生態学のデータ解析 - R 講習

森林総研 2008

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/ce/RwsFfpri2008.html

講釈: 久保拓弥 kubo@ees.hokudai.ac.jp

この二日間の R 講習で説明したいこと

- 自然科学のデータ解析ってのは統計モデリングというモデル作りである
- R の上で自由にデータを操作したり作図するのは統計モデリングにおいてとても重要
- 統計モデリング勉強の手はじめとして一般化線形モデル (GLM) を使えるようになればよいだろう

(予定はこちらに)

http://hosho.ees.hokudai.ac.jp/~kubo/ce/RwsFfpri2008.html

まあ，皆さんのデータ解析のとりかかりになれば，と ……

前口上的なハナシ: 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 ?」

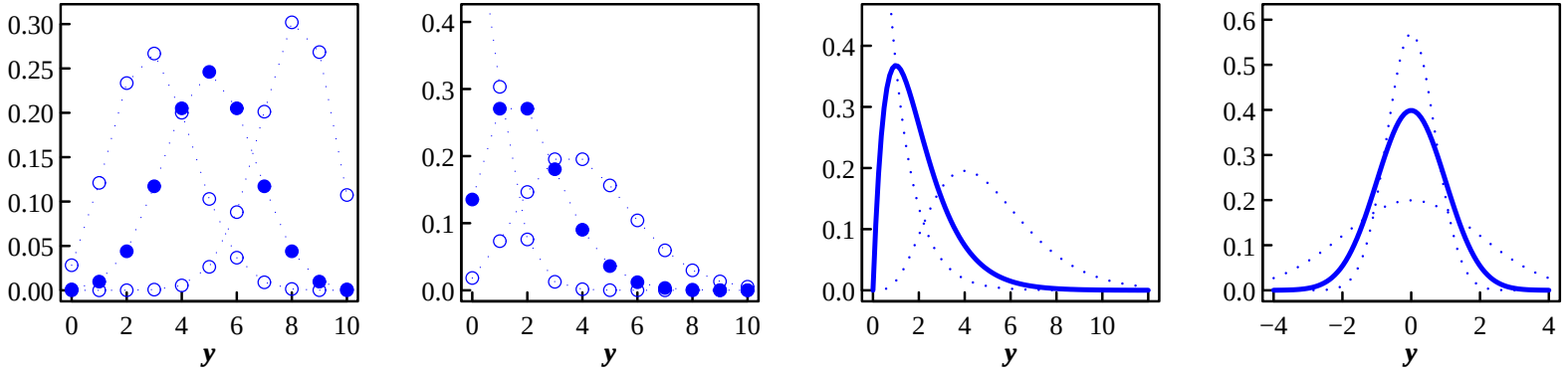
どういうふうに使えて，どう勉強すればいいか

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

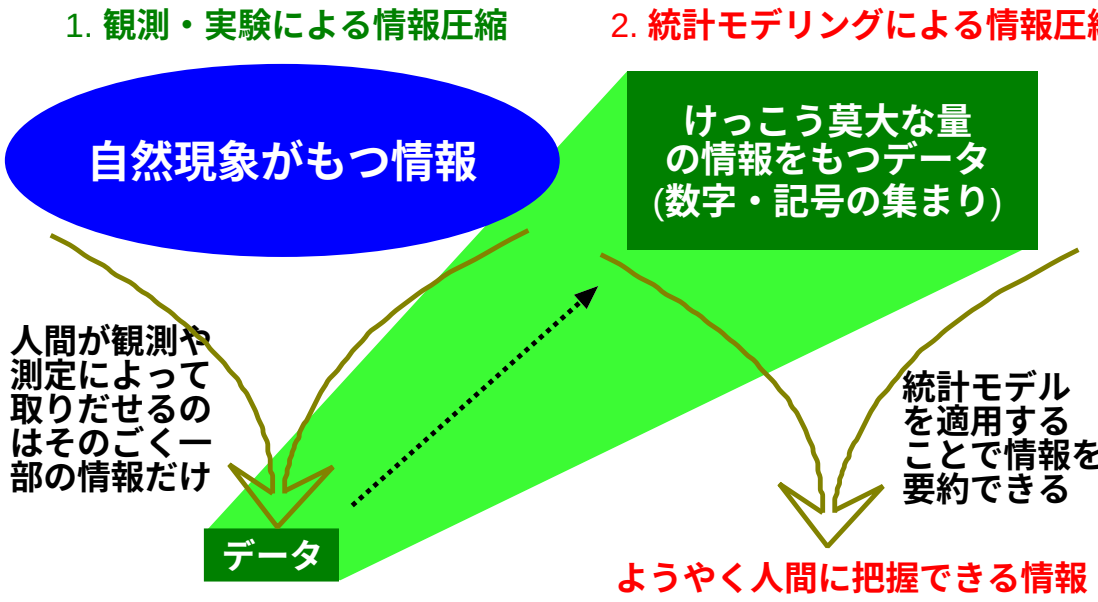


1. とりあえずの「統計学って何 ?」

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

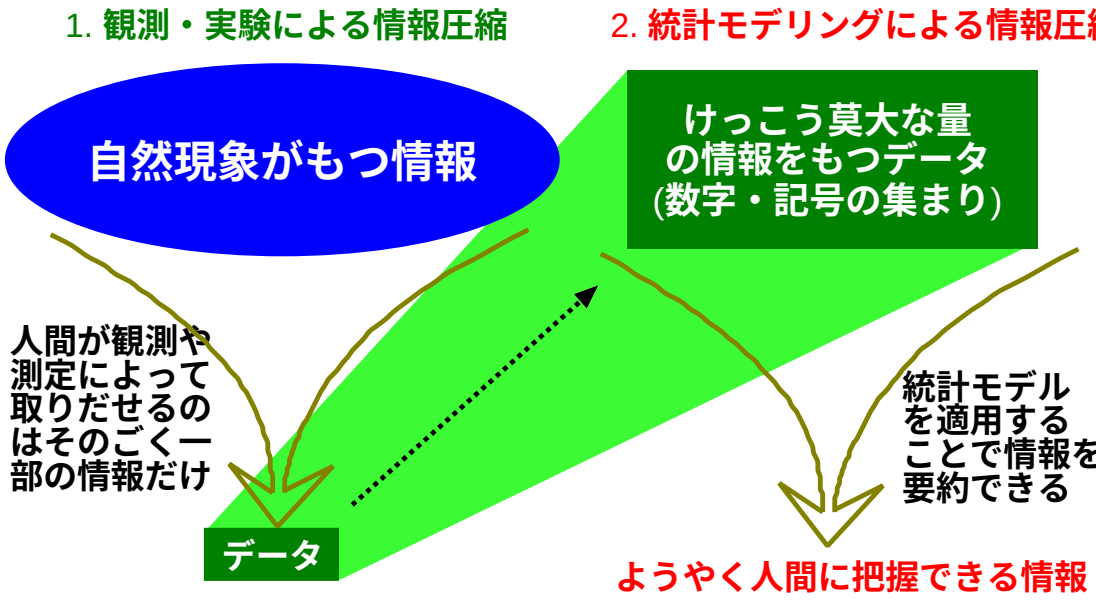
第一段階: 自然現象 → 数値データ



- 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

自然科学研究における二段階の情報損失

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)

- スジのとお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

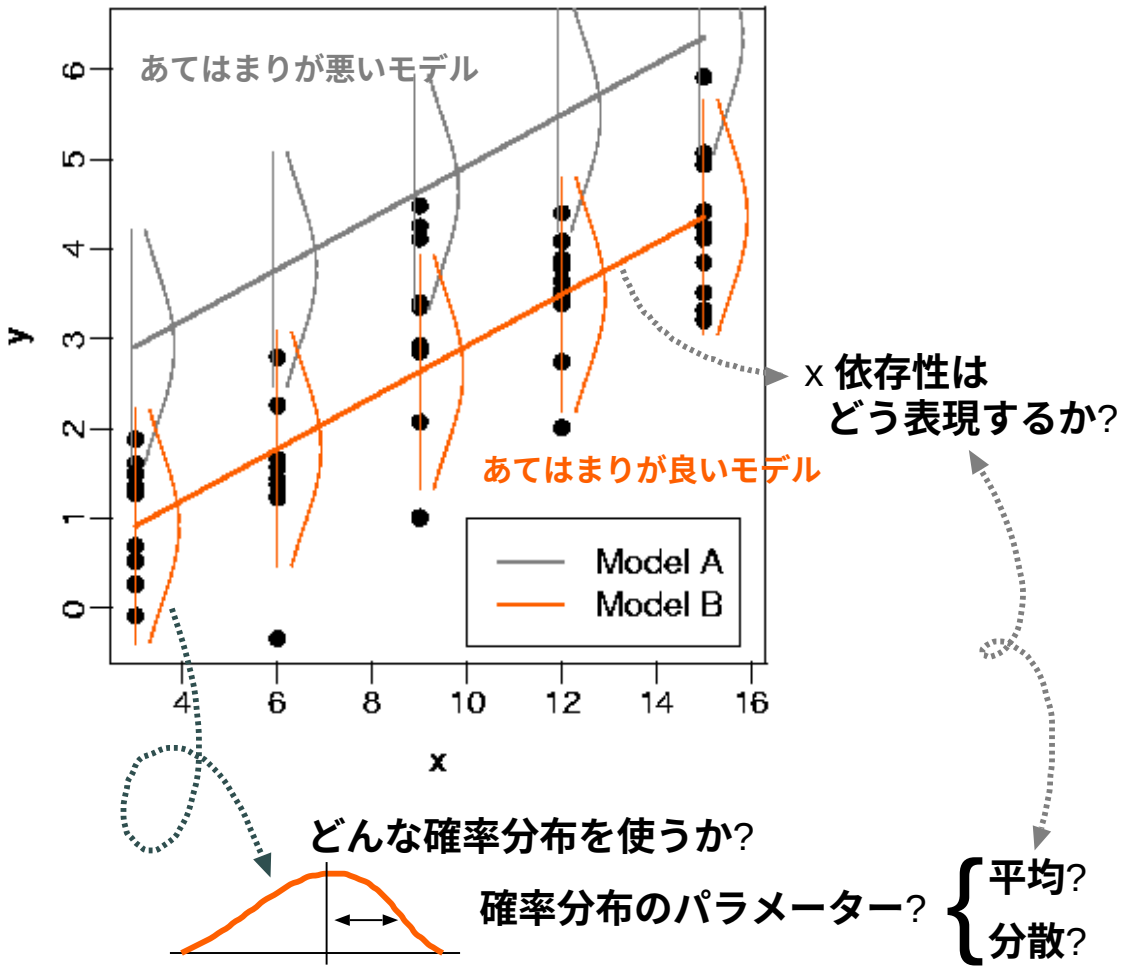
問: 統計モデリングとはどういうものか?

統計モデリングとは何か?

データ解析とは統計モデリングのことだ

- 統計モデルは (解析したい) 観測データと対象に関する先験的な知識・情報にもとづいて構築される
- 統計モデルは観測データのパターンをうまく**説明**できるようなモデル
- 統計モデルの基本的な部品は**確率分布**，確率分布のカタチはパラメーターによって決まる
- 観測データをうまく説明できるようにパラメーターの値を決めることを「統計モデルの**あてはめ**」または「統計モデルによる**推定**」という

統計モデル ← 確率分布 ← パラメーター ← 説明変数など



観測データと統計モデルの「比較」 ⇒ パラメーターの推定

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**

統計モデリングわざを修得するため，とりあえず

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

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 による保健医療データ解析演習」 中澤港 (2007)
  - 「The R-Tips」 舟尾暢男 (2005)
  - "Statistics: An Introduction Using R" M. Crawley (2005)
  - ネット上のあちこち

植生解析・群落動態な R package あれこれ

- `library(vegan)`
  - Jari Oksanen たちによる．Ordination tools (CCA, RDA, DCA, CA, PCA) などが含まれる
- `library(VGAM)`
  - Thomas W. Yee の package．Vector Generalized Linear and Additive Models，より新しい群集モデリングの道具
- `library(CTFS)`
  - The Center for Tropical Forest Science (CTFS) がバロコロラド島 50ha 調査地のデータを解析するときに使うツール群

Open で free な統計モデリングの道具

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

(線形な部分をもつ) 統計モデルの地図

古典的な統計学的方法から最新の統計モデリングへのつながりを示す

[尤度をあつかう統計モデル]

パラメーターを確率分布として表現する Bayes 統計学

階層 Bayes モデル など

[最尤推定法 であつかう統計モデル]

パラメーターの推定値を点推定する，random effects もあつかえる

経験 Bayes 法: 一般化線形混合モデル (GLMM) 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, fixed effects のみ

[最小二乗法 であつかう統計モデル]

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

よい教科書?? とりあえず久保「講義のーと」

http://hosho.ees.hokudai.ac.jp/~kubo/ce/EesLecture2007.html

1. 生態学データ解析の統計モデリングとは?
2. さまざまな確率分布と最尤推定
3. 一般化線形モデル (GLM) 1
4. 一般化線形モデル (GLM) 2
5. 検定とモデル選択
6. 一般化線形混合モデル (GLMM)
7. 階層ベイズモデル

統計モデリング重視 + R での実例を重視
興味があるヒトは download してください

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?

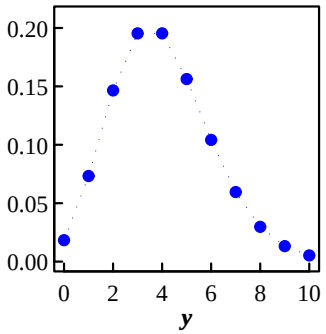
統計学の中核概念

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, $\lambda = 3$)



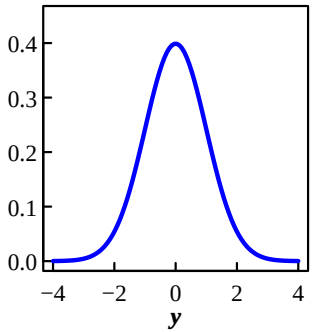
→

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

正規分布

R の関数:

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



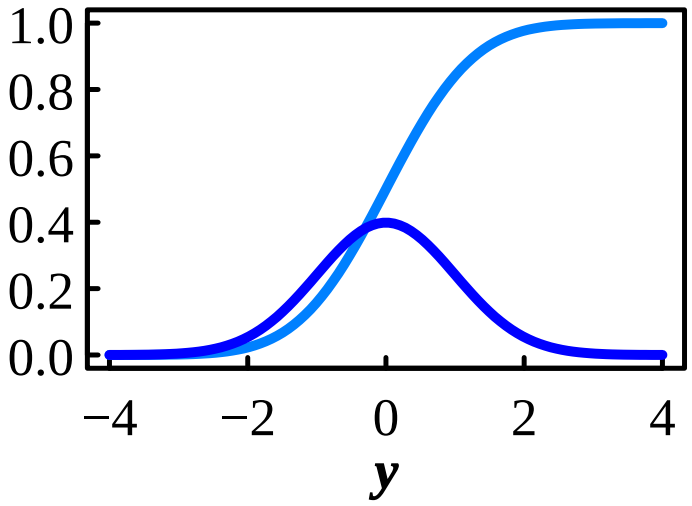
→

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

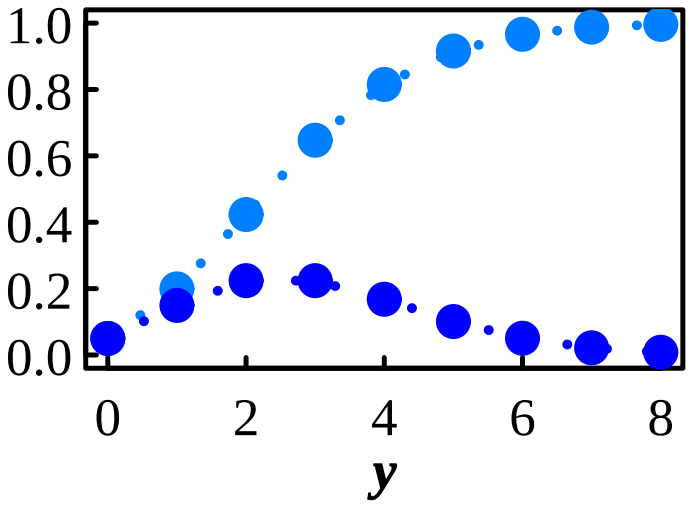
確率分布: 確率分布 (関数) と確率密度分布 (関数)

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

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

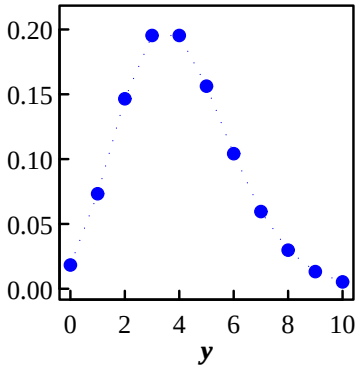
分散: ばらつき

$$\mathrm{Var} = \int_{-\infty}^{\infty} (y-m)^2 \, \mathrm{d}f(y) \qquad \mathrm{Var} = \sum_{0}^{\infty} (y-m)^2 f(y)$$

じゃあ推定ってのは何なの? → 乱数生成の逆

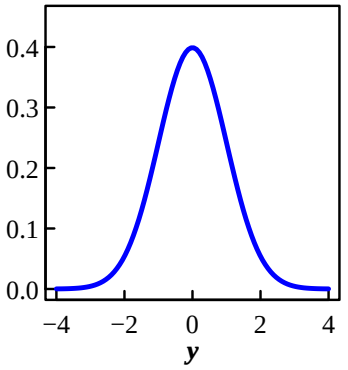
ポアソン分布の推定

5 4 3 2 4 2 4 1 7 1 →



正規分布の推定

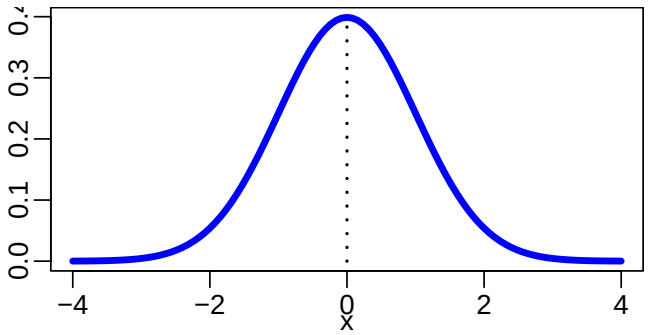
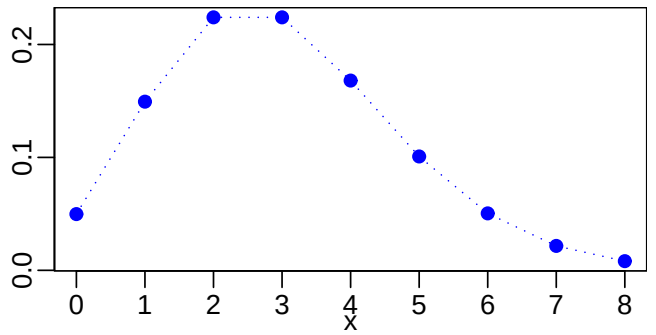
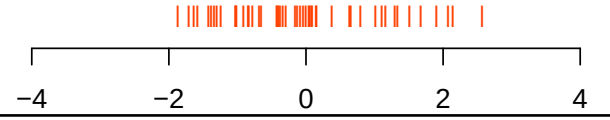
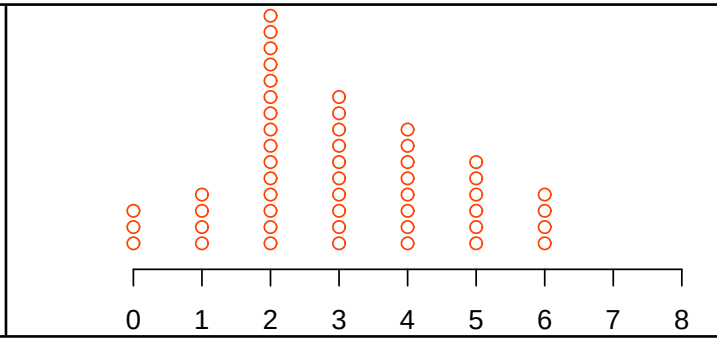
```
 1.4851004 -0.9912880 -0.1092131 →
-2.1752314 -0.3779424  1.1360432
 1.2493592 -1.2405408 -0.4425550
```



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!

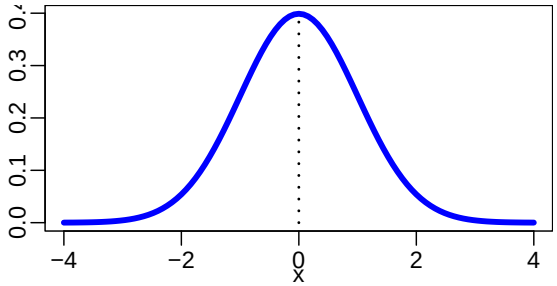
(よびかた)	[連続確率密度分布]	[離散確率密度分布]
• モデル • 確率分布 • 母集団		
• データ • 乱数 • 標本集団		

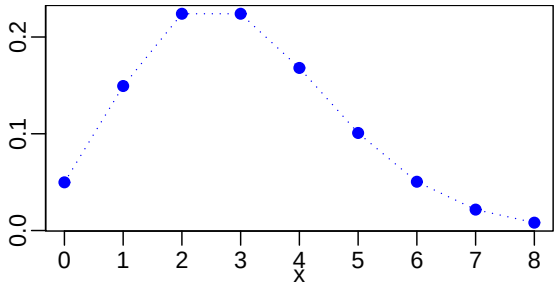
サンプリング ⇓ ⇑ (パラメーター) 推定

推定については後で例を示す

ここまでで言いたいこと: 乱数と自然現象

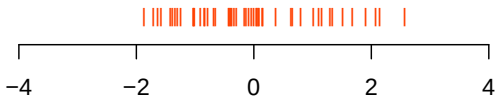
- モデル
- 確率分布
- 母集団

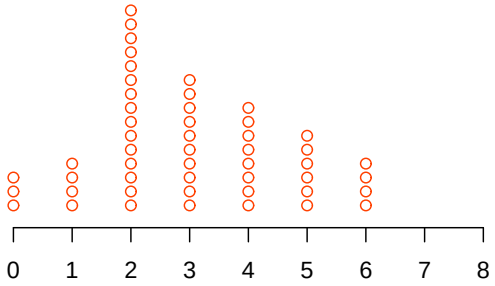




サンプリング ⇓ ⇑ (パラメーター) 推定

- 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……

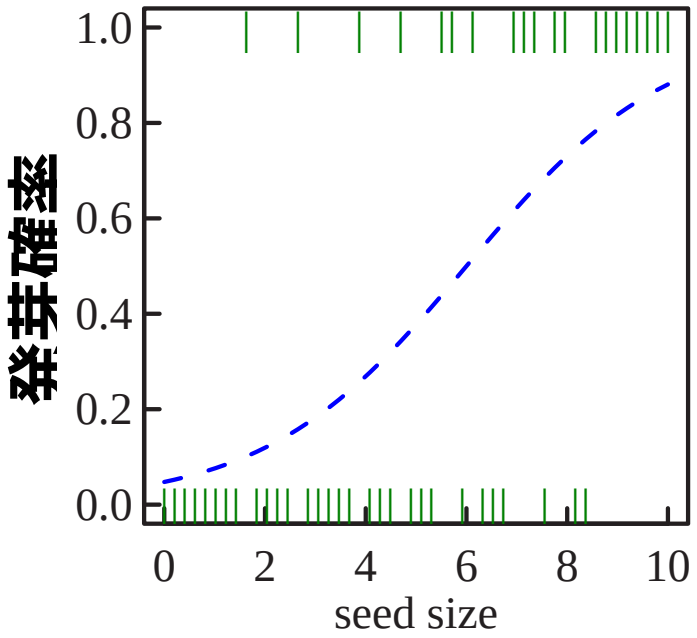
3. ダメ解析と良い解析

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

| は観測データ (1 = 発芽した)

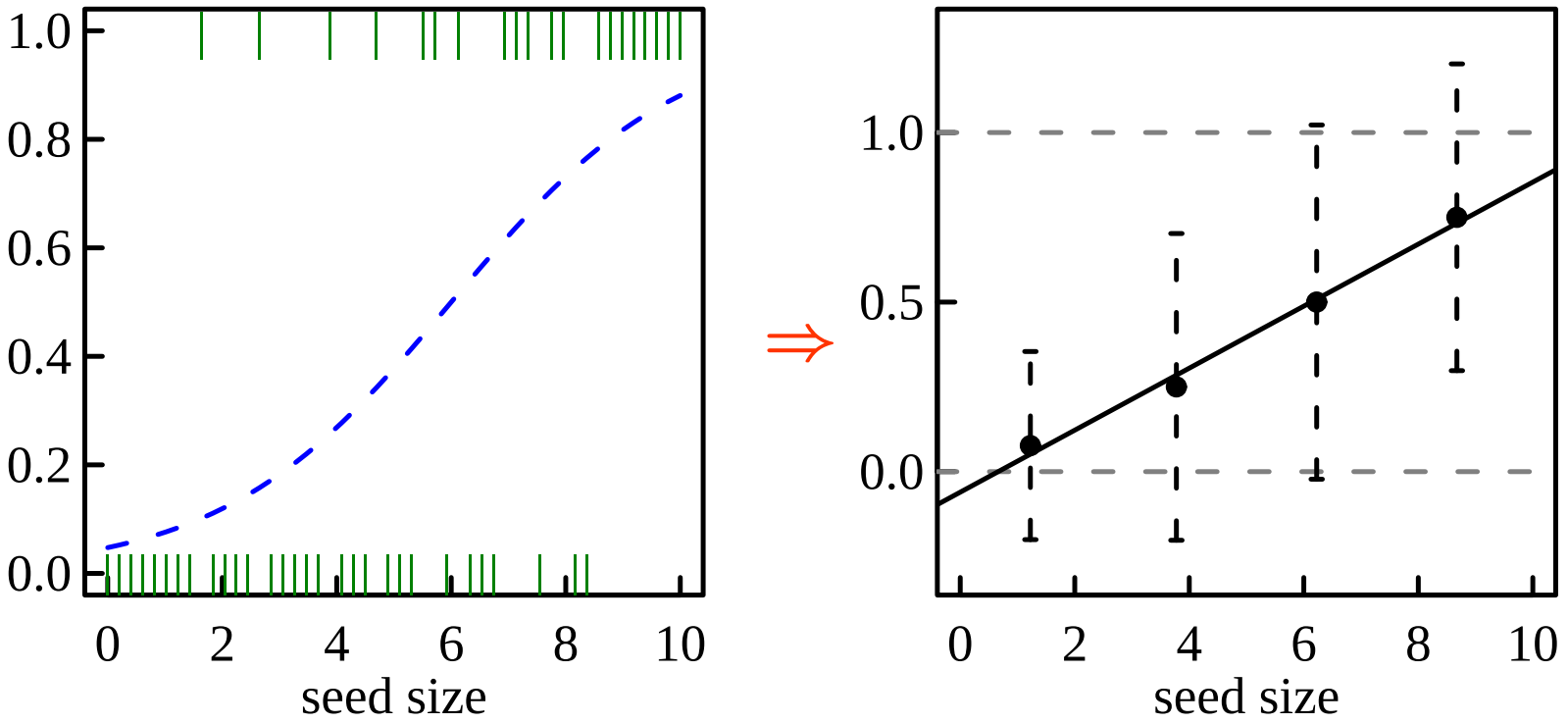


["神" の立場で知ってるコト]

- 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

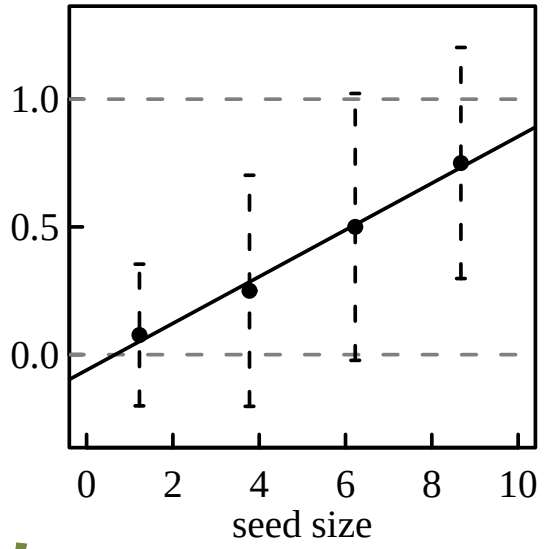
区画はてきとー

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

[尤度をあつかう統計モデル]

パラメーターを確率分布として表現する Bayes 統計学

階層 Bayes モデル など

[最尤推定法 であつかう統計モデル]

パラメーターの推定値を点推定する，random effects もあつかえる

経験 Bayes 法: 一般化線形混合モデル (GLMM) 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, fixed effects のみ

[最小二乗法 であつかう統計モデル]

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

とりあえず，の一般化線形モデル (GLM; R の glm() 関数)

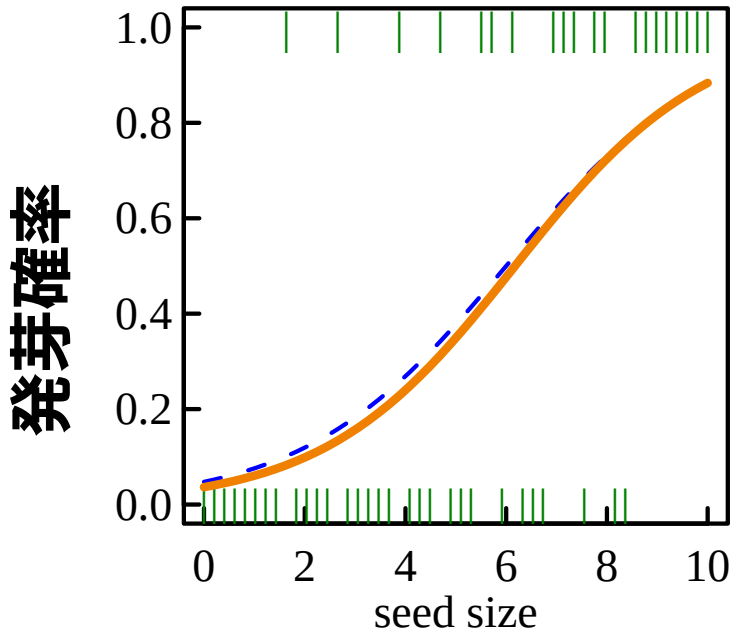
いろいろな確率分布に適用できる推定計算手法

	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	`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

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

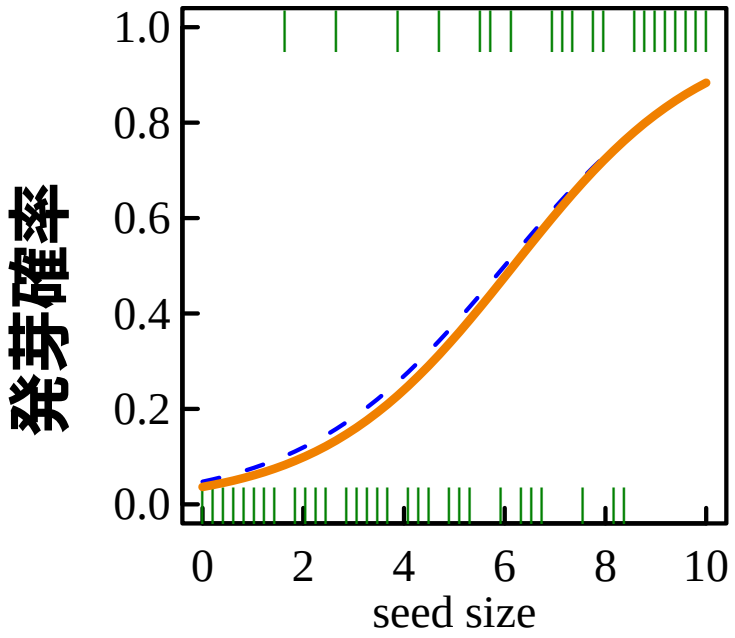
$$p = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}$$

(logistic 式)

- パラメーター β_0 と β_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) と推定の関係をおさえる

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