第 60 回日本生態学会大会 自由集会 (W30)

データ解析で出会う統計的問題 — その「割算」あぶなくない?

1. 「割算」やめて統計モデルで対処しよう

- 久保拓弥 (北海道大・地球環境) kubo@ees.hokudai.ac.jp

2. 連続的な量が分子にくる「割算」の場合

- 粕谷英一 (九州大・理)

この自由集会の web site の URL

http://goo.gl/E1cjA

ファイルや資料などをダウンロードできます

メモ! メモ!

久保のハナシ

- データどうしの「割算」をやめて統計モデルの工夫を!
- カウントデータの場合: ポアソン分布をうまく使ってみる
- 連続値データの場合: ガンマ分布を使えばいいのか……?

粕谷さんのハナシ

- 連続量同士の「割算」した量の素性は計算して確かめよう
- 正規分布っぽい連続量同士の「割算」した量は本当に必要かよく考えた方がいい
- 分母も分子も正規分布なら、そこには魔物が棲んでいる — コーシー分布

つまり要約すると……

久保のハナシ

- 「わりざん」はこわいからからにげろー!!

粕谷さんのハナシ

- 「わりざん」のこわい世界をのぞいてみたい……

(久保独白)

「割算値」のハナシは
もう何回くりかえしたことか……!

ということで総集篇的にといいますか

今回は「つかいまわし」がおおいです

今日の久保のハナシ

- 初級篇: 「データわるデータ」作法の問題点と簡単な対処
  - ポアソン回帰の **offset** 項わざ，ロジスティック回帰
- 中級篇: 複数の「割合」をポアソン分布で統計モデル化
  - 二項分布・多項分布のモデルをポアソン分布で
- ややめんどう篇: 連続値データの「比率」はどうする?
  - ガンマ分布で何とかなるか?
  - 他に何か工夫はないか?

「こういう自由集会は初めて」なヒトのための

とりあえず GLM 基礎知識

ハナシの前提: 統計ソフトウェア R

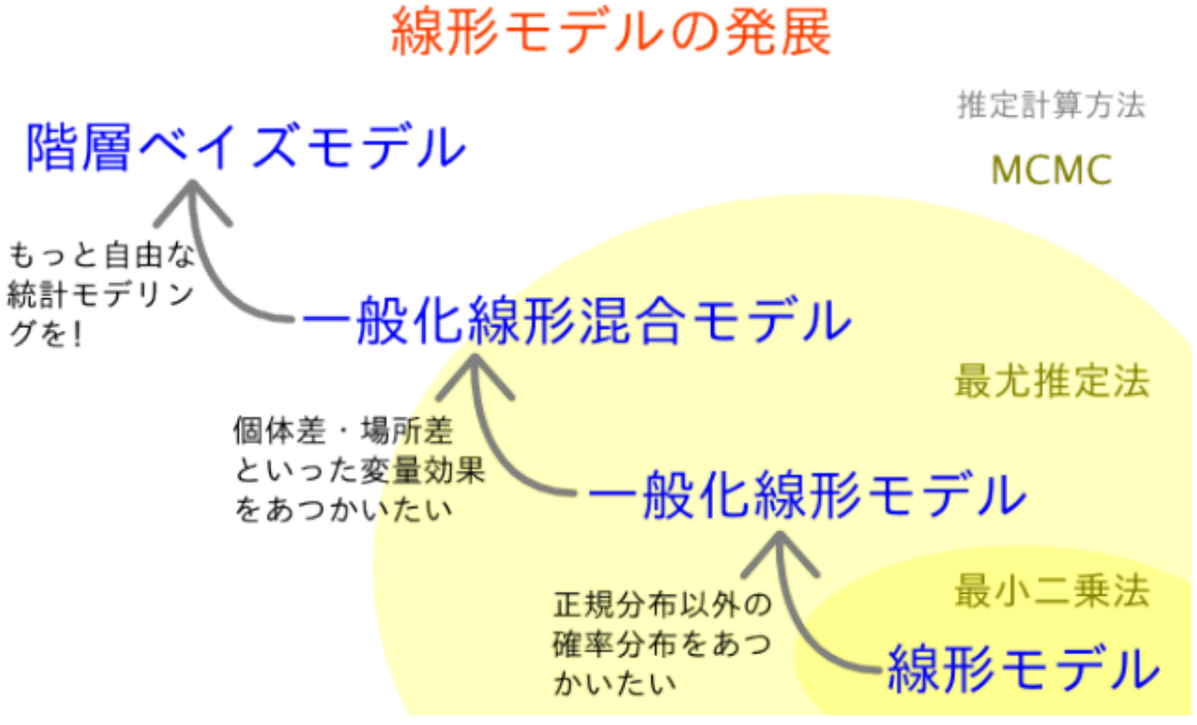
http://www.r-project.org/

- いろいろな OS で使える **free software**
- 使いたい機能が充実している
- 作図機能も強力
- S 言語によるプログラミング可能
- よい教科書が出版されつつある
  - たとえば，ですが……
  - 統計モデリング入門 http://goo.gl/Ufq2
  - R グラフィックス http://goo.gl/CKi2h

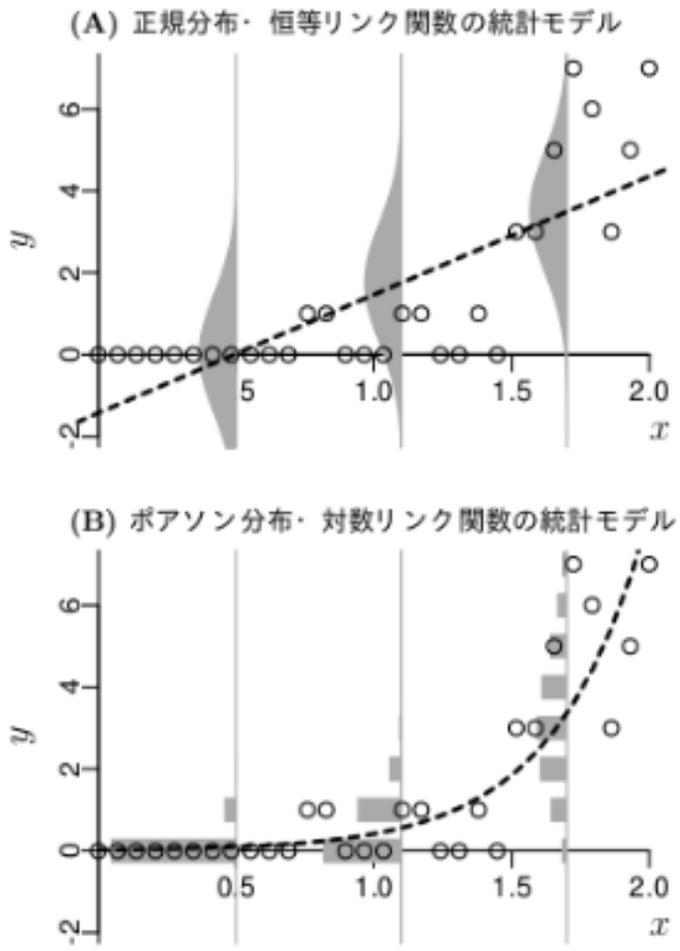
生態学での R の普及 - GLM

一般化線形モデル（GLM）を簡単にあつかえるソフトウェアとして R が注目されるようになった

私もこういう教科書かいてみました



GLM - ばらつきをよく見る



線形モデルの発展

推定計算方法

階層ベイズモデル

MCMC

もっと自由な統計モデリングを!

一般化線形混合モデル

最尤推定法

個体差・場所差といった変量効果をあつかいたい

一般化線形モデル

正規分布以外の確率分布をあつかいたい

最小二乗法

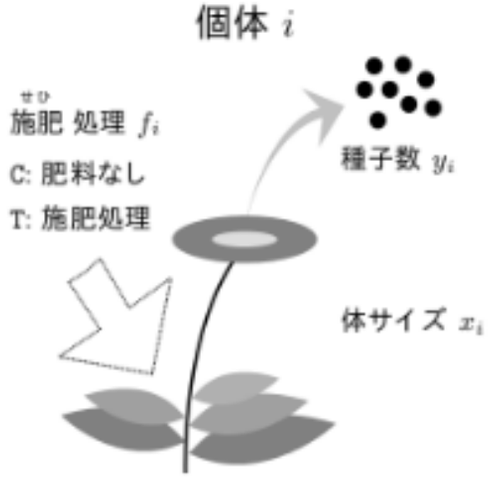
線形モデル

0 個，1 個，2 個と数えられる種子数が「正規分布」なわけないだろ!!

3.9 回帰モデルと確率分布の関係．また別の架空データに対して GLM をあてはめた例．破線は x とともに変化する平均値．グレイで

R でデータを読みこむ

→ data.frame



この例題に登場する架空植物の第 i 番目の個体．この
イズ（ 個体の大きさ） x_i と肥料をやる施肥処理 f_i が種
う影響しているのかを知りたい．

web サイト (まえがき末尾を参照) からダウンロードできます．データのファイル名は data3a.csv であるとしましょう．R では

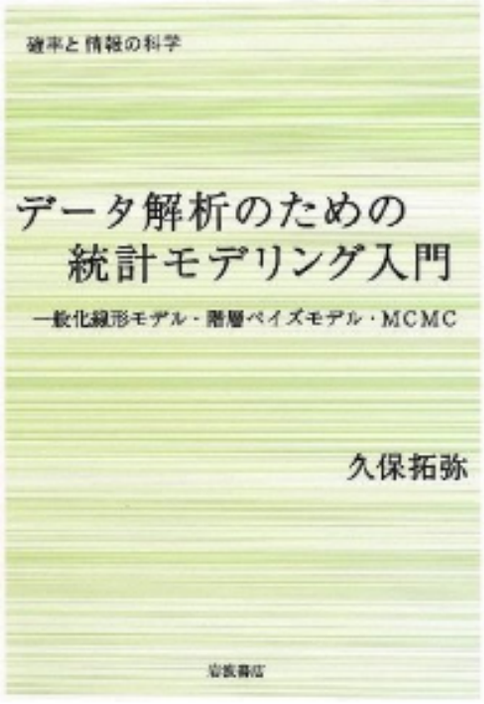
```
> d <- read.csv("data3a.csv")
```

と命じるだけでファイルを読みこんで，その内容を格納したデータフレイムに d という名前が付けられます．このデータフレイムというデータ構造は，とりあえず「 表 (table) のようにあつかえるデータ構造」であると考えてください．R のコマンドプロンプトで d あるいは print(d) とすると，全 100 個体ぶんのデータがディスプレイ上に表示されます[*4]．

```
> 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
```

このように d というオブジェクトには，全 100 個体ぶんのデータがあたかも 100 行 3 列の行列のような形式で格納されているように見えます．このデータ

R でデータをよくながめる!



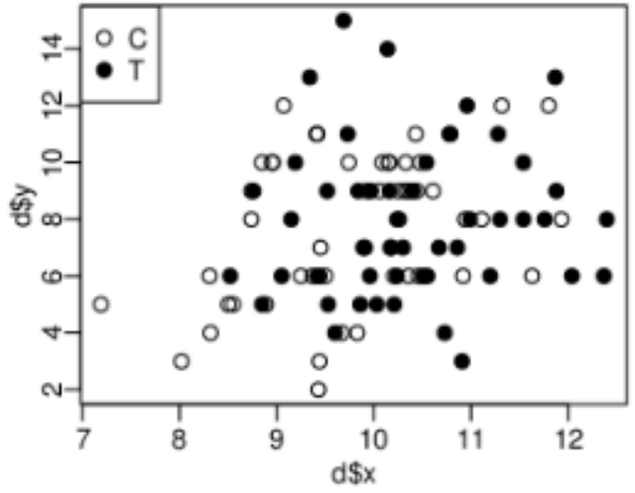


図 3.2 例題の架空データの図示．植物の種子数 y_i と，体サイズ x_i や施肥処理 f_i の関係を示している．白丸は施肥処理なし（処理 C），黒丸は施肥処理あり（処理 T）．

タのばらつきかたを視覚的に把握するようにしましょう[*6]．

観測データに余計な手を加えないで，データ全体をよく見るには plot() 関数などを使うと便利でしょう．

```
> plot(d$x, d$y, pch = c(21, 19)[d$f])
```

GLM をデータにあてはめる

glm() 関数

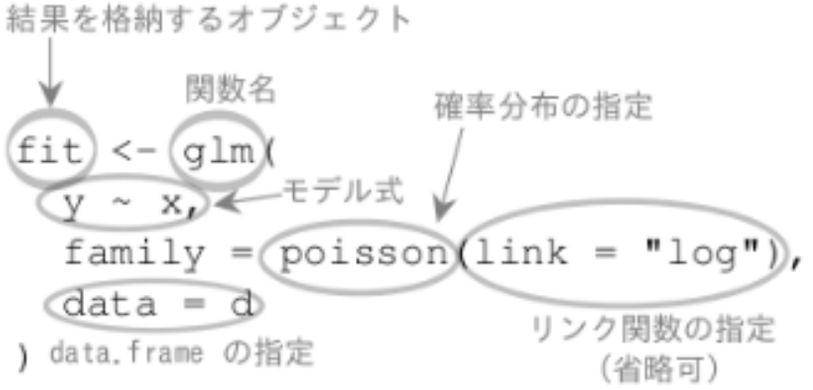


図 3.5　glm() 関数の引数の指定方法.

時間があれば，ちょっと実演します

R で統計モデルの予測を図示

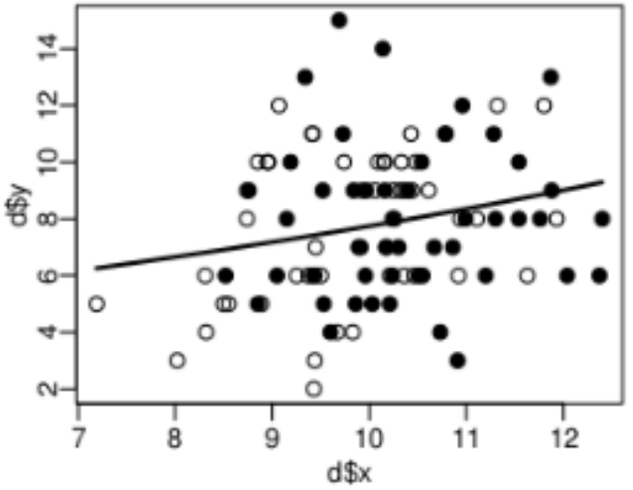
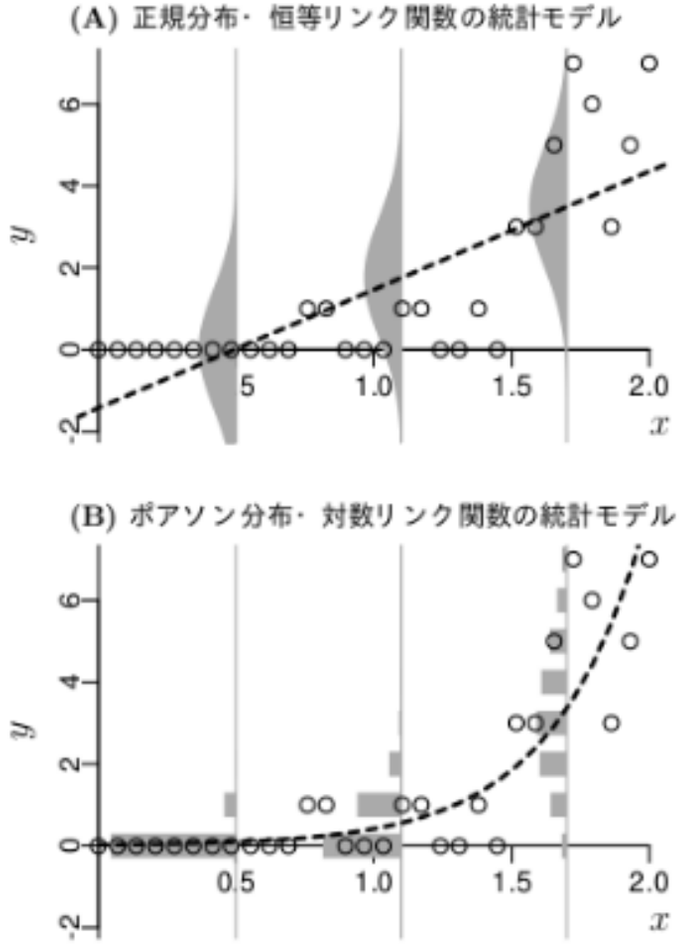


図 3.7 平均種子数 λ の予測．図 3.2 に λ の予測値（実線）を上がきしたもの．

植物の体のサイズが大きくなると
平均種子数が増えます……という
ことをあらわしている予測

GLM による改善，その次の改善



3.9　回帰モデルと確率分布の関係．また別の架空データに対して GLM をあてはめた例．破線は x とともに変化する平均値．グレイで

線形モデルの発展

推定計算方法

階層ベイズモデル

MCMC

もっと自由な統計モデリングを！

一般化線形混合モデル

最尤推定法

個体差・場所差といった変量効果をあつかいたい

一般化線形モデル

最小二乗法

正規分布以外の確率分布をあつかいたい

線形モデル

「なんでもかんでもセンをひけばよかろう」というドグマから脱出するための GLM

統計モデリング: 観測データのモデル化

- 統計モデルは観測データのパターンをうまく説明できるようなモデル
- 基本的部品: 確率分布 (とそのパラメーター)
- データにもとづくパラメーター推定，あてはまりの良さを定量的に評価できる

初級篇 (久保講義資料 http://goo.gl/76c4i)

「データわるデータ」作法の問題点と簡単な対処の例

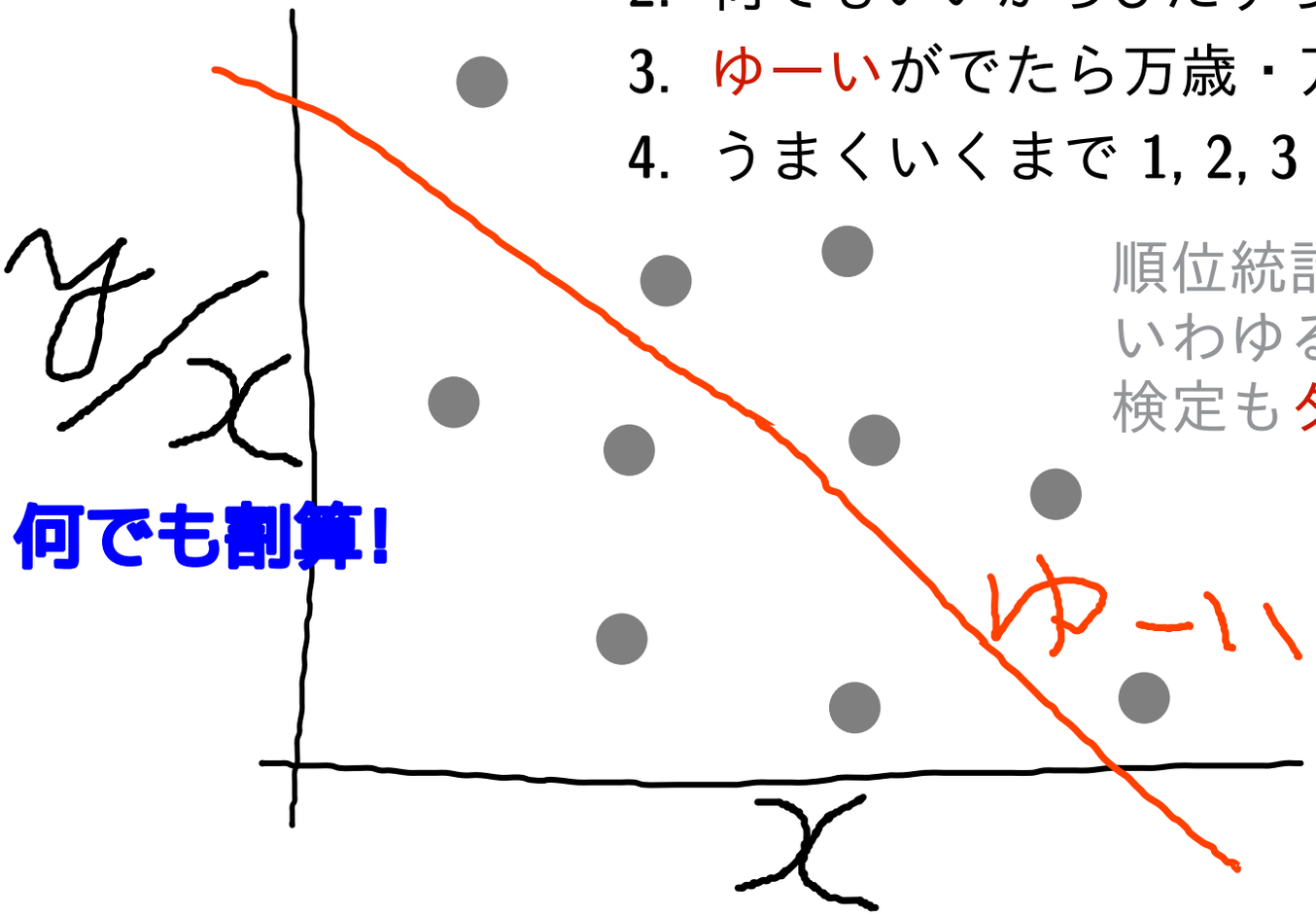
「データわるデータ」なんて必要ないでしょ?

対策: **offset** 項ポアソン回帰とロジスティック回帰

この悪しき「割算」な統計学

世間でよくみかけるおススめできない作法の例

1. データどんどん割算・割算
2. 何でもいいからひたすらセンをひく
3. ゆーいがでたら万歳・万歳
4. うまくいくまで 1, 2, 3 ぐるぐる



順位統計量つかった
いわゆるノンパラメトリック
検定もダメですよ!

ちなみにこれは x と $\circ/x$ を比較してるんだから，反比例みたいな偽「負の相関」がでるのはあたりまえ

割算値ひねくるデータ解析はなぜよくないのか?

- 観測値 / 観測値 はどんな確率分布にしたがうのか?

  — これは見とおしが悪く，めんどうのもとになりがち

- 情報が失われる: 「10 打数 3 安打」と「200 打数 60 安打」，「どちらも 3 割バッター」と言ってよいのか?

- 割算値を使わないほうが見とおしのよい，合理的なデータ解析ができる (今回の授業の主題)

- したがって割算値を使ったデータ解析は不利な点ばかり，そんなことをする必要性はどこにもない

避けられるわりざん，避けにくいわりざん

- 避けられる割算値
  - 密度などの指数

    例: 人口密度，specific leaf area (SLA) など

    初歩的な対策: **offset** 項わざ
  - 確率

    例: N 個のうち k 個にある事象が発生する確率

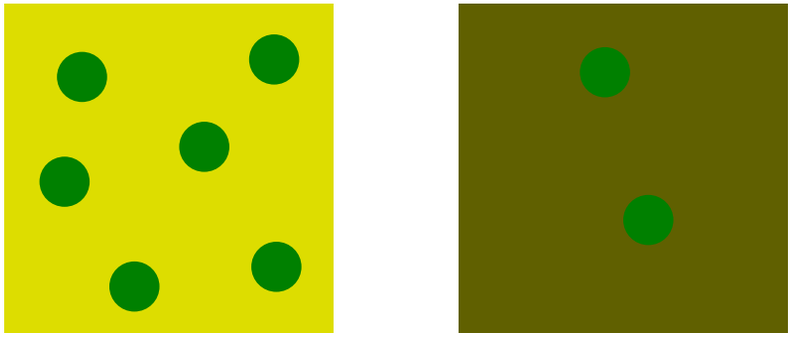
    初歩的な対策: ロジスティック回帰など二項分布モデルで
- 避けにくい割算値
  - 測定機器が内部で割算した値を出力する場合
  - 割算値で作図せざるをえない場合があるかも

初級わざ (1)「脱」割算の offset 項わざ

ポアソン回帰を強めてみる

例題: 調査区画内の個体密度は明るさで変わるか?

- 何か架空の植物個体の密度が「明るさ」 x に応じて どう変わるかを知りたい
- 明るさは $\{0.1, 0.2, \cdots, 1.0\}$ の 10 段階で観測した

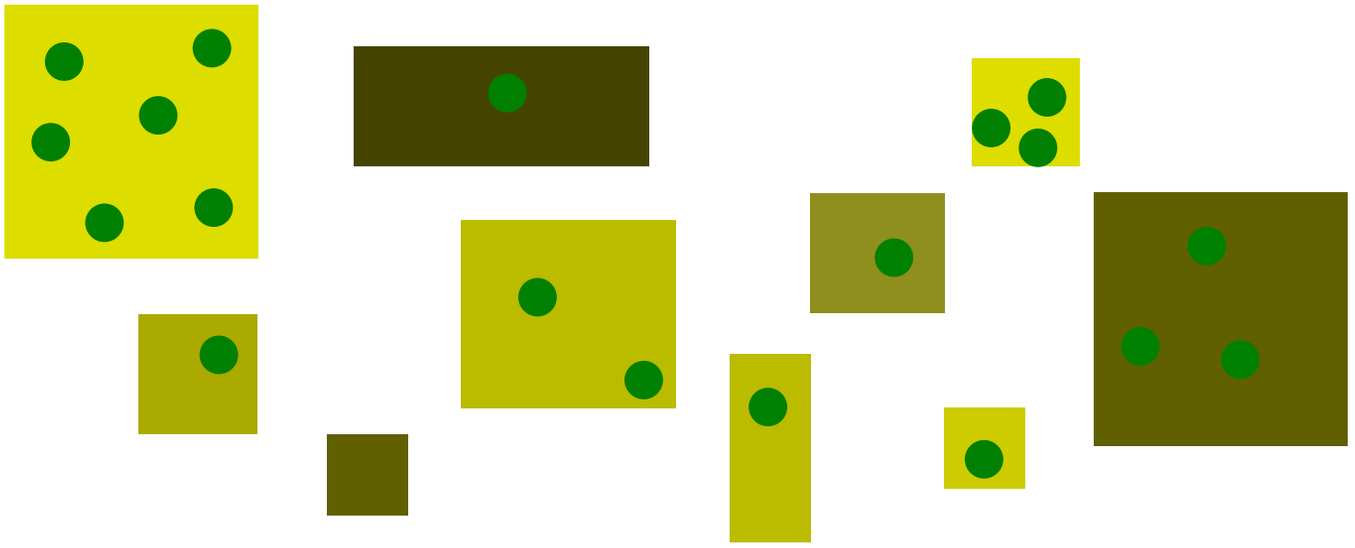


x 大
明るい

x 小
暗い

これだけなら単純に `glm(..., family = poisson)` すればよいのだが ……

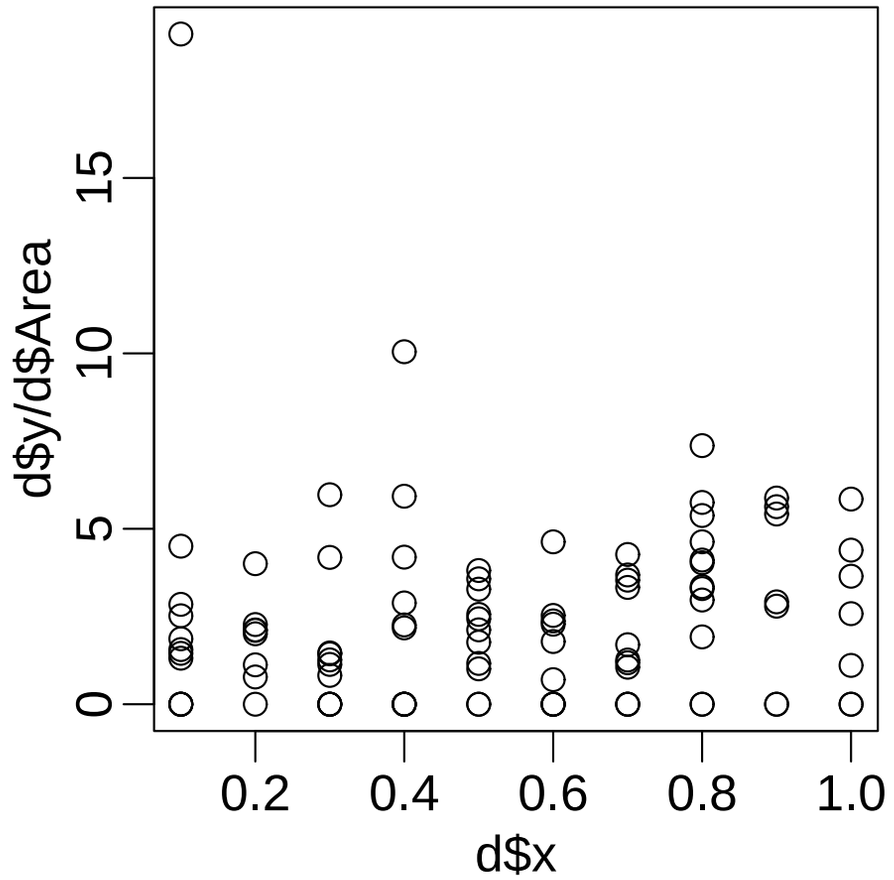
「場所によって調査区の面積を変えました」?!!



- 明るさ x と面積 A を同時に考慮する必要あり
- ただし「密度 = 個体数 / 面積」といった割算値解析はやらない!
- `glm()` の `offset` 項わざでうまく対処できる
- ともあれその前に観測データを図にしてみる

明るさ vs 割算値図の図

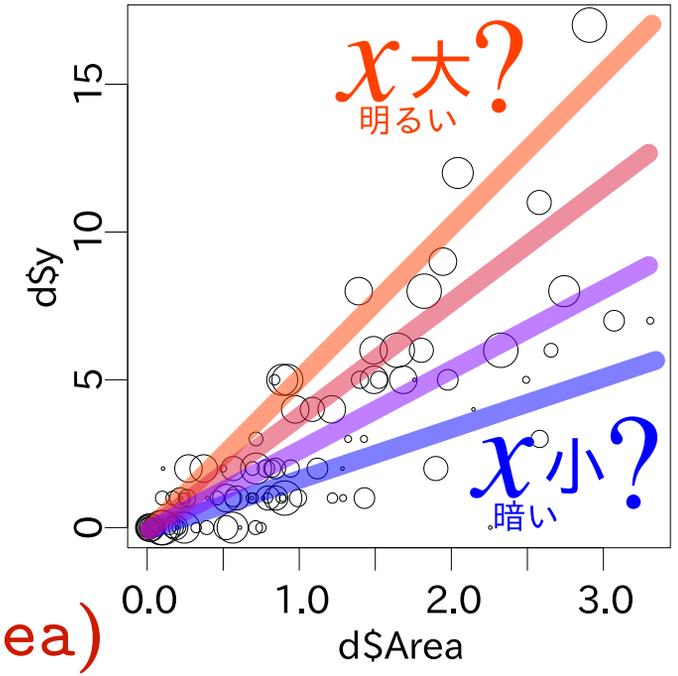
```
plot(d$x, d$y / d$Area)
```



- いまいちよくわからない……

この問題は GLM であつかえる!

- family: poisson, ポアソン分布
- link 関数: "log"
- モデル式 : y ~ x
- offset 項の指定 : log(Area)



- 線形予測子 $z = a + b\,\mathrm{x} +$ log(Area)

  a, b は推定すべきパラメーター
- 応答変数の平均値を λ とすると $\log(\lambda) = z$

  つまり $\lambda = \exp(z) = \exp(a + b\,\mathrm{x} +$ log(Area)$)$
- 応答変数 は平均 λ のポアソン分布に従う:

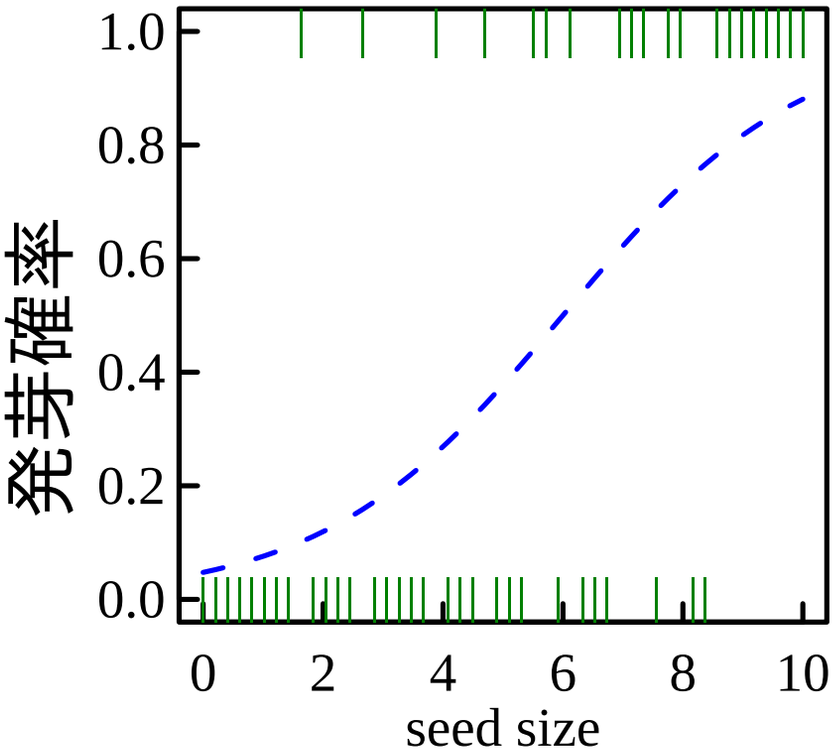
初級わざ (2) ロジスティック回帰

おススめできない解析と対比しつつ

架空植物の発芽実験データ

種子サイズと発芽確率の関係を調べる実験やってみた

| は観測データ (1 = 発芽した)

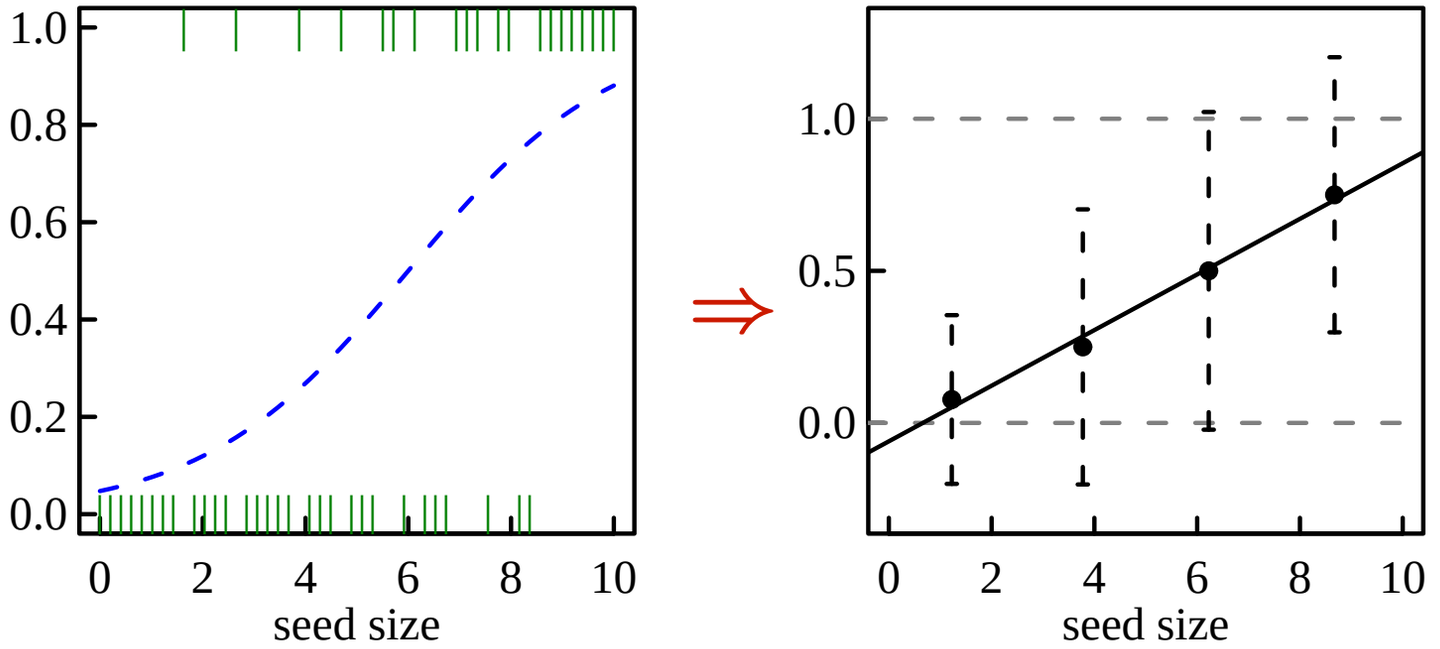


[「ホント」の発芽パターン]

- 種子が大きいほど発芽確率が高い
- 発芽確率は青破線 で示されているように上昇する

データから 青破線 (つまり真のモデル・母集団) を推定したい

(よく見かける) おススめできない解析の一例



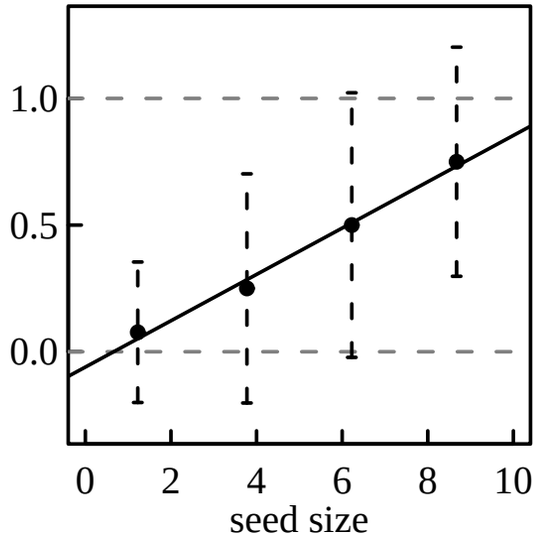
1. てきとーに種子サイズの区画を取る (上の例だと 4 区画)
2. 区画ごとに縦横の平均値など計算; $\{0, 1\}$ データを割算値に
3. 何も考えずに統計ソフトウェアにほうりこむ

(直線回帰する or 「分散分析」する or 「検定」& 多重比較する)

なぜよろしくないか? データの特徴を無視

区画はてきとー

区画のとりかたで結果は変わる



割算すると情報が失われる

1 / 2 と 100 / 200 は違う!

等分散でもなければ正規分布でもない

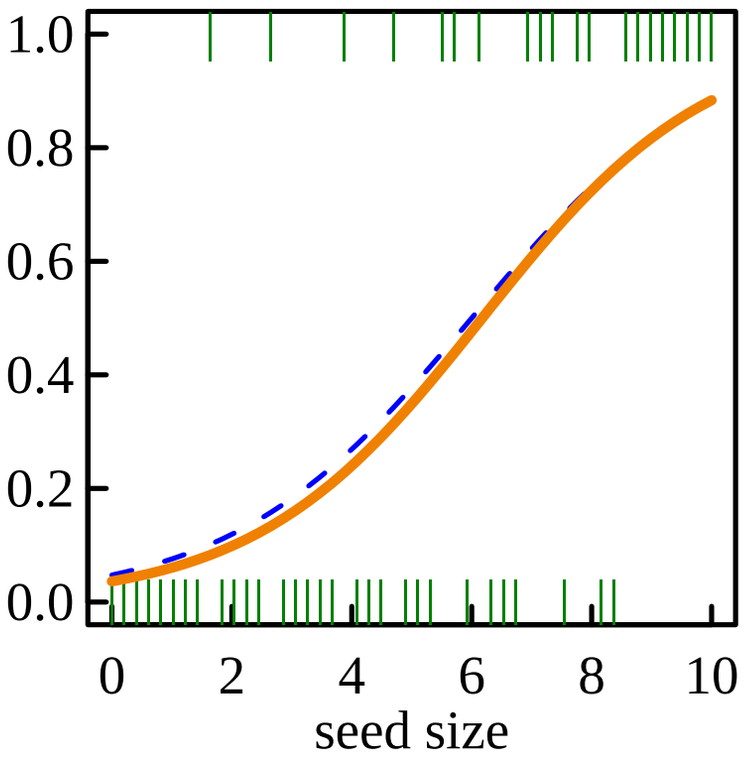
ということで直線回帰も分散分析も使えん — さらに，いわば母分散が異なる状況なので，ノンパラメトリック検定のたぐいもだめ

何を推定してるのだろうか?

発芽する確率がマイナスになったり，1 をこえたりするモデルってのは ……? (変数変換すればいいって? そのワザは呪われてる)

R の glm() で推定: ロジスティック回帰の例

発芽する・しないが二項分布にしたがうと仮定している



- 各種子について，そのサイズ (x) と“発芽した or しなかった”の対応をみる
- 発芽確率 q を以下のように仮定

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

(logistic 式)

- パラメーター a と b の推定値を最尤推定法で計算する
- ここでは R の glm() 関数を使った (上の図の赤線が推定結果)

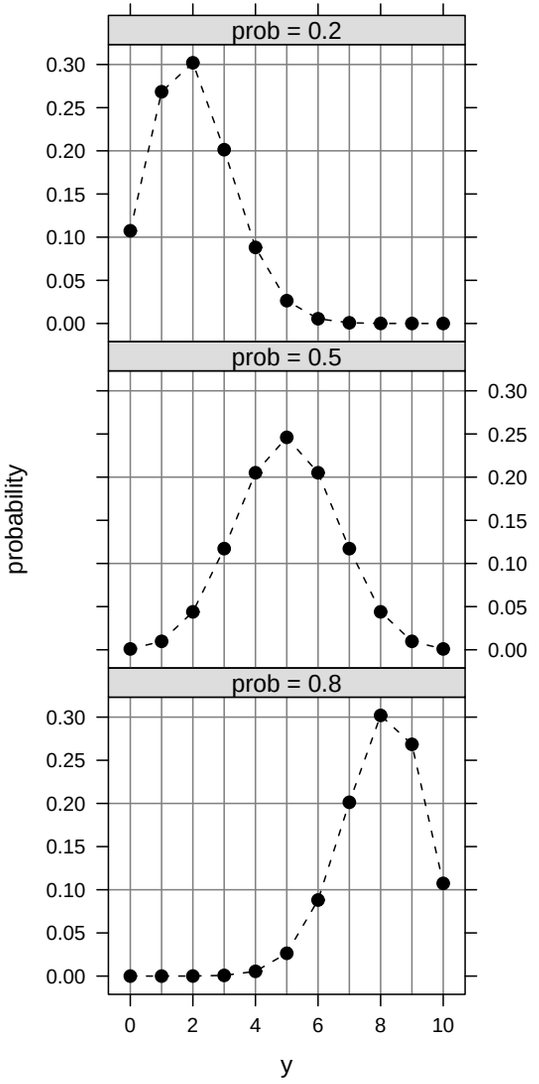
二項分布 (binomial distribution) とは何か?

- 離散分布 $y_i \in \{0, 1, 2, \cdots, N\}$
- 確率分布 (paramter: q, N)

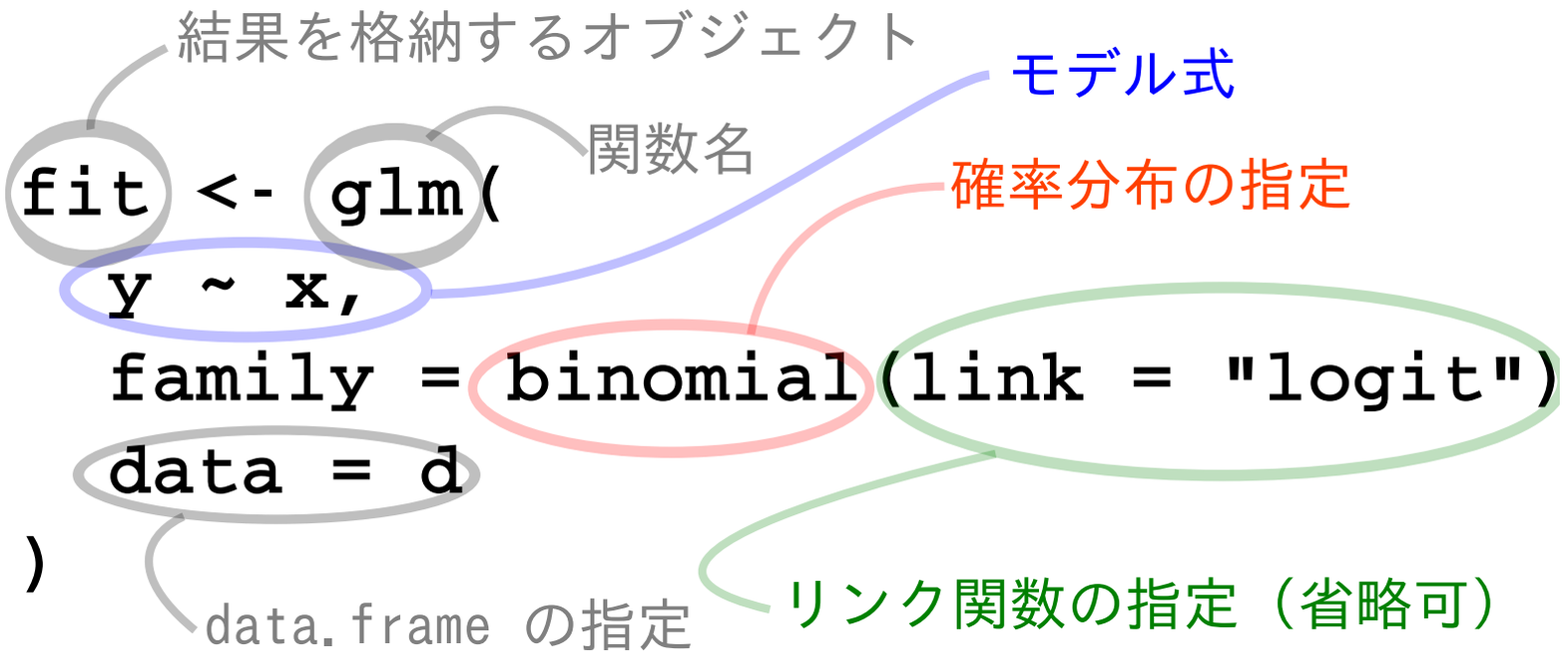
$$\binom{N}{y} q^y (1-q)^{N-y}$$

- 平均 Nq, 分散 $Nq(1-q)$

- 上限のあるカウントデータに
- 例: N 個体中 y 個体に反応があった，死亡した，など



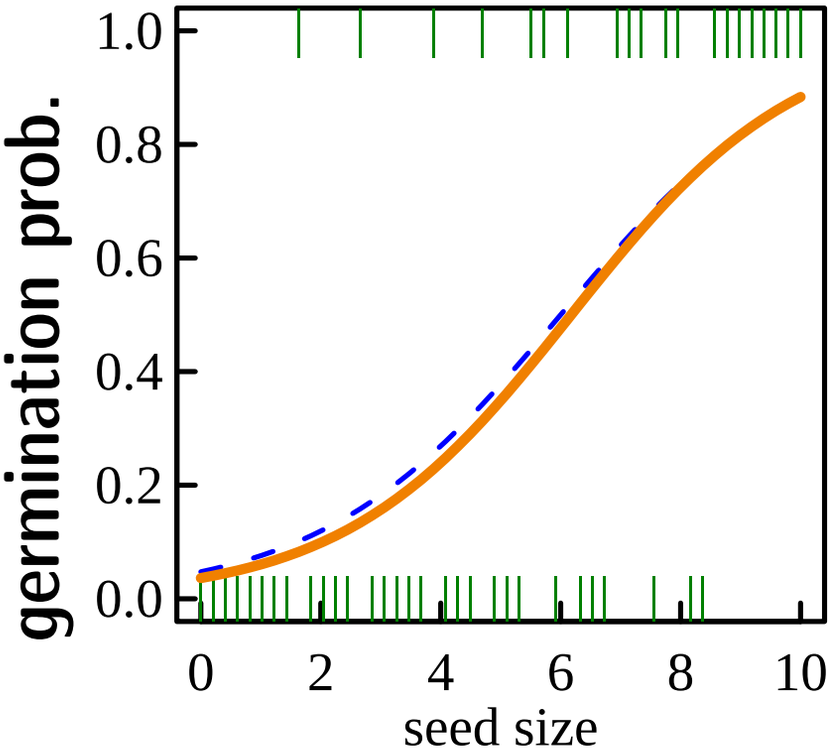
R の glm() 関数: 何を指定すればいい?



- モデル式 (線形予測子 z): 種子重 x が説明変数
- link 関数: logit リンク関数
- family: binomial, 二項分布

良い推定 (データ $\rightarrow$ モデル) をめざして Ending

おススめできないデータ解析を回避するための注意点



- むやみに 区画わけしない！
- 何でも 割り算するな！
- たくさん 図を描く
- 「観測データを説明する 確率分布は何か?」 を考える

コツ: 不自然にデータをこねくりまわさない

データの性質・構造にあったモデリングを!

中級篇 (ESJ59 自由集会 http://goo.gl/qQ1Ok)

「割合」をポアソン分布であつかってみる

対策: ポアソン回帰 glm() の使いかたを工夫

…… 二項分布はともかく多項分布はしんどい ……

ここで説明したいこと: ポアソン分布だけでなんとかしたい

二項分布はともかく，多項分布は回避したい

Related distributions

- If $X_1 \sim \mathrm{Pois}(\lambda_1)$ and $X_2 \sim \mathrm{Pois}(\lambda_2)$ are independent, then the difference $Y = X_1 - X_2$ follows a Skellam distribution.
- If $X_1 \sim \mathrm{Pois}(\lambda_1)$ and $X_2 \sim \mathrm{Pois}(\lambda_2)$ are independent, then the distribution of X_1 conditional on $X_1 + X_2$ is a binomial distribution. Specifically, given $X_1 + X_2 = k$, $X_1 \sim \mathrm{Binom}(k, \lambda_1/(\lambda_1 + \lambda_2))$. More generally, if $X_1, X_2, \ldots, X_n$ are independent Po random variables with parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$ then given $\sum_{j=1}^{n} X_j = k, X_i \sim \mathrm{Binom}\left(k, \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}\right)$. In fact, $\{X_i\} \sim \mathrm{Multinom}\left(k, \left\{\frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}\right\}\right)$.
- If $X \sim \mathrm{Pois}(\lambda)$ and the distribution of Y, conditional on X = k, is a binomial distributio $Y|(X = k) \sim \mathrm{Binom}(k, p)$, then the distribution of Y follows a Poisson distributio $Y \sim \mathrm{Pois}(\lambda \cdot p)$. In fact, if $\{Y_i\}$, conditional on X = k, follows a multinomial distributic $\{Y_i\}|(X = k) \sim \mathrm{Multinom}(k, p_i)$, then each Y_i follows an independent Poisson distribution $Y_i \sim \mathrm{Pois}(\lambda \cdot p_i), \rho(Y_i, Y_j) = 0$.
- The Poisson distribution can be derived as a limiting case to the binomial distribution as the number of trials goes to infinity and the expected number of successes remains fixed — see l rare events below. Therefore it can be used as an approximation of the binomial distribution

ポアソン分布で OK というハナシ `http://en.wikipedia.org/wiki/Poisson_distribution`

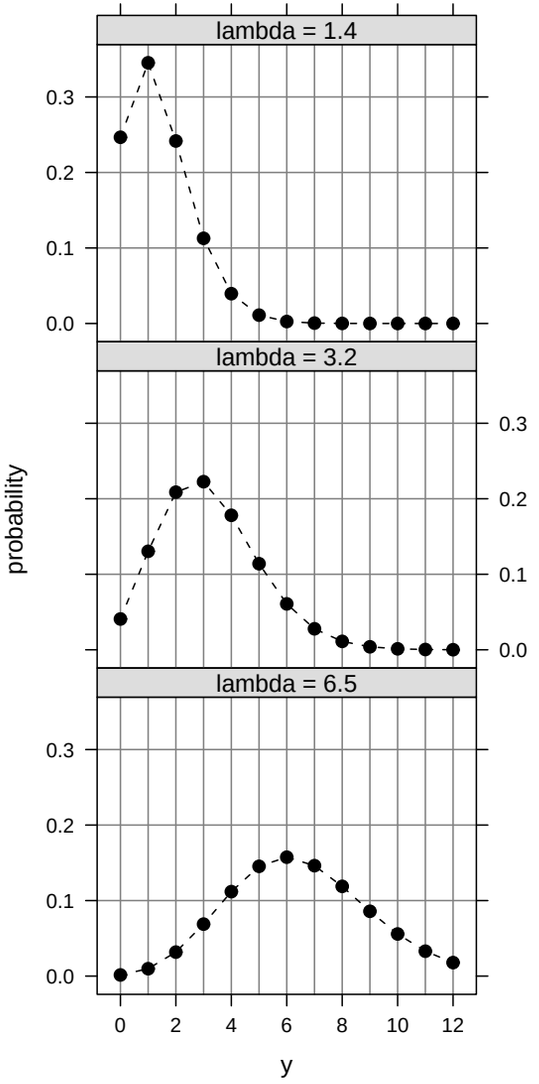
ポアソン分布 (Poisson distribution) とは何か?

- 離散分布 $y \in \{0, 1, 2, \cdots, \infty\}$
- 確率密度関数 (paramter: λ)

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

- 期待値 λ, 分散 λ

- 上限を設定できないカウントデータに
- 例: 産卵数・種子数・個体数……



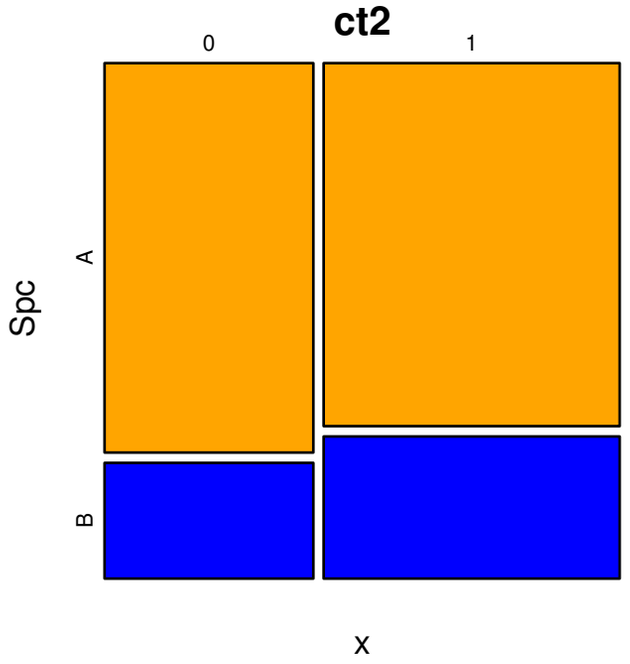
2×2 分割表の統計モデル

ポアソン回帰の `glm()` を使ってみる

対数線形モデル **log-linear model**

前回の自由集会の架空例題データ: 種ごとの個体数

```
> d2 <- read.csv("d2.csv") # データよみこみ
> (ct2 <- xtabs(y ~ x + Spc, data = d2))
   Spc
x     A   B
  0 286  85
  1 378 148
> plot(ct2, col = c("orange", "blue"))
```



ポアソン分布の GLM (分割方式) — A

ふつーに考えるとロジスティック回帰 (二項分布)
を使って解析したいデータだけど……
ここであえてポアソン回帰

$$y_{A,x} \sim \mathrm{Pois}(\lambda_{A,x})$$
$$\log(\lambda_{A,x}) = \alpha_A + \beta_A x$$

```
> # SpcA だけ
> summary(glm(y ~ x, data = d2[d2$Spc == "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) — B

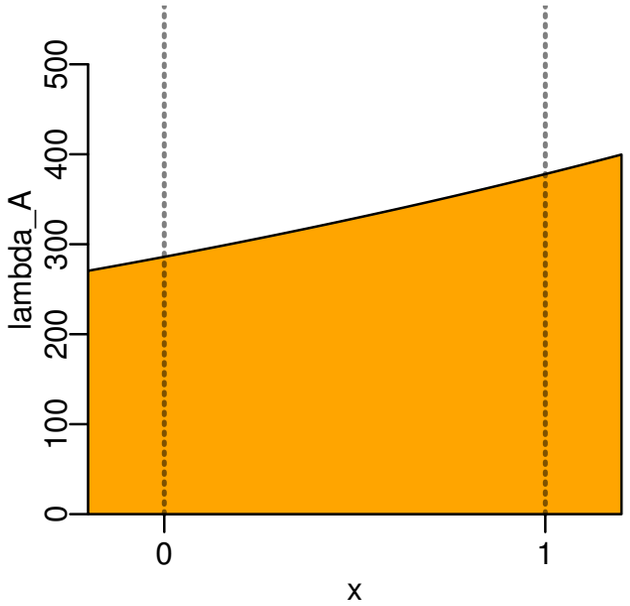
$$
\begin{aligned}
y_{\mathrm{B},x} &\sim \mathrm{Pois}(\lambda_{\mathrm{B},x}) \\
\log(\lambda_{\mathrm{B},x}) &= \alpha_{\mathrm{B}} + \beta_{\mathrm{B}} x
\end{aligned}
$$

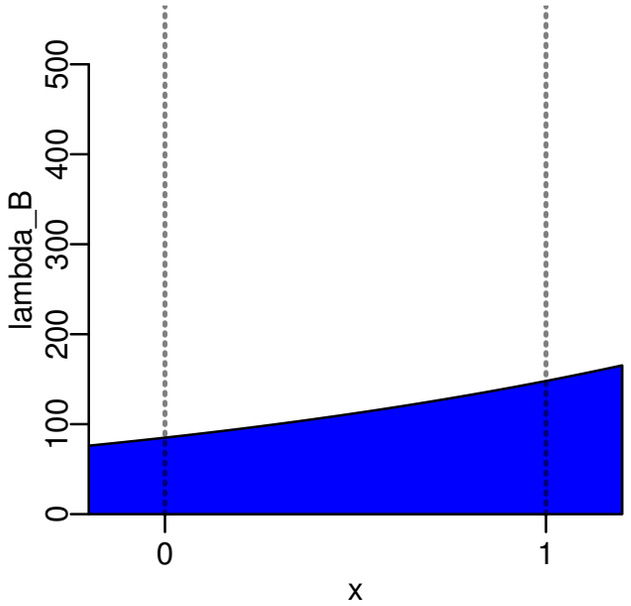
```
> # SpcB だけ
> summary(glm(y ~ x, data = d2[d2$Spc == "B",], family = poisson)
(... 略...)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.443      0.108   40.96  < 2e-16
x               0.555      0.136    4.07  4.6e-05
```

こんなことをやって何か意味があるのか?

ポアソン回帰の推定にもとづく予測

$\log(\lambda_{A,x}) = 5.66 + 0.279x$ $\log(\lambda_{B,x}) = 4.44 + 0.555x$





モデル	AIC	モデル	AIC
$\lambda_{A,x} = \alpha_A + \beta_A x$	19.3	$\lambda_{B,x} = \alpha_B + \beta_B x$	17.1
$\lambda_{A,x} = \alpha_A$	30.1	$\lambda_{B,x} = \alpha_B$	32.4

ロジスティック回帰との対応関係を調べてみよう!

ポアソン分布 GLM・二項分布 GLM のつながり

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

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

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

$$\lambda_{\mathsf{A},x} = \exp(\alpha_{\mathsf{A}} + \beta_{\mathsf{A}}x)$$

$$\lambda_{\mathsf{B},x} = \exp(\alpha_{\mathsf{B}} + \beta_{\mathsf{B}}x)$$

……「A の割合」は?

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

$$= \frac{1}{1+\exp[\alpha_{\mathsf{B}} - \alpha_{\mathsf{A}} + (\beta_{\mathsf{B}} - \beta_{\mathsf{A}})x]}$$

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}	$=$	$\beta_{\mathrm{A}} - \beta_{\mathrm{B}}$

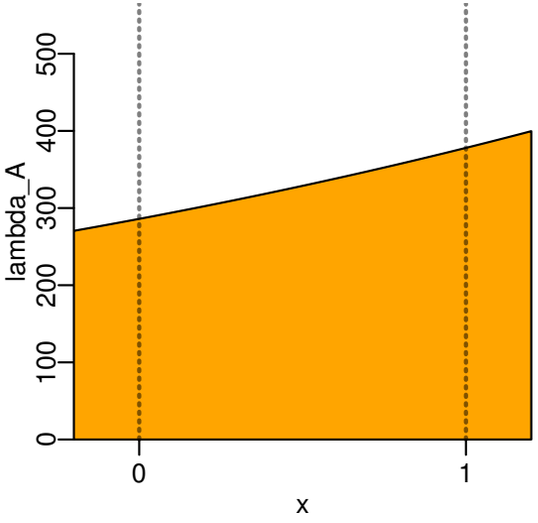
比較: 二項分布とポアソン分布の GLM

二項分布 GLM		ポアソン分布 GLM
$a_A = 1.213$	=	$\alpha_A - \alpha_B$
$b_A = -0.276$	=	$\beta_A - \beta_B$

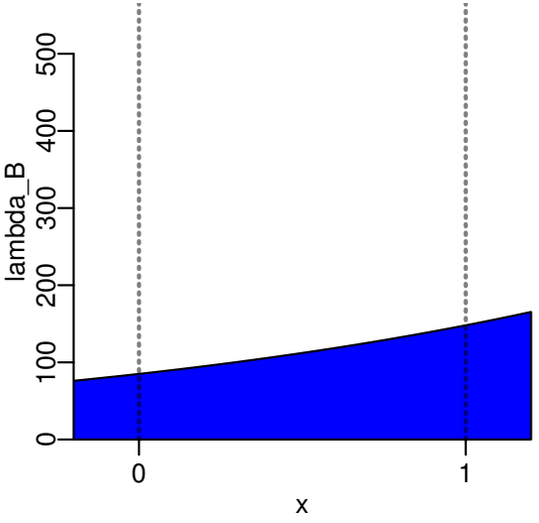
```
> 二項分布 GLM (A 種の比率)
> glm(ct2 ~ c(0, 1), data = d2, family = binomial)
(Intercept)       c(0, 1)
      1.213        -0.276
> ポアソン分布 GLM (A 種の比率)
> glm(y ~ x, data = d2[d2$Spc == "A",], family = poisson)
(Intercept)             x
      5.656         0.279
> ポアソン分布 GLM (B 種の比率)
> glm(y ~ x, data = d2[d2$Spc == "B",], family = poisson)
(Intercept)             x
      4.443         0.555
```

図解: ポアソン分布 GLM・二項分布 GLM のつながり

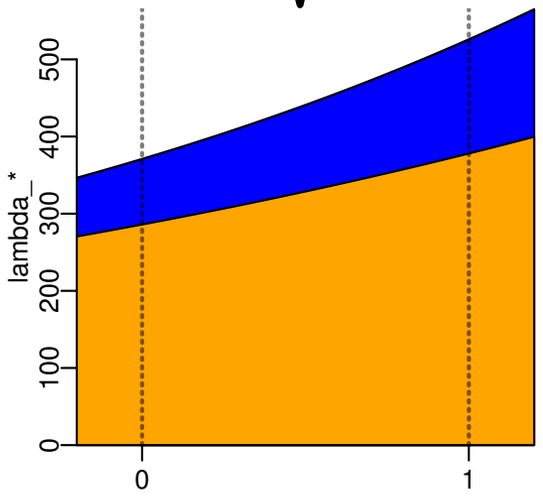
ポアソン分布の GLM (A)



ポアソン分布の GLM (B)

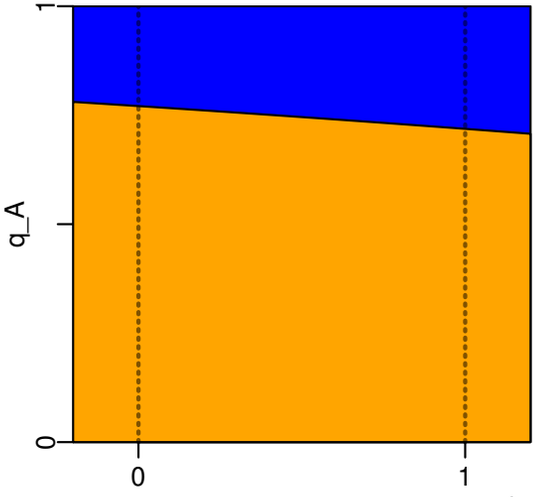


つみあげる



たいらに
押しつぶす

二項分布の GLM
(A + B)



2×3 分割表の統計モデル

これもポアソン回帰の `glm()` で

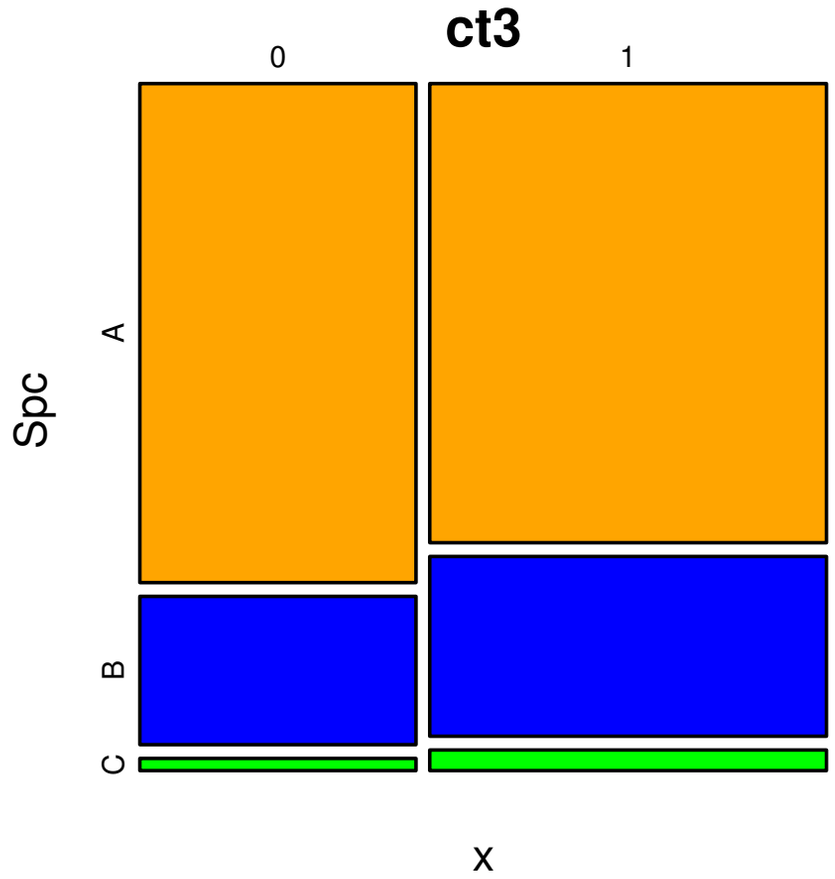
多項分布 GLM はめんどうだから

R には多項分布ロジスティック回帰のための

package があるけれど，あえてそれは使わない

次の架空例題データ: 2 種から 3 種にふやす

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



ポアソン分布の GLM (一括方式)

```
> glm(y ~ x * Spc, data = d3, family = poisson)
(... 略...)
Coefficients:
(Intercept)           x        SpcB        SpcC     x:SpcB     x:SpcC
      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……

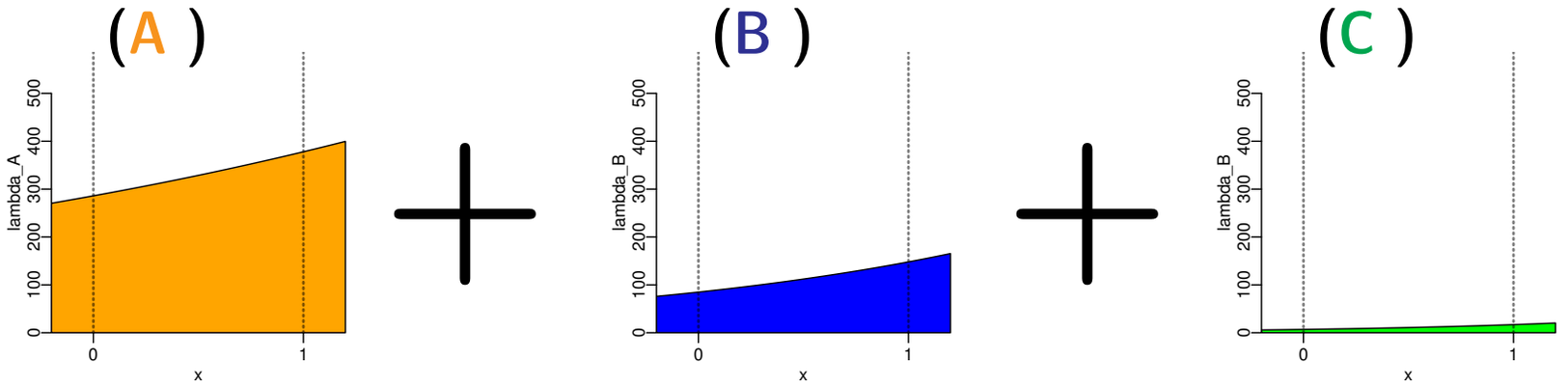
$$
\begin{aligned}
y_{\mathrm{A},x} &\sim \mathrm{Pois}(\lambda_{\mathrm{A},x}) \\
\log(\lambda_{\mathrm{A},x}) &= \alpha_{\mathrm{A}} + \beta_{\mathrm{A}} x
\end{aligned}
$$

$\alpha_{\mathrm{A}} = 5.66$ $\quad$ $\alpha_{\mathrm{B}} = 5.66 - 1.21$ $\quad$ $\alpha_{\mathrm{C}} = 5.66 - 3.71$

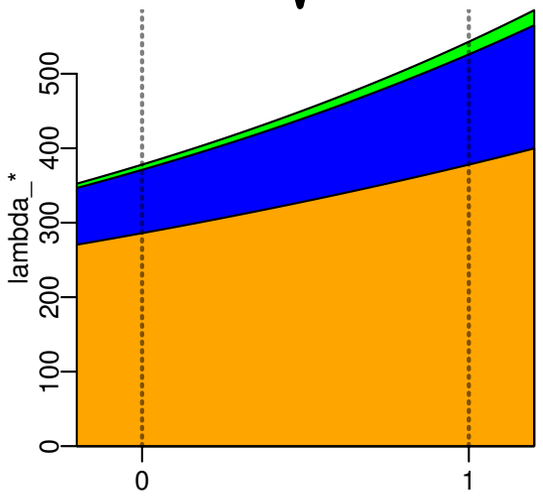
$\beta_{\mathrm{A}} = 0.279$ $\quad$ $\beta_{\mathrm{B}} = 0.279 + 0.276$ $\quad$ $\beta_{\mathrm{C}} = 0.279 + 0.608$

ポアソン分布 GLM の予測など

ポアソン分布の GLM

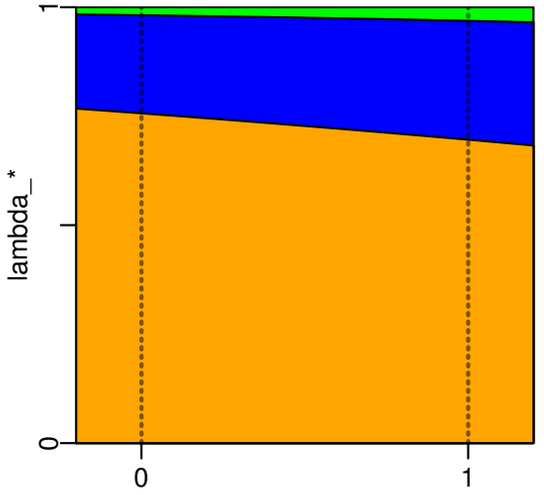


つみあげる



たいらに
押しつぶす

三項分布の GLM



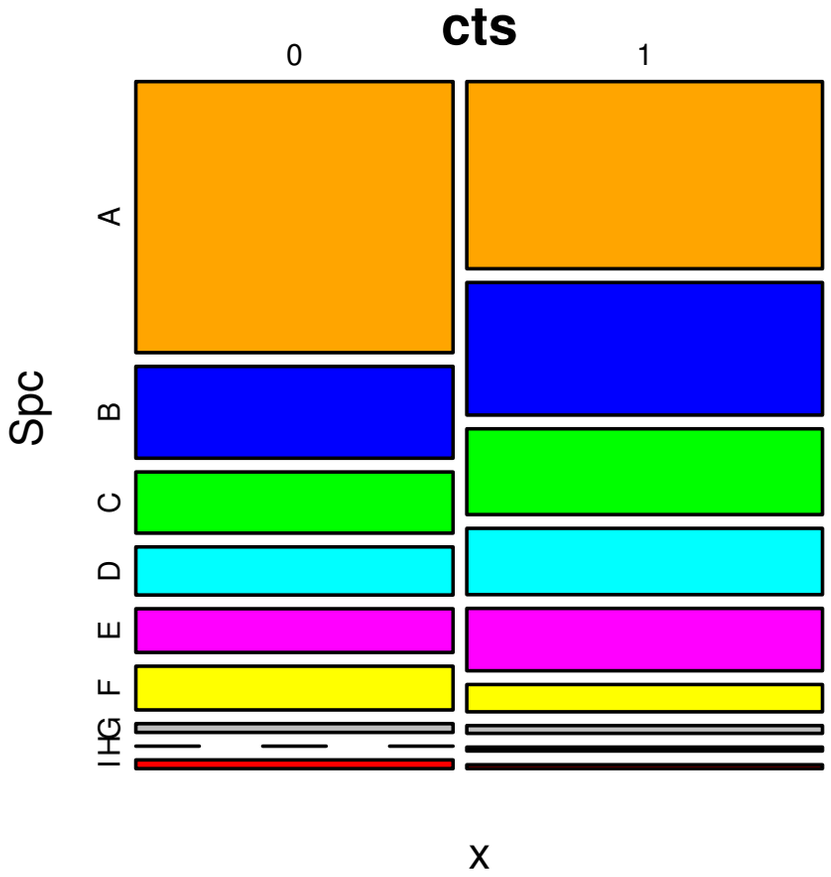
2×9 分割表の統計モデル

…… そろそろ単なる **GLM** ではしんどい

ということで **GLMM** や階層ベイズモデルを!

また別のデータ: 種数が 3 から 9 に増えた!

```
    Spc
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(ct9, col = c(ごちゃごちゃと指定))
```



ポアソン分布の GLM (一括方式)

```
> ct9
   Spc
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 * Spc, data = d9, family = poisson))

            (Intercept)                   x                SpcB                SpcC
                  4.127              -0.256              -1.083              -1.488
                   SpcD                SpcE                SpcF                SpcG
                 -1.729              -1.825              -1.825              -3.434
                   SpcH                SpcI             x:SpcB              x:SpcC
                -26.430              -3.434               0.738               0.708
                x:SpcD              x:SpcE              x:SpcF              x:SpcG
                 0.691               0.726              -0.101               0.256
                x:SpcH              x:SpcI
                22.559              -0.437
```

H 種 の推定値がかなりヘン!

「なんでも glm()」方針の問題点と改良案

- 分割表がでかくなったときに，独立に推定されるパラメーター数がどんどん増えてしまう
- そのような場合，とくにゼロデータなどがあると，パラメーター推定が難しくなる
- そこで **GLMM** を使ってみよう!
  - え? なんで?! — Spc の推定に制約をいれたい
  - 制約とは?: 9 種には「似ている」ところがある

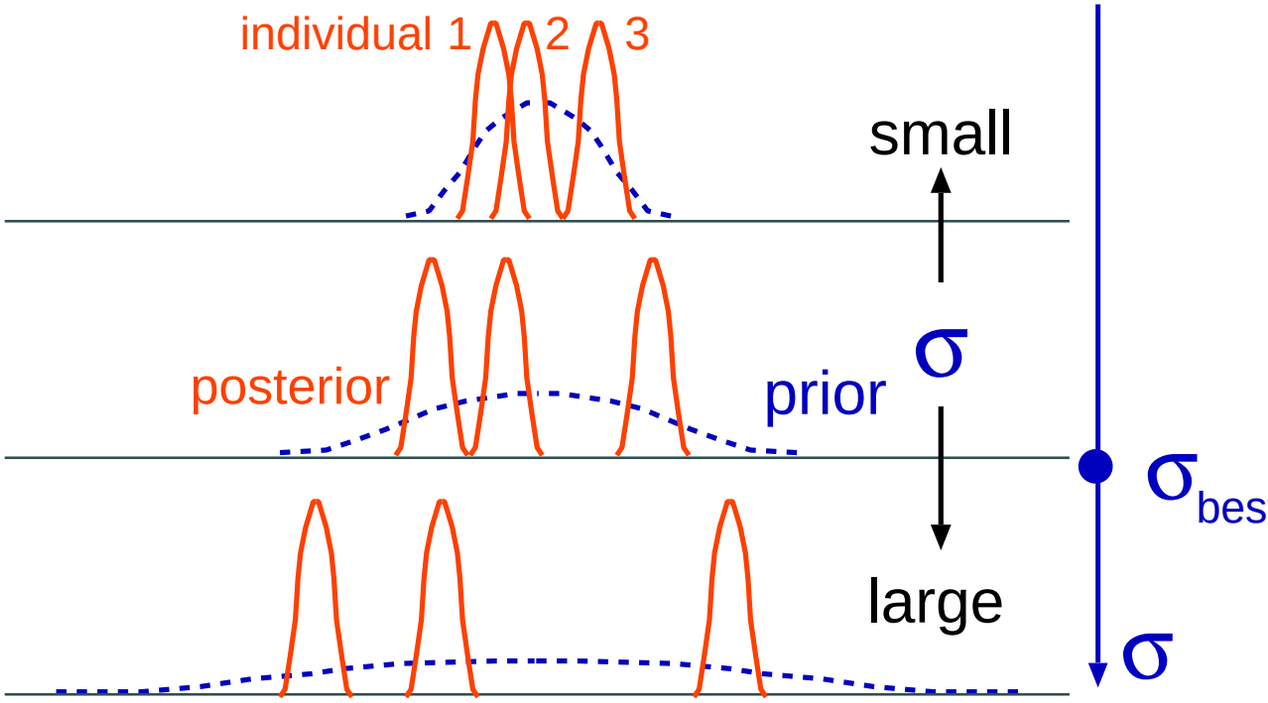
    → Spc の値は (対数スケイルの世界で) 正規分布にしたがう

ポアソン分布の GLMM で「似ている種差」を推定

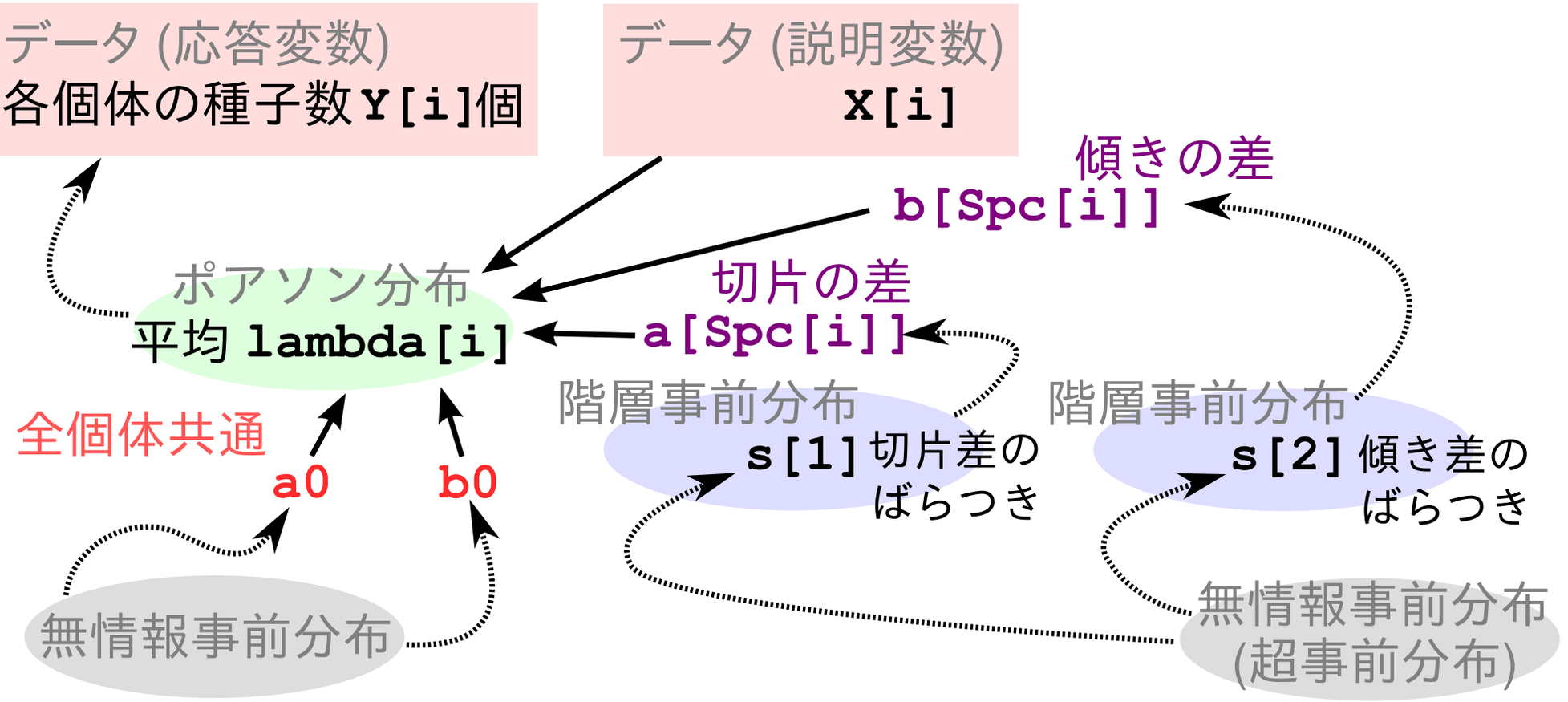
```
> (fit.glmm <- glmmML(y ~ x, data = d9, cluster = Spc, family = poisson))
```

(略)

```
> 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 (線形ポアソン回帰)



- このようにデータを設計する
- WinBUGS を使ってパラメーター推定 (MCMC 法) をしたいので，BUGS コードで書く
- WinBUGS を R の下っぱとして使う

それでは連続値データの場合は どういう確率分布を……?

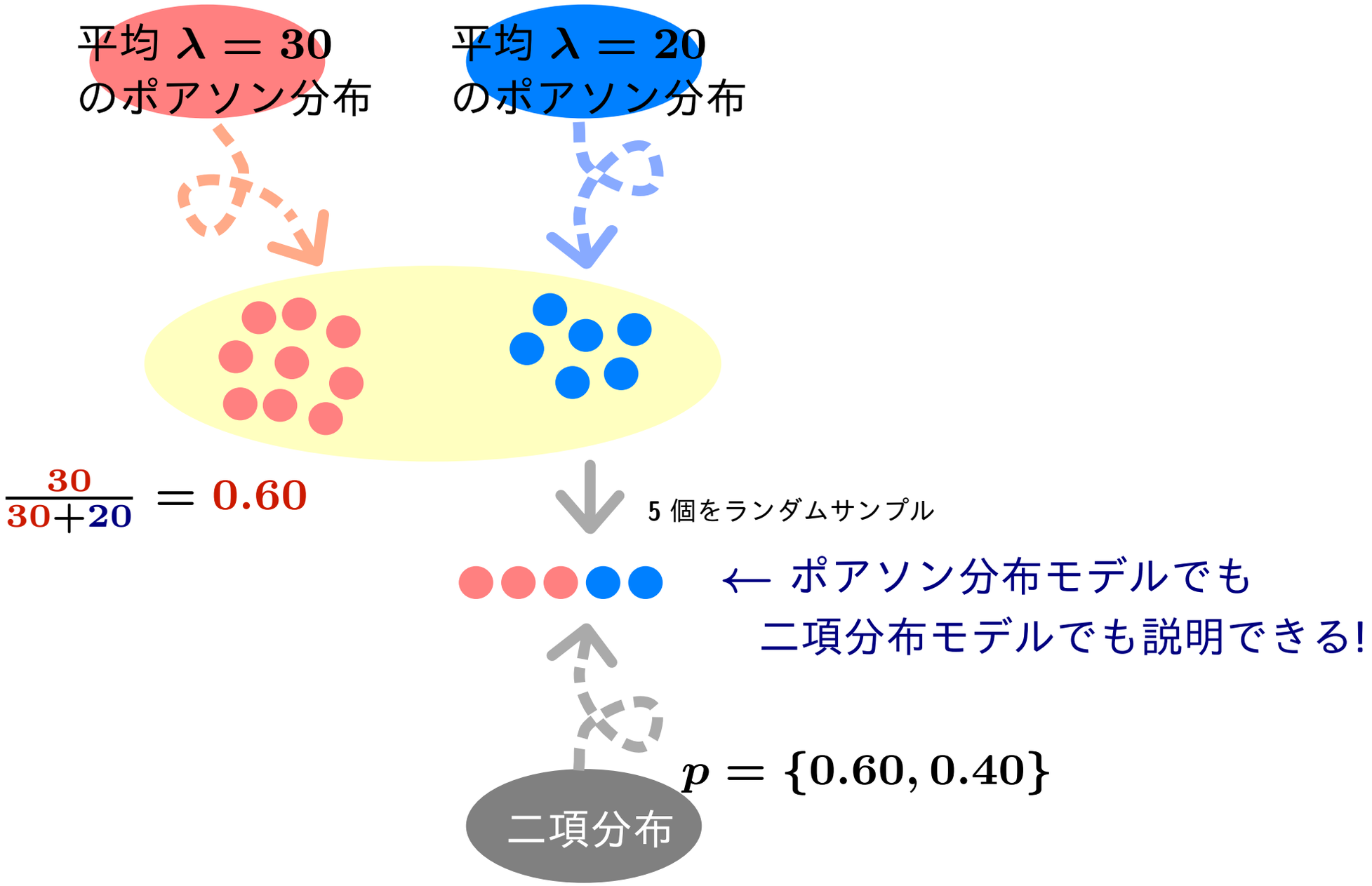
とりあえず「正の連続値」をあつかう

ガンマ分布が使えそう?

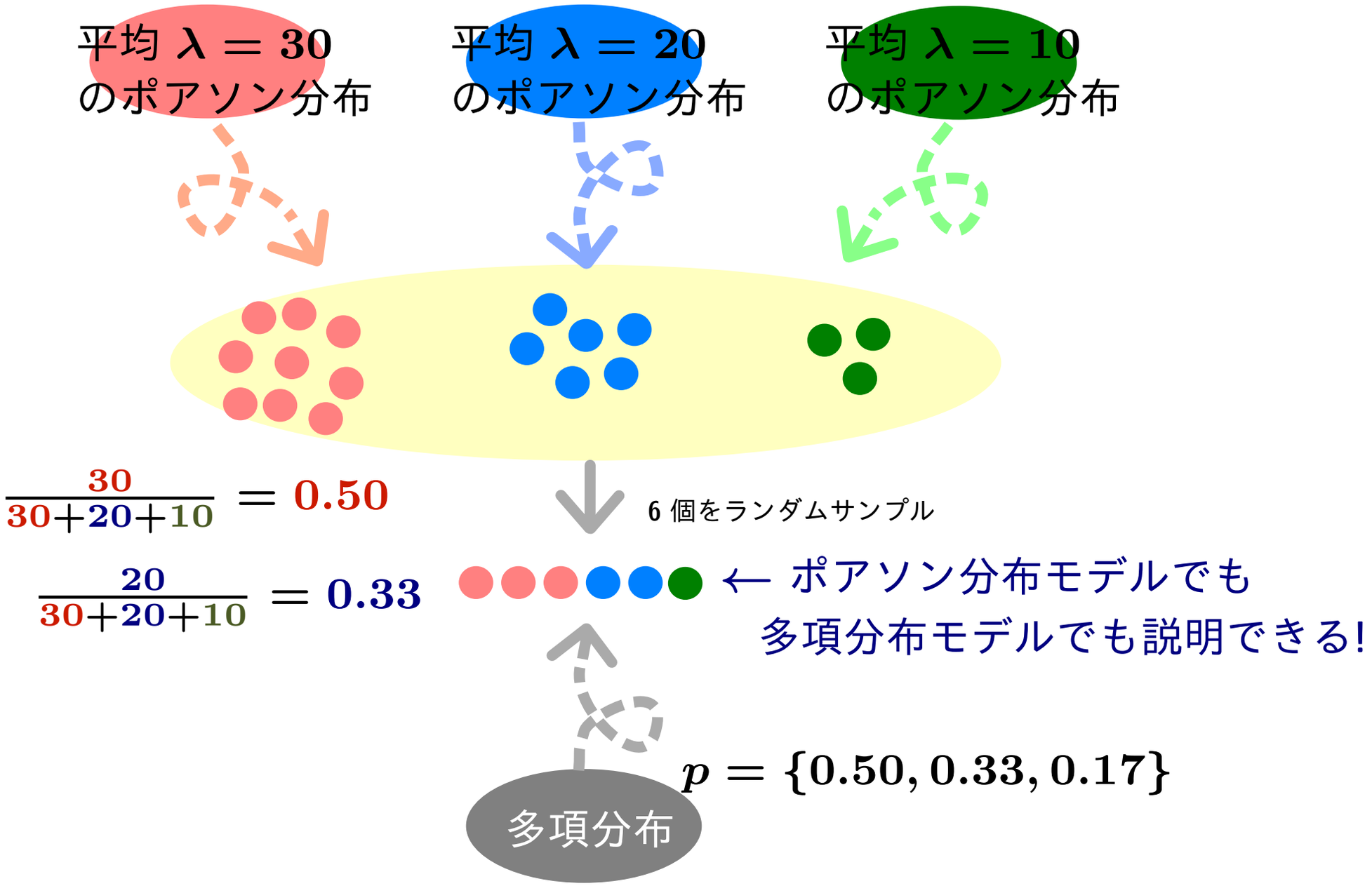
ベータ分布やディリクレ分布といった

「比率」の確率分布との対応を考える

もいちど説明: ポアソン分布と二項分布



もいちど説明: ポアソン分布と多項分布



ガンマ分布 (Gamma distribution) とは何か?

- 連続値の確率分布 $y \geq 0$
- 確率密度関数 (paramter: r, s)

$$p(y \mid s, r) = \frac{r^s}{\Gamma(s)} y^{s-1} \exp(-ry)$$

- 期待値 s/r, 分散 s/r^2

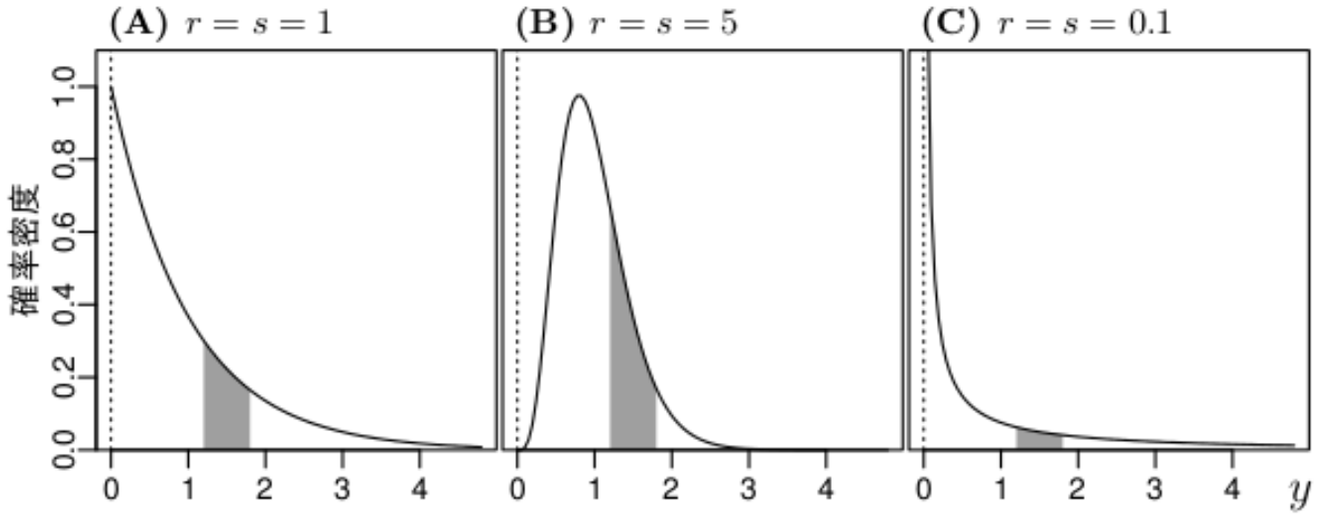
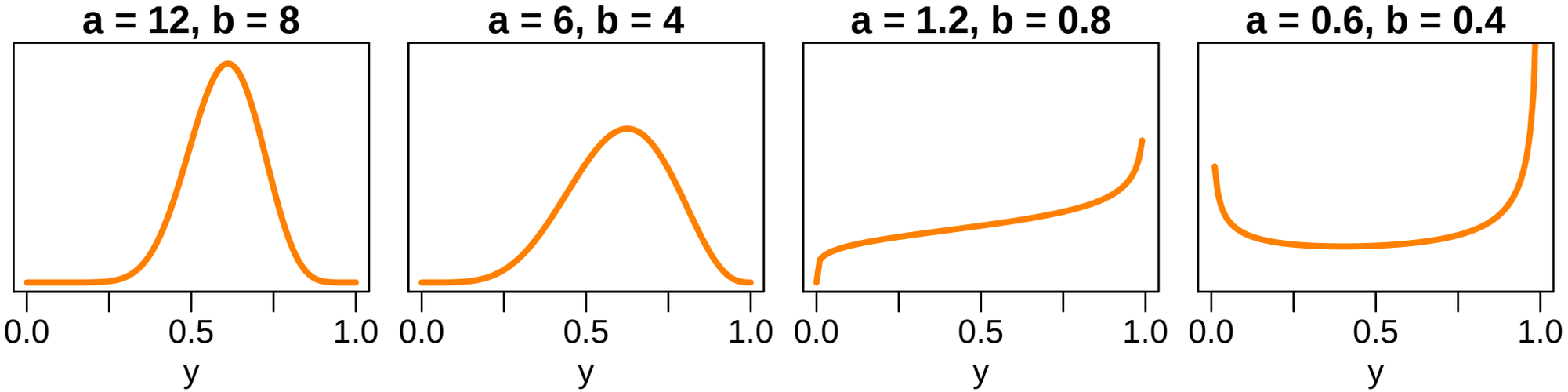


図 6.12 ガンマ分布の確率密度関数．横軸は確率変数 y，縦軸は確率密度．グレイの領域の面積は $1.2 \leqq y \leqq 1.8$ となる確率をあらわす．

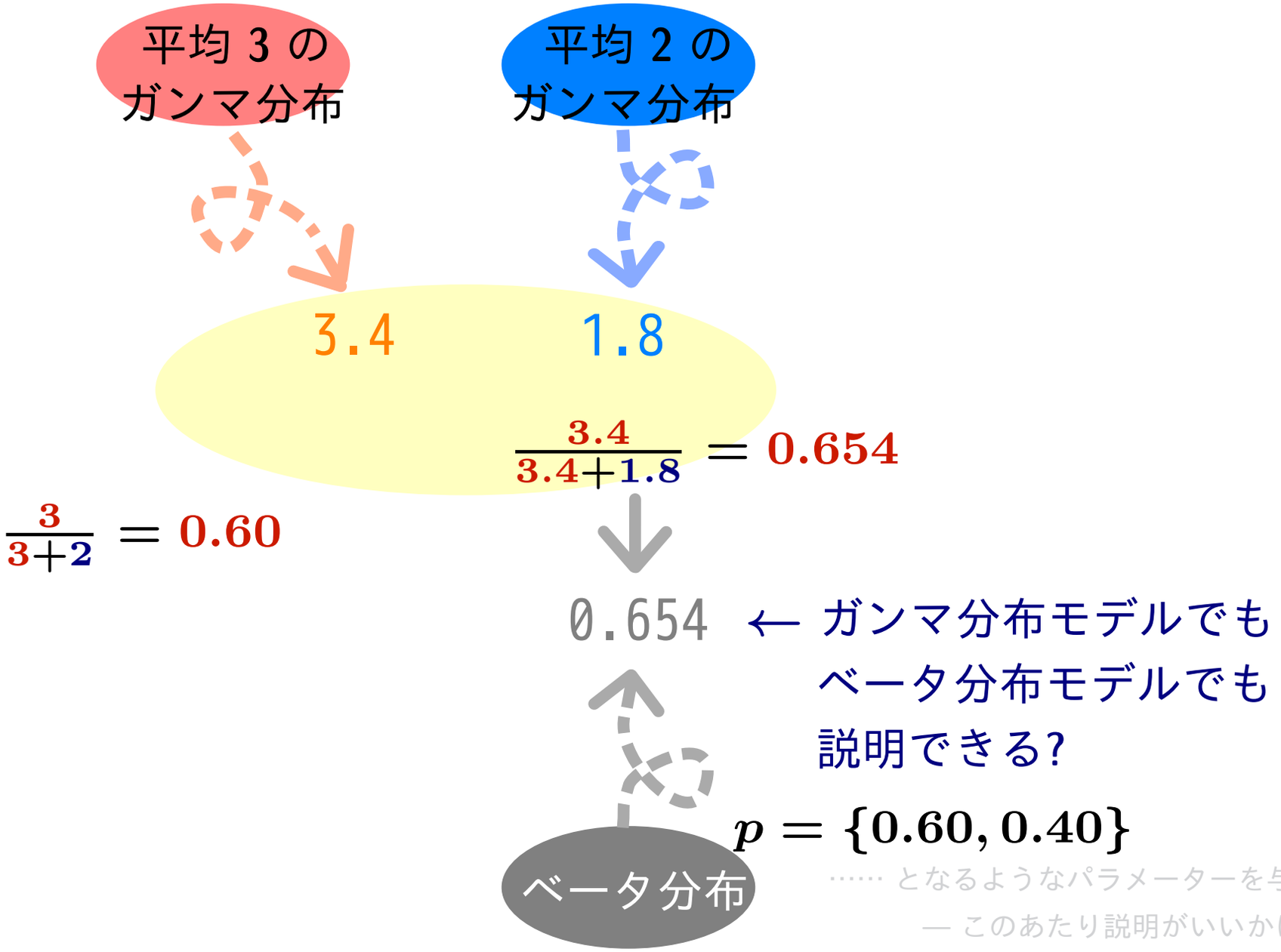
「比率」の確率分布: ベータ分布とディリクレ分布

- どちらも連続値の確率分布
- ただし「比率」の確率分布なので「足して 1 になる正の連続値」を乱数として生成する
- ベータ分布: $(0.6, 0.4)$ とか $(0.325, 0.675)$ など
- ディリクレ分布: $(0.5, 0.3, 0.1)$ とか $(0.28, 0.54, 0.18)$ など

期待値が 0.6 となるベータ分布の例



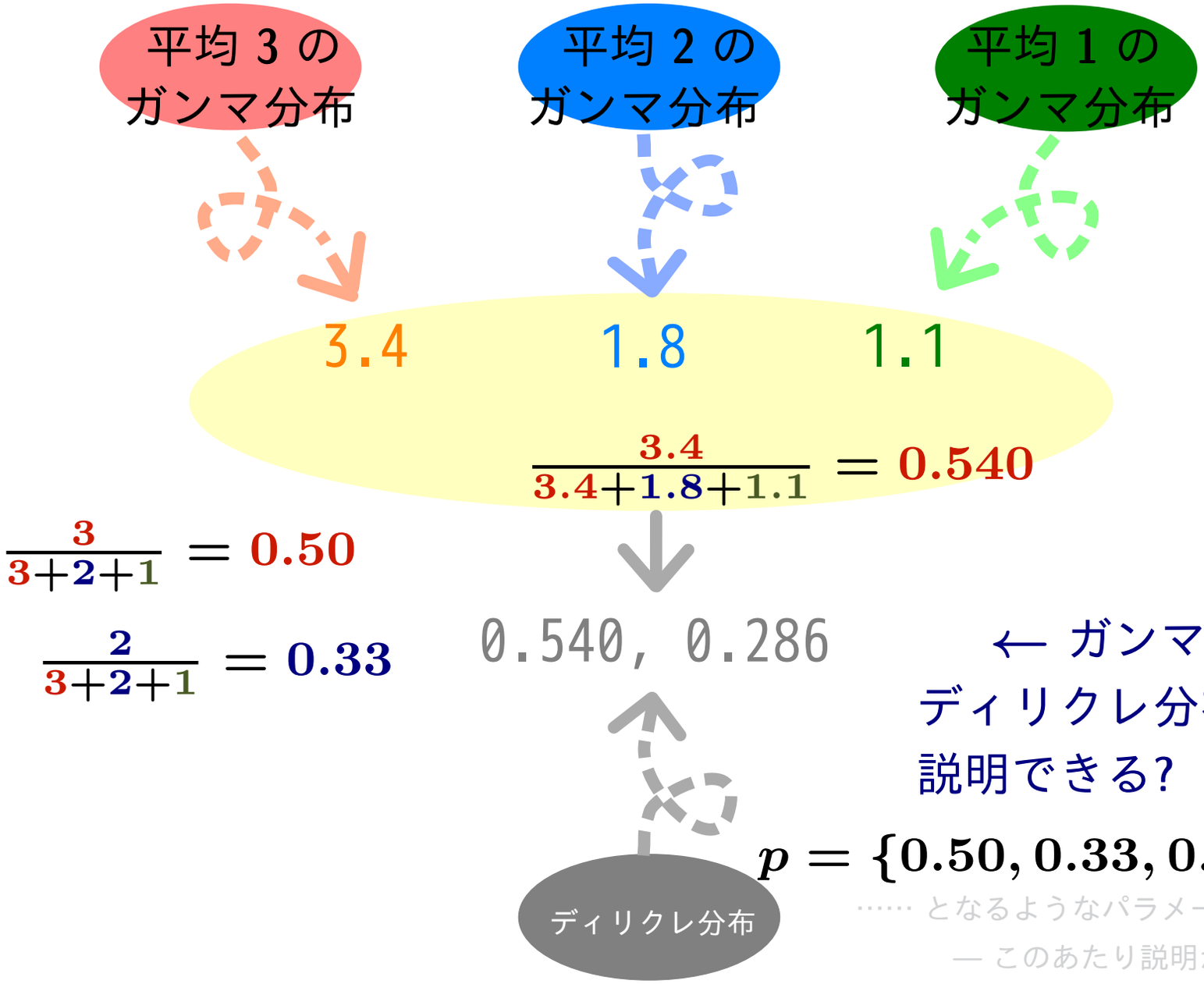
連続値ならこれで OK? ガンマ分布とベータ分布



…… となるようなパラメーターを与える

— このあたり説明がいいかげんです

連続値ならこれで OK? ガンマ分布とディリクレ分布



← ガンマ分布モデルでも
ディリクレ分布モデルでも
説明できる?

…… となるようなパラメーターを与える
— このあたり説明がいいかげんです

ガンマ分布だけでなんとか……なりそう??

- ベータ分布はディレクレ分布の特殊なカタチであり，さらに……

Related distributions [edit]

If, for $i \in \{1, 2, \ldots, K\}$,

$$Y_i \sim \text{Gamma}(\text{shape} = \alpha_i, \text{scale} = \theta) \text{ independently},$$

then [8]

$$V = \sum_{i=1}^{K} Y_i \sim \text{Gamma}(\text{shape} = \sum_{i=1}^{K} \alpha_i, \text{scale} = \theta),$$

and

$$X = (X_1, \ldots, X_K) = (Y_1/V, \ldots, Y_K/V) \sim \text{Dir}(\alpha_1, \ldots, \alpha_K).$$

Although the X_is are not independent from one another, they can be seen to be generated from a set of K independent gamma random variables (see [9] for proof). Unfortunately, since the sum V is lost in forming X, it is not possible to recover the original gamma random variables from these values alone. Nevertheless, because independent random variables are simpler to work with, this reparametrization can still be useful for proofs about properties of the Dirichlet distribution.

ガンマ分布で OK というハナシ `http://en.wikipedia.org/wiki/Dirichlet_distribution`

- 問題点: R の `glm(y ~ x, family = Gamma)` は 使えない

  (理由): $\theta = 1/r =$ 定数，という制約を満たしていないから

R の library(DirichletReg) という package がある

ディリクレ分布を使って GLM っぽいことができる

```
> ALake <- ArcticLake
> ALake$Y <- DR_data(ALake[,1:3])
> DirichReg(Y ~ depth + I(depth^2), ALake)
--------------------------------------------
Coefficients for variable no. 1: sand
(Intercept)         depth    I(depth^2)
   1.436197    -0.007238     0.000132
--------------------------------------------
Coefficients for variable no. 2: silt
(Intercept)         depth    I(depth^2)
  -0.025970     0.071745     -0.000268
--------------------------------------------
Coefficients for variable no. 3: clay
(Intercept)         depth    I(depth^2)
  -1.793149     0.110791     -0.000487
--------------------------------------------
```

ただし，この library(DirichletReg) も万能ではなく……

多項分布ロジスティック回帰と同じ問題が発生するでしょう

- 応答変数の項目数が 3 とかではなく 9 とか 100 になった場合
- 個体差・場所差といった random effects を考慮しなければならない場合
- ゼロデータがたくさんある場合

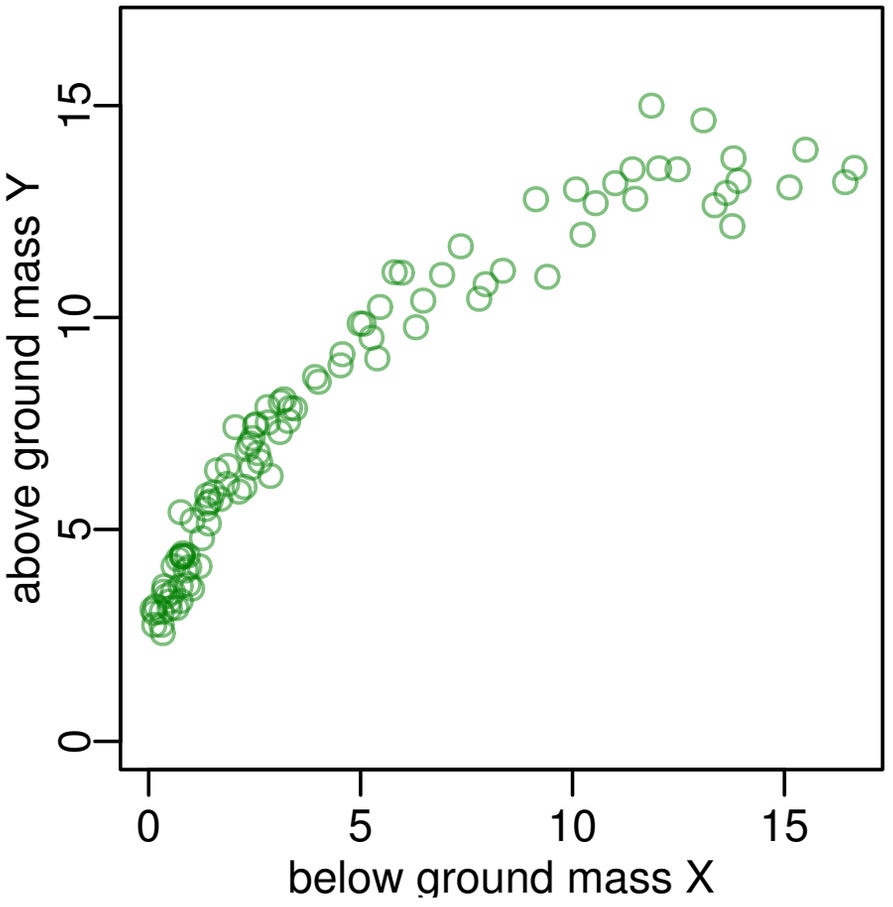
じつは「比率」の確率分布はいらない，かも……?

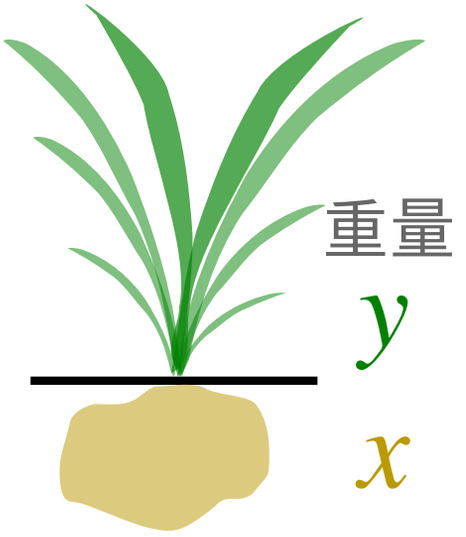
ガンマ分布を使わない統計モデリングを考える

「比率の世界」と「観測できる世界」をつなぐ

ここでもまた …… めんどくさくなったら階層ベイズモデルで!

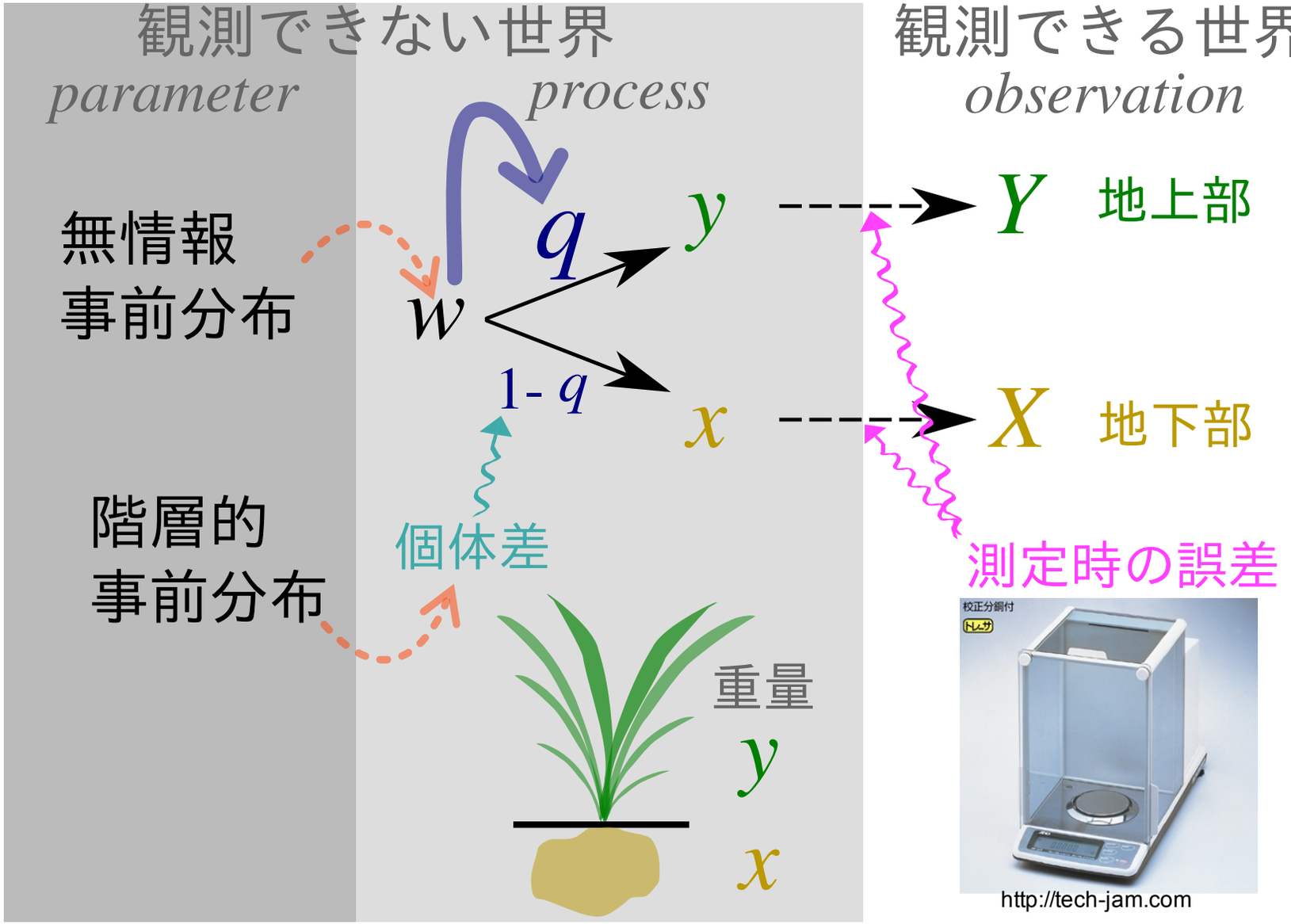
架空データ: 重量増大とともに分配が変化





- 小さいときには地上部重量を大きくする
- 総重量が大きくなってくれば地下部を大きくする

重量分割モデルの改造: q を w 依存にするだけ



重量分配モデルを BUGS code で (process の部分のみ)

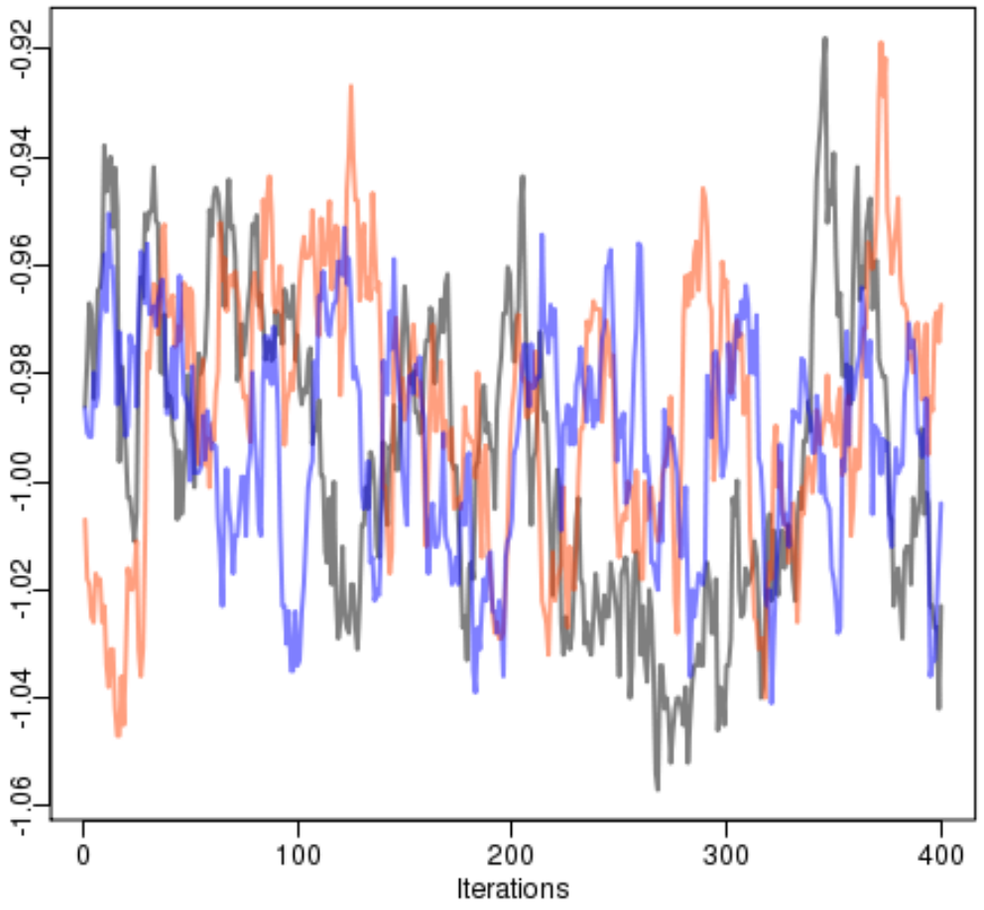
```
for (i in 1:N) {
  Y[i] ~ dnorm(y[i], Tau.err) # 地上部の重量
  X[i] ~ dnorm(x[i], Tau.err) # 地下部の重量
  y[i] <- q[i] * w[i]
  x[i] <- (1 - q[i]) * w[i]
  # 下のように説明変数の効果と個体差をくみあわせる
  logit(q[i]) <- a + b * log.w[i] + re[i]
  w[i] <- exp(log.w[i])
  # log.w[i] は地上部 + 地下部の重量
  log.w[i] ~ dnorm(0, Tau.noninformative) # !!
  (以下，省略，また実際のコードでは中央化その他が必要)
  ......
```

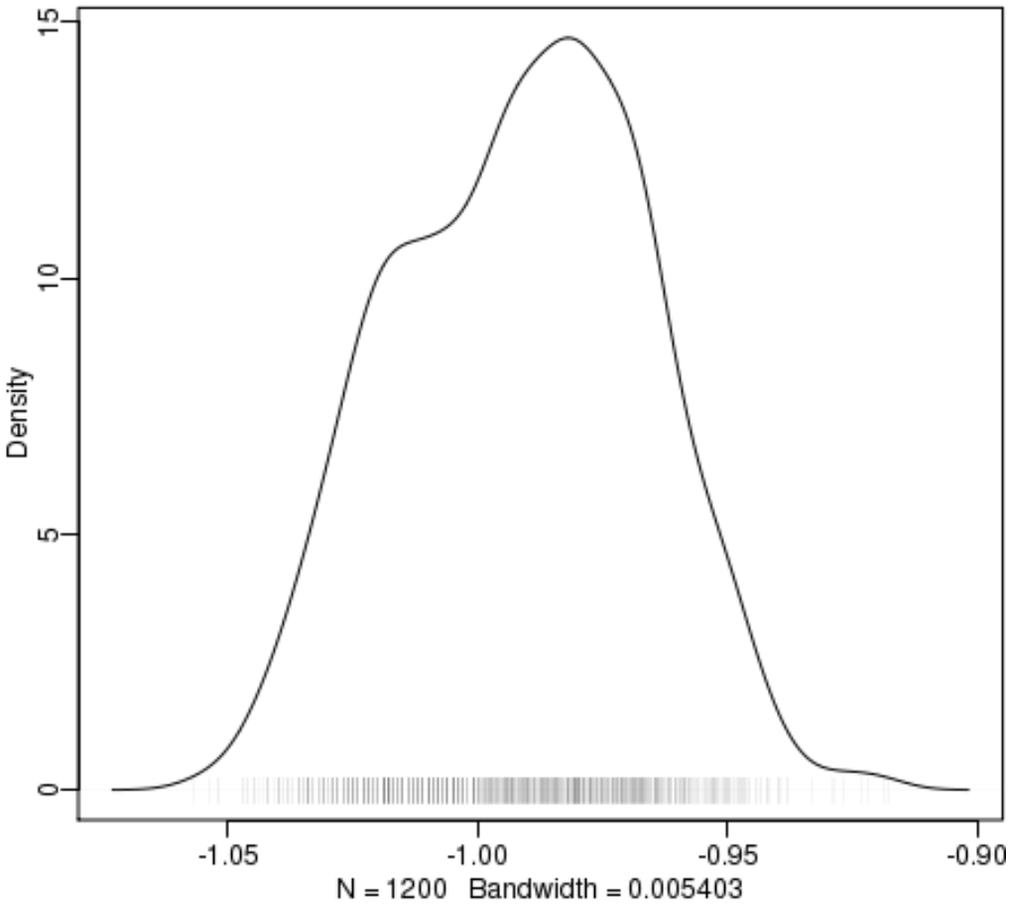
階層ベイズモデルのパラメーター推定: MCMC

1. BUGS code で重量分割モデルを記述する (`model1.txt`)
2. これにデータを渡したりする R スクリプトを書く (`runbus1.R`)
3. R で `runbus1.R` を実行 (`source("runbugs1.R")`)
4. R 内から `library(R2WinBUGS)` によって WinBUGS が起動
5. WinBUGS 内で Markov chain Monte Carlo (MCMC) サンプリング
6. 事後分布からのサンプリング結果が R に渡される

必要なファイルは自由集会サイトからダウンロードできます

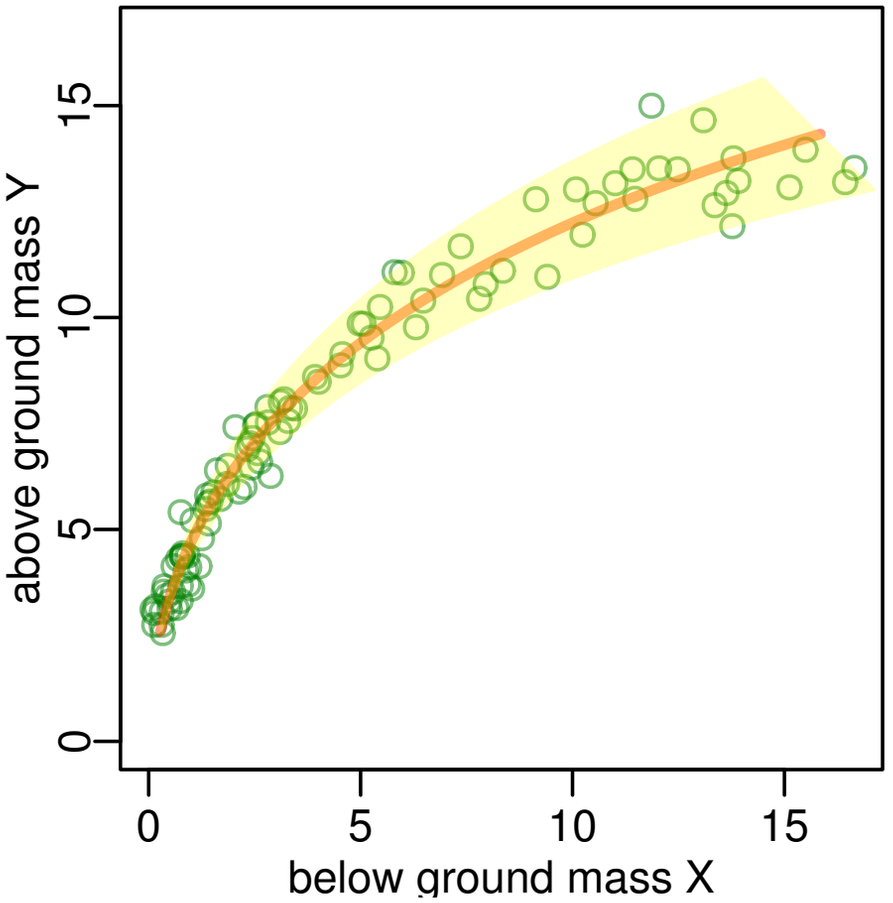
推定結果: 総重量増大 → 地上部への分配減少

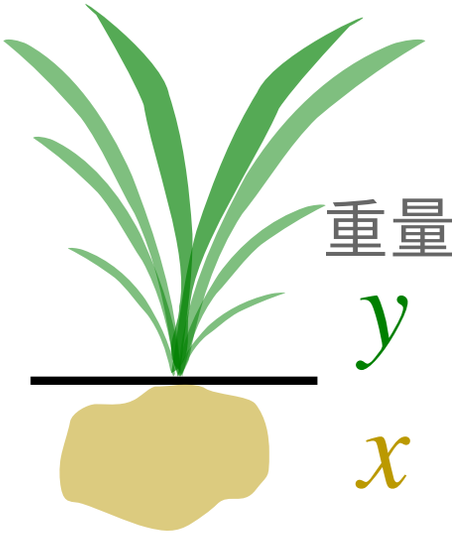




- 総重量 w 依存のパラメーター b はマイナス
- こういう問題は MCMC 収束が遅い

このモデルで複雑な重量分配を表現できる





- オレンジ色の線は中央値 (median)
- 黄色の領域は個体差による予測のばらつき (95% CI)

さてさて

このあたりで本日は終了

今回とりあげなかった問題あれこれ ……

- 測定機器が吐き出す割算値 — **C-N** だの，同位体比だの
- 「見た目」被度 **(coverage)** — 植生なアレ
- 説明変数としての割算値 — **GIS** 「同心円」被度とか

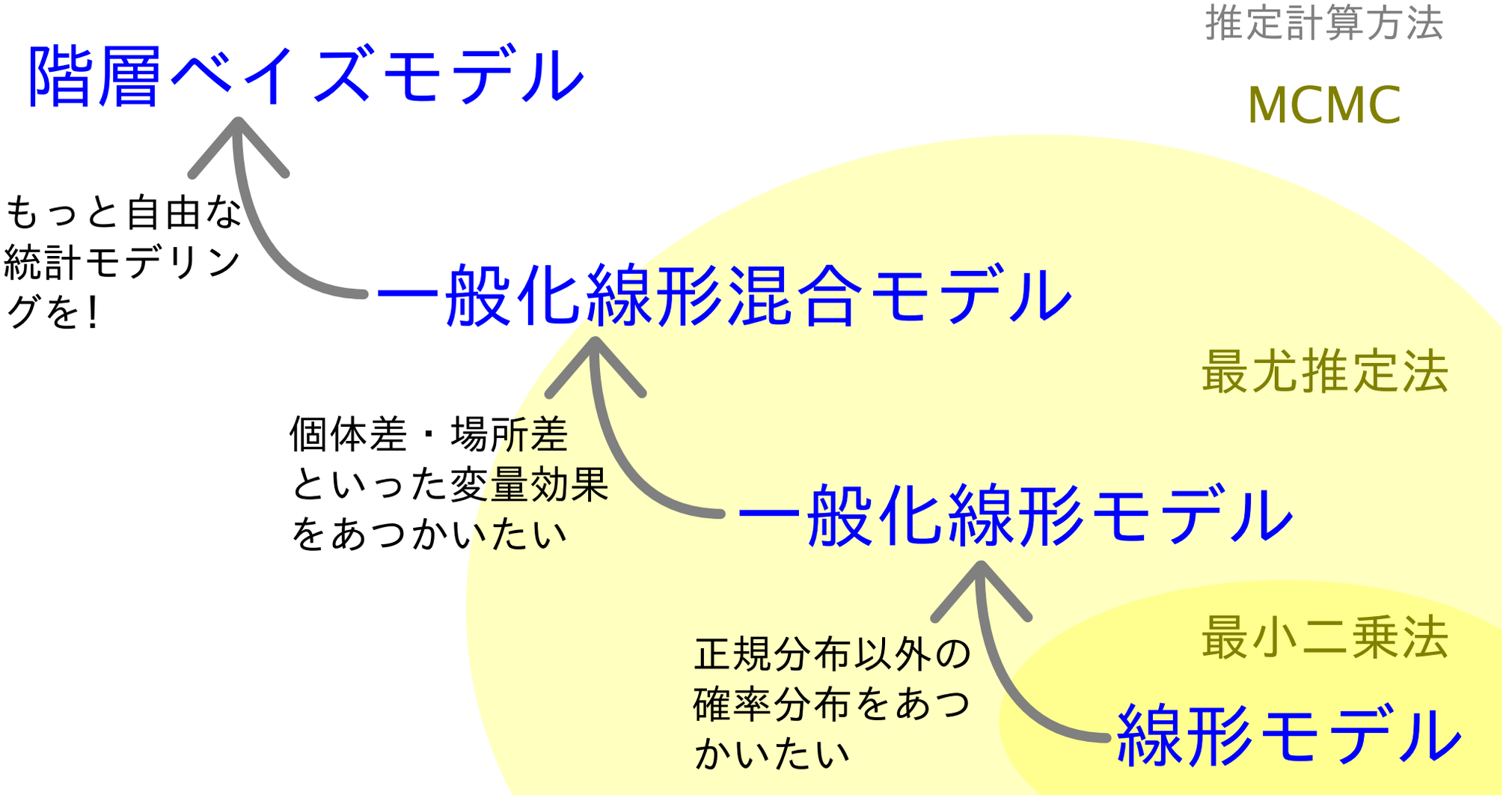
…… ほかにもいろいろあるでしょう ……

今日の久保のハナシ

- 初級篇: 「データわるデータ」作法の問題点と簡単な対処
  - ポアソン回帰の **offset** 項わざ，ロジスティック回帰
- 中級篇: 複数の「割合」をポアソン分布で統計モデル化
  - 二項分布・多項分布を使うよりも，ポアソン分布の階層ベイズモデルなんかが便利かもね
- ややめんどう篇: 連続値データの「比率」はどうする?
  - ベータ分布やディレクレ分布といった「比率」の確率分布だの，それをガンマ分布で代替するのはめんどうそう
  - 階層ベイズモデルの工夫で「簡単」になるのでは?

問題の難しさに応じて線形モデルを発展させる

線形モデルの発展



1. 「割算」やめて統計モデルで対処しよう

- 久保拓弥 (北海道大・地球環境) kubo@ees.hokudai.ac.jp

2. 連続的な量が分子にくる「割算」の場合

- 粕谷英一 (九州大・理)

この自由集会の web site の URL

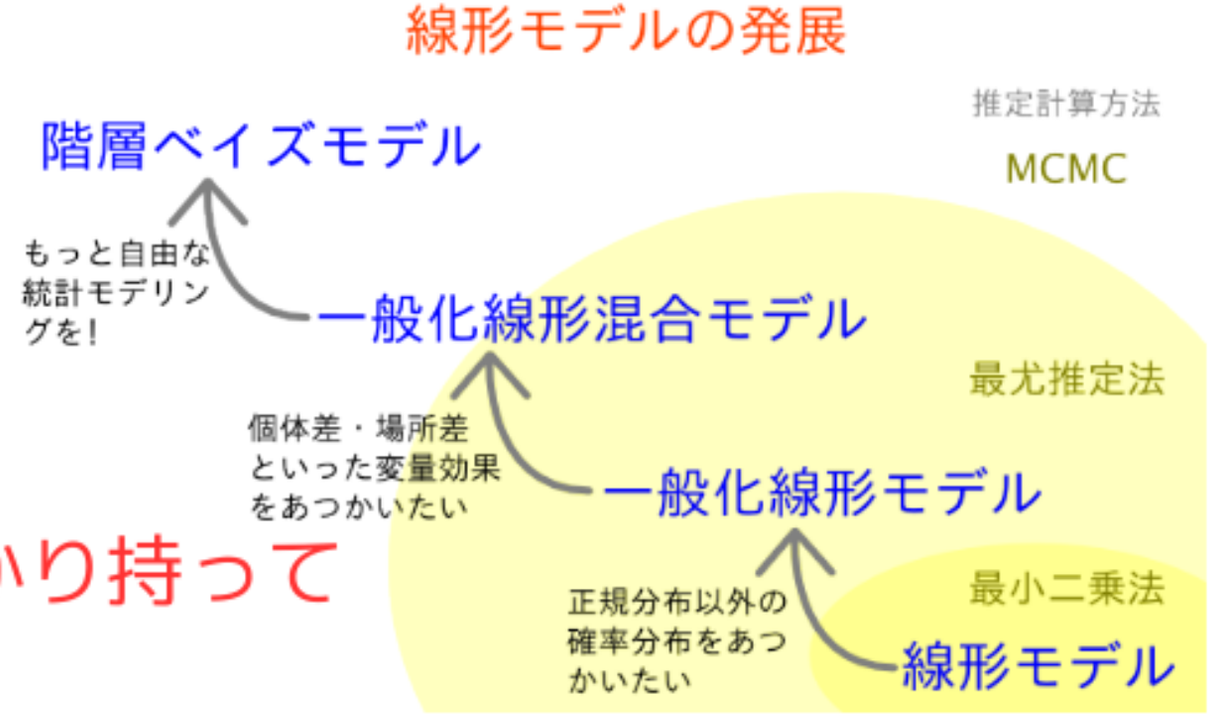
http://goo.gl/E1cjA

ファイルや資料などをダウンロードできます

メモ! メモ!

「統計モデリング入門」わりびき販売!

3990 円 → 3000 円



今回も 2 冊ばかり持ってきています

後日の郵送も可!