GLMM の紹介

GLM → GLMM → 階層ベイズモデル

久保拓弥 kubo@ees.hokudai.ac.jp, @KuboBook

日本社会心理学会 春セミナー http://goo.gl/5aNIZz

2015–03–25

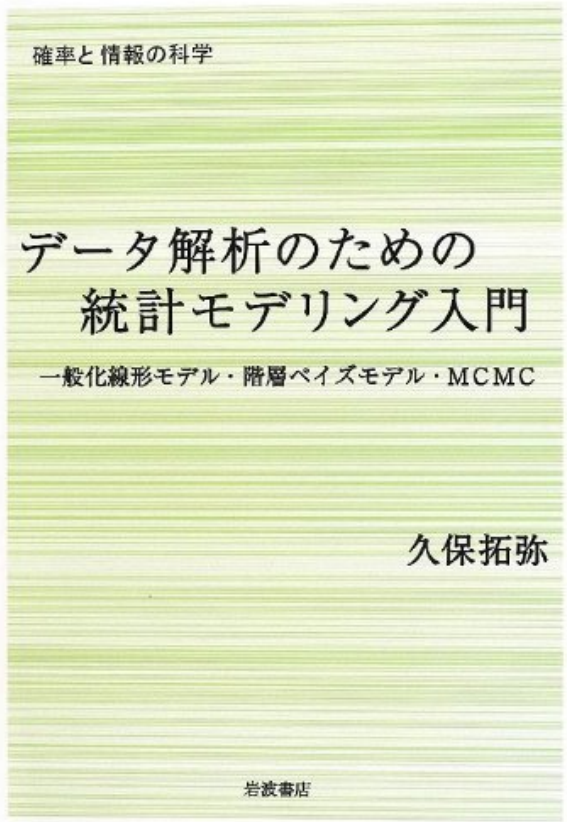
ファイル更新時刻: 2015–03–25 15:41

“みどりぼん” の紹介: 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!

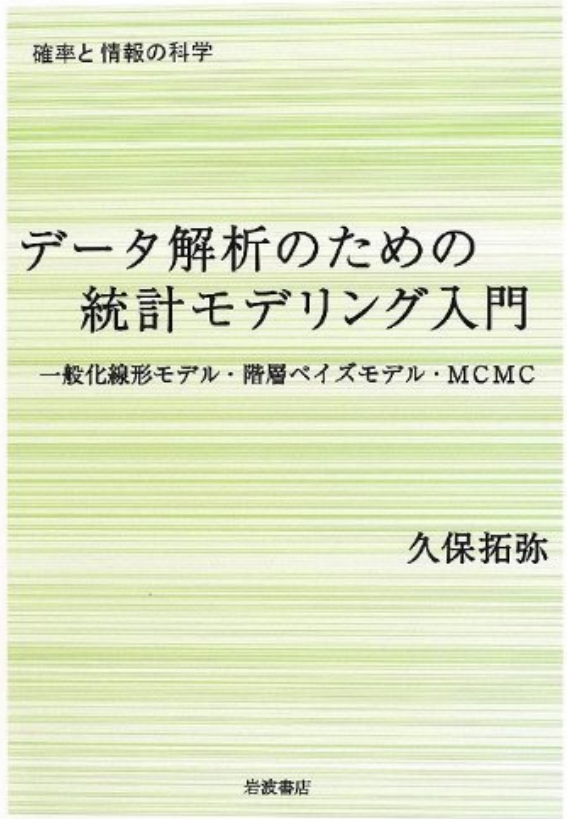


なんで，そんな本なんか書いたの?!

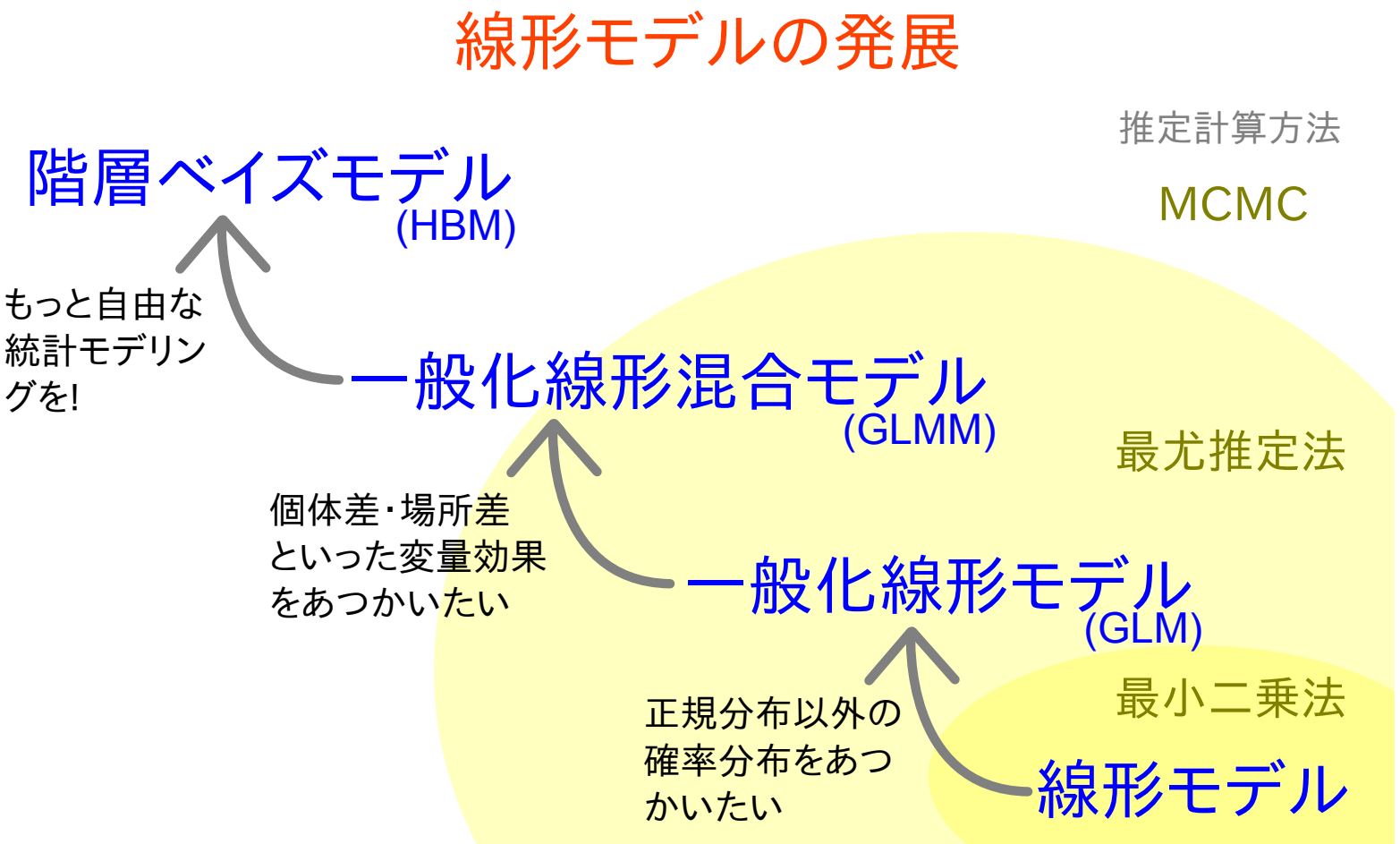
生態学の統計解析はあまりおもしろくなかった

- 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
- この本は “統計モデリング” を意識したデータ解析のススめ

この本ではブラックボックス統計学として批判



“統計モデリング入門” に登場する統計モデル



データの特徴にあわせて線形モデルを改良・発展させる

Q. 今日のハナシは “GLMM の紹介” なのに
なぜ GLM から始めるのか?

A. “GLMM よくわからない” というヒトは
じつは GLM も**よくわかっていない**から

Q. 今日のハナシは “GLMM の紹介” なのに
なぜ階層ベイズモデル (HBM) なんかも
説明するのか?

A. “研究の道具” として使うためには
GLMM を**ベイズモデル化**した
HBM が必要になるから
(あとで説明)

今日のハナシ: いずれも例題 driven なかんじで

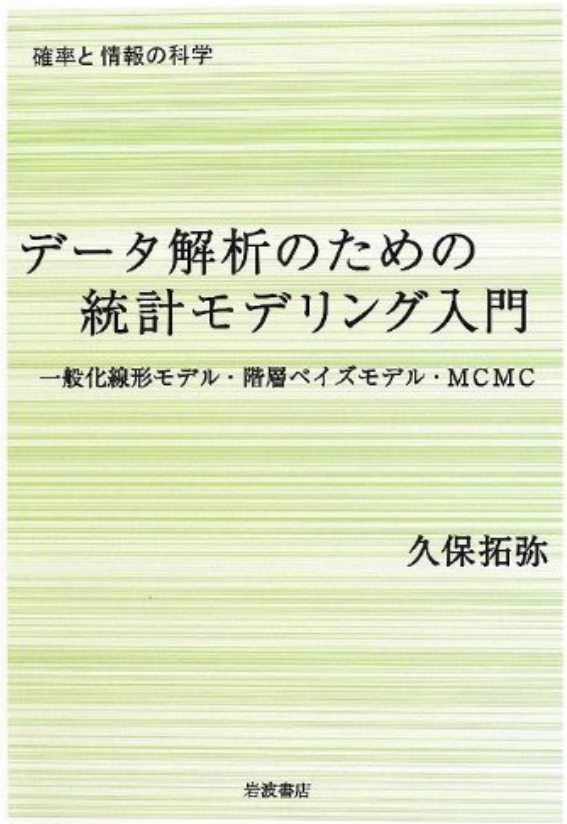
1. **統計モデル・確率分布・最尤推定**
2. **ポアソン分布の一般化線形モデル** (GLM)
3. **二項分布の** GLM **と** GLMM
4. MCMC **と階層ベイズモデル**

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

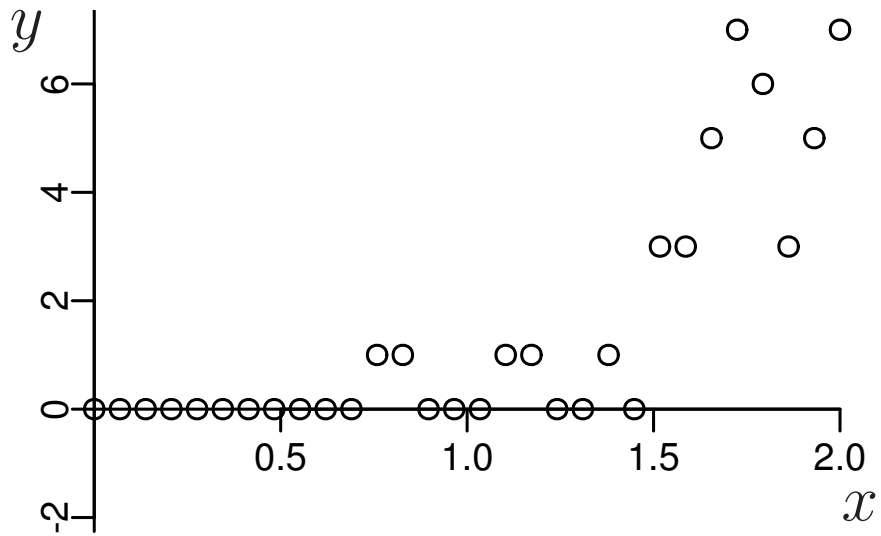
- 観察によって**データ化された現象**を説明するために作られる
- **確率分布**が基本的な部品であり，これはデータにみられるばらつきを表現する手段である
- **データとモデルを対応づける手つづき**が準備されていて，モデルがデータにどれぐらい良くあてはまっているかを定量的に評価できる



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**$)$

応答変数



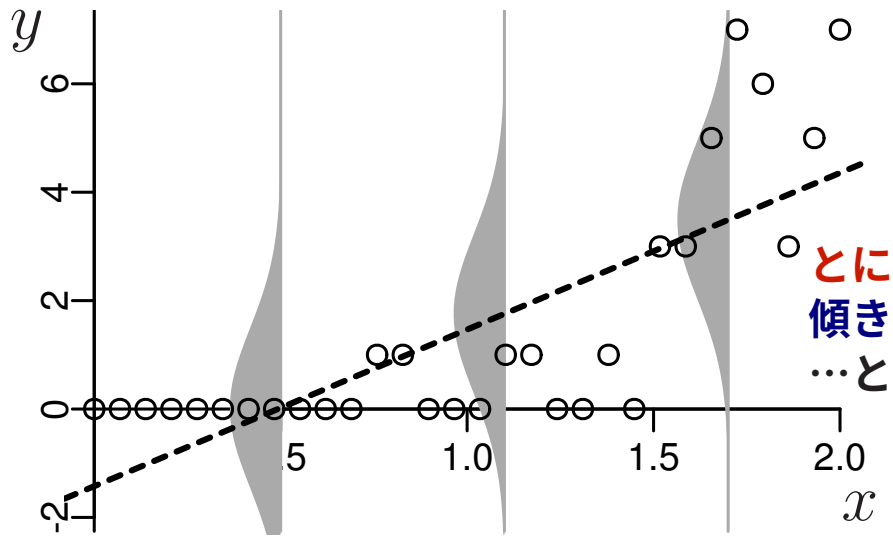
説明変数

- (たとえば) x は “動物の体の大きさ” (単位不詳)
- y はその動物が一年間に産んだ子どもの数
- この現象を表現する統計モデルは?

“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

応答変数



NO!

とにかくセンひきゃいいんでしょ
傾き「ゆーい」ならいいんでしょ
…という安易な発想のデータ解析

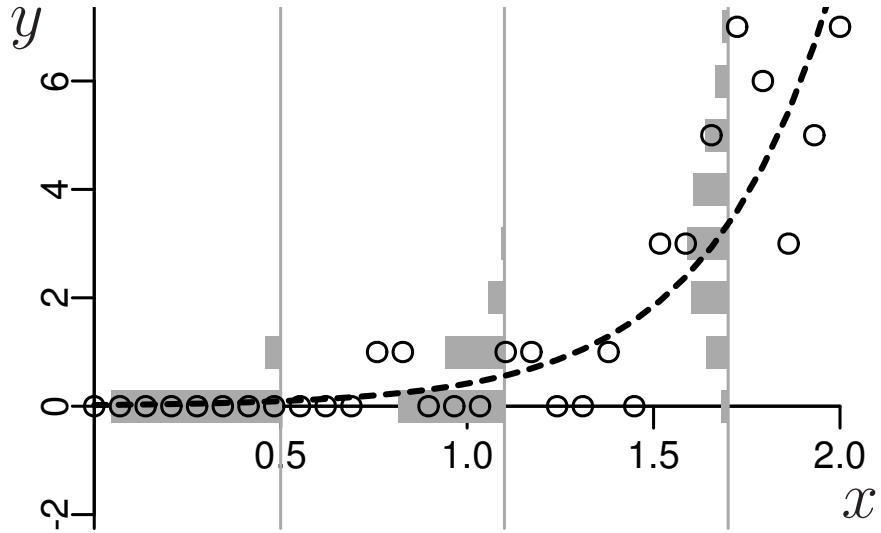
説明変数

- 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

応答変数



YES!

説明変数

- 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

1. 例題: カウントデータの統計モデリング

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

こんなデータ (架空) があったとしましょう

100 秒間に算数の問題をいくつ回答できるか? そういう調査をやったとします

児童 i

全 50 児童

$i \in \{1, 2, 3, \cdots, 50\}$

100 秒間の回答数 y_i

この $\{y_i\}$ が観測データ!

$\{y_i\} = \{y_1, y_2, \cdots, y_{50}\}$

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
```

統計ソフトウェア R

統計学の勉強には良い統計ソフトウェアが必要!

- 無料で入手できる
- 内容が完全に公開されている
- 多くの研究者が使っている
- 作図機能が強力

この教科書でも R を
使って問題を解決する
方法を説明しています

R でデータの様子をながめる

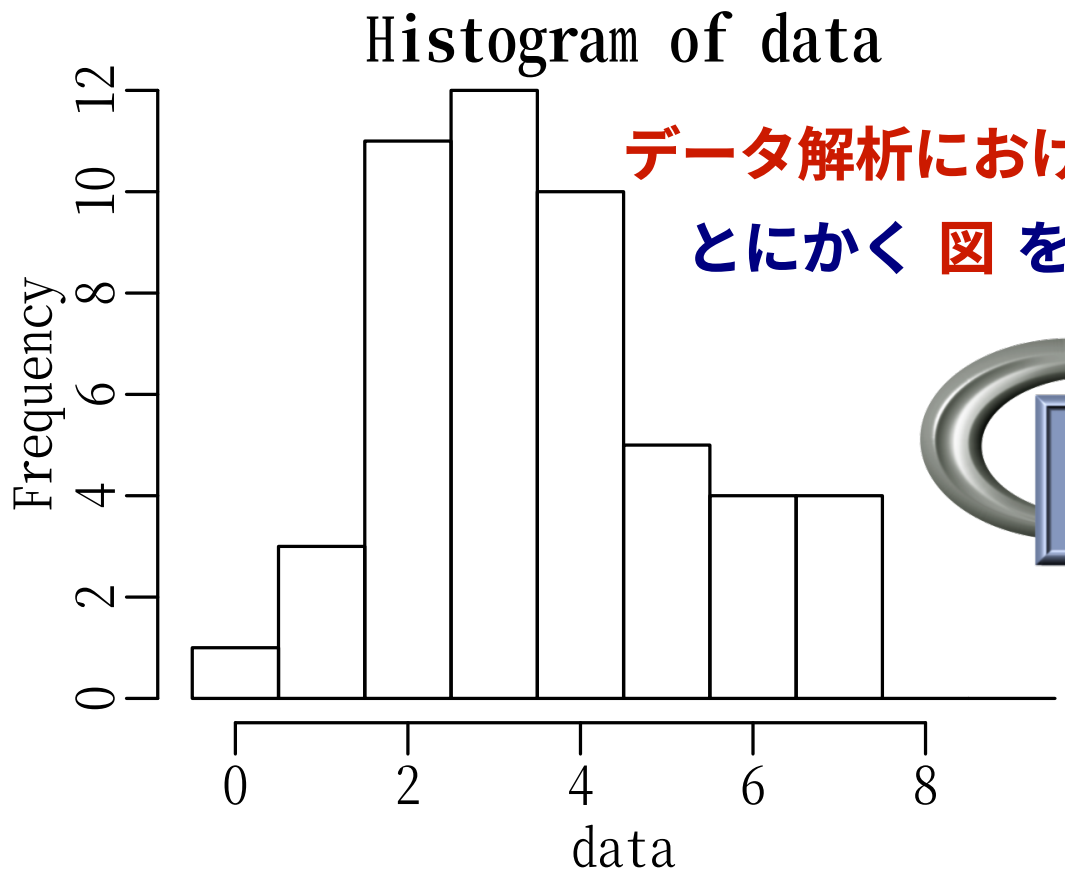

の table() 関数を使って回答数の頻度を調べる

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

(回答数 5 は 5 児童，回答数 6 は 4 児童 ……)

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))
```

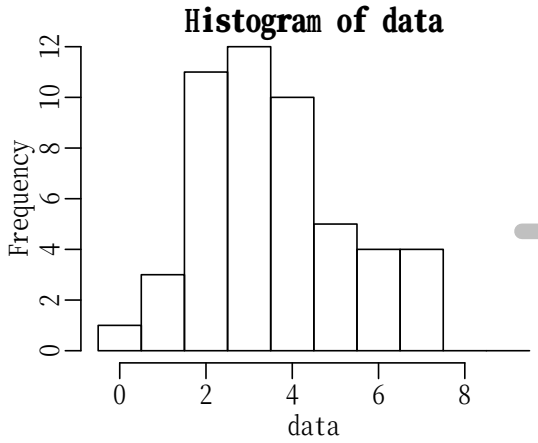


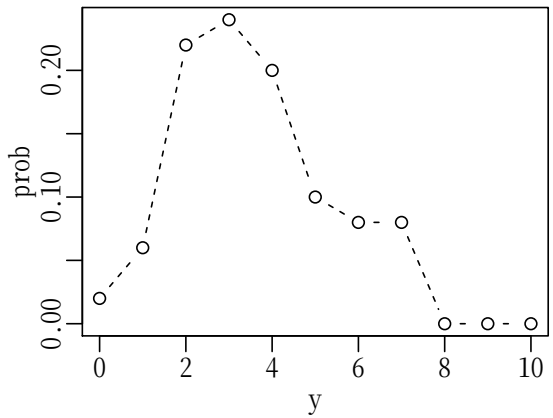
データ解析における最重要事項

とにかく 図 を描く!

“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**

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**?
- **回答数** $y = \{0, 1, 2, \cdots\}$ **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)

確率分布 (ポアソン分布) を数式で決めてしまう

回答数が y である確率は以下のように決まる，と考えている

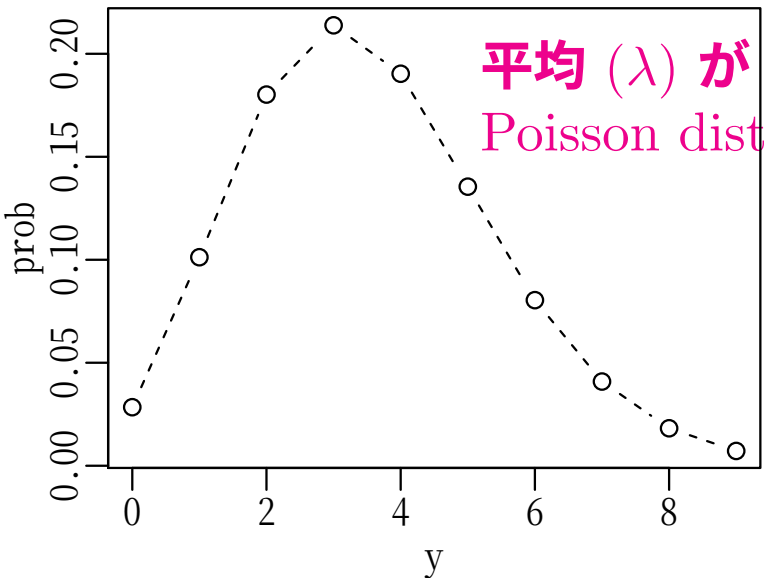
$$p(y \mid \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}$$

- **$y!$ 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**

このデータの標本平均は 3.56 だったので，とりあえず……

とりあえず $\lambda = 3.56$ のポアソン分布を R で作図してみる

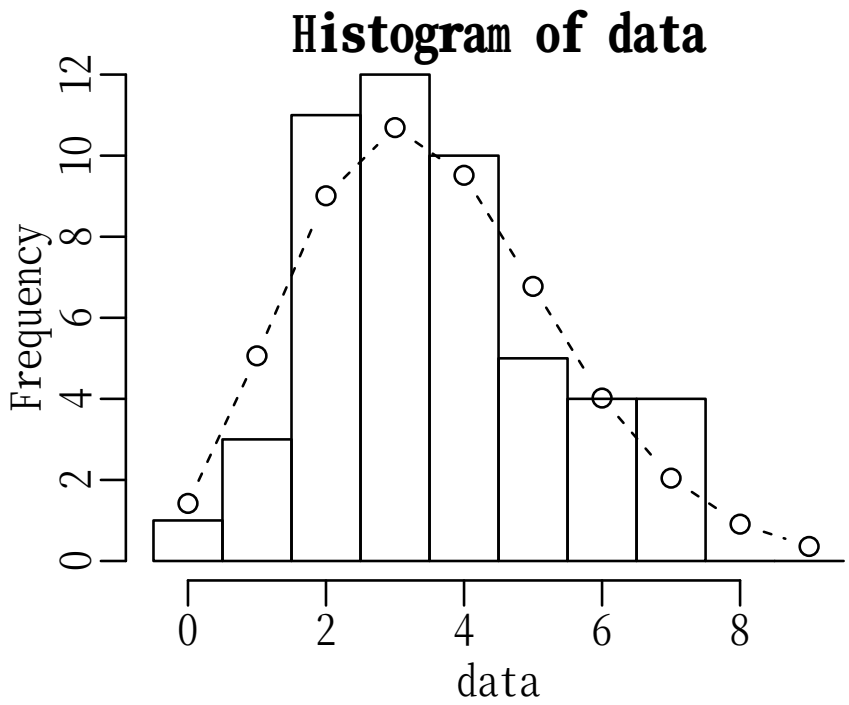
```
> y <- 0:9                                  # これは回答数（確率変数）
> prob <- dpois(y, lambda = 3.56)           # ポアソン分布の確率の計算
> plot(y, prob, type = "b", lty = 2)
```



平均 (λ) が 3.56 である
Poisson distribution

```
> # 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
```

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
```

どういう場合にポアソン分布を使う?

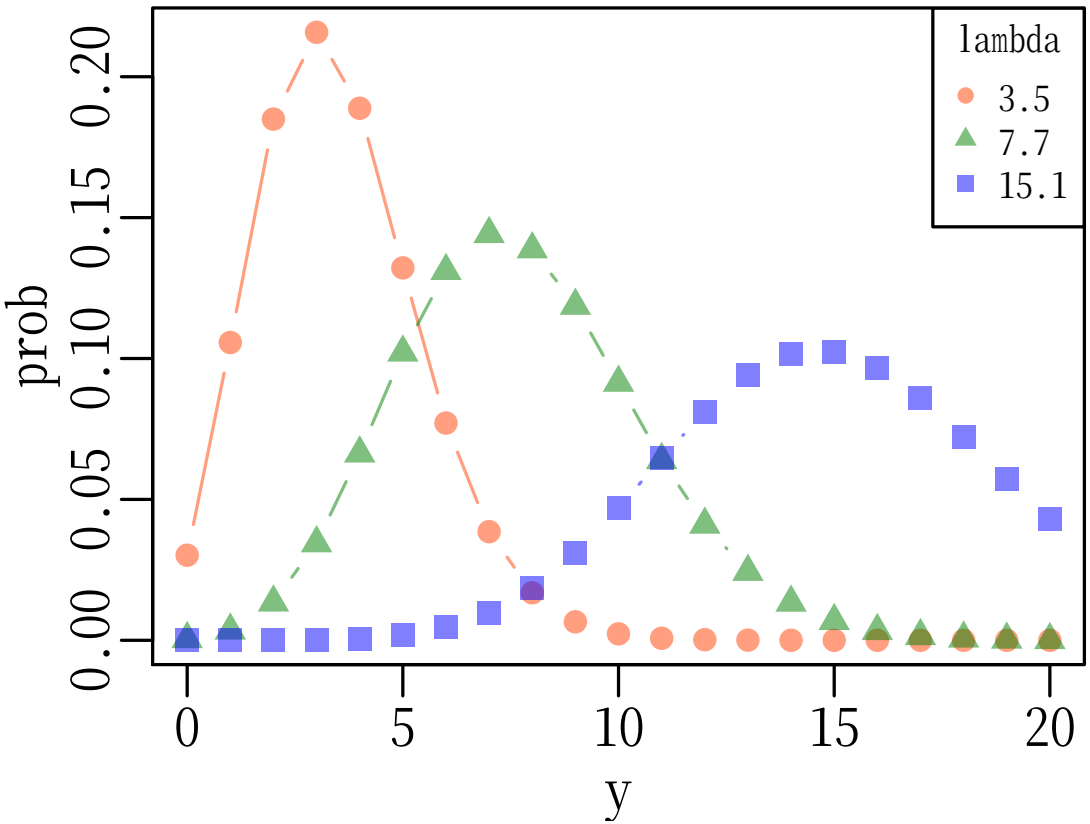
統計モデルの部品としてポアソン分布が選んだ理由:

- **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**
  - **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**

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 は平均をあらわすパラメーター



2. ポアソン分布のパラメーターの最尤推定

もっともらしい推定?

尤度 (likelihood) とは何か?

- 最尤推定法では，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」が良くなる λ を見つけたい
- たとえば，いまデータが 3 児童ぶん，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

尤度 $L(\lambda)$ はパラメーター λ の関数

この例題の尤度:

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

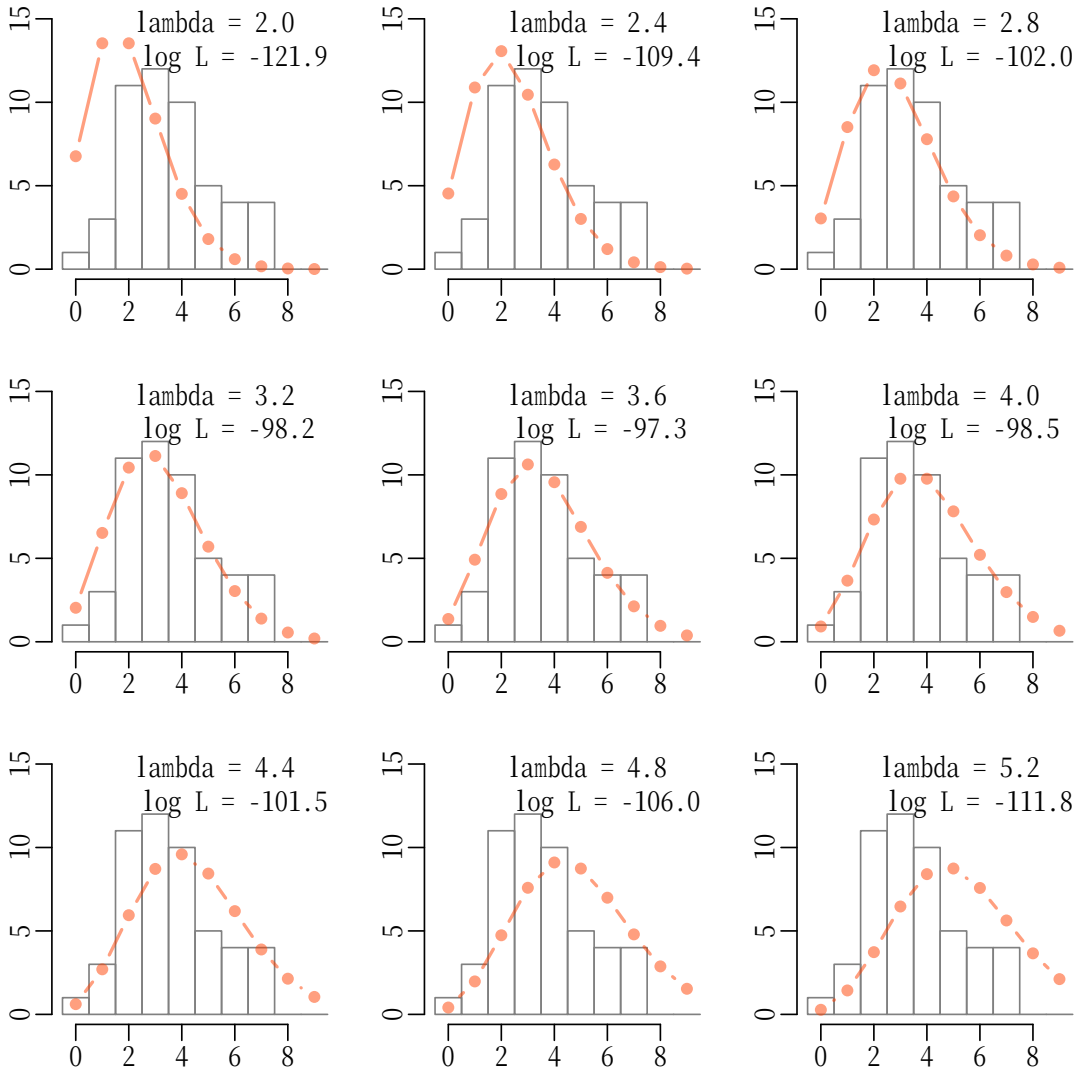
尤度はしんどいので対数尤度を使う

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

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

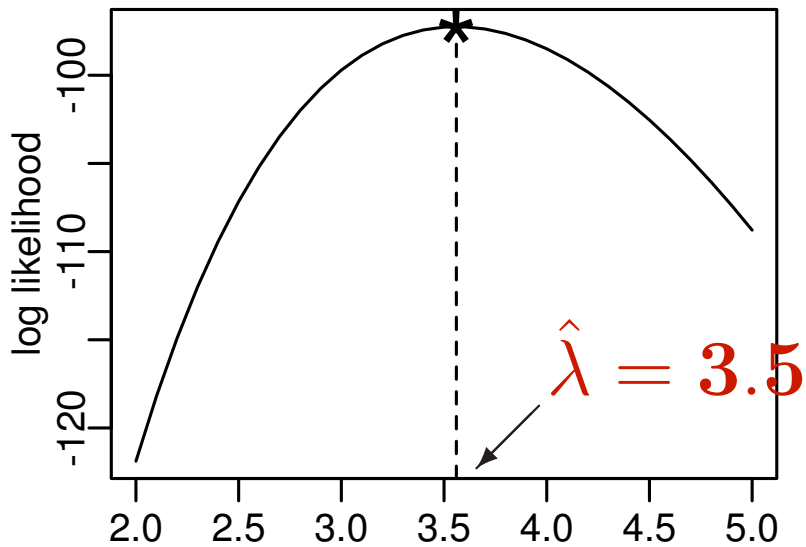
対数尤度 $\log L(\lambda)$ の最大化は尤度 $L(\lambda)$ の最大化になるから
まずは，平均をあらわすパラメーター λ を変化させていったときに，ポアソン分布のカタチと対数尤度がどのように変化するのかを調べてみましょう

λ を変えるとあてはまりの良さが変わる



対数尤度を最大化する $\hat{\lambda}$ をさがす

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

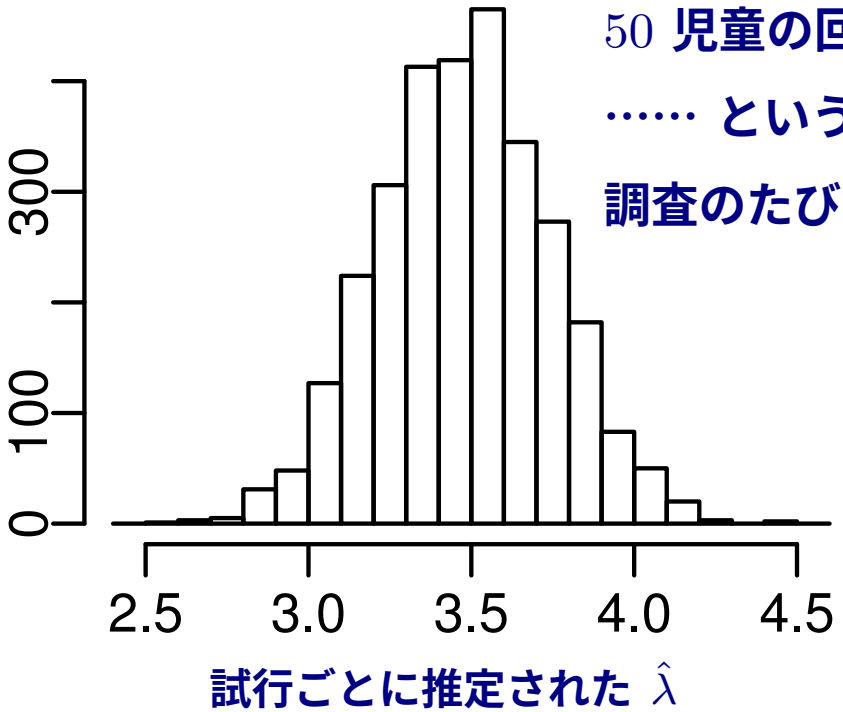


$\frac{d \log L}{d\lambda} = 0$ より

- **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**

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 児童の回答数を調べる

…… 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

3. ポアソン回帰の例題: 架空の計算問題テスト

被験者の属性，あるいは実験処理が回答数に影響?

年齢と実験処理の効果を調べる例題

- 7 – 12 歳ぐらいの児童 100 人
- 100 秒間にいくつ計算問題に回答できるか調べた
- 50 人は「教育法 F」，50 人は無処理 (コントロール)

- **応答変数**: 100 秒間の回答数 $\{y_i\}$
- **説明変数**:
  - 年齢 $\{x_i\}$
  - 教育法 F $\{f_i\}$
- 無処理 ($f_i = \mathtt{C}$): 50 sample ($i \in \{1, 2, \cdots 50\}$)
- 教育法 F ($f_i = \mathtt{T}$): 50 sample ($i \in \{51, 52, \cdots 100\}$)

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")
```

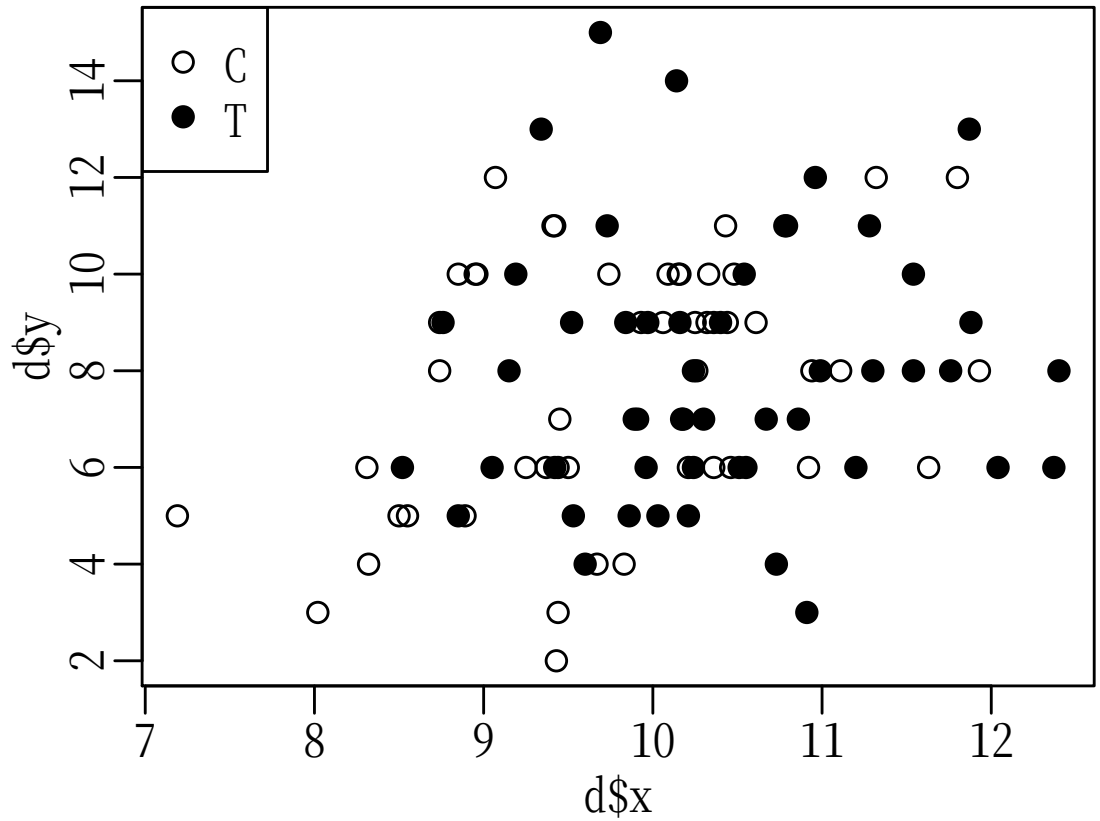
データは d と名付けられた data frame (表みたいなもの) に格納される

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
```

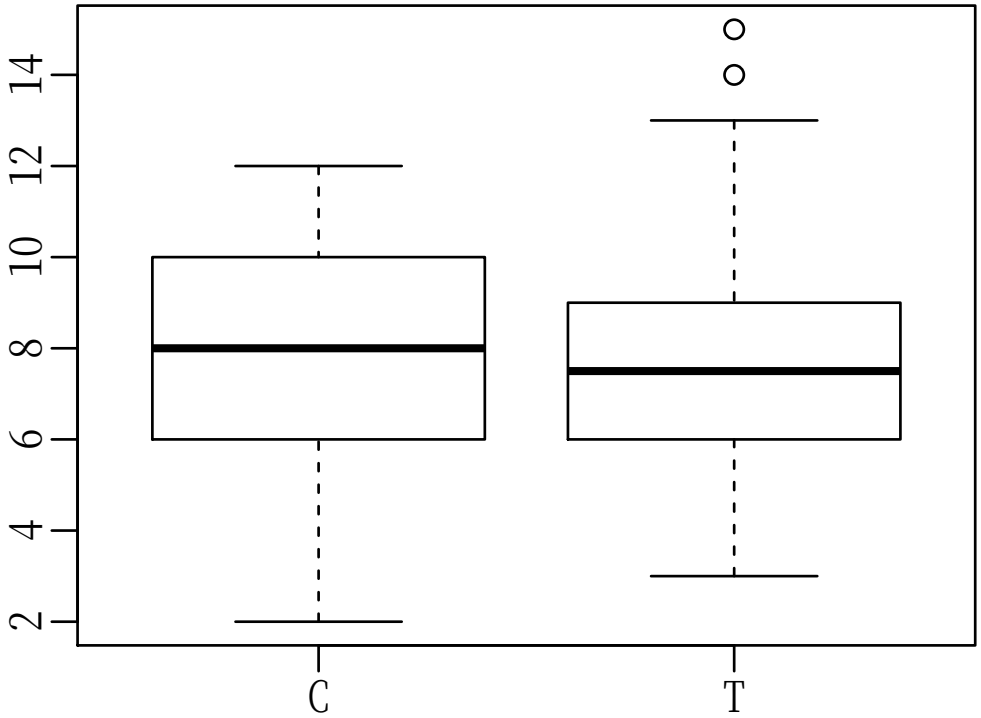
データはとにかく図示する!

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



教育法F をあらわす f を横軸とした図

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



4. GLM の詳細を指定する

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

一般化線形モデル (GLM) って何だろう?

一般化線形モデル (Generalized Linear Model)

- **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)
- ……

一般化線形モデルを作る

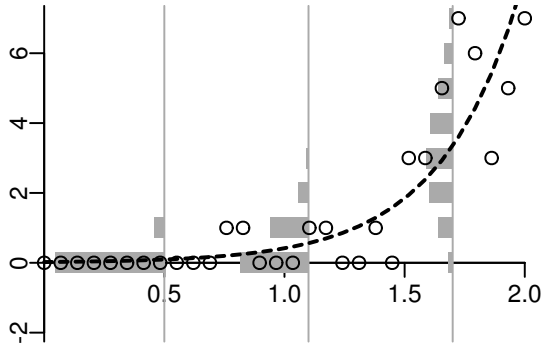
一般化線形モデル (GLM)

- 確率分布は?
- 線形予測子は?
- リンク関数は?

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

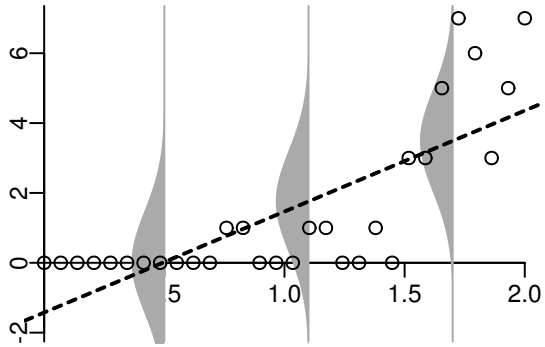
- **確率分布**: **ポアソン分布**
- **線形予測子**: $\mathrm{e.g.,}\ \beta_1 + \beta_2 x_i$
- **リンク関数**: **対数リンク関数**



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

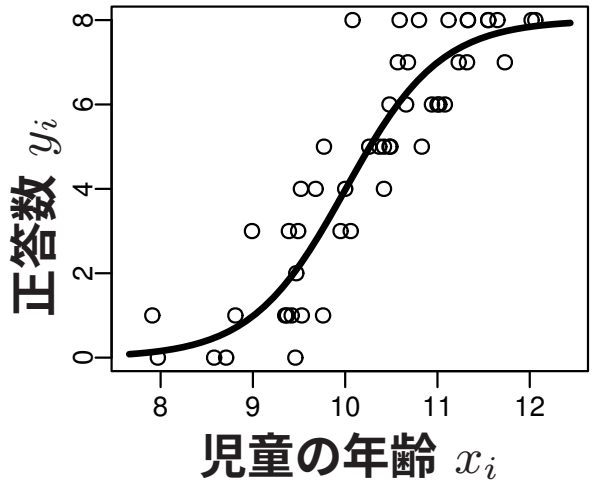
- 確率分布: 正規分布
- 線形予測子: $\mathrm{e.g.,}\ \beta_1 + \beta_2 x_i$
- リンク関数: 恒等リンク関数



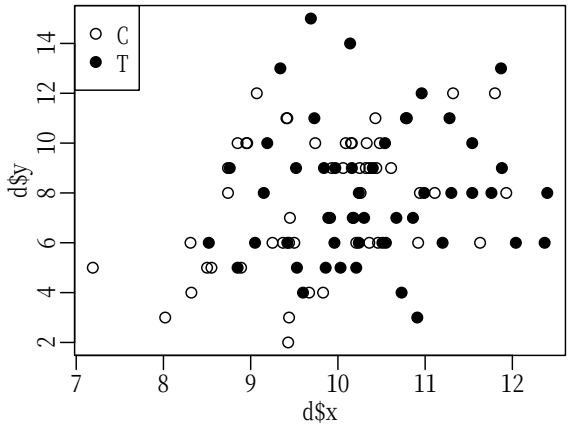
GLM のひとつである logistic 回帰モデルを指定する

ロジスティック回帰のモデル

- **確率分布**: **二項分布**
- **線形予測子**: e.g., $\beta_1 + \beta_2 x_i$
- **リンク関数**: logit **リンク関数**



さてさて，この例題にもどって



回答数 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

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

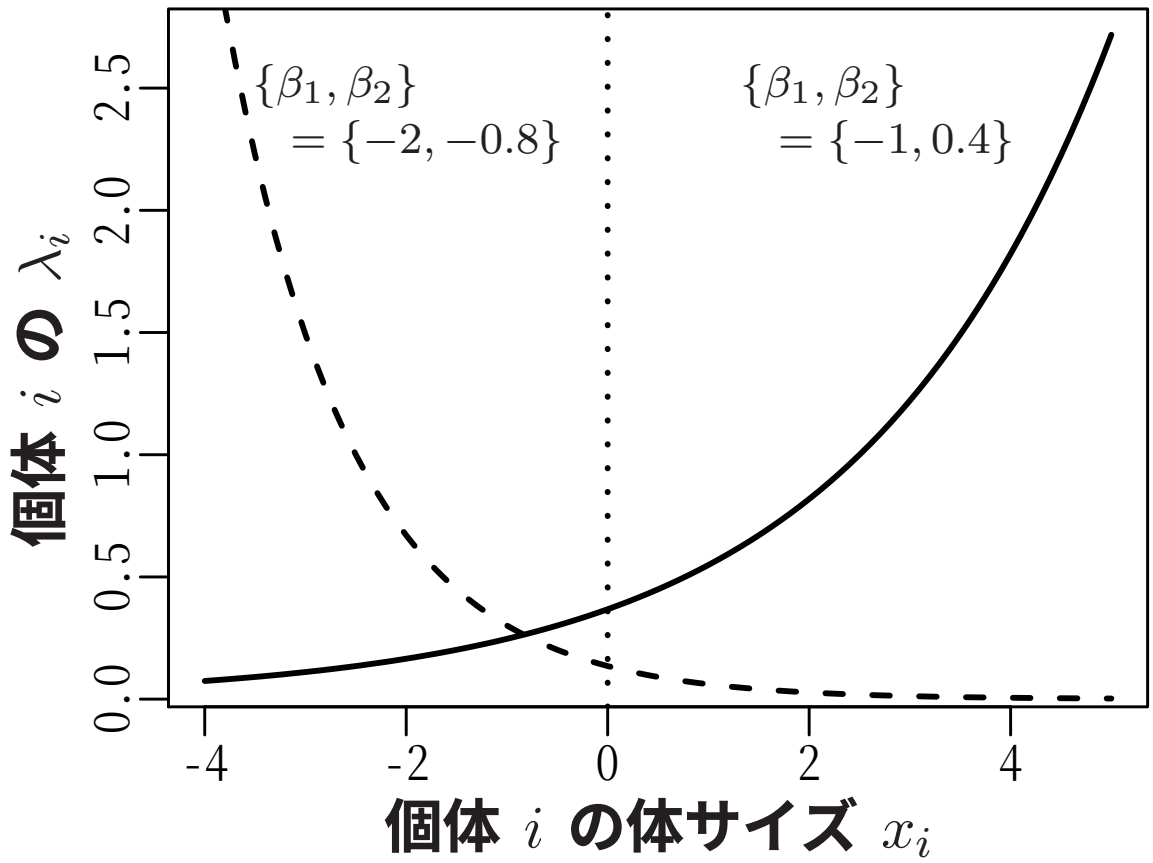
児童 i の平均 λ_i を以下のようにおいてみたらどうだろう……?

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

- **β_1 と β_2 は係数 (パラメーター)**
- **x_i は児童 i の年齢，f_i はとりあえず無視**

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

児童 i の回答数の平均は λ_i

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

$$\Updownarrow$$

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

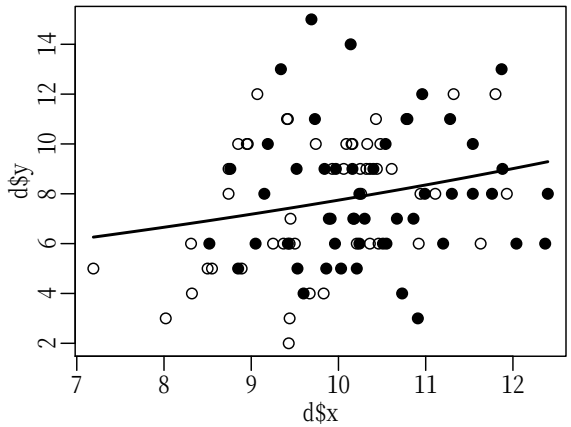
$$\log(\textbf{平均}) = \textbf{線形予測子}$$

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**

この例題のための統計モデル

ポアソン回帰のモデル

- 確率分布: **ポアソン分布**
- **線形予測子**: $\beta_1 + \beta_2 x_i$
- **リンク関数**: **対数リンク関数**

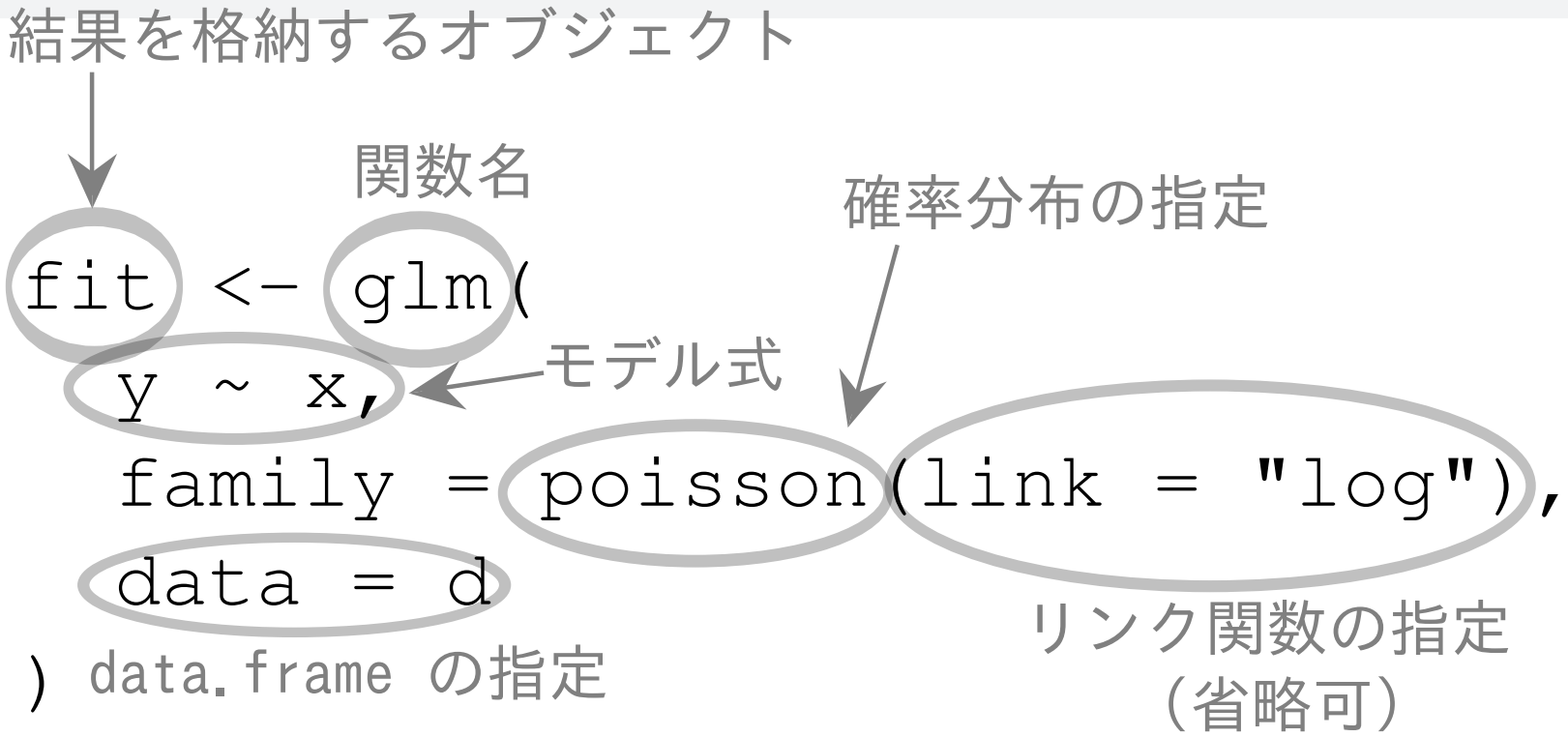


5. R で GLM のパラメーターを推定

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

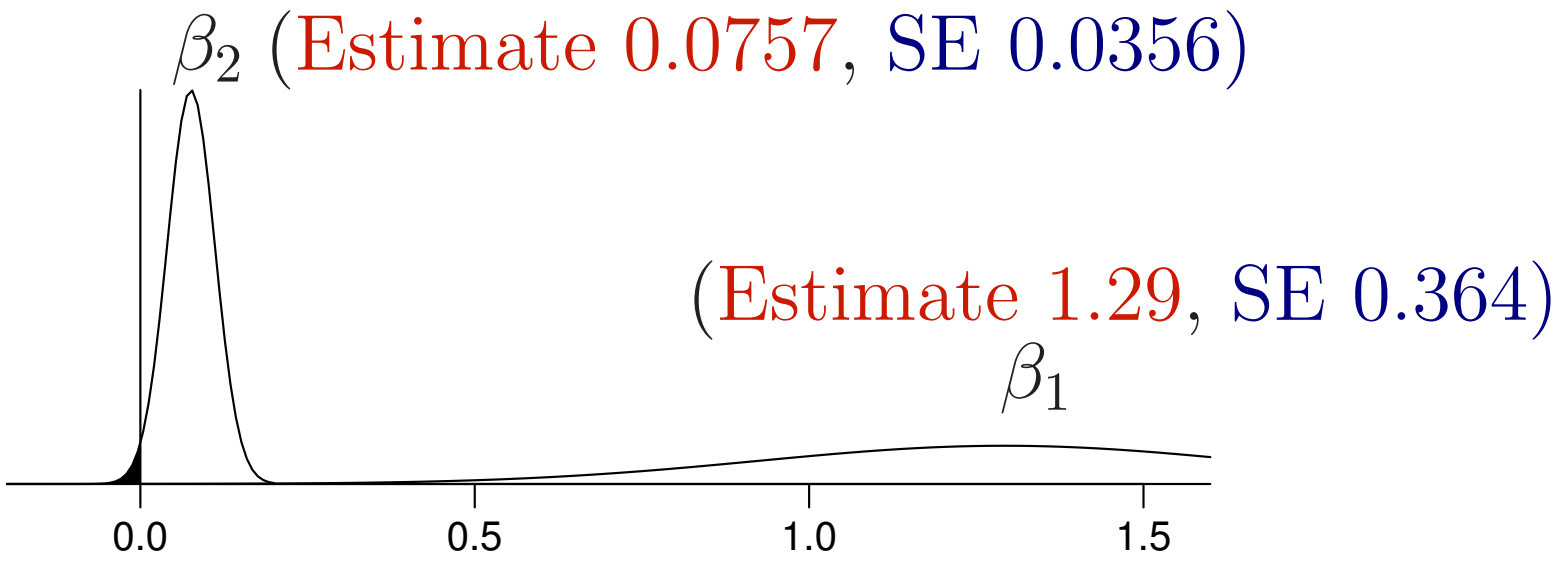
```
> summary(glm(y ~ x, family = poisson, data = d))
Call:
glm(formula = y ~ x, family = poisson, data = d)

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) ……

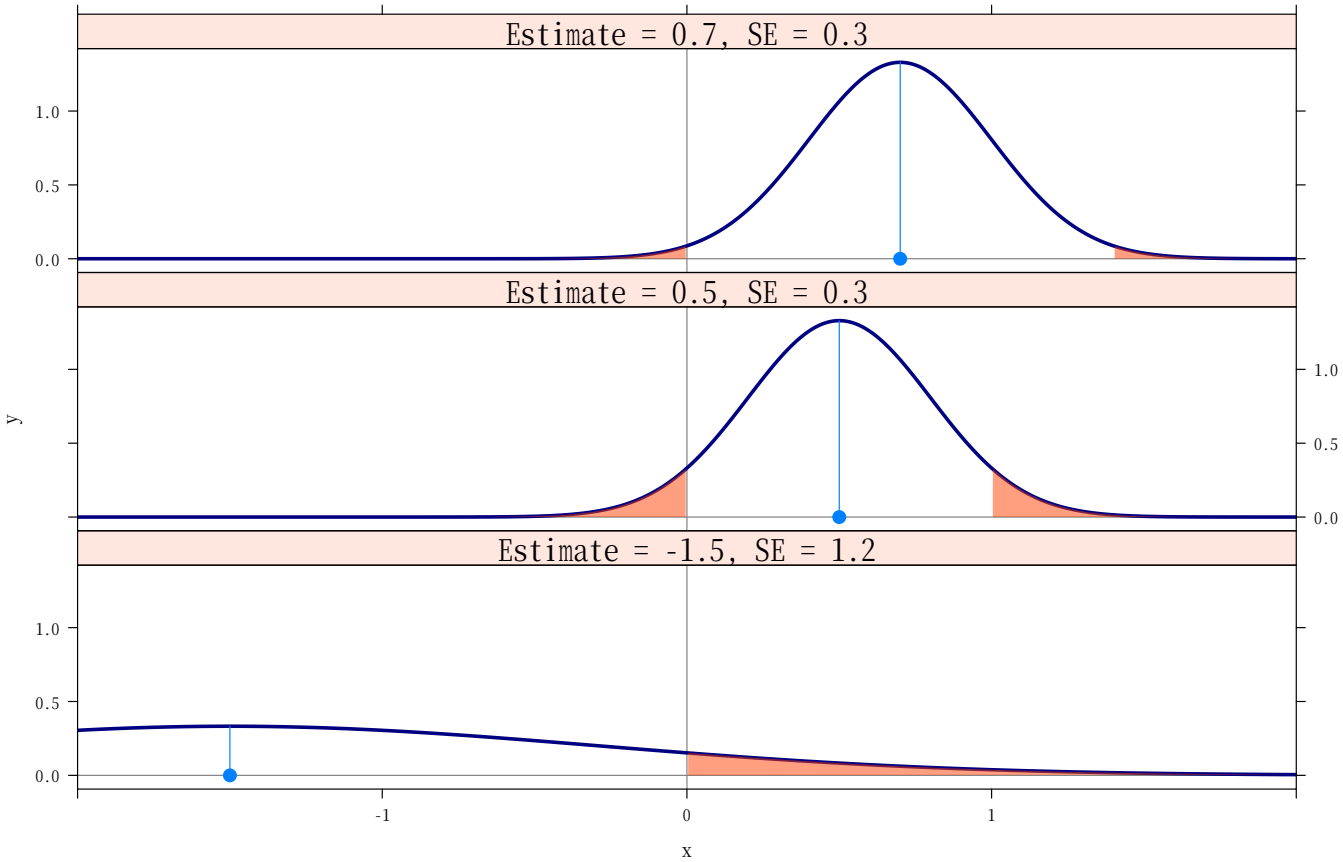
推定値と標準誤差



この図の要点:

- 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

p は “ゼロからの近さ” をあらわす



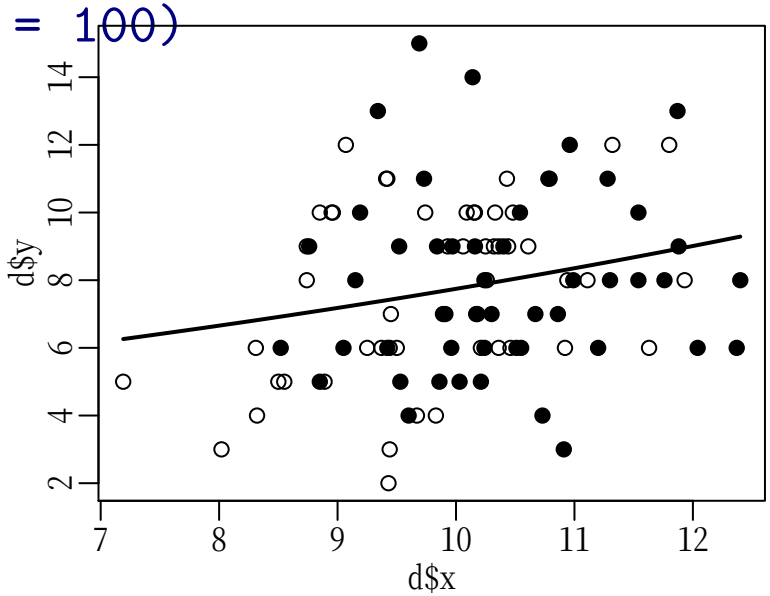
- p がゼロに近いほど **推定値** $\hat{\beta}$ はゼロから離れている
- p が 0.5 に近いほど **推定値** $\hat{\beta}$ はゼロに近い

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))
```

ここでは観測データと予測の関係を見ているだけ，なのだが



6. “N 問のうち k 個が正答” 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

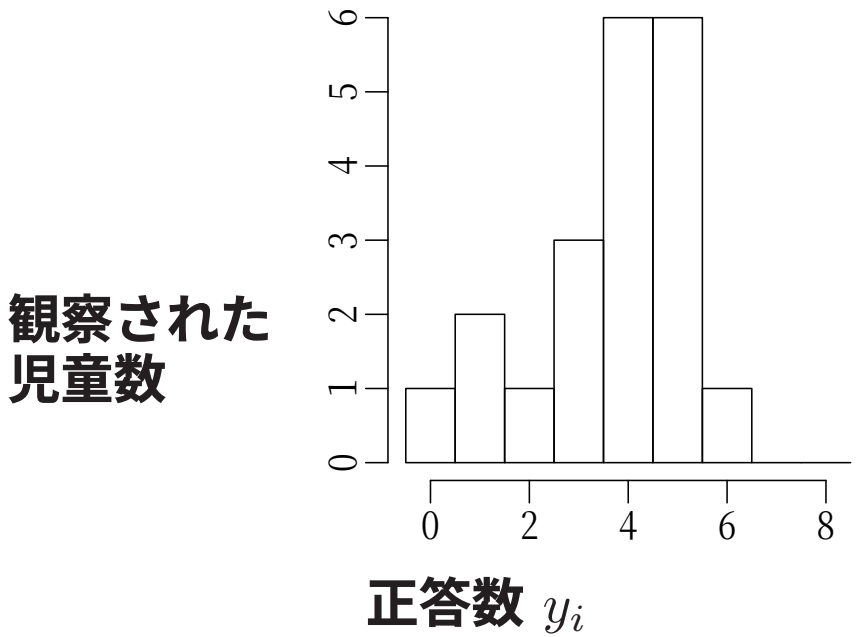
また別の例題: 算数の計算問題の正解確率

- 算数の計算問題の正解数の割合を調べたい
- **回答数**: 問題紙上の計算問題数
  - どの児童にも 8 個 の計算問題を問いてもらう
- **正解確率**: ある回答が正答である確率

- データ: 20 児童，合計 160 回答の正解数を調べた
- 73 回答が正解だった — 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

児童ごとの正解数	0	1	2	3	4	5	6	7	8
観察された児童数	1	2	1	3	6	6	1	0	0



これは児童差なしの均質な集団

正解確率 q と二項分布の関係

- 正解確率を推定するために**二項分布** という確率分布を使う
- 児童 i の N_i 回答中 y_i 個が正解する確率

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

- ここで仮定していること
  - **児童差はない**
  - つまり **どの児童も同じ正解確率** q

(尤度などについてはのちほど)

ただし書き: 現実にはこんなに簡単ではない

- 100 秒あたりの計算問題回答数 $y_i \to$ **ポアソン分布**
- 計算問題 8 問中 y_i 問正答 $\to$ **二項分布**

しかしながら実際のところは…

- こういう “試験問題” みたいなデータの確率分布は**簡単ではない**
- 作問かなり工夫した場合，ポアソン分布・二項分布で近似できる場合も?
- あとから登場する “児童差” はまた**別の問題**

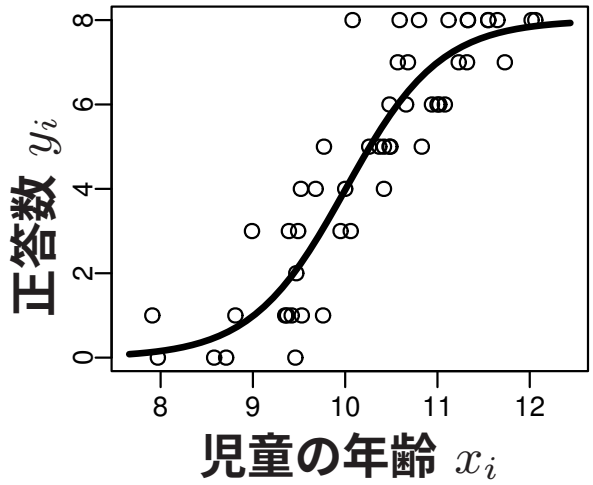
7. ロジスティック回帰のモデル

もっともよく使われる 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**: **二項分布**
- **線形予測子**: e.g., $\beta_1 + \beta_2 x_i$
- **リンク関数**: logit **リンク関数**



また架空の例題: 正答確率は年齢・教育法で変わるか?

8 問の計算問題のうち y 個が正答だった，というデータ

児童 i

回答数 $N_i = 8$

処理 f_i

C: 処理なし

T: 処理あり

正答数 $y_i = 3$

年齢 x_i

データファイルを読みこむ 

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")
```

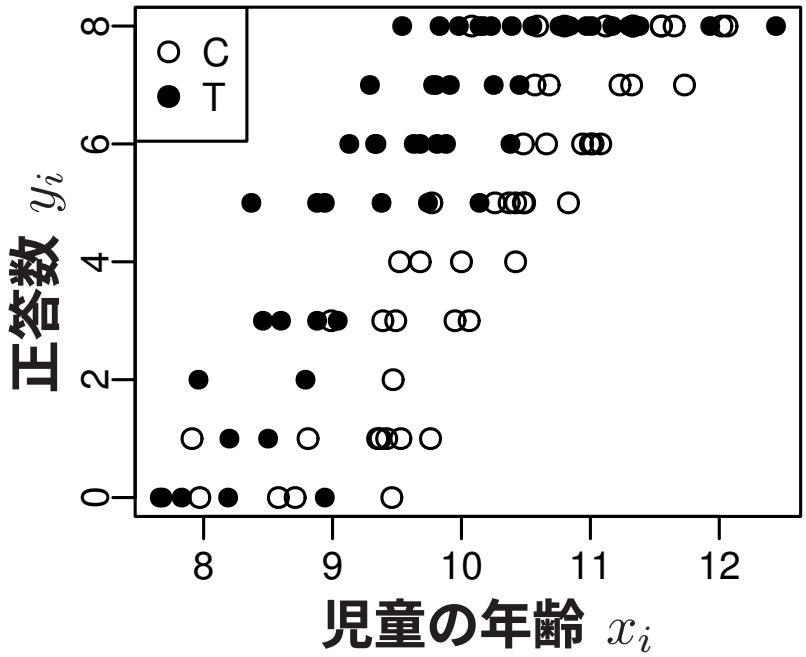
データは d と名付けられた data frame (表みたいなもの) に格納される

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))
```



今回は教育法 F がきいている?

8. ロジスティック回帰の部品

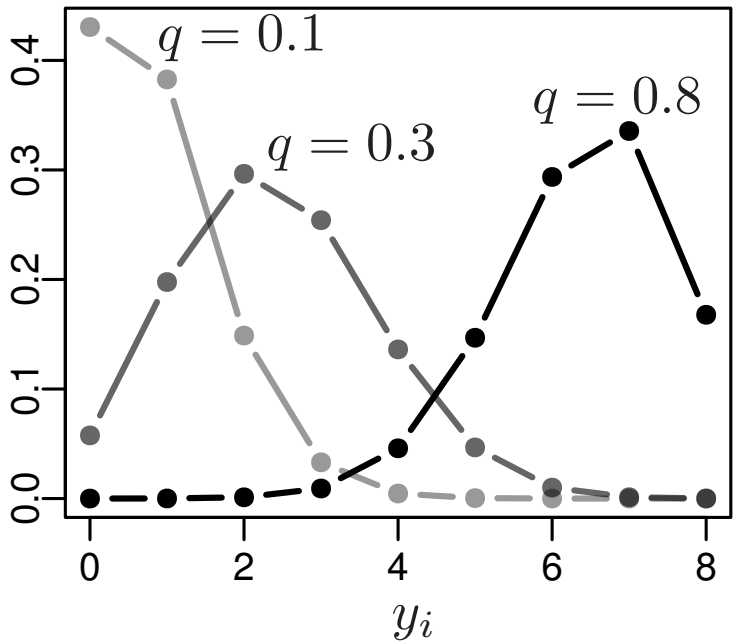
二項分布と logit link function

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 個の問題中 y 個の正答する場合の数

確率 $p(y_i \mid 8, q)$

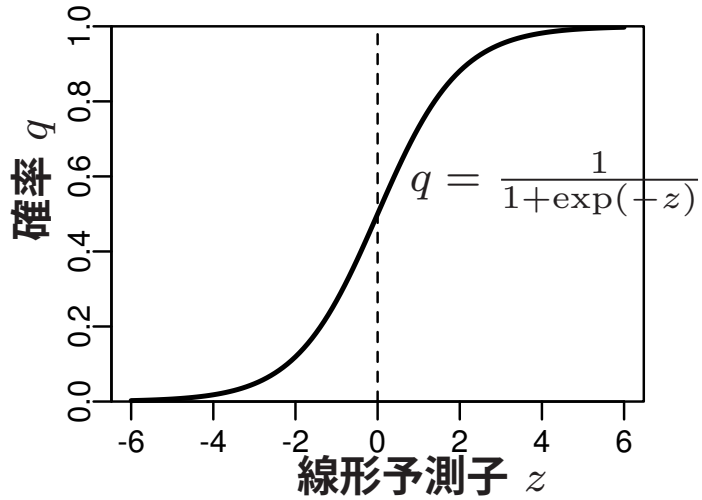


ロジスティック曲線とはこういうもの

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

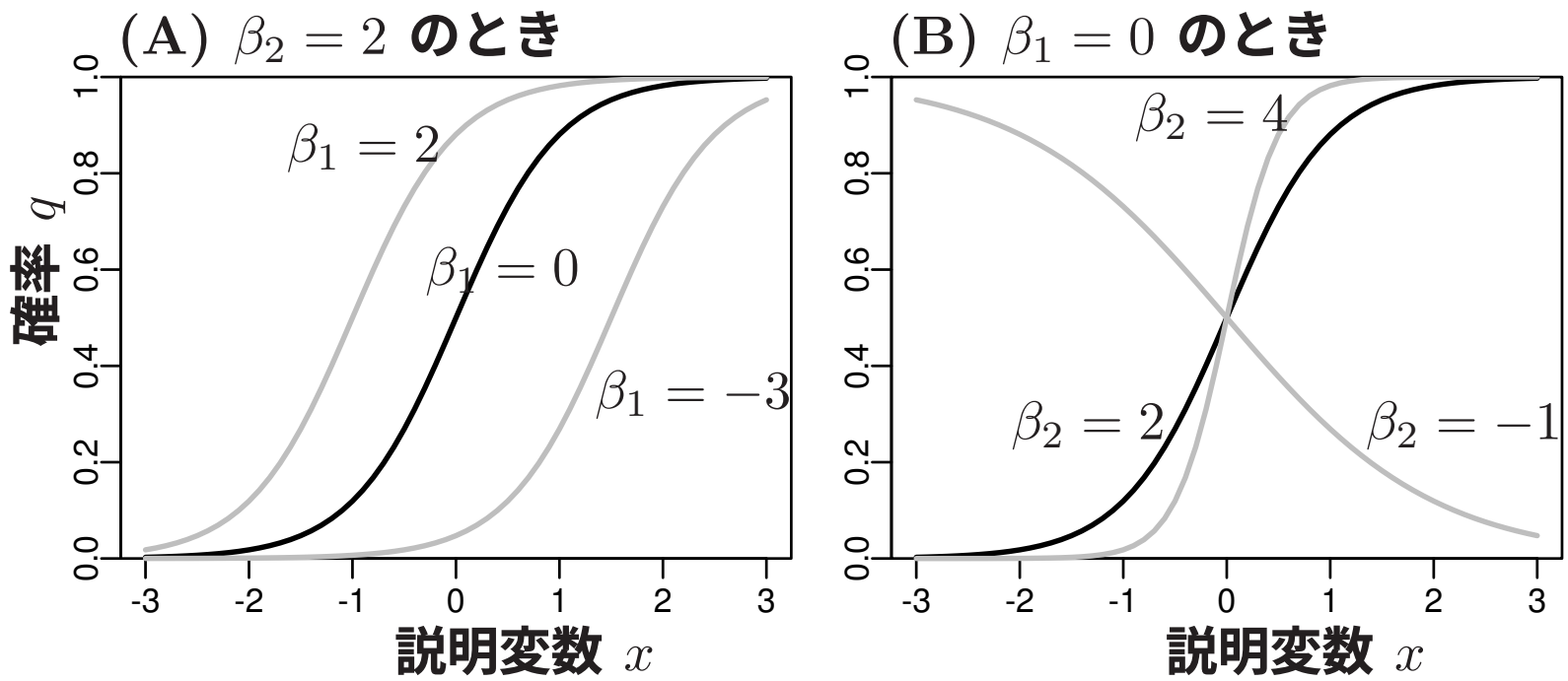
$$q_i = \mathrm{logistic}(z_i) = \frac{1}{1 + \exp(-z_i)}$$

```
> 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**

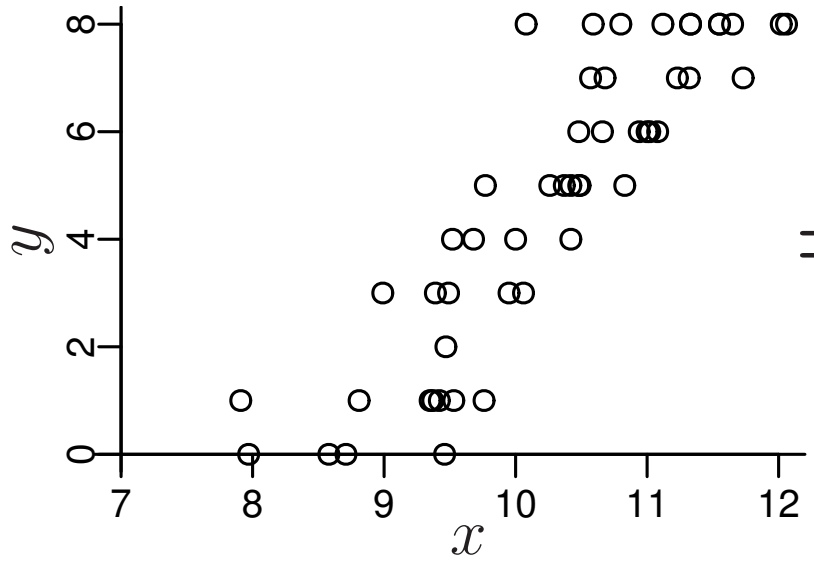
$$\mathrm{logit}(q) = \log \frac{q}{1-q} = \beta_1 + \beta_2 x$$

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

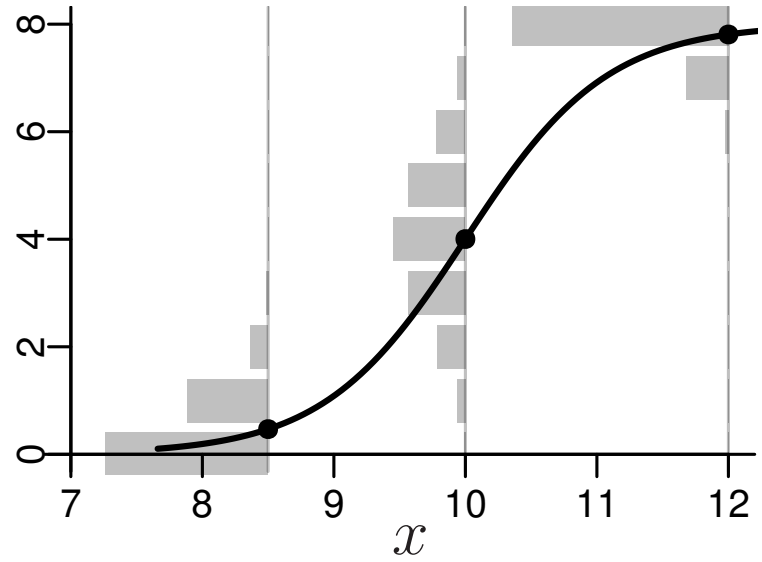
logit is the inverse function of logistic function, vice versa

R でロジスティック回帰 — β_1 と β_2 の最尤推定

(A) 例題データの一部（f_i =C）

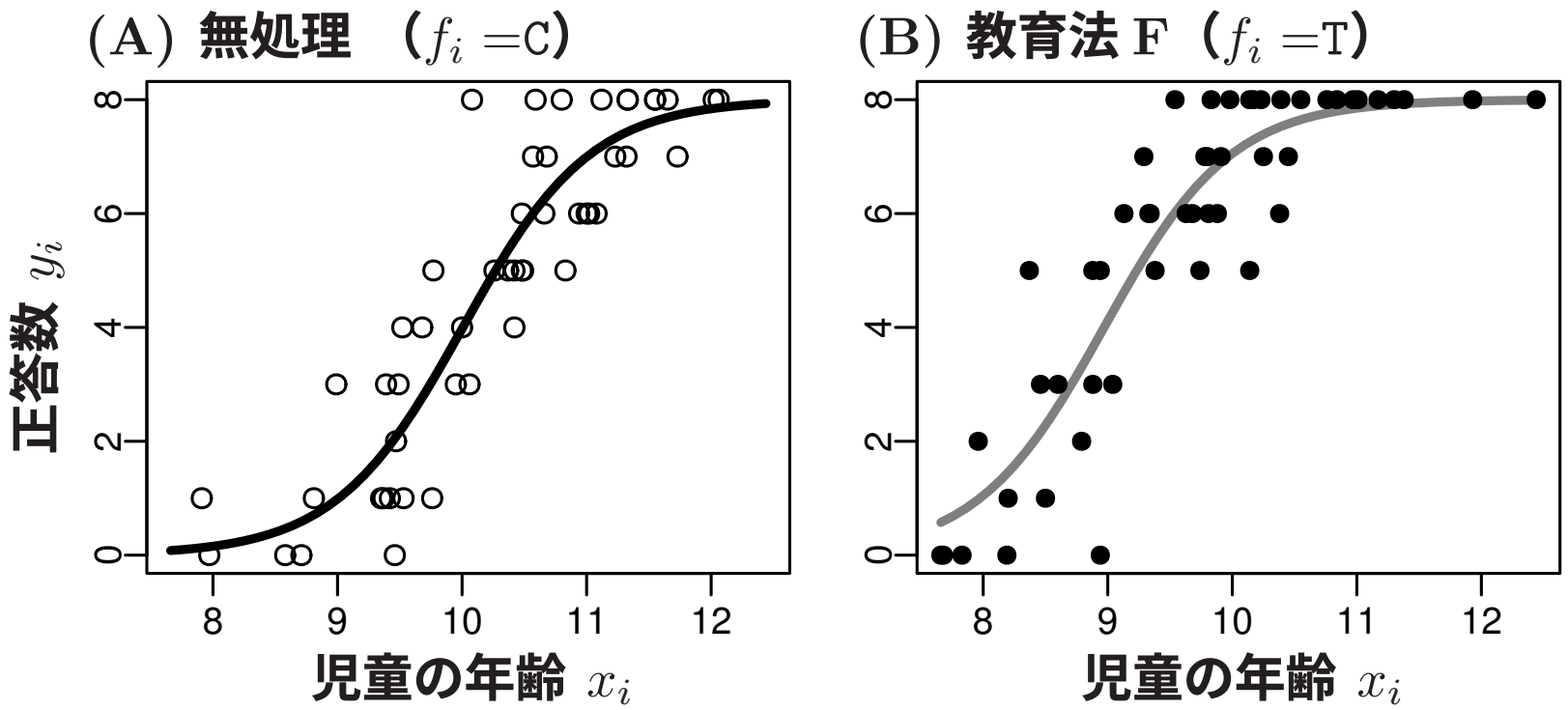


(B) 推定されるモデル




```
> 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: 教育法 F によって応答が違う

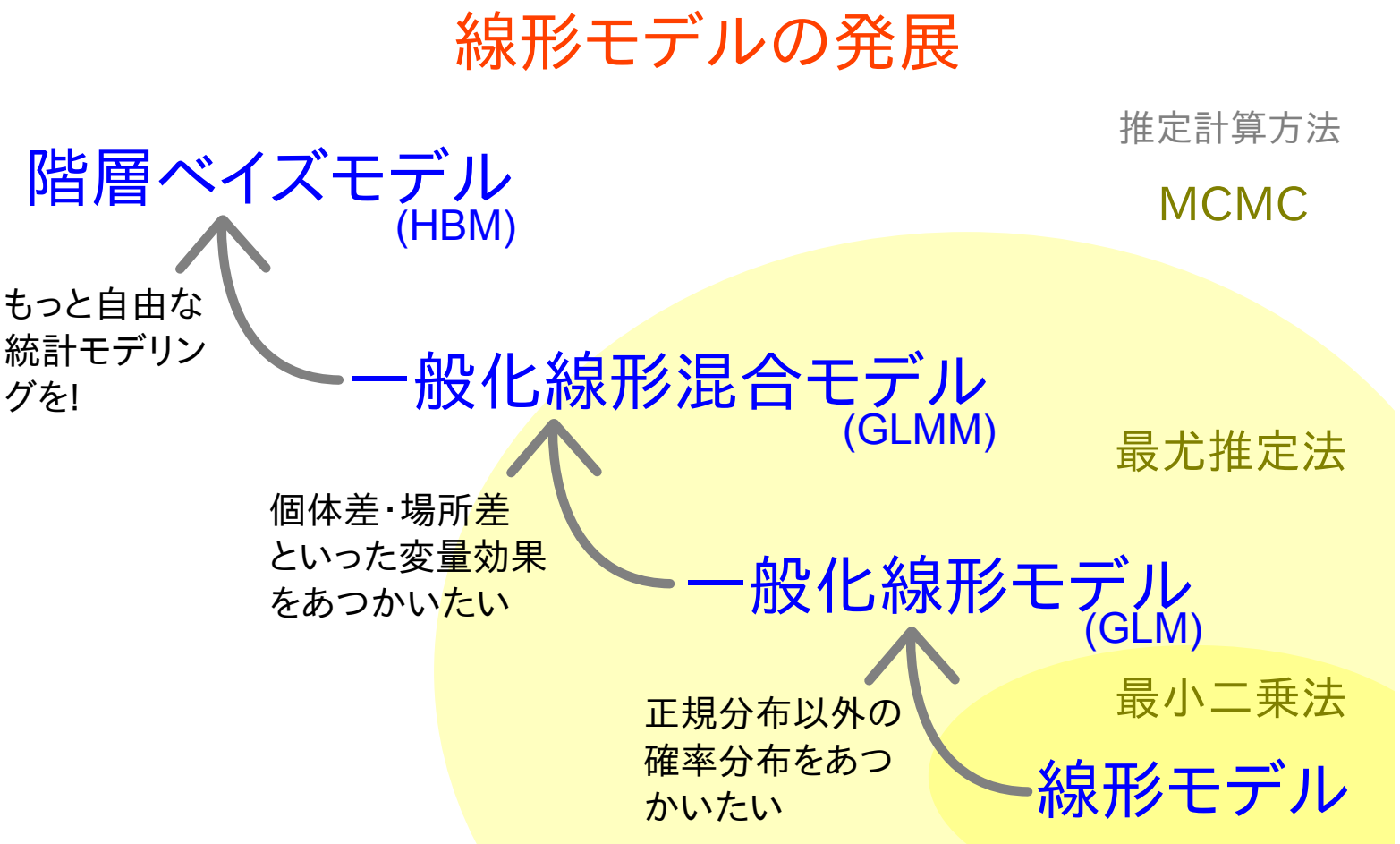


9. GLM だけでは実際のデータ解析はできない

一般化線形混合モデル (GLMM) 登場!

GLM は「個体差」などを無視しているところが問題

“統計モデリング入門” に登場する統計モデル



データの特徴にあわせて線形モデルを改良・発展させる

計算問題の正解確率の GLMM

(A) 児童 i で観測されたデータ

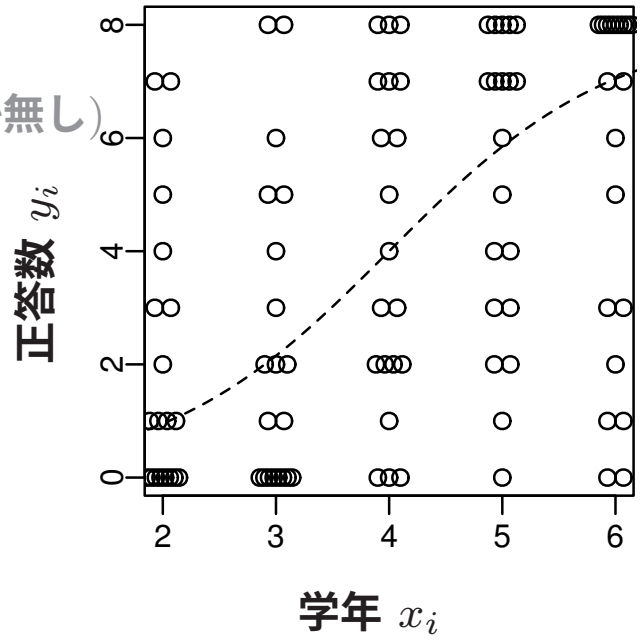
問題数 $N_i = 8$

正答数 $y_i = 3$

(この例題は「教育法」とか無し)

学年 $x_i \in \{2, 3, 4, 5, 6\}$

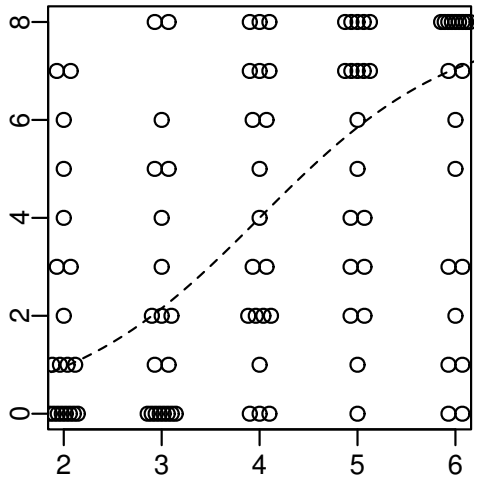
(B) 全 100 児童の x_i と y_i



“N 個中の y 個” というデータ $\rightarrow$ ロジスティック回帰?

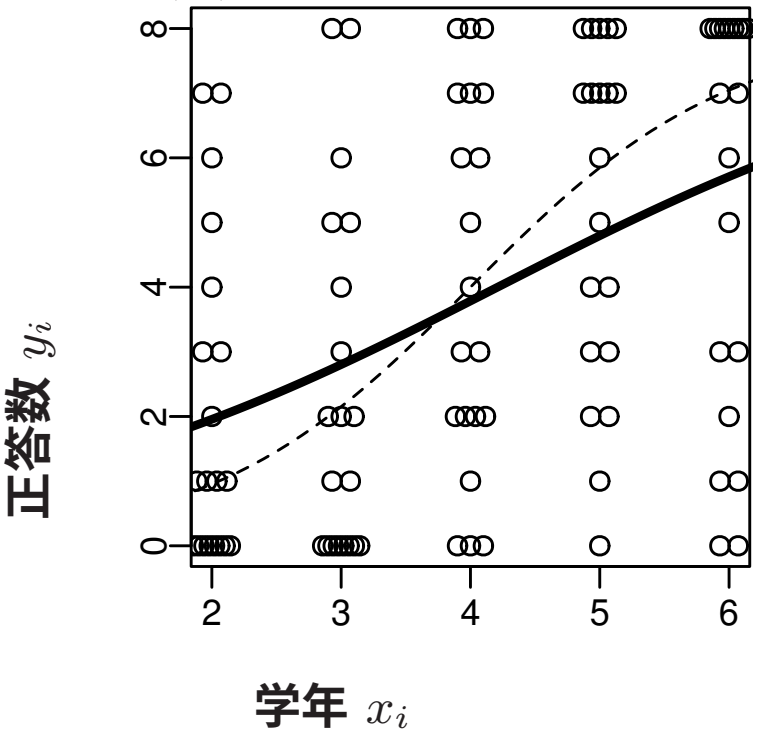
ロジスティック回帰のモデル

- **確率分布**: **二項分布**
- **線形予測子**: $\beta_1 + \beta_2 x_i$
- **リンク関数**: logit **リンク関数**

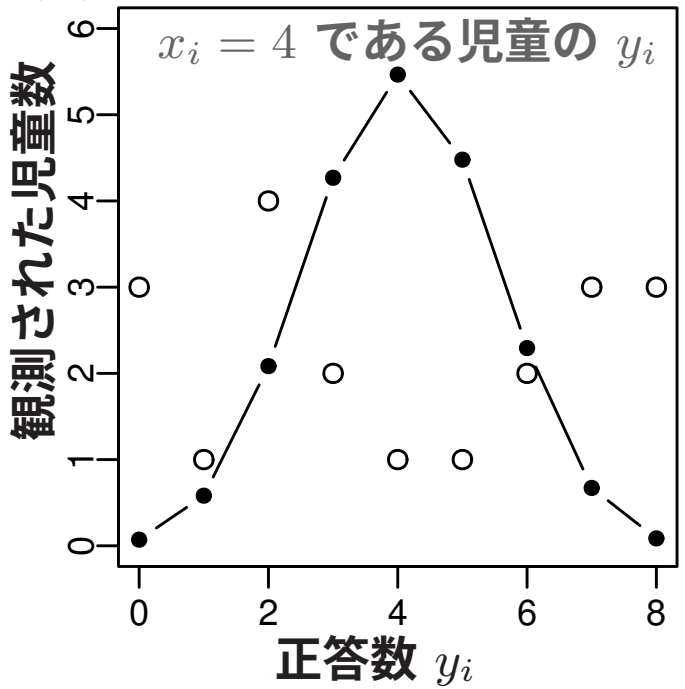


GLM では説明できないばらつき!

(A) 傾き β_2 が過小推定



(B) ぜんぜん二項分布じゃない!

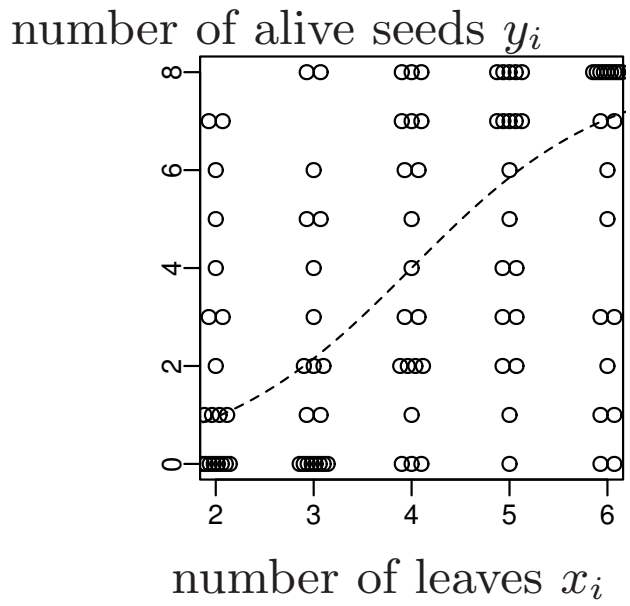


ロジスティック回帰やポアソン回帰といった GLM では全サンプルの均質性を仮定している

(例) $\mathrm{logit}(q_i) = \beta_1 + \beta_2 x_i$

現実のカウントデータは，多くの場合「過分散」

ロジスティック回帰のモデルを改良する



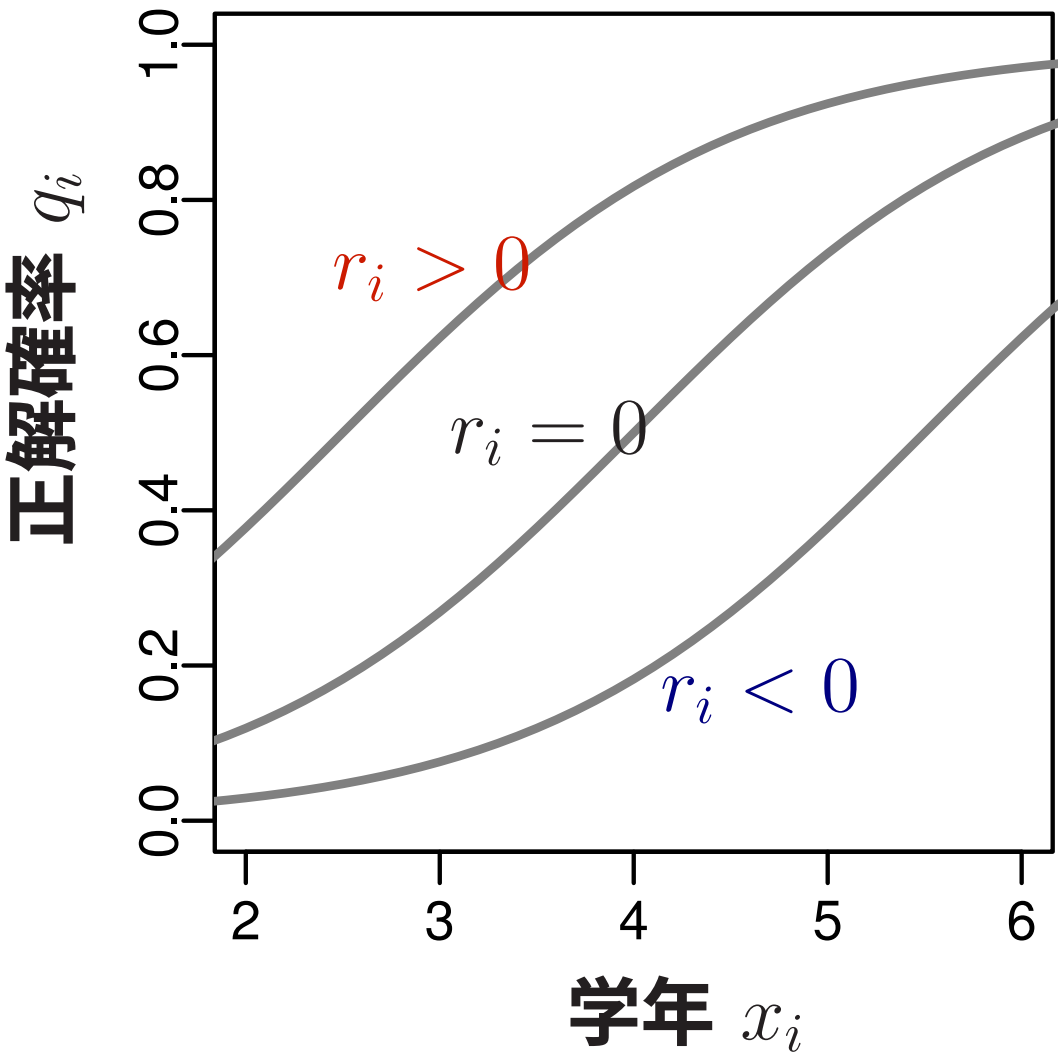
ロジスティック回帰のモデル

- 確率分布: 二項分布
- 線形予測子: $\beta_1 + \beta_2 x_i + r_i$
- リンク関数: logit リンク関数

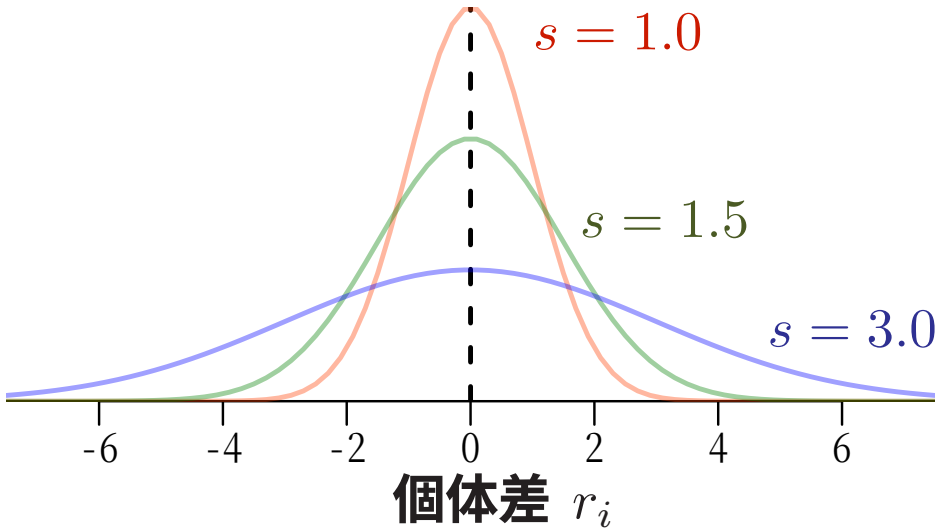
なぜ? $+r_i$ とかするの……?

…… このあとおいおい説明してみます

児童 i の “差” を r_i とするとどうなる?



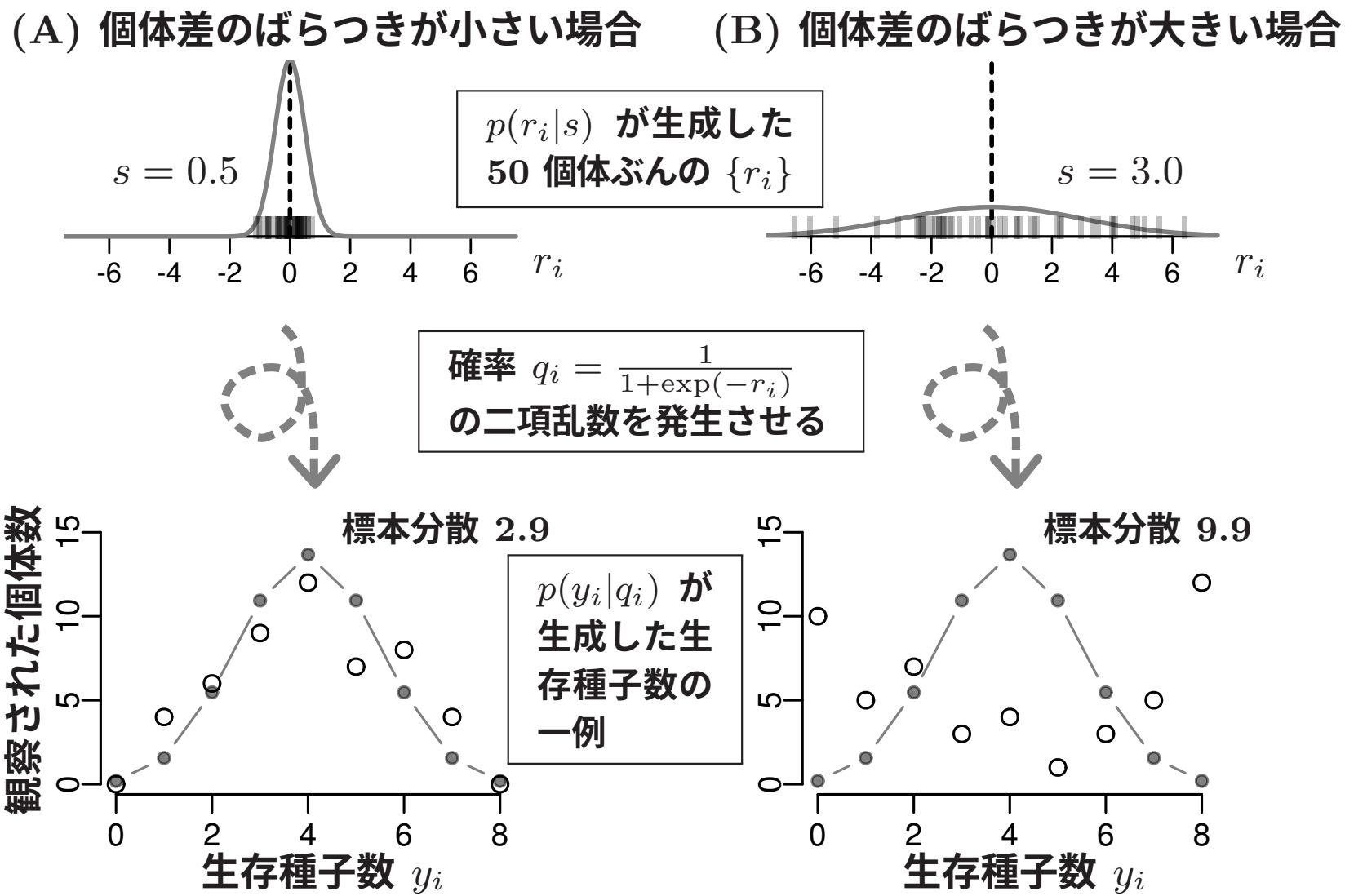
$\{r_i\}$ のばらつきは正規分布だと考えてみる



$$p(r_i \mid s) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{r_i^2}{2s^2}\right)$$

この確率密度 $p(r_i \mid s)$ は r_i の「出現しやすさ」をあらわしていると解釈すればよいでしょう．r_i がゼロにちかい児童はわりと「ありがち」で，r_i の絶対値が大きな児童は相対的に「あまりいない」.

児童差 r_i の分布と過分散の関係



固定効果 と ランダム効果

“伝統的” なんだけど，もうヤメてもらいたい呼びかた

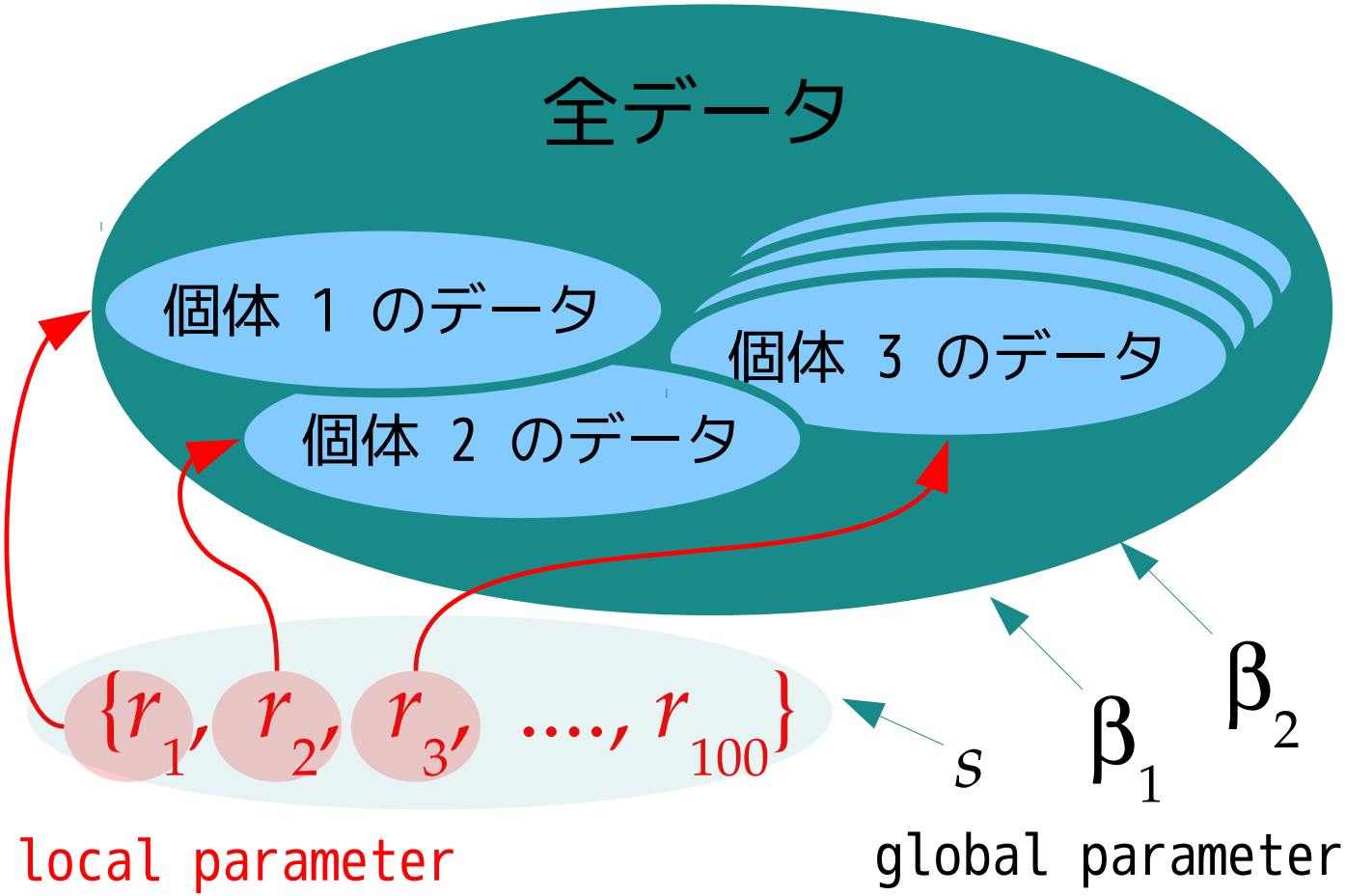
Generalized Linear Mixed Model (GLMM)

で使う Mixed **な 線形予測子**: $\beta_1 + \beta_2 x_i + r_i$

- fixed effects: $\beta_1 + \beta_2 x_i$
- random effects: $+r_i$

fixed? random? **よくわからん……?**

統計モデルの大域的・局所的なパラメーター



データのどの部分を説明しているのか?

global parameter と local parameter

Generalized Linear Mixed Model (GLMM)

で使う Mixed **な 線形予測子**: $\beta_1 + \beta_2 x_i + r_i$

- fixed effects: $\beta_1 + \beta_2 x_i$
  - global parameter — **全児童に共通**
- **全児童のばらつき** s **も** global parameter
- random effects: $+r_i$
  - local parameter — **児童** i **だけを説明**

10. 一般化線形混合モデル (GLMM) を作って推定

個体差 r_i を積分して消す尤度方程式

global parameter と local parameter

Generalized Linear Mixed Model (GLMM)

で使う Mixed **な 線形予測子**: $\beta_1 + \beta_2 x_i + r_i$

- global parameter **は最尤推定できる**
  - fixed effects: β_1, β_2
  - **全児童のばらつき**: s
- local parameter **は最尤推定できない**
  - random effects: $\{r_1, r_2, \cdots, r_{100}\}$

個々の r_i **を最尤推定するには**
データが少なすぎるから

だったら尤度関数の中で r_i を積分してしまえよ!

データ y_i のばらつき — 二項分布

$$p(y_i \mid \beta_1, \beta_2) = \binom{8}{y_i} q_i^{y_i} (1 - q_i)^{8 - y_i}$$

児童差 r_i のばらつき — 正規分布

$$p(r_i \mid s) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{r_i^2}{2s^2}\right)$$

児童 i の 尤度 — r_i を消す

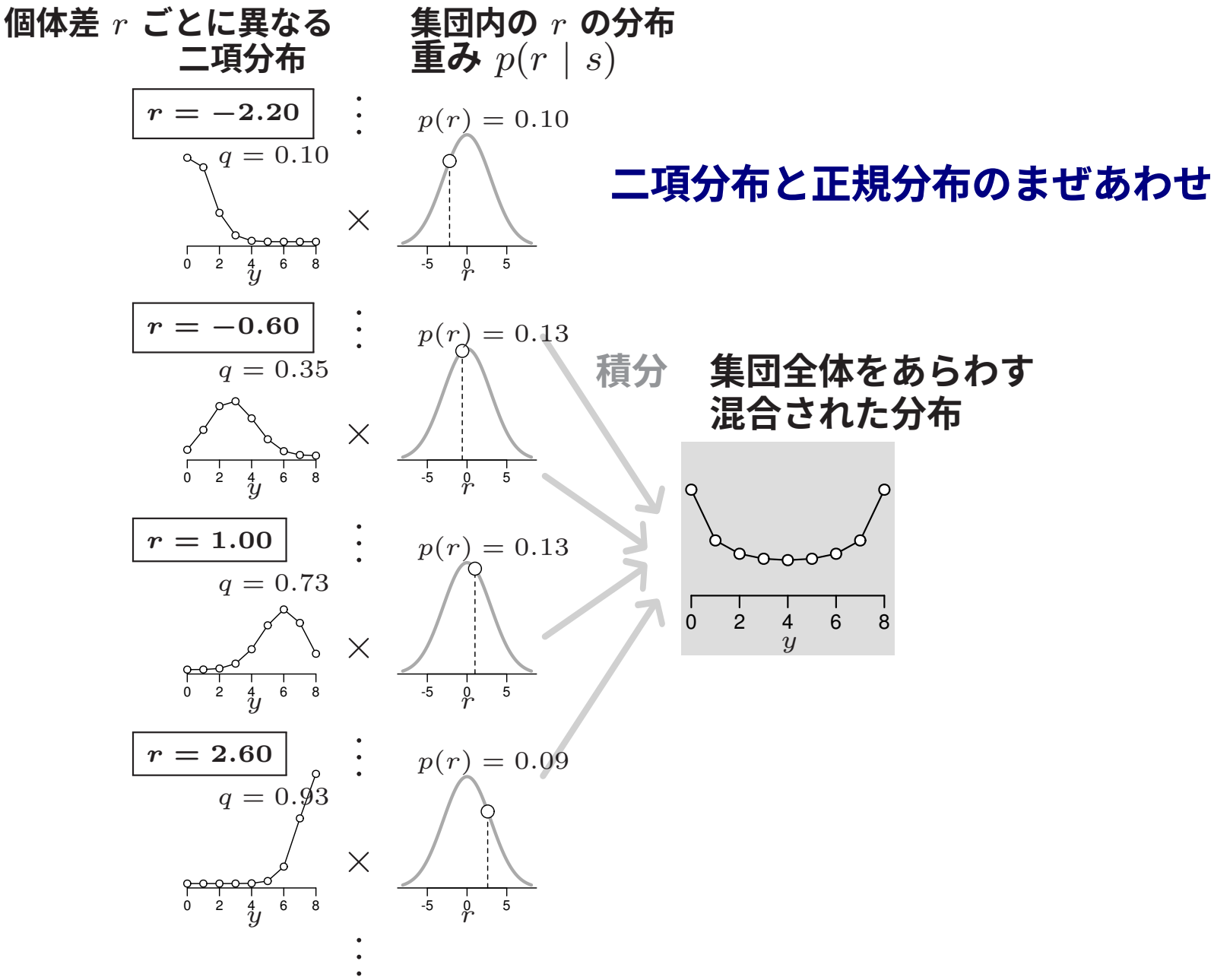
$$L_i = \int_{-\infty}^{\infty} p(y_i \mid \beta_1, \beta_2, r_i)\ p(r_i \mid s) dr_i$$

全データの尤度 — β_1, β_2, s の関数

$$L(\beta_1, \beta_2, s) = \prod_i L_i$$

児童差 r_i について積分するということは二項分布と正規分布をまぜあわせること

Integral of $r_i \to$ mixture distribution of the binomial and Gaussian distributions



二項分布と正規分布のまぜあわせ

glmmML package を使って GLMM の推定

```
> install.packages("glmmML") # if you don't have glmmML
> library(glmmML)
> glmmML(cbind(y, N - y) ~ x, data = d, family = binomial
+ cluster = id)

> d <- read.csv("data.csv")
> head(d)
  N y x id
1 8 0 2  1
2 8 1 2  2
3 8 2 2  3
4 8 4 2  4
5 8 1 2  5
6 8 0 2  6
```

GLMM の推定値: $\hat{\beta}_1, \hat{\beta}_2, \hat{s}$

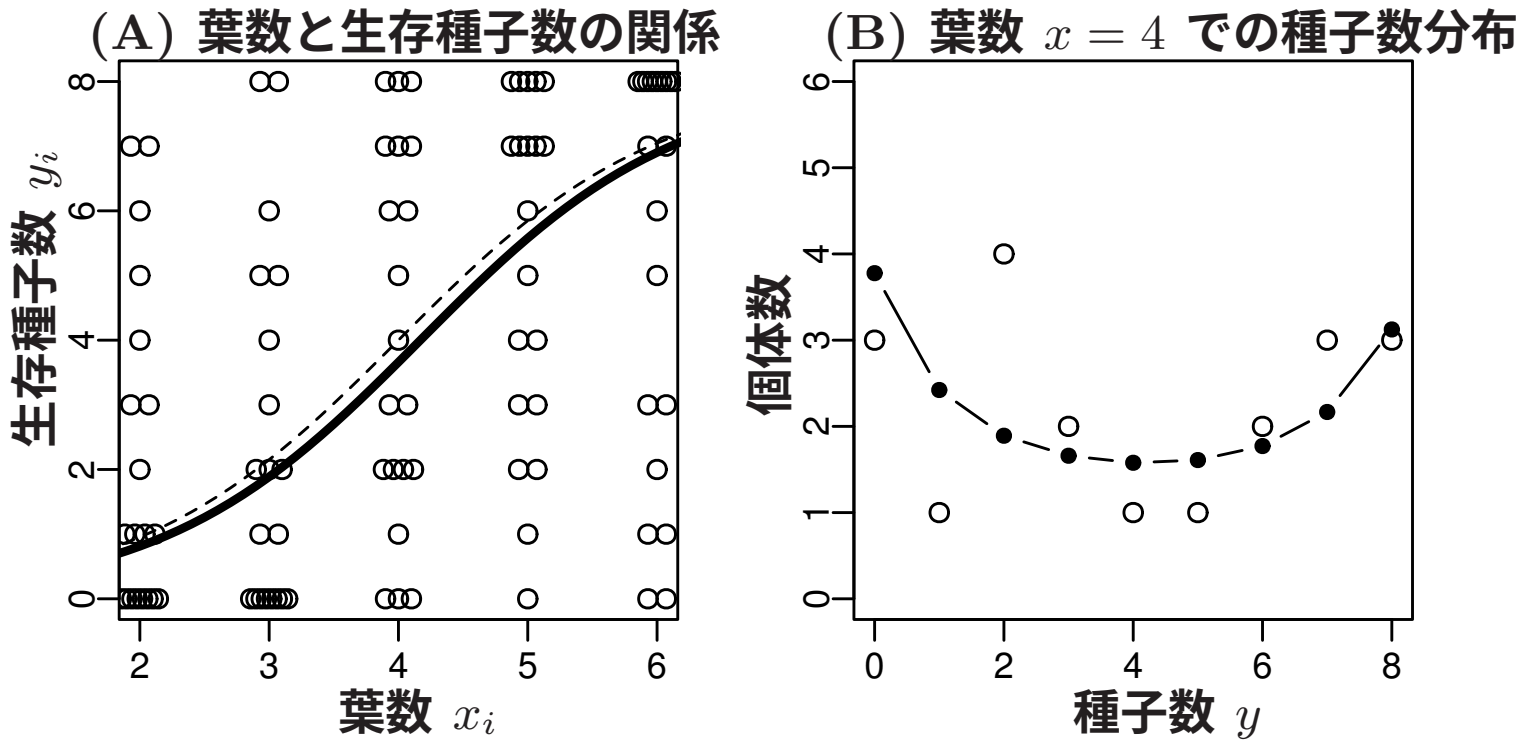
```
> glmmML(cbind(y, N - y) ~ x, data = d, family = binomial,
+ cluster = id)
...(snip)...
             coef se(coef)     z Pr(>|z|)
(Intercept) -4.13    0.906 -4.56  5.1e-06
x            0.99    0.214  4.62  3.8e-06

Scale parameter in mixing distribution:  2.49 gaussian
Std. Error:                              0.309

Residual deviance: 264 on 97 degrees of freedom  AIC: 270
```

$$\hat{\beta}_1 = -4.13, \hat{\beta}_2 = 0.99, \hat{s} = 2.49$$

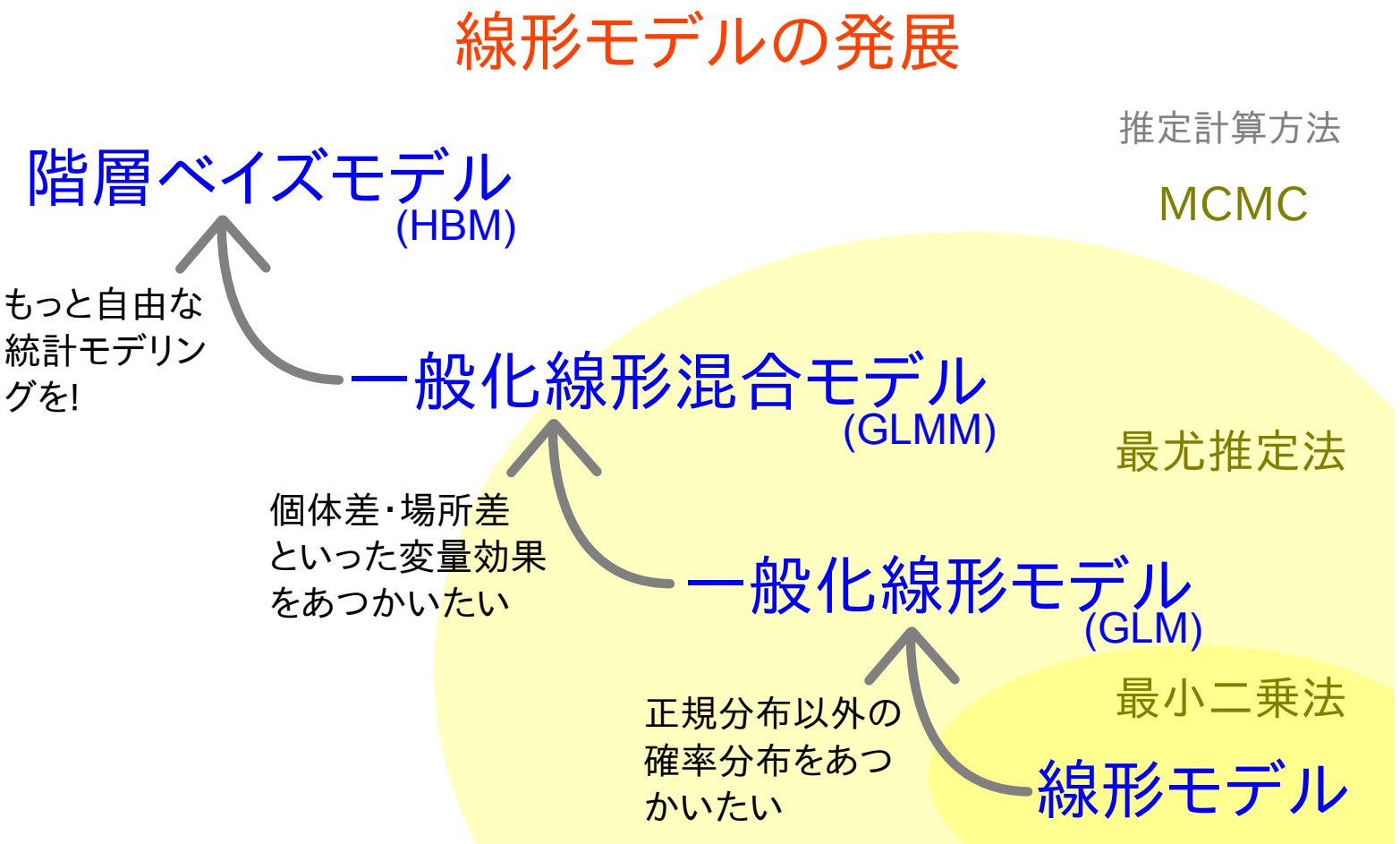
推定された GLMM を使った予測



11. 現実のデータ解析には GLMM が必要

個体差・グループ差を考えないといけないから

“統計モデリング入門” に登場する統計モデル



データの特徴にあわせて線形モデルを改良・発展させる

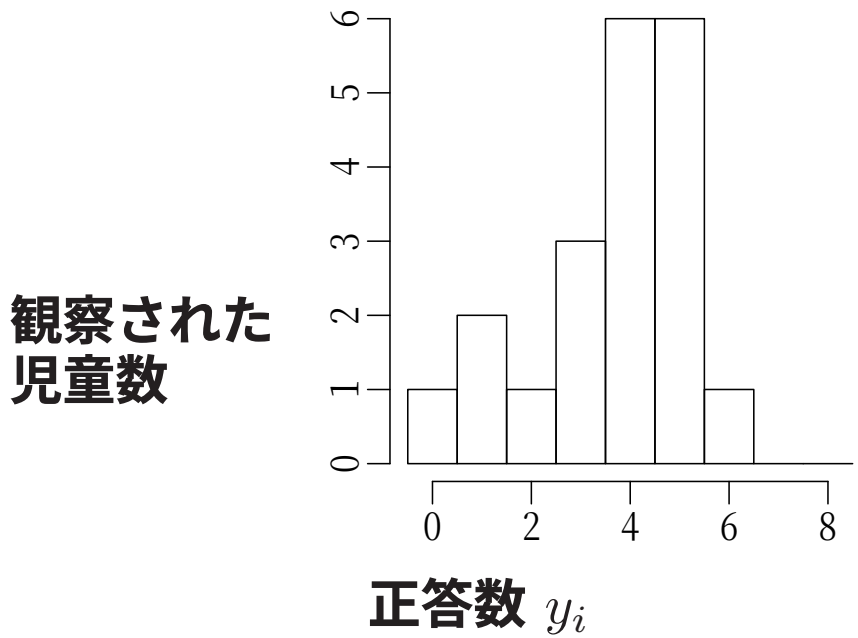
12. MCMC サンプリングのための例題

二項分布のパラメーターを最尤推定 (同じもの再掲)

前の時間の最初のハナシを少しくりかえします

簡単すぎる例題: 正解確率は全児童で同じ (「児童差」なし)

児童ごとの正解数	0	1	2	3	4	5	6	7	8
観察された児童数	1	2	1	3	6	6	1	0	0



これは児童差なしの均質な集団

正解確率 q と二項分布の関係

- 正解確率を推定するために**二項分布** という確率分布を使う
- 児童 i の N_i 回答中 y_i 個が正解する確率

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

- ここで仮定していること
  - **児童差はない**
  - つまり **すべての児童で同じ正解確率** q

最尤推定 (MLE) とは何か

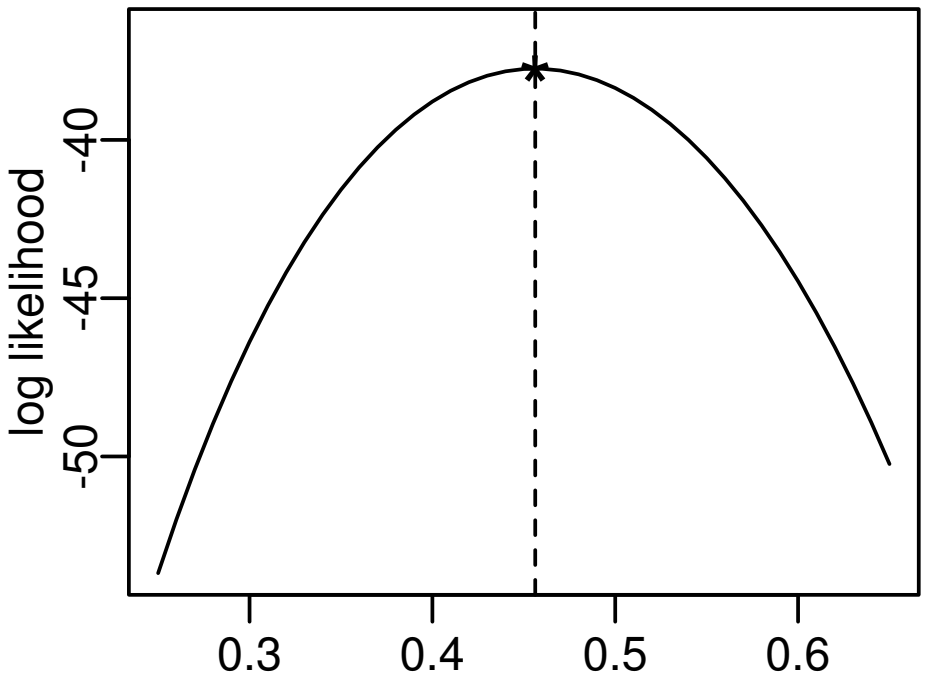
- **対数尤度 $L(q \mid$ データ$)$ が最大になるパラメーター q の値をさがしだすこと**

- **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**

$$\partial \log L(q \mid \text{データ})/\partial q = 0$$

- **正解確率 q が全児童共通の場合の最尤推定量・最尤推定値は**

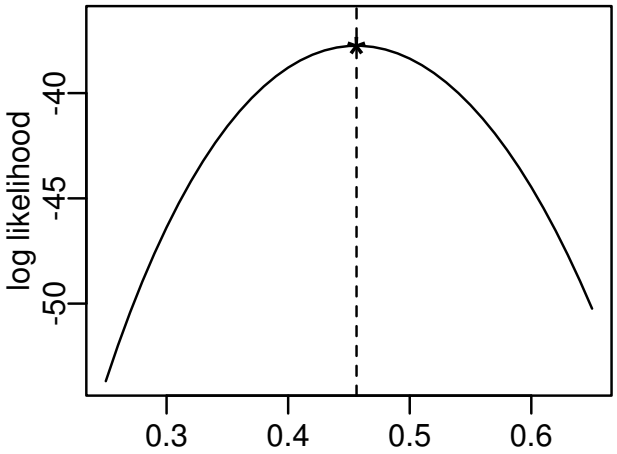
$$\hat{q} = \frac{\text{正答数}}{\text{問題数}} = \frac{73}{160} = 0.456 \text{ ぐらい}$$



最尤推定と MCMC のちがいは何か?

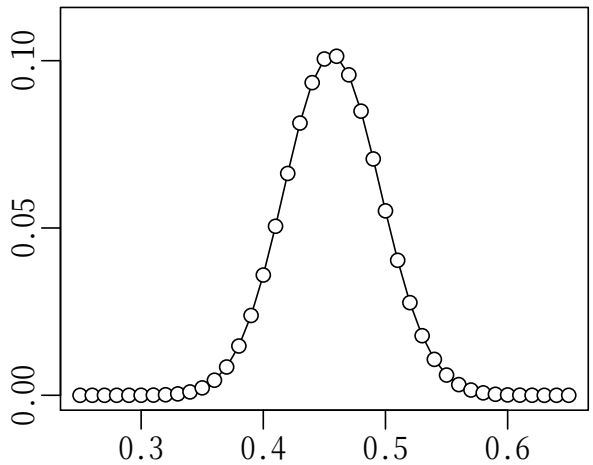
ひたすら “山のぼり” する最尤推定

対数尤度 $\log L(q)$

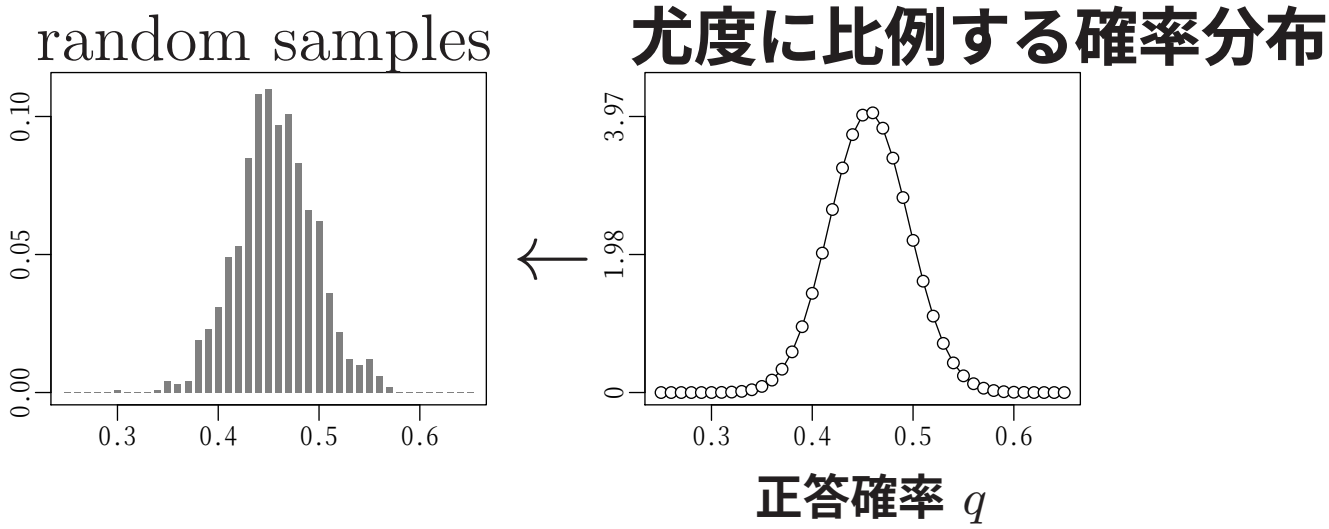


尤度 $L(q)$ **に比例する確率分布**

$p(q)$



MCMC: 尤度に比例する確率分布からのランダムサンプリング



- データ + 統計モデル → (MCMC)

  → $p(q)$ からのランダムサンプル
- このランダムサンプルをもとに，q の平均や 95% 区間などがわかる— 便利じゃないか!

さてさて …… MCMC という**推定方法**
“パラメーター q の確率分布”
というちょっと奇妙な考えかたが
でてきた ……

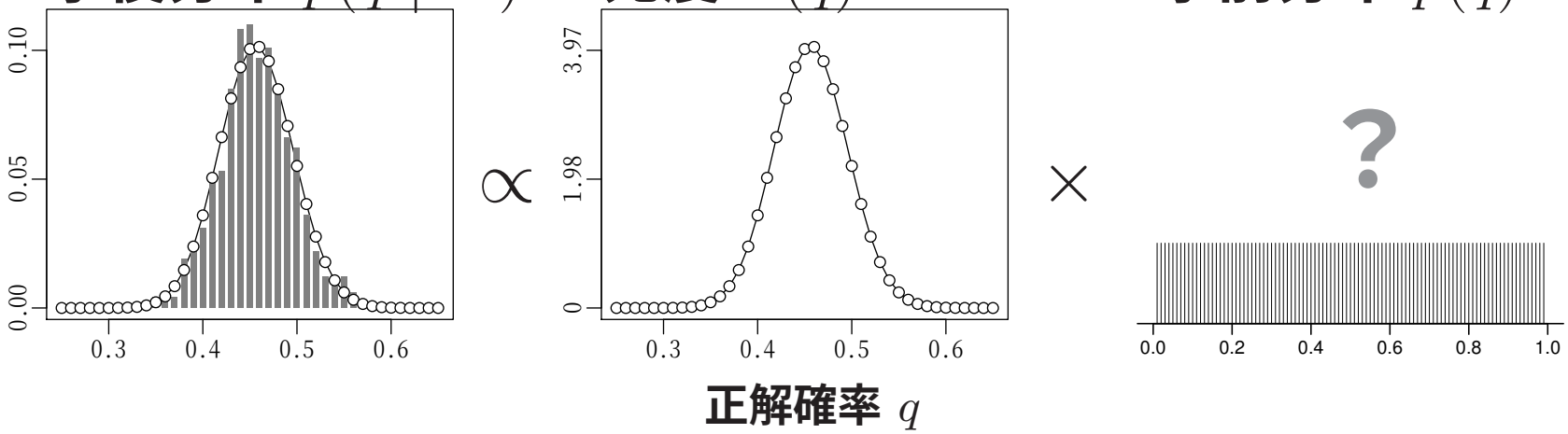
“ふつう”の統計学では
“パラメーターの確率分布”といった
考えかたはしない，しかし ……

ベイズ統計学なら “パラメーターの確率分布” はぜんぜん 自然な考えかただ

ベイズ統計にむりやりこじつけてみると?

q の事前分布は一様分布，と考えるとつじつまがあう?

事後分布 $p(q \mid \mathrm{Y})$ $\propto$ 尤度 $L(q)$ $\times$ 事前分布 $p(q)$



$$(\text{事後分布}) \propto \frac{\text{尤度} \times \text{事前分布}}{(\text{データが得られる確率})}$$

$$\propto \text{尤度} \times \text{事前分布}$$

以上の説明は，

“MCMC によって得られる結果”

は

“ベイズ統計でいうパラメーターの事後分布”

と考えると解釈しやすいかも

といったことを

ばくぜんかつなんとなく対応づける

ひとつのこころみでありました……

厳密な正当化とかそういったものではありません

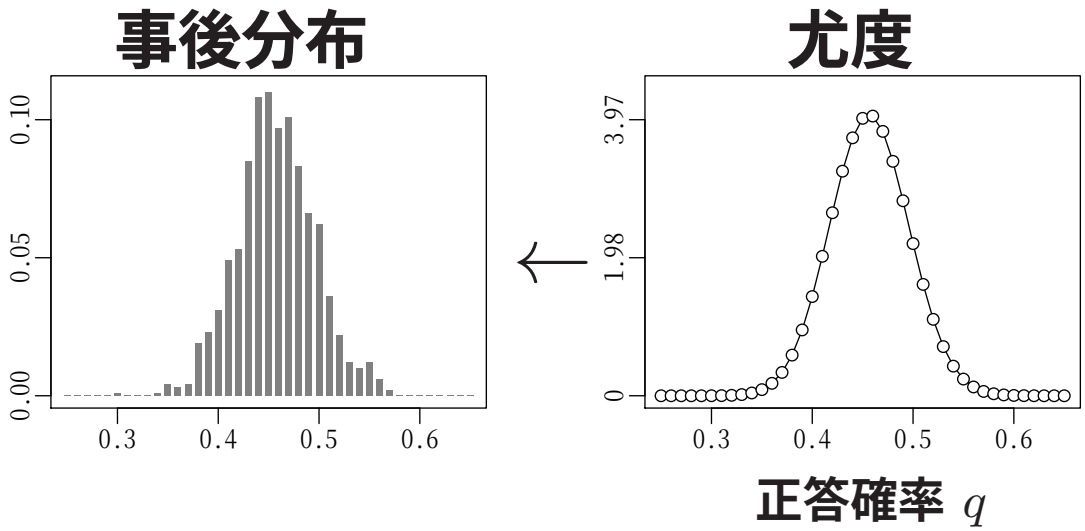
13. MCMC のためのソフトウェア

“Gibbs sampling” などが簡単にできるような……

事後分布から効率よくサンプリングしたい

統計モデルがややこしい，最尤推定がしんどい

→ MCMC で事後分布からサンプリング



めんどくさそう

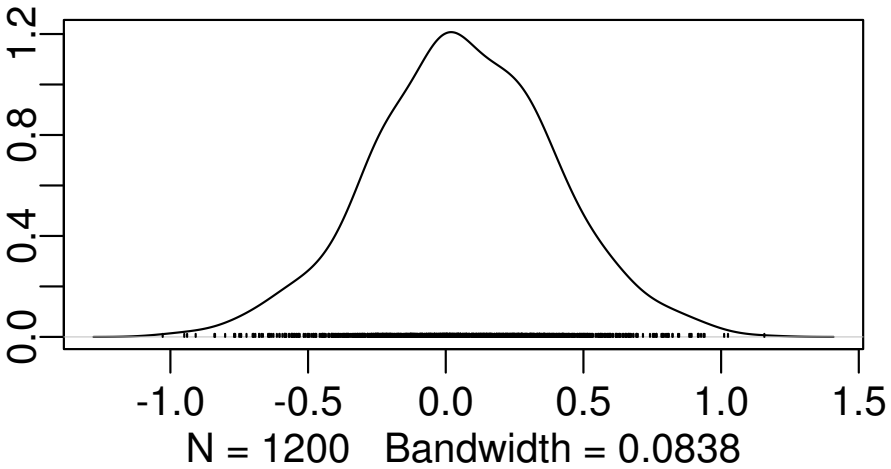
→ こういうことをやってくれるソフトウェアあります

再確認: "事後分布からのサンプル" って何の役にたつの?

こういう乱数のあつまりが得られたときに ……

```
 -0.7592 -0.7689 -0.9008 -1.0160 -0.8439 -1.0380 -0.8561 -0.9837
 -0.8043 -0.8956 -0.9243 -0.9861 -0.7943 -0.8194 -0.9006 -0.9513
 -0.7565 -1.1120 -1.0430 -1.1730 -0.6926 -0.8742 -0.8228 -1.0440
... (以下略) ...
```

これらのサンプルの平均値・中央値・95% 区間を
調べることで事後分布の概要がわかる

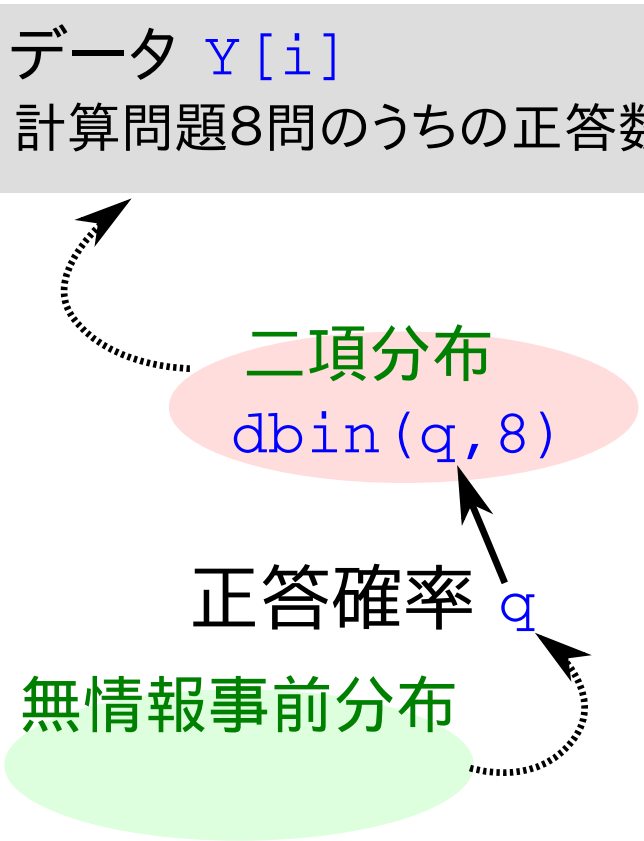


便利な “BUGS” 汎用 Gibbs sampler たち

- BUGS 言語 (+ っぽいもの) でベイズモデルを記述できるソフトウェア
  - WinBUGS — ありがとう，さようなら?
  - OpenBUGS — 予算が足りなくて停滞?
  - **JAGS** — お手軽で良い，どんな OS でも動く
  - Stan — たぶん “次” はこれ
    — 今日は紹介しませんが ……
- リンク集: `http://hosho.ees.hokudai.ac.jp/~kubo/ce/BayesianMcmc.html`

えーと……BUGS 言語って何?

このベイズモデルを BUGS 言語で記述したい



BUGS 言語コード

```
for (i in 1:N.sample) {
  Y[i] ~ dbin(q, 8)
}
q ~ dunif(0.0, 1.0)
```

矢印は手順ではなく，依存関係をあらわしている

BUGS 言語: **ベイズモデルを記述する言語**

Spiegelhalter et al. 1995. BUGS: Bayesian Using Gibbs Sampling version 0.50.

いろいろな OS で使える JAGS3.4.0

- R core team のひとり Martyn Plummer さんが開発
  - Just Another Gibbs Sampler
- C++ で実装されている
  - R がインストールされていることが必要
- Linux, Windows, Mac OS X バイナリ版もある
- ぢりぢりと開発進行中
- R からの使う: `library(rjags)`

14. GLMM と階層ベイズモデル

GLMM のベイズモデル化

階層ベイズモデルとなる

Q. 今日のハナシは “GLMM の紹介” なのに
なぜ階層ベイズモデル (HBM) なんかも
説明するのか?

A. “研究の道具” として使うためには
GLMM を**ベイズモデル化**した
HBM が必要になるから

補足 A1. GLMM + 最尤推定は “児童差” はあつかえるが “児童差 + 学校差 + ⋯” は難しい
最尤推定やめて頑健な MCMC を使いたい
MCMC 使うなら GLMM ベイズモデル化

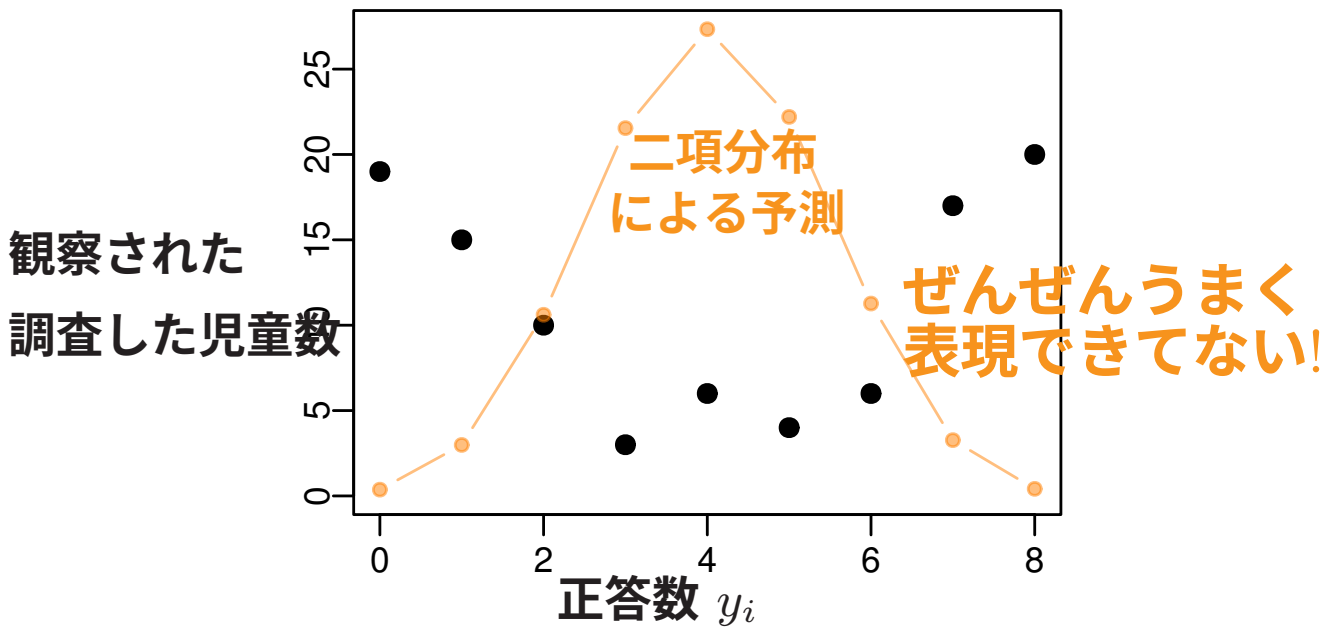
補足 A2. GLMM + **最尤推定は “学校差” など積分してしまう……**

“学校差” の推定もしたい

階層ベイズモデル + MCMC **なら** OK!

二項分布では説明できない観測データ!

100 児童 × 8 問 → 合計 800 問中 403 個の正答だったので，平均正答確率は 0.50 と推定されたが……



“個体差” があると overdispersion が生じる

モデリングやりなおし: 個体差を考慮する

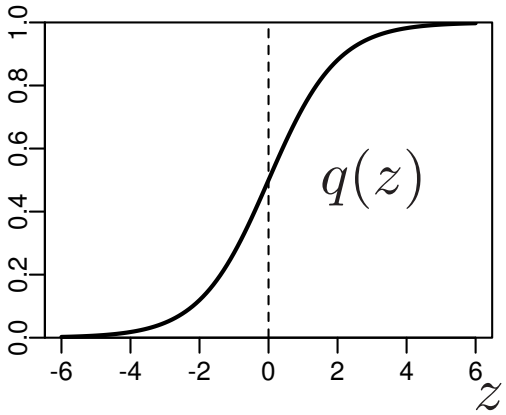
- 正答確率を推定するために 二項分布という確率分布を使う
- 児童 i の回答 N_i 問のうち y_i 個が正解である確率は二項分布

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

- ここで仮定していること
  - 個体差があるので児童ごとに正答確率 q_i が異なる

GLM わざ: ロジスティック関数で表現する正答確率

- **正答確率** $q_i = q(z_i)$ **をロジスティック関数** $q(z) = 1/\{1 + \exp(-z)\}$ **で表現**



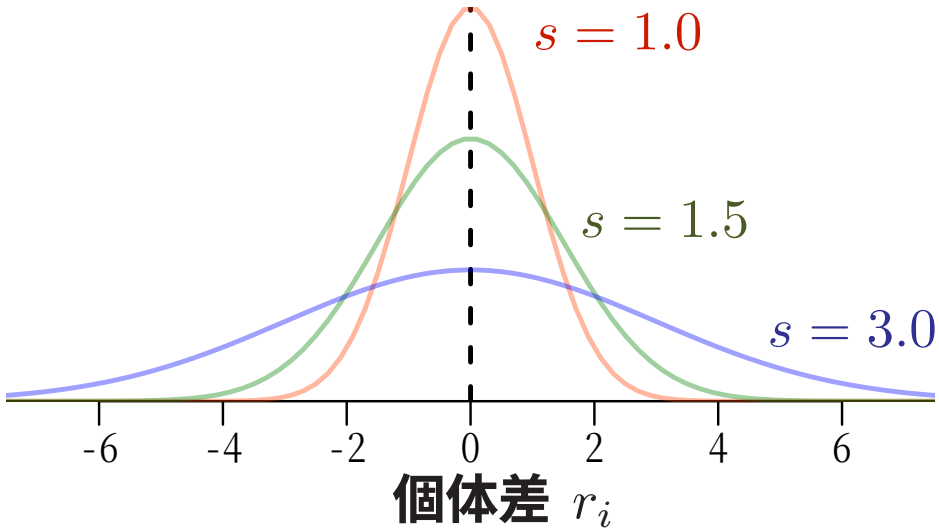
- **線形予測子** $z_i = a + r_i$ **とする**
  - **パラメーター** a**: 全体の平均**
  - **パラメーター** r_i**: 個体** i **の個体差 (ずれ)**

個々の個体差 r_i を最尤推定するのはまずい

- 100 個体の正答確率を推定するためにパラメーター 101 個 (a と $\{r_1, r_2, \cdots, r_{100}\}$) を推定すると……
- 個体ごとに正答数 / 問題数を計算していることと同じ! (「データのよみあげ」と同じ)

そこで，次のように考えてみる

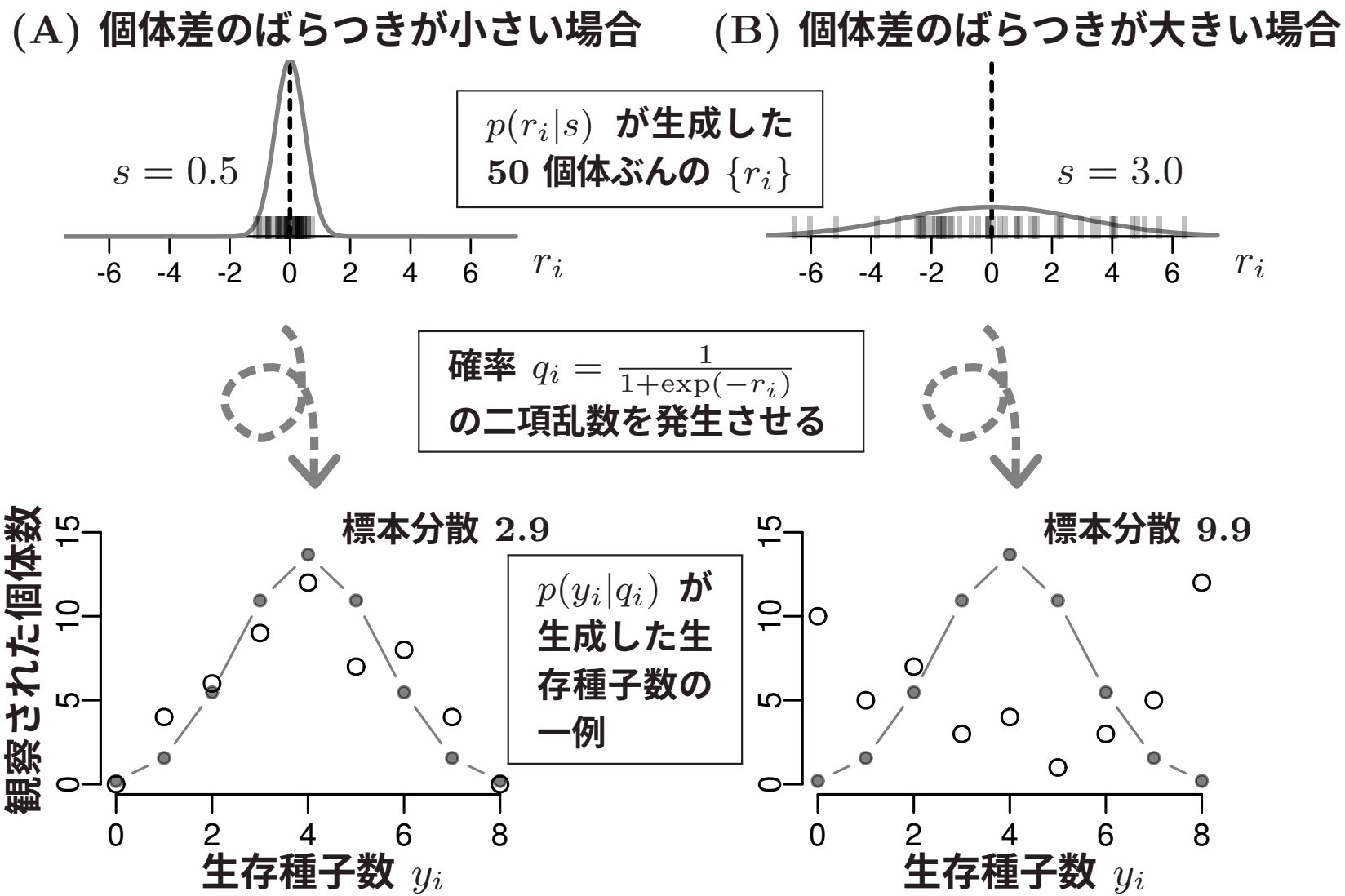
$\{r_i\}$ のばらつきは正規分布だと考えてみる



$$p(r_i \mid s) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{r_i^2}{2s^2}\right)$$

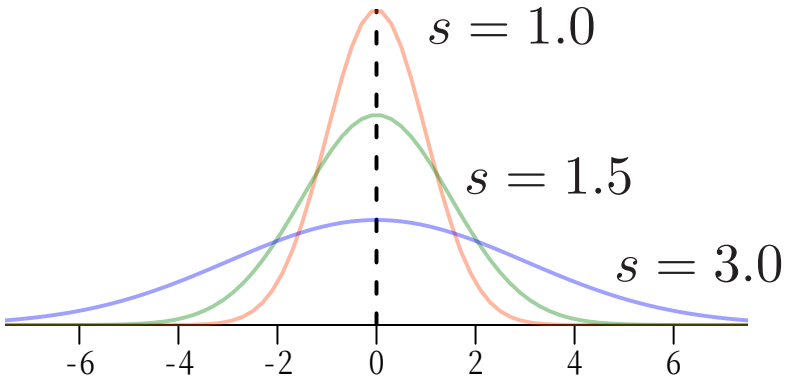
この確率密度 $p(r_i \mid s)$ は r_i の「出現しやすさ」をあらわしていると解釈すればよいでしょう．r_i がゼロにちかい個体はわりと「ありがち」で，r_i の絶対値が大きな個体は相対的に「あまりいない」．

ひとつの例示: 個体差 r_i の分布と過分散の関係



これは r_i の事前分布の指定，ということ

前回の授業で $\{r_i\}$ は正規分布にしたがうと仮定したが
ベイズ統計モデリングでは 「100 個の r_i たちに
共通する事前分布として正規分布 を指定した」
ということになる

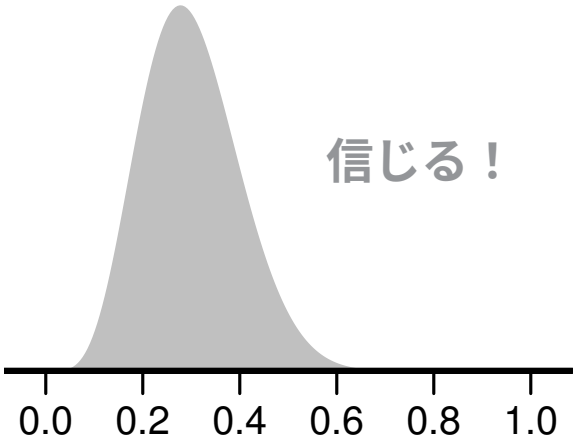


$$p(r_i \mid s) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{r_i^2}{2s^2}\right)$$

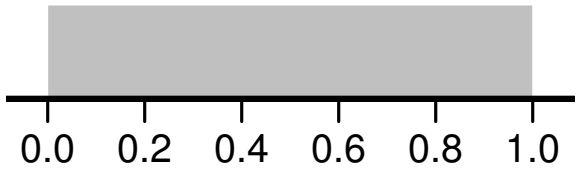
ベイズ統計モデルでよく使われる三種類の事前分布

たいていのベイズ統計モデルでは，ひとつのモデルの中で複数の種類の事前分布を混ぜて使用する．

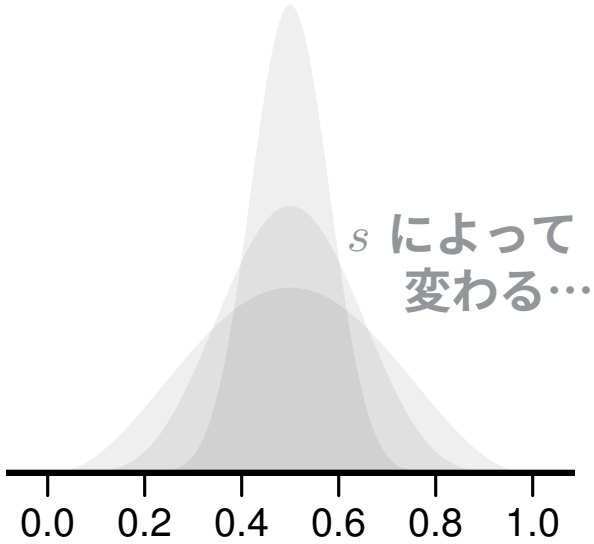
(A) **主観的な事前分布**
(できれば使いたくない!)



(B) **無情報事前分布**



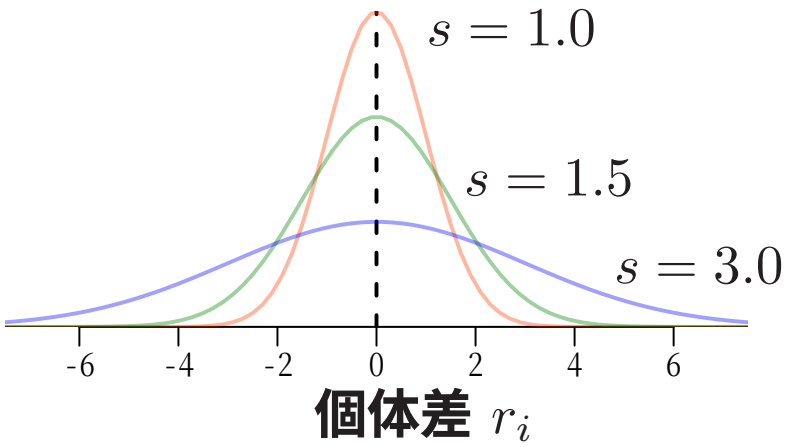
(C) **階層事前分布**



r_i の事前分布として階層事前分布を指定する

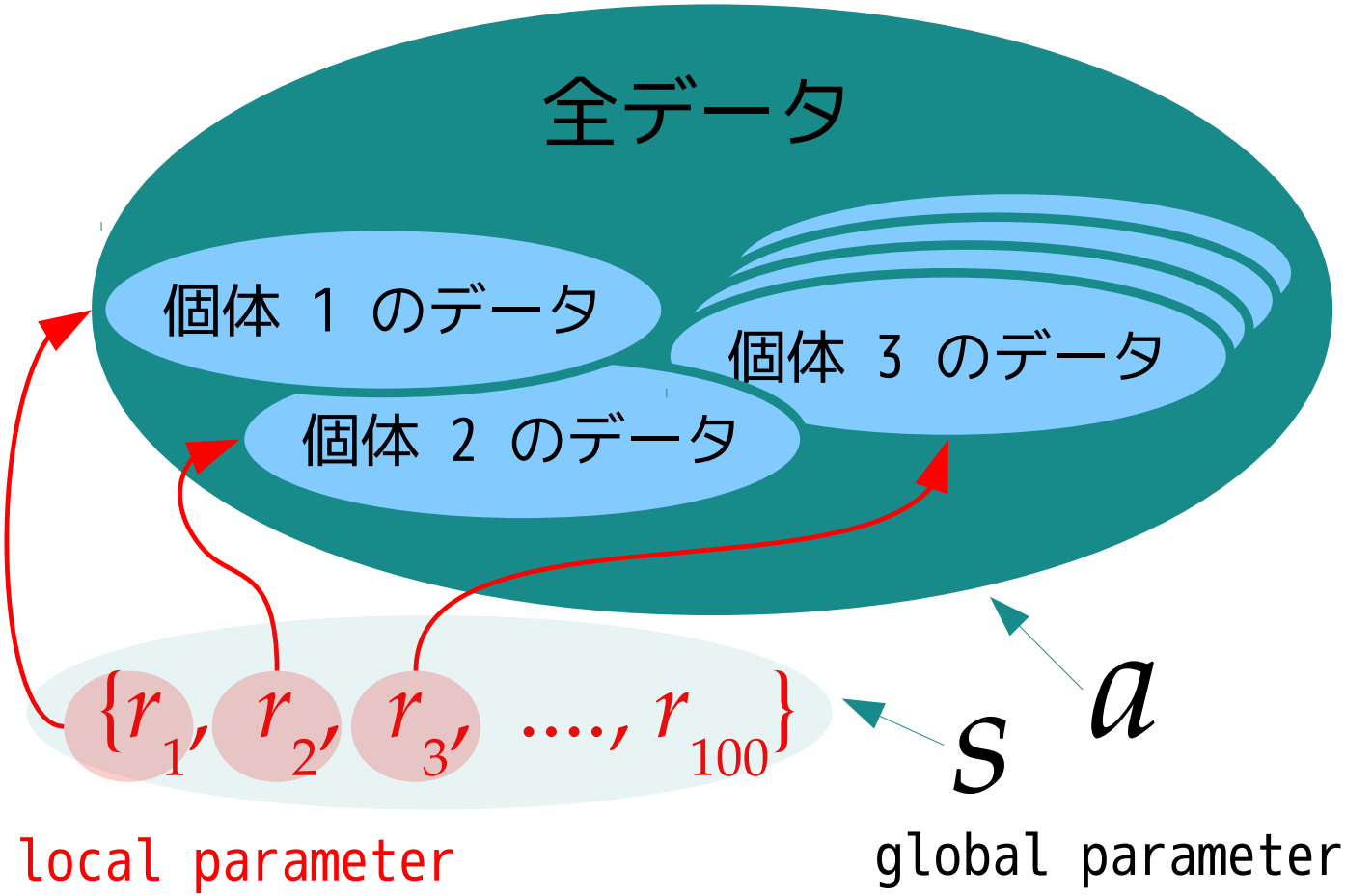
階層事前分布の利点

「データにあわせて」事前分布が変形!



$$p(r_i \mid s) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{r_i^2}{2s^2}\right)$$

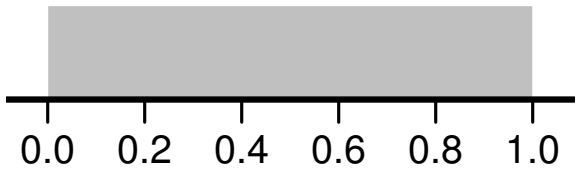
統計モデルの大域的・局所的なパラメーター



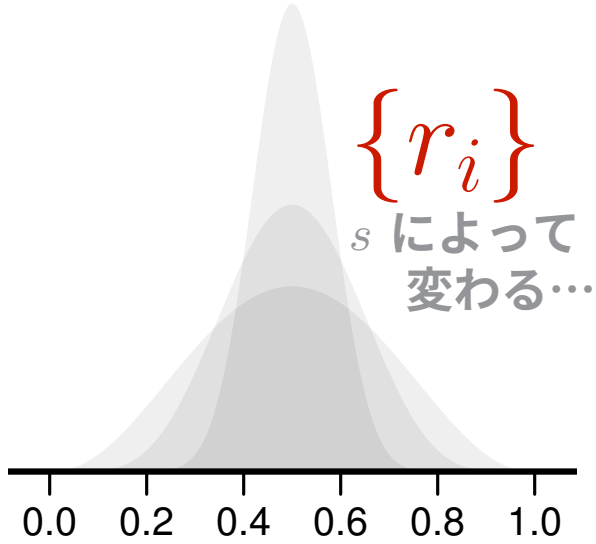
データのどの部分を説明しているのか?

パラメーターごとに適切な事前分布を選ぶ

(B) 無情報事前分布

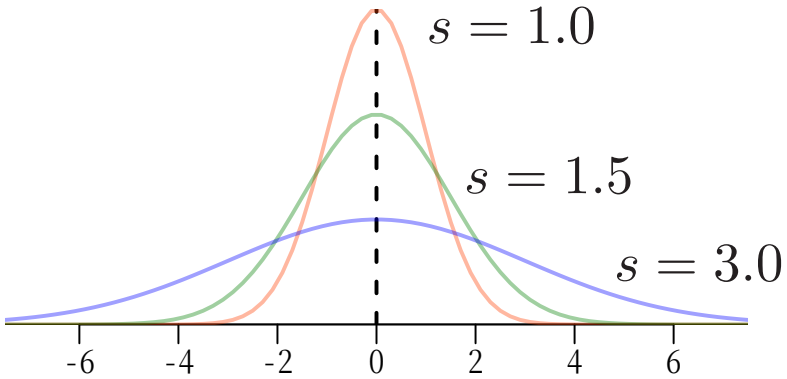


(C) 階層事前分布



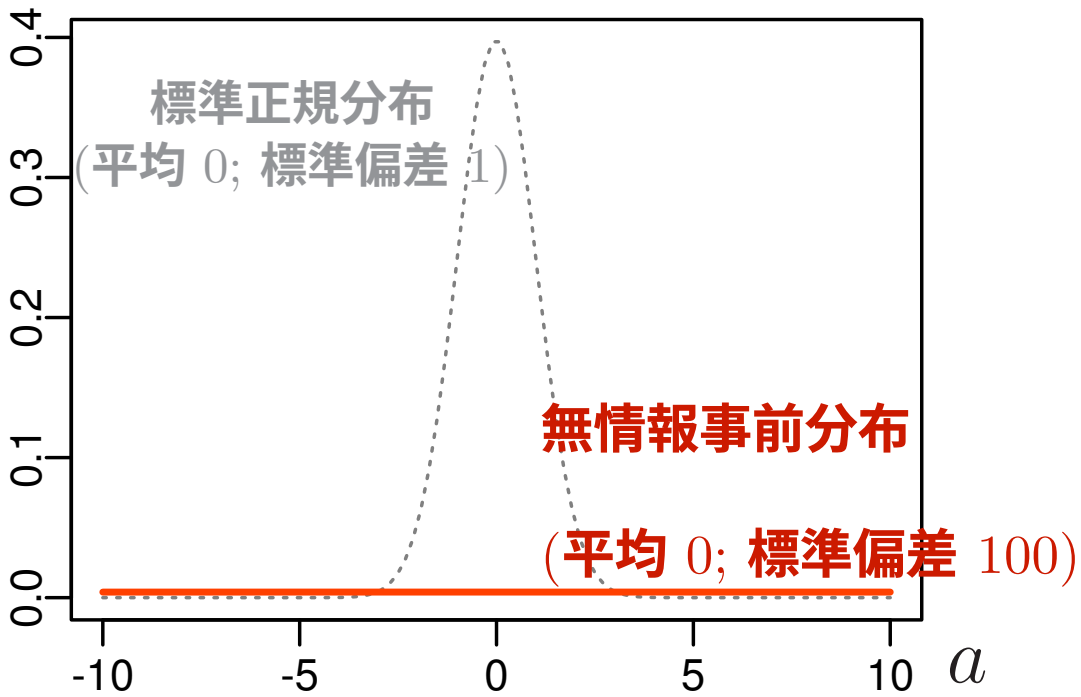
パラメーターの種類	説明する範囲	事前分布
全体に共通する平均・ばらつき	大域的	無情報事前分布
個体・グループごとのずれ	局所的	階層事前分布

個体差 $\{r_i\}$ のばらつき s の無情報事前分布



- s はどのような値をとってもかまわない
- そこで s の事前分布は **無情報事前分布** (non-informative prior) とする
- たとえば一様分布，ここでは $0 < s < 10^4$ の一様分布としてみる

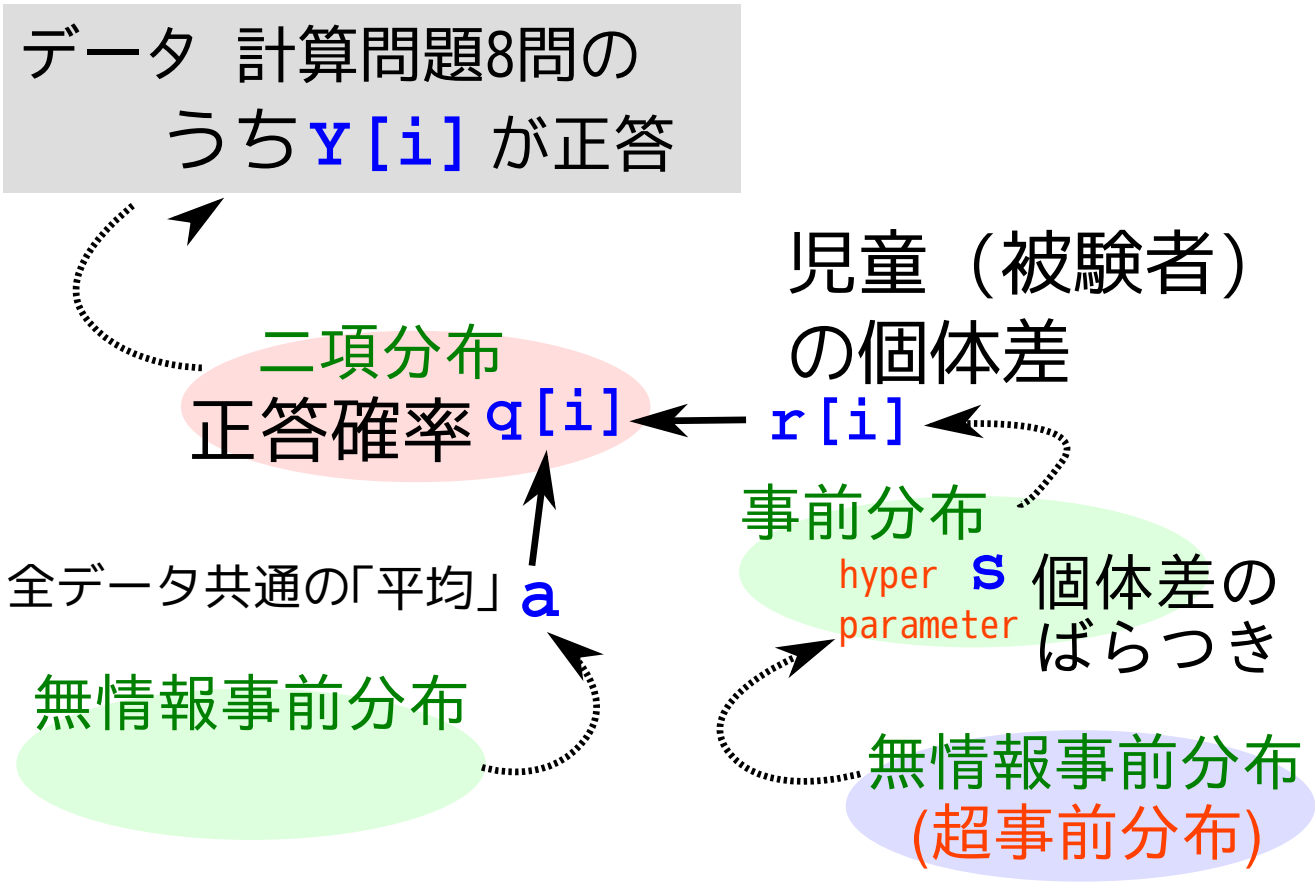
全個体の「切片」 a の無情報事前分布



「正答確率の (logit) 平均 a は何でもよい」と表現している

階層ベイズモデル: 事前分布の階層性

超事前分布 → 事前分布という階層があるから



矢印は手順ではなく，依存関係をあらわしている

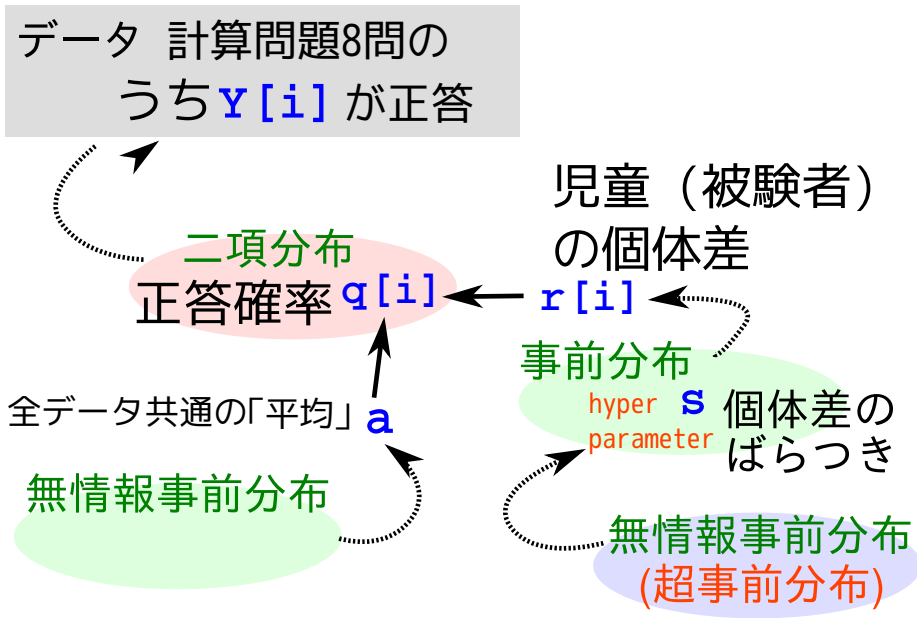
15. 階層ベイズモデルの推定

ソフトウェア JAGS を使ってみる

R の “したうけ” として JAGS を使う

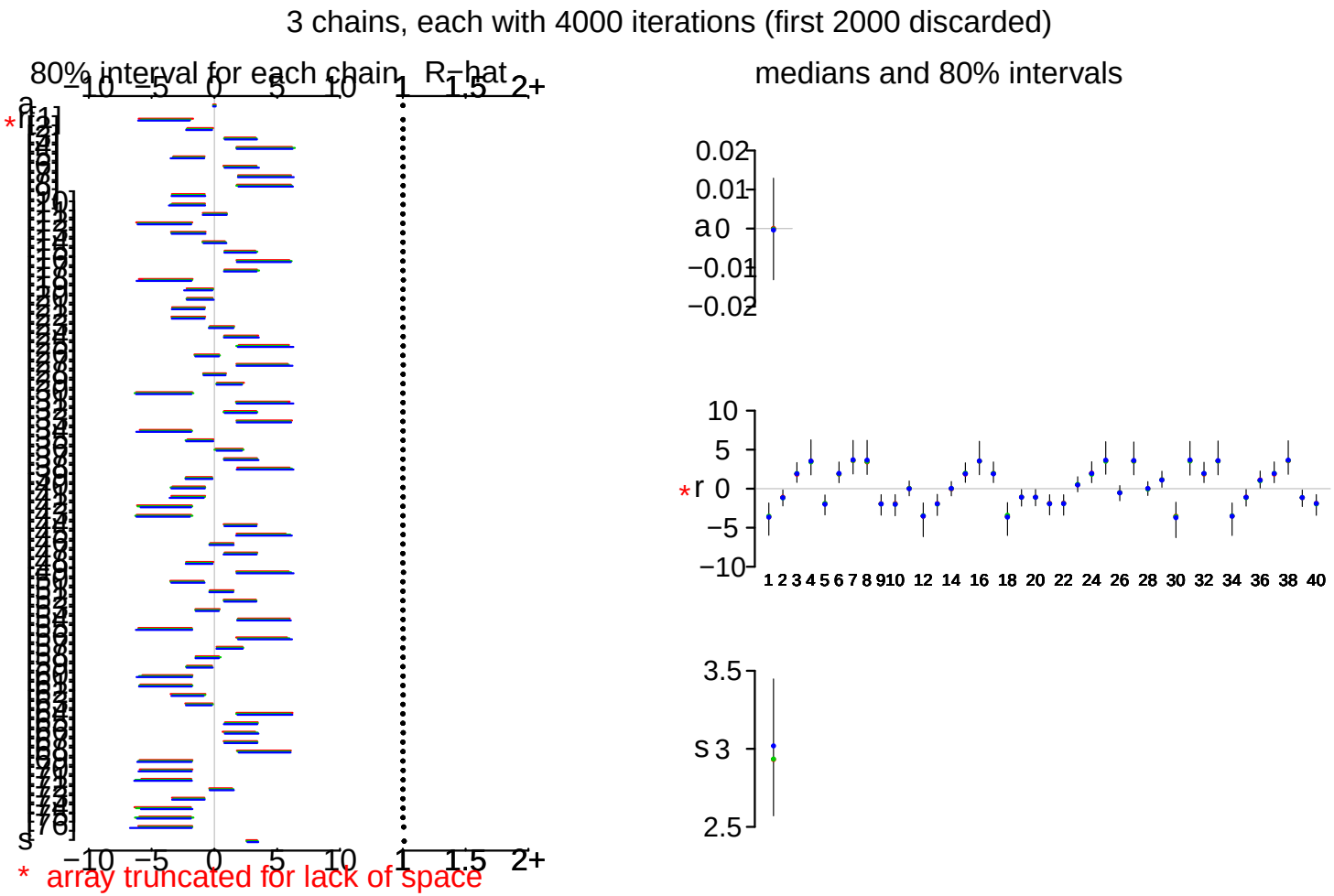
階層ベイズモデルを BUGS コードで記述する

```
model
{
  for (i in 1:N.data) {
    Y[i] ~ dbin(q[i], 8)
    logit(q[i]) <- a + r[i]
  }
  a ~ dnorm(0, 1.0E-4)
  for (i in 1:N.data) {
    r[i] ~ dnorm(0, tau)
  }
  tau <- 1 / (s * s)
  s ~ dunif(0, 1.0E+4)
}
```



JAGS で得られた事後分布サンプルの要約

```
> source("mcmc.list2bugs.R")                    # なんとなく便利なので…
> post.bugs <- mcmc.list2bugs(post.mcmc.list) # bugs クラスに変換
```

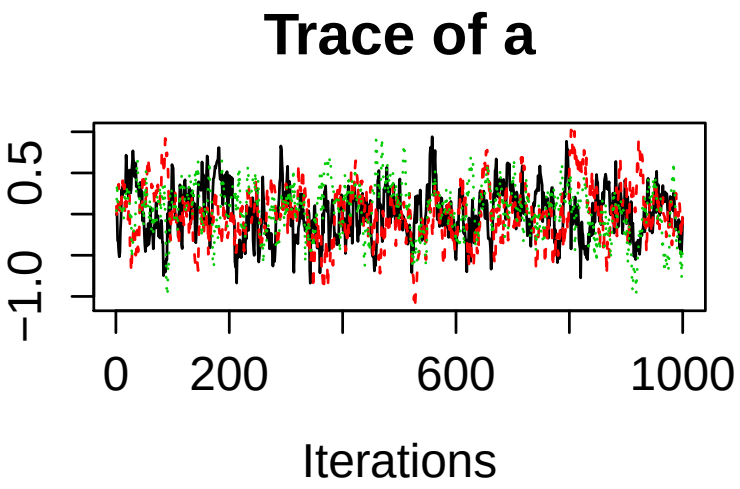


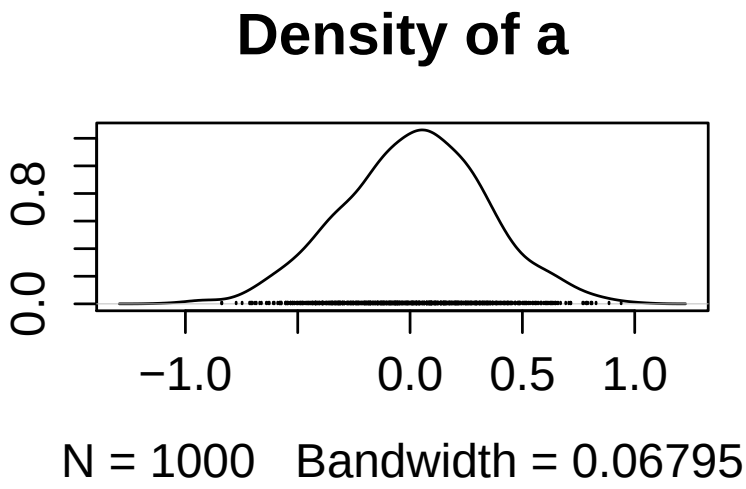
bugs オブジェクトの post.bugs を調べる

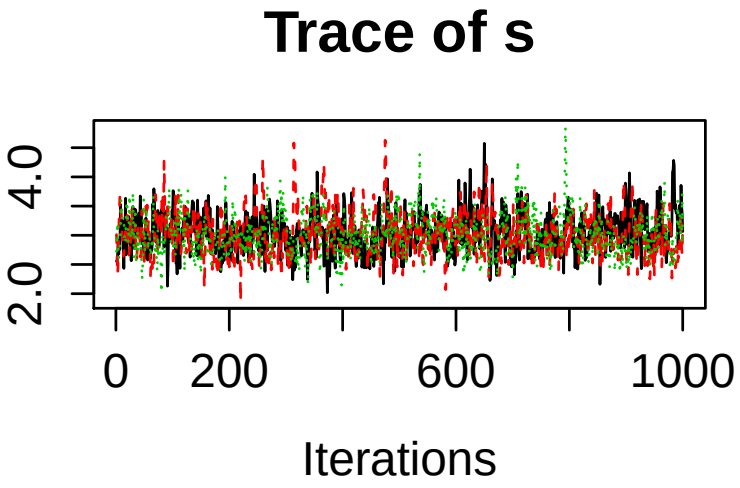
- print(post.bugs, digits.summary = 3)
- **事後分布の 95% 信頼区間などが表示される**

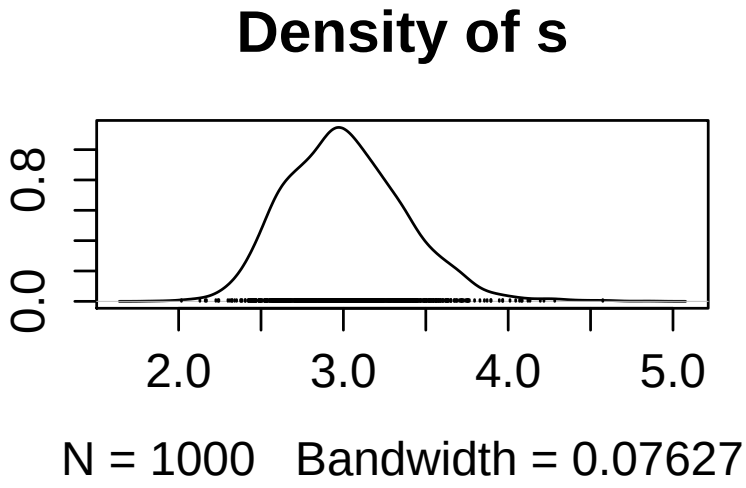
```
 3 chains, each with 4000 iterations (first 2000 discarded), n.thin = 2
 n.sims = 3000 iterations saved
         mean    sd   2.5%    25%    50%    75%  97.5%  Rhat n.eff
a       0.020 0.321 -0.618 -0.190  0.028  0.236  0.651 1.007   380
s       3.015 0.359  2.406  2.757  2.990  3.235  3.749 1.002  1200
r[1]   -3.778 1.713 -7.619 -4.763 -3.524 -2.568 -1.062 1.001  3000
r[2]   -1.147 0.885 -2.997 -1.700 -1.118 -0.531  0.464 1.001  3000
r[3]    2.014 1.074  0.203  1.282  1.923  2.648  4.410 1.001  3000
r[4]    3.765 1.722  0.998  2.533  3.558  4.840  7.592 1.001  3000
r[5]   -2.108 1.111 -4.480 -2.775 -2.047 -1.342 -0.164 1.001  2300
... (中略)
r[99]   2.054 1.103  0.184  1.270  1.996  2.716  4.414 1.001  3000
r[100] -3.828 1.766 -7.993 -4.829 -3.544 -2.588 -1.082 1.002  1100
```

各パラメーターの事後分布サンプルを R で調べる



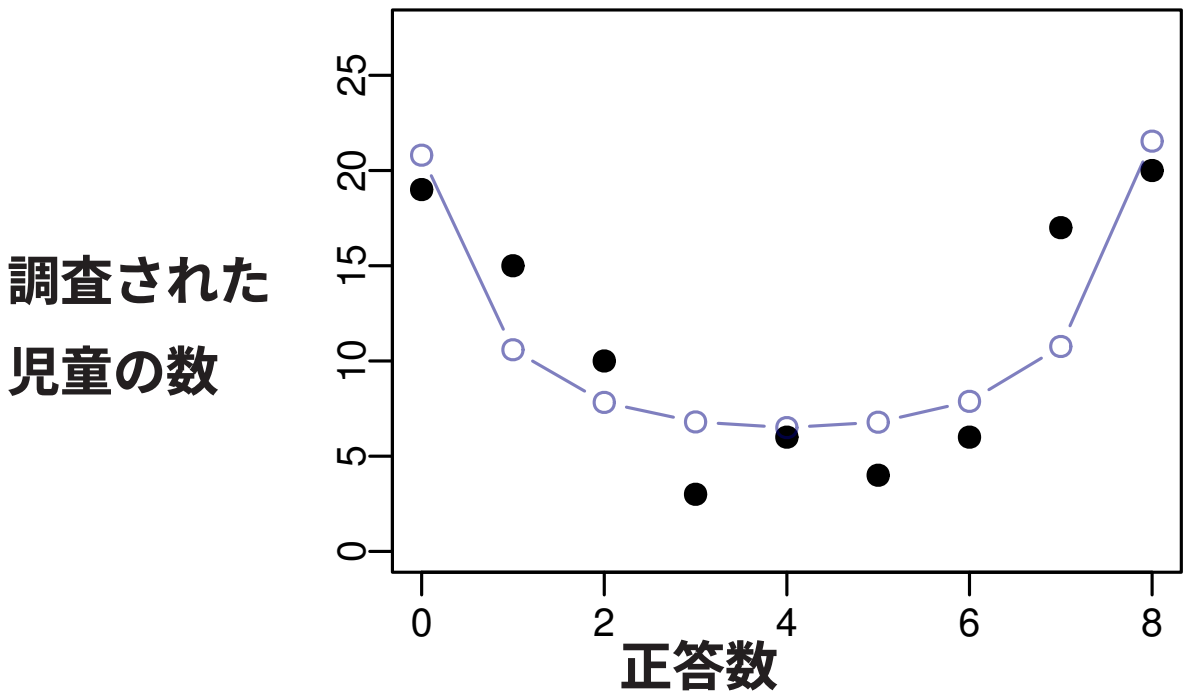






得られた事後分布サンプルを組みあわせて予測

- `post.mcmc <- to.mcmc(post.bugs)`
- **これは `matrix` と同じようにあつかえるので，作図に便利**



16. 複数ランダム効果の階層ベイズモデル

個体差 + グループ差，など

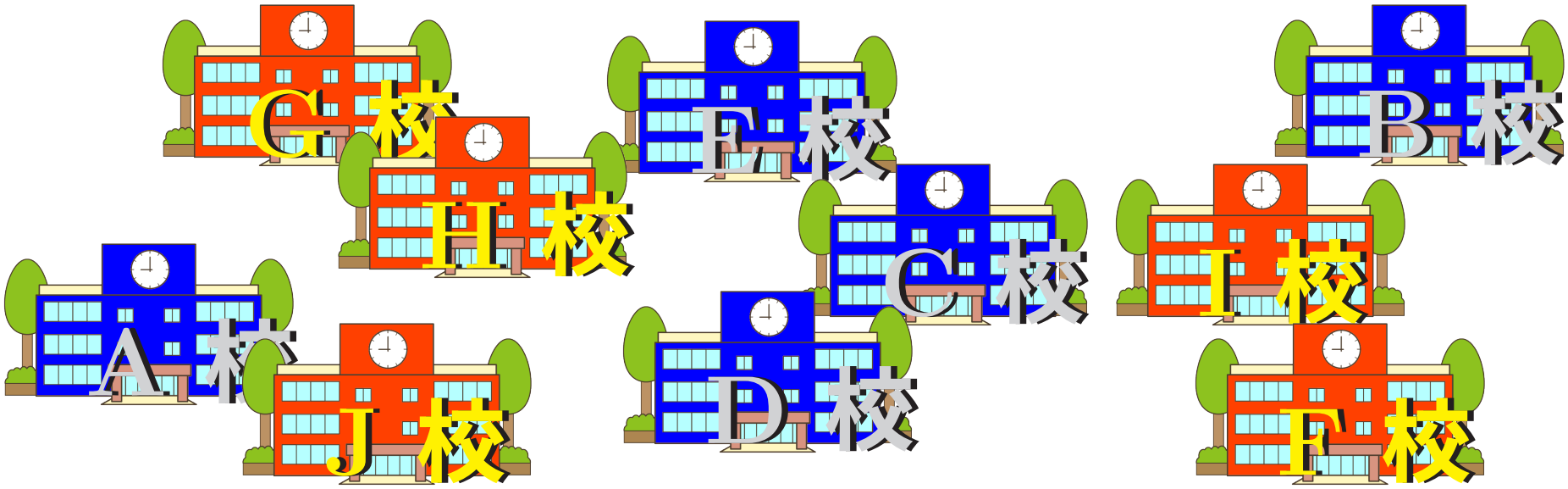
そして“てぬき”モデリングの危なさについて

データはこのように格納されている

```
> d <- read.csv("d1.csv")
> head(d)
  id pot f  y
1  1   A C  6
2  2   A C  3
3  3   A C 19
4  4   A C  5
5  5   A C  0
6  6   A C 19
```

- id **列**: **児童番号** $\{1, 2, 3, \cdots, 100\}$
- pot **列**: **学校名** $\{\mathrm{A}, \mathrm{B}, \mathrm{C}, \cdots, \mathrm{J}\}$
- f **列**: **無処理** C, **教育法** F T
- y **列**: **回答数** (**応答変数**)

架空教育法の例題: 複数の学校で計算能力を調査

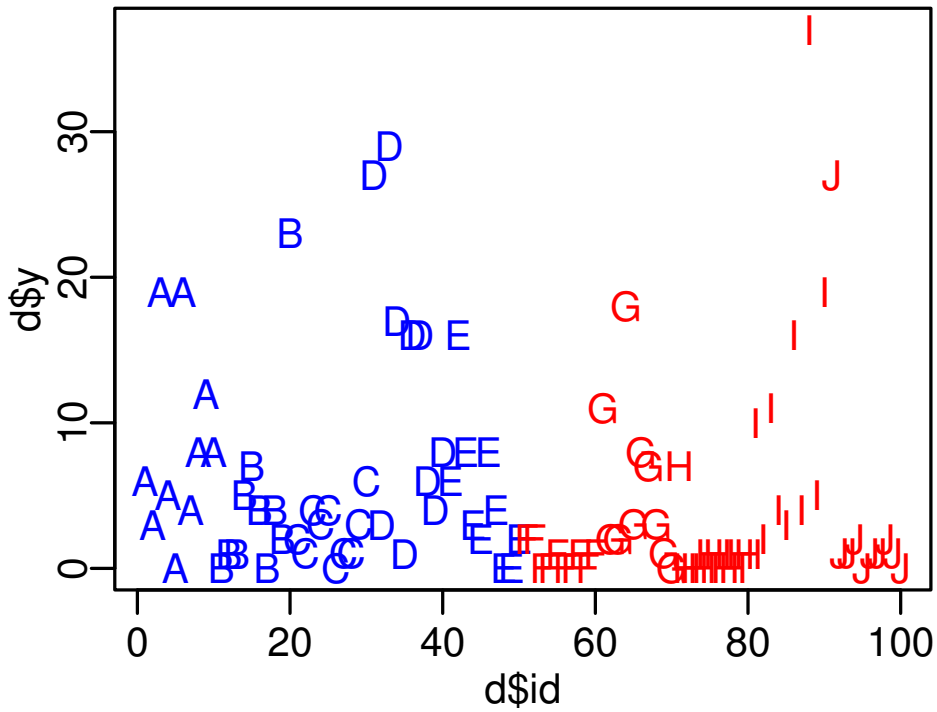


- 教育法 F によって児童の計算速度 y_i が速くなるかを調べたい
- 小学校 10 校，各校に 10 人の児童に算数のテスト (合計 100 児童)
  - コントロール ($f_j = \mathrm{C}$) 5 校 (合計 50 児童)
  - 教育法 F 実施 ($f_j = \mathrm{T}$) 5 校 (合計 50 児童)

この架空データを生成したときには，処理の効果はゼロと設定

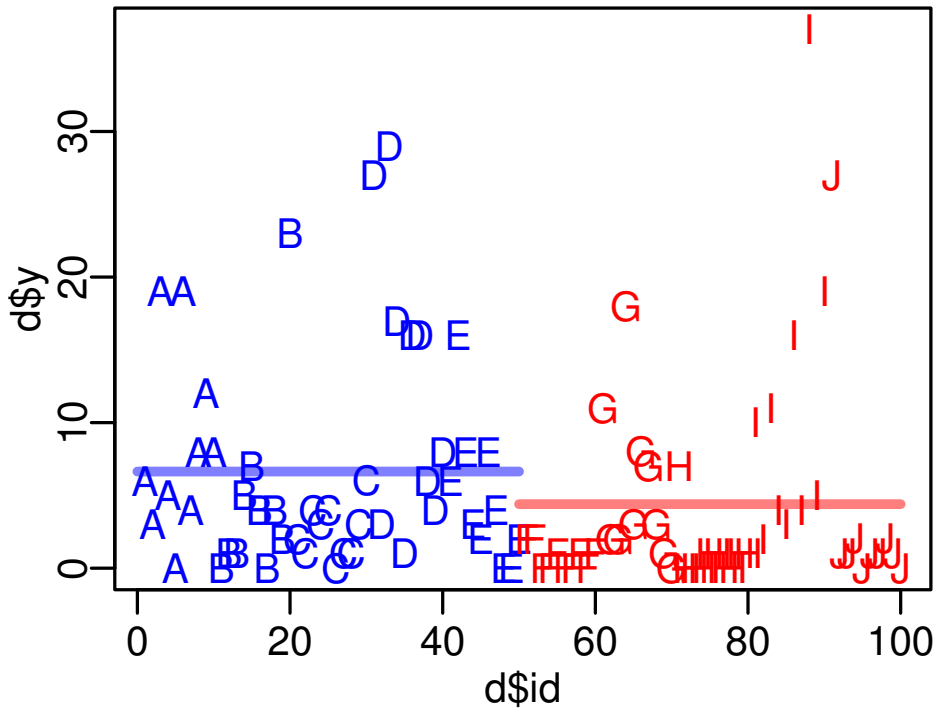
「教育法 F の効果はゼロ」をただしく推定できるか……?

とりあえず，データはわかりやすく図示



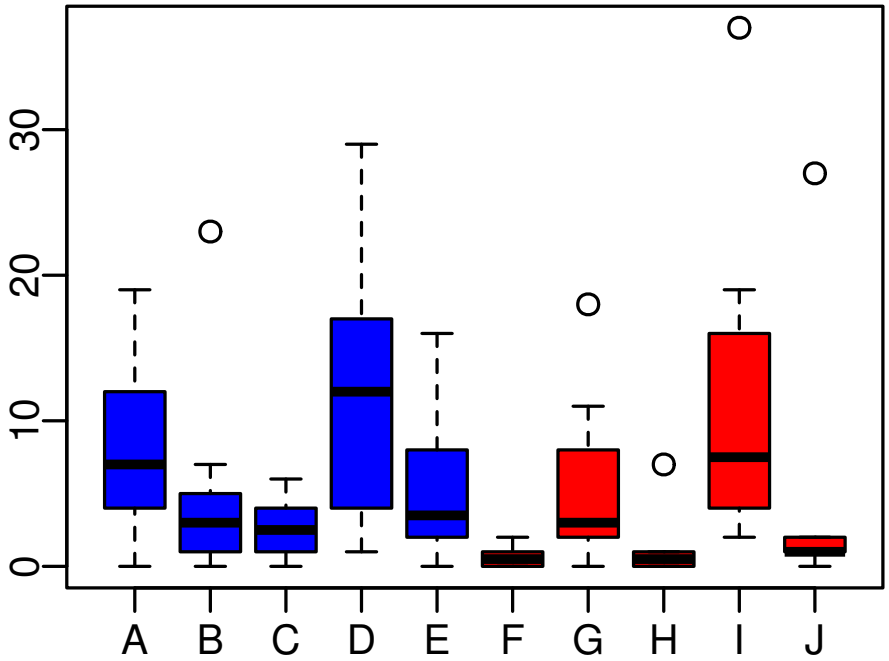
- `plot(d$id, d$y, pch = as.character(d$pot), ...)`
- **コントロール・処理 でそんなに差がない**?

処理ごとの平均も図に追加してみる



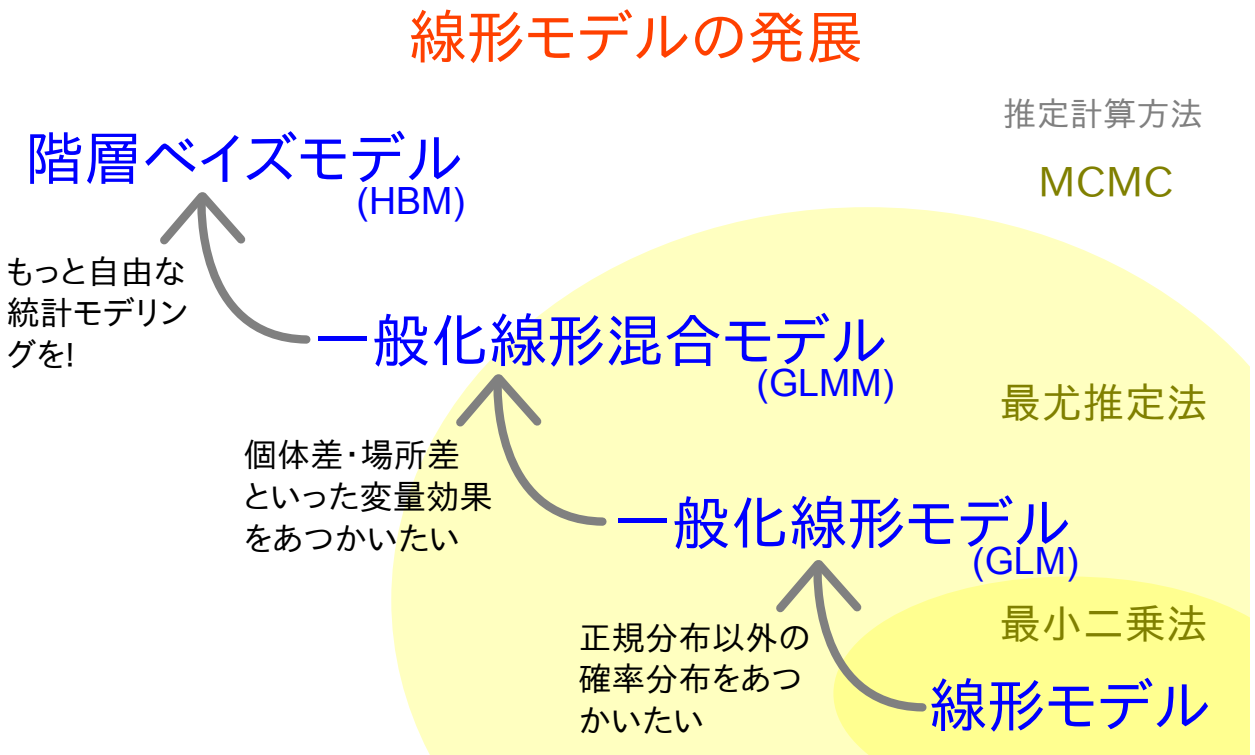
- むしろ **処理** のほうが平均回答数が低い?
- (注) この架空データは **処理の効果はゼロ** と設定して生成した

児童差だけでなく学校差もある



- `plot(d$pot, d$y, col = rep(c("blue", "red"), each = 5))`
- **学校由来の** random effects **みたいなものはブロック差と呼ばれる**

一般化線形モデルのわくぐみで，とりあえず考えてみる



GLM: 児童差も学校差も無視

```
> summary(glm(y ~ f, data = d, family = poisson))
...(略)...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.8931     0.0549   34.49  < 2e-16
fT           -0.4115     0.0869   -4.73  2.2e-06
...(略)...
```

- 教育法 F の “おかげ” で平均回答数が低下
- $p = 0.000002$ …… “確信” を持って間違っている!

GLMM: 児童差だけ考慮，学校差は無視

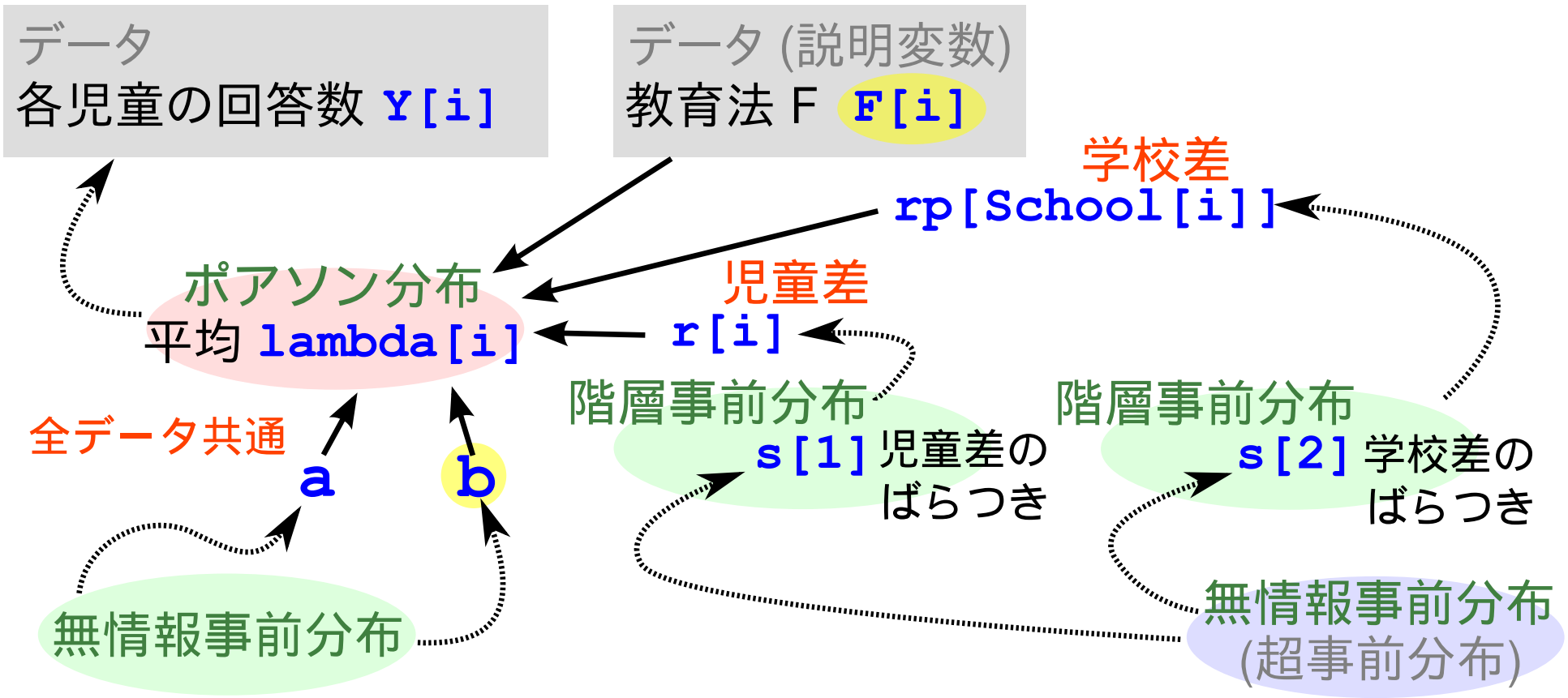
```
> library(glmmML)
> summary(glmmML(y ~ f, data = d, family = poisson,
+ cluster = id))
...(略)...
                coef se(coef)     z Pr(>|z|)
(Intercept)  1.351    0.192  7.05  1.8e-12
fT          -0.737    0.280 -2.63  8.4e-03
...(略)...
```

- GLM と同様の結果
- $p = 0.0084$ …… “確信” を持って間違っている!

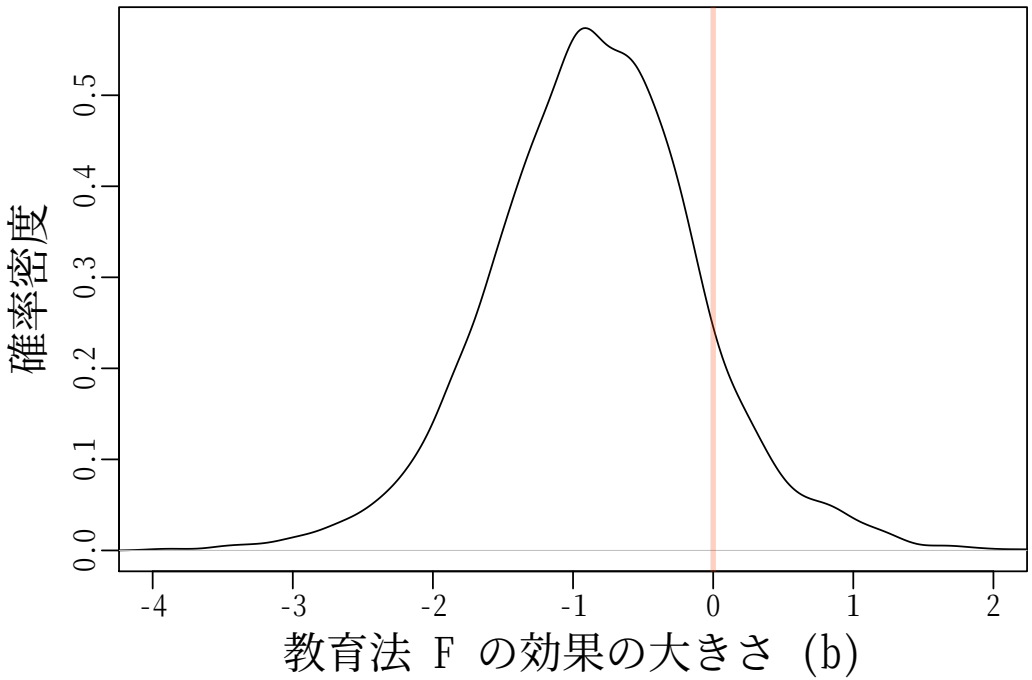
児童差 + 学校差を考える階層ベイズモデル

- ここでは log リンク関数を使う
- 平均の対数 $\log(\lambda_i) = a + bf_i +$ (児童差) + (学校差)
- 事前分布の設定
  - 切片 a と f_i の係数 b は無情報事前分布 (すごく平らな正規分布)
  - 児童差と学校差は階層的な事前分布 (それぞれ標準偏差 σ_1, σ_2 の正規分布，平均はゼロ)
  - 標準偏差 σ_* は無情報事前分布 ($[0, 10^4]$ の一様分布)

児童差 + 学校差を考慮している階層ベイズモデル



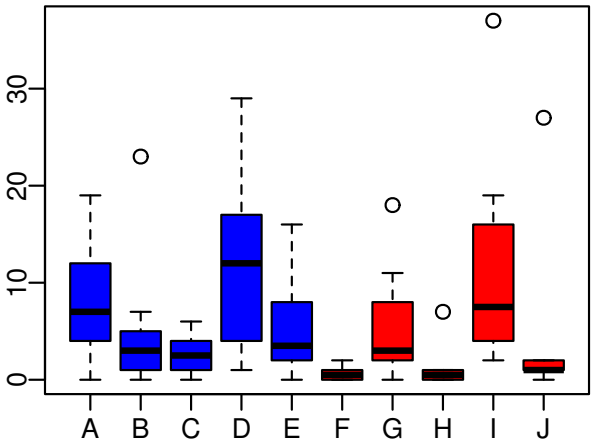
処理の効果 (パラメーター b) はなさそう?

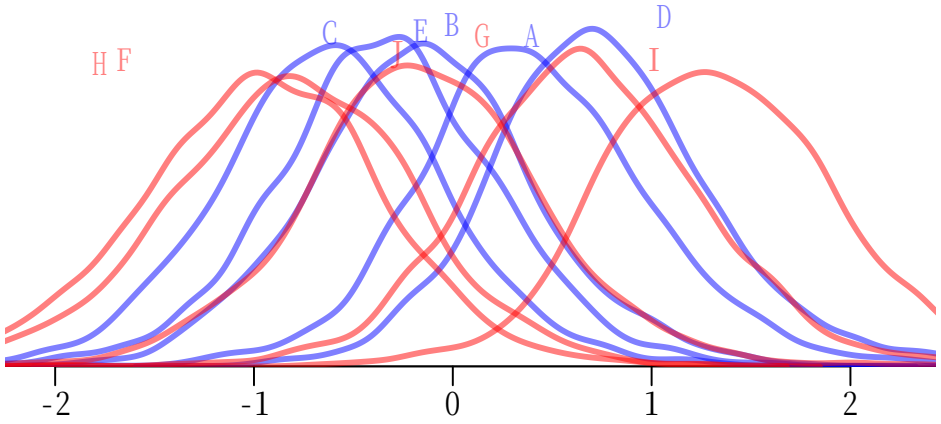


	mean	sd	2.5%	25%	50%	75%	97.5%	Rha
a	1.501	0.529	0.482	1.157	1.493	1.852	2.565	1.03
b	-1.016	0.706	-2.436	-1.476	-0.993	-0.565	0.395	1.01
sigma[1]	1.020	0.114	0.822	0.939	1.014	1.089	1.265	1.00

...(略)...

学校差の事後分布 — これを推定できるのが利点のひとつ





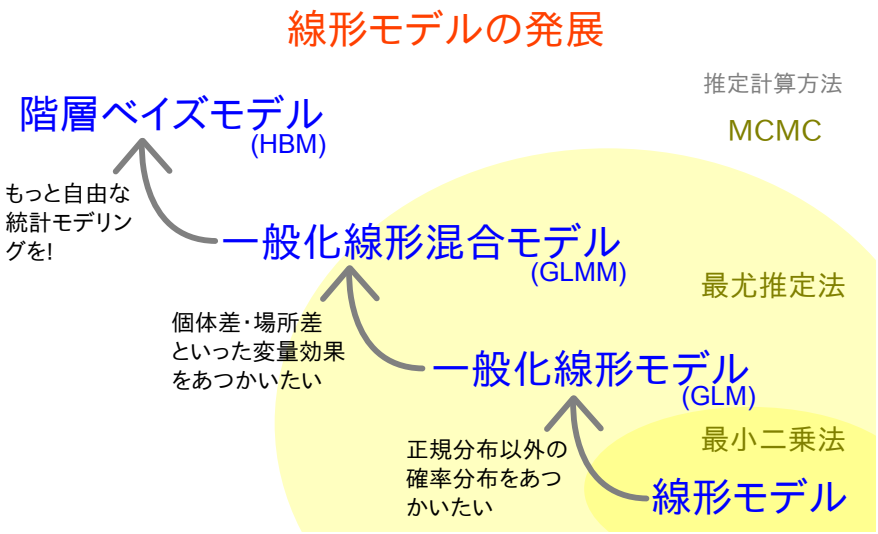
学校差

- GLMM + 最尤推定だと“学校差”などは見えなくなる
- “学校間の差”なども確率分布として得られる

統計モデリングの手ぬきは危険!

- random effects の影響が大きいときには，fixed effects の大きさが見えにくくなる
- “手ぬき” モデルである GLM・GLMM は……
  - “確信をもって” まちがえるので非常に危険!
  - 児童差・学校差の階層ベイズモデルが必要!

階層ベイズモデルと GLMM の関係は?

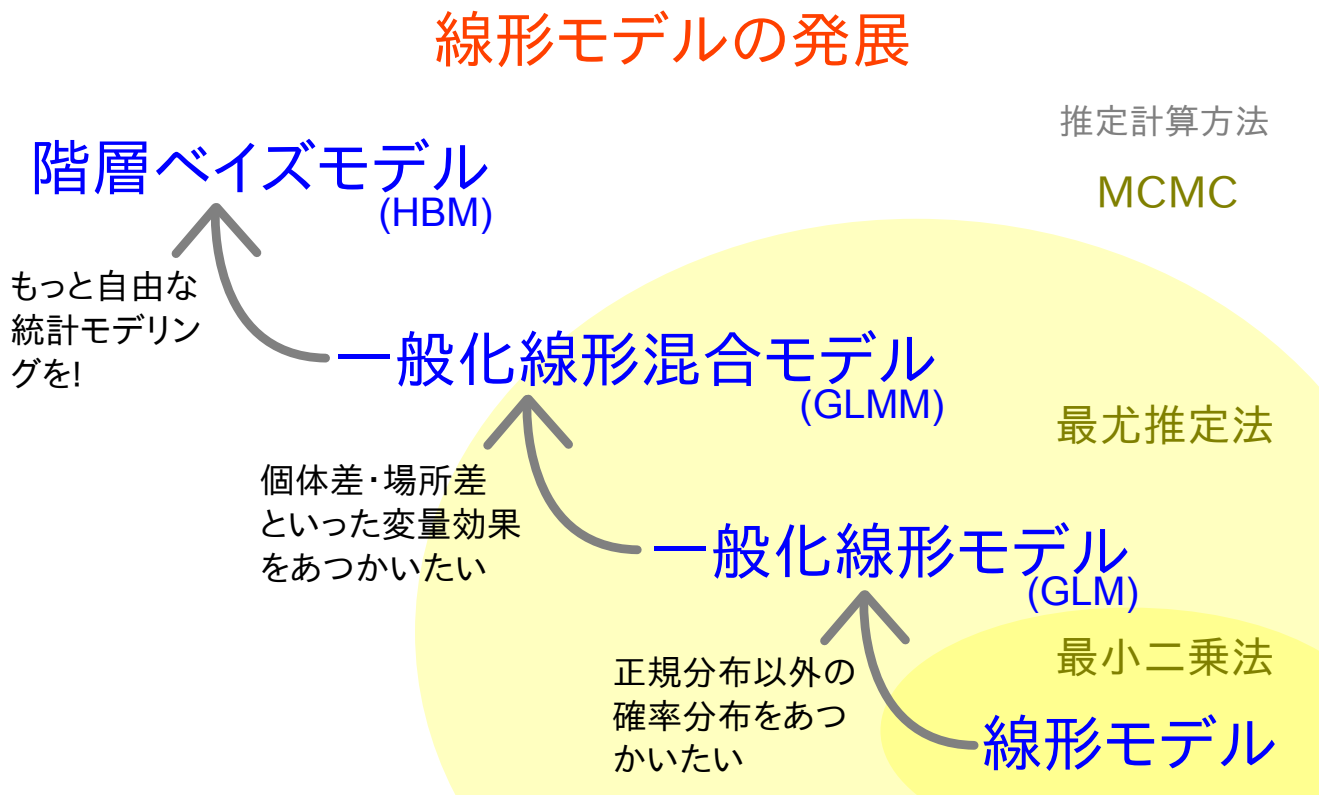


一般化線形混合モデル (Generalized Linear Mixed Model) **は階層ベイズモデルのひとつ**

- GLMM **では児童差・学校差といった** local parameter **は積分して消去してしまう**
- **階層ベイズモデルでは，何もかも事後分布として推定してしまう**

ここでこの 統計モデリング入門 は終了

- **一般化線形モデル** → **階層ベイズモデル**
- **最尤推定** → Markov chain Monte Carlo (MCMC)



“ブラックボックス” ではない 幸せなデータ解析を!!