統計モデリングの基礎 (1)

統計モデル・確率分布・最尤推定

久保拓弥 kubo@ees.hokudai.ac.jp

生態学基礎論

2019–01–21

ファイルのダウンロード: http://goo.gl/76c4i
ファイル更新時刻: 2018–12–12 15:29

1. はじめに

とりあえず，全体のながれなど

簡単な自己紹介その他あれこれ

とりあえず簡単な自己紹介: 久保拓弥 (北大・環境科学)

研究: 生態学データの統計モデリング

- 自分ではデータをとらない (野外調査・実験などをやらない) で，他のみなさんのデータ解析をすることが専門です
- これではあまりにも寄生者的なので，ときどきデータ解析に必要な統計モデリングの解説みたいなことをしております……

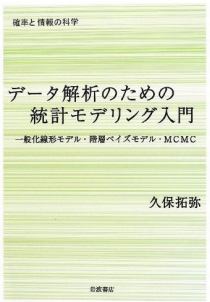
統計モデリングの教科書も書きました!

なんで，そんな本なんか書いたの?!

生態学の統計解析はあまりおもしろくなかった

- 他人の論文の method section を読んで，内容を理解しないまま同じソフトウェアを使って，$p < 0.05$ なら何でも OK といった作業になりがち
- 統計ソフトウェアが何をやっているのかわかっていないので，誤用が多い
- こういう発想は，計算環境が貧弱だった昔の遺物

この本ではブラックボックス統計学として批判

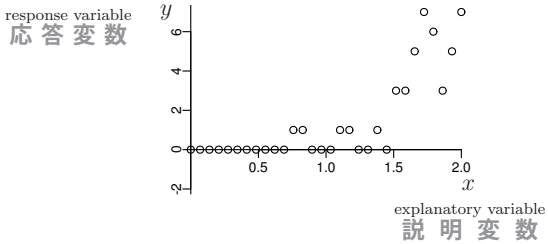


カタチ だけまねをするデータ解析
何がよくないのか?
例をあげて考えてみましょう

suppose that you have a “count data” set ...

架空の例題: 0 個, 1 個, 2 個と数えられるデータ

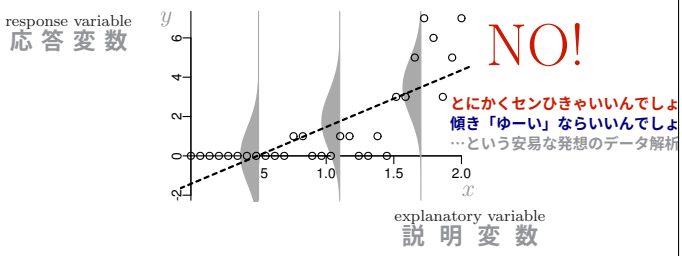
カウントデータ ($y \in \{0, 1, 2, 3, \cdots\}$ なデータ)



- たとえば x は植物個体の大きさ，y はその個体の花数
- 体サイズが大きくなると花数が増えるように見えるが……

“何でもかんでも直線あてはめ” という安易な発想……はギモン

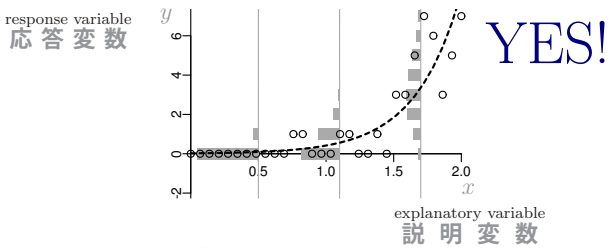
正規分布・恒等リンク関数の統計モデル



- タテ軸のばらつきは「正規分布」なのか?
- y の値は 0 以上なのに ……
- 平均値がマイナス?

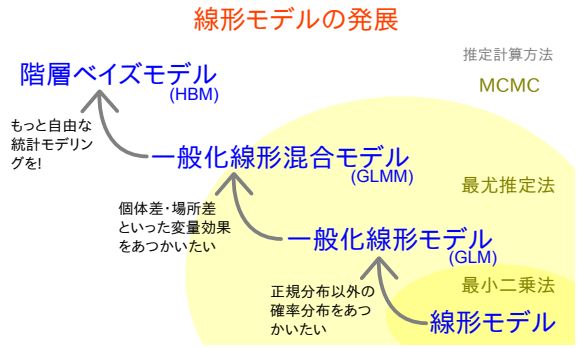
データにあわせた “統計モデル” つかうとマシかもね?

ポアソン分布・対数リンク関数の統計モデル



- タテ軸に対応する「ばらつき」
- 負の値にならない「平均値」
- 正規分布を使ってるモデルよりましだね

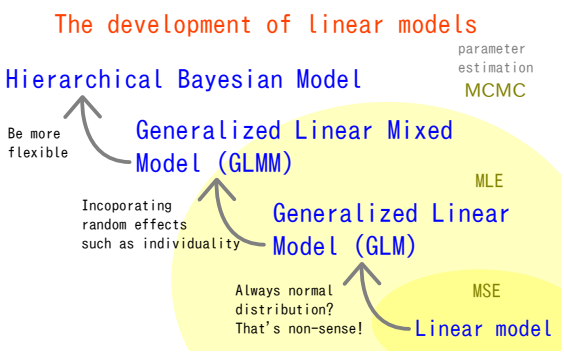
この講義で勉強する統計モデル



ひとことでいうと 「直線あてはめ」をどんどん改善する

statistaical models appeared in the class

この講義であつかう統計モデル



“See the evolution of linear-model family!”

この講義の流れ: 例題を考えながら理解する

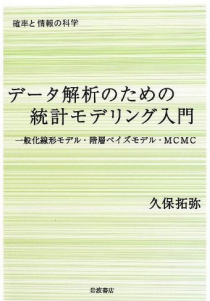
1. 統計モデル・確率分布・最尤推定
2. ポアソン分布の一般化線形モデル (GLM)
3. 二項分布の GLM
4. MCMC と階層ベイズモデル

単純化した例題にそって統計モデルを説明

統計モデルって何?

どんな統計解析においても統計モデルが使用されている

- 観察によってデータ化された現象を説明するために作られる
- 確率分布が基本的な部品であり，これはデータにみられるばらつきを表現する手段である
- データとモデルを対応づける手つづきが準備されていて，モデルがデータにどれぐらい良くあてはまっているかを定量的に評価できる



もくじ

この時間に説明したいこと

例題: 種子数の統計モデリング　まあ，かなり単純な例から始めましょう

1. 例題: 種子数の統計モデリング

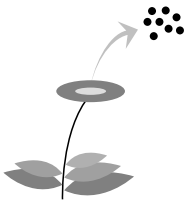
まあ，かなり単純な例から始めましょう

R でデータをあつかいつつ

例題: 種子数の統計モデリング　まあ，かなり単純な例から始めましょう

a simplified data set, easy to understand

この授業では架空植物の架空データをあつかう



理由: よけいなことは考えなくてすむので

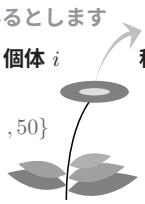
現実のデータはどれも授業で使うには難しすぎる……

例題: 種子数の統計モデリング　まあ，かなり単純な例から始めましょう

number of seeds per plant individual

こんなデータ (架空) があったとしましょう

まあ，なんだかこういうヘンな
植物を調査しているとします



このデータ $\{y_i\}$ がすでに R という統計ソフトウェアに
格納されていた，としましょう

```
> data
 [1] 2 2 4 6 4 5 2 3 1 2 0 4 3 3 3 3 4 2 7 2 4 3 3 3 4
[26] 3 7 5 3 1 7 6 4 6 5 2 4 7 2 2 6 2 4 5 4 5 1 3 2 3
```


例題: 種子数の統計モデリング　まあ，かなり単純な例から始めましょう

apply table() to categorize data

R でデータの様子をながめる


の table() 関数を使って種子数の頻度を調べる

```
> table(data)
 0  1  2  3  4  5  6  7
 1  3 11 12 10  5  4  4
```

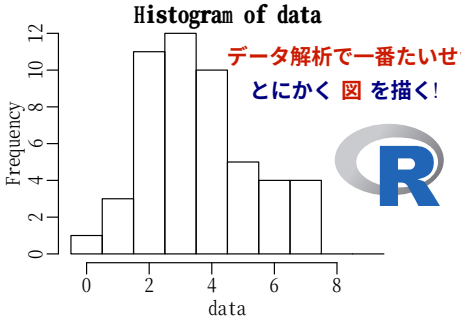
(種子数 5 は 5 個体，種子数 6 は 4 個体 ……)

例題: 種子数の統計モデリング　まあ，かなり単純な例から始めましょう

start with data plotting, always

とりあえずヒストグラムを描いてみる

```
> hist(data, breaks = seq(-0.5, 9.5, 1))
```

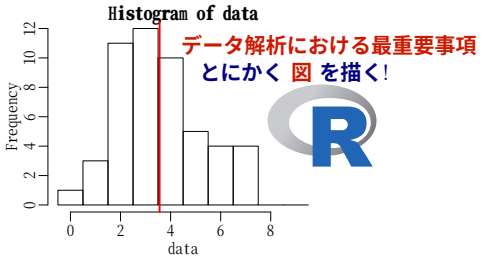


データ解析で一番たいせつなこと
とにかく 図 を描く!

例題: 種子数の統計モデリング　まあ，かなり単純な例から始めましょう

How to evaluate mean value using R?

```
> mean(data)
[1] 3.56
> abline(v = mean(data))
```



データ解析における最重要事項
とにかく 図 を描く!

例題: 種子数の統計モデリング　まあ，かなり単純な例から始めましょう

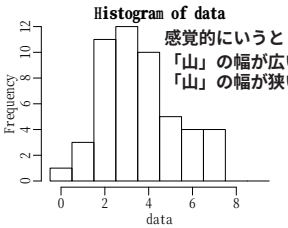
statistics to represent dispersion
「ばらつき」の統計量

あるデータの ばらつき をあらわす標本統計量の例: sample variance 標本分散

```
> var(data)
[1] 2.9861
```

sample standard deviation
標本標準偏差 とは標本分散の平方根 $(\mathrm{SD} = \sqrt{\mathrm{variance}})$

```
> sd(data)
[1] 1.7280
> sqrt(var(data))
[1] 1.7280
```



感覚的にいうと
「山」の幅が広い: 分散が大きい
「山」の幅が狭い: 分散が小さい

データと確率分布の対応　probability distribution, the core of statistical model

2. データと確率分布の対応

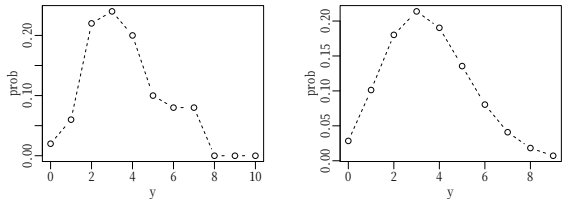
probability distribution, the core of statistical model

確率分布は統計モデルの重要な部品

データと確率分布の対応　probability distribution, the core of statistical model

Empirical VS Theoretical Distributions

統計モデルの部品である 確率分布 には
“データそのまま” な 経験分布 (cf. サイコロ) と
数式で定義される理論的な分布 がある

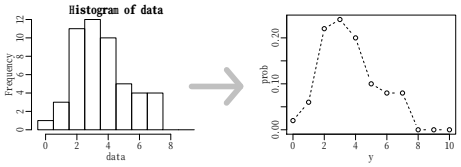


データと確率分布の対応　probability distribution, the core of statistical model

empirical distribution
“データそのまま” な経験分布

```
> data.table <- table(factor(data, levels = 0:10))
> cbind(y = data.table, prob = data.table / 50)
```

	y	prob
0	1	0.02
1	3	0.06
2	11	0.22
3	12	0.24
4	10	0.20
5	5	0.10
6	4	0.08
7	4	0.08
8	0	0.00
9	0	0.00
10	0	0.00



- 確率分布とは 発生する事象 と 発生する確率 の対応づけ
- “たまたま手もとにある” データから “発生確率” を決める確率分布が経験分布

データと確率分布の対応　probability distribution, the core of statistical model

なるほど経験分布は “直感的” かもしれないが……

- データが変わると確率分布が変わる?
- 種子数 $y = \{0, 1, 2, \cdots\}$ となる確率が，個々におたがい無関係に決まる?
- パラメーターは $\{p_0, p_1, p_2, \cdots, p_{99}, p_{100}, \cdots\}$ 無限個ある?

道具として使うには，ちょっと不便かもしれない……

なにか理論的に導出された確率分布のほうが便利ではないか?

- 少数のパラメーターで分布の “カタチ” が決まる
- “なめらかに” 確率が変化する
- いろいろと数理的な道具が準備されている (パラメーター推定方法など)

Mathematical expression of the Poisson distribution

確率分布 (ポアソン分布) を数式で決めてしまう

種子数が y である確率 (probability) は以下のように決まる，と考えている

$$p(y \mid \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}$$

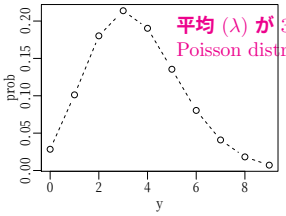
- $y!$ は y の階乗 (factorial) で，たとえば $4!$ は $1 \times 2 \times 3 \times 4$ をあらわしています．
- $\exp(-\lambda) = e^{-\lambda}$ のこと $(e = 2.718\cdots)$
- ここではなぜポアソン分布の確率計算が上のようになるのかは説明しません— まあ，こういうもんだと考えて先に進みましょう

the Poisson distribution

数式で決められたポアソン分布?

とりあえず R で作図してみる

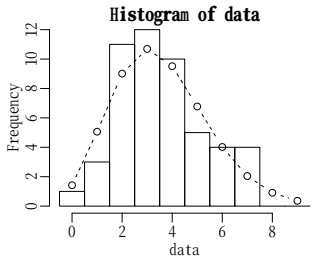
```
> y <- 0:9                                  # これは種子数（確率変数）
> prob <- dpois(y, lambda = 3.56)  # ポアソン分布の確率の計算
> plot(y, prob, type = "b", lty = 2)
```




```
> # cbind で「表」作り
> cbind(y, prob)
    y        prob
1   0 0.02843882
2   1 0.10124222
3   2 0.18021114
4   3 0.21385056
5   4 0.19032700
6   5 0.13551282
7   6 0.08040427
8   7 0.04089132
9   8 0.01819664
10  9 0.00719778
```

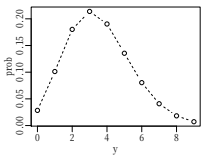

the Poisson distribution represent data?

データとポアソン分布を重ね合わせる




```
> hist(data, seq(-0.5, 8.5, 0.5))      # まずヒストグラムを描き
> lines(y, prob, type = "b", lty = 2) # その「上」に折れ線を描く
```


パラメーター λ はポアソン分布の平均




```
> # cbind で「表」作り
> cbind(y, prob)
    y        prob
1   0 0.02843882
2   1 0.10124222
3   2 0.18021114
4   3 0.21385056
5   4 0.19032700
6   5 0.13551282
7   6 0.08040427
8   7 0.04089132
9   8 0.01819664
10  9 0.00719778
```

- 平均 λ はポアソン分布の唯一のパラメーター
- 確率分布の平均は λ である ($\lambda \geq 0$)
- 分散と平均は等しい: $\lambda =$ 平均 $=$ 分散
- $y \in \{0,\ 1,\ 2,\ \cdots,\ \infty\}$ の値をとり，すべての y について和をとると 1 になる

$$\sum_{y=0}^{\infty} p(y \mid \lambda) = 1$$

どういう場合にポアソン分布を使う?

統計モデルの部品としてポアソン分布が選んだ理由:

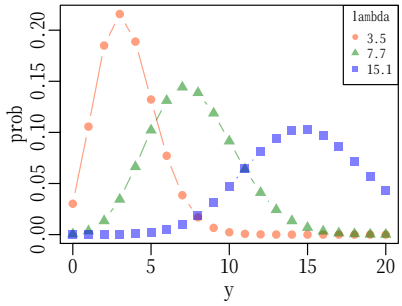
- データに含まれている値 y_i が $\{0,\ 1,\ 2,\ \cdots\}$ といった非負の整数である (カウントデータ (count data) である)
- y_i に下限 (ゼロ) はあるみたいだけど上限はよくわからない
- この観測データでは平均と分散がだいたい等しい ($\text{mean} \approx \text{variance}$)
  - このだいたい等しいがあやしいのだけど，まあ気にしないことにしましょう

λ changes the shape of distribution

ポアソン分布の λ を変えてみる

$$p(y \mid \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}$$

λ は平均(mean)をあらわすパラメーター



maximum likelihood estimation of parameter λ

3. ポアソン分布のパラメーターの最尤推定

もっとももっともらしい推定?

"fitting" = "parameter estimation"

「あてはめる」ことは推定すること

尤度 (likelihood) とは何か?

- 最尤推定法(maximum likelihood estimation)では，尤度という**あてはまりの良さ**をあらわす統計量に着目
- 尤度は**データが得られる確率**をかけあわせたもの
- この例題の場合，パラメーター λ を変えると尤度が変わる
- もっとも「あてはまり(goodness of fit)」が良くなる λ を見つけたい
- たとえば，いまデータが 3 個体ぶん，たとえば，$\{y_1,\ y_2,\ y_3\} = \{2,\ 2,\ 4\}$，これだけだった場合，尤度はだいたい $0.180 \times 0.180 \times 0.19 = 0.006156$ といった値になる

likelihood $L(\lambda)$ depends on the value of mean, λ

尤度 $L(\lambda)$ はパラメーター λ の関数

この例題の尤度:

$$\begin{aligned} L(\lambda) &= (y_1\text{が } 2 \text{ である確率}) \times (y_2\text{が } 2 \text{ である確率}) \\ &\quad \times \cdots \times (y_{50}\text{が } 3 \text{ である確率}) \\ &= p(y_1 \mid \lambda) \times p(y_2 \mid \lambda) \times p(y_3 \mid \lambda) \times \cdots \times p(y_{50} \mid \lambda) \\ &= \prod_i p(y_i \mid \lambda) = \prod_i \frac{\lambda^{y_i} \exp(-\lambda)}{y_i!}, \end{aligned}$$

evaluate not likelihood, but log likelihood!

尤度はしんどいので対数尤度を使う

尤度は確率 (あるいは確率密度) の積であり，あつかいがふべん (大量のかけ算!)

そこで，パラメーターの最尤推定では，**対数尤度関数** (log likelihood function) を使う

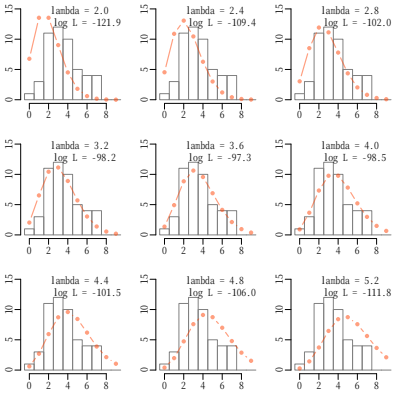
$$\log L(\lambda) = \sum_i \left(y_i \log \lambda - \lambda - \sum_k^{y_i} \log k \right)$$

対数尤度 $\log L(\lambda)$ の最大化は尤度 $L(\lambda)$ の最大化になるから

まずは，平均をあらわすパラメーター λ を変化させていったときに，ポアソン分布のカタチと対数尤度がどのように変化するのかを調べてみましょう

λ changes the log likelihood, i.e., goodness of fit

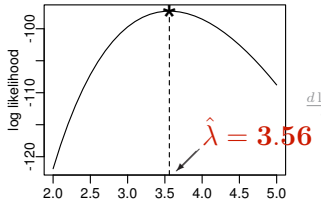
λ を変えるとあてはまりの良さが変わる



seek the maximum likelihood estimate, $\hat{\lambda}$

対数尤度を最大化する $\hat{\lambda}$ をさがす

対数尤度 $\log L(\lambda) = \sum_i \left(y_i \log \lambda - \lambda - \sum_k^{y_i} \log k\right)$



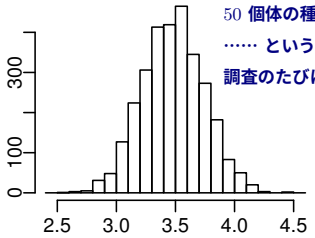
$\frac{d \log L}{d\lambda} = 0$ より

- **最尤推定量** (ML estimator): $\sum_i y_i/50$ 標本平均値!
- **最尤推定値** (ML estimate): $\hat{\lambda} = 3.56$ ぐらい

no one knows "the true λ" based on finite size data

最尤推定を使っても真の λは見つからない

真の λ が 3.5 の場合



50 個体の種子数を調べる
…… ということを 3000 回くりかえし
調査のたびに $\hat{\lambda}$ を最尤推定した

データは有限なので真の λ はわからない

一般化線形モデルって何だろう?

Generalized Linear Model

一般化線形モデル (GLM)

- **ポアソン回帰** (Poisson regression)
- **ロジスティック回帰** (logistic regression)
- **直線回帰** (linear regression)
- ……

4. ポアソン回帰の例題: 架空植物の種子数データ

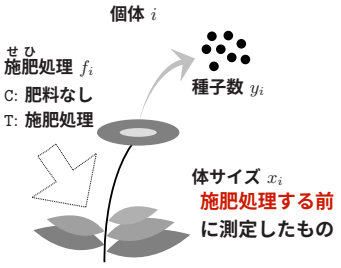
植物個体の属性，あるいは実験処理が種子数に影響?

まずはデータの概要を調べる

body size x and fertilization f change seed number y?

個体サイズと実験処理の効果を調べる例題

- response variable 応答変数: seed number 種子数 $\{y_i\}$
- explanatory variable 説明変数:
  - body size 体サイズ $\{x_i\}$
  - fertilization 施肥処理 $\{f_i\}$



sample size 標本数

- control 無処理 (f_i = C): 50 sample ($i \in \{1, 2, \cdots 50\}$)
- treated 施肥処理 (f_i = T): 50 sample ($i \in \{51, 52, \cdots 100\}$)

Reading data file

データファイルを読みこむ

data3a.csv は CSV (comma separated value) format file なので，R で読みこむには以下のようにする:

```
> d <- read.csv("data3a.csv")
```

データは d と名付けられた data frame (表みたいなもの) に格納される

とりあえず data frame d を表示

```
> d
     y     x  f
1    6  8.31  C
2    6  9.44  C
3    6  9.50  C
... (中略) ...
99   7 10.86  T
100  9  9.97  T
```

data frame d を調べる: 連続値と整数値

```
> d$x
  [1]  8.31  9.44  9.50  9.07 10.16  8.32 10.61 10.06
  [9]  9.93 10.43 10.36 10.15 10.92  8.85  9.42 11.11
... (中略) ...
 [97]  8.52 10.24 10.86  9.97

> d$y
  [1]  6  6  6 12 10  4  9  9  9 11  6 10  6 10 11  8
 [17]  3  8  5  5  4 11  5 10  6  6  7  9  3 10  2  9
... (中略) ...
 [97]  6  8  7  9
```


data frame d を調べる: “因子型” のデータ

施肥処理の有無をあらわす f 列はちょっと様子がちがう

```
> d$f
  [1] C C C C C C C C C C C C C C C C C C C C C C C C C
 [26] C C C C C C C C C C C C C C C C C C C C C C C C C
 [51] T T T T T T T T T T T T T T T T T T T T T T T T T
 [76] T T T T T T T T T T T T T T T T T T T T T T T T T
Levels: C T
```

data type: factor
因子型データ: いくつかの levels 水準をもつデータ
ここでは C と T の 2 levels 水準

R のデータのクラスとタイプ

data type and class

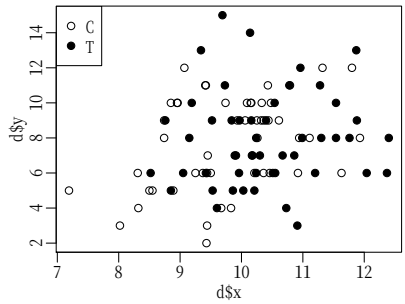
```
> class(d) # d は data.frame クラス
[1] "data.frame"
> class(d$y) # y 列は整数だけの integer クラス
[1] "integer"
> class(d$x) # x 列は実数も含むので numeric クラス
[1] "numeric"
> class(d$f) # そして f 列は factor クラス
[1] "factor"
```


data frame の summary()

```
> summary(d)
       y               x          f
 Min.   : 2.00   Min.   : 7.190   C:50
 1st Qu.: 6.00   1st Qu.: 9.428   T:50
 Median : 8.00   Median :10.155
 Mean   : 7.83   Mean   :10.089
 3rd Qu.:10.00   3rd Qu.:10.685
 Max.   :15.00   Max.   :12.400
```


データはとにかく図示する!

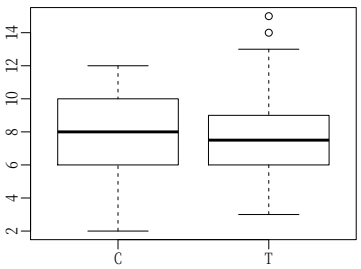
```
> plot(d$x, d$y, pch = c(21, 19)[d$f])
> legend("topleft", legend = c("C", "T"), pch = c(21, 19))
```



散布図

施肥処理 f を横軸とした図

```
> plot(d$f, d$y)
```



箱ひげ図

how to specify GLM
5. GLM の詳細を指定する

probability distribution, linear predictor and link function
確率分布・線形予測子・リンク関数

ポアソン回帰では log link **関数を使うのが便利**

how to specify GLM
一般化線形モデルを作る

Generalized Linear Model
一般化線形モデル (GLM)

probability distribution
- **確率分布は**?

linear predictor
- **線形予測子は**?

link function
- **リンク関数は**?

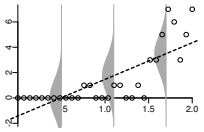
how to specify linear regression model, a GLM
GLM のひとつである**直線回帰モデル**を指定する

直線回帰のモデル

probability distribution Gaussian distribution
- **確率分布**: **正規分布**
- **線形予測子**: e.g., $\beta_1 + \beta_2 x_i$

直線の式: (**切片**) + (**傾き**) $\times x_i$

link function identity link function
- **リンク関数**: **恒等リンク関数**



結果 ← 原因 (かも?) を表現する線形モデル

- **結果**: **応答変数** (response variable)
- **原因**: **説明変数** (explanatory variable)
- **線形予測子** (linear predictor):

(**応答変数の平均**) = **定数** (**切片**, intercept)
+ (**係数** 1) × (**説明変数** 1)
+ (**係数** 2) × (**説明変数** 2)
+ (**係数** 3) × (**説明変数** 3)
+ ···

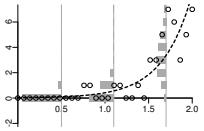
how to specify Poisson regression model, a GLM
GLM のひとつである**ポアソン回帰**モデルを指定する

ポアソン回帰のモデル

probability distribution Poisson distribution
- **確率分布**: **ポアソン分布**

linear predictor
- **線形予測子**: e.g., $\beta_1 + \beta_2 x_i$

link function log link function
- **リンク関数**: **対数リンク関数**



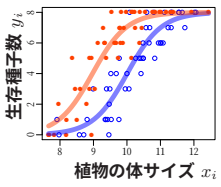
how to specify logistic regression model, a GLM
GLM のひとつである**logistic 回帰モデル**を指定する

ロジスティック回帰のモデル

probability distribution binomial distribution
- **確率分布**: **二項分布**

linear predictor
- **線形予測子**: e.g., $\beta_1 + \beta_2 x_i$

link function
- **リンク関数**: logit **リンク関数**

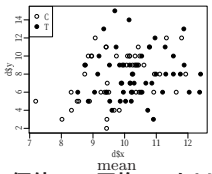


R で一般化線形モデル (GLM) の推定を……

	probability distribution 確率分布	random number generation 乱数発生	GLM fitting GLM あてはめ
(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() in library(MASS)
(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 は直線回帰・重回帰・分散分析・ポアソン回帰・ロジスティック回帰その他の「よせあつめ」と考えてもよいかも

さてさて，種子数の例題にもどって



seed number y_i follows the Poisson distribution

種子数 y_i は平均 λ_i のポアソン分布にしたがう としましょう

$$p(y_i \mid \lambda_i) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}$$

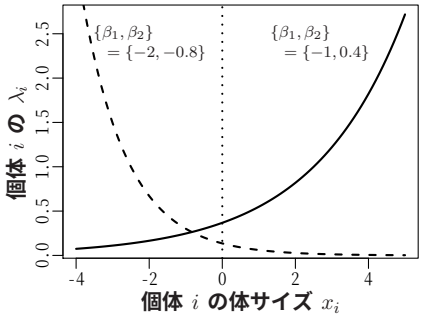
個体 i の平均 λ_i を以下のようにおいてみたらどうだろう……?

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i)$$

- β_1 と β_2 は係数 (coefficient) (パラメーター (parameter))
- x_i は個体 i の体サイズ (body size)，f_i はとりあえず無視 (no f_i, for simplicity)

exponential function
指数関数ってなんだっけ?

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i)$$



GLM のリンク関数と線形予測子 ← (直線の式)

個体 i の平均 (mean) λ_i

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i)$$

$$\Updownarrow$$

log link function = linear predictor

$$\log(\lambda_i) = \beta_1 + \beta_2 x_i$$

log link function = linear predictor

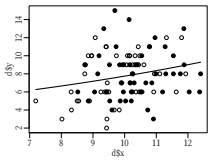
$$\log(\text{平均}) = \text{線形予測子}$$

log リンク関数とよばれる理由は，上のようになっているから

a statistical model for this example
この例題のための統計モデル

ポアソン回帰のモデル

- probability distribution 確率分布: Poisson distribution ポアソン分布
- linear predictor 線形予測子: $\beta_1 + \beta_2 x_i$
- link function リンク関数: log link function 対数リンク関数



6. R で GLM のパラメーターを推定

あてはまりの良さは対数尤度関数で評価

推定計算はコンピューターにおまかせ

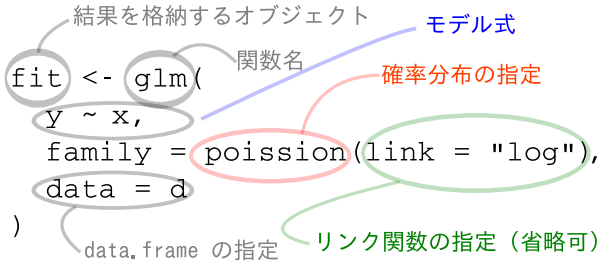
glm() 関数の指定 (function)

```
> d
    y     x  f
1   6  8.31  C
2   6  9.44  C
3   6  9.50  C
... (中略) ...
99  7 10.86  T
100 9  9.97  T
```

Is that all?
これだけ!

```
> fit <- glm(y ~ x, data = d, family = poisson)
```


glm() 関数の指定の意味



- モデル式 (線形予測子 z): どの説明変数を使うか?
- link 関数: z と応答変数 (y) 平均値 の関係は?
- family: どの確率分布を使うか?

glm() 関数の出力 (output)

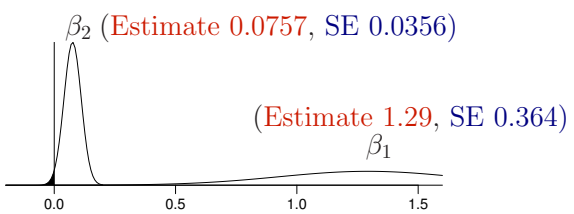
```
> fit <- glm(y ~ x, data = d, family = poisson)

all:  glm(formula = y ~ x, family = poisson, data = d)

Coefficients:
(Intercept)            x
     1.2917       0.0757

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:^^I    89.5
Residual Deviance: 85 ^^IAIC: 475
```


glm() 関数のくわしい出力

```
> summary(fit)
Call:
glm(formula = y ~ x, family = poisson, data = d)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.368  -0.735  -0.177   0.699   2.376

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.2917     0.3637    3.55  0.00038
x             0.0757     0.0356    2.13  0.03358
```

…… (以下，省略) ……

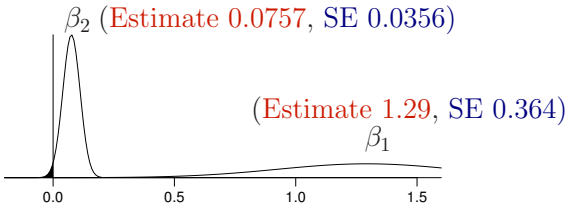
推定値と標準誤差のいめーじ (かなりいいかげんな説明)



- 確率 p は ゼロからの距離 をあらわしている
- p がゼロに近いほど 推定値 $\hat{\beta}$ はゼロから離れている
- p が 0.5 に近いほど 推定値 $\hat{\beta}$ はゼロに近い

(注: 頻度主義的な信頼区間の正しい解釈はもっとめんどくさい)

推定値と標準誤差のいめーじ (何がめんどくさいの?)

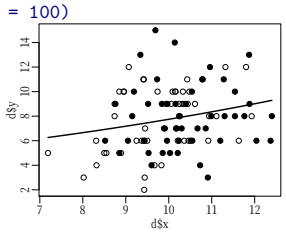


- 区間 95% 内に「ゼロ」があるとしよう → 「だから何?」
- 多数のパラメーターがある場合には?
- 授業の後半であつかうベイズ統計モデルでの解釈は 簡単 ……になるはず……

R で GLM のパラメーターを推定 あてはまりの良さは対数尤度関数で評価

model prediction
モデルの予測

```
> fit <- glm(y ~ x, data = d, family = poisson)
...
Coefficients:
(Intercept)            x
     1.2917       0.0757

> plot(d$x, d$y, pch = c(21, 19)[d$f]) # data
> xp <- seq(min(d$x), max(d$x), length = 100)
> lines(xp, exp(1.2917 + 0.0757 * xp))
```



the figure shows the relationship
ここでは観測データと予測の関係
between model prediction and data
を見ているだけ，なのだが

処理をした・しなかった 効果も統計モデルに入れる GLM の因子型説明変数

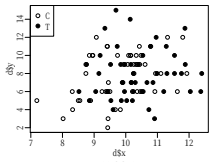
7. 処理をした・しなかった 効果も統計モデルに入れる

factor type
GLM の因子型説明変数

数量型 + 因子型 という組み合わせで

処理をした・しなかった 効果も統計モデルに入れる GLM の因子型説明変数

incorporate the fertilization effects in GLM
肥料の効果 f_i もいれましょう



seed number y_i follows the Poisson distribution
種子数 y_i は平均 λ_i のポアソン分布にしたがう としましょう

$$p(y_i \mid \lambda_i) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}$$

個体 i の平均 λ_i を次のようにする

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i + \beta_3 d_i)$$

- β_3 は施肥処理の効果 の 係 数 (fertilization effects, coefficient)
- f_i のダミー変数 (dummy variable)

$$d_i = \begin{cases} 0 & (f_i = \mathtt{C} \text{ の場合}) \\ 1 & (f_i = \mathtt{T} \text{ の場合}) \end{cases}$$

処理をした・しなかった 効果も統計モデルに入れる GLM の因子型説明変数

glm(y ∼ x + f, ...) の出力 (output)

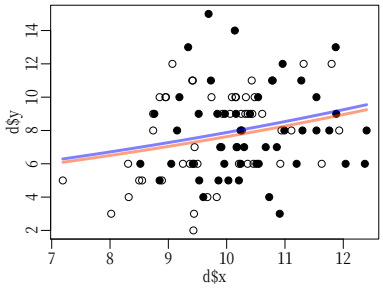
```
> summary(glm(y ~ x + f, data = d, family = poisson))
...(略)...

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.2631     0.3696    3.42  0.00063
x             0.0801     0.0370    2.16  0.03062
fT           -0.0320     0.0744   -0.43  0.66703
```

……（以下，省略）……

処理をした・しなかった 効果も統計モデルに入れる GLM の因子型説明変数

model prediction
x + f モデルの予測

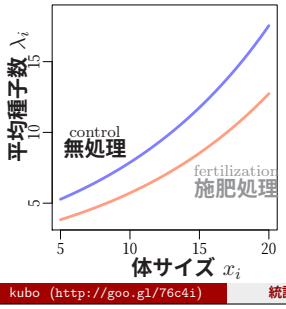
```
> plot(d$x, d$y, pch = c(21, 19)[d$f]) # data
> xp <- seq(min(d$x), max(d$x), length = 100)
> lines(xp, exp(1.2631 + 0.0801 * xp), col = "blue", lwd = 3) # C
> lines(xp, exp(1.2631 + 0.0801 * xp - 0.032), col = "red", lwd = 3) # T
```



処理をした・しなかった 効果も統計モデルに入れる GLM の因子型説明変数

multiple explanatory variables
複数の説明変数をいれた場合の統計モデル

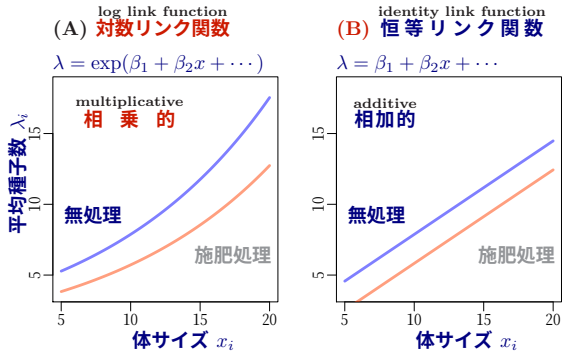
- $f_i = \mathtt{C}$: $\lambda_i = \exp(1.26 + 0.0801 x_i)$
- $f_i = \mathtt{T}$: $\lambda_i = \exp(1.26 + 0.0801 x_i - 0.032)$
  $= \exp(1.26 + 0.0801 x_i) \times \exp(-0.032)$



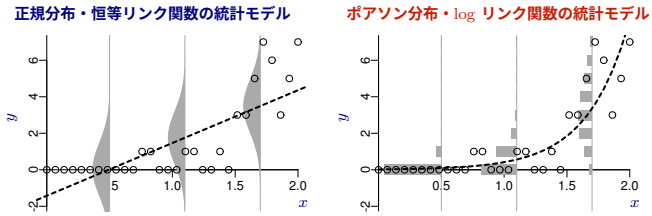
施肥効果である $\exp(-0.032)$ は**かけ算**できくことに注意!

model interpretation depends on link function

リンク関数が違うとモデルの解釈が異なる

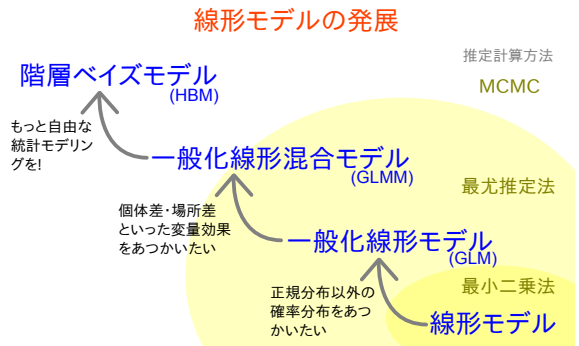


GLM: 適切な 確 率 分 布 とリンク関数を選ぶ



statistaical models appeared in the class

この講義であつかう統計モデルたち



データの特徴にあわせて線形モデルを改良・発展させる

8. “N 個のうち k 個が生きてる” タイプのデータ

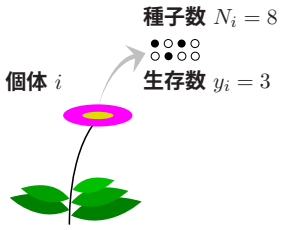
上限のあるカウントデータ

ポアソン分布ではなく二項分布で

example seed survivorship, again

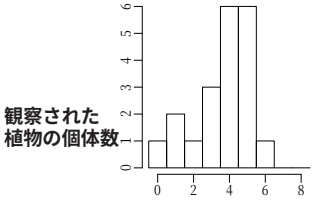
例 題： 植物の種子の生存確率

- 架空植物の種子の生存を調べた
- 種子: 生きていれば発芽する
  - どの個体でも 8 個 の種子を調べた
- 生存確率: ある種子が生きている確率
- データ: 植物 20 個体，合計 160 種子の生存の有無を調べた
- 73 種子が生きていた — このデータを統計モデル化したい



たとえばこんなデータが得られたとしましょう

個体ごとの生存数	0	1	2	3	4	5	6	7	8
観察された個体数	1	2	1	3	6	6	1	0	0



これは個体差なしの均質な集団

生存確率 q と 二 項 分 布 (binomial distribution) の関係

- 生存確率を推定するために**二項分布** という確率分布を使う
- 個体 i の N_i 種子中 y_i 個が生存する確率

$$p(y_i \mid q) = \binom{N_i}{y_i} q^{y_i}(1-q)^{N_i - y_i},$$

- ここで仮定していること
  - 個体差はない
  - つまり すべての個体で同じ生存確率 q

尤度 (ゆうど): 20 個体ぶんのデータが観察される確率

- 観察データ $\{y_i\}$ が確定しているときに
- パラメータ q は値が自由にとりうると考える
- 尤 度 (likelihood) は 20 個体ぶんのデータが得られる確率の積，パラメータ q の関数として定義される

$$L(q|\{y_i\}) = \prod_{i=1}^{20} p(y_i \mid q)$$

個体ごとの生存数	0	1	2	3	4	5	6	7	8
観察された個体数	1	2	1	3	6	6	1	0	0

対数尤度方程式と最尤推定

- この尤度 $L(q \mid$ データ$)$ を最大化するパラメータの推定量 $\hat{q}$ を計算したい
- 尤度を対数尤度になおすと

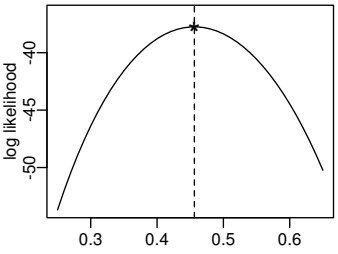
$$\log L(q \mid \text{データ}) = \sum_{i=1}^{20} \log \binom{N_i}{y_i} + \sum_{i=1}^{20} \{y_i \log(q) + (N_i - y_i)\log(1-q)\}$$

- この対数尤度を最大化するように未知パラメーター q の値を決めてやるのが**最尤推定**

最 尤 推 定 (maximum likelihood estimation; MLE) とは何か

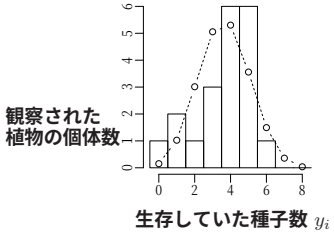
- 対数尤度 $L(q \mid$ データ$)$ が最大になるパラメーター q の値をさがしだすこと
- 対数尤度 $\log L(q \mid$ データ$)$ を q で偏微分して 0 となる $\hat{q}$ が対数尤度最大
  $\partial \log L(q \mid \text{データ})/\partial q = 0$
- 生存確率 q が全個体共通の場合の最尤推定量・最尤推定値は

$$\hat{q} = \frac{\text{生存種子数}}{\text{調査種子数}} = \frac{73}{160} = 0.456 \text{ ぐらい}$$



二項分布で説明できる 8 種子中 y_i 個の生存

$\hat{q} = 0.46$ なので $\binom{8}{y} 0.46^y\ 0.54^{8-y}$

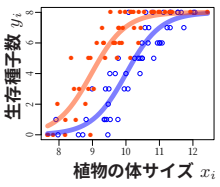


how to specify logistic regression model, a GLM

GLM のひとつである**logistic 回帰モデル**を指定する

ロジスティック回帰のモデル

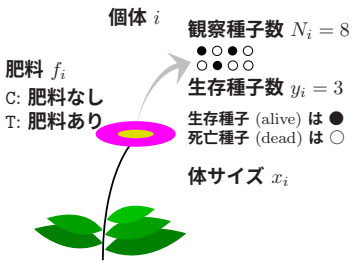
- 確率分布 (probability distribution): **二 項 分 布** (binomial distribution)
- **線形予測子** (linear predictor): e.g., $\beta_1 + \beta_2 x_i$
- **リンク関数** (link function): logit **リンク関数**



N 個のうち y 個で何かが生起した…データ

8 個の種子のうち y 個が **発芽可能** だった!

…… という **"わりあい"** みたいなデータ



Reading data file

例題のデータファイル

data4a.csv は CSV (comma separated value) format file なので，R で読みこむには以下のようにする:

```
> d <- read.csv("data4a.csv")
```

or

```
> d <- read.csv(
+ "http://hosho.ees.hokudai.ac.jp/~kubo/stat/2014/Fig/binomial/data4a.csv")
```

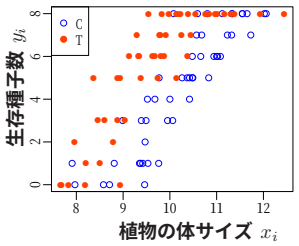
データは d **と名付けられた** data frame (**表みたいなもの**) **に格納される**

data frame d を調べる

```
> summary(d)
       N          y               x          f
 Min.   :8   Min.   :0.00   Min.   : 7.660   C:50
 1st Qu.:8   1st Qu.:3.00   1st Qu.: 9.338   T:50
 Median :8   Median :6.00   Median : 9.965
 Mean   :8   Mean   :5.08   Mean   : 9.967
 3rd Qu.:8   3rd Qu.:8.00   3rd Qu.:10.770
 Max.   :8   Max.   :8.00   Max.   :12.440
```


まずはデータを図にしてみる

```
> plot(d$x, d$y, pch = c(21, 19)[d$f])
> legend("topleft", legend = c("C", "T"), pch = c(21, 19))
```



fertilization effective
今回は施肥処理 がきいている?

logistic regression

9. ロジスティック回帰の部品

二項分布 binomial distribution **と** logit link function

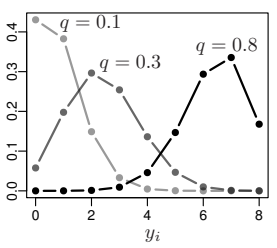
binomial distribution

二 項 分 布 : N 回のうち y 回，となる確率

$$p(y \mid N, q) = \binom{N}{y} q^y (1-q)^{N-y}$$

$\binom{N}{y}$ は N 個の観察種子の中から y 個の生存種子を選びだす場合の数

確率 $p(y_i \mid 8, q)$

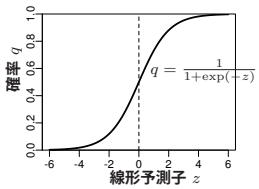


logistic curve
ロジスティック曲線とはこういうもの

ロジスティック関数の関数形 (z_i: 線形予測子, e.g. $z_i = \beta_1 + \beta_2 x_i$)

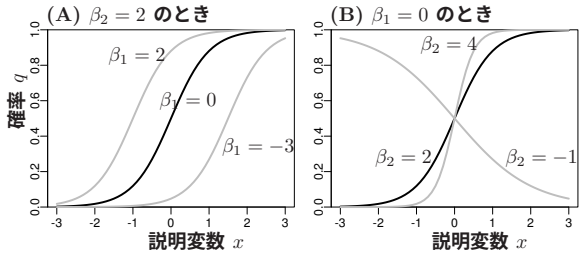
$$q_i = \mathrm{logistic}(z_i) = \frac{1}{1+\exp(-z_i)}$$

```
> logistic <- function(z) 1 / (1 + exp(-z)) # 関数の定義
> z <- seq(-6, 6, 0.1)
> plot(z, logistic(z), type = "l")
```



パラメーターが変化すると……

黒い曲線は $\{\beta_1, \beta_2\} = \{0, 2\}$**.** (A) $\beta_2 = 2$ **と固定して** β_1 **を変化させた場合.** (B) $\beta_1 = 0$ **と固定して** β_2 **を変化させた場合.**



パラメーター $\{\beta_1, \beta_2\}$ **や説明変数** x **がどんな値をとっても確率** q **は** $0 \leq q \leq 1$ **となる便利な関数**

logit link function

- logistic **関数**

$$q = \frac{1}{1+\exp(-(\beta_1 + \beta_2 x))} = \mathrm{logistic}(\beta_1 + \beta_2 x)$$

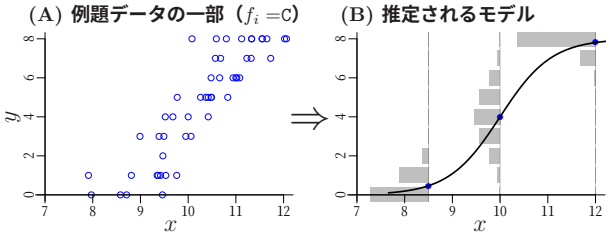
- logit **変換**

$$\mathrm{logit}(q) = \log \frac{q}{1-q} = \beta_1 + \beta_2 x$$

logit は logistic の逆関数, logistic は logit の逆関数

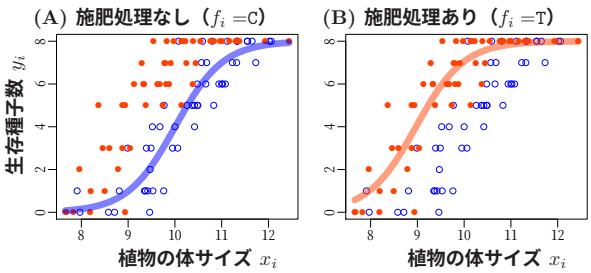
logit is the inverse function of logistic function, vice versa

R でロジスティック回帰 — β_1 と β_2 の最尤推定




```
> glm(cbind(y, N - y) ~ x + f, data = d, family = binomial)
...
Coefficients:
(Intercept)           x          fT
    -19.536       1.952       2.022
```


統計モデルの予測: 施肥処理によって応答が違う



この講義の流れ: 例題を考えながら理解する

1. **統計モデル・確率分布・最尤推定**
2. **ポアソン分布の一般化線形モデル** (GLM)
3. **二項分布の** GLM
4. MCMC **と階層ベイズモデル**

単純化した例題にそって統計モデルを説明