1.2405408 -0.4425550
```

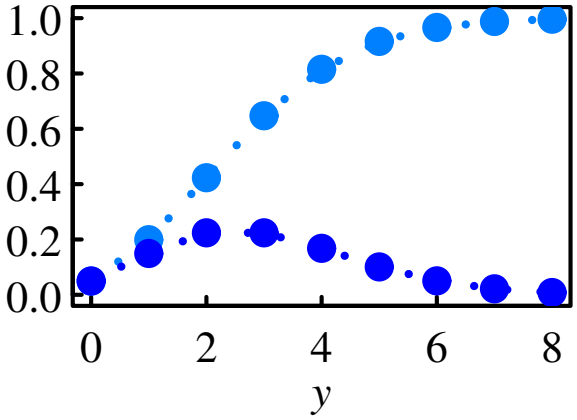
確率分布: 確率分布 (関数) と確率密度分布 (関数)

確率分布関数 $F(y)$ と確率分布密度関数 $f(y)$ の関係

連続関数の例: 正規分布



離散関数の例: ポアソン分布



カタチを決めるパラメーター

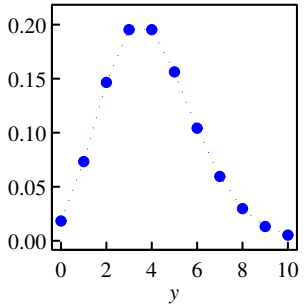
平均: 重心 $m = \int_{-\infty}^{\infty} y \, \mathrm{d}f(y) = \sum_{0}^{\infty} y f(y)$

分散: ばらつき $\mathrm{Var} = \int_{-\infty}^{\infty} (y - m)^2 \, \mathrm{d}f(y) = \sum_{0} (y - m)^2 f(y)$

じゃあ推定ってのは何なの?

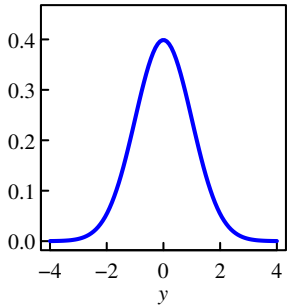
ポアソン分布の推定

5 4 3 2 4 2 4 1 7 1 $\longrightarrow$



正規分布の推定

1.4851004 -0.9912880 -0.1092131 $\longrightarrow$
-2.1752314 -0.3779424 1.1360432
1.2493592 -1.2405408 -0.4425550

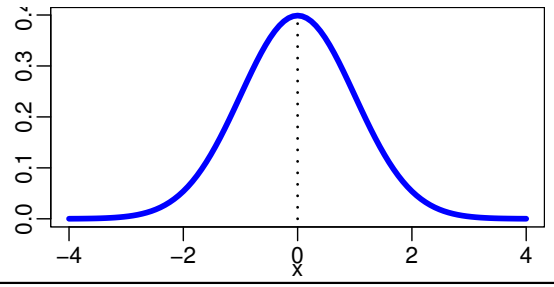
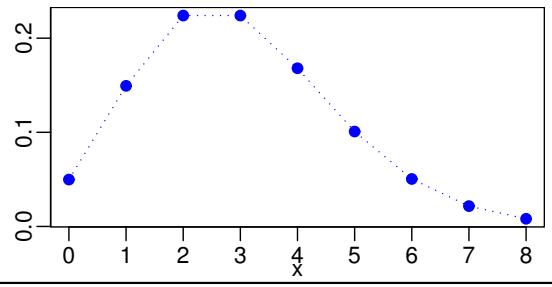


乱数とみなされる標本集団
$\longrightarrow$ 母集団すなわち確率分布
のパラメーターを決めてやる技法

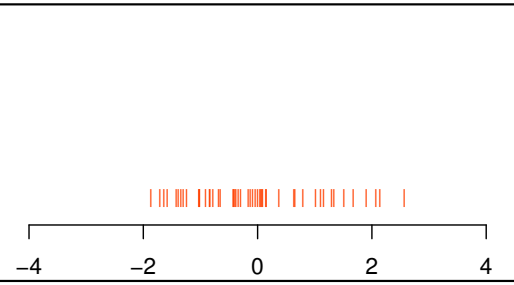
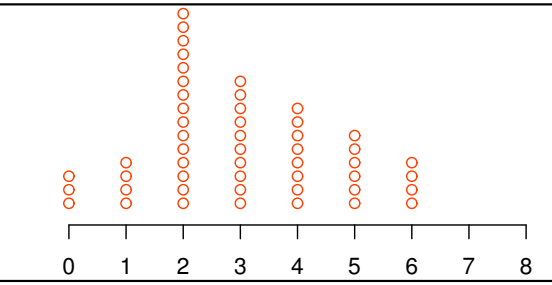
(分布がわかってないと推定できない)

統計学とは結局これ: 確率分布，乱数と推定

今日はこの関係さえ理解してもらえればそれで OK!

(よびかた)	[連続確率密度分布]	[離散確率密度分布]
• モデル • 確率分布 • 母集団		

サンプリング ⇓ ⇑ **(パラメーター) 推定**

• データ • 乱数 • 標本集団		

推定については後で例を示す

統計学勉強における乱数利用のススメ

(;_;) **統計学がわからない**

→ **ひたすら考える・わからぬまま使う**

(^_^) **統計学がわからない**

→ **とりあえず「実験」してみる**

(^o^) **その「実験」結果を考える・利用する**

乱数を手軽に生成できる
R は画期的なソフトウェア

R で乱数生成 $\longrightarrow$ 推定のやり方を「実験」

線形モデル: $\mu(x) = \beta_0 + \beta_1 x$

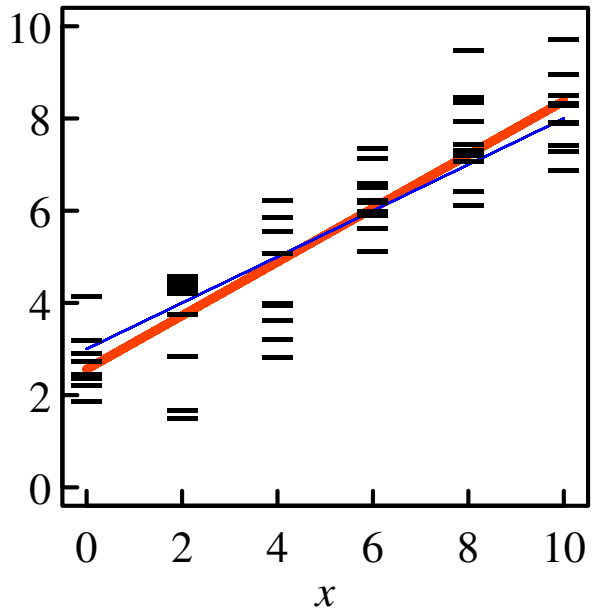
- **まず平均値 $\mu(x)$ の** vector **をつくる**

```
beta0 <- 3
beta1 <- 0.5
x <- seq(x.min, x.max, by = 2.0)
x <- rep(x, 10) # 各 x に 10 個ずつ
mu <- beta0 + beta1 * x
```

- **正規乱数を** rnorm **で**

```
sample <- rnorm(length(mu),
                mean = mu, sd = 1)
```

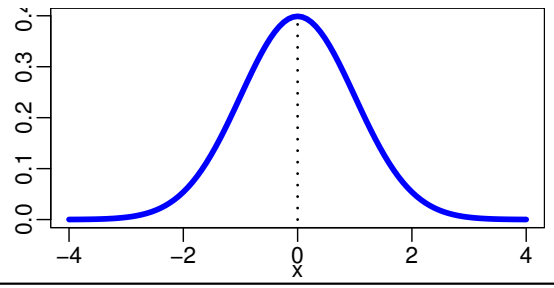
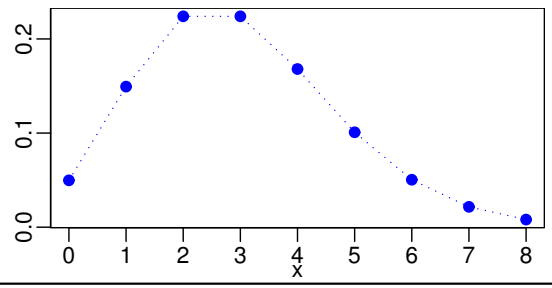
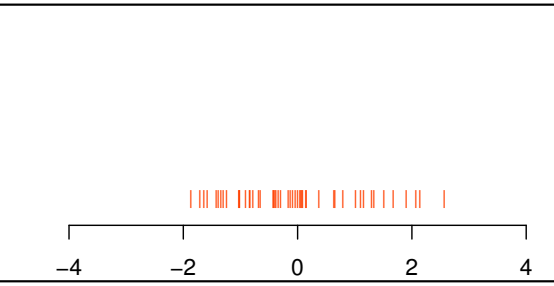
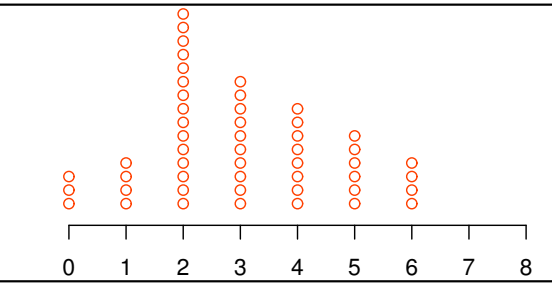
青: ホントの $\mu(x)$, 赤: 推定された $\hat{\mu}(x)$



- R **の** glm **関数による推定** (**この場合は** glm **でなくて** lm() **でいいんだが**)

```
result <- glm(sample ~ 1 + x, family = gaussian)
```

ここまでで言いたいこと: 乱数と自然現象

(よびかた)	[連続確率密度分布]	[離散確率密度分布]
• モデル • 確率分布 • 母集団		
	サンプリング ⇓ ⇑ (パラメーター) 推定	
• データ • 乱数 • 標本集団		

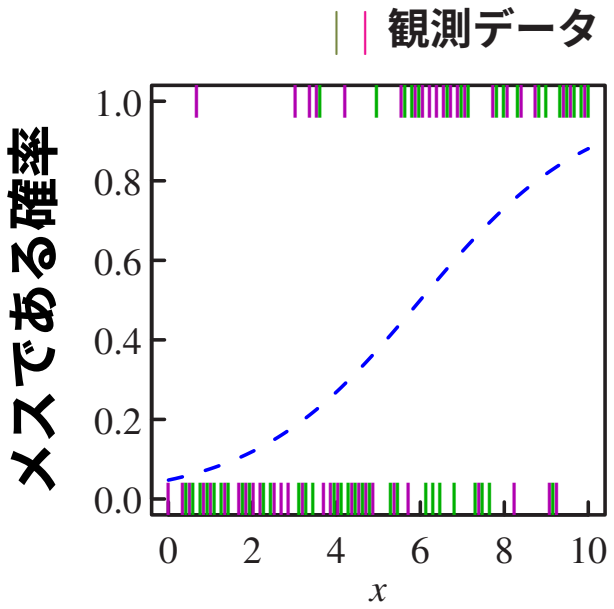
- 自然科学者は何か ばらつきのある自然現象をみたときにそれが確率論的モデルによって生成された，と仮定する → 現象を簡単な数式で書ける
- このばらつきのあるデータから確率論的モデルのカタチを特定してやることがパラメーター推定である → モデル選択や検定につながっていく (次回)

生態学方面でよく見かける……

3. ダメ推定と良い推定

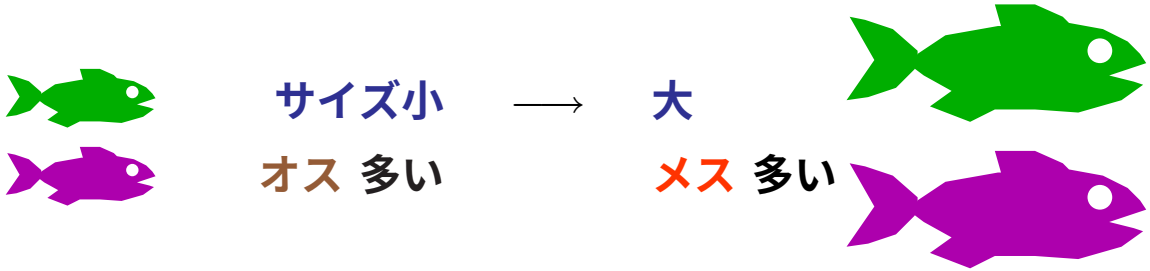
架空だけど具体的な例をながめつつ

架空生物の観測データ: 性転換するサカナ



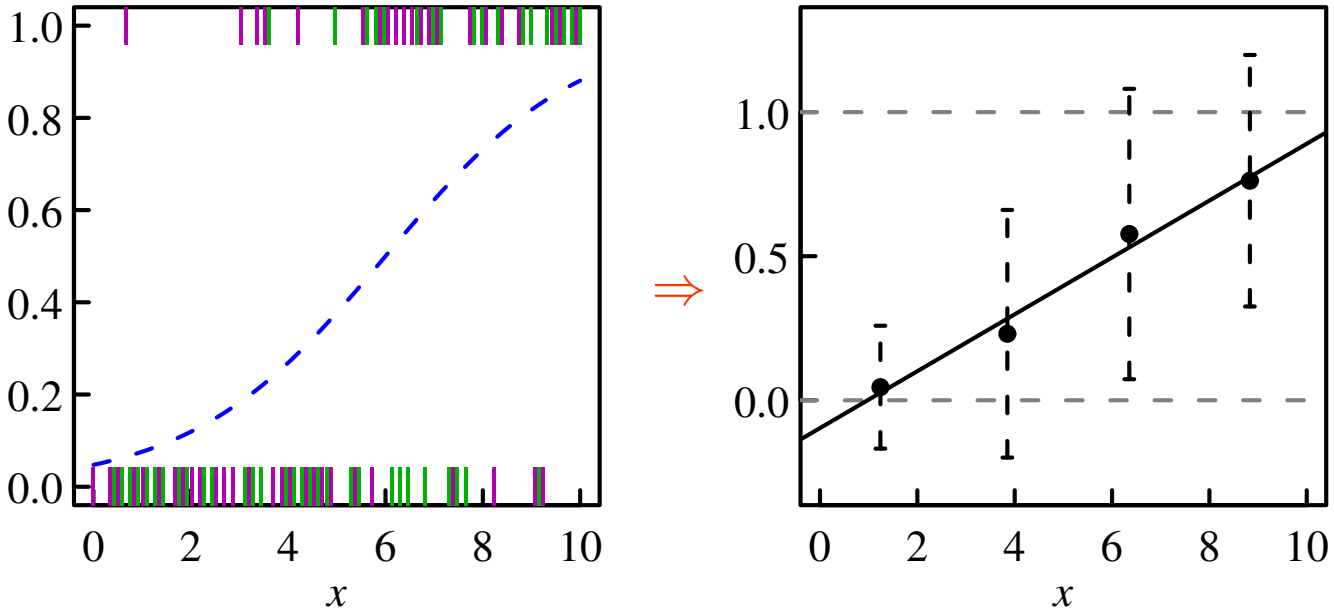
[“神” の立場で知ってるコト]

- メスである確率は 青破線で示されているように上昇する
- この確率は，魚の色 (緑 と 紫) とは 無関係
- サンプリングに少し偏りあり



青破線(つまり真のモデル・母集団) を推定したい

(よく見かける) ダメ推定の一例



1. てきとーにサイズ (x) の「区画」を取る (上の例だと 4 区画)
2. 区画ごとに縦横の平均値など計算; $\{0, 1\}$ データを割算値に
3. 統計ソフトウェアにほうりこむ
   (直線回帰する or 「分散分析」する or 「検定」& 多重比較する)

2004.11.24

なぜよろしくないか? データの特徴を無視

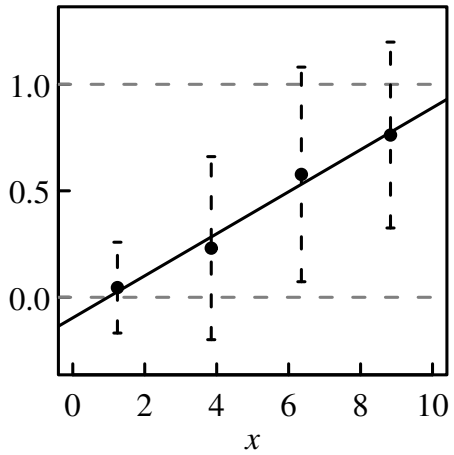
区画はてきとー

区画のとりかたで結果は変わる

割算すると情報が失われる

1 / 2 と 100 / 200 は違う!

— 十円玉なげの例で考えてみよ



等分散でもなければ正規分布でもない

ということで直線回帰も分散分析も使えん— さらに，いわば母分散が異なる状況なので，ノンパラメトリック検定のたぐいもだめ

何を推定してるのだろうか?

メスになる確率がマイナスになったり，1 をこえたりするモデルってのは……? (変数変換すればいいって? そのワザは呪われてる)

確率分布を推定する方法たちの階層性

「なんでもかんでも正規分布」という 思いこみを克服するための地図

[最尤推定法で扱えるモデル]
何でもいいから確率分布があるモデル
一般化線形混合モデルなどなど

[一般化線形モデル (GLM)]
指数関数族の確率分布 + 線形モデル
ロジスティック回帰，ポアソン回帰などなど

[最小二乗法的に考えるモデル]
等分散正規分布 + 線形モデル
直線回帰，いわゆる「分散分析」などなど

「いろいろな方法あるじゃないか!」

とりあえず，の一般化線形モデル (R の glm() 関数)

いろいろな確率分布の推定ができてしまう，たいへん「使い勝手のよい」推定計算手法

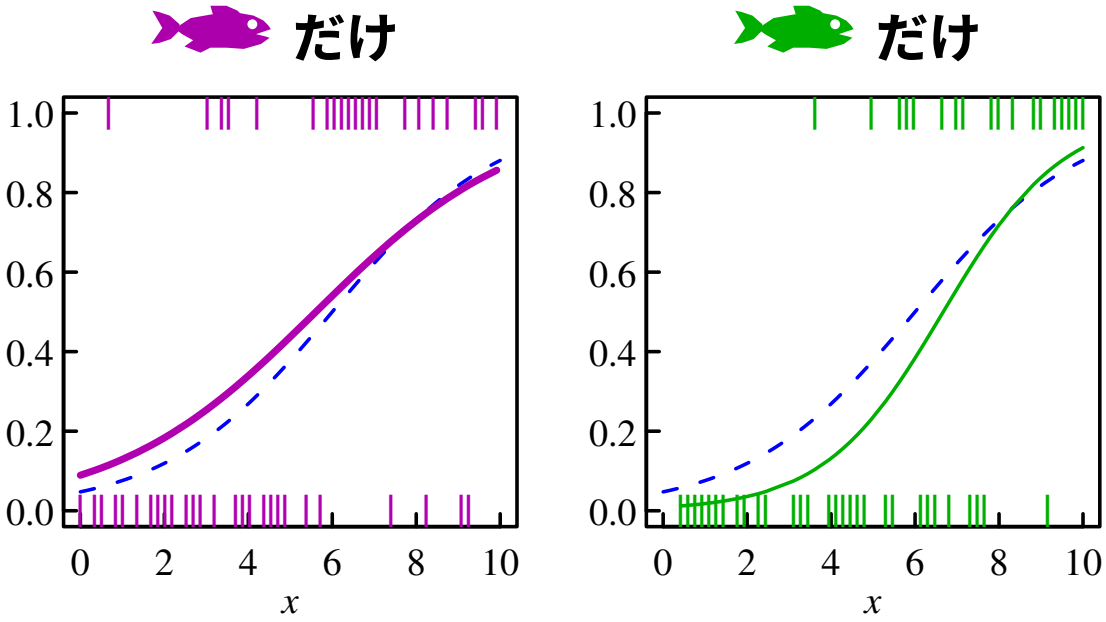
	確率分布	乱数生成	パラメーター推定
(離散)	ベルヌーイ分布	rbinom()	glm(family = binomial)
	二項分布	rbinom()	glm(family = binomial)
	ポアソン分布	rpois()	glm(family = poisson)
	負の二項分布	rnbinom()	glm.nb()
(連続)	ガンマ分布	rgamma()	glm(family = gamma)
	正規分布	rnorm()	glm(family = gaussian)

- glm で使える確率分布は上記以外もある
- glm.nb() は MASS library 中，またここには rnegbin() なども含まれる

R の glm で推定: ロジスティック回帰の例

ロジスティックモデルは一般化線形モデルの一部であり，R の glm でパラメーター推定 (つまりロジスティック回帰) できる

ここでは推定結果を紫・緑の曲線で示している

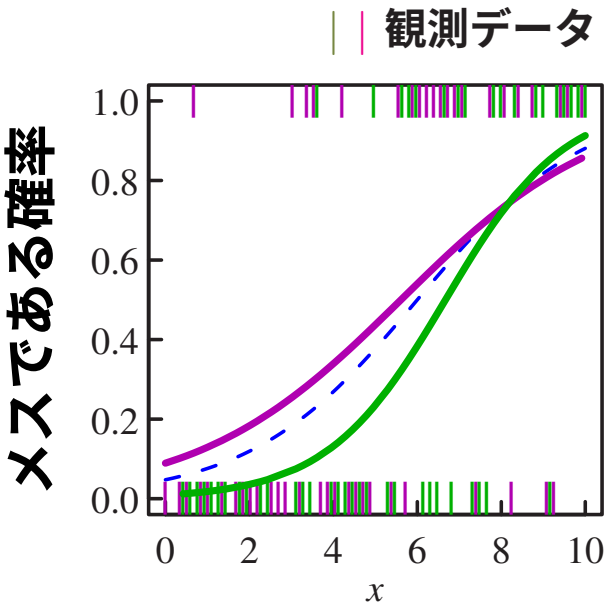


…… 魚の種類によってモデルが違うときは? ……

(この 2 回だけの授業ではこの問題の解きかたは説明しない)

良い推定をめざして

統計学的なデータ解析のコツ



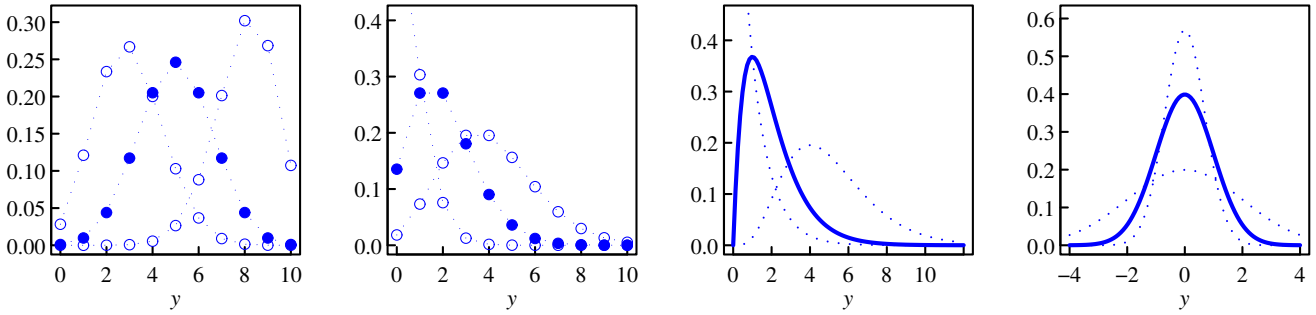
- むやみに **区画わけしない**！
- むやみに **割り算しない**！
- たくさん **図を描く**
- 「観測データを説明する**確率分布は何か**?」を考える (初心者は `glm` との対応を検討する)

自分で作ったウソ (人造ニセ相関など) に自分自身が騙されないために……

2004.11.24

今日のまとめ: 「わかる」データ解析のために

1. **「統計学って何 ?」を理解する**
   - 数理モデルを使ってると意識する
2. **乱数 (標本) と推定の関係をおさえる**
   - ばらつきを見る → 確率分布
3. **ダメ推定を避ける**
   - 割り算値解析しない，図を描く，データにあわせる



次回予告

2004.11.29

生物多様性論 I: 「統計モデルの基礎」

全部で 2 回講義の 2

「検定」の使われかたを観察しよう

— 「検定」はそんなにエラいのか? —

“「ゆーい差」至上主義の世界を見物”

http://hosho.ees.hokudai.ac.jp/~kubo/stat/2004/