漁業統計検討会 (清水)

「統計モデリングセミナー」 (2012 年 12 月) 投影資料

全部で 7 回中の 6 回目

階層ベイズモデルの応用

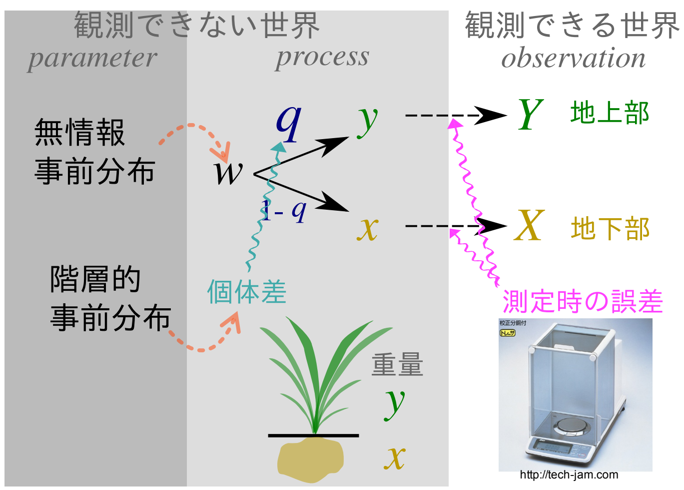
資源分配など割合をあつかう

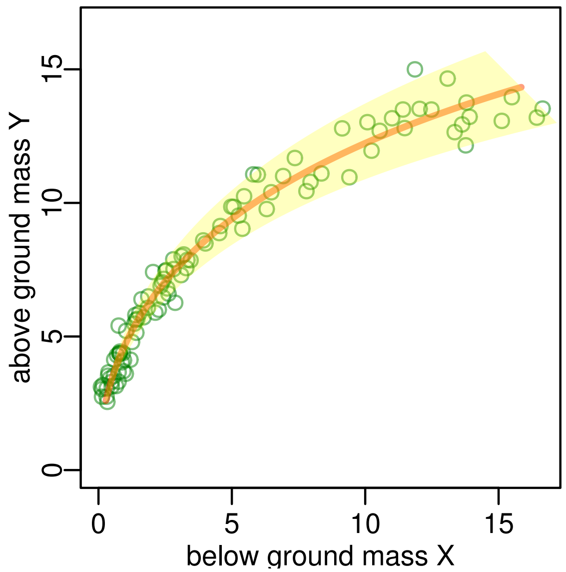
久保拓弥 kubo@ees.hokudai.ac.jp

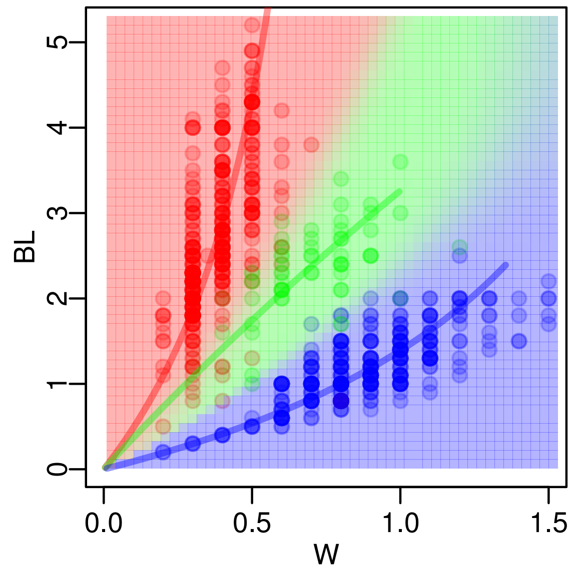
http://goo.gl/OyB2k

今日のハナシ: 連続値の割算値をやっつける

- 例題 1. 植物の地上部・地下部の重量比 (基本篇)
- 例題 2. 植物の地上部・地下部の重量比 (発展篇)
- 例題 3. 植物の葉のタテヨコ比 (カタチ篇)
- 例題 4. 植物の葉のタテヨコ比 (種識別篇)





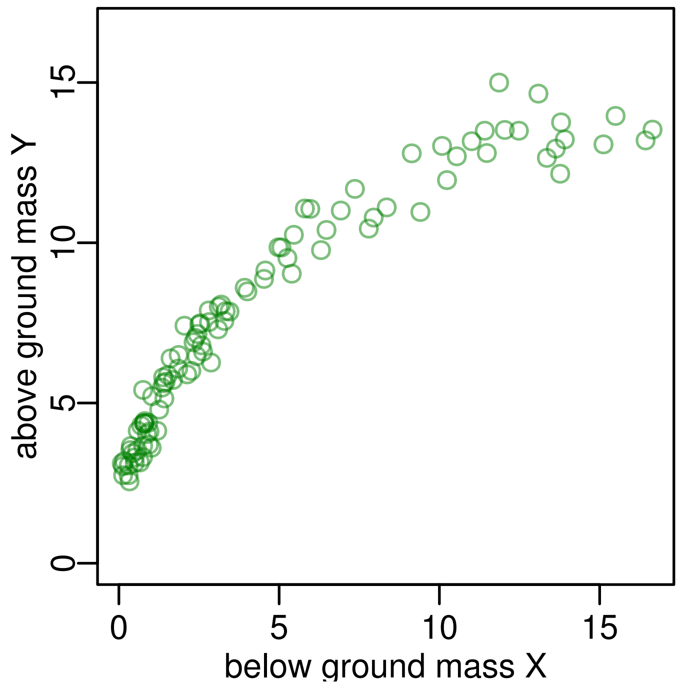


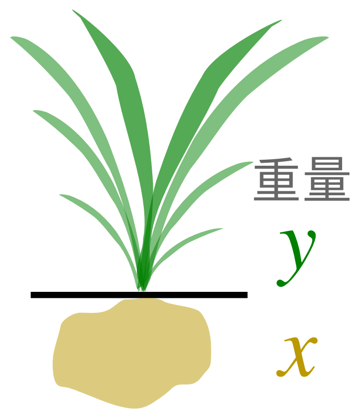
前半: 植物の地上部・地下部の重量比のハナシ

1. まずは $\{X, Y\}$ 誤差ありデータ解析における，ありがちな回帰適用お作法の問題点について検討
2. 解決策のひとつになりうる重量分割モデルの紹介
   - とても簡単な架空データ例題を使って
3. 重量分割モデルの拡張方法を検討
   - もう少し現実的な架空データ例題を使って
   - モデルの発展や他の問題への応用を考える

たとえばこういうデータがあったとしましょう

架空植物の地下部 (X) と地上部 (Y) の重量





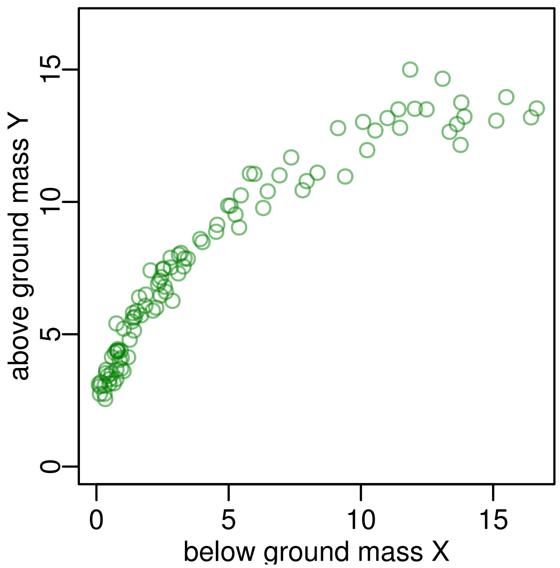
X と Y の関係を調べたい

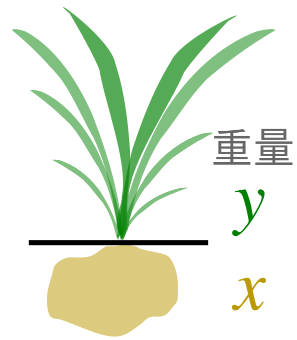
生態学における $\{X, Y\}$ 誤差ありデータ解析の典型

(今日はずっとこのたぐいの架空植物例題ばかり)

「あろめとらー」たちのお作法!

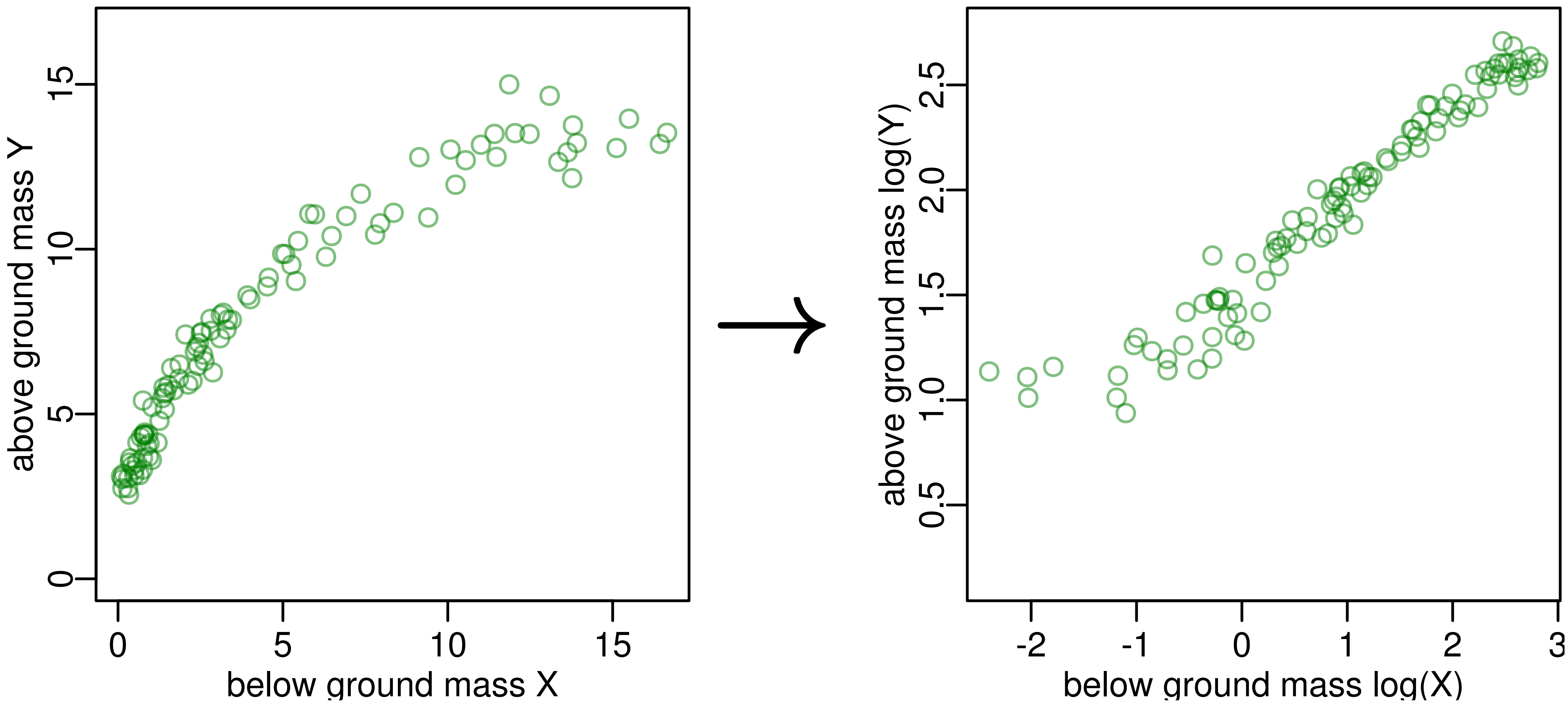
1. $\boldsymbol{X}$ も $\boldsymbol{Y}$ もすぐに対数変換してしまう
2. $\log \boldsymbol{Y} \sim \boldsymbol{N}(\boldsymbol{a} + \boldsymbol{b} \log \boldsymbol{X}, \boldsymbol{\sigma}^2)$ な回帰をやっちゃう
3. 「ゆーい」とか「せつめい力 $\boldsymbol{R}^2$」とか言ってみる



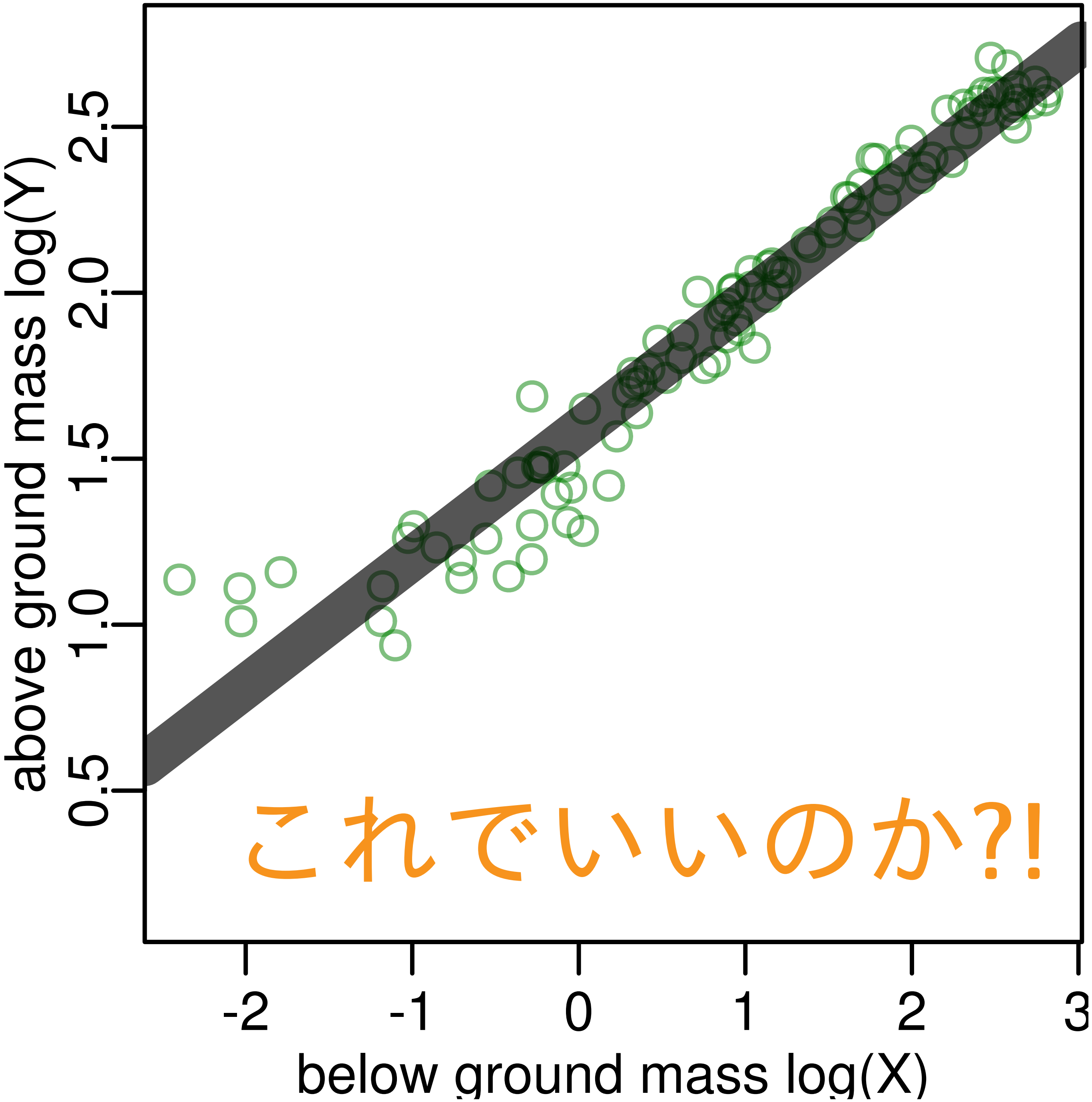


定義 あろめとらー: 上記のようなことをするヒトたち

つまりこのように対数変換してしまって

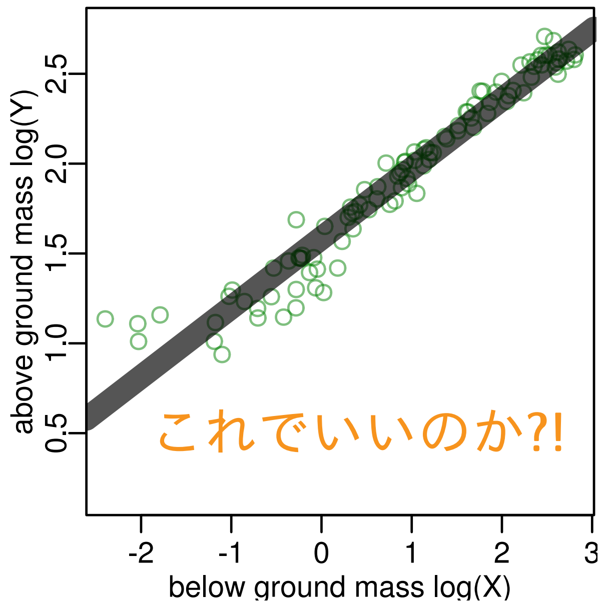


「センをひっぱる」なる行動をします，と……


above ground mass log(Y)
0.5
1.0
1.5
2.0
2.5
-2
-1
0
1
2
3
below ground mass log(X)
これでいいのか?!

あろめとらーなお作法の問題点

- それって地下部重 (X) が原因で地上部重 (Y) が結果なのか?
  - 因果関係がわからんのに回帰してよいの?
- なぜ $E(\log Y) = a + b \log X$?
  - これって何を表現しているモデル?
- X の測定時の誤差はどこにいった?
  - なぜ Y 軸方向にだけ「誤差」?



こういう現象をあつかうもうひとつのモデル?

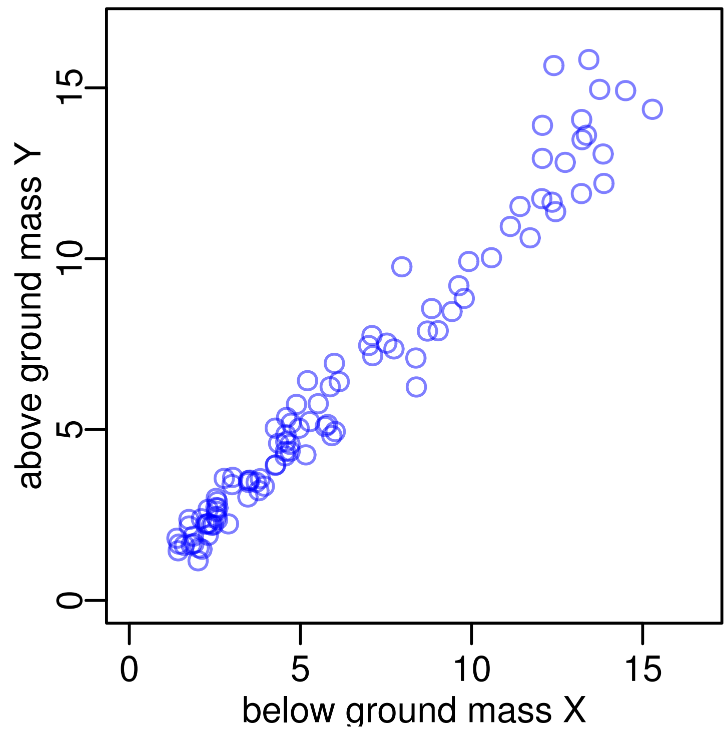
重量分割モデル: 階層ベイズモデルとして定式化

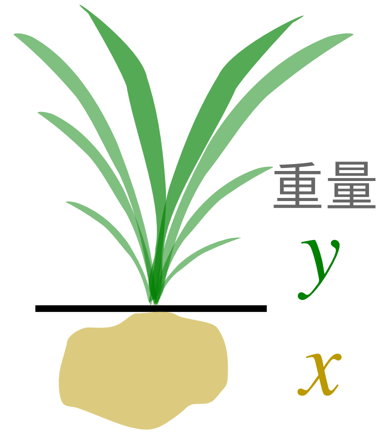
- X も Y も結果である，原因ではない
  - (バイオマスの) 重量分割という現象の結果
- 全重量 $\longrightarrow$ 地下部 (X) + 地上部 (Y) と考える
  - これも「近似的な」現象のとらえかたではあるけれど
- X と Y の測定時の誤差を明示的にあつかう
  - そして「個体差」由来のばらつき(random effects) も考慮する

例題 1

「アロメトリーな回帰」はヤメて 重量分割モデルを作ってみよう

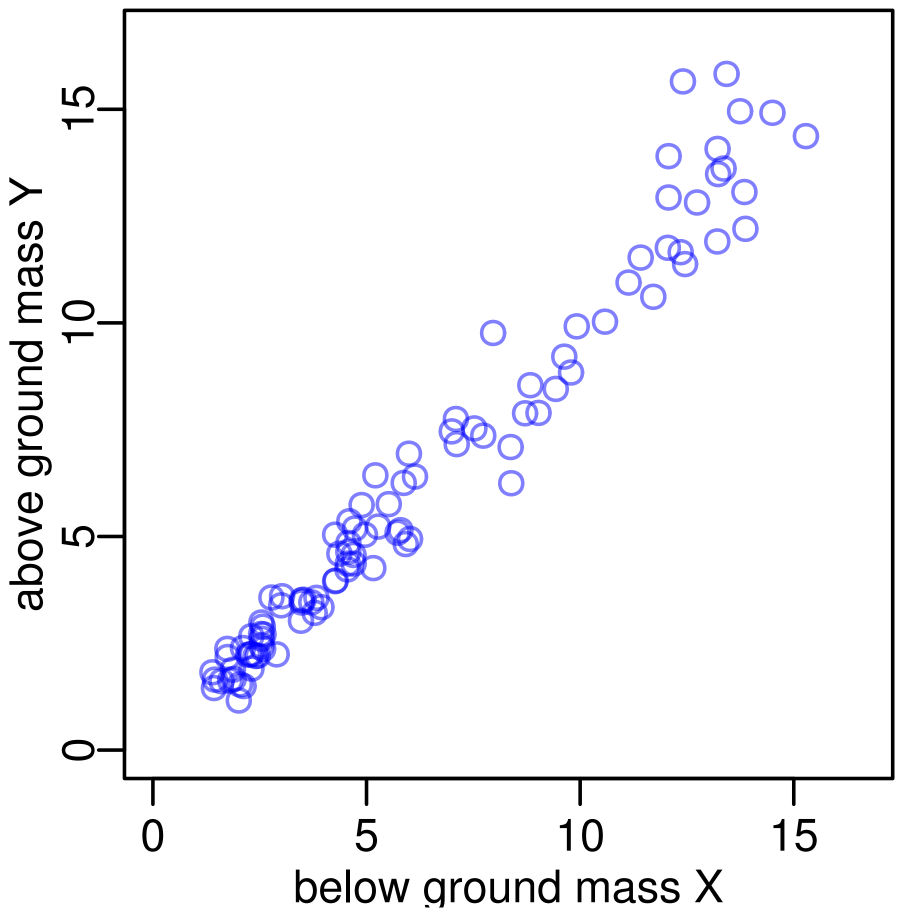
架空データ 1: 地下部・地上部の重量

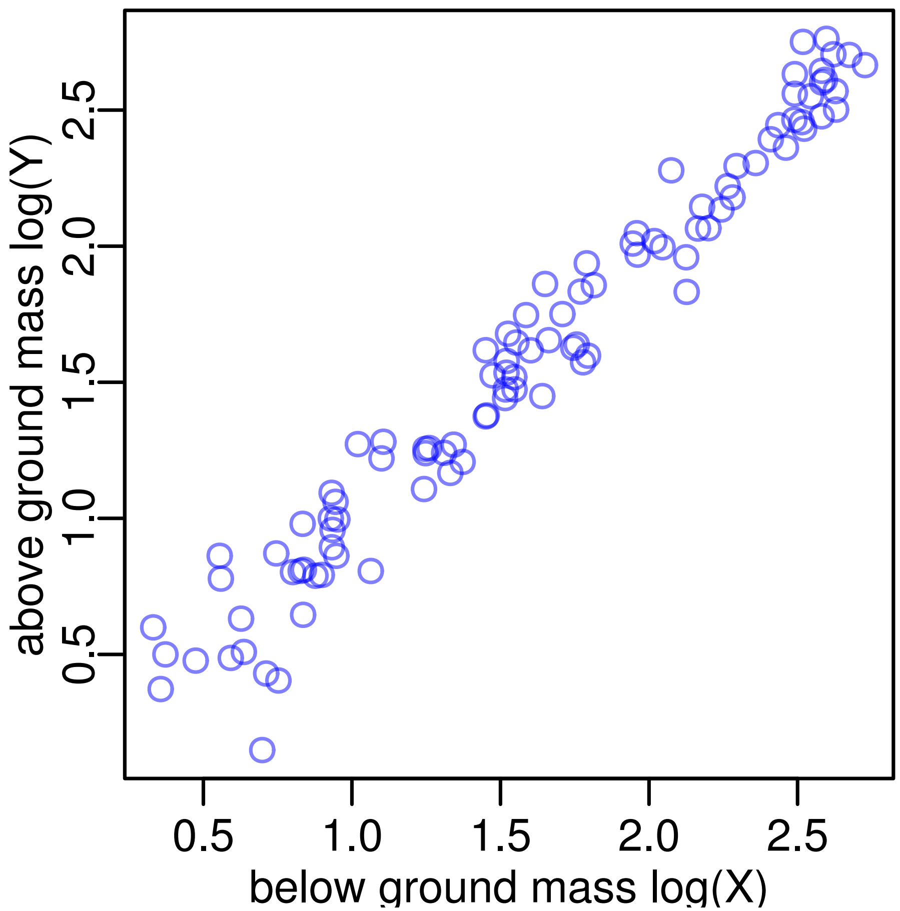




- 対数をとらなくても「直線」にのってる?
- つまり，むしろアイソメトリック (isometric)?
- 重量が重くなるとばらつきが大きくなる?

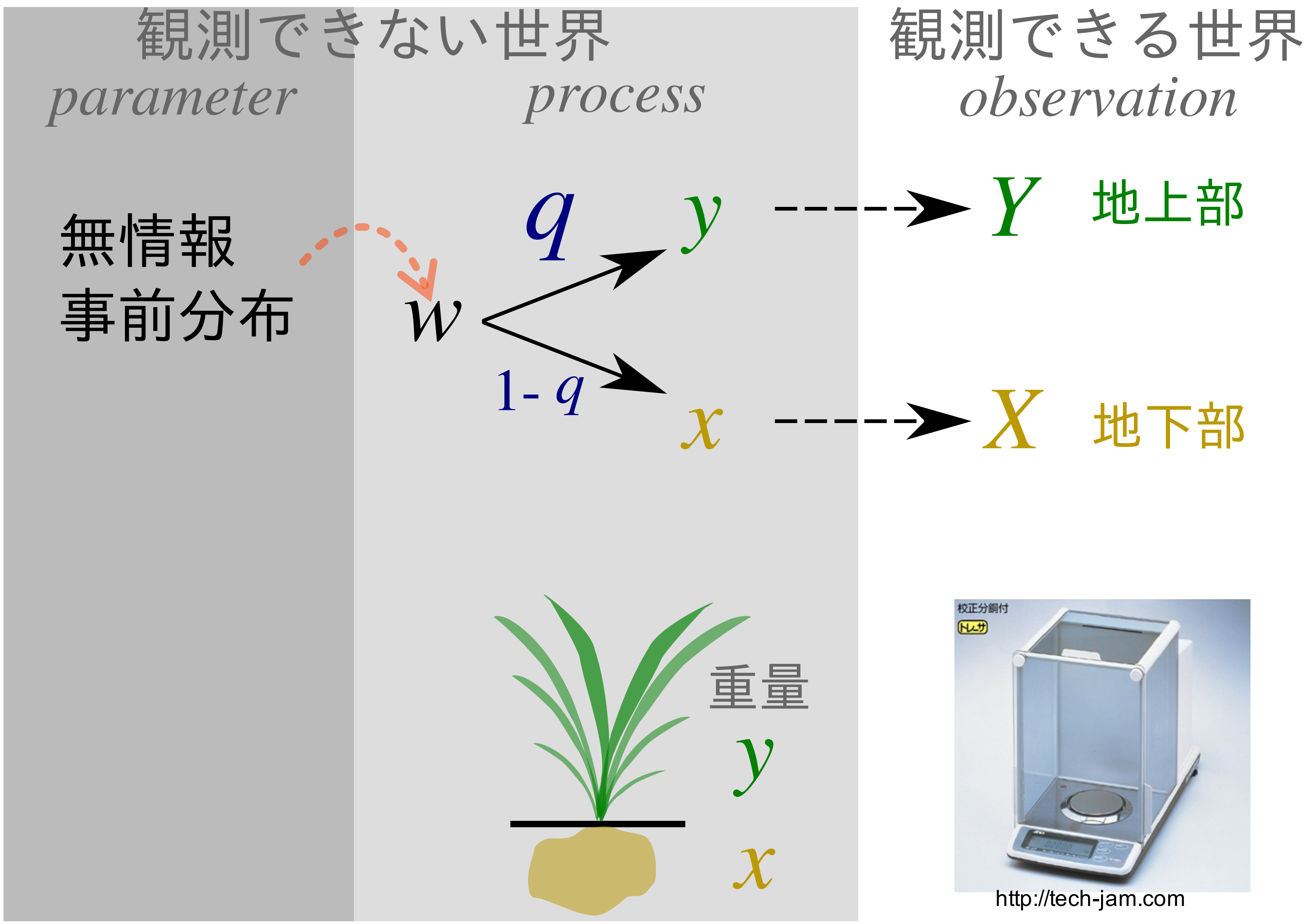
対数スケイルでみるとこうなってます



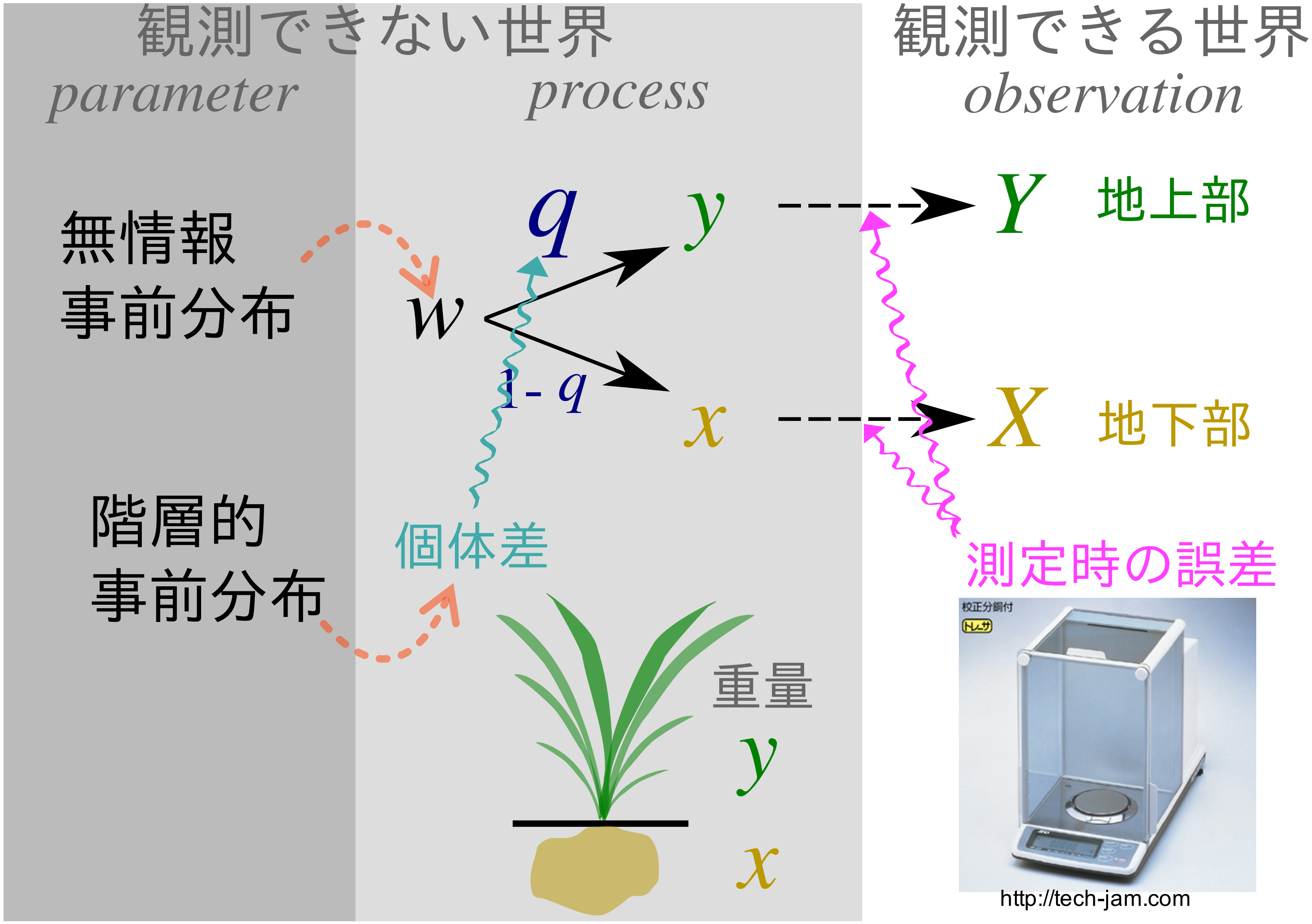


- とはいえ対数世界で「センをひっぱる」わけではない

重量分割モデル (階層ベイズモデル): そのプロセス



重量分割モデル (階層ベイズモデル): 「誤差」の入りかた



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]
  logit(q[i]) <- a + re[i]
  w[i] <- exp(log.w[i])
  log.w[i] ~ dnorm(0, Tau.noninformative) # !!
}# log.w[i] は地上部 + 地下部の重量
```

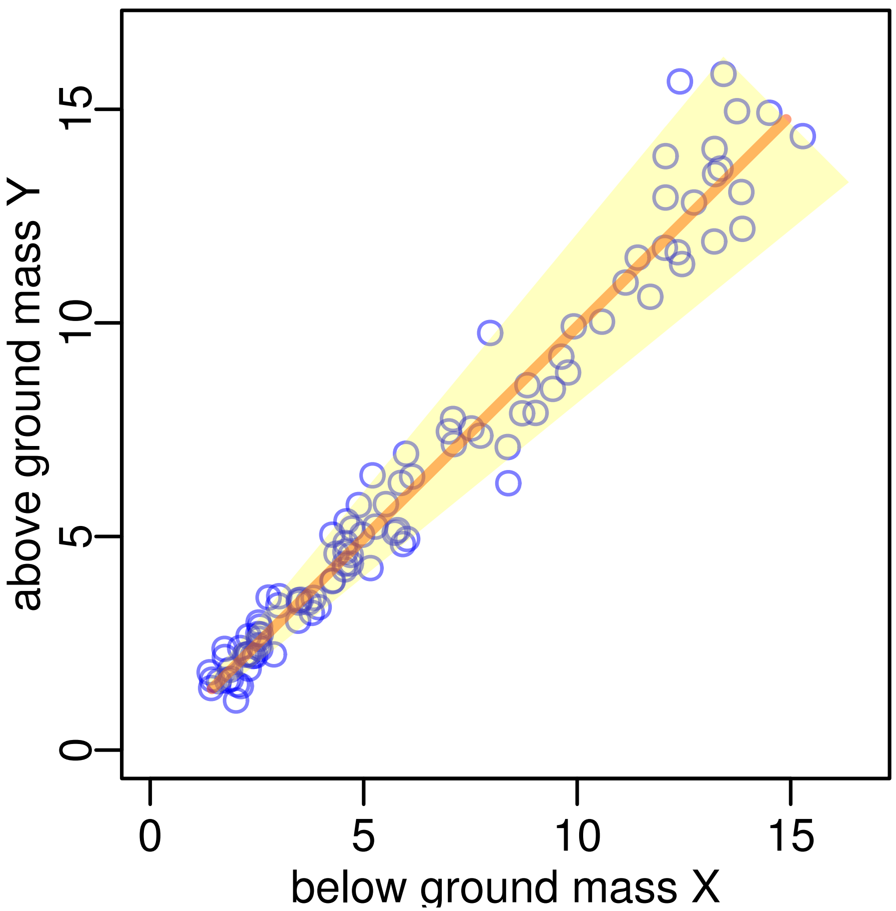
このように明示的にモデルを記述できる!

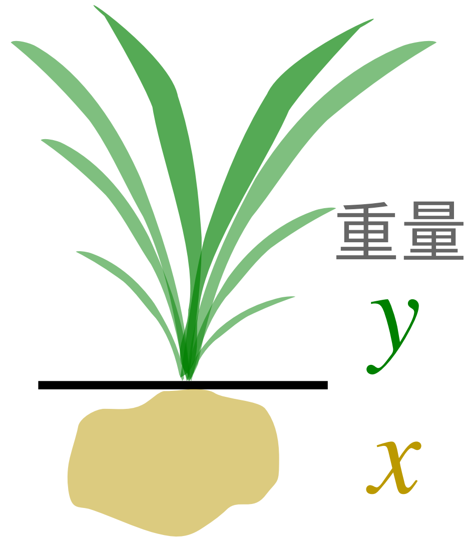
階層ベイズモデルのパラメーター推定: MCMC

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

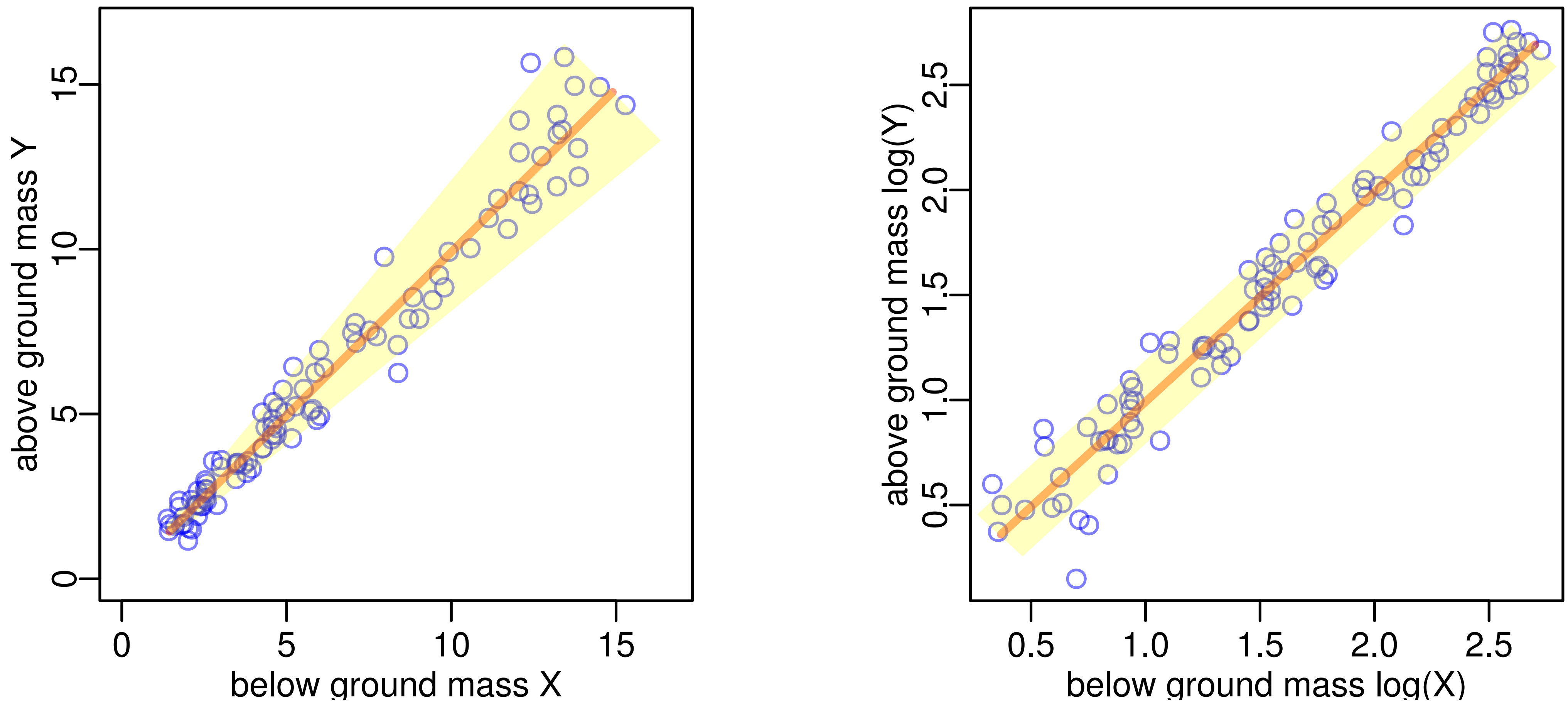
推定結果を組みあわせた予測





- 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**)

個体差によるばらつき，そして測定時の誤差



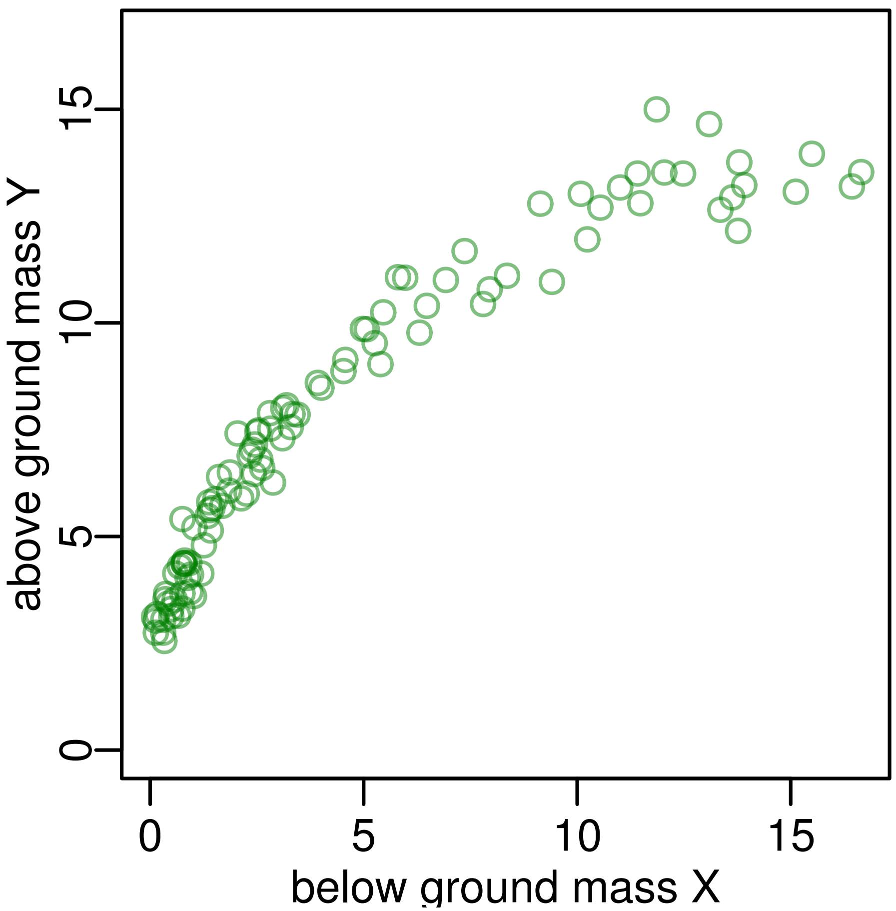
- 総重量が小さいときには測定時の誤差が相対的に大きく
- 総重量が大きくなると個体差が占める割合が大きくなる

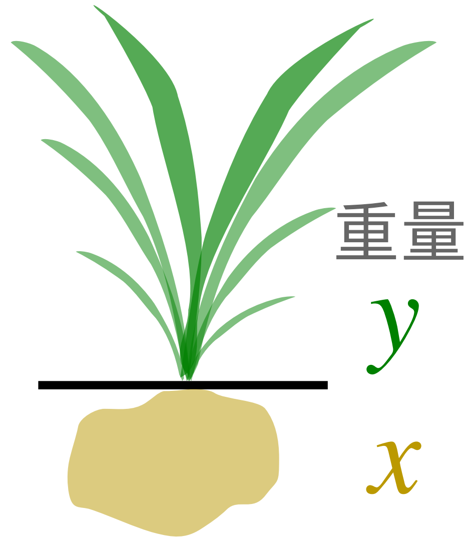
このモデルの問題点

- 測定時の誤差は正規分布と仮定していいのか?
  - そうですね，重量は非負の値なのに……
- 測定時の誤差の大きさはどうやって推定したの?
  - 今回は「真の値」をほうりこみました
  - 実際には，測定機器のカタログとか見ながら「てきとー」に決めるしかないのかも? (主観的な事前分布)
  - ひとつの観測対象に対して，複数の測定値が得られていれば，階層ベイズモデルで測定時の誤差の大きさを推定できます
- 状況がちょっと単純すぎない?
  - それでは次の例題を……

例題 2
もうちょっと複雑な
重量分割モデル

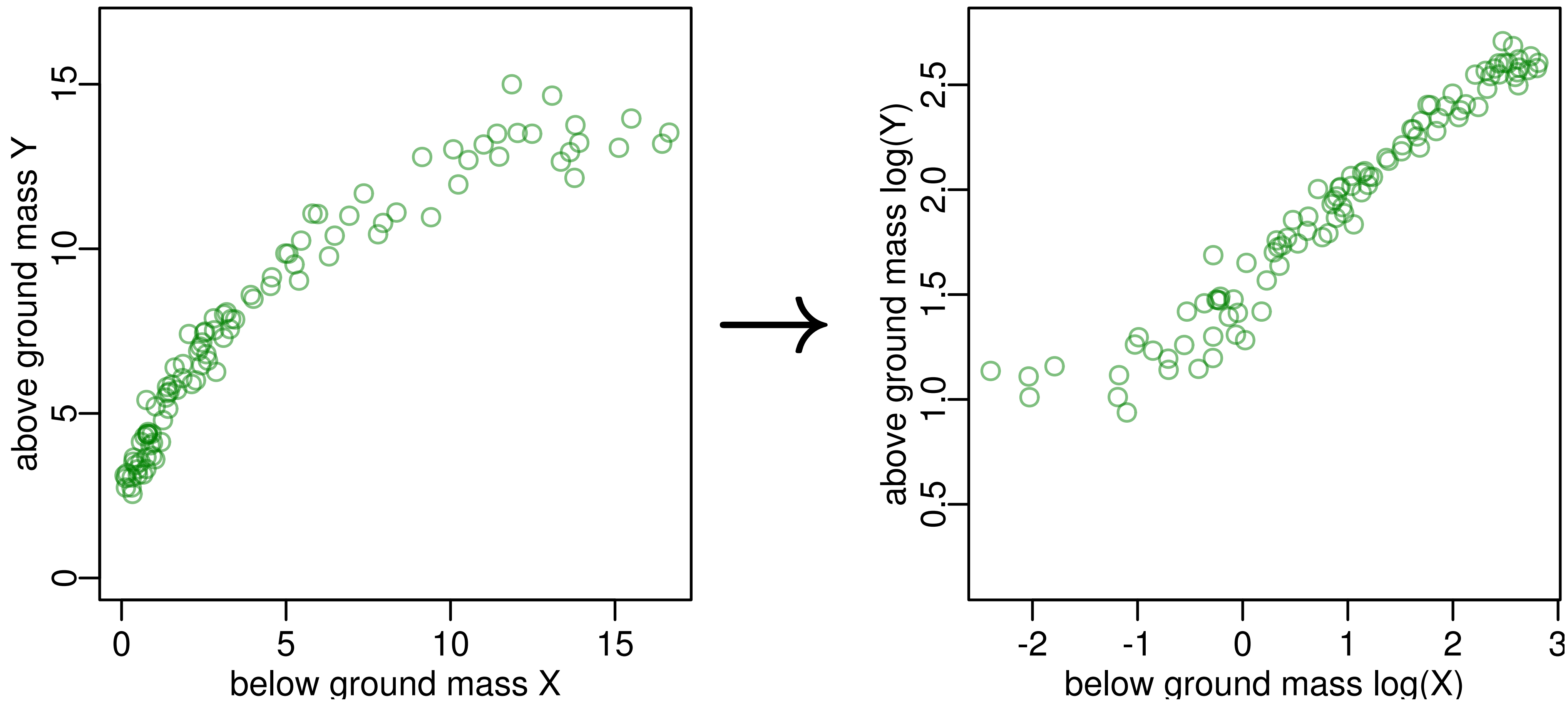
架空データ 2: 重量増大とともに分配が変化





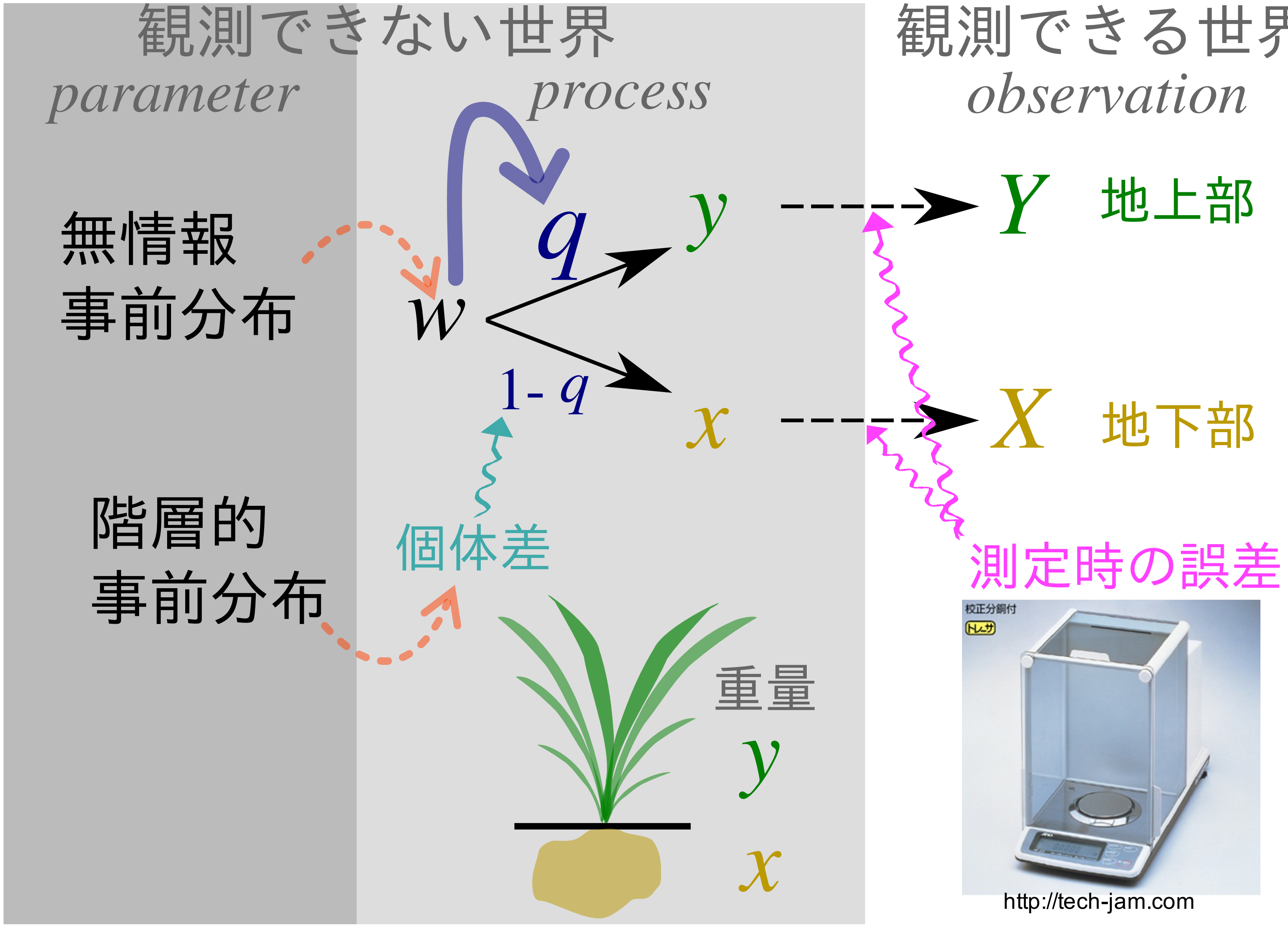
- 小さいときには地上部重量を大きくする
- 総重量が大きくなってくれば地下部を大きくする

これはアロメトリーな問題なのだろうか?



- 両対数で直線になっているのか?
- ま，それはあとで考えることにして……

重量分割モデルの改造: q を w 依存にするだけ



BUGS code の変更点

- 先ほどの簡単な例では (切片) + (個体差) だったが

```
logit(q[i]) <- a + re[i]
```

- ここを以下のように**総重量 (w) 依存**に変更するだけ

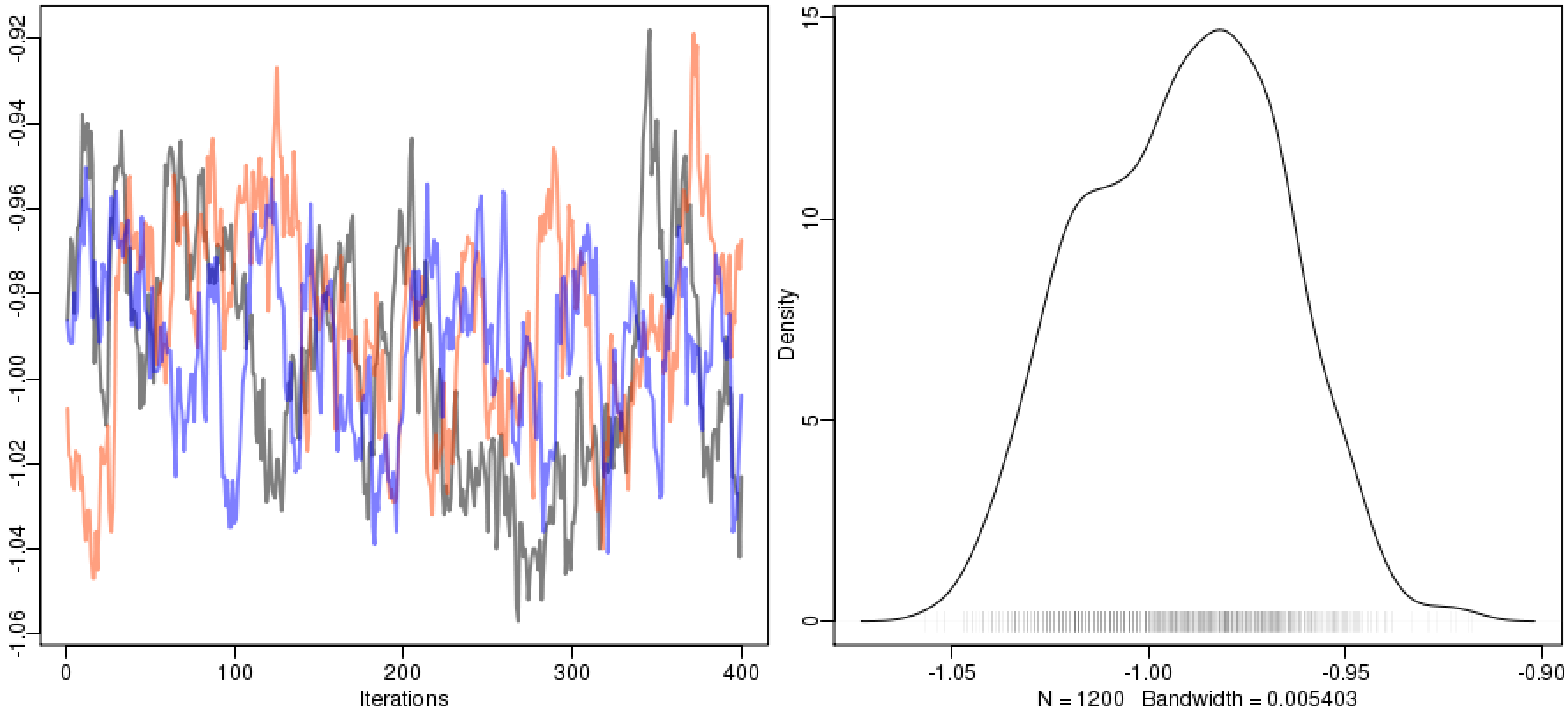
(b ∼ dnorm(0, Tau.noninformative) 追加も必要だけど)

```
logit(q[i]) <- (
  a + b * (log.w[i] - Mean.log.w) + re[i]
)
```

Mean.log.w うんぬんは WinBUGS に必須な中央化ワザ

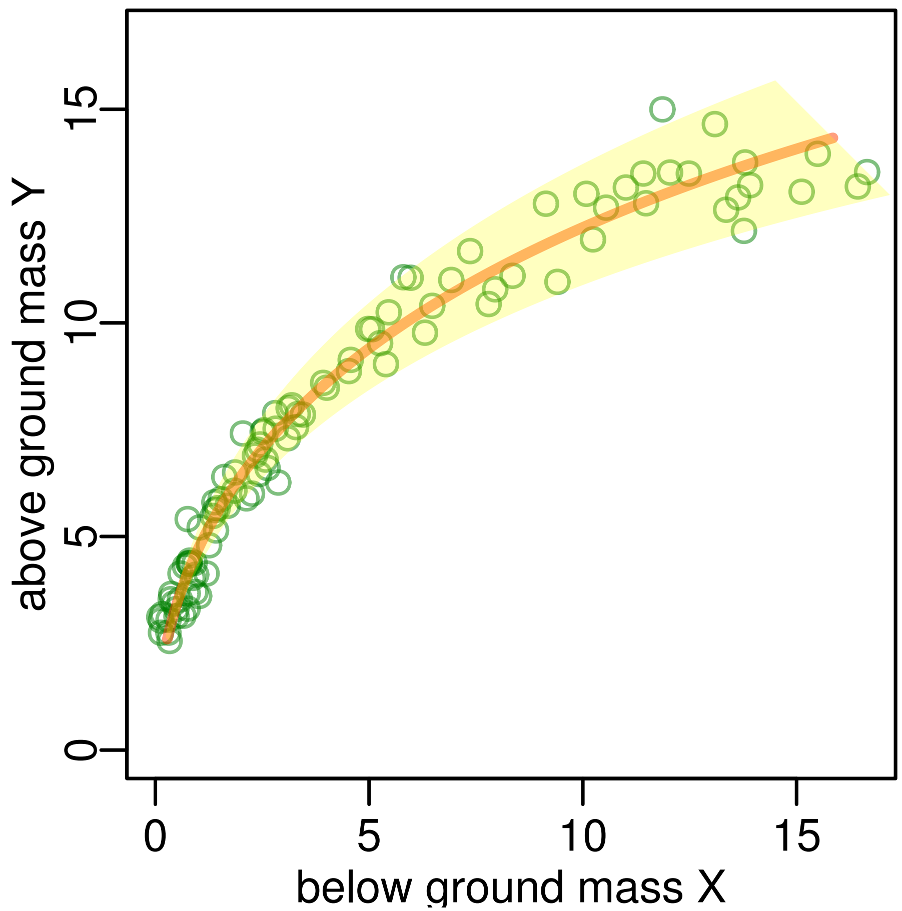
- あとは R と WinBUGS で MCMC するだけ

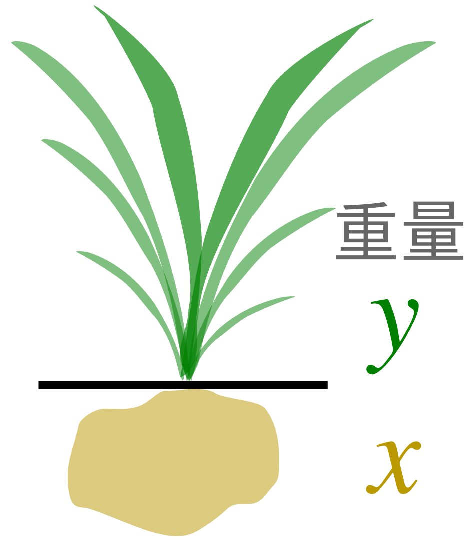
推定結果: 総重量増大 $\longrightarrow$ 地上部への分配減少



- 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

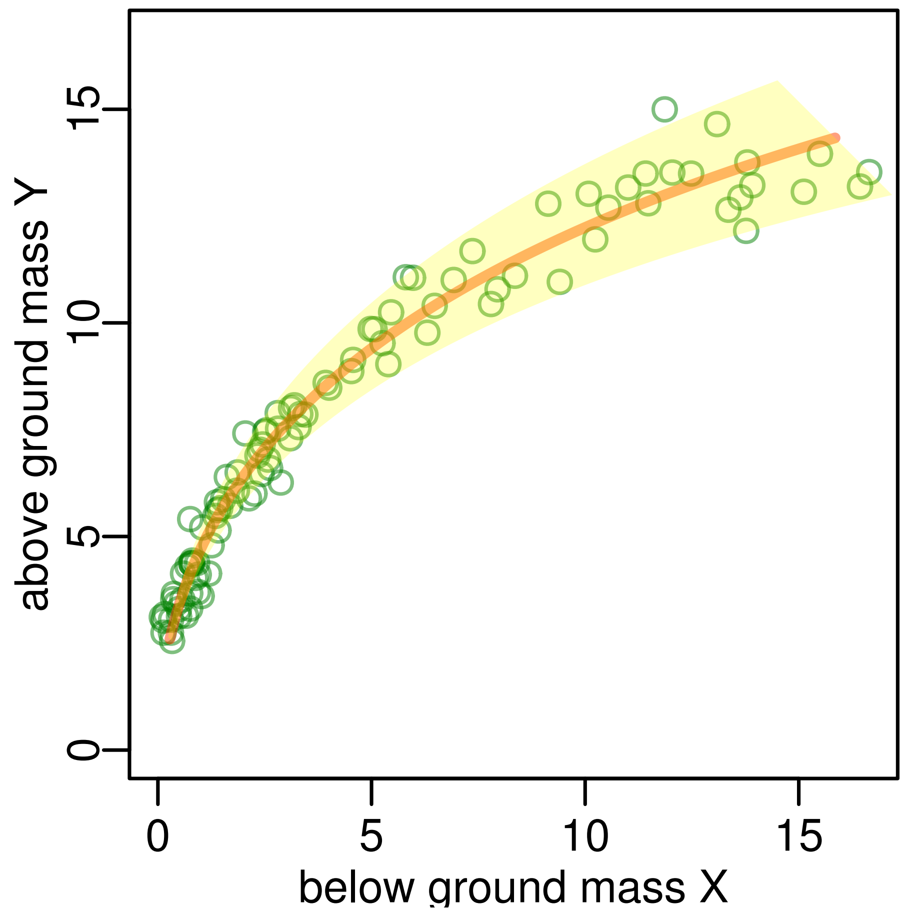
このモデルで複雑な重量分配を表現できる

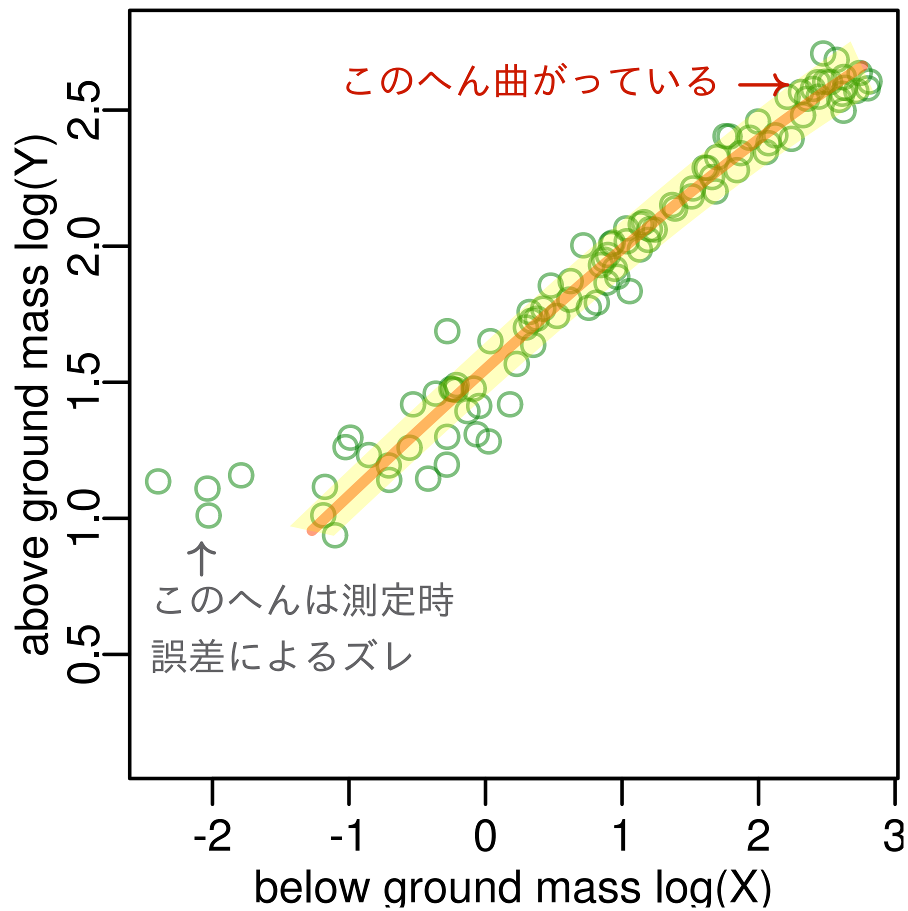




- オレンジ色の線は中央値 (**median**)
- 黄色の領域は個体差による予測のばらつき (**95% CI**)

対数の世界でも曲がっている (アロメトリーじゃない!)





- 「両対数表示でも曲がっている」状況でも重量分配モデルは柔軟に対応できる (さらに改訂するのも簡単)

連続数量の分配モデルの応用例

- **Iijima & Shibuya. 2010. J. For. Res 15:46–54.**
- 雌雄同株の植物個体内でのオス・メス繁殖器官への資源分配
- 植物個体内での資源分配を調べる安定同位体の存在比の解析
- 動物の行動観察記録に見られる「時間の分配」のモデル化…… ただし時系列構造の考慮も必要
- ほかにもいろいろあるかも?

今日のハナシのながれ

1. まずは $\{X, Y\}$ 誤差ありデータ解析における，ありがちな回帰適用お作法の問題点について検討
   - 因果関係なさそうなのに無理に回帰するのはヤメよう!
2. 解決策のひとつになりうる重量分割モデルの紹介
   - X と Y はどちらも結果だ
   - できればいくつかの X と Y の複数回測定を!
3. 重量分割モデルの拡張方法を検討
   - さらに生物学的な過程をとりこんだ改造も可能

例題 3

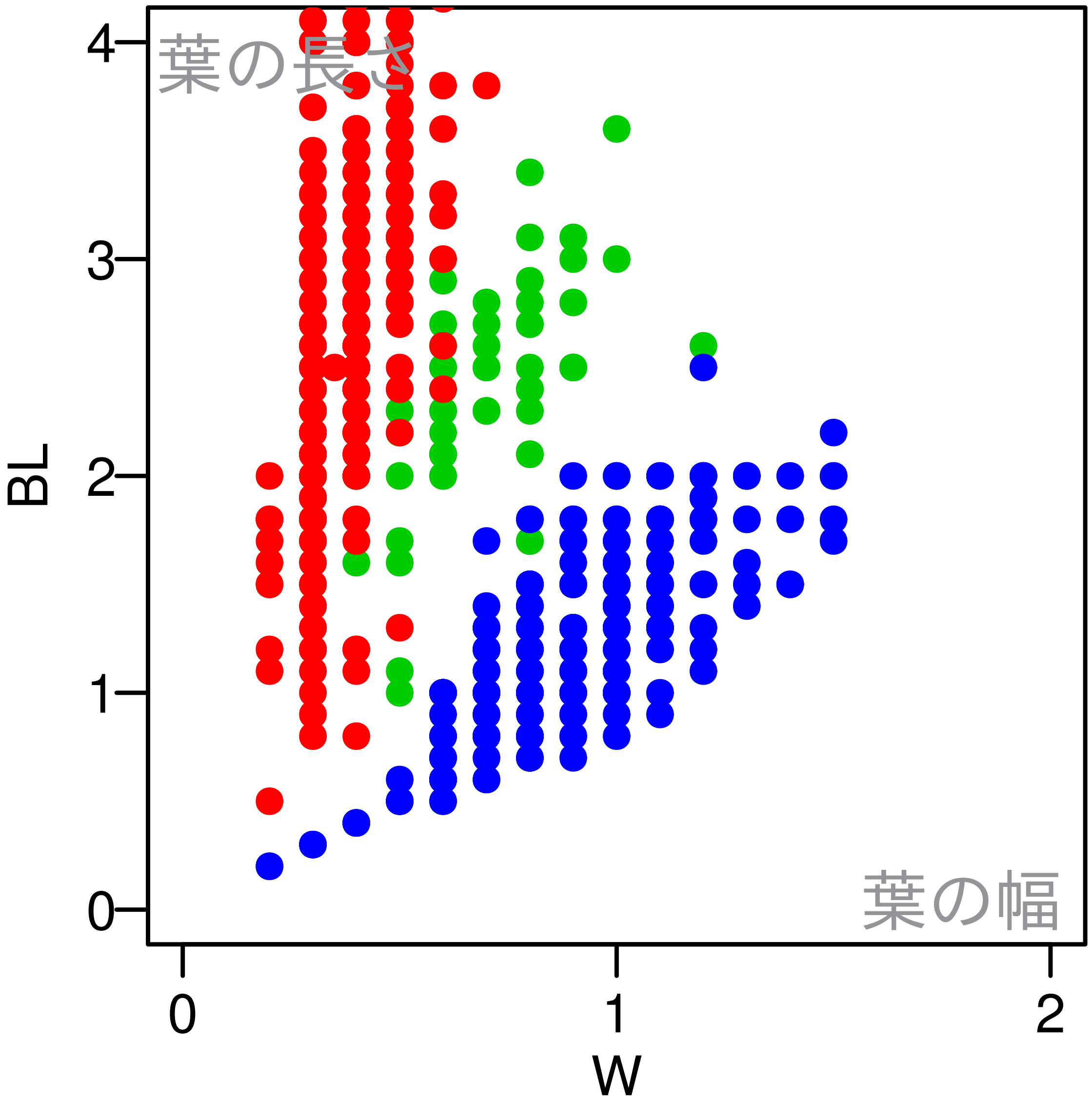
モウセンゴケの葉のカタチ
「面積」分割モデルを作ってみよう

例題 3, 4 のデータ

ここで保要さんにモウセンゴケ研究の概要

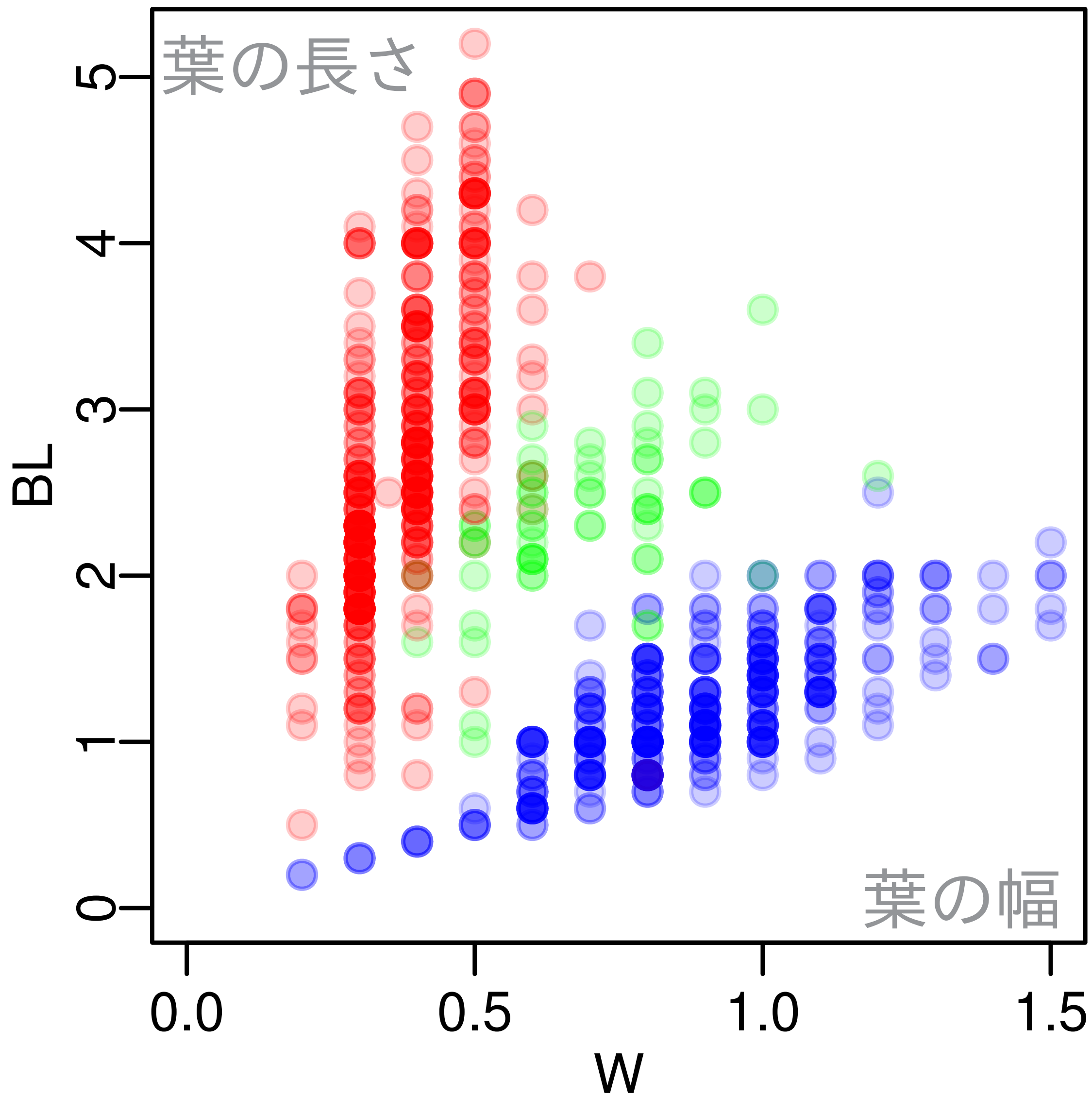
とそのデータについて解説をお願いします

さてさて，まずは作図についての検討から



- うーむ，R の作図機能をもっと使ってみよう!

半透過色指定で見えなかったデータを見る



- それから図のタテヨコの range (`xlim`, `ylim`) 指定も正確に!

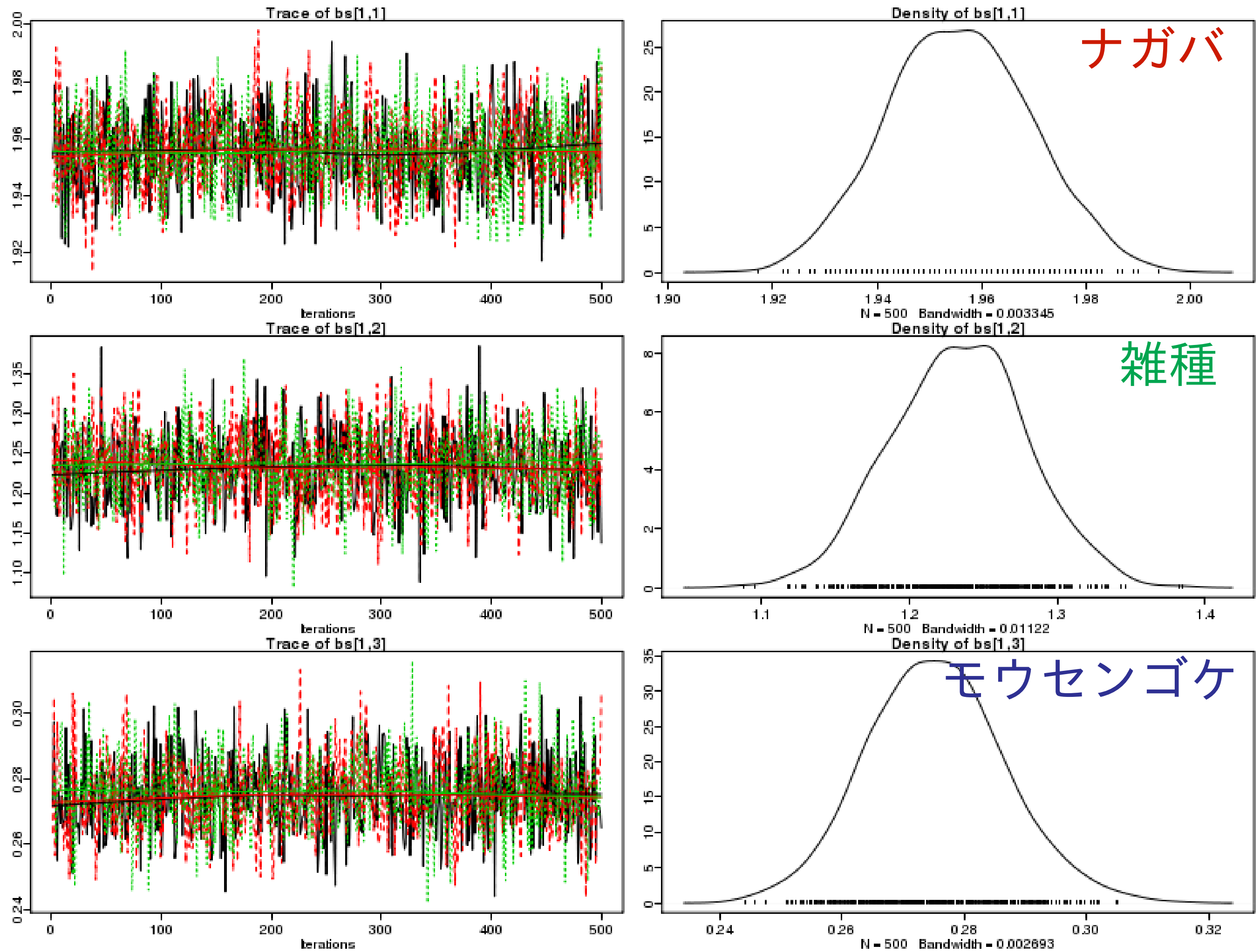
統計モデリング: 単位長さとタテヨコ比

- 単位長さ $= \sqrt{\text{タテの長さ} \times \text{ヨコ幅}}$
  - つまり 単位長さ $= \sqrt{\text{葉面積みたいなもの}}$
- タテヨコ比 c = タテの長さ / ヨコ幅，という概念
  - タテの長さ = 単位長さ $\times \sqrt{c}$
  - ヨコ幅 = 単位長さ / $\sqrt{c}$
  - $c > 0$
- これらの事前分布を考える
  - log(単位長さ): 無情報事前分布 $N(0, 10^2)$
  - $\log(c)$: 無情報事前分布 $N(0, 10^2)$

葉のタテヨコモデルを BUGS code で (process の部分のみ)

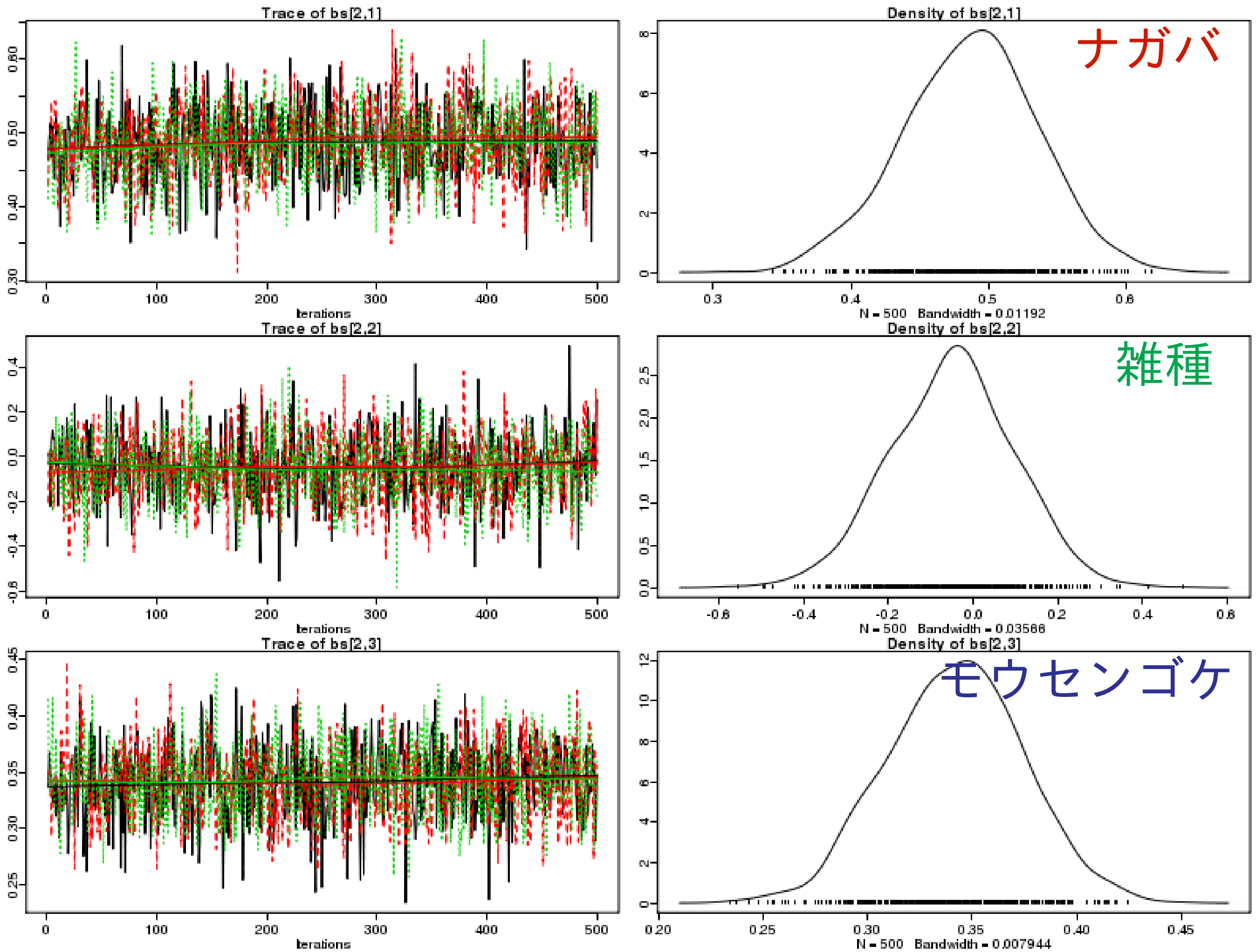
```
Y[i] ~ dnorm(y[i], Tau.err)
X[i] ~ dnorm(x[i], Tau.err)
y[i] <- unit.length[i] * sqrt(rxy[i])
x[i] <- unit.length[i] / sqrt(rxy[i])
rxy[i] <- exp(log.rxy[i])
log.rxy[i] ~ dnorm(mean.log.rxy[i], tau[Spc[i]])
mean.log.rxy[i] <- (
  bs[1, Spc[i]]
  + bs[2, Spc[i]] * (unit.length[i] - Mean.ul)
)
unit.length[i] <- exp(log.unit.length[i])
log.unit.length[i] ~ dnorm(0, Tau.noninformative)
```

切片 β_1 の事後分布



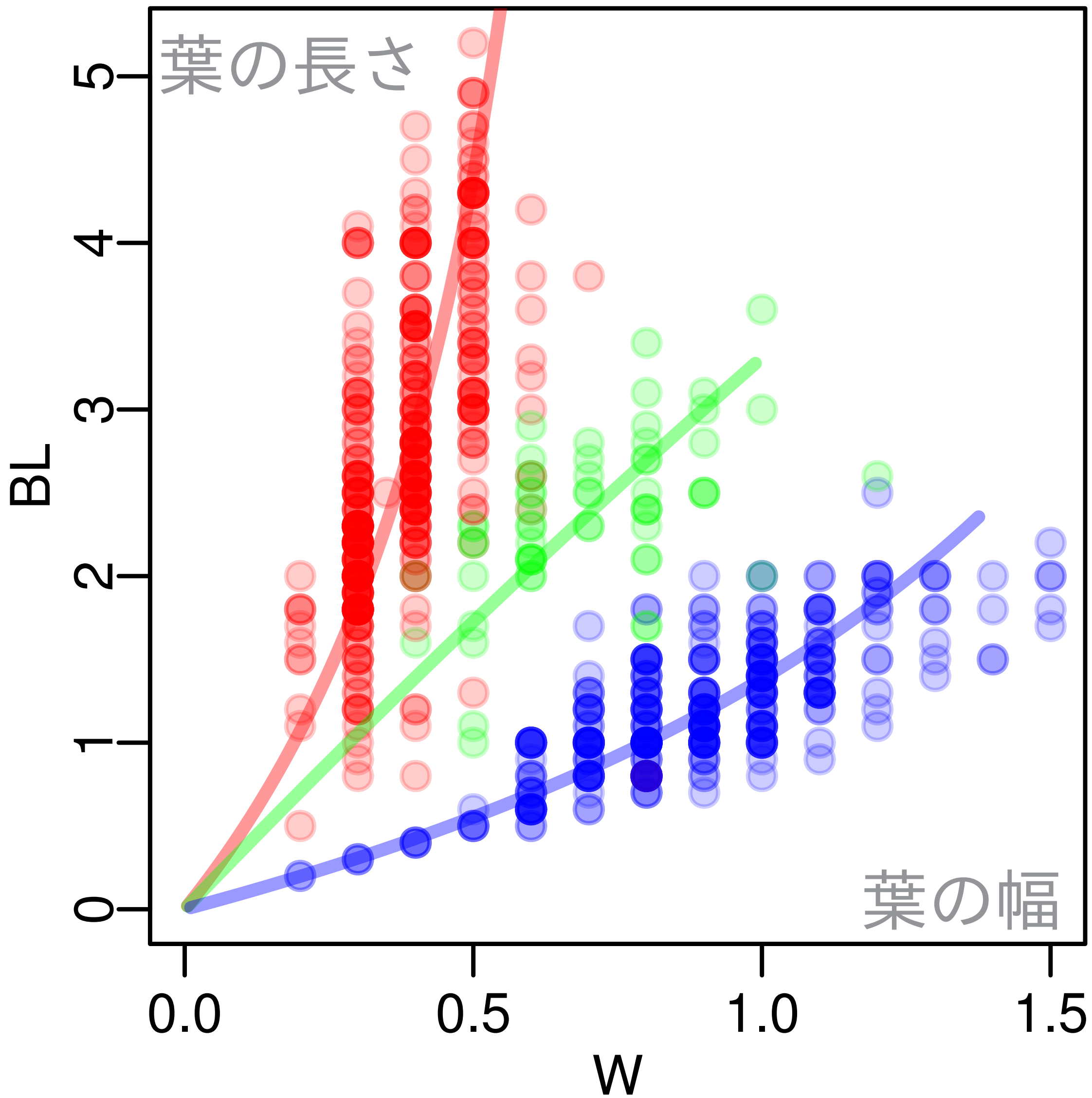
- 平均的な単位長さでの長さ / 幅

傾き β_2 の事後分布



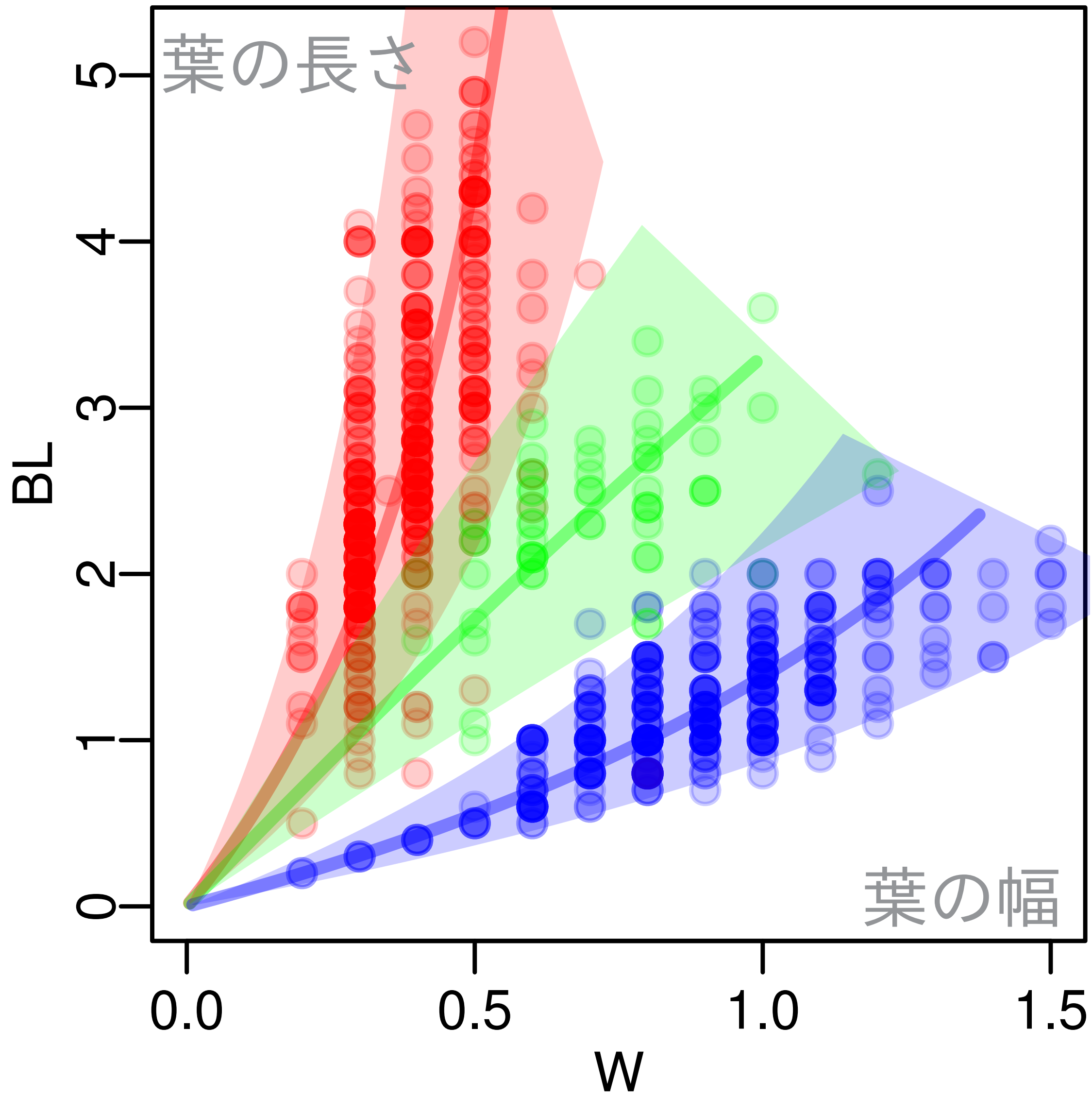
- 単位長さが増大すると細長くなるか，幅が広くなるのか?

中央値の変化: 成長とともに葉のカタチは変わる



- ナガバ・モウセンゴケ は次第に長細くなる
- 雑種 は同じカタチのままなのかな?

個体差によるばらつきを図示してみる

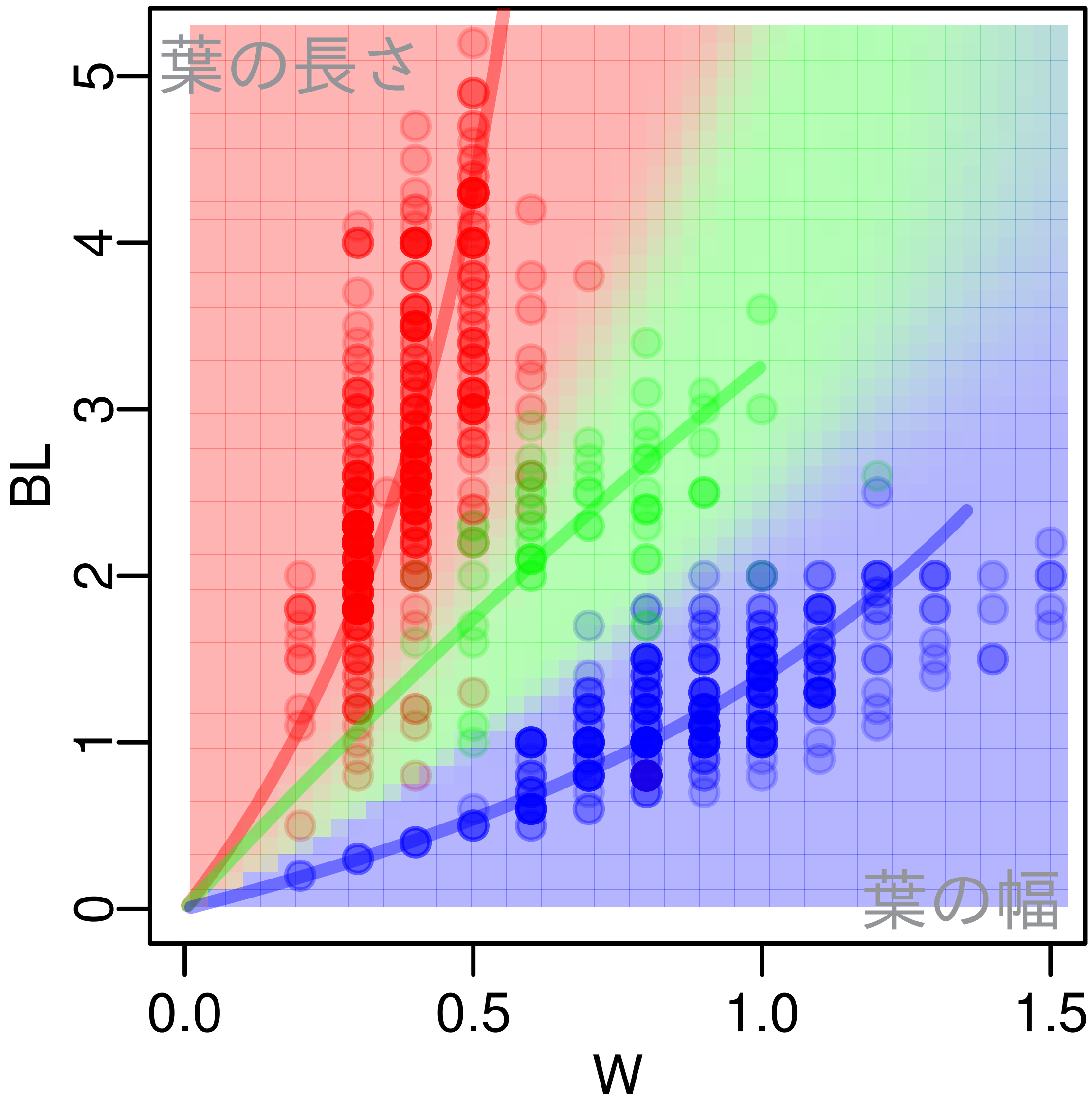


- 分布が重なっている?

例題 4

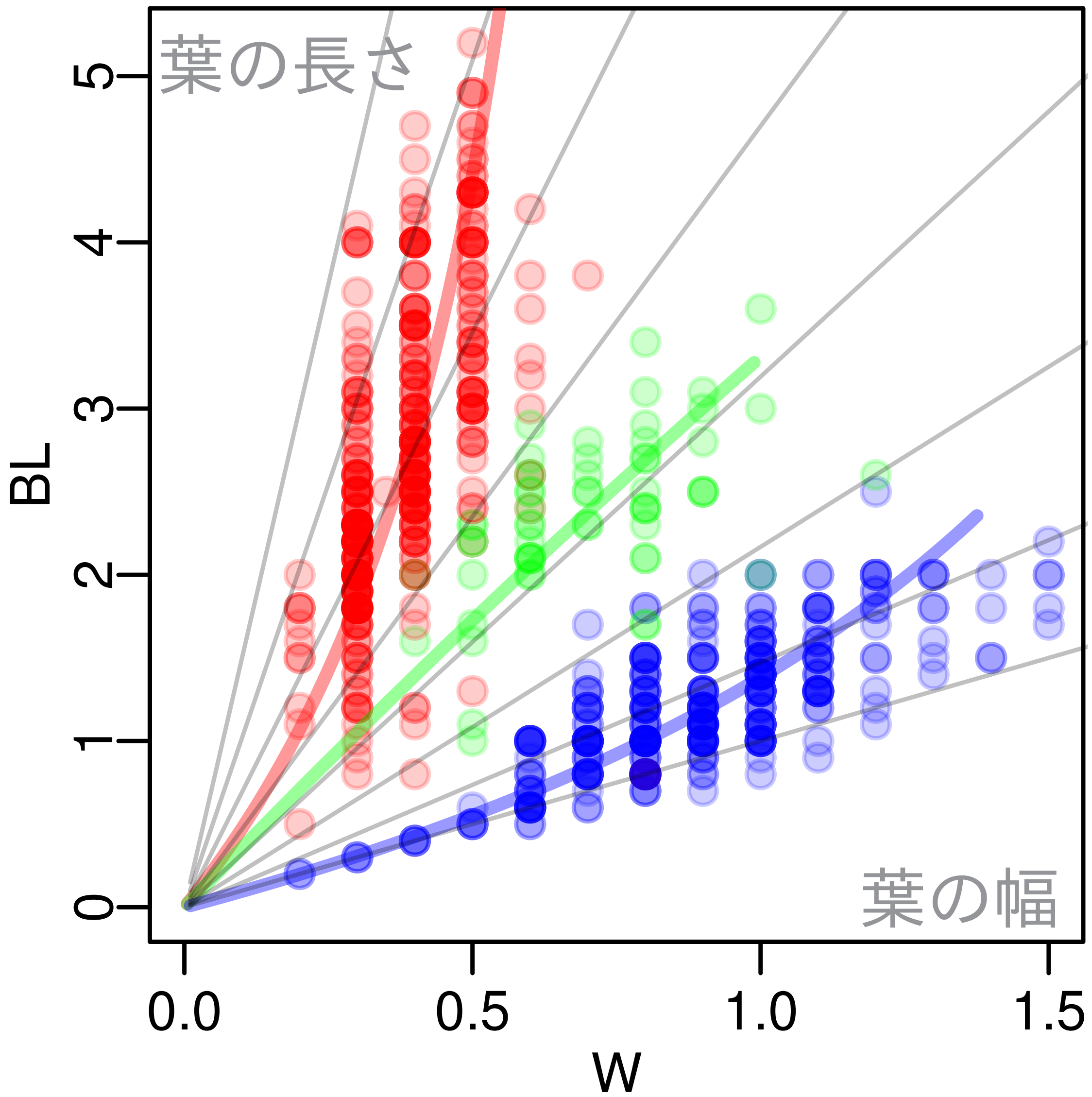
モウセンゴケの葉のカタチにもとづく種識別のベイズモデル

Q. 種の識別って? A. こんな図を作りたい



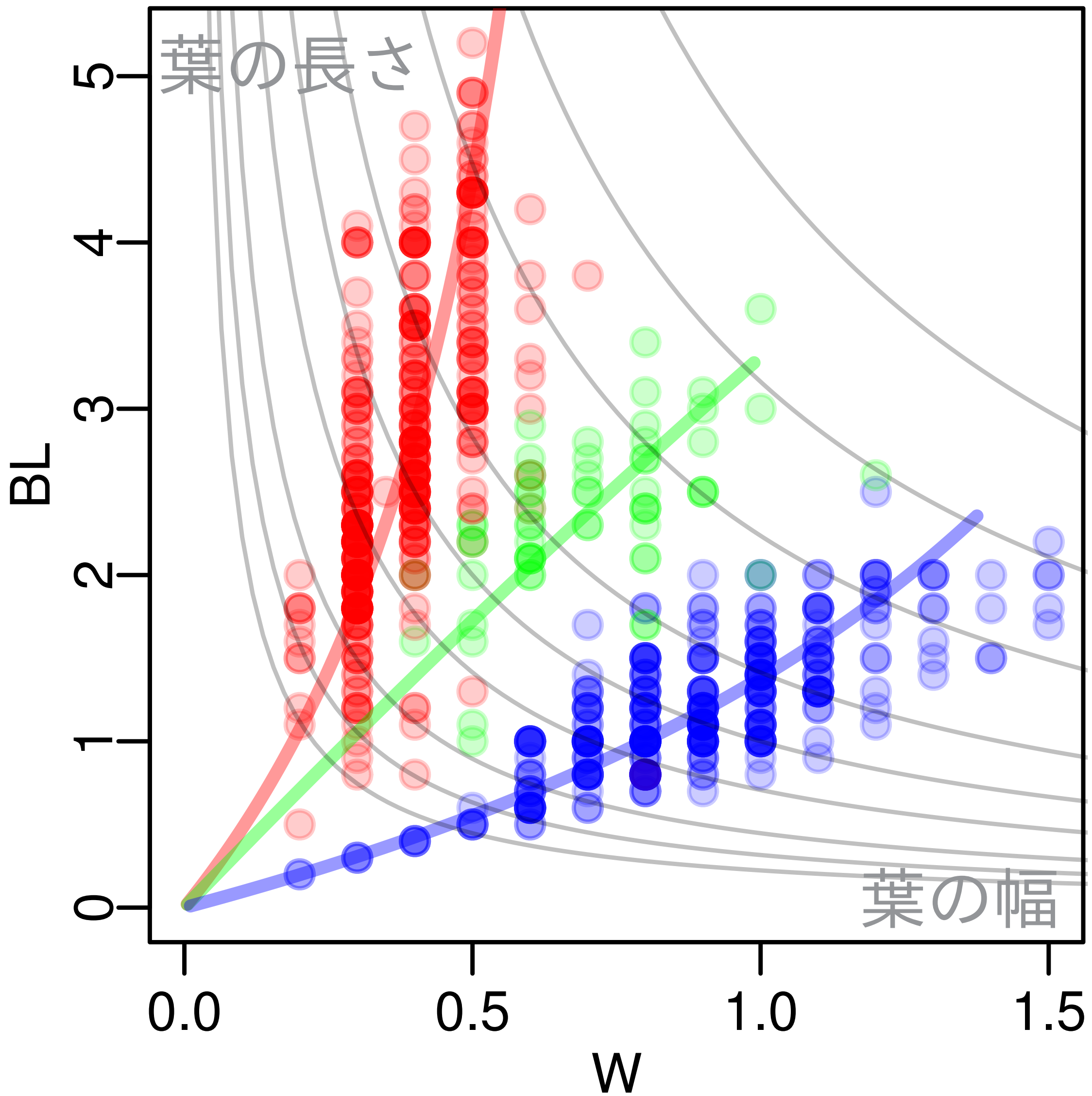
- ある葉の幅・長さが得られたときに，その情報にもとづいて，その葉はどの種のものであるかを知りたい

観測値どうしのタテわるヨコといった割算値はダメ!



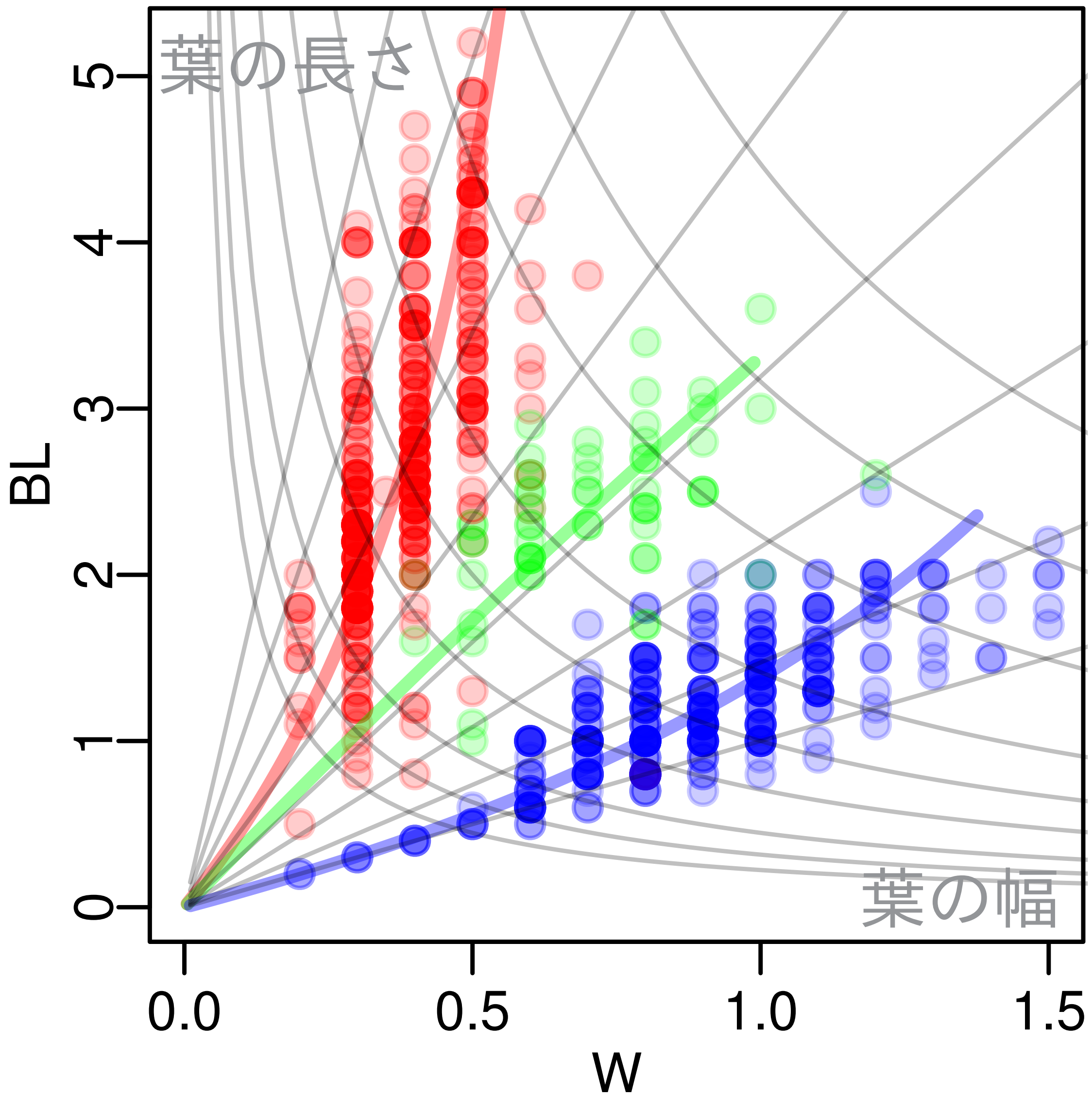
- 葉のサイズが変わると，タテヨコ比も変わるから
- さきほど得た推定結果，曲がる中央値

ある「面積」の葉が与えられたときに……と考える



- ある双曲線は同じ「面積 (単位長さ)」の葉をあらわす
- タテ × ヨコ = 面積, つまり $y =$ 面積$/x$

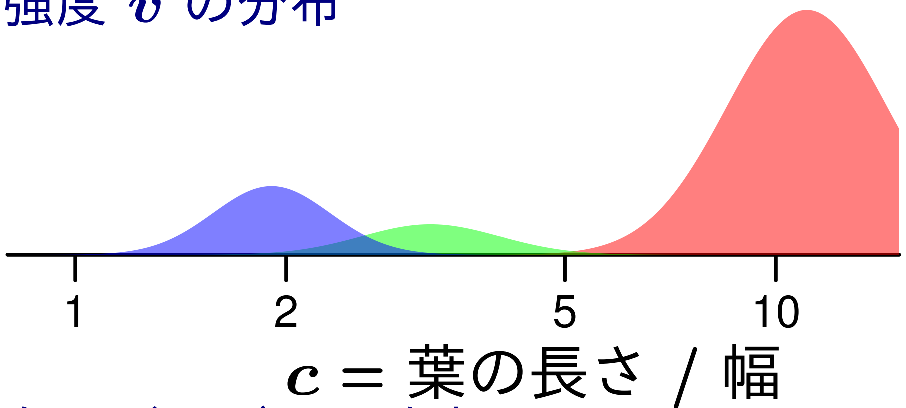
ある単位長さにおけるタテヨコ比 c で種を判別



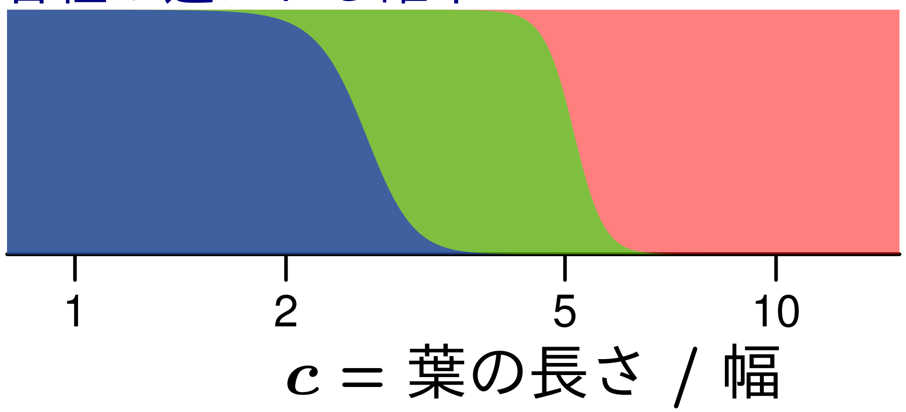
- この曲線上にさまざまなタテヨコ比の葉があると考える
- ただし観測値を割算するわけではない!

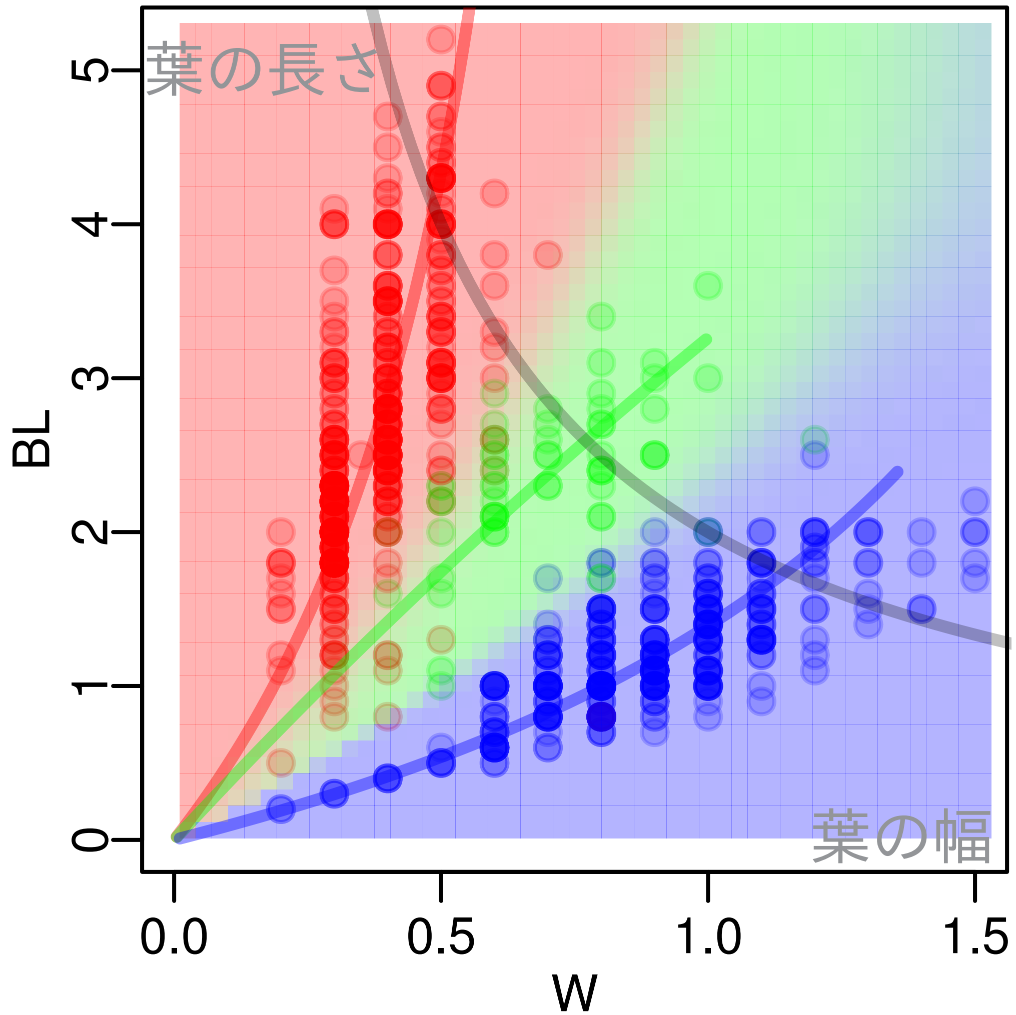
種ごとの強度 v という概念を導入する

単位長さが $\sqrt{2}$ のときの
強度 v の分布



各種が選ばれる確率



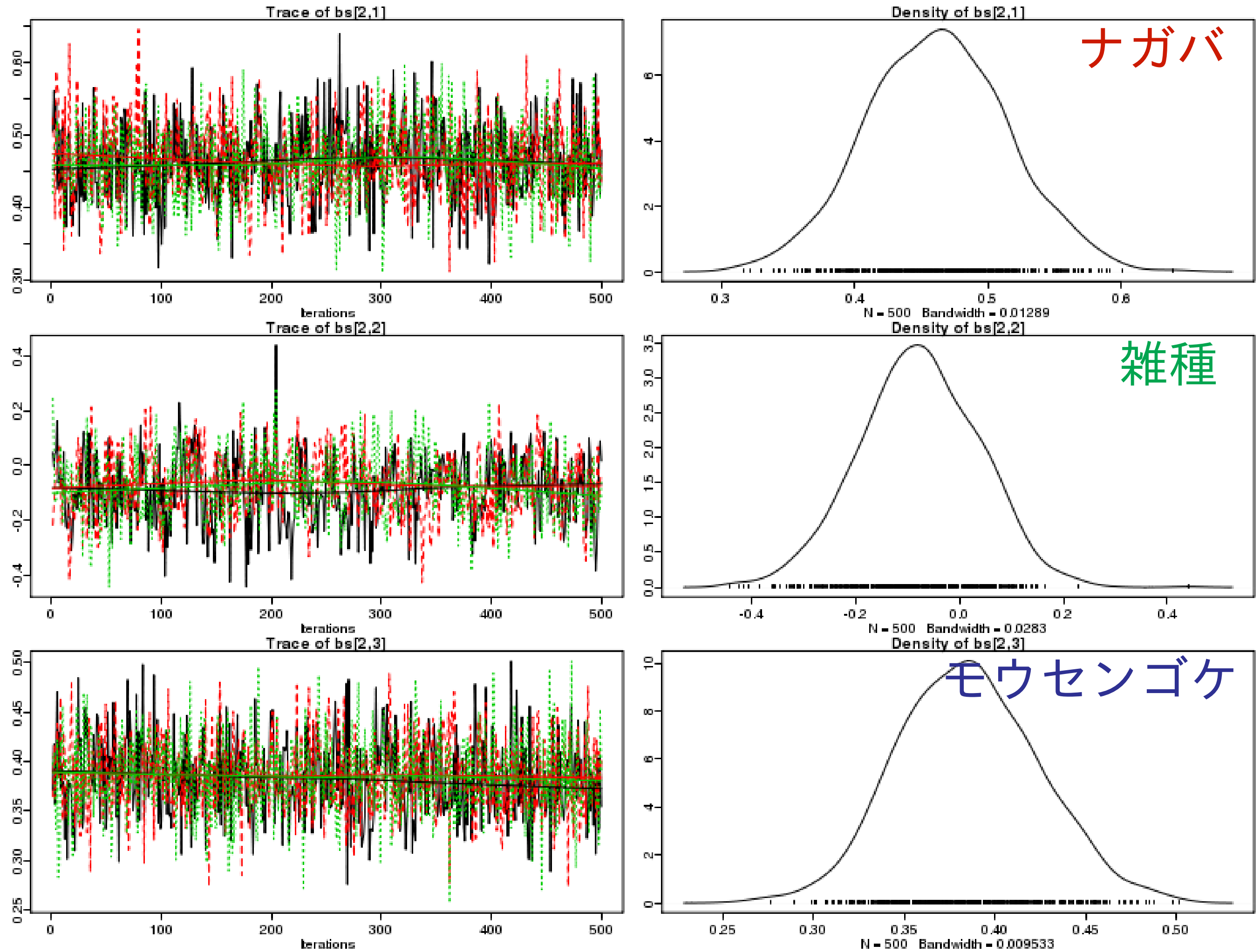


- データから種ごとの v の分布を推定する
- 3 種の v の相対値で識別の確率を評価する

葉のタテヨコモデルを BUGS code で (追加部分周辺)

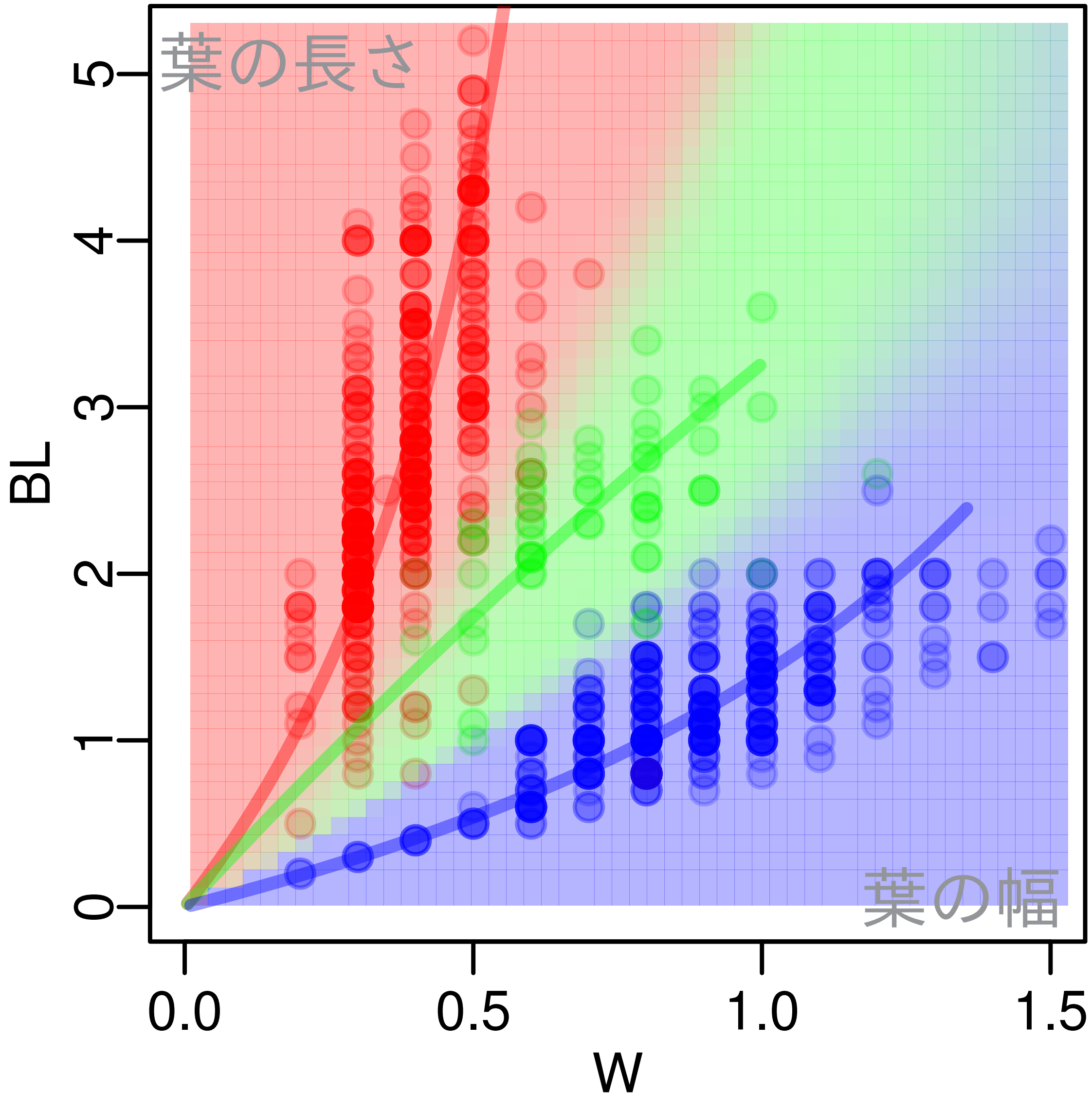
```
spc[i] ~ dcat(q[i,]) # 計算させながら変化させる種
Spc[i] ~ dcat(q[i,]) # データ（正解）へのあてはめ
q[i, 1] <- v[i, 1] * w[1] / total.v[i]
q[i, 2] <- v[i, 2] * w[2] / total.v[i]
q[i, 3] <- v[i, 3] / total.v[i]
total.v[i] <- v[i, 1] * w[1] + v[i, 2] * w[2] + v[i,
for (s in 1:N.spc) {
  v[i, s] <- exp(
    -pow(log.rxy[i] - mean.log.rxy[i, s], 2) * tau[s]
  ) / sigma[s]
}
```

また WinBUGS $\longrightarrow$ 傾き β_2 などの事後分布



- 単位長さが増大すると細長くなるか，幅が広くなるのか?
- 先にやったカタチのあてはめモデルとほぼ同じ結果

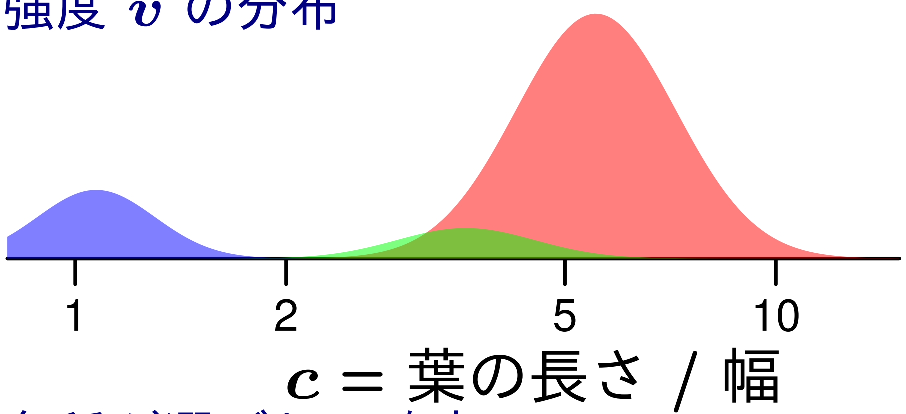
推定結果: 種を識別する領域



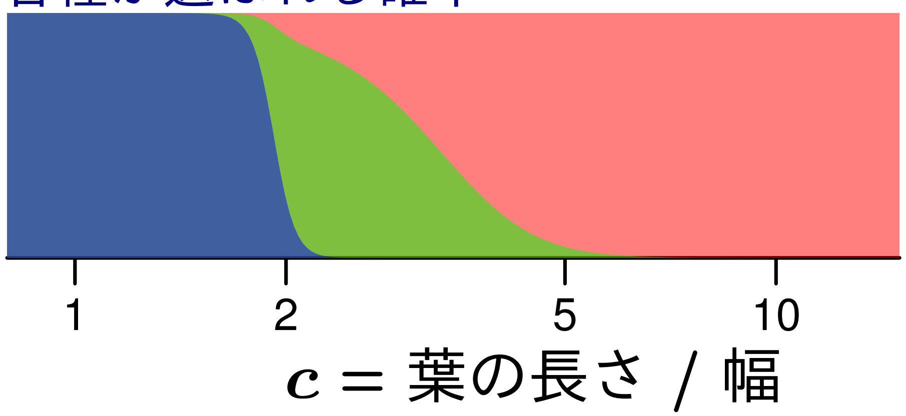
葉が小さい (単位長さが短い) ときの種の識別

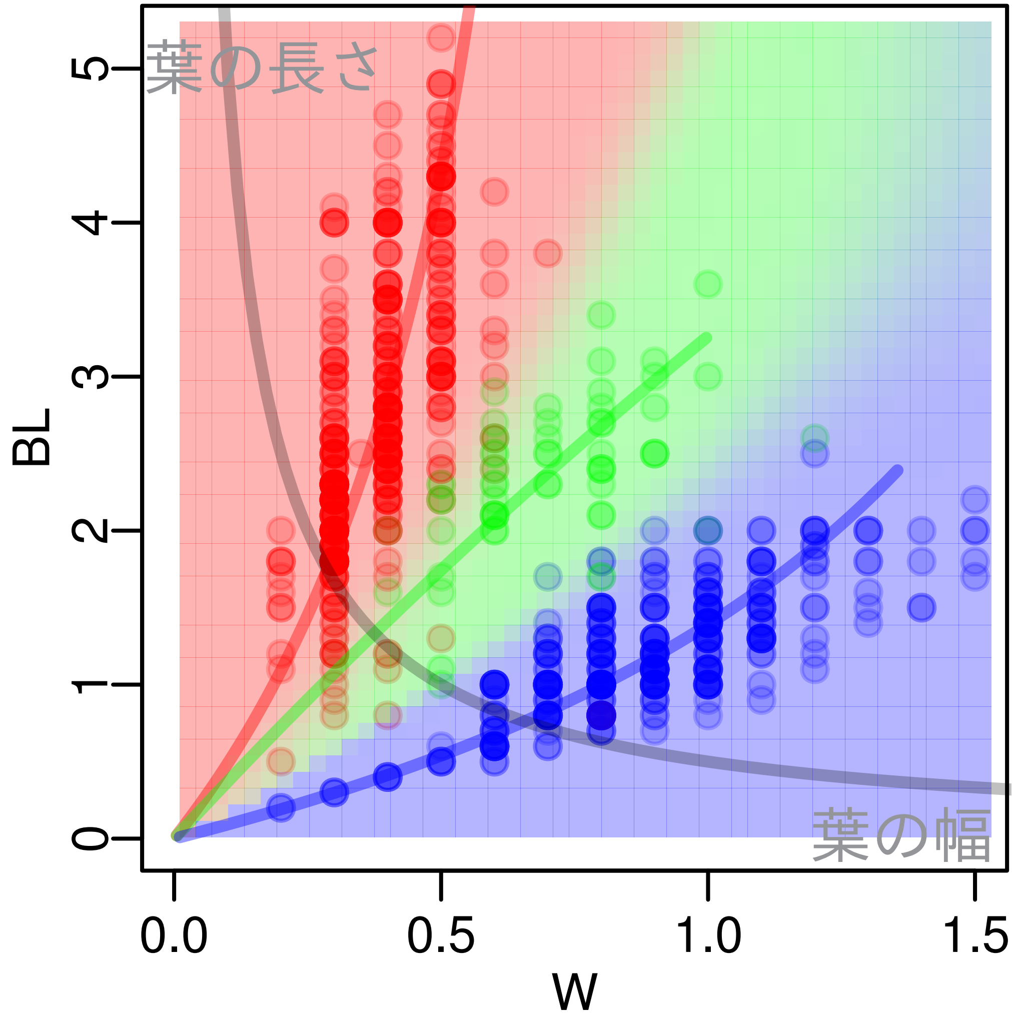
単位長さが $\sqrt{\mathbf{0.5}}$ のときの

強度 $\boldsymbol{v}$ の分布



各種が選ばれる確率



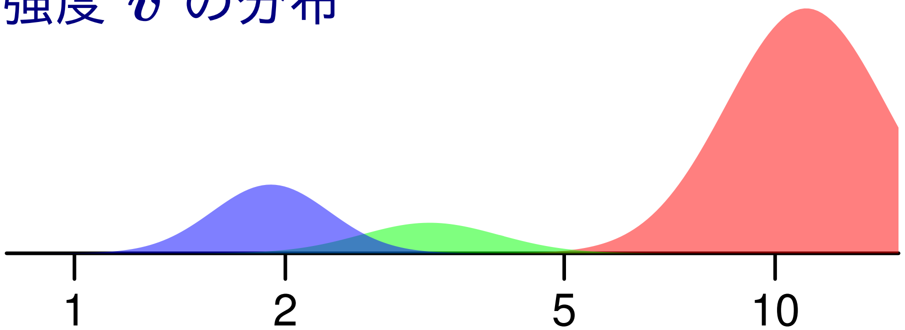


- ??ガバ・雑種 の識別が難しい

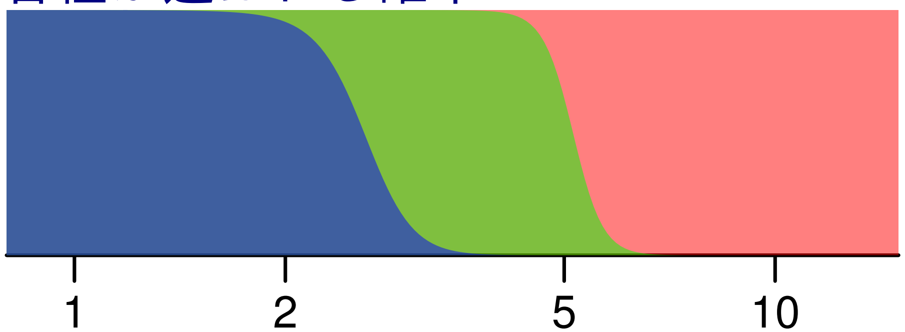
葉が大きい (単位長さが長い) ときの種の識別

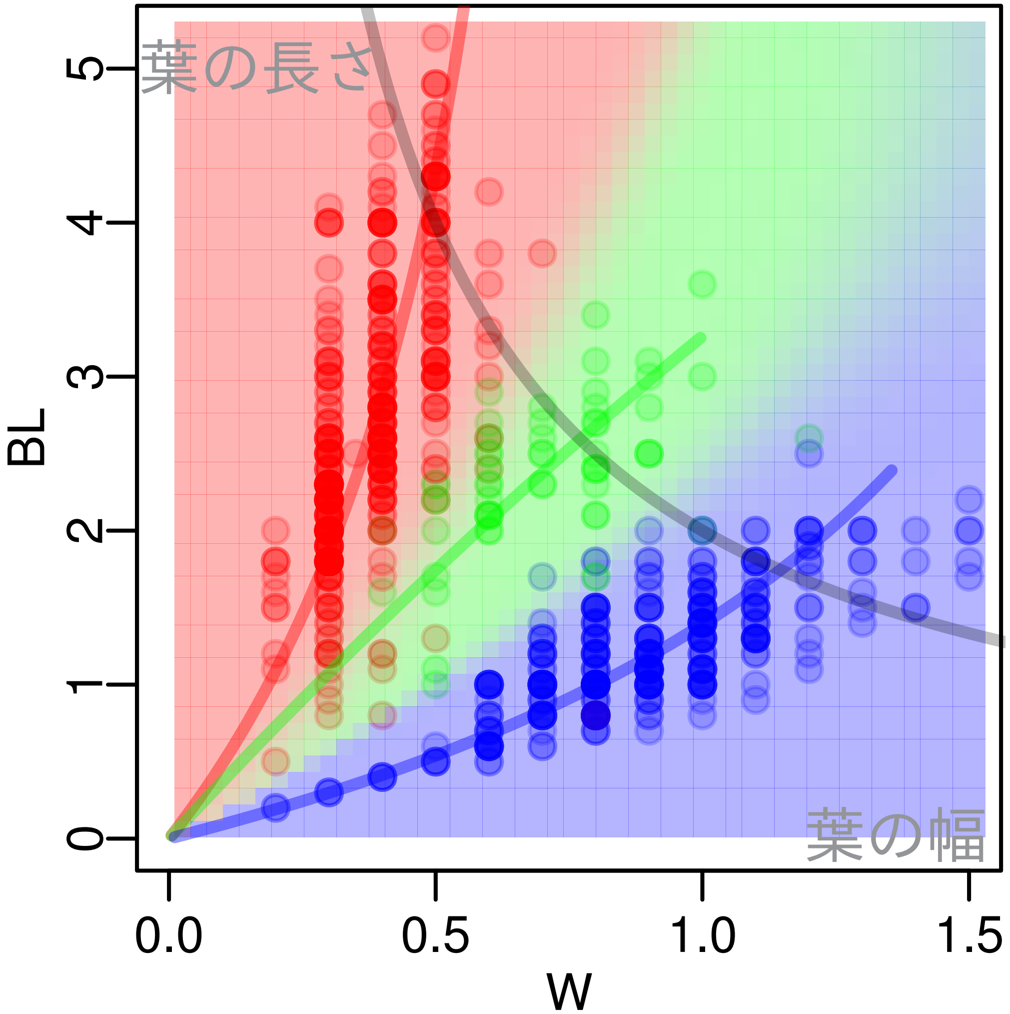
単位長さが $\sqrt{2}$ のときの

強度 v の分布



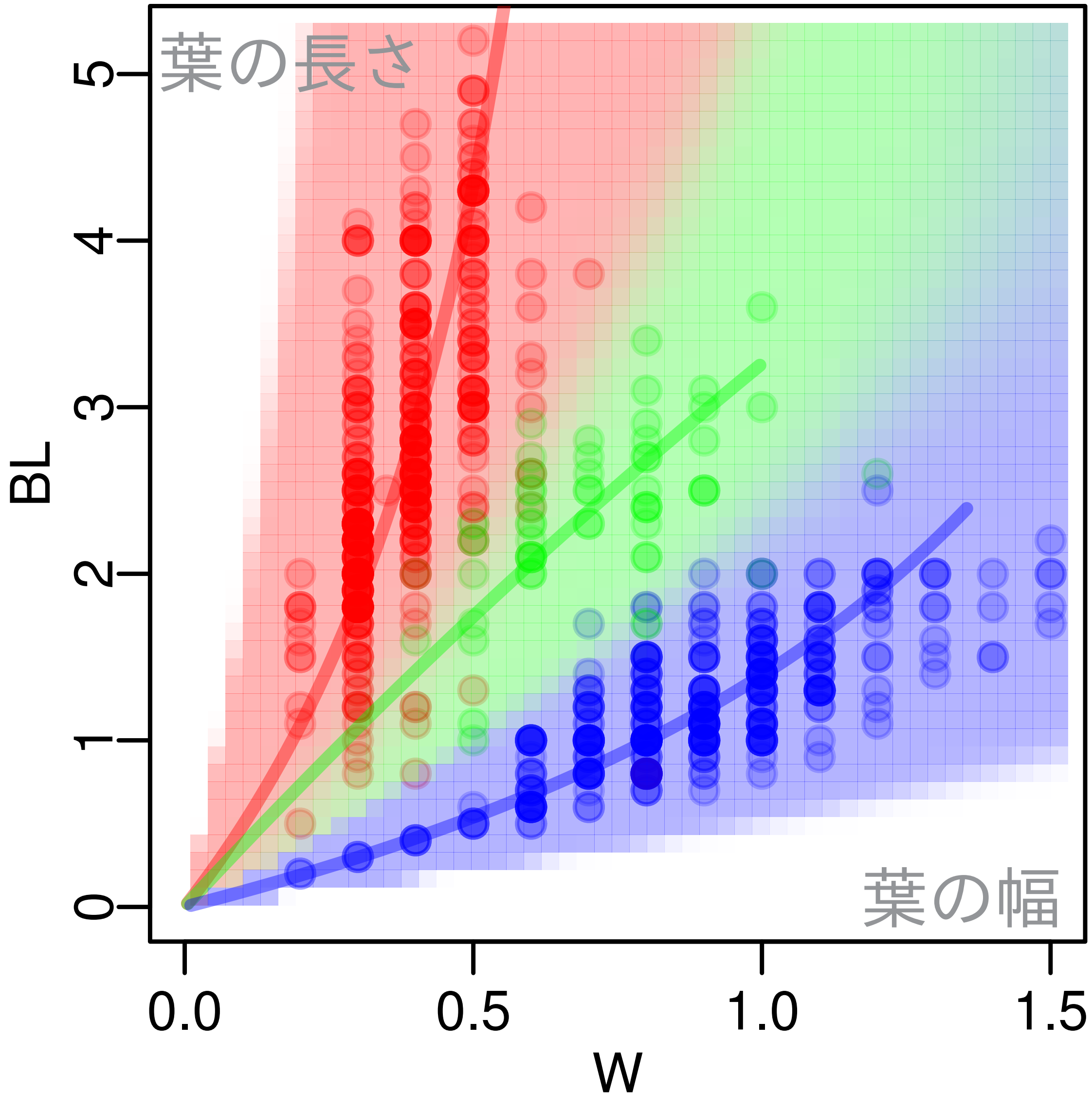
各種が選ばれる確率





- 大きな葉では，3 種の識別は比較的容易

データがないところは除外してみる



- v が小さいところでは識別する能力が低くなるような図示

このあたりで終了