

植物生態学特論 II 代講

全部で 2 回講義の 1

# 生態学研究で得られたデータを 解析するための統計モデリング

— 理解できる統計学めざして —

<http://hosho.ees.hokudai.ac.jp/~kubo/stat/2005/>

講釈: 久保拓弥 [kubo@ees.hokudai.ac.jp](mailto: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: 2005.07.04 (月)  
生態学研究で得られたデータを解析するための統計モデリング  
— 理解できる統計学めざして
  - 第 2 回: 2005.07.06 (水)  
「検定」の使われかたを観察してみる  
— 「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?」

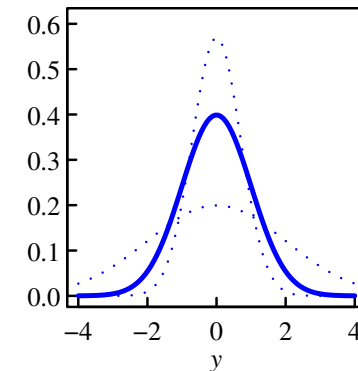
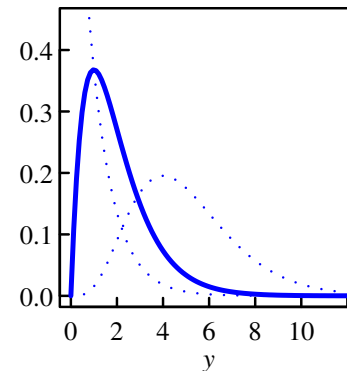
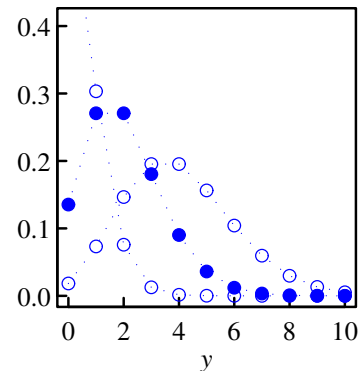
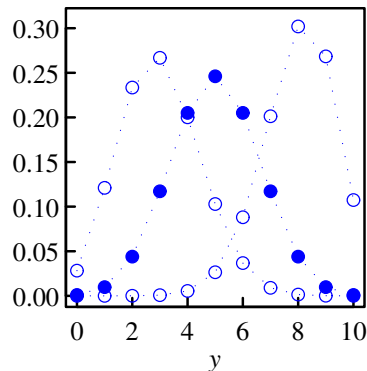
どういうふうに使えて, どう勉強すればいいか

## 2. 乱数 (標本) と推定

今日はこれさえわかれば OK

## 3. ダメ解析と良い解析

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

# 自然科学研究における二段階の情報損失

## 第一段階: 自然現象 → 数値データ

- 観察・実験による情報損失
- 人間が自然現象からとりだせる数値データはごくわずか
- (とくに野外調査では) 厳密に「同じ」データを再びとれない

## 第二段階: 数値データ → 解析結果

- 統計解析による情報損失
- 人間のアタマは大量の数値データも把握できない
- この情報損失過程には再現性がある(「客観的」に検討できる)

ここでは第二段階での改善について考える

## 「数値データ → 解析結果」過程の現状と理想

### 生態学研究まわりにおける現状

- 軽視されている (授業でも適切な方法を教えない)
- そもそも何やってるかわかってないヒトたちが多い
- まちがっている方法に固執する (指摘すると逆ぎれ)

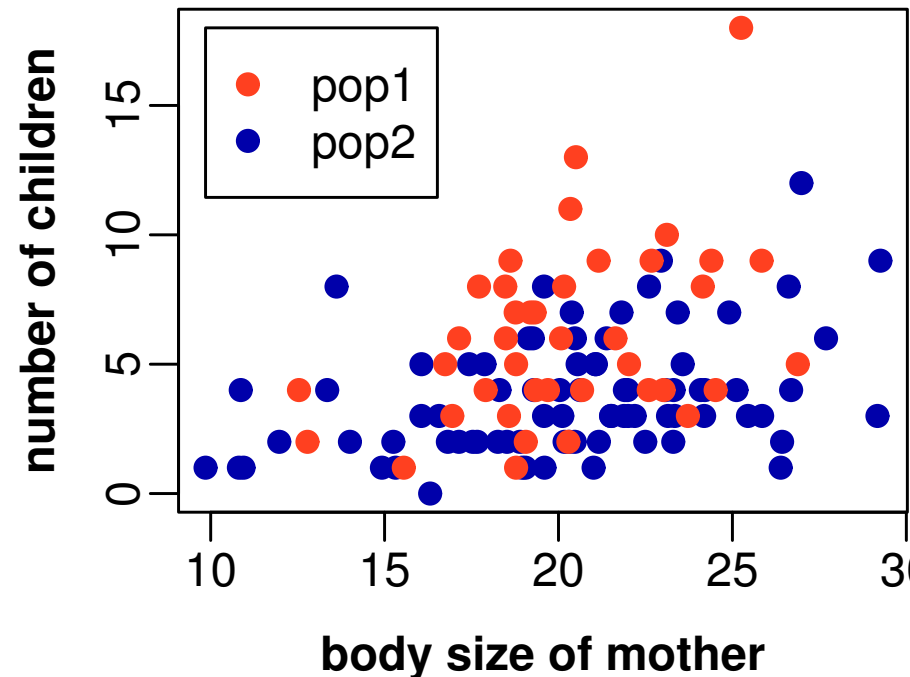
### 理想 — 情報をうまく圧縮する (ムダなく・わかりやすく)

- スジのとあった合理的な統計解析をやりたい
- データの性質・構造によくあった手法 (データの有効利用)
- 自然現象うまく説明できるモデリングになってれば

問: 統計学でどういうことができたらいいか?

# 統計学「こんなことができたらなあ」例 1

なにか植物の地上部 vs 種子数のデータ (架空) あったときに.....

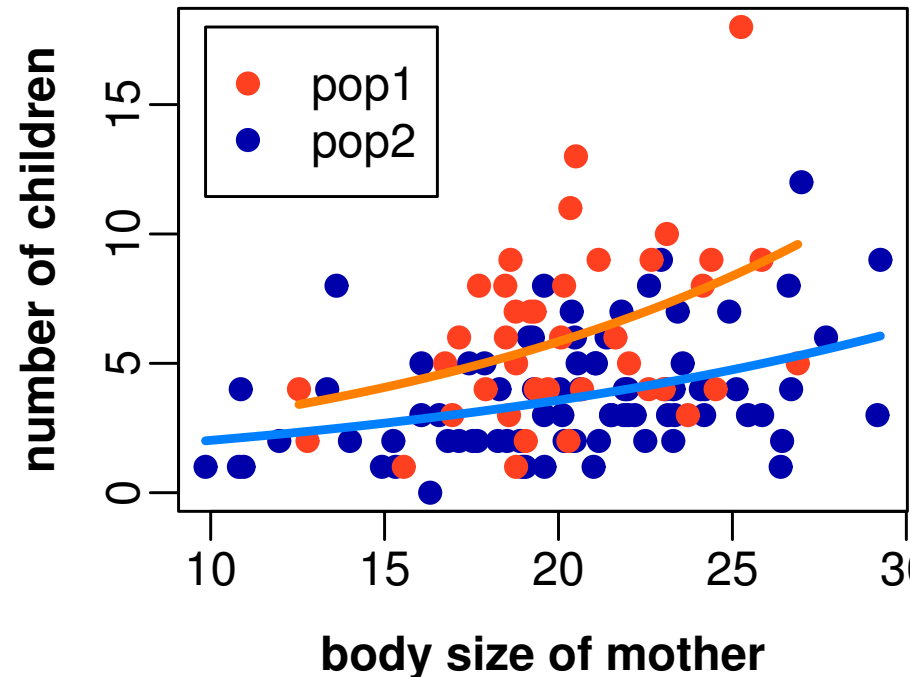


この観測データみていると疑問がいろいろと出てくる:

- 地上部サイズと種子数はどういう関係なんだろうか?
- 集団 1 と 2 で種子生産量のサイズ依存性は同じか?

# 統計学「こんなことができたらなあ」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.....



- (平均種子数) =  $\exp(\alpha + \beta (\text{母サイズ}))$
- 集団 1 と 2 の種子生産は別のモデルで説明される

..... てなことを結論できたらなあ .....

(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「こんなことができたならなあ」例 2

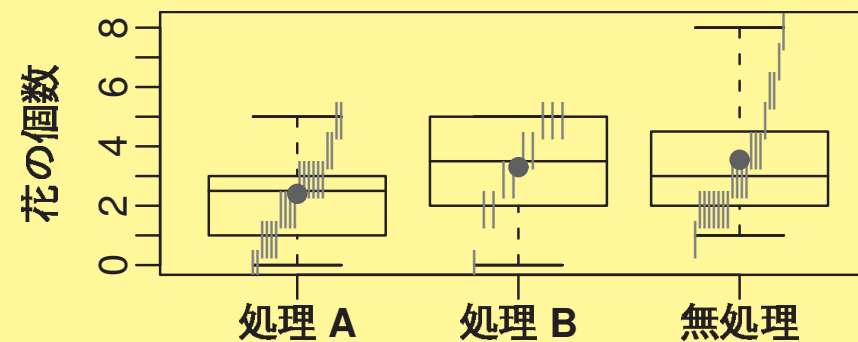
架空の実験: 何か処理をして花の個数の変化をみる

(架空実験のデータシート)

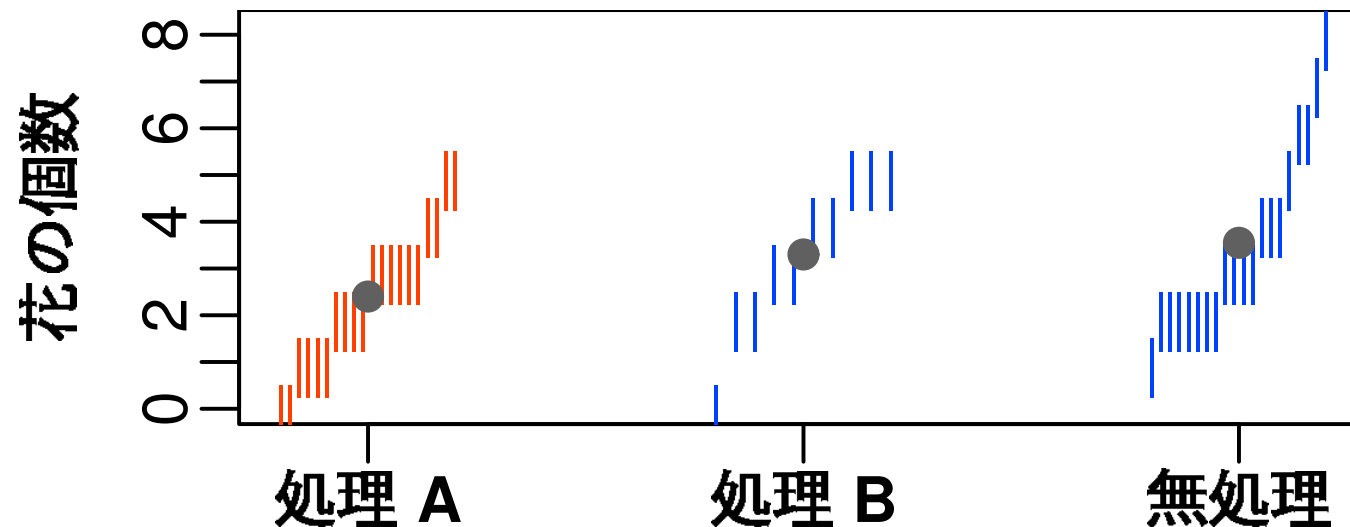
level	n.flower
A	3
A	0
...	...
B	4
B	1
...	...
C	5
C	4
...	...

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
  - $n.flower \in \{0, 1, 2, \dots\}$



## 統計学「こんなことができたならなあ」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った

..... てなことを結論できたらなあ .....

(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「ちょっとヘン」なので勉強に**時間かかる**かも
- 疑いぶかく— あなたも私もいつでもどこかで何かを誤解
- そもそも世の中は統計学の**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
  - 「The R-Tips」 舟尾暢男 (2005)
  - “Statistics: An Introduction Using R” M. Crawley (2005)
  - “Introductory Statistics with R” P. Dalgaard (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?

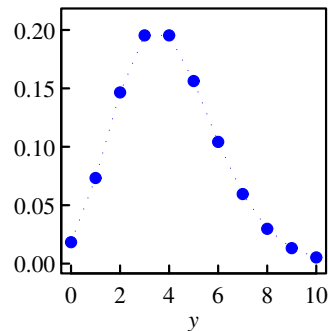
## 統計学の中核概念

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)`



→

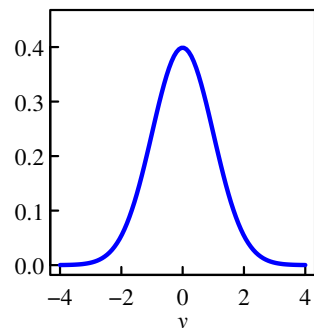
```
> rpois(10, lambda = 3)
```

```
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)
```

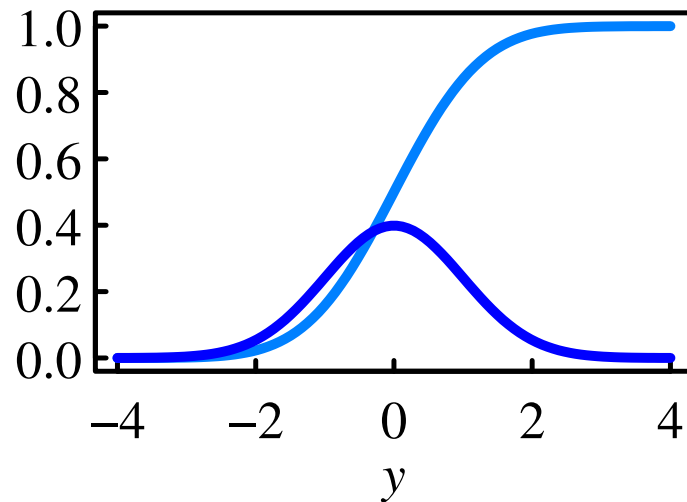
```
1.4851004 -0.9912880 -0.1092131  
-2.1752314 -0.3779424 1.1360432  
1.2493592 -1.2405408 -0.4425550
```



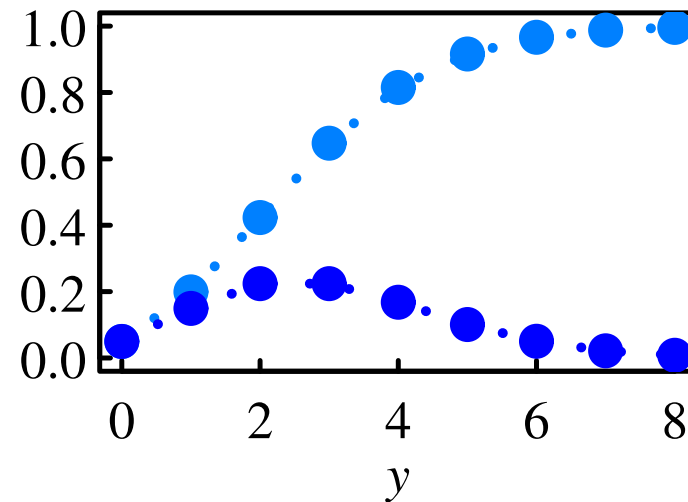
# 確率分布: 確率分布 (関数) と確率密度分布 (関数)

## 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  $m = \int_{-\infty}^{\infty} y \, df(y)$

$$m = \sum_0^{\infty} y f(y)$$

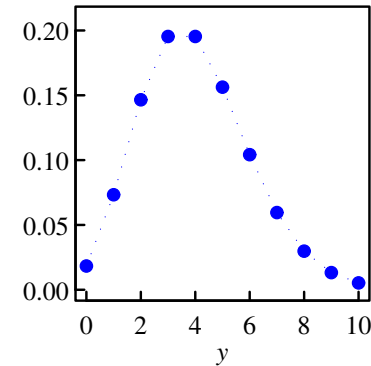
分散: ばらつき  $\text{Var} = \int_{-\infty}^{\infty} (y - m)^2 \, df(y)$

$$\text{Var} = \sum_0^{\infty} (y - m)^2 f(y)$$

# じゃあ推定ってのは何なの? → 乱数生成の逆

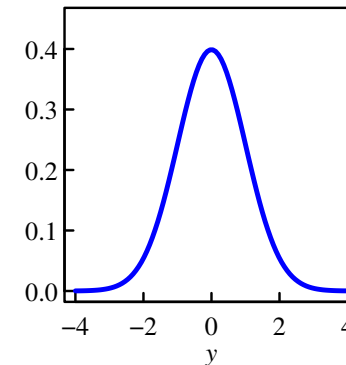
## ポアソン分布の推定

5 4 3 2 4 2 4 1 7 1



## 正規分布の推定

1.4851004 -0.9912880 -0.1092131 →  
-2.1752314 -0.3779424 1.1360432  
1.2493592 -1.2405408 -0.4425550



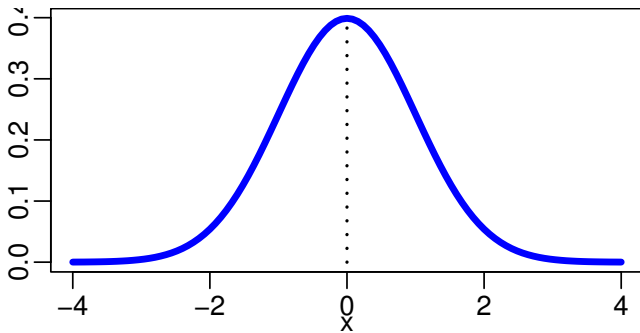
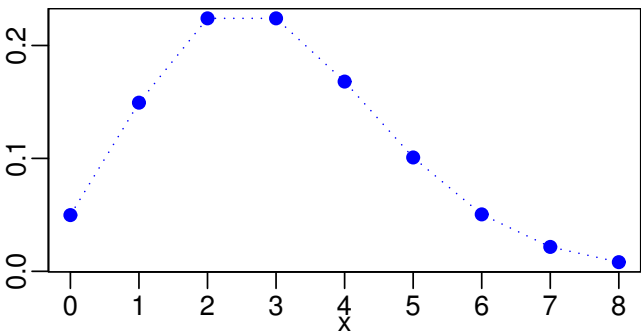
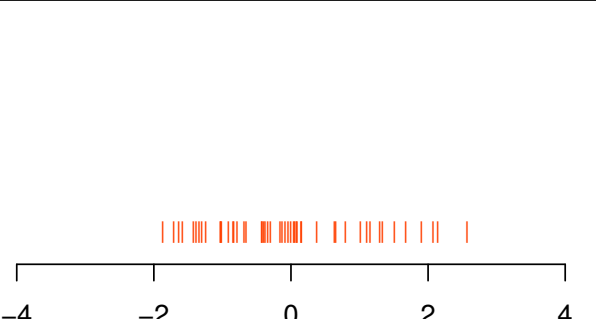
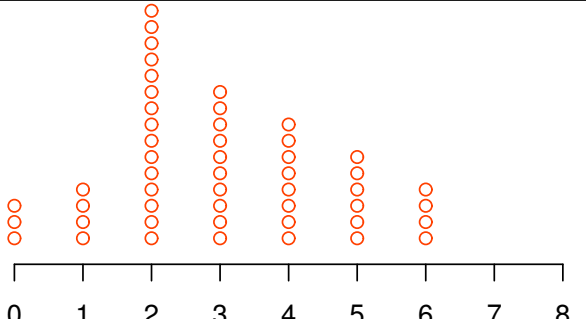
**乱数**とみなされる標本集団

→ 母集団すなわち確率分布を決め

そのパラメーターを決めてやる技法

# 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]
<ul style="list-style-type: none"> <li>● モデル</li> <li>● 確率分布</li> <li>● 母集団</li> </ul>		
<p>サンプリング ↓ ↑ (パラメーター) 推定</p>		
<ul style="list-style-type: none"> <li>● データ</li> <li>● 乱数</li> <li>● 標本集団</li> </ul>		

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 → 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
```

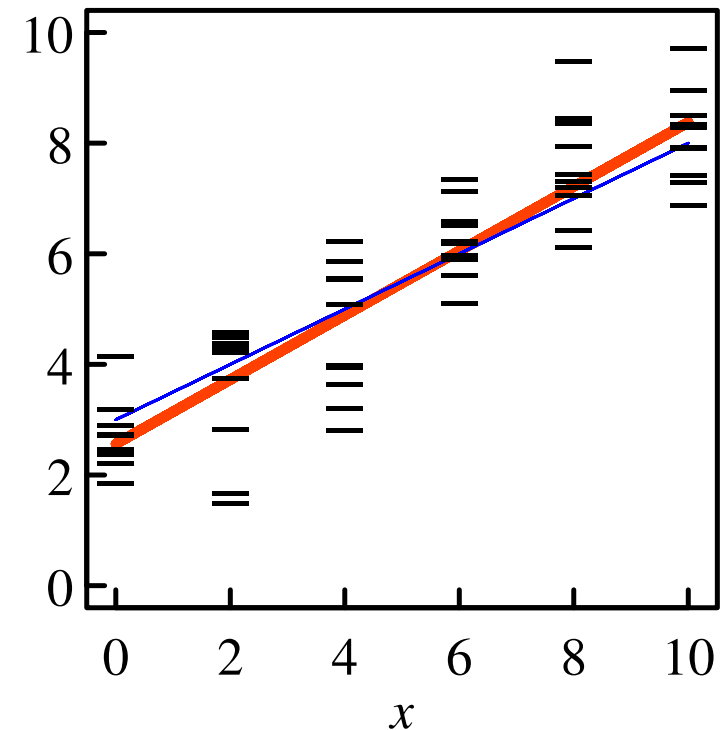
- 正規乱数を rnorm で

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

- 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)

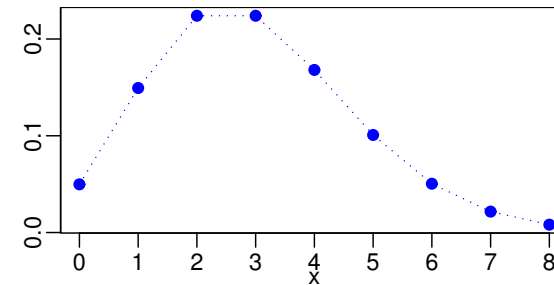
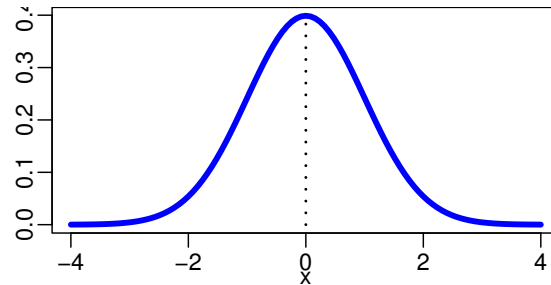
```
result <- glm(sample ~ 1 + x, family = gaussian)
```

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



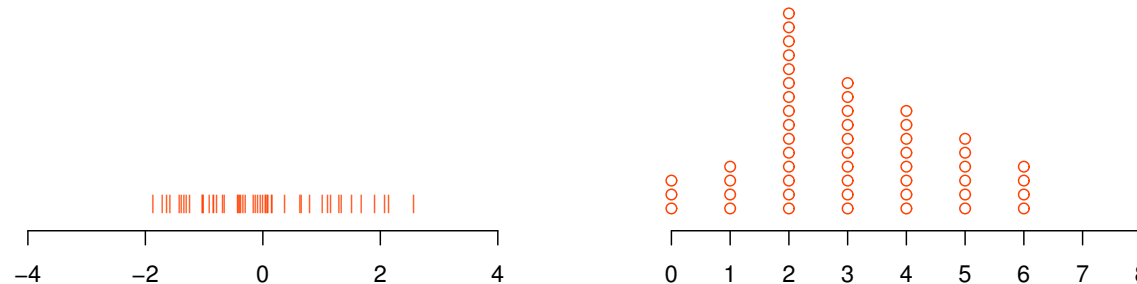
# ここまでで言いたいこと: 乱数と自然現象

- 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 → モデルによる**単純化**
- 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**や検定につながる

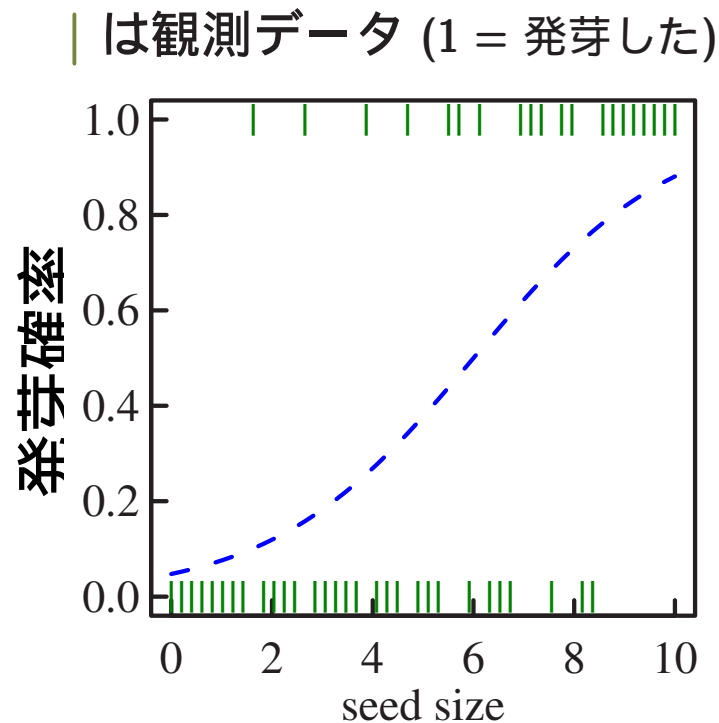
生態学方面でよく見かける.....

### 3. ダメ解析と良い解析

架空だけど具体的な例をながめつつ

# 架空植物の発芽実験データ: 種子サイズと発芽確率

種子サイズと発芽確率の関係をしらべる実験やってみた



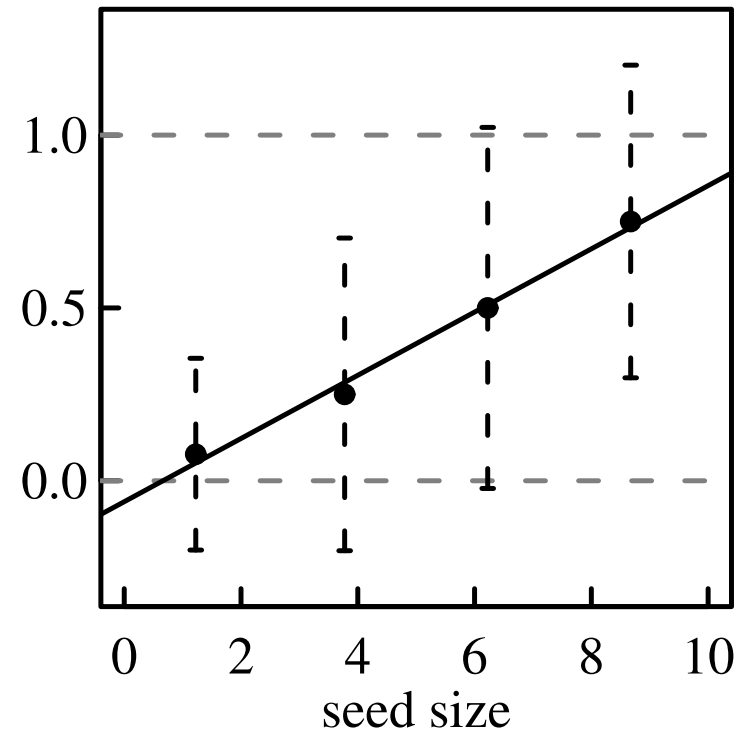
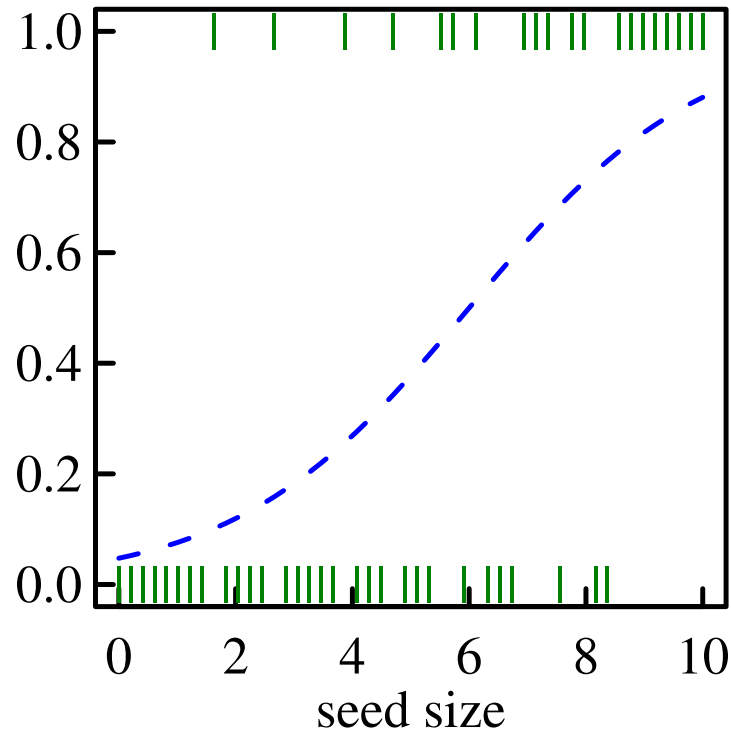
[“神” の立場で知ってるコト]

- 種子が大きいほど発芽確率が高い
- 発芽確率は青破線で示されているように上昇する

データから青破線(つまり真のモデル・母集団) を推定したい



## (よく見かける) ダメ解析の一例



1. てきとーに種子サイズの区画を取る (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. 何も考えずに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)

# なぜよろしくないか? データの特徴を無視

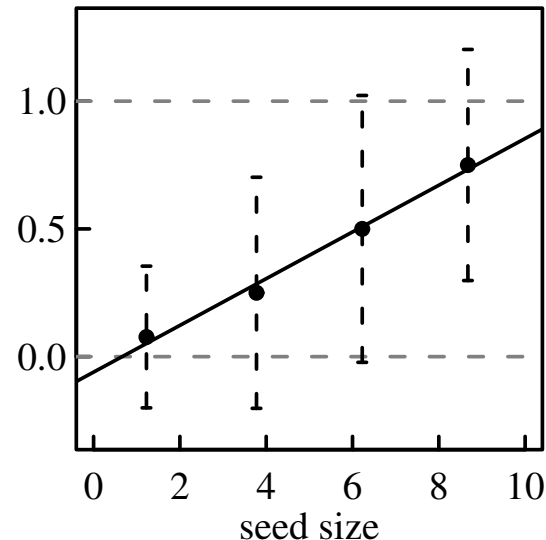
## 区画はてきとー

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?

発芽す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]

確率分布で表現できるモデルたち

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 (GLM; R の `glm()` 関数)

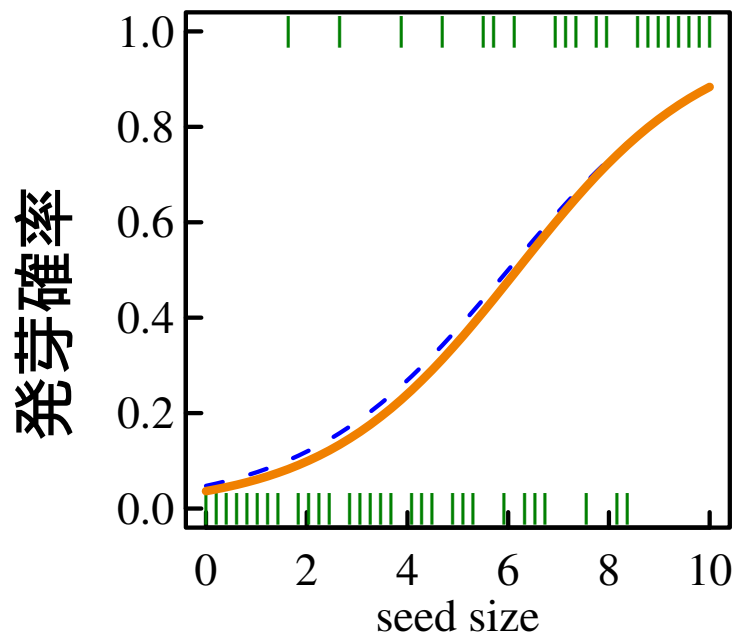
## いろいろな確率分布に適用できる推定計算手法

	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	<code>rbinom()</code>	<code>glm(family = binomial)</code>
	二項分布	<code>rbinom()</code>	<code>glm(family = binomial)</code>
	ポアソン分布	<code>rpois()</code>	<code>glm(family = poisson)</code>
	負の二項分布	<code>rnbinom()</code>	<code>glm.nb()</code>
(連続)	ガンマ分布	<code>rgamma()</code>	<code>glm(family = gamma)</code>
	正規分布	<code>rnorm()</code>	<code>glm(family = gaussian)</code>

- `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

発芽する・しないが二項分布にしたがうと仮定している



- 各種子について, そのサイズ ( $x$ ) と “発芽した or しなかった” の対応をみる
- 発芽確率  $p$  を以下のように仮定

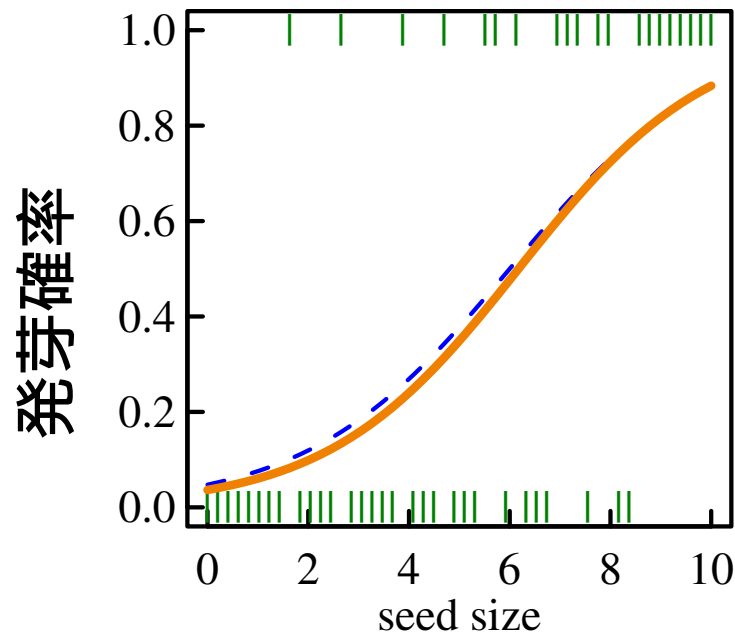
$$p = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}$$

(logistic 式)

- パラメーター  $\beta_0$  と  $\beta_1$  の推定値を最尤推定法で計算する
- ここでは R の glm() 関数を使った (上の図の赤線が推定結果)

# 良い推定 (データ → モデル) をめざして

でたらめな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 (初心者は `glm()` との対応を検討する)

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!

# 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

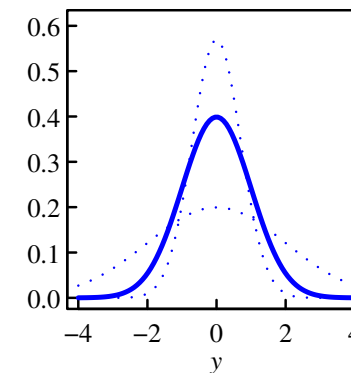
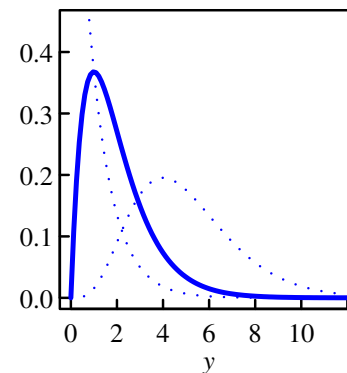
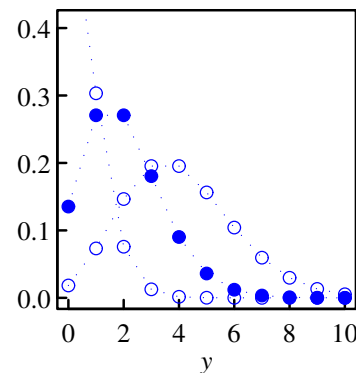
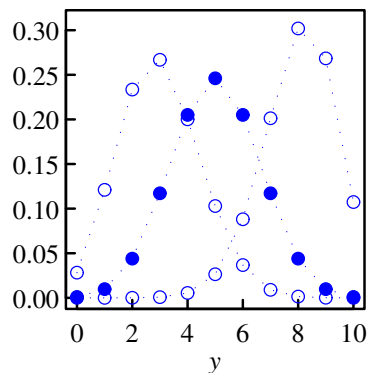
データ解析とはモデリングによる情報圧縮

## 2. 乱数 (標本) と推定のおさえる

データのばらつきをよく見る → 確率分布

## 3. ダメ解析を避ける

割り算値解析しない, 図を描く, データにあわせた手法を



# 次回予告

植物生態学特論 II 代講 2005.07.06

全部で 2 回講義の 2

## 「検定」の使われかたを観察してみる

— 「検定」はそんなにエラいのか? —

「ゆーい差」重視なヒトたちの論法

<http://hosho.ees.hokudai.ac.jp/~kubo/stat/2005/>