生態学基礎論 (生物多様性論 II)

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

「数えられる (カウント)」データの統計解析・統計モデリング

観測データを一般化線形モデル (GLM) 化しよう

http://hosho.ees.hokudai.ac.jp/~kubo/stat/2006/

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」を少しだけ下げる こと

- **第 1 回**: 2007-01-22 (**月**)
  「数えられる」データの統計解析・統計モデリング

  観測データを一般化線形モデル (GLM) 化しよう
- **第 2 回**: 2007-01-24 (**水**)
  「個体差」を階層ベイズモデルであつかう

  個体差・ブロック差の random effects

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 ?」

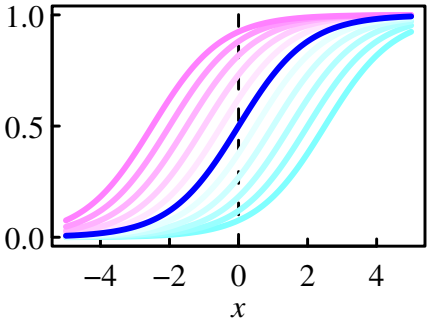
   どういうふうに使えて，どう勉強すればいいか

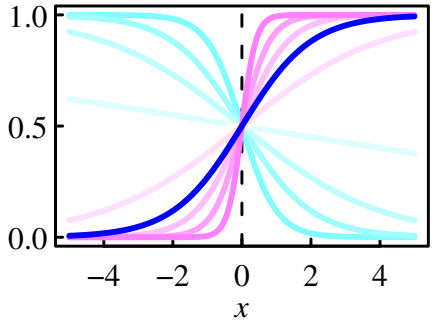
2. 最尤推定法とロジスティック回帰

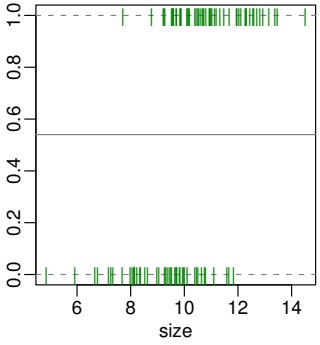
   カウントデータ解析の基本中の基本

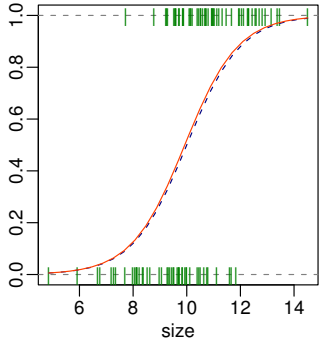
3. さらに強めるロジスティック回帰わざ

   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, ISBN: 4-8299-2123-4
- 発行年月: 1998.3

われわれがいかにたやすく統計学の使いかたや
データ解析のやりかたをまちがってしまうのか
思い知らせてくれる希有な一冊

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

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

生態学のデータ解析の手法: 旧来 vs これから

「旧来」の (というよりすごく古い) 手法

- なんでもかんでも正規分布 → むりやりこじつけるので難解
- なんでもかんでも「ゆーい差」検定
- なぜこんなことを? 数値計算の部分がラクだから

「これから」の (というより現在の常識的な) 手法

- 計算なんて R に全部まかせればよい
- データにあわせて確率分布を選ぶ → わかりやすい
- 現象の統計モデル化，モデル選択，ベイズ統計……

古くさい手法を勉強するのはしんどい
新しくて無理のない手法をすっきりと

「数値データ → 解析結果」過程の現状と理想

(再度) 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)

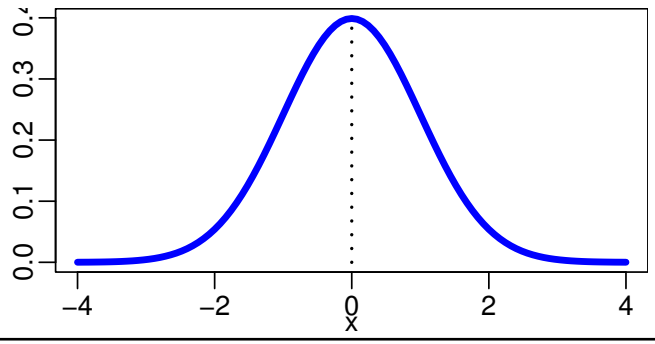
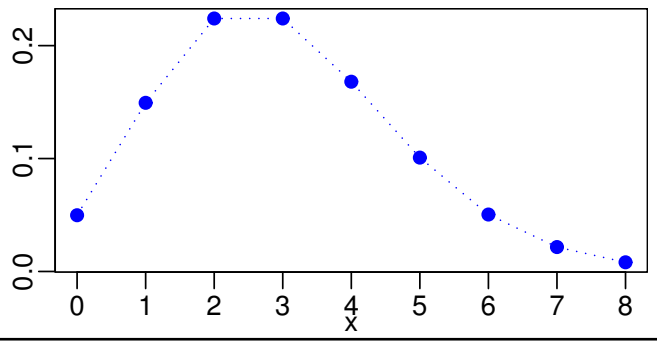
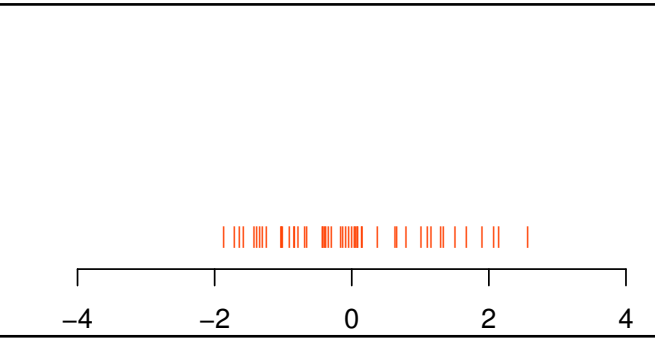
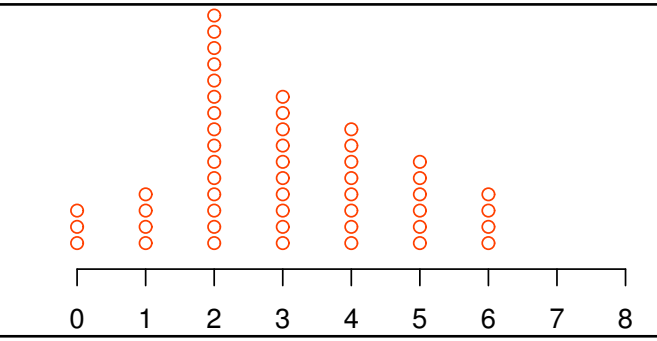
理想 — 情報をうまく圧縮する (ムダなく・わかりやすく)

- 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: そもそも統計学って何なのか?

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

- 「ばらつき」のある観測は「統計モデル」で表現できるだろう
- 「統計モデル」は確率分布を主要な部品とする数理モデル

乱数とは何か?

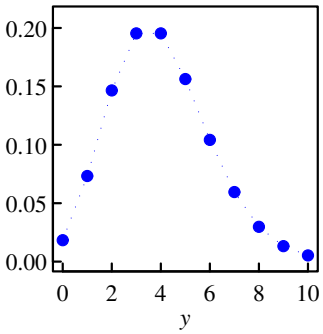
統計学の中核概念

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:

$\texttt{dpois}(y, \lambda = 3)$



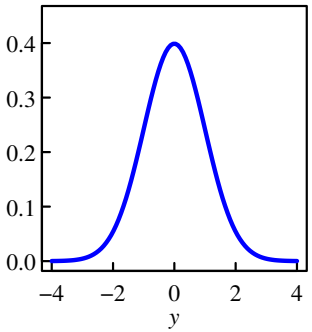
→

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

正規分布

R の関数:

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



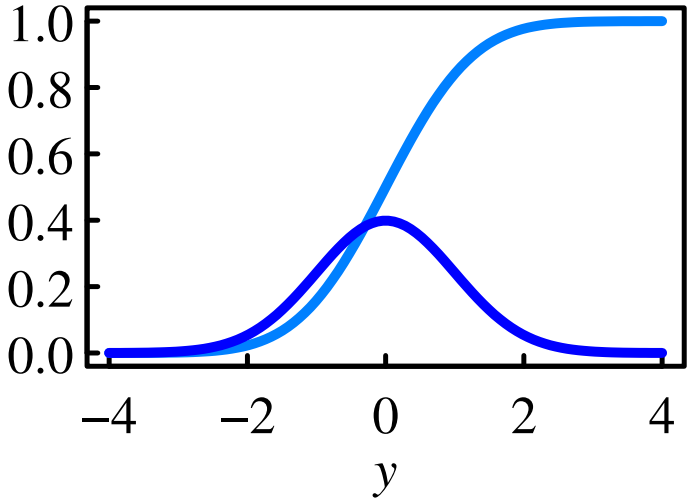
→

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

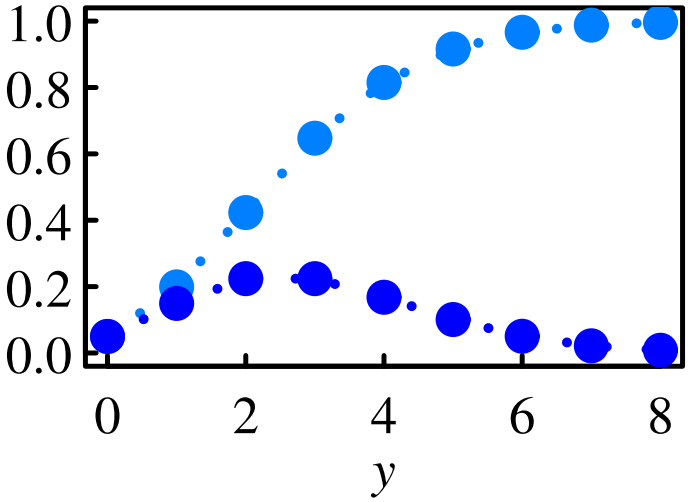
確率分布: 確率分布 (関数) と確率密度分布 (関数)

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

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

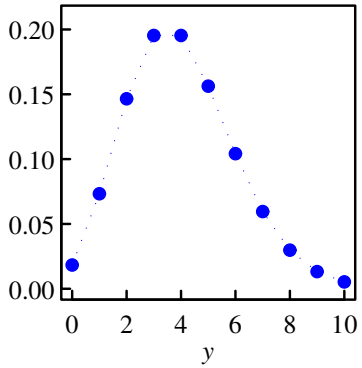
分散: ばらつき

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

じゃあ推定ってのは何なの? → 乱数生成の逆

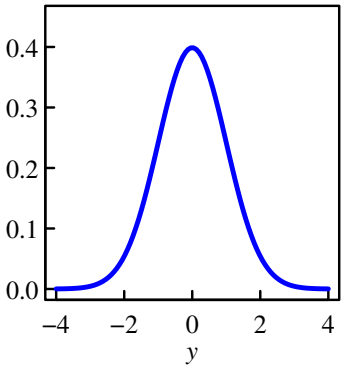
ポアソン分布の推定

5 4 3 2 4 2 4 1 7 1 →



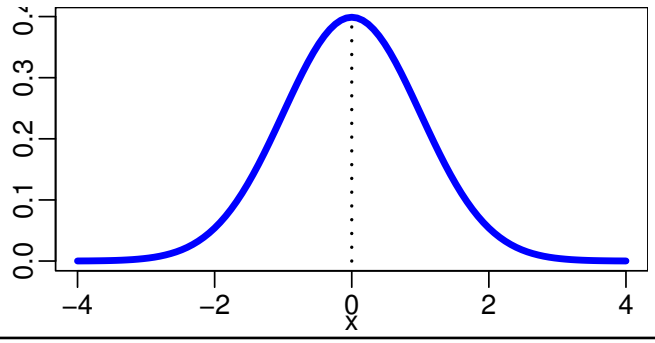
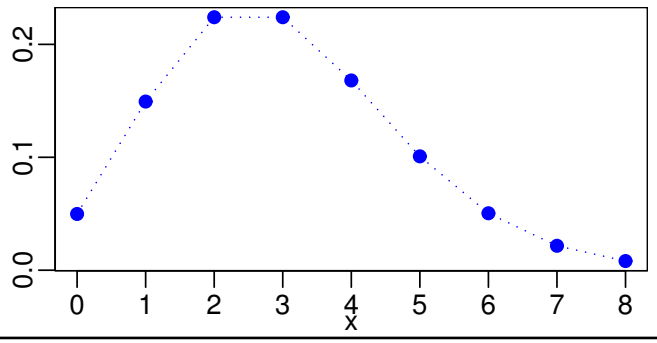
正規分布の推定

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

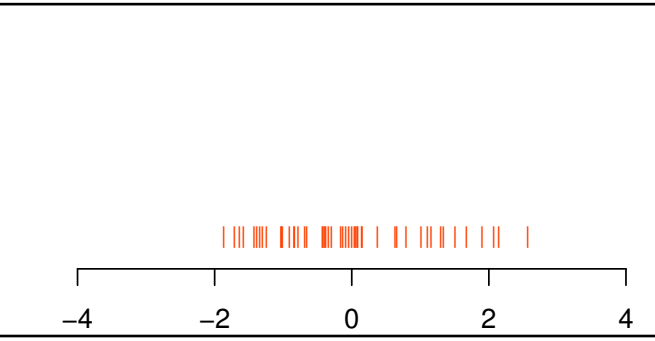
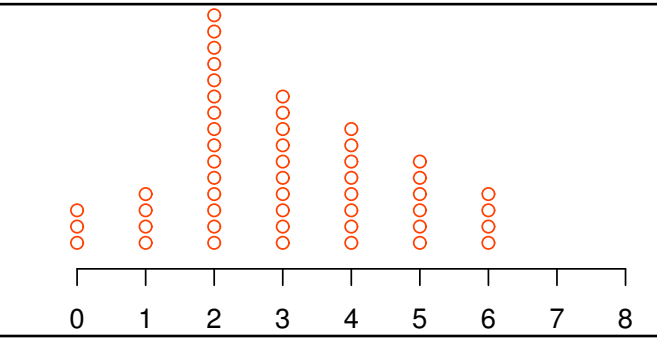


乱数とみなされる標本集団
→ 母集団すなわち確率分布を決め
そのパラメーターを決めてやる技法

統計学とは結局これ: 確率分布，乱数と推定

(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 → 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

統計学勉強における R 実験のススメ

(;_;) 統計学がわからない

→ ひたすら考える・わからぬまま使う

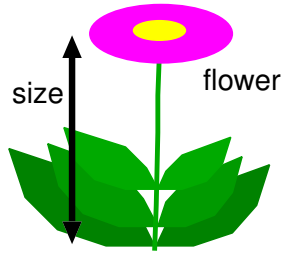
(^_^) 統計学がわからない

→ 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

「あった」「なかった」現象はこう解析しよう

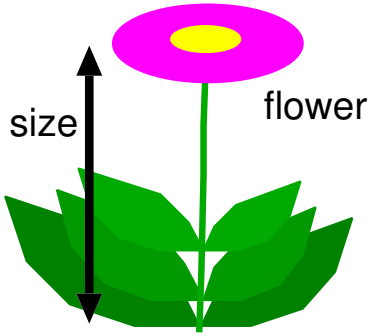


2. 最尤推定法とロジスティック回帰

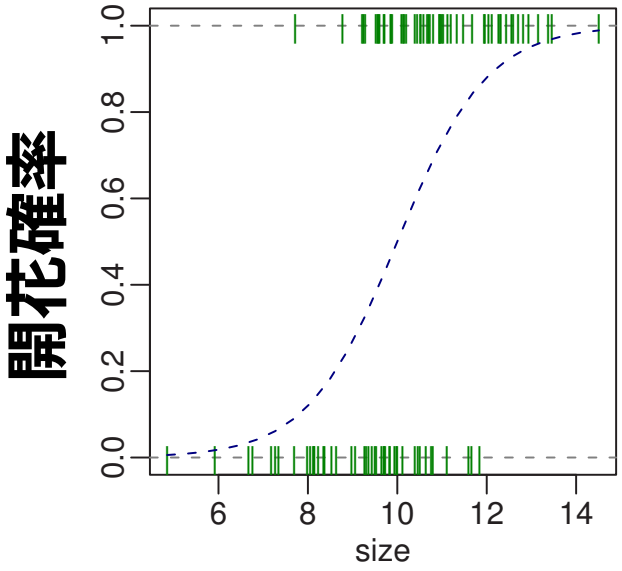
カウントデータ解析の基本中の基本

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]

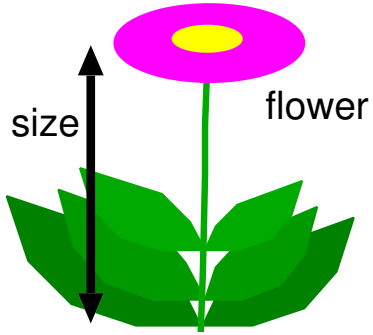
- サイズが大きいほど開花確率が高い
  (しかしこれはしばらく放置する)

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

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

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

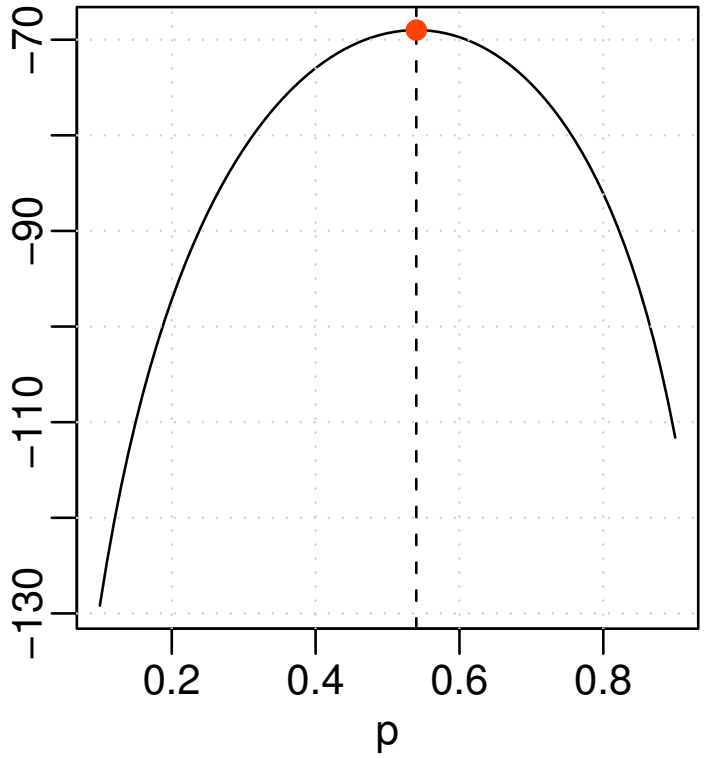
で考える ($\log L(p)$ は $L(p)$ の単調増加関数なんで)．

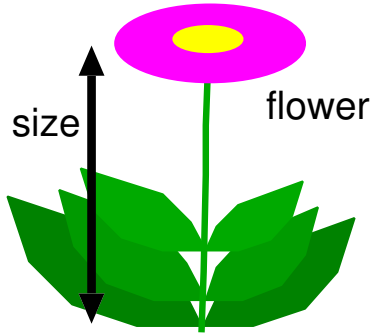
R で計算してみよう．

```
> 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

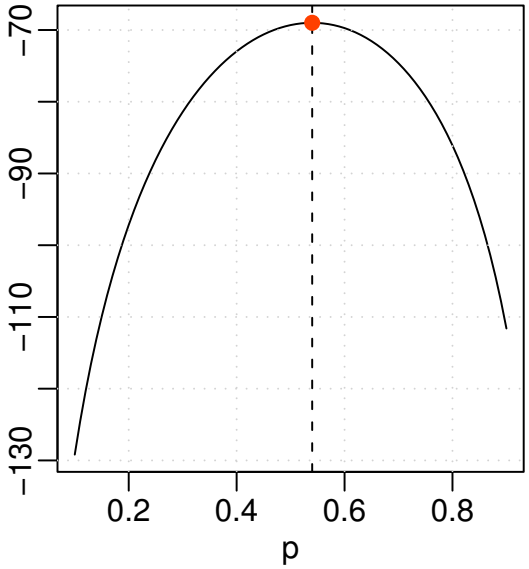
対数尤度

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

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?

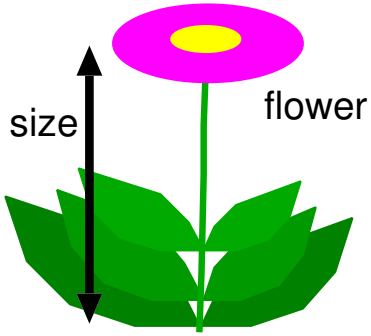
対数尤度 $\log L(p)$ を p で偏微分して

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

$\frac{\partial \log L(p)}{\partial p} = 0$ 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?

$\rightarrow$ ロジスティック (logistic) モデル化

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

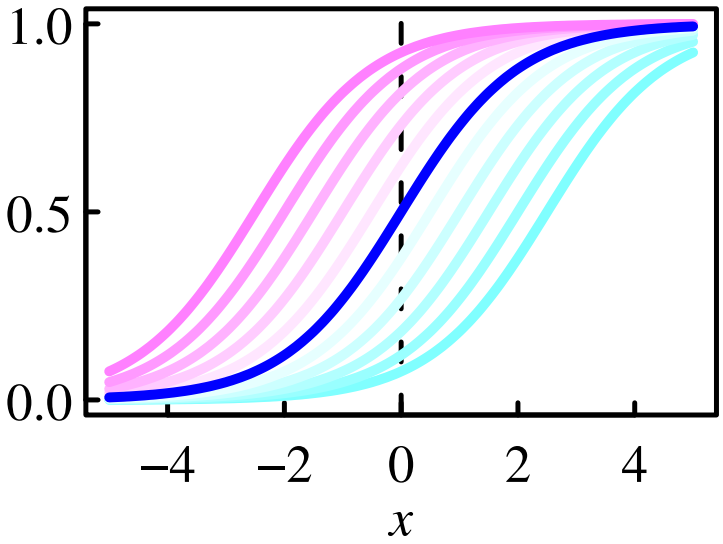
$$
\begin{aligned}
L(a) &= p(a)^{54}(1-p(a))^{46} \\
&= \left\{\frac{1}{1+\exp(-a)}\right\}^{54}\left\{\frac{\exp(-a)}{1+\exp(-a)}\right\}^{46}
\end{aligned}
$$

その「ロジスティック関数」って何なの?

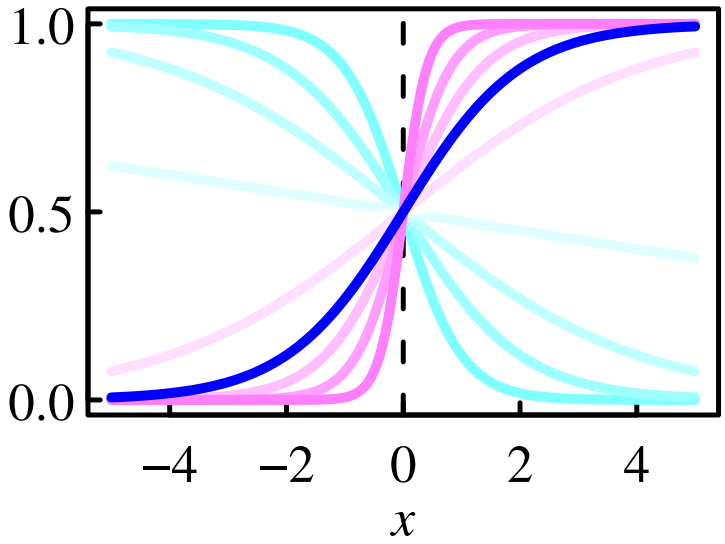
$$p(x) = \frac{1}{1 + \exp(-(a + bx))}$$

($\exp(Z) = \mathrm{e}^Z$ のこと)

a だけ変化させる



b だけ変化させる



つまり $p(a) = \dfrac{1}{1 + \exp(-a)}$ 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

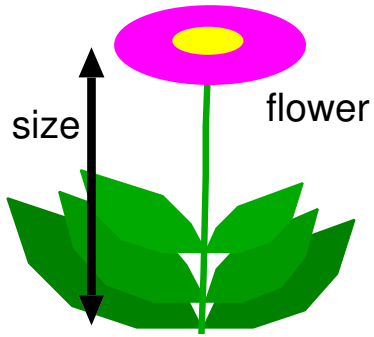
$$\hat{a} = \log\frac{54}{46} \approx 0.16$$ …… これは対数オッズ (log odds)

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よみこむ)

```
> 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, familiy = 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) $\times 1$ で説明しろ
- familiy = 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)

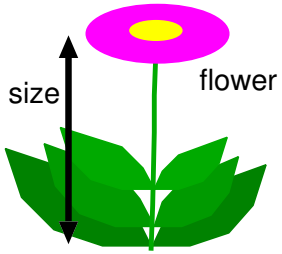
ロジスティックモデルの最尤推定，glm() で (3)

glm() の推定結果を読む

```
> summary(glm(flower ~ 1, familiy = 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

一般化線形モデル (GLM) って論文で見たんですけど……

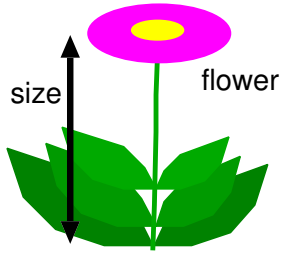


3. さらに強めるロジスティック回帰わざ

説明変数とそのパラメーター推定

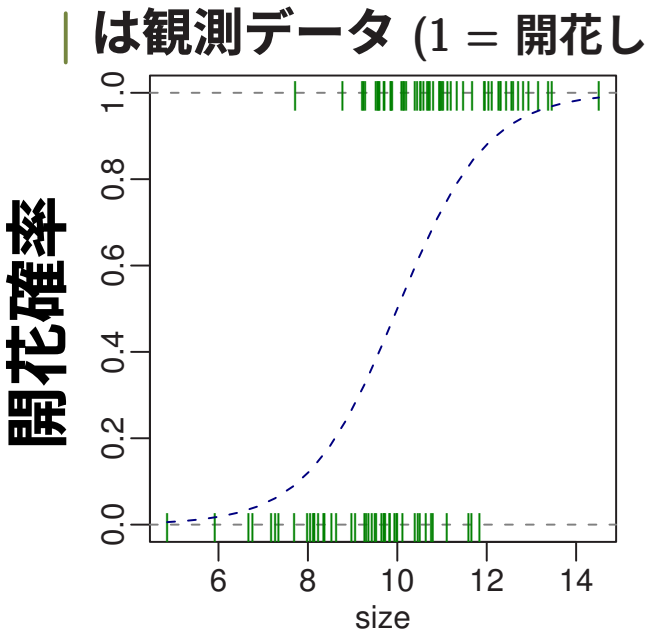
開花確率とサイズの関係を調べたい

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]

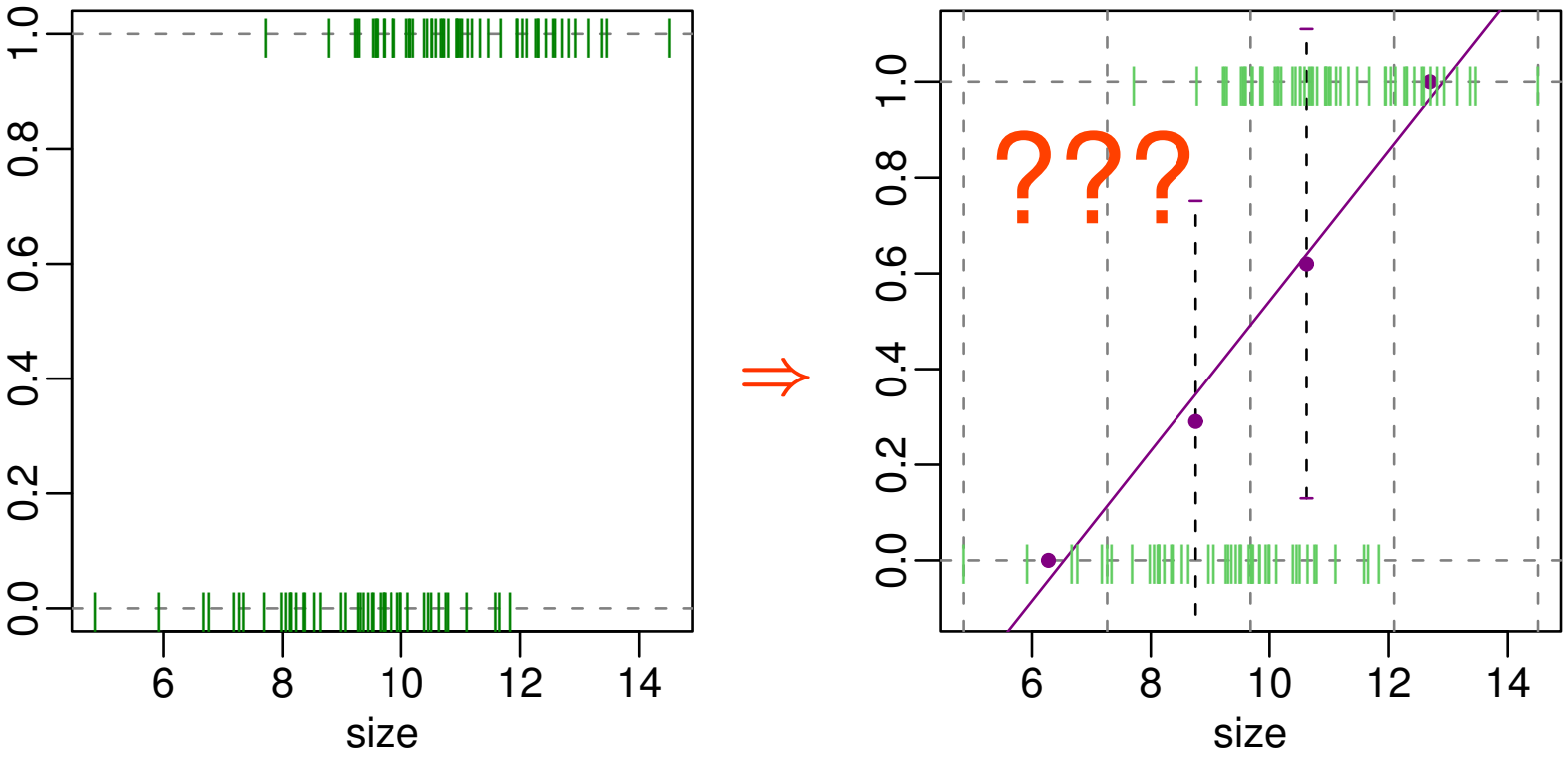
- サイズ (x) が大きいほど開花確率が高い

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

$a = -10$ 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

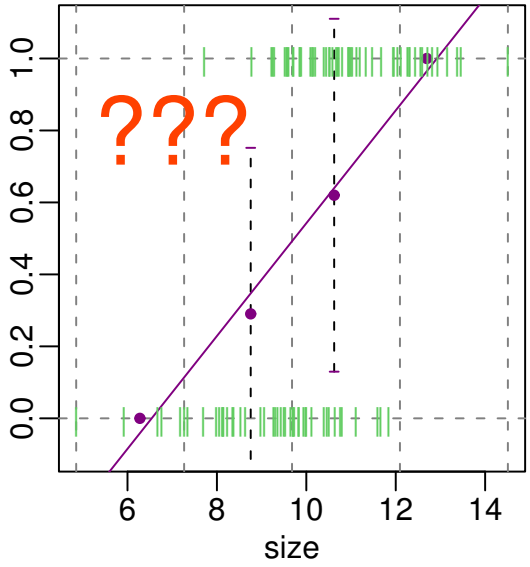
区画はてきとー

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)

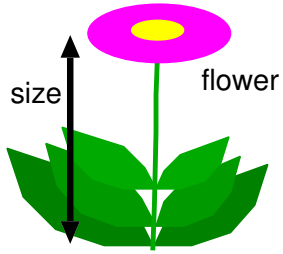
いろいろな確率分布に適用できる R の推定計算手法

	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 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()` なども含まれる

(再掲) 開花確率とサイズの関係を調べたい

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

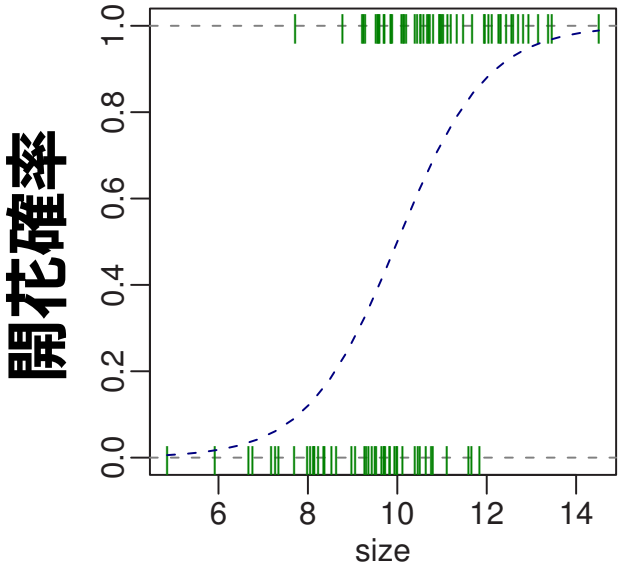
[“神” の立場で知ってるコト]

- 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

| は観測データ ($1 =$ 開花した)



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

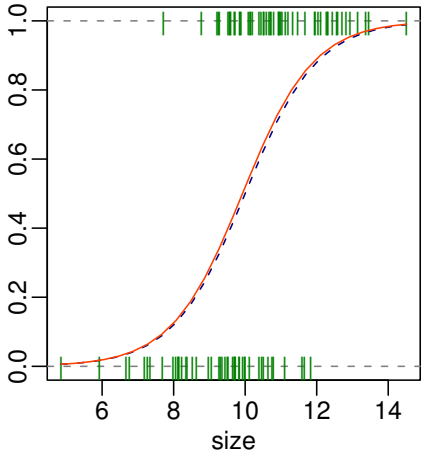
glm() で flower ~ 1 + size モデルのパラメーター推定

```
> summary(glm(flower ~ 1 + size, familiy = 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，よけーな割り算とか使わずに「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

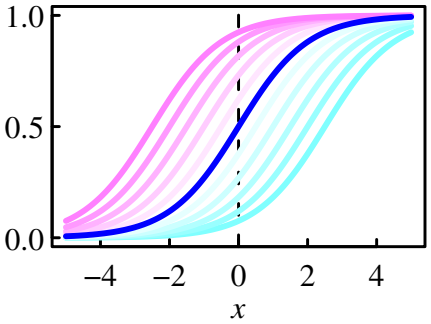
データ解析とはモデリングによる情報圧縮

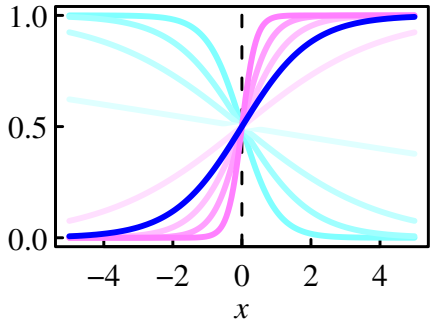
2. 最尤推定法とロジスティック回帰

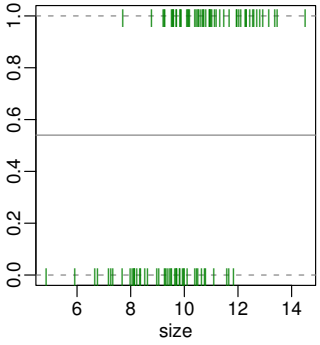
カウントデータは，まず `glm()` で!

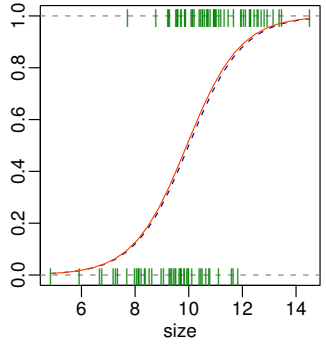
3. さらに強めるロジスティック回帰わざ

割り算値解析しない，データにあわせたばらつき (確率分布) を









次回予告

生態学基礎論 (生物多様性論 II) 2007-01-24

全部で 2 回講義の 2

「個体差」を階層ベイズモデルであつかう

— 個体差・ブロック差の random effects —

「個体差」って何なの?

http://hosho.ees.hokudai.ac.jp/~kubo/stat/2006/