カウントデータの統計モデリング入門

え? いまさら分割表 !? — GLM 化と階層ベイズ化

久保拓弥 kubo@ees.hokudai.ac.jp

統計的言語研究の現在 (国立国語研究所)

2015–09–04

投影資料おき場 http://goo.gl/HQbeoh

ファイル更新時刻: 2015–11–17 09:41

はじめに

簡単な自己紹介と全体のながれなど

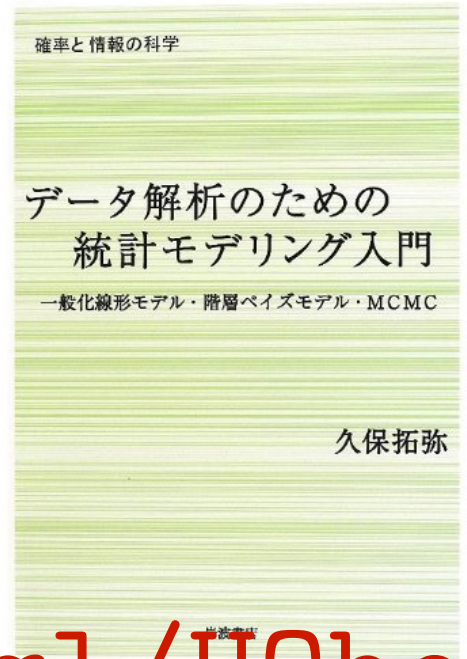
投影資料おき場 http://goo.gl/HQbeoh
(とりあえず版)

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

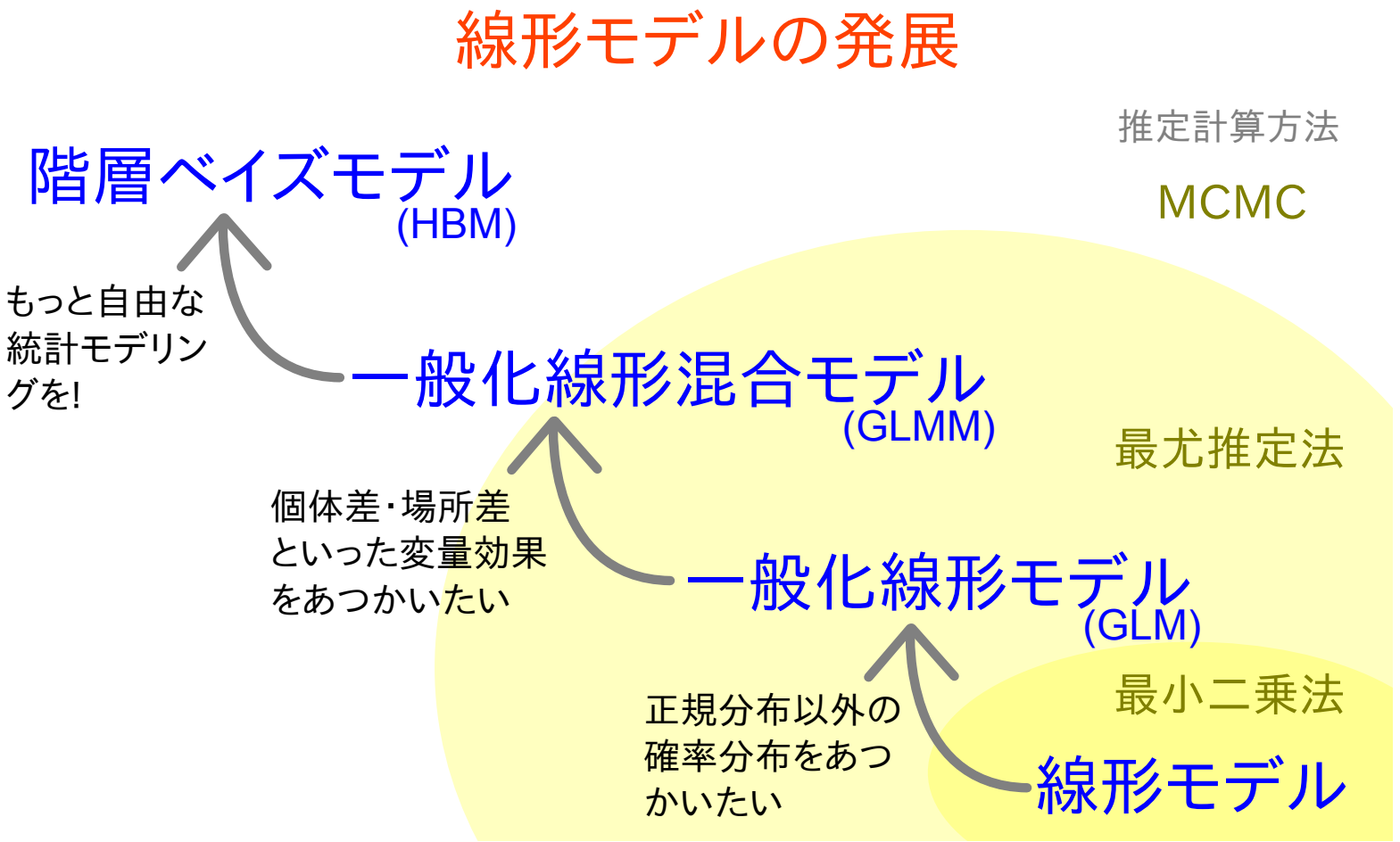
- 自分ではデータをとらない・野外調査・実験やらない
- 他のみなさんのデータ解析をすることが専門
- これではあまりにも**寄生者**的→統計モデルの教科書を書きました

統計モデリングの教科書



投影資料おき場 http://goo.gl/HQbeoh
(とりあえず版)

分割表の統計モデル: GLM から階層ベイズモデル

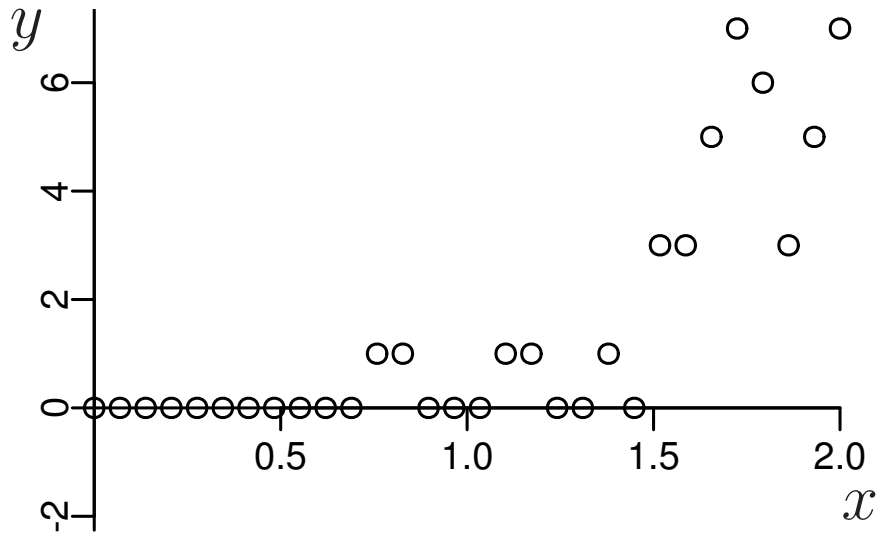


どうして これらの統計モデルを勉強するのか……?

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



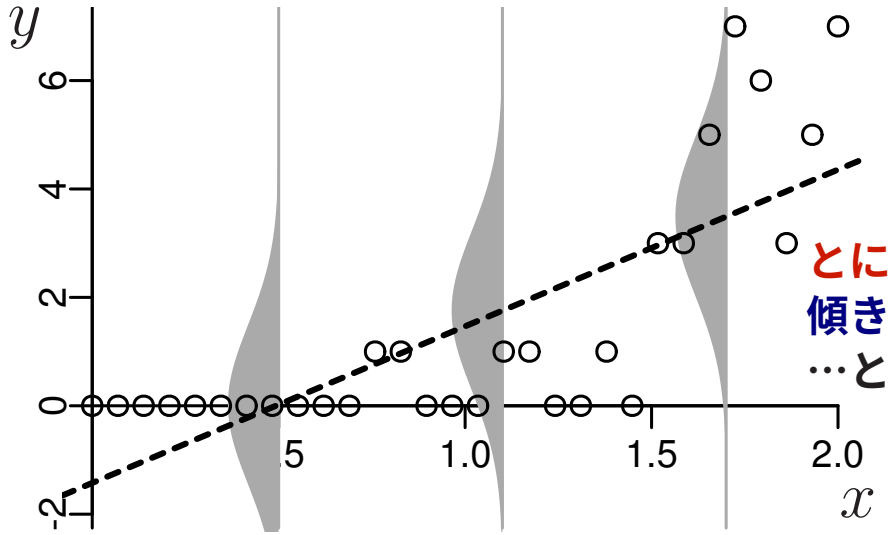
説明変数

- **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……**
- **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



ちょっと無理

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

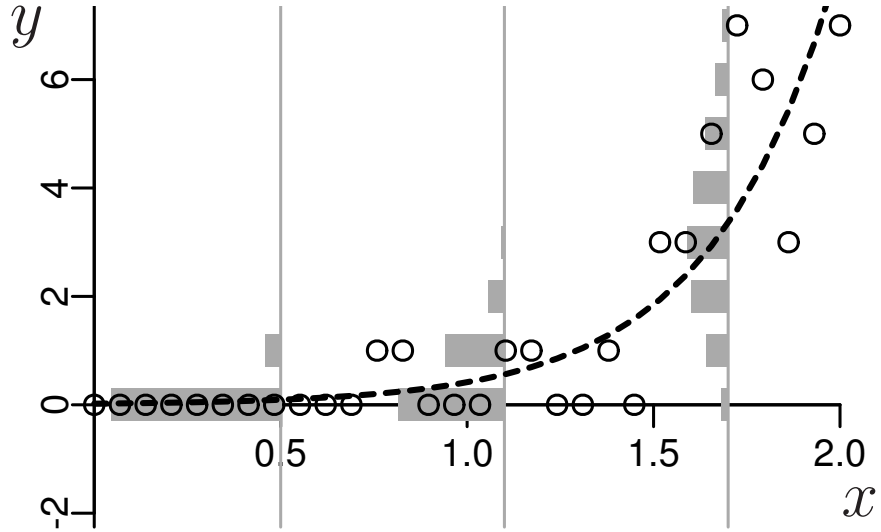
説明変数

- 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

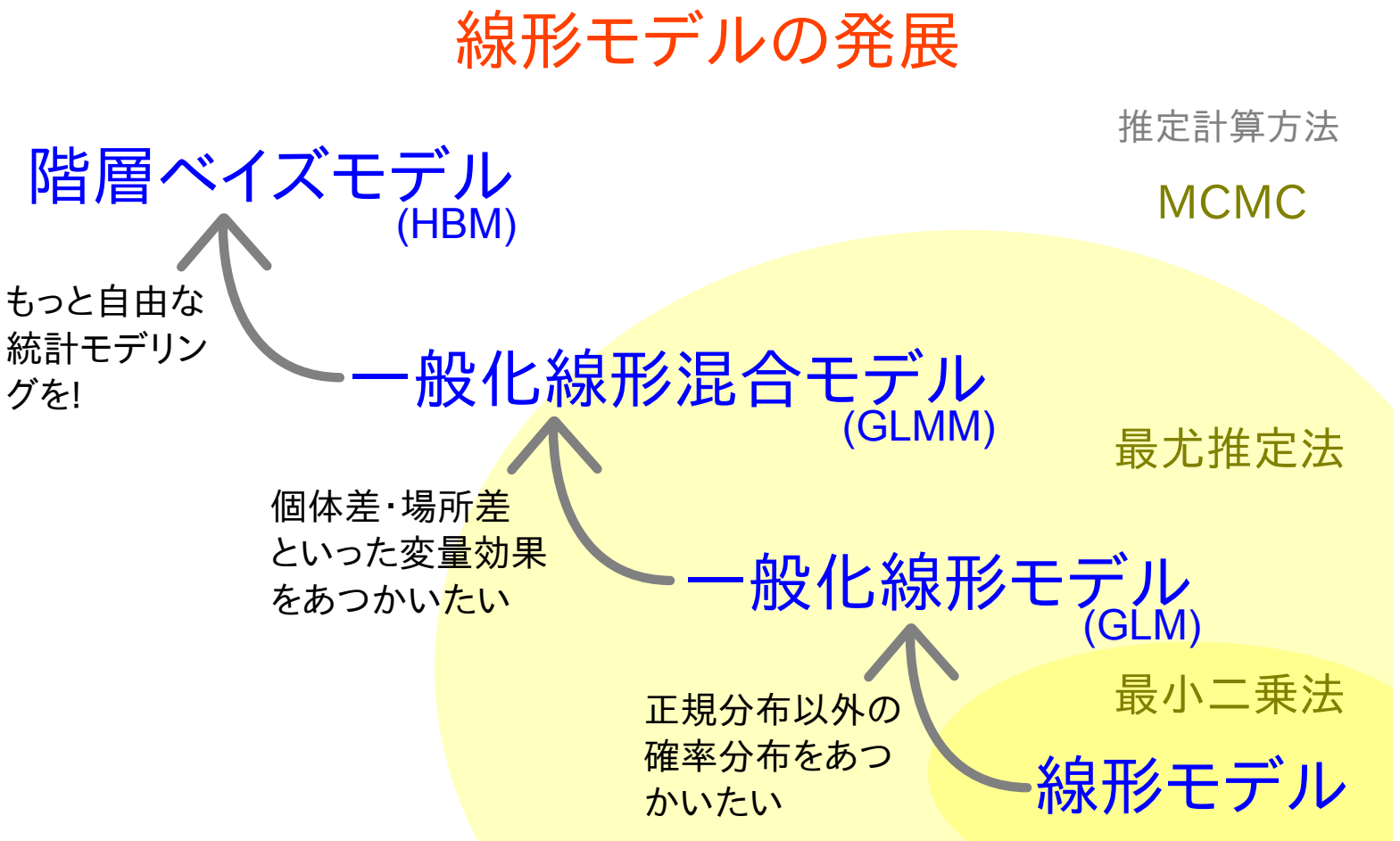


まだマシだよね

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

この授業であつかう統計モデルたち

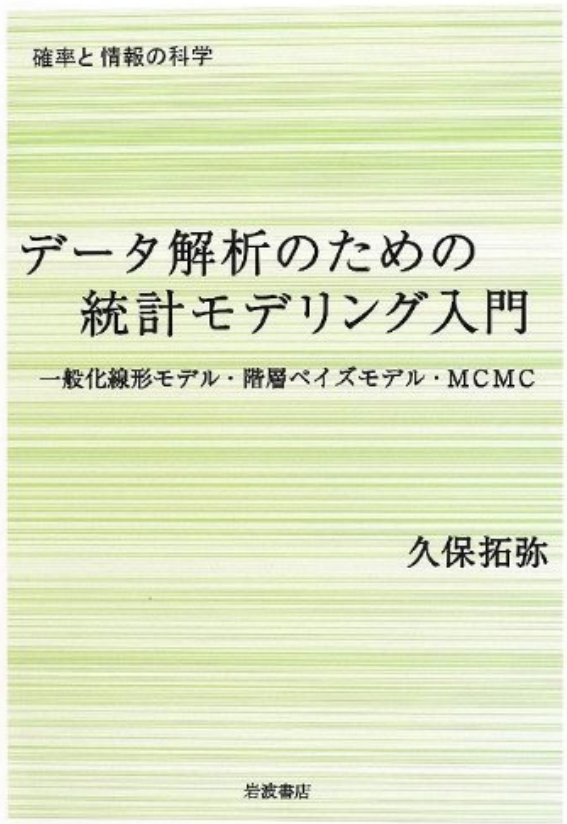


データの特徴にあわせて線形モデルを改良・発展させる

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



この時間に説明したいこと

今日登場する分割表と統計モデル

- 2×2 **分割表** — GLM (**ポアソン分布**, **二項分布**)
- 2×3 **分割表** — GLM (**ポアソン分布**, **多項分布**)
- 2×9 **分割表** — **階層ベイズ** GLM (**ポアソン分布** + **正規分布**)

ちょっと批判してみたい，よくみかけるお作法

X	Ctgr A	B	C	D	E	F	G	H	I
0	62	21	14	11	10	10	2	0	2
1	48	34	22	17	16	7	2	1	1

こういう 2×9 分割表などをみたときに何も考えずに……

- 「表の検定」 だからカイ二乗検定やればいいやー
  - 「なんでも検定」かよー
  - データ解析 $\neq$ 検定 !! 検定はデータ解析の一部 !!
- データを捨てれば検定できる– !!
  - 捨てるな !!
- なんでもかんでも多変量解析すればいいよ!
  - 古典的なやつは，いろいろ問題ありそうですな……

分割表であれ，どんなデータであれ

- まず「どんな統計モデルで説明できるか」を考える
- カウントデータの場合は，とりあえず GLM で説明できないか考えてみる
- 次の項目をきちんと区別しよう
  - **データを発生させうる統計モデル**
    (例: GLM や階層ベイズモデル)
  - **統計モデルのパラメーター推定方法**
    (例: 最尤推定法や MCMC 法)
  - **推定結果の比較方法**
    (例: Neyman-Pearson な検定，モデル選択，信用区間)

(Section 1)

2 × 2 の分割表

もっとも簡単な

今日の例題の構造 (すみません，てきとうにつくりました)

コーパスごとに異なる品詞カテゴリー出現の頻度?

Ctgr

X	A	B	C	D	E	F	G	H	I
0	62	21	14	11	10	10	2	0	2
1	48	34	22	17	16	7	2	1	1

- コーパス X0 と X1 がある
- 単語の品詞をカテゴリー化: {A, B, C, ...}
- 知りたいこと: コーパスによって，

  品詞カテゴリーの **組成** (出現割合) は変化するか?

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 で分割表をあつかう: 2×9 は難しいので，まずは 2×3

	A	B	C
1	y	x	Spc
2	286	0	A
3	85	0	B
4	378	1	A
5	148	1	B
6			

- 「CSV」として保存 (**脱** ゑくせる!)
- d2.csv というファイル名にする
- d2.csv の内容

```
y,X,Ctgr
286,0,A
85,0,B
378,1,A
148,1,B
```

データを R によみこみ，data.frame に変換

```
> d2 <- read.csv("d2.csv")
> d2  # d2 という data.frame を表示
    y X   Ctgr
1 286 0    A
2  85 0    B
4 378 1    A
5 148 1    B
```

Xtabs: 分割表をあつかう R のクラス

```
    y X   Ctgr
1 286 0    A
2  85 0    B
4 378 1    A
5 148 1    B
```

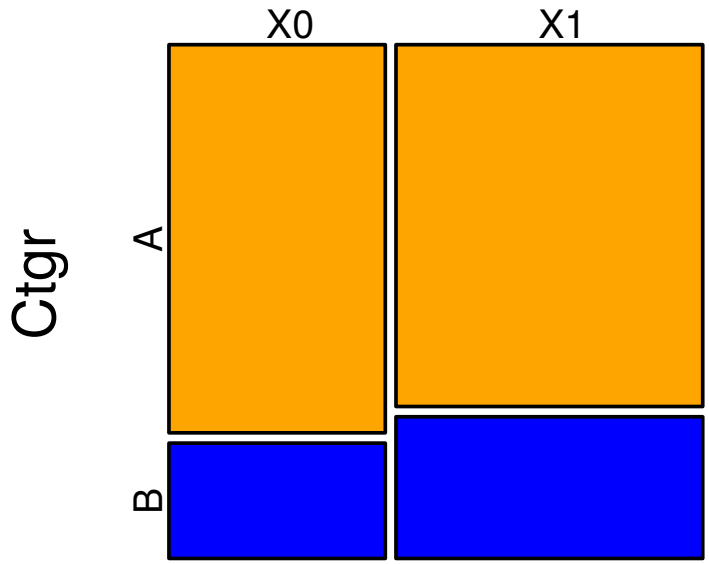
```
> (ct2 <- xtabs(y ~ X + Ctgr, data = d2))
    Ctgr
 X     A   B
  0  286  85
  1  378 148
```

xtabs: 自由自在に集計できる

```
> xtabs(y ~ X, data = d2)
X
  0   1
371 526
> xtabs(y ~ Ctgr, data = d2)
Ctgr
   A   B
 664 233
> xtabs(y ~ Ctgr + X, data = d2)
        X
Ctgr     0   1
     A 286 378
     B  85 148
```

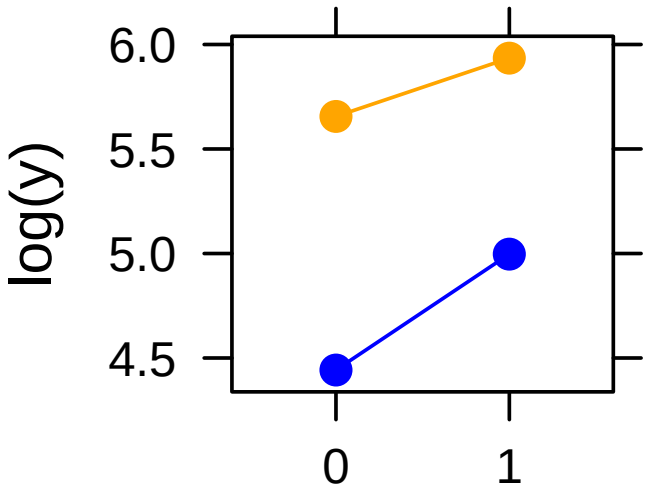
xtabs: 分割表の図示

```
    Ctgr
X      A   B
  0  286  85
  1  378 148
> plot(ct2, col = c("orange", "blue"))
```



library(lattice) を使った図示

```
    Ctgr
  X    A   B
   0 286  85
   1 378 148
> library(lattice)
> xyplot(log(y) ~ factor(X), data = d2, groups = Ctgr, type = "b")
```



(Section 2)

2 × 2 分割表の統計モデル

まずは二項分布の GLM から

ロジスティック回帰 logistic regression

一般化線形モデルを作る

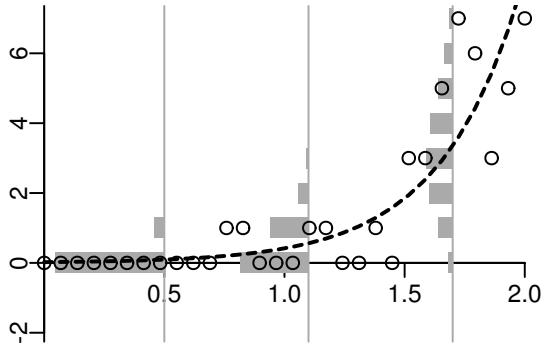
一般化線形モデル (GLM)

- 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

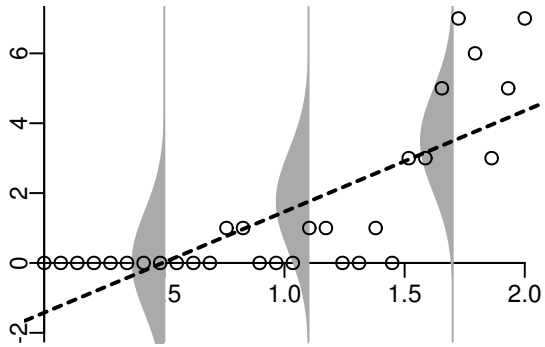
- 確率分布: ポアソン分布
- 線形予測子: $\mathrm{e.g.,}\ \beta_1 + \beta_2 x_i$
- リンク関数: 対数リンク関数
- 対数線形モデル とよばれることもある



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

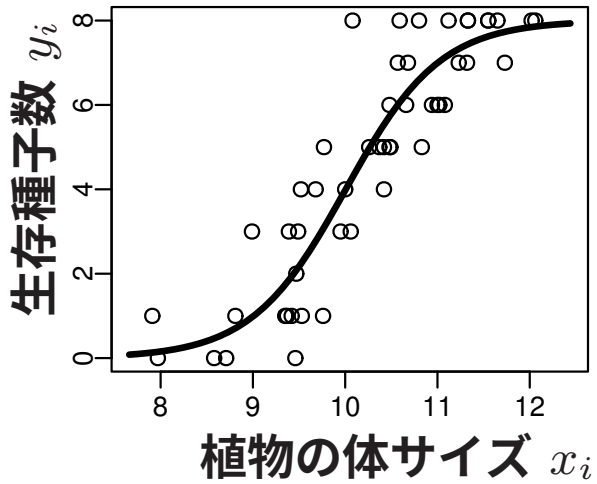
- 確率分布: 正規分布
- 線形予測子: $\mathrm{e.g.,}\ \beta_1 + \beta_2 x_i$
- リンク関数: 恒等リンク関数



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: $\mathrm{e.g.,}\ \beta_1 + \beta_2 x_i$
- リンク関数: logit リンク関数
- 割り算をしないで 割合を調べる統計モデル

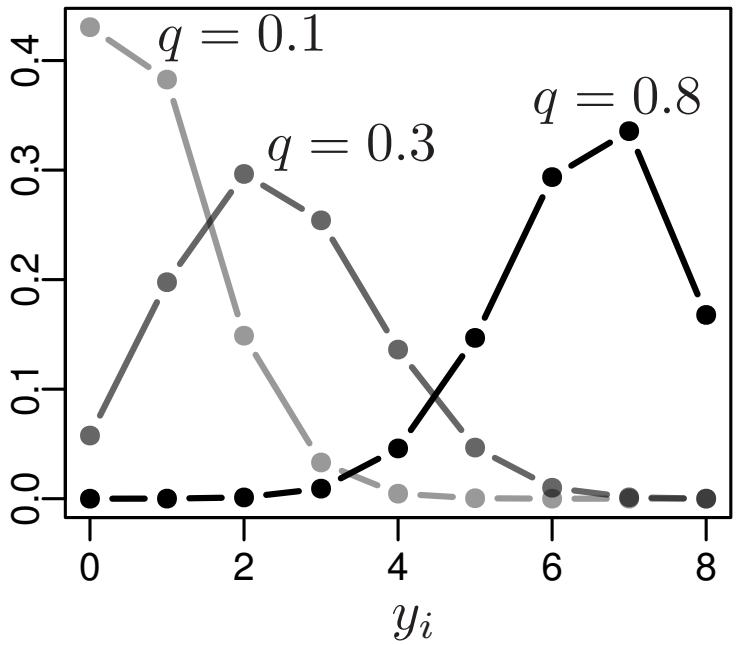


二項分布: N 回のうち y 回，となる確率

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

$\binom{N}{y}$ 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 $p(y_i \mid 8, q)$

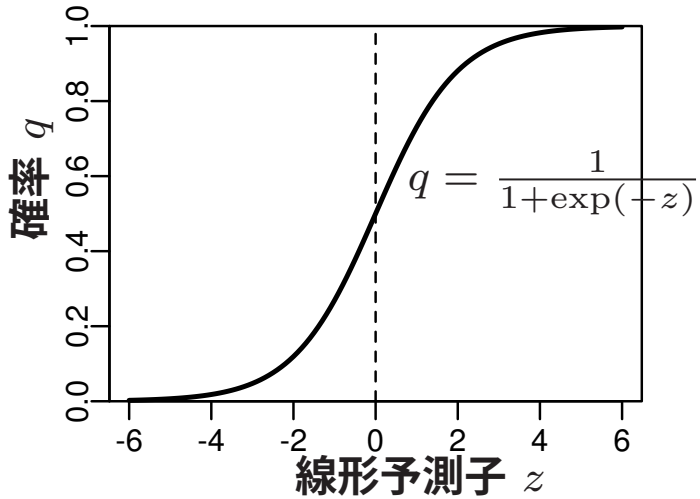


ロジスティック曲線とはこういうもの

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")
```



logit link function

- logistic **関数**

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

- logit **変換**

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

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

logit is the inverse function of logistic function, vice versa

二項分布の GLM を適用してみる

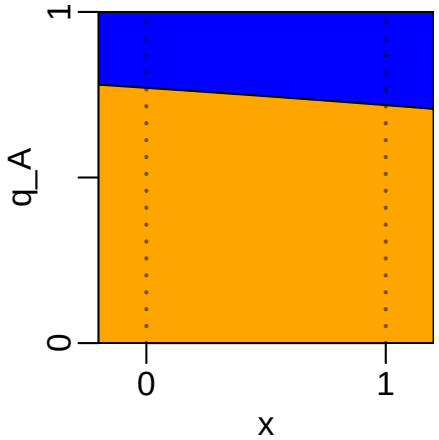
```
    Ctgr
X      A   B
  0 286  85
  1 378 148
```

$$
\begin{aligned}
y_{\mathrm{A},x} &\sim \mathrm{Binom}(q_{\mathrm{A},x},\ y_{\mathrm{A},x} + y_{\mathrm{B},x}) \\
\mathrm{logit}(q_{\mathrm{A},x}) &= a_{\mathrm{A}} + b_{\mathrm{A}} X
\end{aligned}
$$

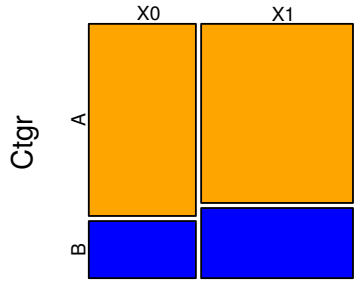
```
> summary(glm(ct2 ~ c(0, 1), data = d2, family = binomial))
(... 略...)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.213      0.124    9.82   <2e-16
c(0, 1)       -0.276      0.157   -1.76    0.079
```

ロジスティック回帰の推定にもとづく予測

$$\mathrm{logit}(q_{\mathrm{A},x}) = 1.213 + (-0.276)X$$



カウントデータ



コーパス X0 と X1 間で…	モデル $(X \in \{0,1\})$	AIC
差がある	$\mathrm{logit}(q_{\mathrm{A},x}) = a_{\mathrm{A}} + b_{\mathrm{B}}X$	16.5
差がない	$\mathrm{logit}(q_{\mathrm{A},x}) = a_{\mathrm{A}}$	17.6

(Section 3)

2 × 2 分割表の統計モデル

次にポアソン分布の GLM であつかってみる

「分割方式」と「一括方式」

ポアソン分布の GLM (分割方式) — Ctgr A だけモデル

X	Ctgr A	Ctgr B
0	286	85
1	378	148

$$
\begin{aligned}
y_{\mathrm{A},x} &\sim \mathrm{Pois}(\lambda_{\mathrm{A},x}) \\
\log(\lambda_{\mathrm{A},x}) &= \alpha_{\mathrm{A}} + \beta_{\mathrm{A}} X
\end{aligned}
$$

```
> # CtgrA だけ
> summary(glm(y ~ X, data = d2[d2$Ctgr == "A",], family = poisson))
(... 略...)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    5.6560     0.0591   95.65  < 2e-16
X              0.2789     0.0784    3.56  0.00037
```

ポアソン分布の GLM (分割方式) — Ctgr B だけモデル

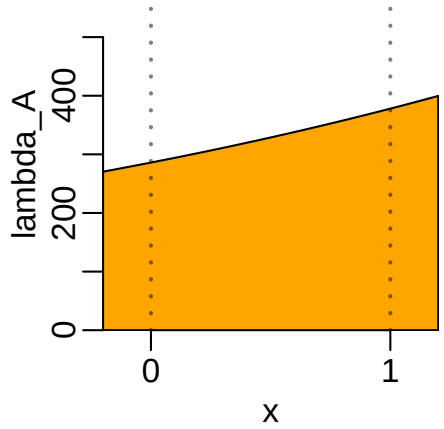
X	Ctgr A	Ctgr B
0	286	85
1	378	148

$$
\begin{aligned}
y_{\mathrm{B},x} &\sim \mathrm{Pois}(\lambda_{\mathrm{B},x}) \\
\log(\lambda_{\mathrm{B},x}) &= \alpha_{\mathrm{B}} + \beta_{\mathrm{B}} X
\end{aligned}
$$

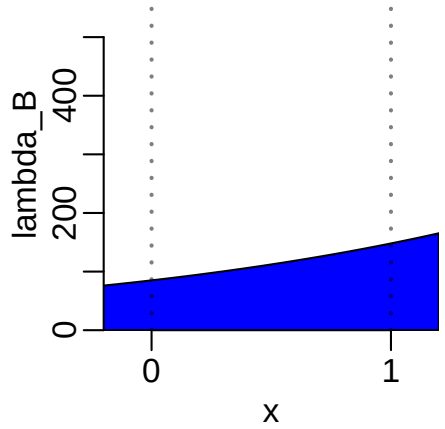
```
> # CtgrB だけ
> summary(glm(y ~ X, data = d2[d2$Ctgr == "B",], family = poisson))
(... 略...)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.443      0.108   40.96  < 2e-16
X               0.555      0.136    4.07  4.6e-05
```

ポアソン回帰の推定にもとづく予測

$\log(\lambda_{\mathrm{A},x}) = 5.66 + 0.279X$



$\log(\lambda_{\mathrm{B},x}) = 4.44 + 0.555X$



X0, X1 **間で…**	**モデル**	AIC	**モデル**	AIC
差がある	$\lambda_{\mathrm{A},x} = \alpha_{\mathrm{A}} + \beta_{\mathrm{A}}X$	19.3	$\lambda_{\mathrm{B},x} = \alpha_{\mathrm{B}} + \beta_{\mathrm{B}}X$	17.1
差がない	$\lambda_{\mathrm{A},x} = \alpha_{\mathrm{A}}$	30.1	$\lambda_{\mathrm{B},x} = \alpha_{\mathrm{B}}$	32.4

じつは

ロジスティック回帰とポアソン回帰

同じ結果になります

つまりどちらを使ってもよい

ポアソン分布 GLM・二項分布 GLM のつながり

- **二項分布** GLM: $\mathrm{logit}(q_{\mathrm{A},x}) = a_{\mathrm{A}} + b_{\mathrm{A}} X$

$$q_{\mathrm{A},x} = \frac{1}{1 + \exp[-(a_{\mathrm{A}} + b_{\mathrm{A}} X)]}$$

- **ポアソン分布**: $\log(\lambda_{\mathrm{A},x}) = \alpha_{\mathrm{A}} + \beta_{\mathrm{A}} X$ **など**

$$\lambda_{\mathrm{A},x} = \exp(\alpha_{\mathrm{A}} + \beta_{\mathrm{A}} X)$$
$$\lambda_{\mathrm{B},x} = \exp(\alpha_{\mathrm{B}} + \beta_{\mathrm{B}} X)$$

……「Ctgr A の割合」は?

$$\frac{\lambda_{\mathrm{A},x}}{\lambda_{\mathrm{A},x} + \lambda_{\mathrm{B},x}} = \frac{\exp(\alpha_{\mathrm{A}} + \beta_{\mathrm{A}} X)}{\exp(\alpha_{\mathrm{A}} + \beta_{\mathrm{A}} X) + \exp(\alpha_{\mathrm{B}} + \beta_{\mathrm{B}} X)} = \frac{1}{1 + \exp[\alpha_{\mathrm{B}} - \alpha_{\mathrm{A}} + (\beta_{\mathrm{B}} - \beta_{\mathrm{A}}) X]}$$

係数の比較: ポアソン分布 GLM・二項分布 GLM のつながり

二項分布の GLM

$$q_{\mathrm{A},x} = \frac{1}{1 + \exp[-(a_{\mathrm{A}} + b_{\mathrm{A}} X)]}$$

ポアソン分布の GLM (**分割方式**)

$$\frac{\lambda_{\mathrm{A},x}}{\lambda_{\mathrm{A},x} + \lambda_{\mathrm{B},x}} = \frac{1}{1 + \exp[\alpha_{\mathrm{B}} - \alpha_{\mathrm{A}} + (\beta_{\mathrm{B}} - \beta_{\mathrm{A}})X]}$$

比較すると……

二項分布 GLM		**ポアソン分布** GLM
a_{A}	$=$	$\alpha_{\mathrm{A}} - \alpha_{\mathrm{B}}$
b_{A}	$=$	$\beta_{\mathrm{A}} - \beta_{\mathrm{B}}$

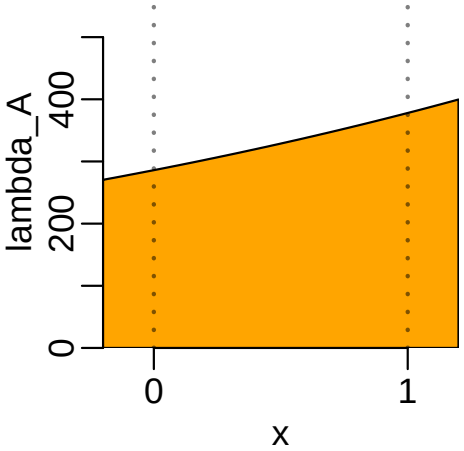
比較: 二項分布とポアソン分布の GLM

二項分布 GLM		ポアソン分布 GLM
$a_{\mathrm{A}} = 1.213$	=	$\alpha_{\mathrm{A}} - \alpha_{\mathrm{B}}$
$b_{\mathrm{A}} = \text{-}0.276$	=	$\beta_{\mathrm{A}} - \beta_{\mathrm{B}}$

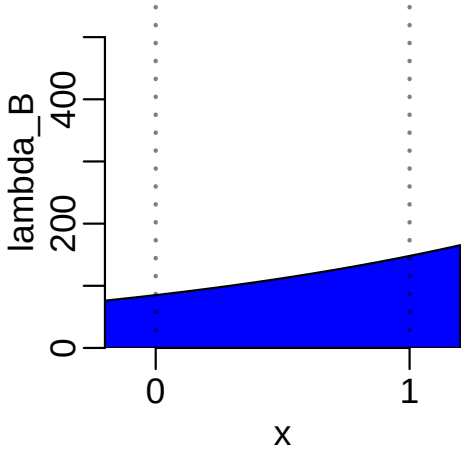
```
> 二項分布 GLM (A 種の比率)
> glm(ct2 ~ c(0, 1), data = d2, family = binomial)
(Intercept)       c(0, 1)
      1.213        -0.276
> ポアソン分布 GLM (A 種の比率)
> glm(y ~ X, data = d2[d2$Ctgr == "A",], family = poisson)
(Intercept)             X
      5.656         0.279
> ポアソン分布 GLM (B 種の比率)
> glm(y ~ X, data = d2[d2$Ctgr == "B",], family = poisson)
(Intercept)             X
```

図解: ポアソン分布 GLM・二項分布 GLM のつながり

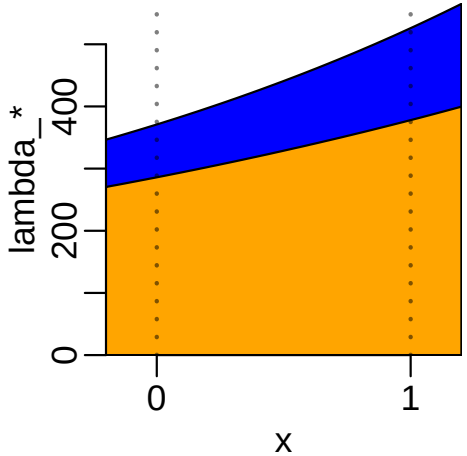
ポアソン分布の GLM (Ctgr A)



ポアソン分布の GLM (Ctgr B)

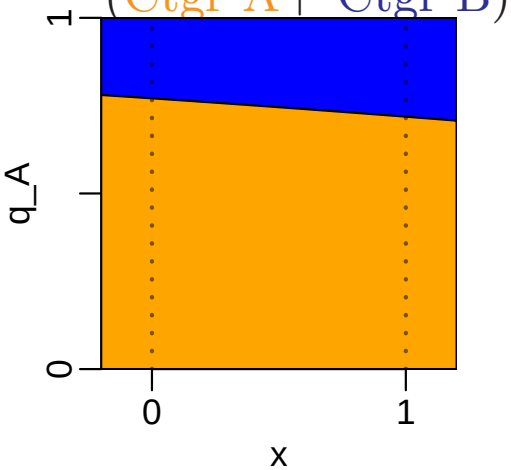


つみあげる



たいらに
押しつぶす

二項分布の GLM
(Ctgr A+ Ctgr B)



2 × 2 分割表の統計モデル

データを分割しないポアソン分布 GLM

「一括方式」(仮称)

ポアソン分布の GLM (一括方式)

交互作用項をうまく利用する

```
> summary(glm(y ~ X * Ctgr, data = d2, family = poisson))
(... 略...)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.6560     0.0591   95.65  < 2e-16
X             0.2789     0.0784    3.56  0.00037
CtgrB        -1.2133     0.1235   -9.82  < 2e-16
X:CtgrB       0.2757     0.1570    1.76  0.07921
(... 略...)
```

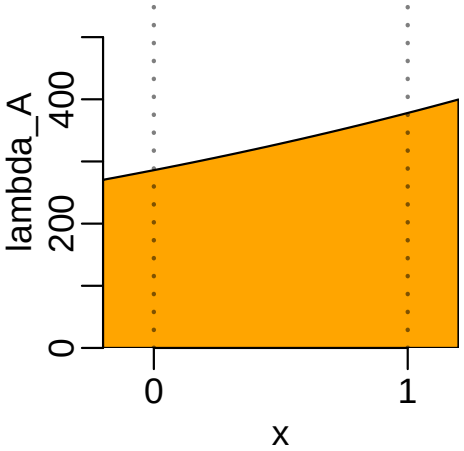
「分割方式」のポアソン分布 GLM と一致 → 二項分布 GLM とも一致

$$\alpha_{\mathrm{A}} = 5.66 \qquad \alpha_{\mathrm{B}} = 5.66 - 1.21$$

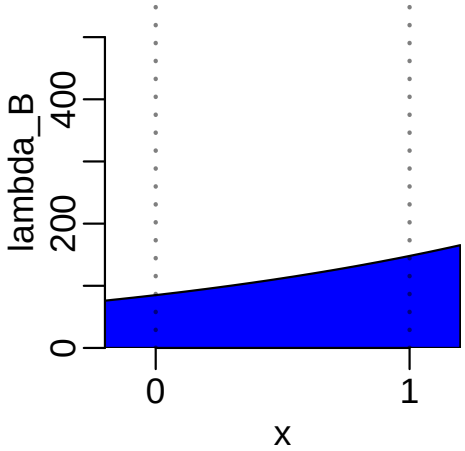
$$\beta_{\mathrm{A}} = 0.279 \qquad \beta_{\mathrm{B}} = 0.279 + 0.276$$

ポアソン・二項分布両 GLM のつながり (再)

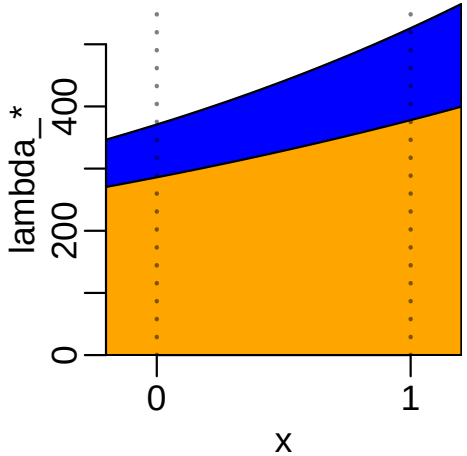
ポアソン分布の GLM (Ctgr A)



ポアソン分布の GLM (Ctgr B)

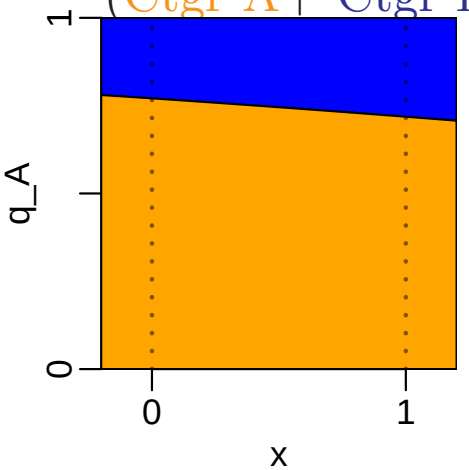


つみあげる



たいらに
押しつぶす

二項分布の GLM
(Ctgr A+ Ctgr B)



ちょっと計算メモ: 条件付きポアソン分布は二項分布になる

平均 λ_x, λ_y である二つのポアソン分布にしたがう確率変数がそれぞれ x, y とする．$x = k$, $y = n - k$ となる同時確率 $p(x = k, y = n - k)$ は，つぎのふたつの書きかたがある．

$$\begin{aligned} p(x = k, y = n - k) &= p(x = k)\ p(y = n - k) \\ &= p(x = k \| x + y = n)\ p(x + y = n) \end{aligned}$$

$$\begin{aligned} p(x = k \| x + y = n) &= \frac{p(x = k)\ p(y = n - k)}{p(x + y = n)} \\ &= \frac{\frac{\lambda_x^k \exp(-\lambda_x)}{k!}\ \frac{\lambda_y^{n-k} \exp(-\lambda_y)}{(n-k)!}}{\sum_{i=0}^{n} \frac{\lambda_x^i \exp(-\lambda_x)}{i!}\ \frac{\lambda_y^{n-i} \exp(-\lambda_y)}{(n-i)!}} = \frac{\lambda_x^k\ \lambda_y^{n-k}}{k!(n-k)!} \Big/ \sum_{i=0}^{n} \frac{\lambda_x^i\ \lambda_y^{n-i}}{i!(n-i)!} \end{aligned}$$

ところで二項定理

$$(\lambda_x + \lambda_y)^n = \sum_{i=0}^{n} \frac{n!}{i!(n-i)!} \lambda_x^i\ \lambda_y^{n-i} \qquad \text{および} \qquad \sum_{i=0}^{n} \frac{\lambda_x^i\ \lambda_y^{n-i}}{i!(n-i)!} = \frac{(\lambda_x + \lambda_y)^n}{n!} \tag{1}$$

これらを使って上の条件付き確率 $p(x = k \| x + y = n)$ を次のように書きなおすことができる．

$$p(x = k \| x + y = n) = \frac{\lambda_x^k\ \lambda_y^{n-k}}{k!(n-k)!} \Big/ \frac{(\lambda_x + \lambda_y)^n}{n!} = \frac{n!}{k!(n-k)!}\ \frac{\lambda_x^k}{(\lambda_x + \lambda_y)^k}\ \frac{\lambda_y^{n-k}}{(\lambda_x + \lambda_y)^{n-k}}$$

ここで $q = \dfrac{\lambda_x}{\lambda_x + \lambda_y}$ とすると，条件付き確率 $p(x = k \| x + y = n)$ は二項分布になっていることがわかる．

$$p(x = k \| x + y = n) = \frac{n!}{k!(n-k)!}\ q^k (1 - q)^{n-k}$$

(おしまい)

(Section 4)

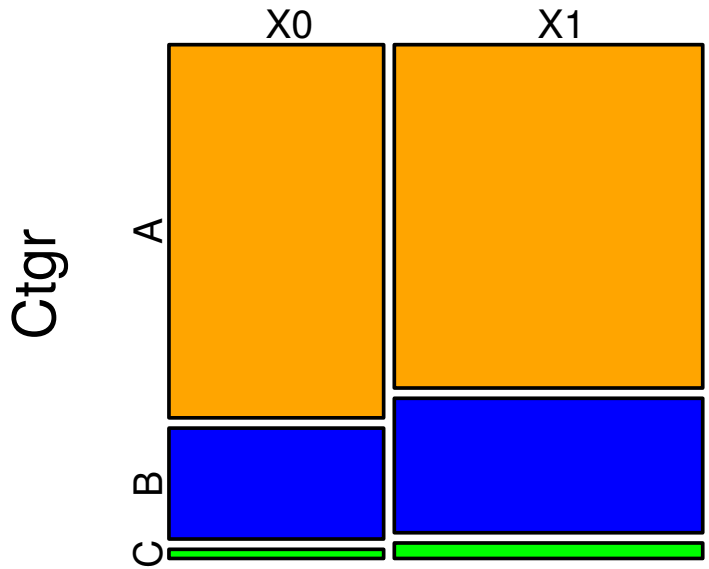
2 × 3 の分割表

多項分布の GLM か?

ポアソン分布の GLM か?

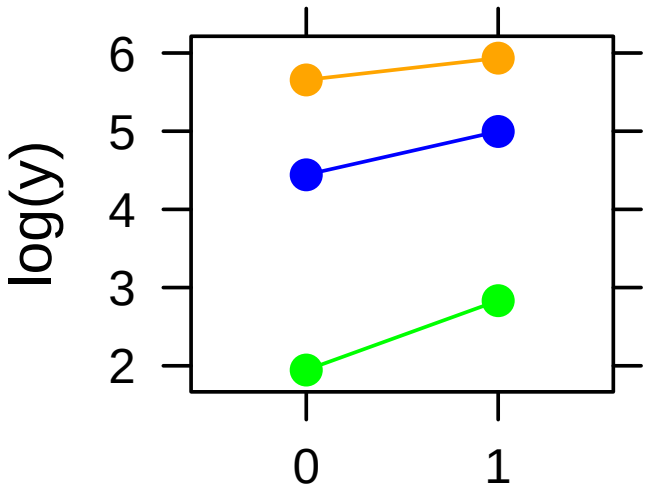
xtabs: 2 × 3 分割表の図示

```
    Ctgr
X      A   B   C
  0  286  85   7
  1  378 148  17
> plot(ct3, col = c("orange", "blue", "green"))
```



library(lattice) を使った図示

```
    Ctgr
X       A   B   C
  0   286  85   7
  1   378 148  17
> library(lattice)
> xyplot(log(y) ~ factor(X), data = d3, groups = Ctgr, type = "b")
```



ポアソン分布の GLM (一括方式)

```
> glm(y ~ X * Ctgr, data = d3, family = poisson)
(... 略...)
Coefficients:
(Intercept)           X      CtgrB      CtgrC    X:CtgrB    X:CtgrC
      5.656      0.279     -1.213     -3.710      0.276      0.608
```

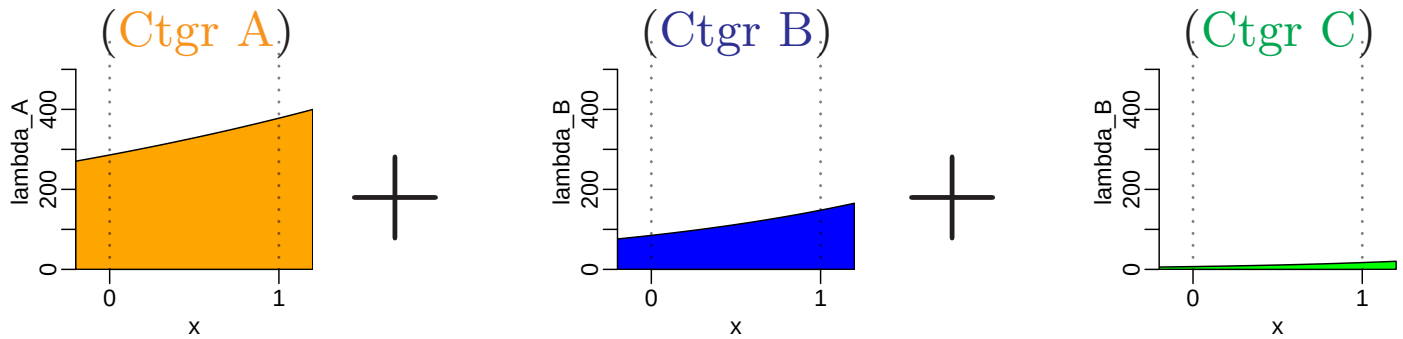
「分割方式」のポアソン分布 GLM のパラメーターで言うと……

$$
\begin{aligned}
y_{\mathrm{A},x} &\sim \mathrm{Pois}(\lambda_{\mathrm{A},x}) \\
\log(\lambda_{\mathrm{A},x}) &= \alpha_{\mathrm{A}} + \beta_{\mathrm{A}} X
\end{aligned}
$$

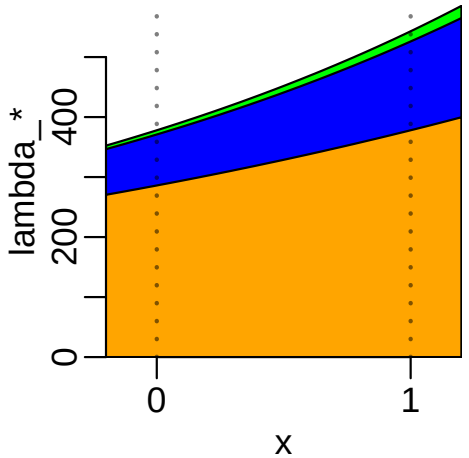
$$
\begin{array}{lll}
\alpha_{\mathrm{A}} = 5.66 & \alpha_{\mathrm{B}} = 5.66 - 1.21 & \alpha_{\mathrm{C}} = 5.66 - 3.71 \\
\beta_{\mathrm{A}} = 0.279 & \beta_{\mathrm{B}} = 0.279 + 0.276 & \beta_{\mathrm{C}} = 0.279 + 0.608
\end{array}
$$

ポアソン分布・三項分布 GLM のつながり

ポアソン分布の GLM

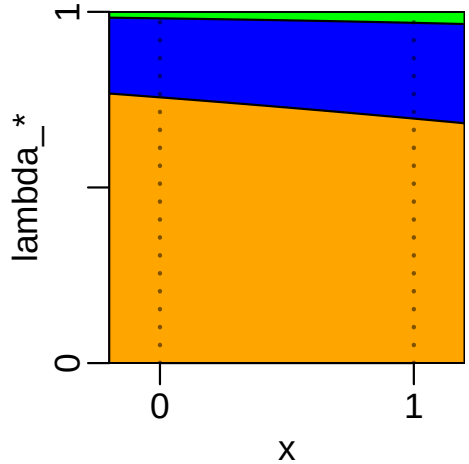


つみあげる



たいらに
押しつぶす

三項分布の GLM



多項分布・ロジスティックな GLM

```
> ct3 # 分割表を表示
   Ctgr
X     A   B   C
  0 286  85   7
  1 378 148  17
> library(nnet) # nnet package よみこみ
> multinom(ct3 ~ c(0, 1))
(... 略...)
Coefficients:
  (Intercept) c(0, 1)
B     -1.2133 0.27552
C     -3.7097 0.60763
```

多項分布・ロジスティック GLM

$$y_{\mathrm{B},x} \sim \mathrm{Multinom}(q_{\mathrm{B},x},\ 3\ \text{種合計数})$$

$$y_{\mathrm{C},x} \sim \mathrm{Multinom}(q_{\mathrm{C},x},\ 3\ \text{種合計数})$$

$$\log\frac{q_{\mathrm{B},x}}{q_{\mathrm{A},x}} = \beta_{\mathrm{B},1} + \beta_{\mathrm{B},2} x$$

$$\log\frac{q_{\mathrm{C},x}}{q_{\mathrm{A},x}} = \beta_{\mathrm{C},1} + \beta_{\mathrm{C},2} x$$

> # ポアソン分布 GLM と同じ推定値!

(Section 5)

2×9 の分割表

単純な GLM では無理 → 階層ベイズモデル

たくさんのパラメーターを制御しながら

また別のデータ: カテゴリ数が 9 個に増えた!

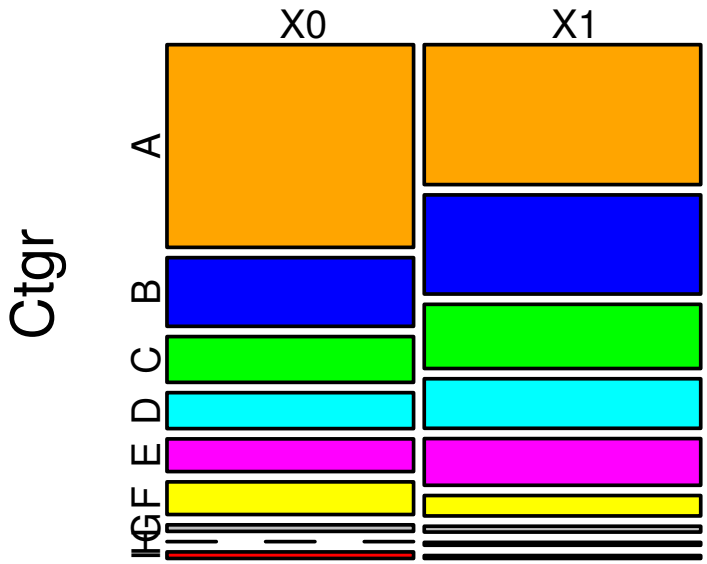
```
> d2x9
    y X Ctgr
1  62 0    A
2  21 0    B
3  14 0    C
4  11 0    D
5  10 0    E
6  10 0    F
7   2 0    G
8   0 0    H
9   2 0    I
10 48 1    A
(... 略...)
15  7 1    F
16  2 1    G
17  1 1    H
```

```
> d2x9
   Ctgr
 X    A  B  C  D  E  F  G  H  I
   0 62 21 14 11 10 10  2  0  2
   1 48 34 22 17 16  7  2  1  1
```

- 種ごとに個体数のばらつきがかなりある
- ゼロデータを含む

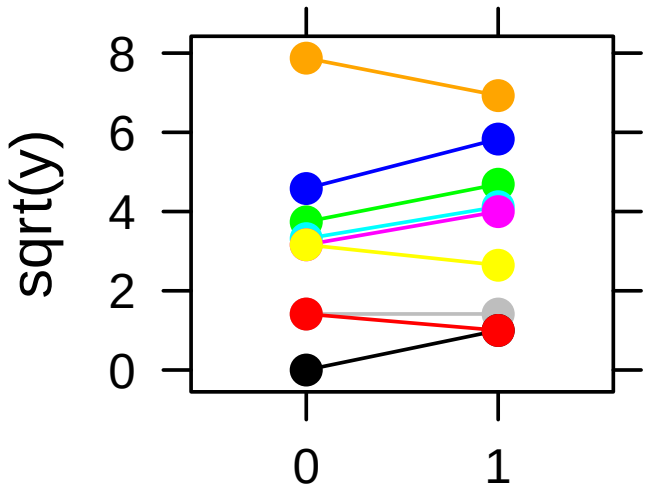
xtabs: 分割表の図示

```
   Ctgr
X     A  B  C  D  E  F  G  H  I
  0 62 21 14 11 10 10  2  0  2
  1 48 34 22 17 16  7  2  1  1
> plot(d2x9, col = c(ごちゃごちゃと指定))
```



library(lattice) を使った図示

```
      Ctgr
   X     A  B  C  D  E  F  G  H  I
     0  62 21 14 11 10 10  2  0  2
     1  48 34 22 17 16  7  2  1  1
> library(lattice)
> xyplot(sqrt(y) ~ factor(X), data = d2x9, groups = Ctgr, type = "b")
```



ポアソン分布の GLM (一括方式)

```
> d2x9
   Ctgr
 X    A  B  C  D  E  F  G  H  I
   0 62 21 14 11 10 10  2  0  2
   1 48 34 22 17 16  7  2  1  1
> summary(glm(y ~ X * Ctgr, data = d2x9, family = poisson))

            (Intercept)                   x          CtgrB          CtgrC
                  4.127              -0.256         -1.083         -1.488
                  CtgrD          CtgrE          CtgrF          CtgrG
                 -1.729         -1.825         -1.825         -3.434
                  CtgrH          CtgrI        x:CtgrB        x:CtgrC
                -26.430         -3.434          0.738          0.708
              x:CtgrD        x:CtgrE        x:CtgrF        x:CtfrG
                  0.691          0.726         -0.101          0.256
              x:CtgrH        x:CtgrI
                 22.559         -0.437
```

H 種 の推定値がかなりヘン!

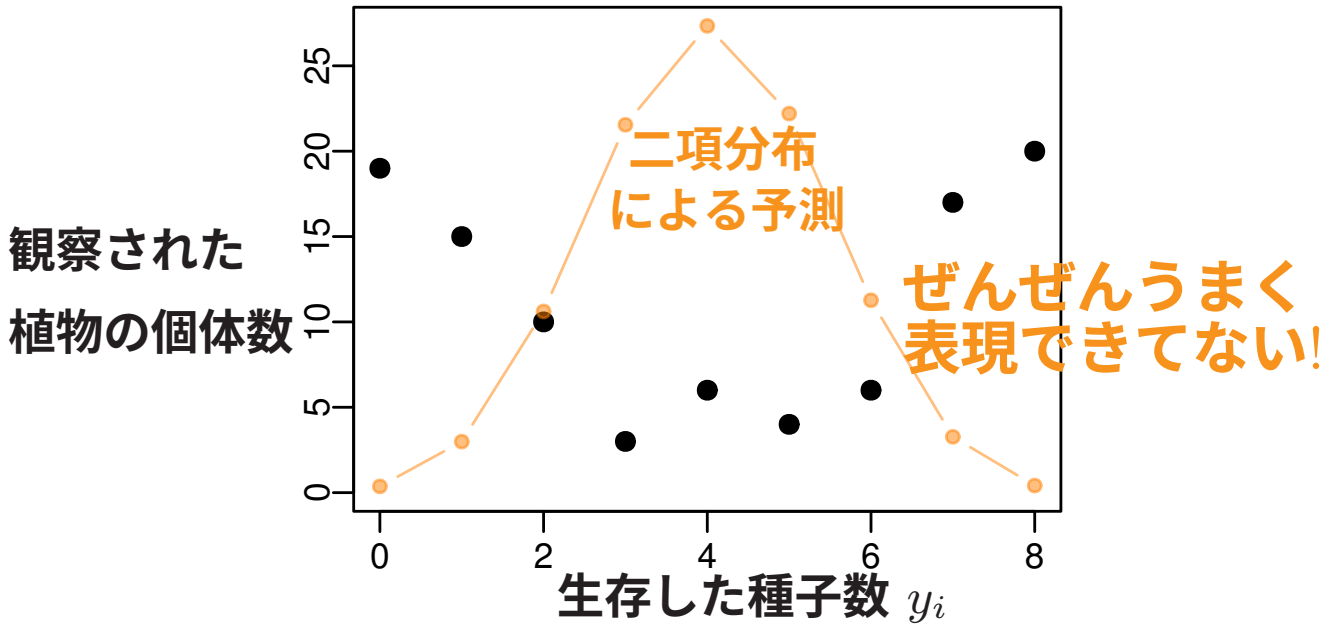
なんでも glm() 方針の問題点

```
> d2x9
   Ctgr
 X     A  B  C  D  E  F  G  H  I
   0 62 21 14 11 10 10  2  0  2
   1 48 34 22 17 16  7  2  1  1
```

- 分割表が大きくなったときに，自由に推定されるパラメーター数が増加
- カウント数の大小で推定値の信頼性がばらばらになる
- とくにゼロデータはこまる

二項分布では説明できない観測データ!

100 個体の植物の合計 800 種子中 403 個の生存が見られたので，平均生存確率は 0.50 と推定されたが……

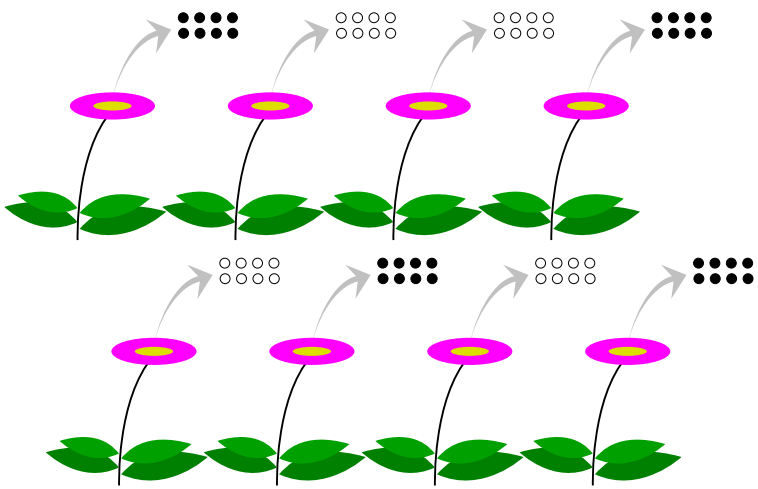


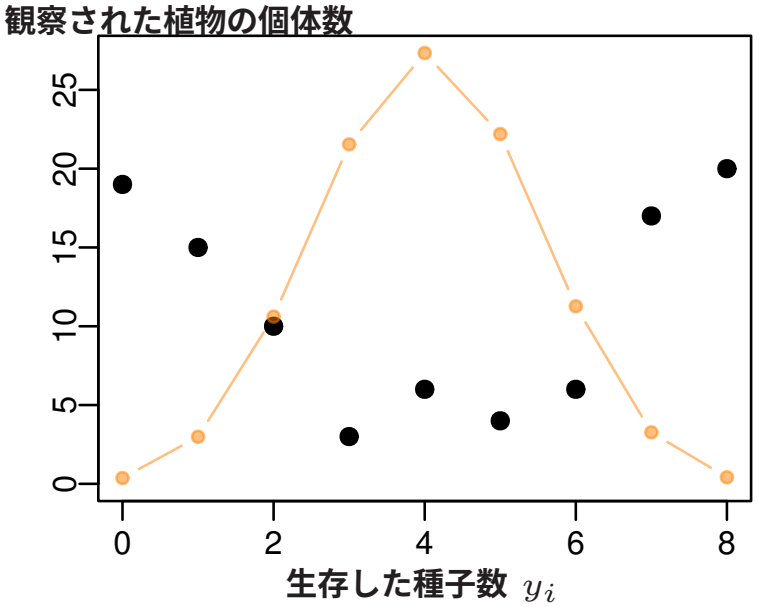
さっきの例題と同じようなデータなのに?

(「統計モデリング入門」第 10 章の最初の例題)

個体差 → 過分散 (overdispersion)

極端な過分散の例





- 種子全体の平均生存確率は 0.5 ぐらいかもしれないが……
- 植物個体ごとに種子の生存確率が異なる: 「個体差」
- 「個体差」があると overdispersion が生じる
- 「個体差」の原因は観測できない・観測されていない

モデリングやりなおし: 個体差を考慮する

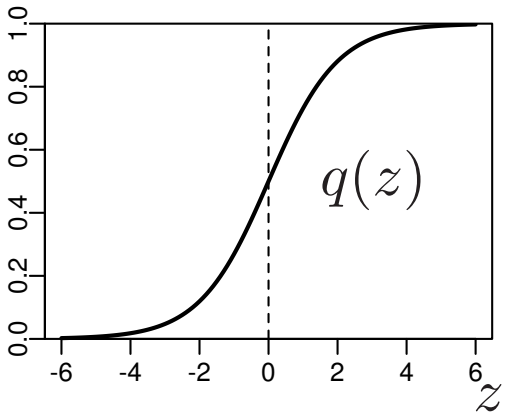
- 生存確率を推定するために 二項分布という確率分布を使う
- 個体 i の N_i 種子中 y_i 個が生存する確率は二項分布

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

- ここで仮定していること
  - 個体差があるので個体ごとに生存確率 q_i が異なる

GLM わざ: ロジスティック関数で表現する生存確率

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



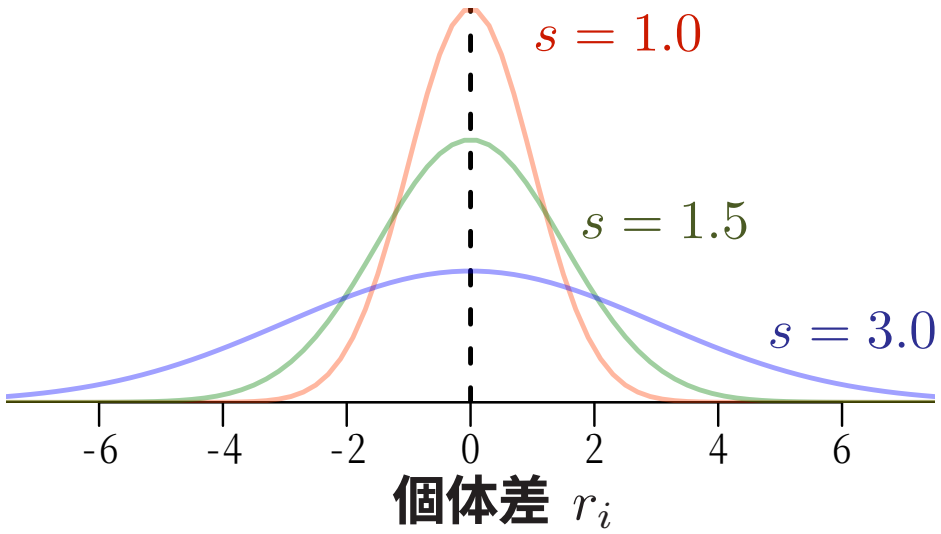
- **線形予測子** $z_i = a + r_i$ **とする**
  - **パラメーター** a**: 全体の平均**
  - **パラメーター** r_i**: 個体** i **の個体差 (ずれ)**

個々の個体差 r_i を最尤推定するのはまずい

- 100 個体の生存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……
- 個体ごとに生存数 / 種子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

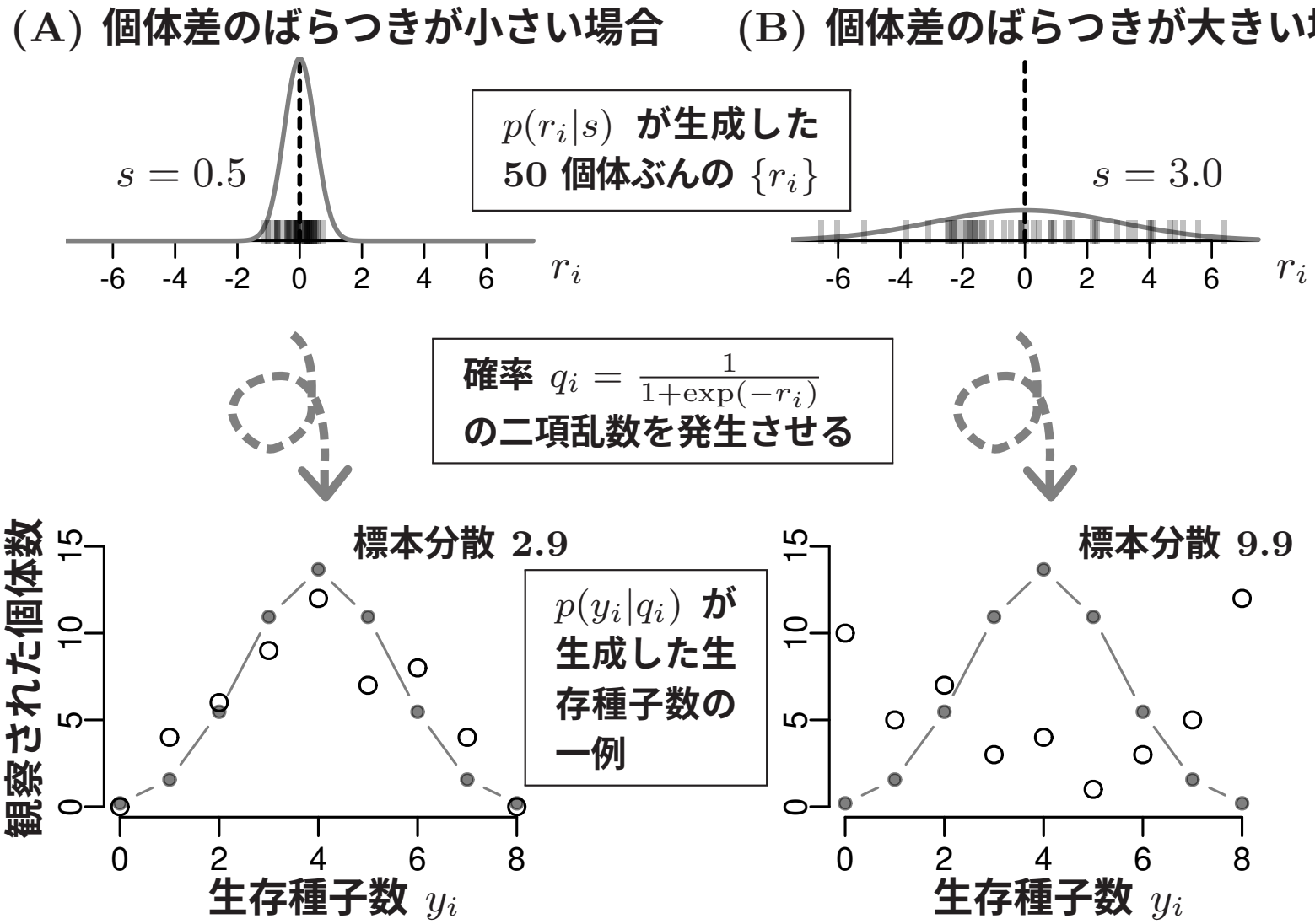
$\{r_i\}$ のばらつきは正規分布だと考えてみる



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

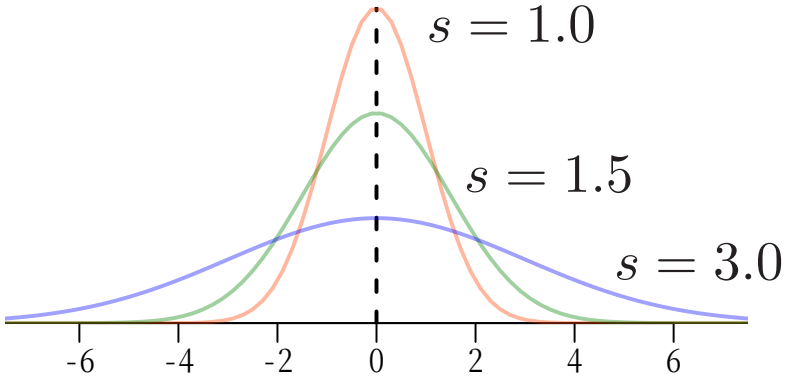
この確率密度 $p(r_i \mid s)$ は r_i の「出現しやすさ」をあらわしていると解釈すればよいでしょう．r_i がゼロにちかい個体はわりと「ありがち」で，r_i の絶対値が大きな個体は相対的に「あまりいない」．

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



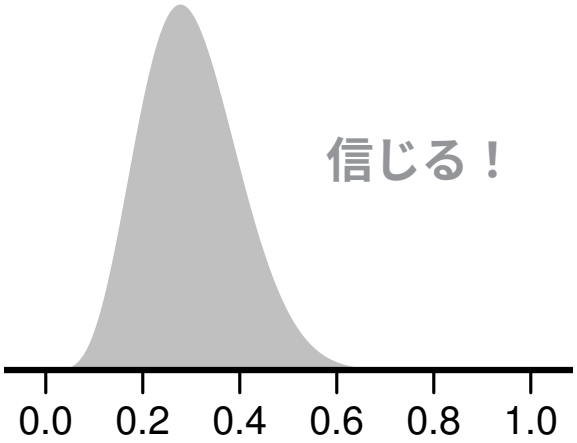
$$p(r_i \mid s) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{r_i^2}{2s^2}\right)$$

ベイズ統計モデルでよく使われる三種類の事前分布

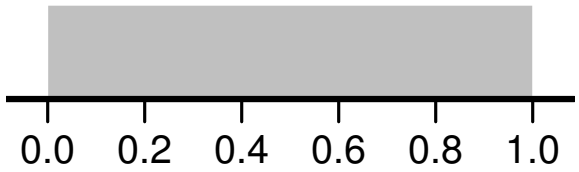
たいていのベイズ統計モデルでは，ひとつのモデルの中で複数の種類の事前分布を混ぜて使用する．

(A) **主観的な事前分布**

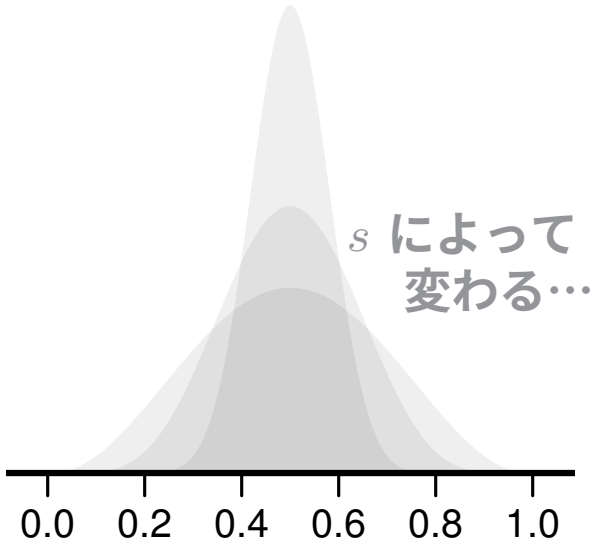
(できれば使いたくない!)



(B) **無情報事前分布**



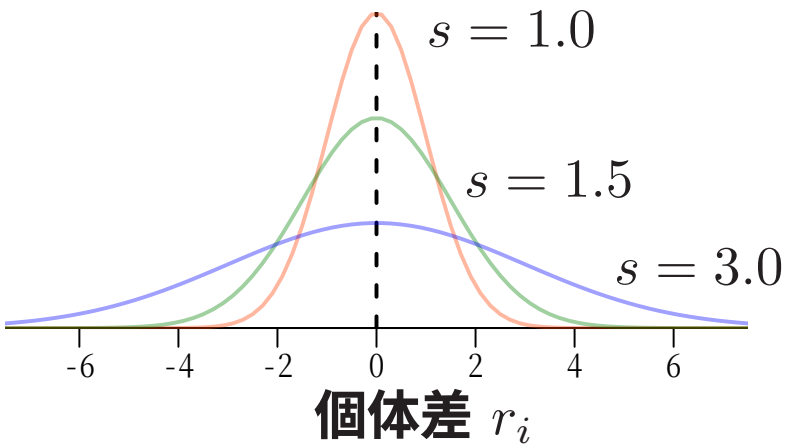
(C) **階層事前分布**



r_i の事前分布として階層事前分布を指定する

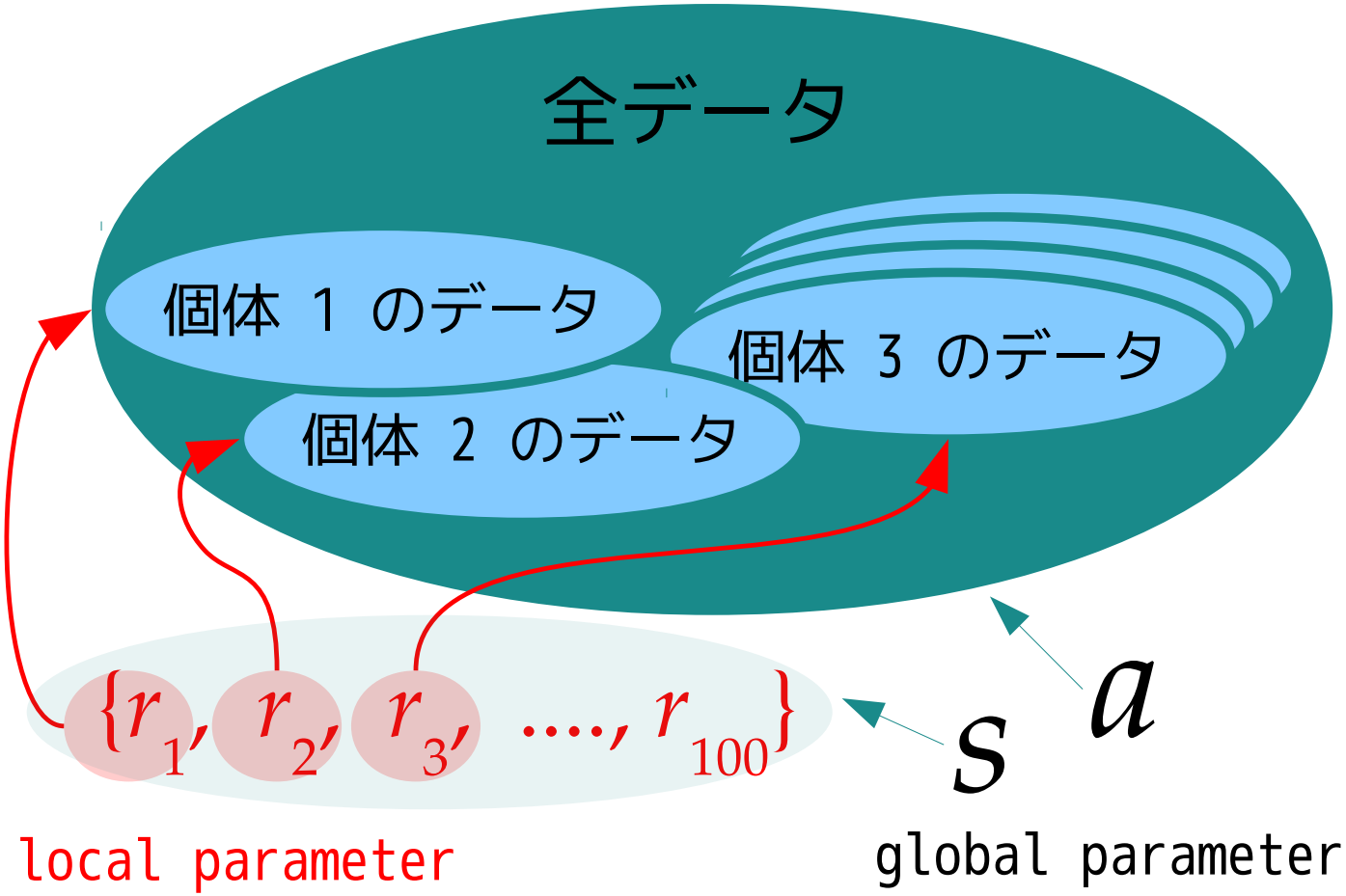
階層事前分布の利点

「データにあわせて」事前分布が変形!



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

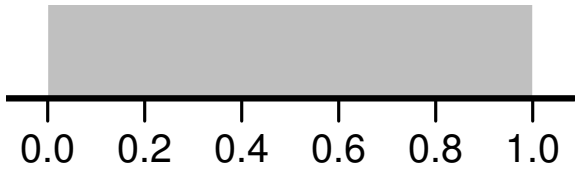
統計モデルの大域的・局所的なパラメーター



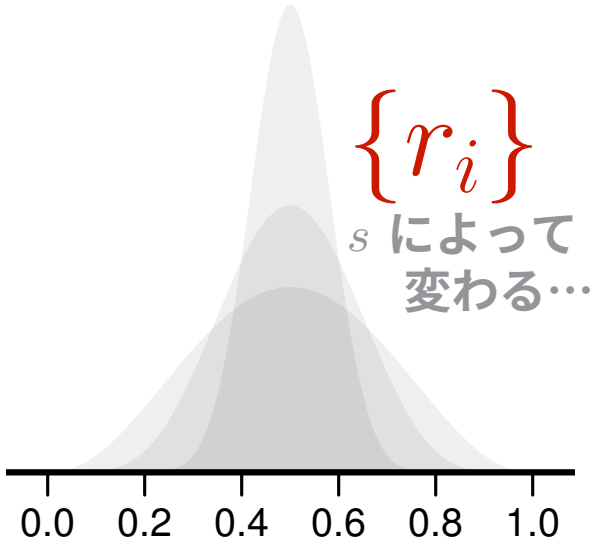
データのどの部分を説明しているのか?

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



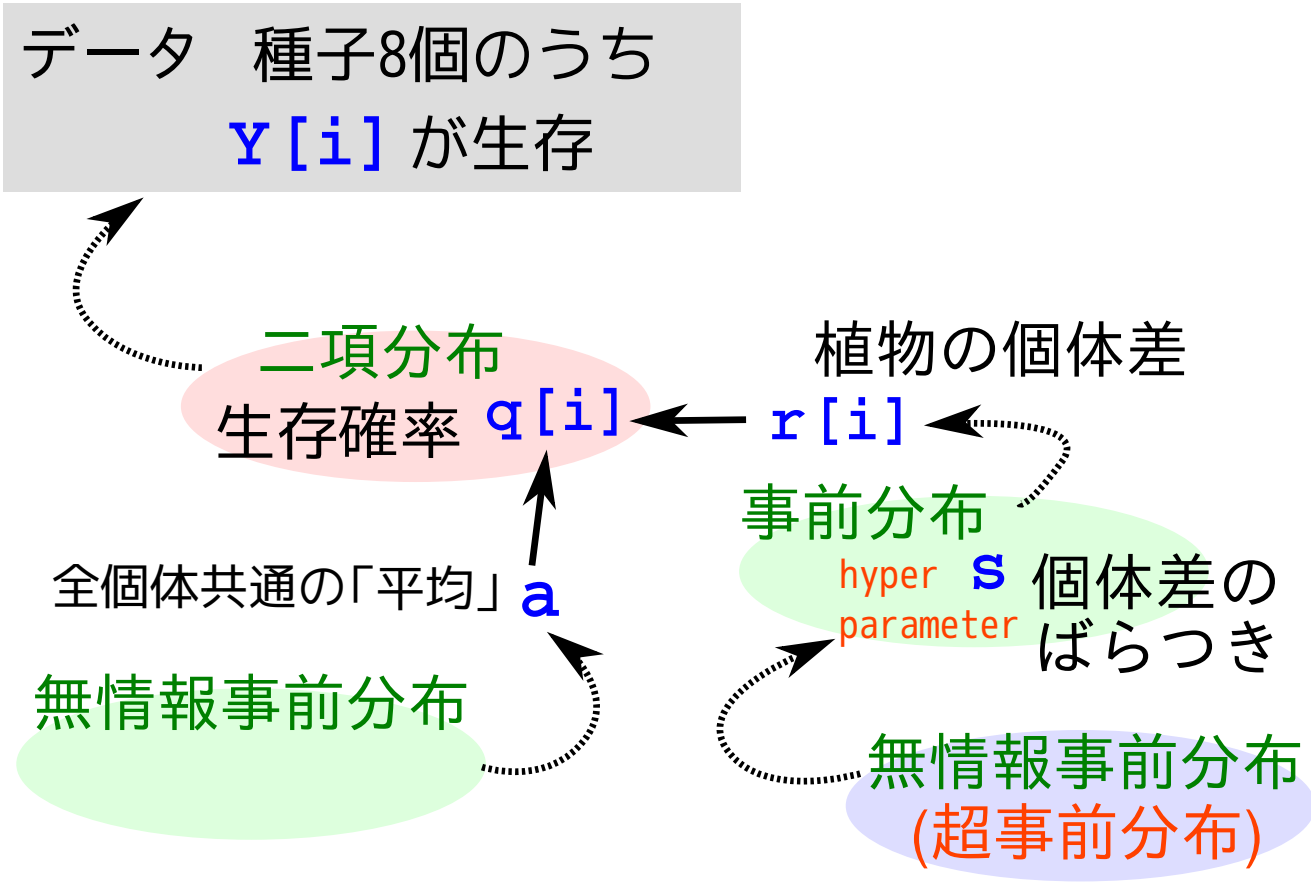
(C) 階層事前分布



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

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

さてさて，分割表のハナシにもどりましょう

	Ctgr								
X	A	B	C	D	E	F	G	H	I
0	62	21	14	11	10	10	2	0	2
1	48	34	22	17	16	7	2	1	1

ポアソン分布の GLMM ならどうだろう?

Ctgr の差だけしかあつかえない – X の効果は?

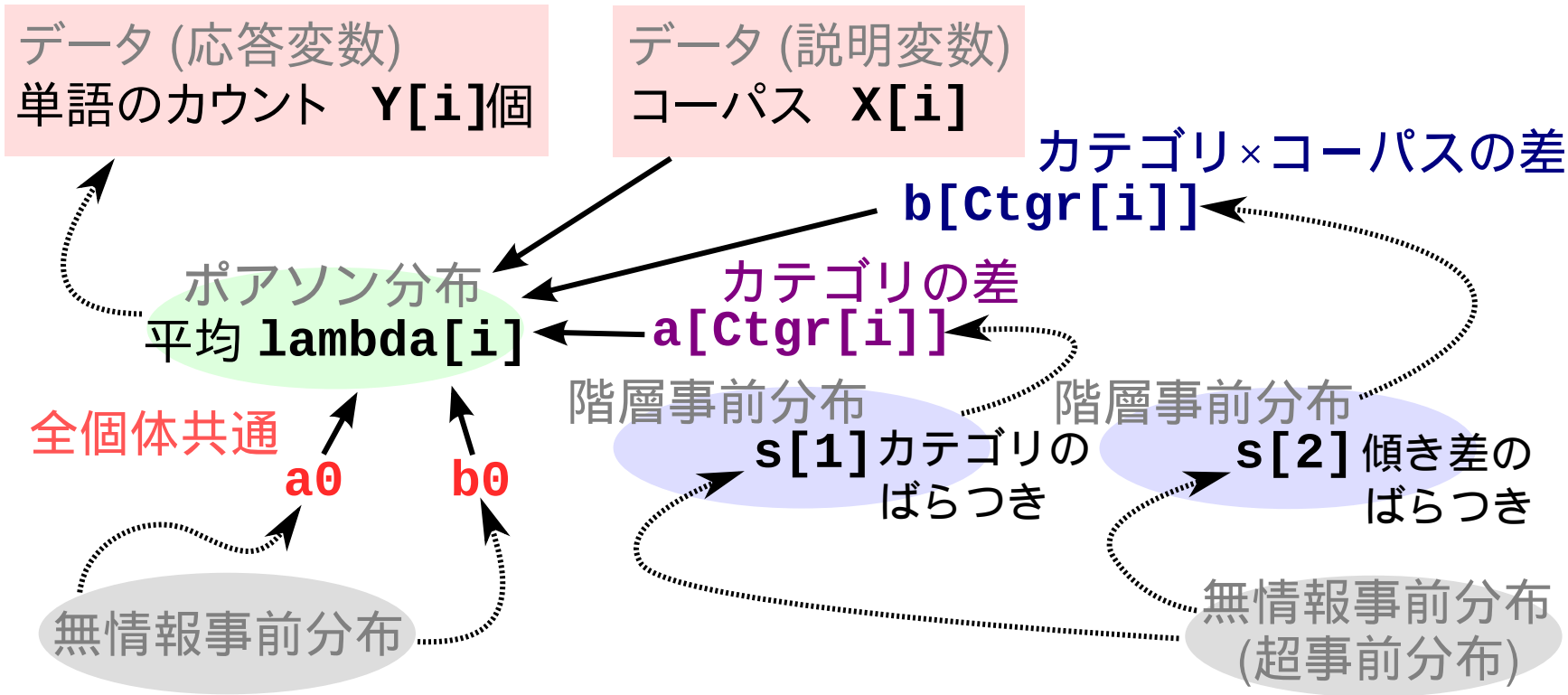
```
> (fit.glmm <- glmmML(y ~ X, data = d2x9,
+ cluster = Ctgr, family = poisson))
```

```
             coef se(coef)     z Pr(>|z|)
(Intercept) 2.012    0.465 4.323  1.5e-05
X           0.114    0.120 0.956  3.4e-01
```

```
> fit.glmm$posterior.modes
[1]  1.926862  1.230935  0.807098  0.557225  0.483854
[6]  0.067305 -1.218522 -2.005846 -1.426968
```

分割表の階層ベイズモデルの設計　線形ポアソン回帰

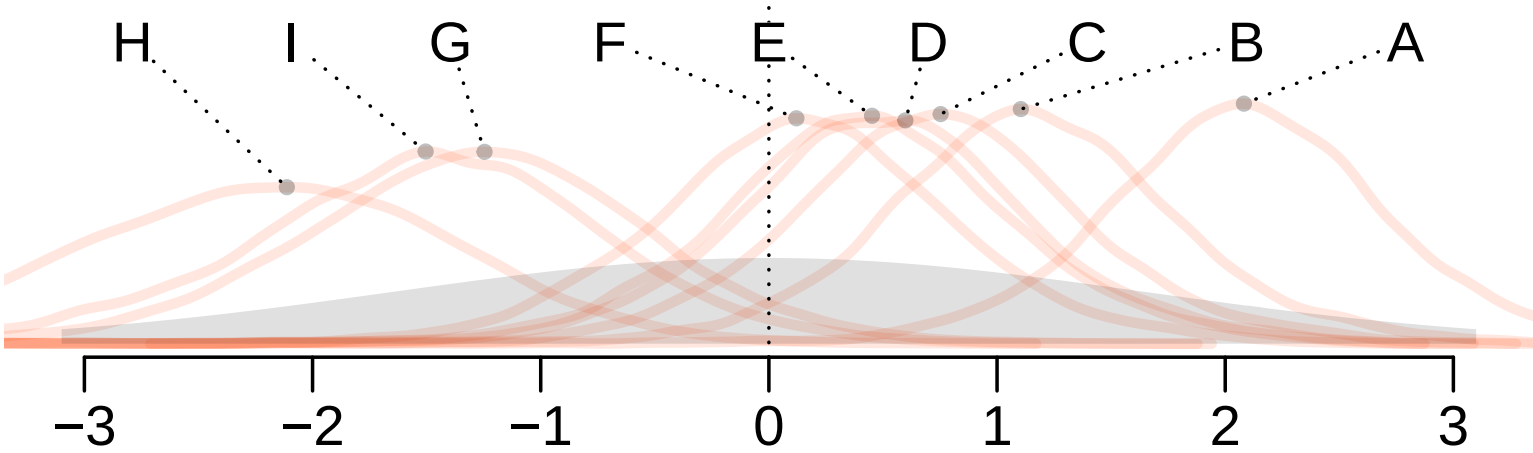
あるいは対数線形モデル



- JAGS を使ってパラメーター推定 (MCMC 法)

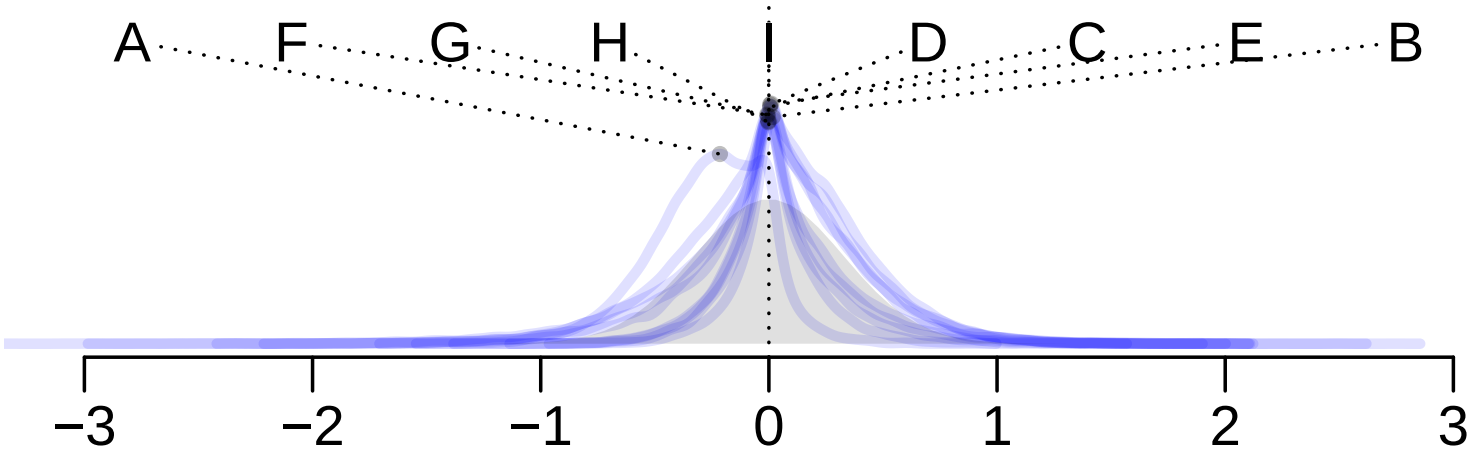
推定された事後分布 — カテゴリの差

	Ctgr								
X	A	B	C	D	E	F	G	H	I
0	62	21	14	11	10	10	2	0	2
1	48	34	22	17	16	7	2	1	1



推定された事後分布 — コーパス × カテゴリの差

	Ctgr								
X	A	B	C	D	E	F	G	H	I
0	62	21	14	11	10	10	2	0	2
1	48	34	22	17	16	7	2	1	1



(Section 6)

おわりに

分割表にデータを格納 — 階層ベイズモデルが必要に

	Ctgr								
X	A	B	C	D	E	F	G	H	I
0	62	21	14	11	10	10	2	0	2
1	48	34	22	17	16	7	2	1	1

- なんでもかんでも GLM でやるのは無理そう
  - とくにゼロデータをふくむ場合
- GLMM でカウント数ゼロ問題はなんとかなるが，複雑な状況に対処できない
  - この例題でいうとカテゴリ差・コーパス差がある場合など
- 階層ベイズモデルを使えば，多くの状況に対処できる