

多様性生物学基礎論 (生物多様性論 I): 生物多様性の現在

5. 生物多様性解析法：統計モデリングの基礎

全部で 2 回講義の 1

カウント

「数えられる」データの 統計解析・統計モデリング

一般化線形モデル (GLM) 入門

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

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
 - しかし多くのユーザーは よくわからん 状態で使ってる
 - この授業の目的はその「わからん度」を少しでも下げること
-
- 第 1 回: 2006.01.23 (月)
「数えられる」データの統計解析・統計モデリング
一般化線形モデル (GLM) 入門
 - 第 2 回: 2006.01.25 (水)
「個体差」を考慮した統計解析・統計モデリング
一般化線形混合モデル (GLMM) 入門

今日のハナシ: 一般化線形モデル (GLM) にさわる

1. てみじかに「統計学って何？」

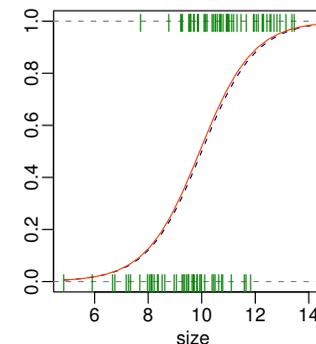
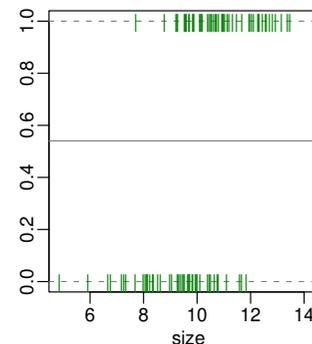
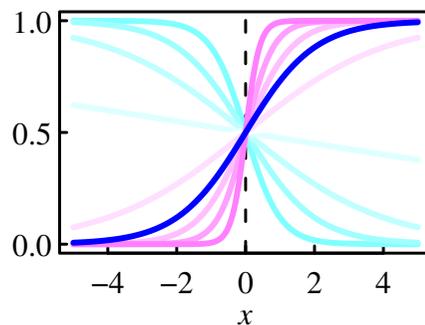
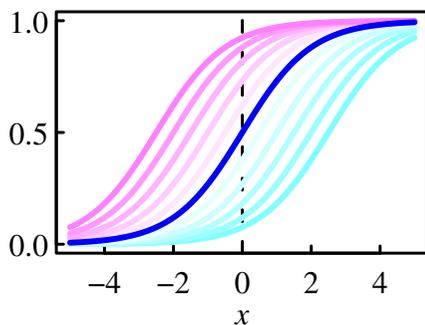
どういうふうに使えて、どう勉強すればいいか

2. 最尤推定法とロジスティック回帰

カウントデータ解析の基本中の基本

3. さらに強めるロジスティック回帰わざ

説明変数とそのパラメーター推定



まずは簡単に



1. てみじかに「統計学って何？」

そして、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」過程の現状と勉強法

生態学研究まわりにおける現状

- 軽視されている
- 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

(この状況下で) 統計学を自分で勉強するためには

- よい教科書が必要
- よい統計ソフトウェアが必要 (実験しつつ勉強するために)
- 相談できる相手をさがす

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

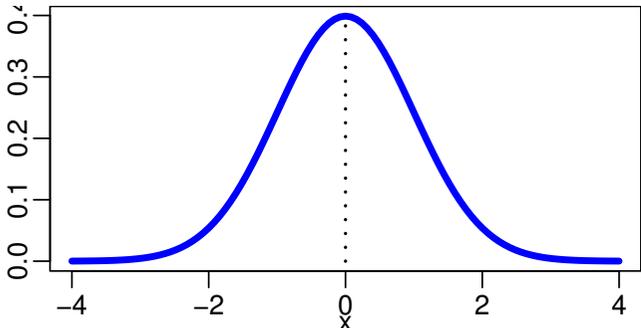
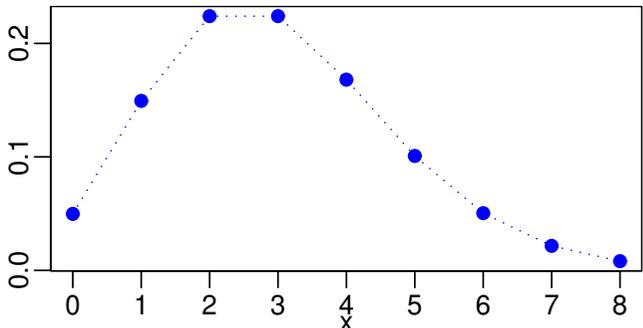
<http://www.r-project.org/>

- 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

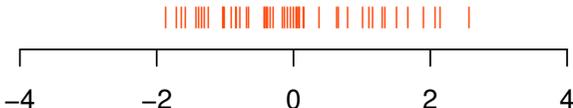
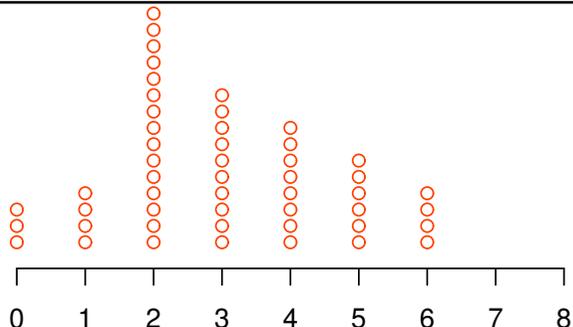


なぜ統計ソフトウェアが必要なのか? — 試行錯誤のため

統計学とは結局これ: 確率分布, 乱数と推定

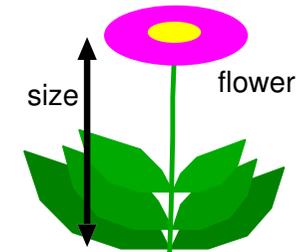
(よびかた)	[連続確率密度分布]	[離散確率密度分布]
<ul style="list-style-type: none"> ● モデル ● 確率分布 ● 母集団 		

サンプリング ↓ ↑ (パラメーター) 推定

<ul style="list-style-type: none"> ● 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

「あった」「なかった」現象はこう解析しよう

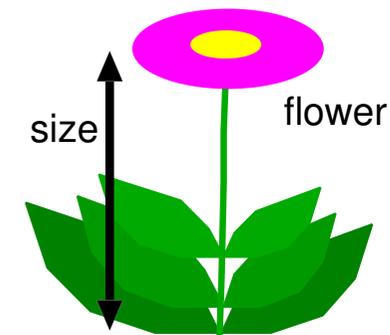


2. 最尤推定法とロジスティック回帰

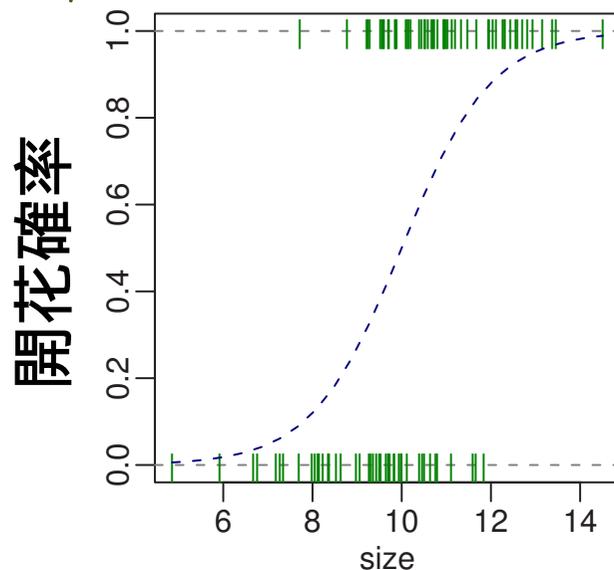
カウントデータ解析の基本中の基本

架空植物の観測データ: 開花現象の統計モデル

「開花する」という現象を統計モデルで表現したい。どうすればよいか?



は観測データ (1 = 開花した)



[観測データ]

- 一個体にひとつの花
- 標本個体数 100, 開花個体数 54

["神" の立場で知ってるコト]

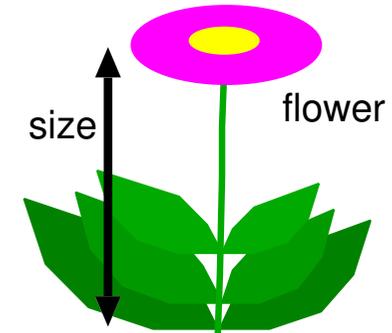
- サイズが大きいほど開花確率が高い (しかしこれはしばらく**放置**する)

まずは「開花した」「しなかった」現象だけ考える (size 無視)
開花確率 = $54 / 100 = 0.54$? なぜそう考えてよいのか?

統計モデルはつねに「最尤法」のわくぐみで考える

ゆーど
尤度 (likelihood) $L(p)$:

あるパラメーター p のもとで観測データが得られる確率 (.....と, とりあえず定義しときましょう)



パラメーター「開花確率」 p (具体的な値はまだわからない)

「標本個体数 100, 開花個体数 54」という観測データが得られる確率は,

$$L(p) = \text{確率} \{ \text{咲いた } 54, \text{ 咲かない } 46 \} = p^{54} (1 - p)^{46}$$

この方程式を**尤度方程式**という. これは (ここでは「個体を区別している」つもりなので) ベルヌーイ分布になる (区別してなければ二項分布).

尤度 (と対数尤度) はパラメーターの関数

尤度方程式 $L(p) = p^{54}(1 - p)^{46}$ はこのままではあつかいづらいので、
対数尤度方程式

$$\log L(p) = 54 \log(p) + 46 \log(1 - p)$$

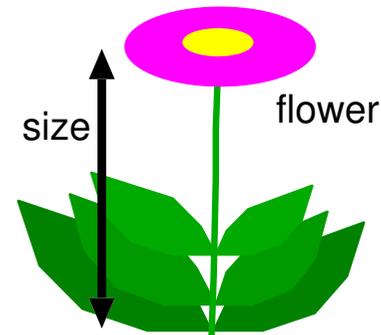
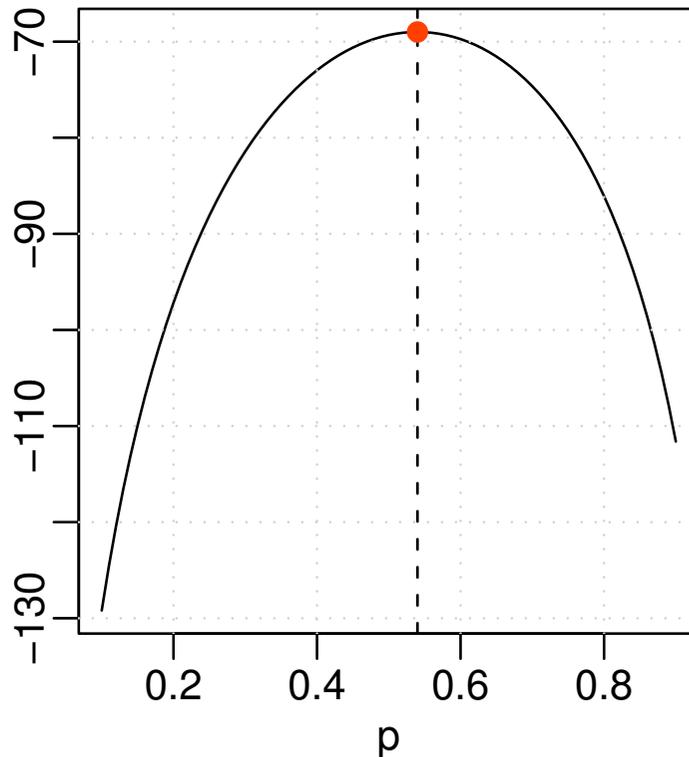
で考える ($\log L(p)$ は $L(p)$ の単調増加関数なんで) .

R で計算してみよう .

```
> for (p in seq(0.3, 0.7, 0.1))  
+ cat("p =", p, ": logL(p) =", 54 * log(p) + 46 * log(1-p), "\n")  
p = 0.3 : logL(p) = -81.422  
p = 0.4 : logL(p) = -72.978  
p = 0.5 : logL(p) = -69.315  
p = 0.6 : logL(p) = -69.734  
p = 0.7 : logL(p) = -74.643
```

尤度を最大化する最尤推定値がある!

p と対数尤度の関係



- 標本個体数 100
- 開花個体数 54

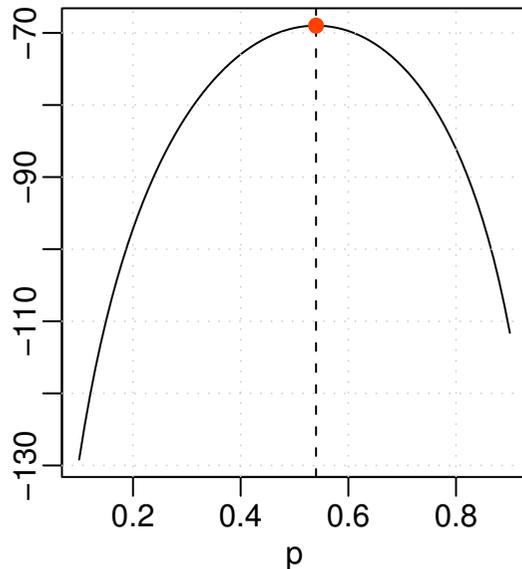
対数尤度

$$\log L(p) = 54 \log(p) + 46 \log(1 - p)$$

を最大化する \hat{p} は? (この推定計算が**最尤推定**)

「開花確率 0.54」は最尤推定値!

p と対数尤度の関係



対数尤度

$$\log L(p) = 54 \log(p) + 46 \log(1 - p)$$

を最大化する \hat{p} は?

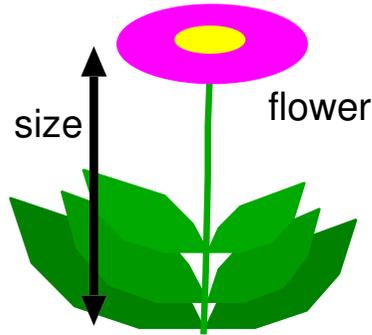
対数尤度 $\log L(p)$ を p で偏微分して

$$\frac{\partial \log L(p)}{\partial p} = \frac{54}{p} - \frac{46}{1 - p}$$

$\frac{\partial \log L(p)}{\partial p} = 0$ となる \hat{p} が最尤推定値 .

$$\hat{p} = \frac{54}{100} = 0.54$$

次の一歩: ロジスティックモデル化



- 標本個体数 100
- 開花個体数 54

- 咲いた・咲かないだけなら
 $L(p) = p^{54}(1 - p)^{46}$ でよい
- しかし size 依存性とか調べたければ?

→ ロジスティック (logistic) モデル化

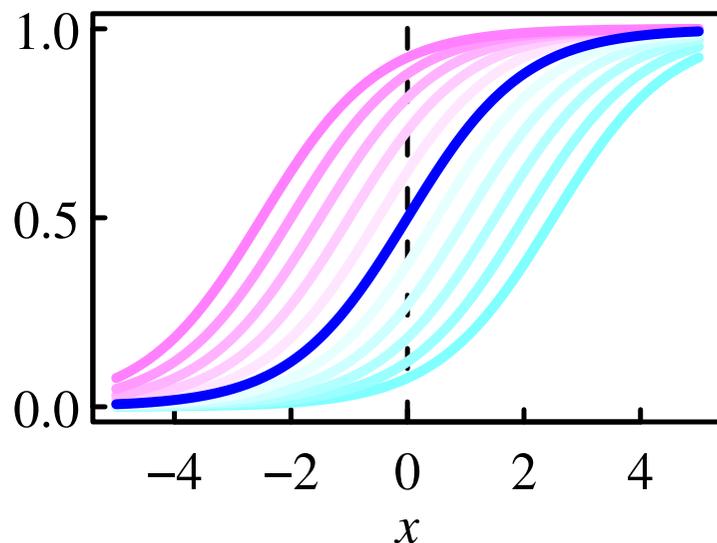
開花確率を $p(a) = \frac{1}{1 + \exp(-a)}$ とおく . 尤度方程式は

$$\begin{aligned} L(a) &= p(a)^{54}(1 - p(a))^{46} \\ &= \left\{ \frac{1}{1 + \exp(-a)} \right\}^{54} \left\{ \frac{\exp(-a)}{1 + \exp(-a)} \right\}^{46} \end{aligned}$$

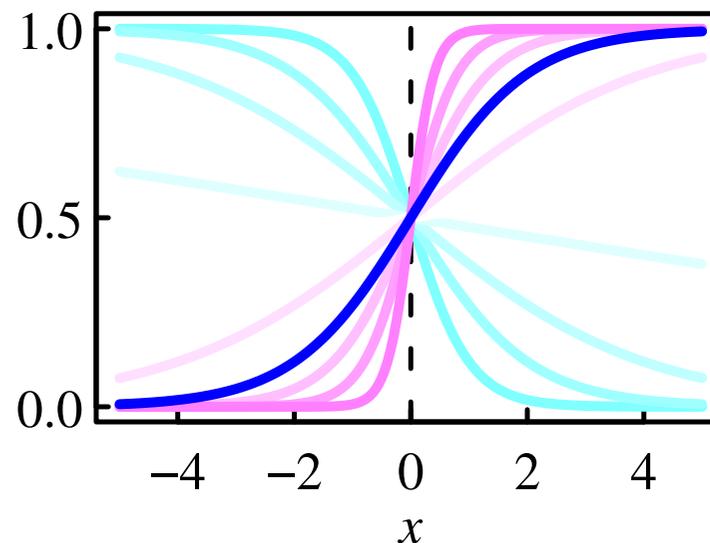
その「ロジスティック関数」って何なの？

$$p(x) = \frac{1}{1 + \exp(-(a + bx))} \quad (\exp(Z) = e^Z \text{ のこと})$$

a だけ変化させる



b だけ変化させる



つまり $p(a) = \frac{1}{1 + \exp(-a)}$ の a がどんな値をとっても
開花確率 $p(a)$ は $0 \leq p(a) \leq 1$ となる便利な関数

ロジスティックモデルの最尤推定も同じように

「標本個体数 100 , 開花個体数 54」観測データの尤度方程式

$$L(a) = \left\{ \frac{1}{1 + \exp(-a)} \right\}^{54} \left\{ \frac{\exp(-a)}{1 + \exp(-a)} \right\}^{46}$$

から対数尤度方程式

$$\log L(a) = -46a - 100 \log(1 + \exp(-a))$$

となり , $\frac{\partial \log L(a)}{\partial a} = 0$ となる最尤推定値 \hat{a} は

$$\hat{a} = \log \frac{54}{46} \approx 0.16 \quad \text{..... これは対数オッズ (log odds)}$$

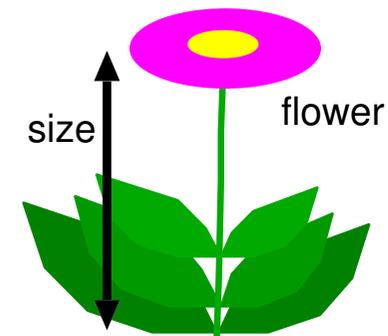
最尤推定値つかった開花確率 $p(\hat{a}) = \frac{1}{1 + \exp(-0.16)} \approx 0.54$

ロジスティックモデルの最尤推定, glm() で (1)

まずはデータの準備 (CSV 形式のテキストファイルを読みこむ)

```
> d <- read.csv("d.csv")  
> head(d) # d の先頭をみる
```

	id	size	flower
1	f001	4.8525	0
2	f002	5.9208	0
3	f003	6.6705	0
4	f004	6.7628	0
5	f005	7.1813	0
6	f006	7.2697	0



`flower` は 0 (開花していない) と 1 (開花している) の値をとる

(CSV: Comma Separated Value, コンマ区切りテキストファイルのこと)

ロジスティックモデルの最尤推定, `glm()` で (2)

`glm()` は R の「一般化線形モデル推定計算」用の関数であり, さまざまな結果を格納した `glm` オブジェクトを返却する

```
> glm(flower ~ 1, family = binomial, data = d)
```

`glm()` の指定のいくつかを簡単に説明すると

- `flower ~ 1` 応答変数 `flower` を (パラメーター) ×1 で説明しろ
- `family = binomial` ばらつきを説明する確率分布は二項分布で (ここは二項分布と指定して問題ない)
- `data = d` データは CSV ファイル読みこんだ `d` を使う (R では `d` のようなデータ構造を `data.frame` とよぶ)

ロジスティックモデルの最尤推定 , glm() で (3)

glm() の推定結果を読む

```
> summary(glm(flower ~ 1, family = binomial, data = d))
```

```
Call:
```

```
glm(formula = flower ~ 1, family = binomial, data = d)
```

```
...(略)...
```

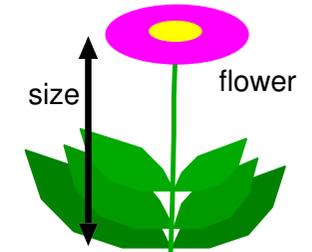
```
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.160	0.201	0.8	0.42

```
...(略)...
```

係数 (Coefficients , ここでいうパラメーターのこと) の
最尤推定値 (Estimate) は $\hat{a} = 0.16$ となった

z value だの Pr(>|z|) だのは z 検定 (Wald 検定) に関するもので , 各推定値が独立と仮定できるときに (ありえん!) , 推定値がゼロから「十分に」離れているかを調べている .



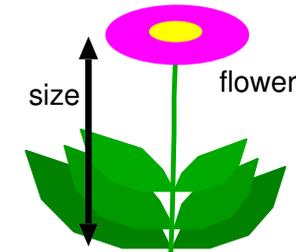
一般化線形モデル (GLM) って論文で見たんですけど.....

3. さらに強めるロジスティック回帰わざ

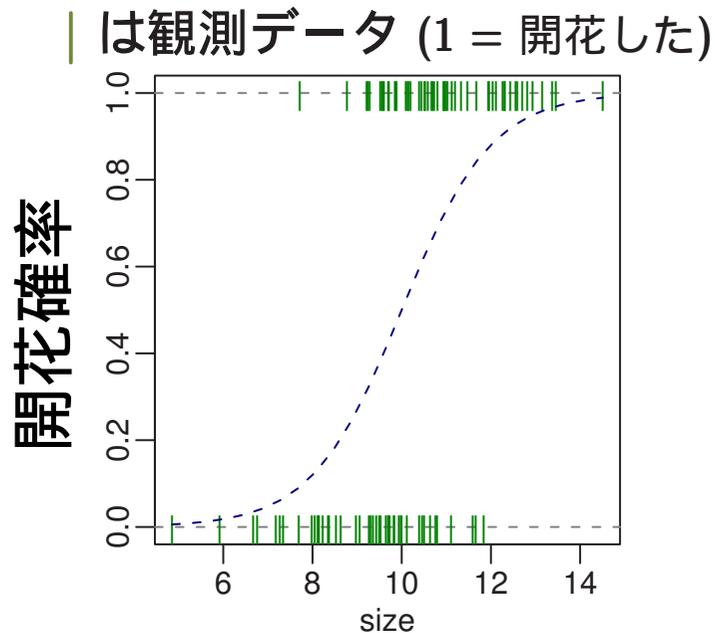
説明変数とそのパラメーター推定

開花確率とサイズの間係を調べたい

開花確率の統計モデル, サイズ依存性を説明変数とするモデルに拡張する



[観測データ]



- 一個体にひとつの花
- 標本個体数 100, 開花個体数 54

["神" の立場で知ってるコト]

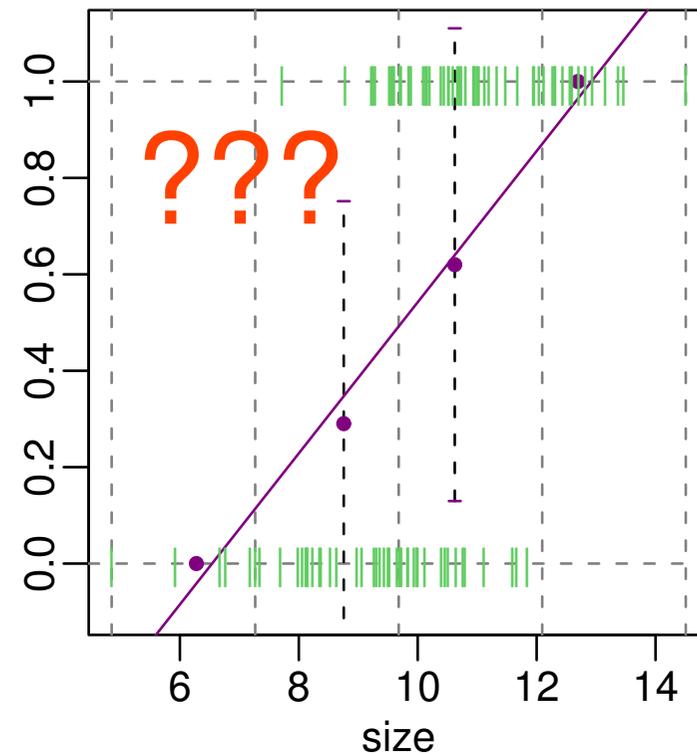
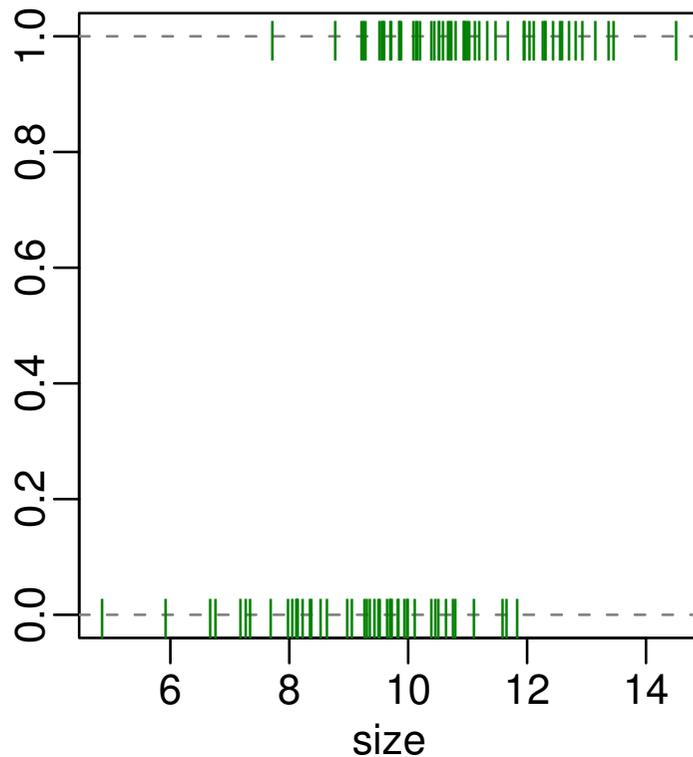
- サイズ (x) が大きいほど開花確率が高い

$$p(a, b) = \frac{1}{1 + \exp(-(a + bx))}$$

$a = -10$ かつ $b = 1$ である

データから青破線 ($a = -10, b = 1$) 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

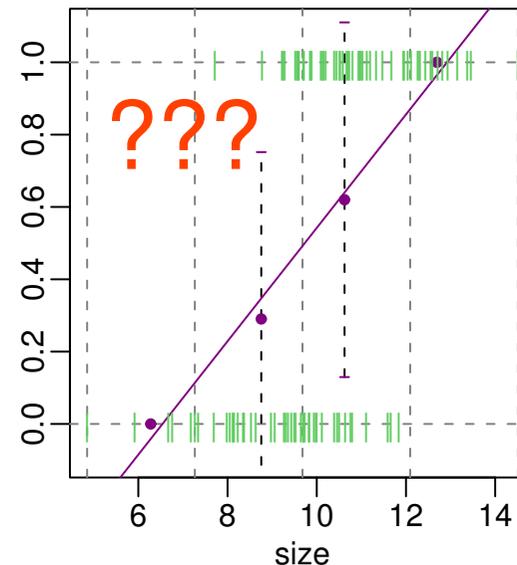
区画はてきとー

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

何を予測してるのだろうか?

開花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)

「なんでもかんでも正規分布でよい」はまちがい

確率分布を推定する方法たちの階層性

[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!」

データにあわせる一般化線形モデル (GLM)

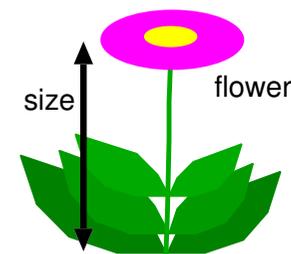
いろいろな確率分布に適用できる推定計算手法

	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>

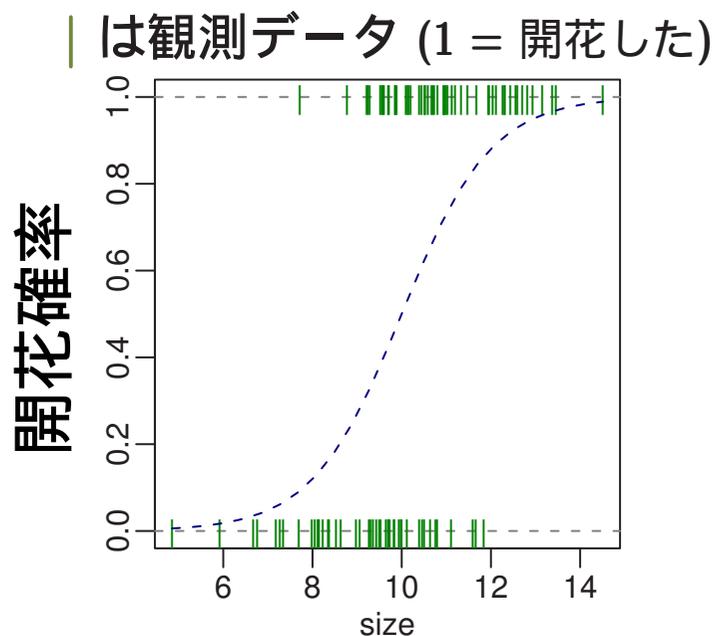
- GLM は Generalized Linear Model の略
- `glm()` で使える確率分布は上記以外もある
- `glm.nb()` は MASS library 中 , またここには `rnegbin()` など含まれる

(再掲) 開花確率とサイズの間係を調べたい

解きたい問題の構造をよく考えて.....



[観測データ]



- 一個体にひとつの花
- 標本個体数 100, 開花個体数 54

["神" の立場で知ってるコト]

- サイズ (x) が大きいほど開花確率が高い

$$p(a, b) = \frac{1}{1 + \exp(-(a + bx))}$$

$a = -10$ かつ $b = 1$ である

データから青破線 ($a = -10, b = 1$) を推定したい

これも glm() ロジスティック回帰で簡単に

glm() で `flower ~ 1 + size` モデルのパラメーター推定

```
> summary(glm(flower ~ 1 + size, family = binomial, data = d))
```

Call:

```
glm(formula = flower ~ size, family = binomial, data = d)
```

...(略)...

Coefficients:

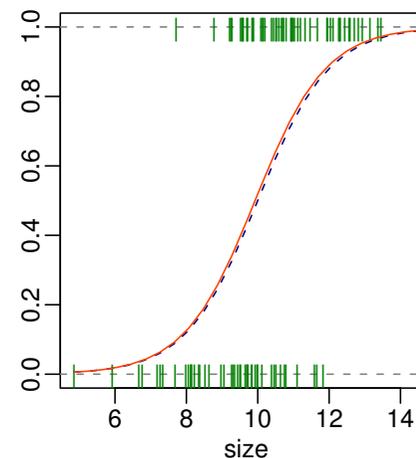
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-10.033	2.239	-4.48	7.4e-06
size	1.011	0.221	4.57	5.0e-06

...(略)...

係数 (Coefficients) の

最尤推定値 (Estimate) は

($\hat{a} = -10.0$, $\hat{b} = -1.01$)



今回，ロジスティック回帰について.....

- 説明してみたこと

- ロジスティック回帰は最尤推定法のひとつで，よけいな割り算とかナシで「あった」「なかった」現象を説明できる
- R で `glm(flower ~ 1 + size, family = binomial, ...)`

- 説明しなかったこと

- `glm()` わざあれこれ
 - * `glm(cbind(開花数, 未開花数) ~ 1 + size, ...)`
 - * 他の family (ポアソン分布など)
- 推定結果，推定値の吟味について
 - * 検定: `anova.glm()` で尤度比検定
 - * **モデル選択** (変数選択): モデル選択規準 (AIC) で `stepAIC()`

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

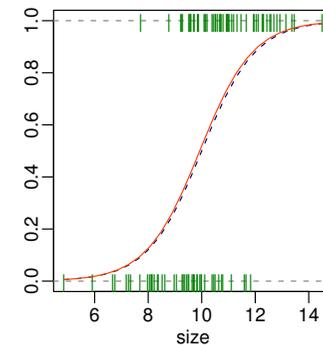
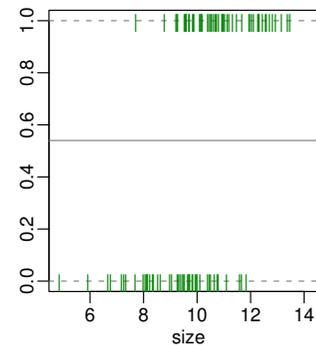
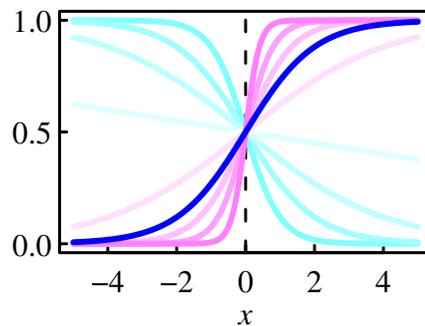
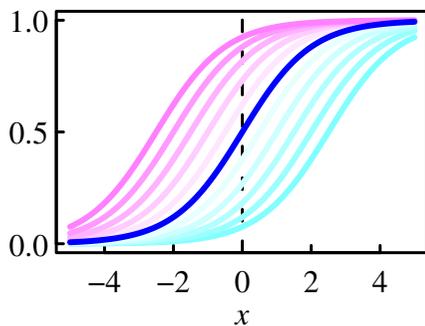
データ解析とはモデリングによる情報圧縮

2. 最尤推定法とロジスティック回帰

カウントデータは, まず `glm()` で!

3. さらに強めるロジスティック回帰わざ

割り算値解析しない, データにあわせてたばらつき (確率分布) を



次回予告

多様性生物学基礎論 (生物多様性論 I) 2006.01.25

全部で 2 回講義の 2

「個体差」を考慮した 統計解析・統計モデリング

— 一般化線形混合モデル (GLMM) 入門 —

「個体差」って何なの？

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