線型モデル、多重比較

2010/10/29

潮雅之

今日の内容

- 線型モデルの基礎を少し
  - One-way ANOVA
  - 単回帰
  - Two-way ANOVA
  - ANCOVA
- 多重比較
- 随時、Rでのplot()の使い方

線型モデル

- $Y_i = \alpha + \beta_1 \times X_i + \varepsilon_i$

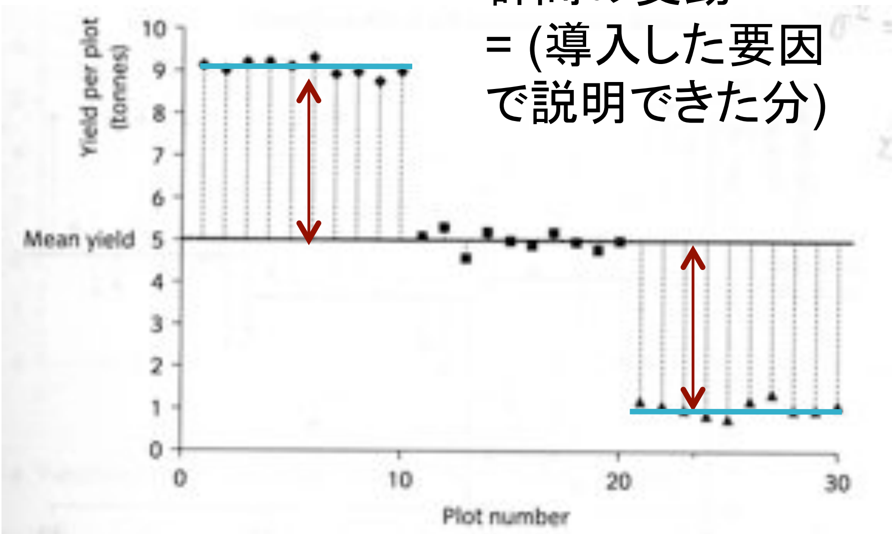
  で表す事のできるものたち
  ただし、ε_i は正規分布

線型モデルの概観

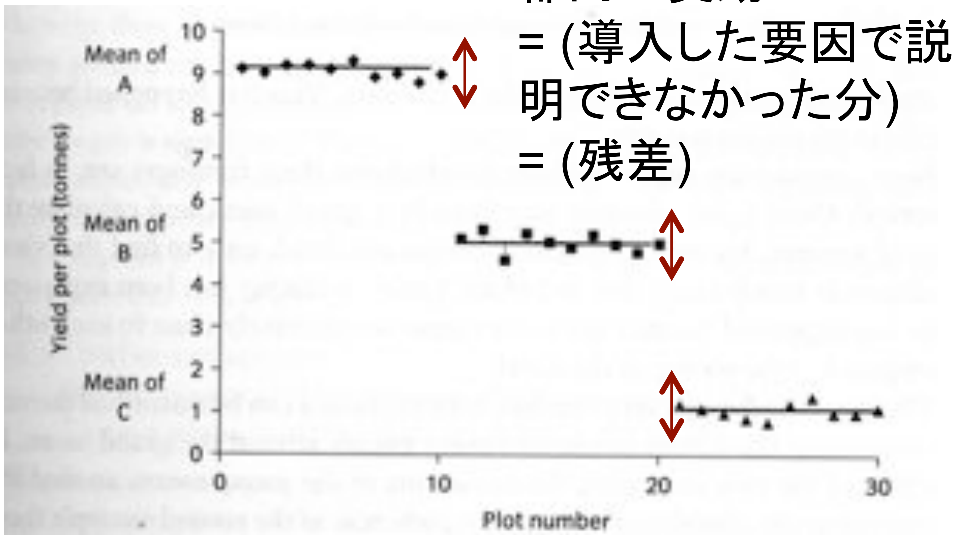
- 被説明変数が連続変数かつ（残差が）正規分布する場合
  - ANOVA (分散分析, Analysis of Variance)
    - 1-way ANOVA (説明変数-1離散変数)
    - 2 (or more)-way ANOVA (説明変数が2変数以上)
  - ANCOVA (共分散分析, 説明変数-1離散変数, 1連続変数)
  - 単回帰 (説明変数が1連続変数)
  - 多重回帰 (説明変数が多数の連続変数)

統計的な背景：ANOVA

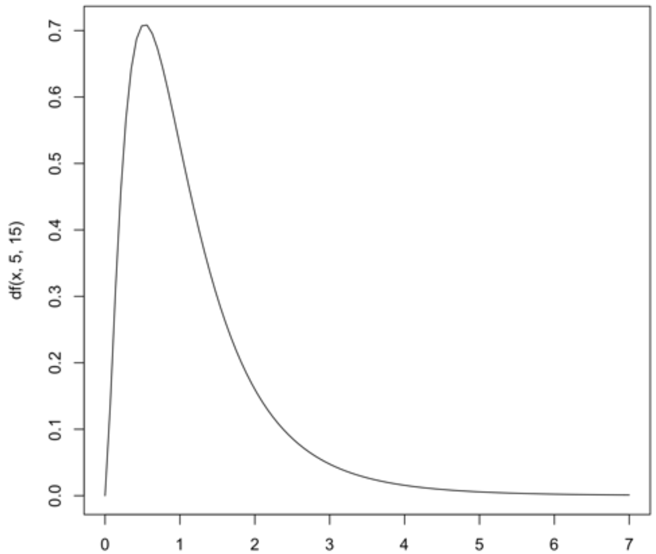
群間の変動
= (導入した要因で説明できた分)



郡内の変動
= (導入した要因で説明できなかった分)
= (残差)



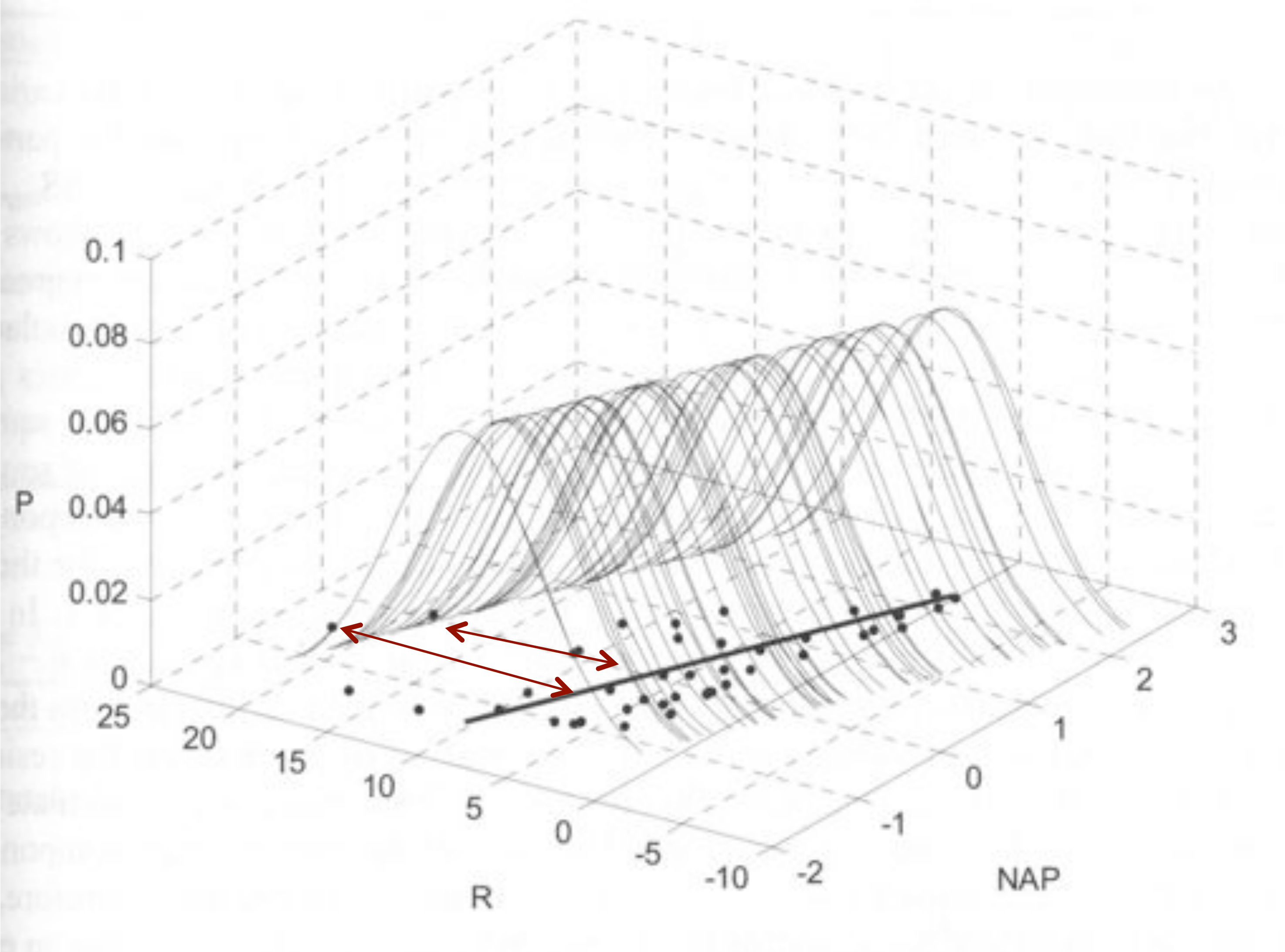
- F値を利用。
- F値 ＝ (群間変動)/(残差)
  →F分布に従う



Grafen & Hails 2002 p.5

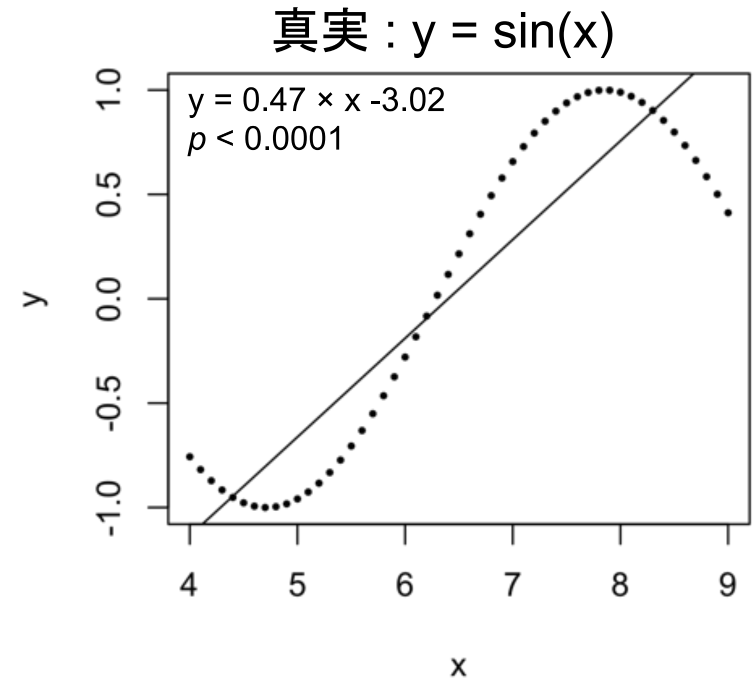
統計的な背景：単回帰

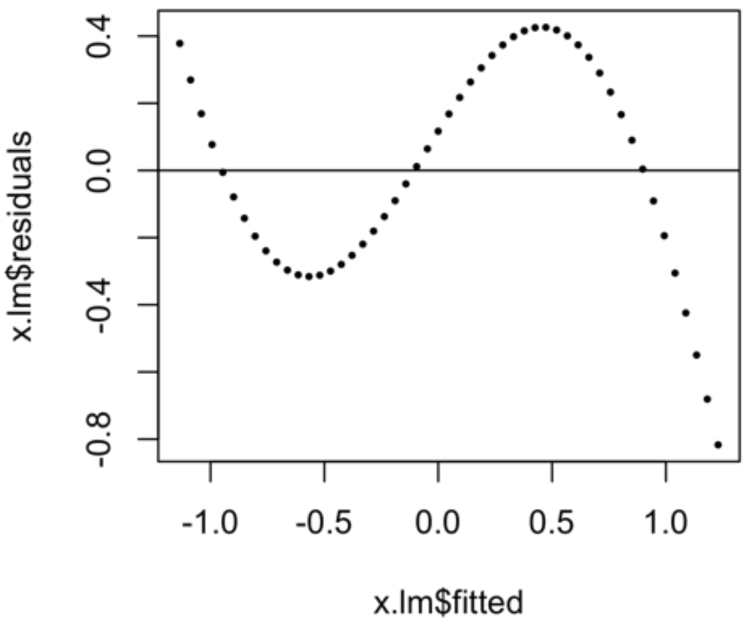
- 残差が正規分布
- （説明できた分散）/（残差）



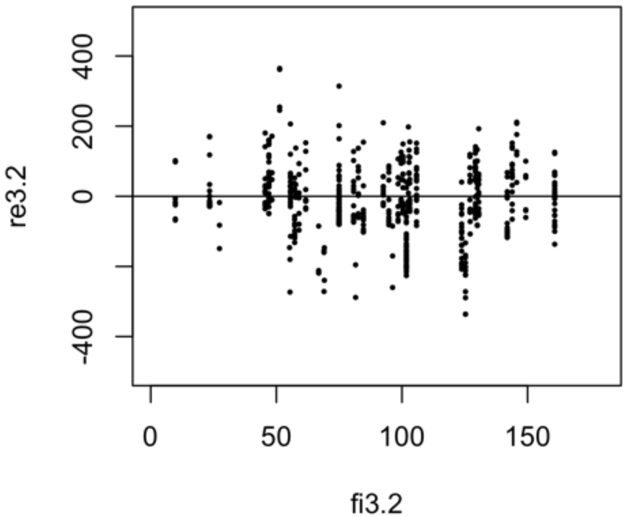
Zuur et al. 2007 p.57

統計的な背景：残差が正規分布



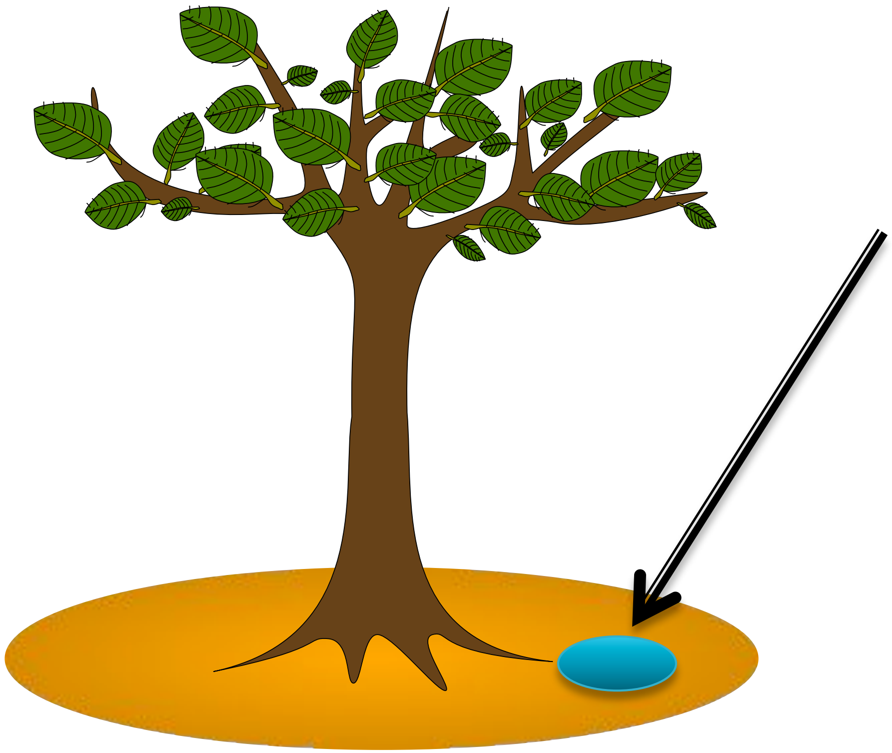


- 残差プロットを見て、パターンが見いだせないことが重要。



で実践

解析例：樹種－土壌特性のデータ



土壌サンプリング

土壌pH
C、N
3種類の酵素活性
(リン酸分解酵素)
(セルロース分解酵素)
(フェノール分解酵素)

解析例：樹種－土壌特性のデータ

針葉樹２種

Dacrycarpus　４反復

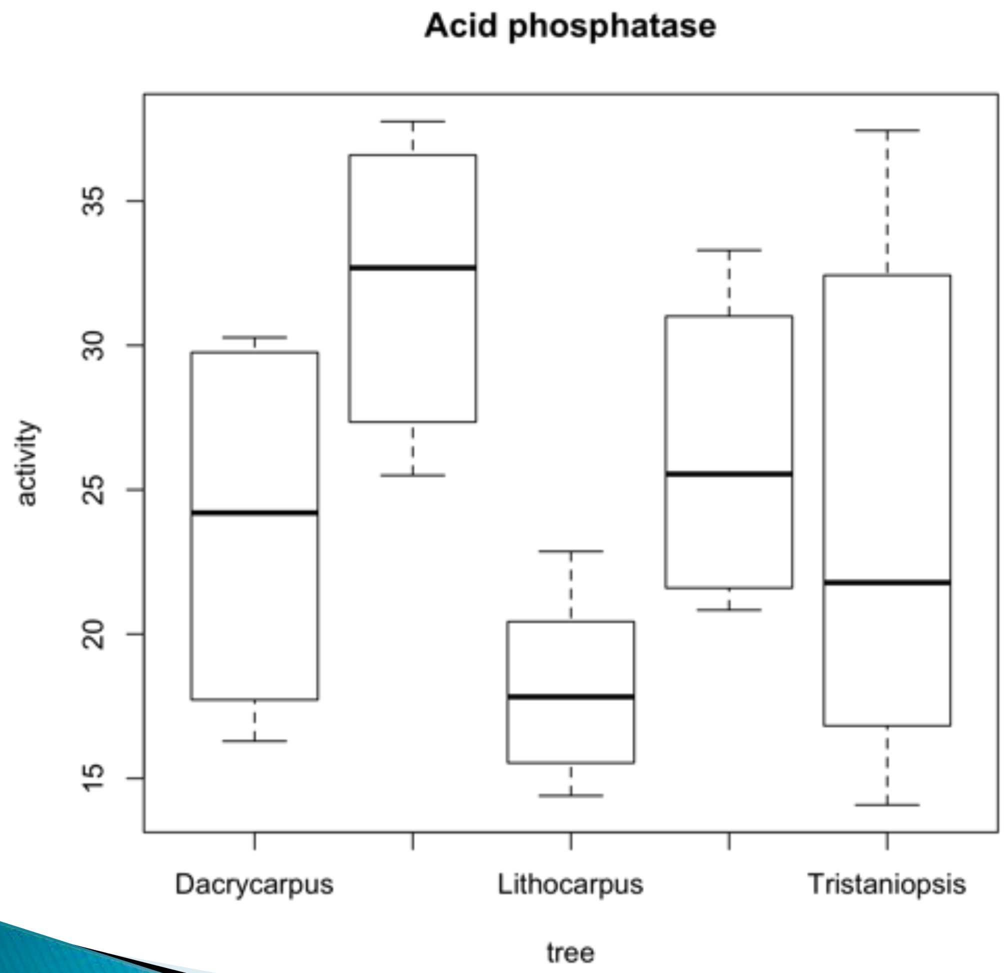
Dacrydium ６反復

解析例：樹種−土壌特性のデータ

広葉樹3種

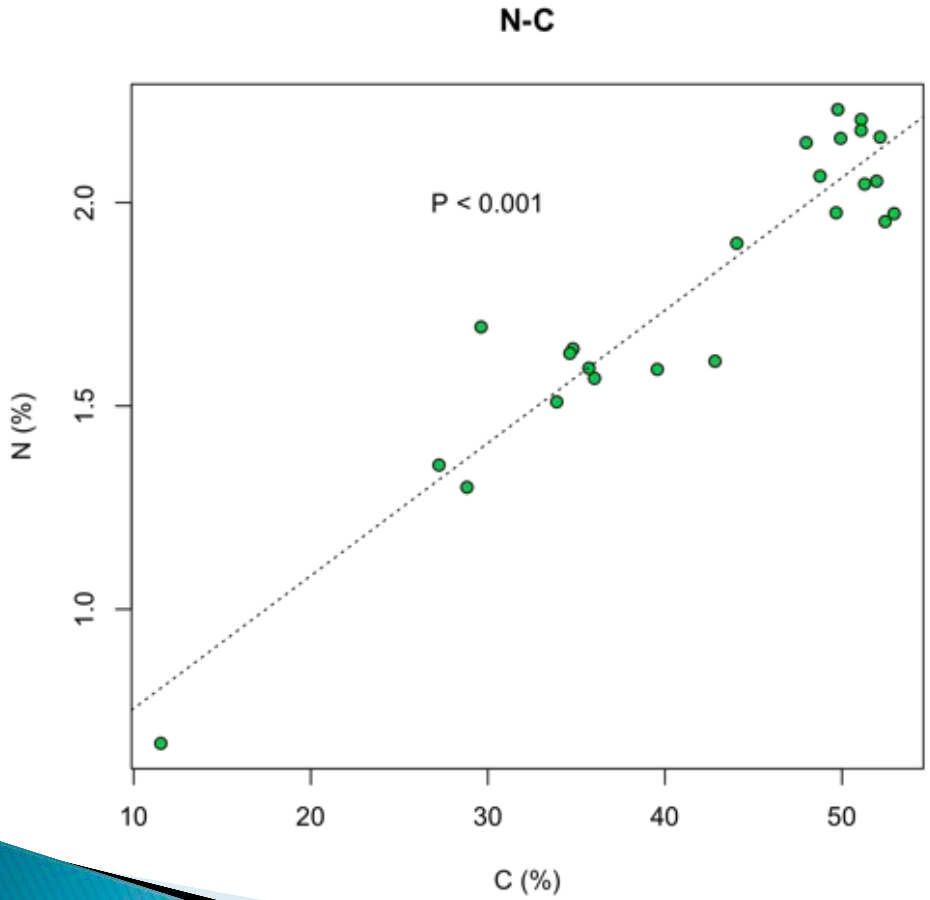
Lithocarpus ５反復　　*Palaquium*　４反復　　*Tristaniopsis*　５反復

1-way ANOVAの例



- 1つの被説明変数を1つの離散変数で説明する。
- F値 ＝ (群間変動)/(群内変動) からp値を算出
- Rでやってみる

単回帰の例



- 1連続変数を1連続変数で説明
- R でやってみる

2-way ANOVAの例

- その前に…
- 交互作用って？

交互作用：数式で書くと、

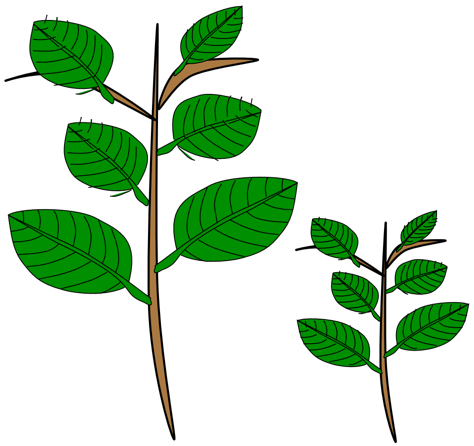
- $Y = \alpha + \beta_1 \times X_1 + \beta_2 \times X_2 + \underline{\boldsymbol{\beta_3 \times X_1 \times X_2}} + \varepsilon_i$

この人のこと

交互作用：言葉で言うと、

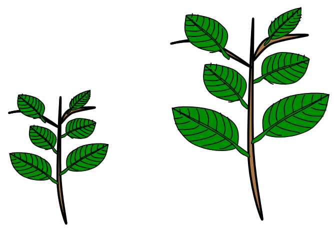
- ある要因によって、もう一つの要因の効果が左右されるか、どうか。
- 例えば、光条件と窒素固定細菌の植物成長への効果。

高光条件下



+根粒菌　-根粒菌

低光条件下



+根粒菌　-根粒菌

中田望さん（首都大学）のポスター発表より

*低窒素濃度の土壌条件下での結果

交互作用：3次以上は説明困難

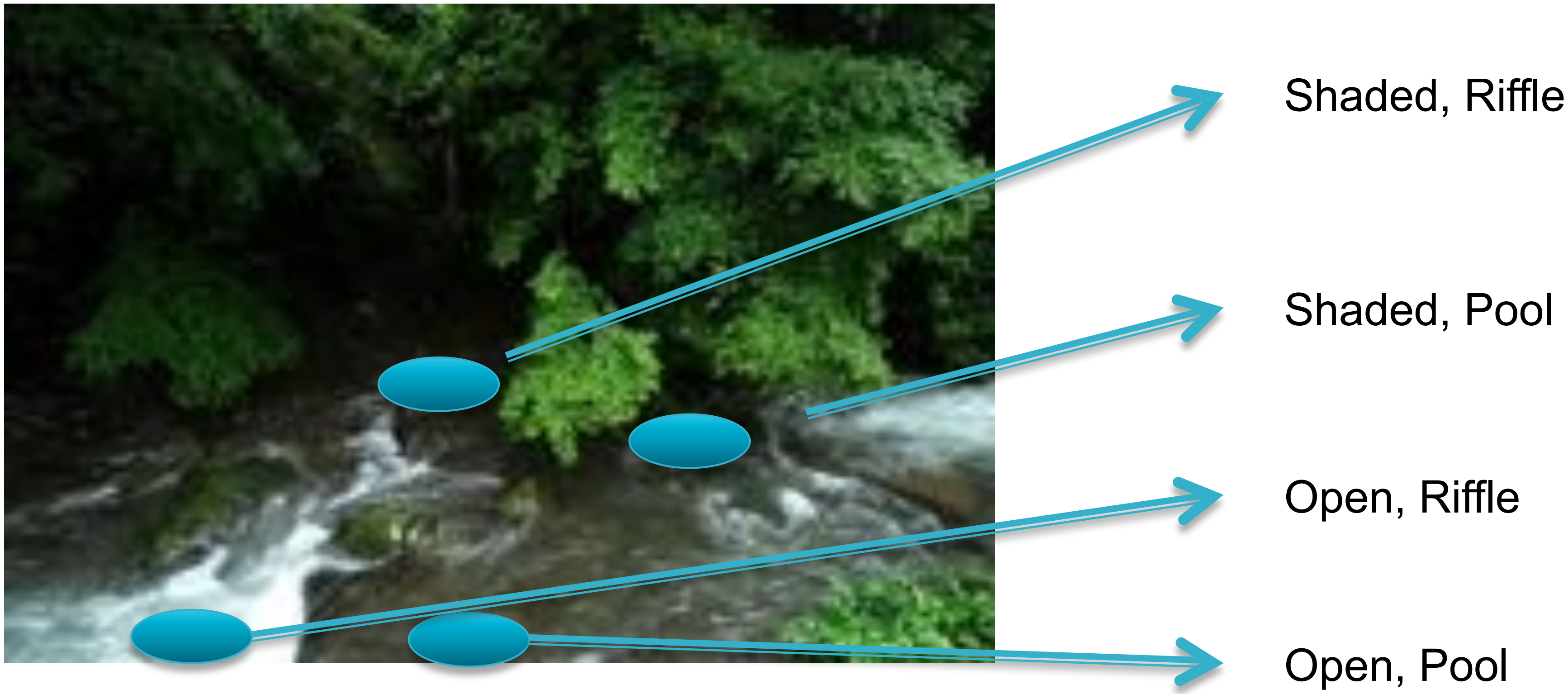
- $Y = \alpha + \beta_1 \times X_1 + \beta_2 \times X_2 + \beta_3 \times X_3$

  $+ \beta_4 \times X_1 \times X_2 + \beta_5 \times X_1 \times X_3 + \beta_6 \times X_2 \times X_3$

  $+ \beta_7 \times X_1 \times X_2 \times X_3 + \varepsilon_i$

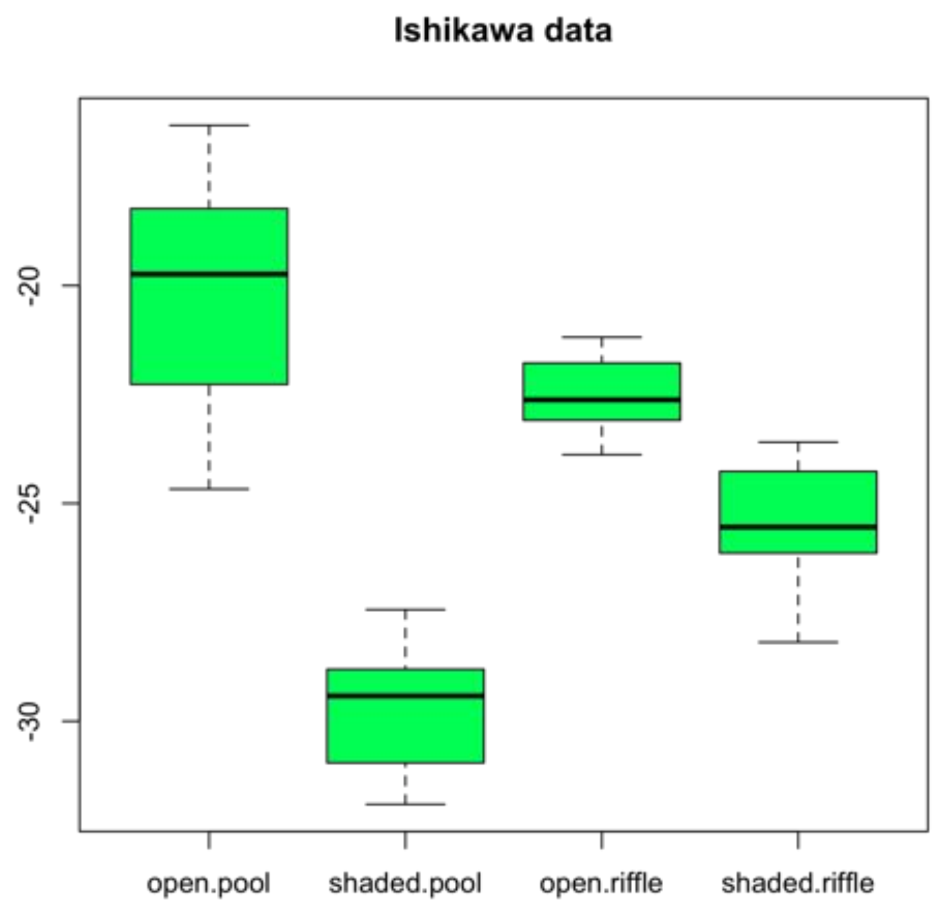
解析例:河川藻類の同位体データ

石川尚人氏よりデータ提供



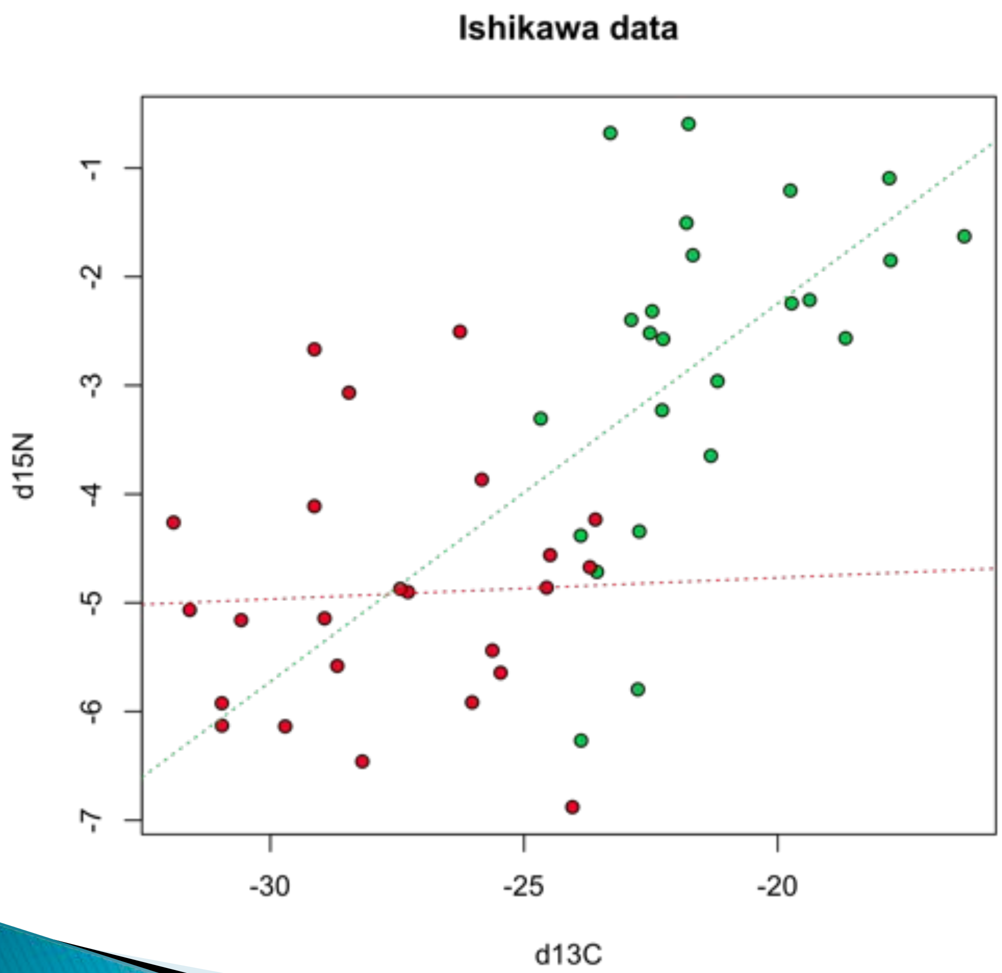
藻類を採取して、δ^{13}C, δ^{15}N, Δ^{14}Cを分析

2-way ANOVA



- 2離散変数で説明。
- R でやってみる。

ANCOVAの例

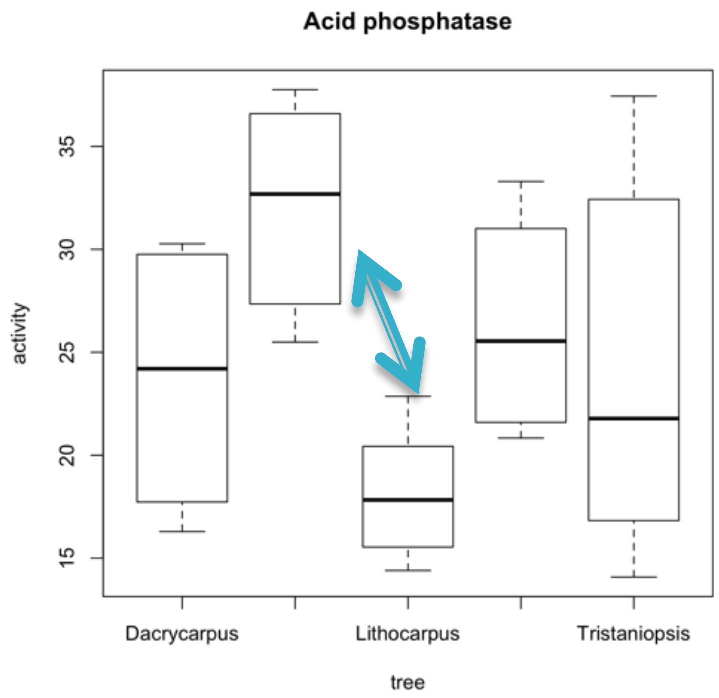


- 1連続変数と1離散変数で説明。
- Rでやってみる。

多重比較

- ANOVAでは、ある要因の効果があるか、ないか、しか分からない。

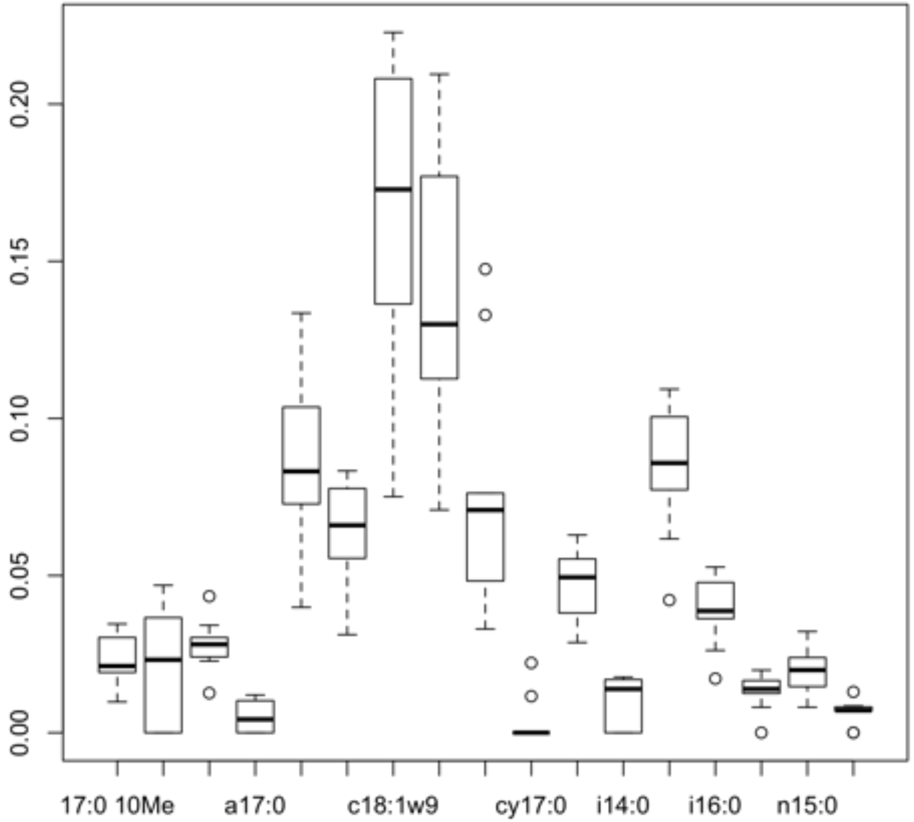
- ある処理区とある処理区の間に差があるかどうかを知りたいときは、多重比較法



なぜ? 検定の多重性の問題

- 同じ検定を何度も繰り返してはだめ。
- 数打ちゃ当たる、はだめ。

- *P* = 0.05を設定して、同じ統計を10回繰り返す。
- 一回でも誤って有意と言えないものを有意とする確率は$1\text{-}0.95^{10} = 0.40$
- という訳で補正が必要。



パラメトリックとノンパラメトリック

- パラメトリック (parametric)
  - 特定の分布を仮定して検定を行なう。パラメトリック（parametric）とは母数（parameter）を使っているということ。母数とは、それがあれば分布の形を決定できる数のことで正規分布の場合は平均と分散
- ノンパラメトリック (non-parametric)
  - 母数はもはや関係ない。分布の形は何でも良い。ただし、一般的に有為差は出にくい。

一番お手軽：Bonferroni補正

- 理論が単純明快。
- 割り算しただけ。
- 普通$p = 0.05$を有意とするが、p値をこなした検定の数だけ割ることで補正する。たとえば10回検定を行なえば、$0.05/10 = 0.005 > p$ のとき有意と判定する。

- いろんな場面で使われる。

- 有為差は<u>出にくい</u>。

Tukey's HSD

- 正規分布、等分散したデータに使用できる。
- 異名がたくさん
  - Tueky-Kramer法、Tukey法、Tukey's HSD
  - Tukey's a, b, qとか。

- もっともよく使用される多重比較法のひとつ。
- わりと有為差が出やすい。
- ちょっとRでやってみる。

多重比較法のいろいろ

条件	分布	分散
パラメトリック		
Tukey’s HSD	正規分布	等分散
Games-Howell	正規分布	不等分散
ノンパラメトリック		
Steel-Dwass	制限なし	制限なし
Scheffe	制限なし	制限なし
Dunnett	制限なし	制限なし

まだまだ色々ある。→『統計的多重比較法の基礎』参照

参考文献、Website

- Analyzing Ecological Data. Zuur et al. (2007)
- Modern Statistics for the Life Sciences. Grafen and Hails (2002)
- 『統計的多重比較法の基礎』 永田靖、吉田道弘　著 (1997)
- 解説PDF (http://www.hs.hirosaki-u.ac.jp/~pteiki/research/stat/multi.pdf)→わりとまとも