環境統計学ぷらす

第2回
分散分析と回帰

高木　俊
shun.takagi@sci.toho-u.ac.jp

2013/10/31

予定

- 第1回：　Rの基礎と仮説検定
- 第2回：　分散分析と回帰
- 第3回：　一般線形モデル・交互作用
- 第4回：　一般化線形モデル・モデル選択
- 第5回：　一般化線形混合モデル
- 第6回：　多変量解析

今日やること

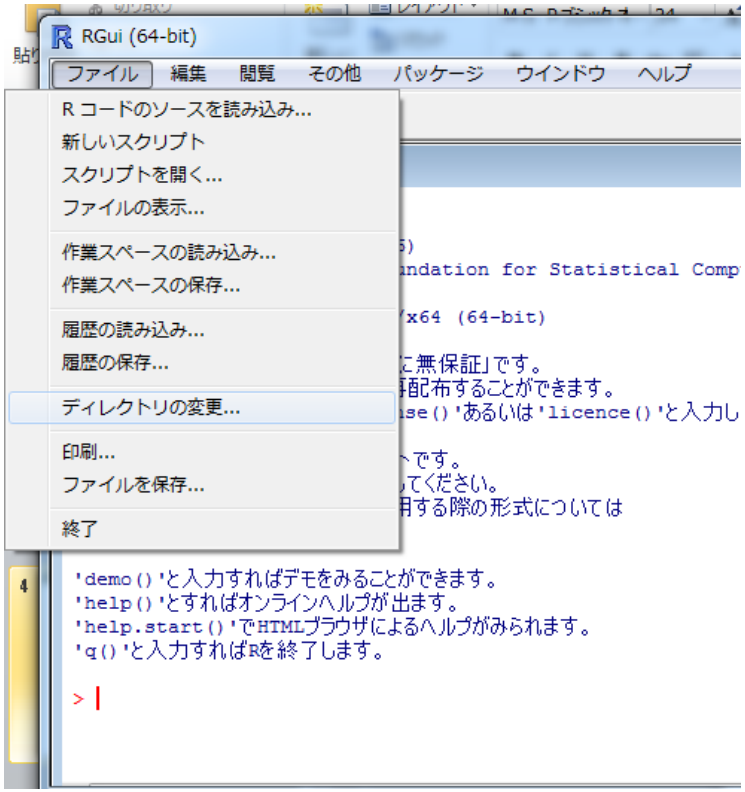
- R操作編
  - RエディタからのRの実行
  - データフレームの操作
- 統計編
  - 分散分析
  - 回帰
- 表現編
  - plotのオプション関数たち
  - エラーバー付き棒グラフ
  - 分散分析表

Rの操作

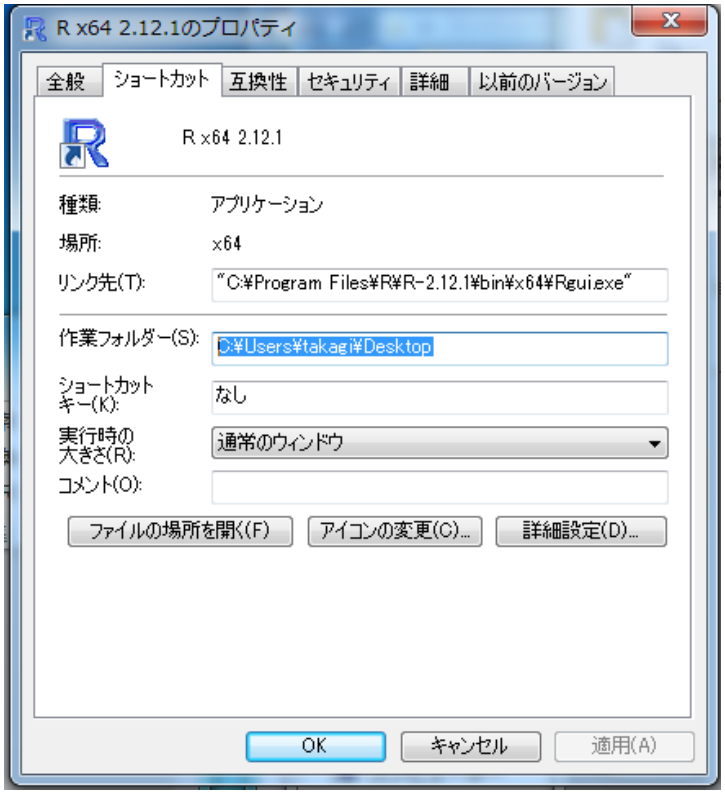
- Rエディタからの実行
- データフレームの操作

作業の前に･･･作業ディレクトリの設定

- Rを起動し、
  ファイル＞ディレクトリの変更



- Rショートカットのプロパティ
  作業フォルダーにパスを入力

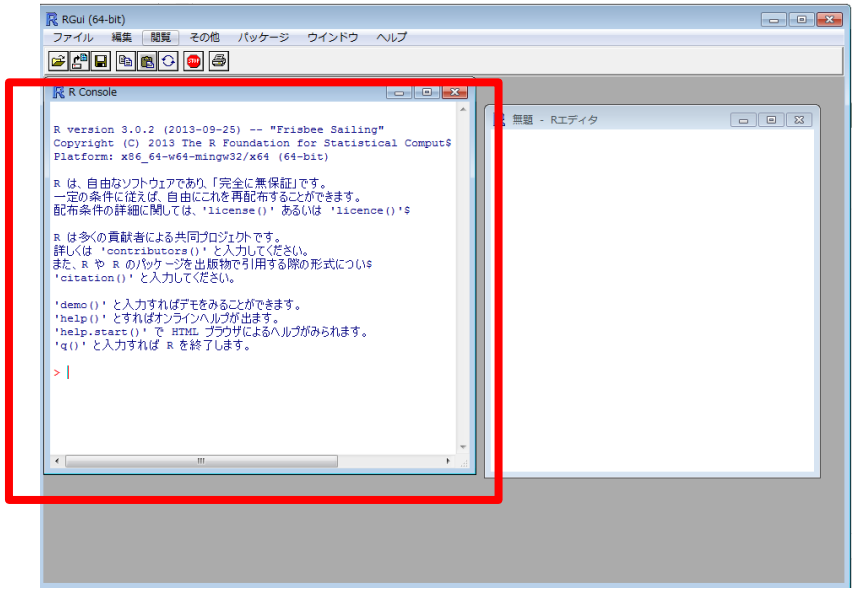


作業ディレクトリの確認は　getwd()　で

Rでのスクリプトの実行

- R上で計算させるには、コンソールウィンドウにコマンドを打ってやれば良い・・・が、

コンソールウィンドウ

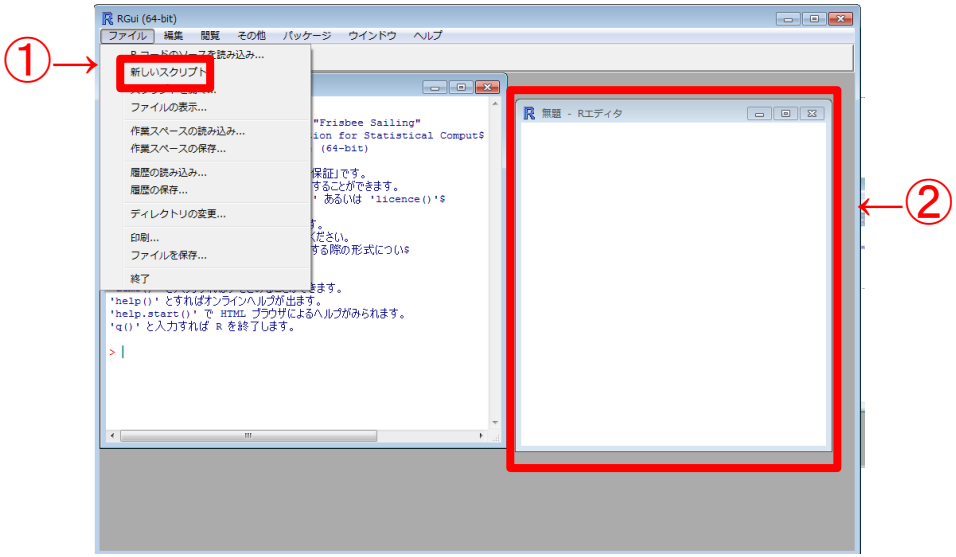


- コマンドが長いとミスしやすい
- 読み返しにくい
- 後から編集しにくい

など非常にストレスがたまる

Rエディタの利用

1. ファイル＞新しいスクリプト　でRエディタを呼び出してやる



2. Rエディタにコマンドを書いて、「Ctrl+R」を押すとコンソールウィンドウにコマンドが送られる
3. カーソル行のコマンドもしくは、選択範囲のコマンドが実行可

スクリプトの保存

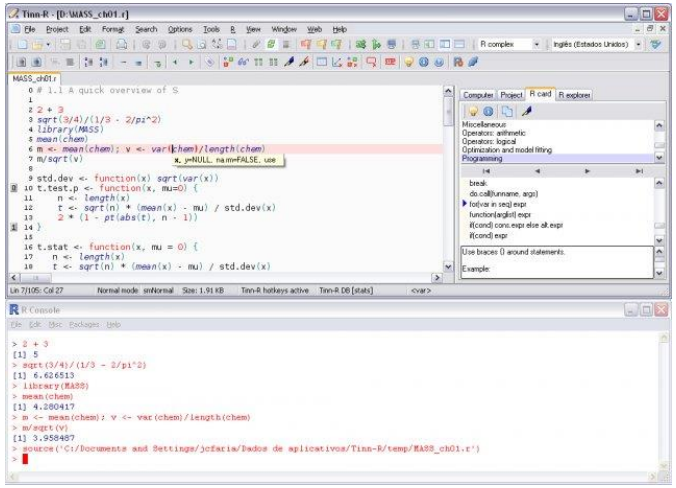
- 書いた内容は保存（拡張子.R）しておき、再解析や修正したいときに呼び出せる

- ファイル名をkekka.new.Rとかにすると、よくわからなくなることが多いので、日付を入れておくのがおすすめ（analysis.20131026.R　など）

テキストエディタの利用

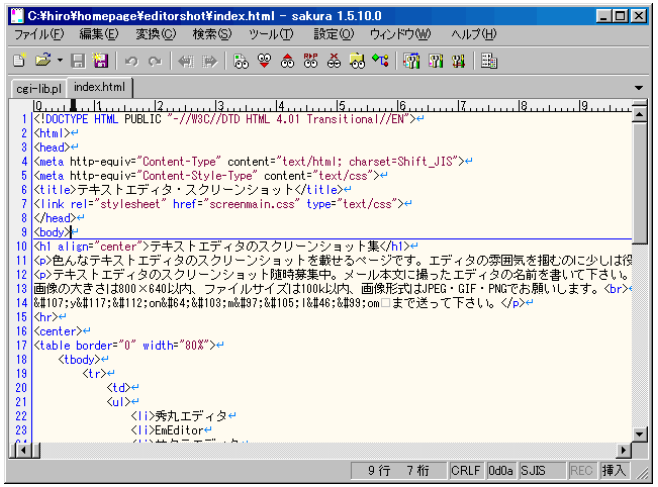
- Rエディタはwindowsのメモ帳程度の残念な機能しかないので、長いスクリプトを書くにはあまり向かない
- 頻繁に使う人は、矩形選択・対カッコ色表示・置換・タブ表示などができるテキストエディタの利用がおすすめ

「Tinn-R」



Rに特化・「Ctrl+R」でR実行可能

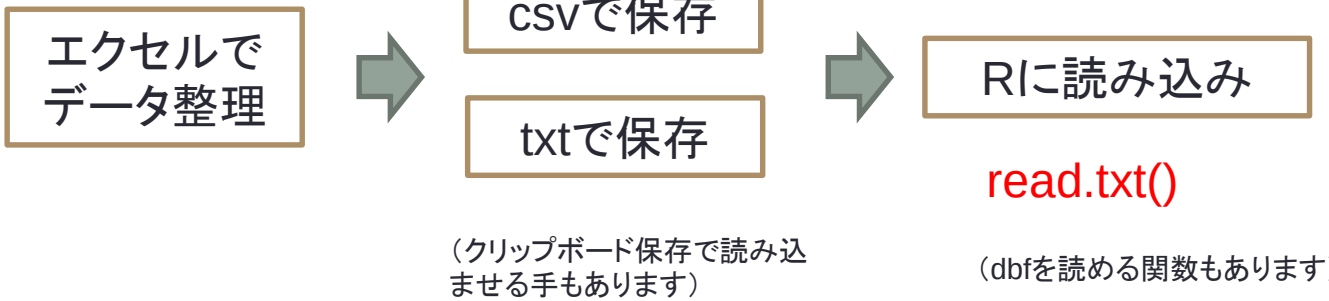
「サクラエディタ」

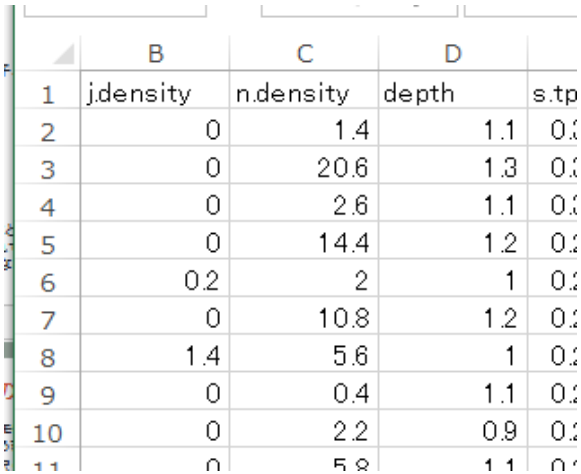


直接のR実行不可だが、機能充実アイコンがおしゃれ

データの読み込み

- エクセルで整理したデータをRに読み込んで解析





	B	C	D	
1	j.density	n.density	depth	s.tp
2	0	1.4	1.1	0.3
3	0	20.6	1.3	0.3
4	0	2.6	1.1	0.3
5	0	14.4	1.2	0.2
6	0.2	2	1	0.2
7	0	10.8	1.2	0.2
8	1.4	5.6	1	0.2
9	0	0.4	1.1	0.2
10	0	2.2	0.9	0.2
11	0	5.8	1.1	0.2

エクセルでのデータのまとめ方

- 1行目にデータの名前
- 2行目以降の各列にデータを縦に入れる
- データにはスペースを入れない
- 空白セルにはNAを入れておく
- データ名は数字で始めない（なるべく）
- データに#を入れない

データフレーム

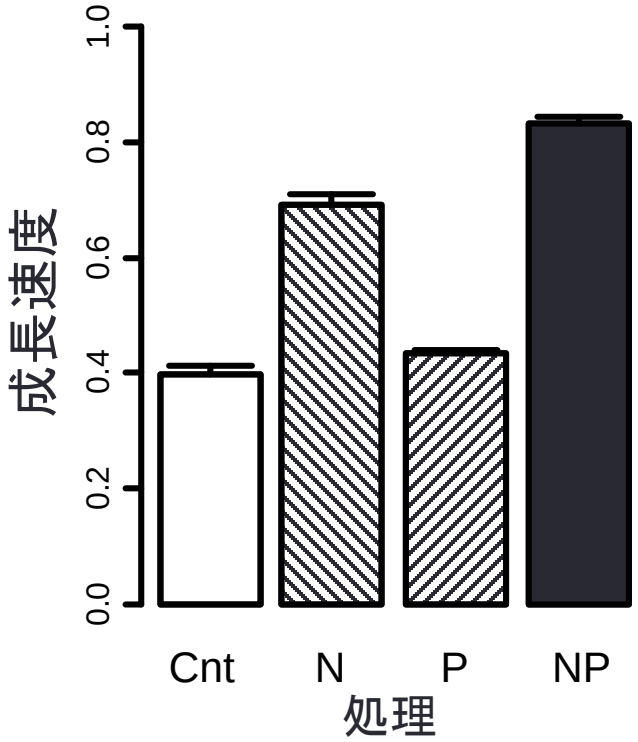
- 読み込んだデータはデータフレームという形で扱われる

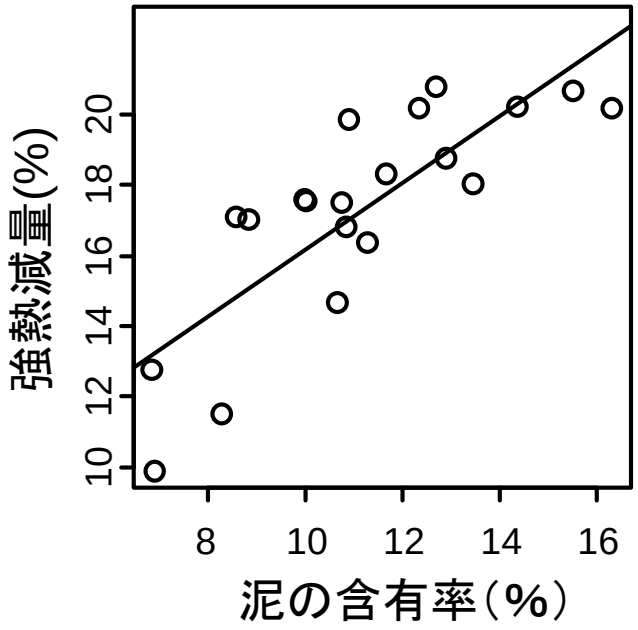
```
data2.1<- read.csv(“data2.1.csv”,header=TRUE)
data2.1<- read.csv(“data2.1.csv”,T)  #実は上と同じように処理される
data2.1
   station.no j.density n.density depth
1           1       0.0       1.4   1.1
2           2       0.0      20.6   1.3
3           3       0.0       2.6   1.1
4           4       0.0      14.4   1.2
5           5       0.2       2.0   1.0 (略)
data2.1$depth  #各列へのアクセスは“データフレーム名$列名”で可能
[1] 1.1 1.3 1.1 1.2 1.0 1.2 1.0 (略)
mean(data2.1$depth)    #depthの平均を求めたければ
[1] 1.06
```

復習：データフレームdata2.1のdepthの標準誤差（SE）を求めるには？

分散分析
回帰

分散分析と回帰





分散分析

説明変数Xの種類（カテゴリー値）によって
応答変数Yがどのような値をとるか

回帰

説明変数Xの変化（連続値）に応じて
応答変数Yがどのように変化するか

説明変数Xのタイプが異なる

3群以上の平均値の差の検定（分散分析）

- 説明変数Xの種類（カテゴリー値）によって応答変数Yがどのような値をとるかを分析
- 普通棒グラフでその関係性を示す（とりあえずの傾向を見るなら箱ひげ図でも）。

Rではplot(Y~X)で実行

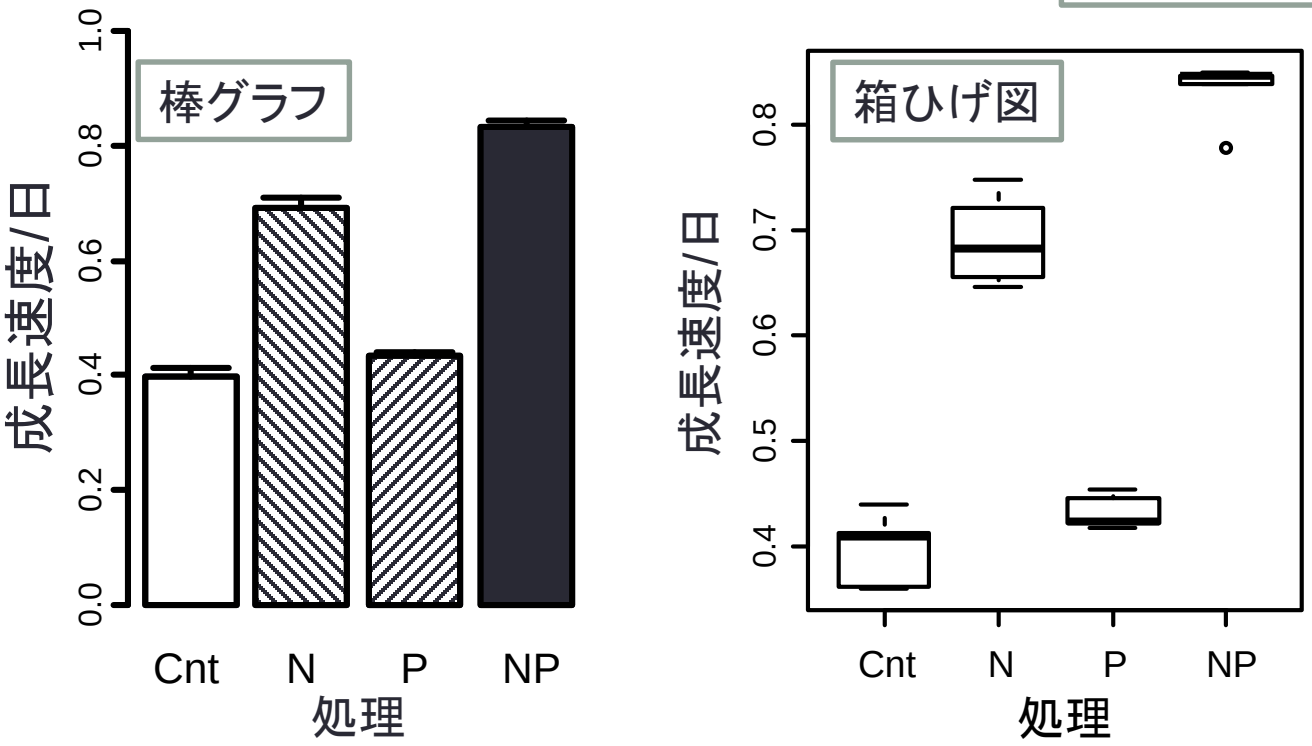


図2. 実験処理ごとの植物プランクトンの成長速度

分散分析の前提条件

t検定と同様に、母集団の正規分布を仮定しているので、以下の条件を満たす必要あり

正規性	各群のデータ（の母集団）が**正規分布に従う**こと
等分散性	各群のデータ（の母集団）の**分散が等しい**こと
独立性	個々のデータは**互いに独立**であること

正規分布を仮定しないノンパラメトリック法における代替法としては、

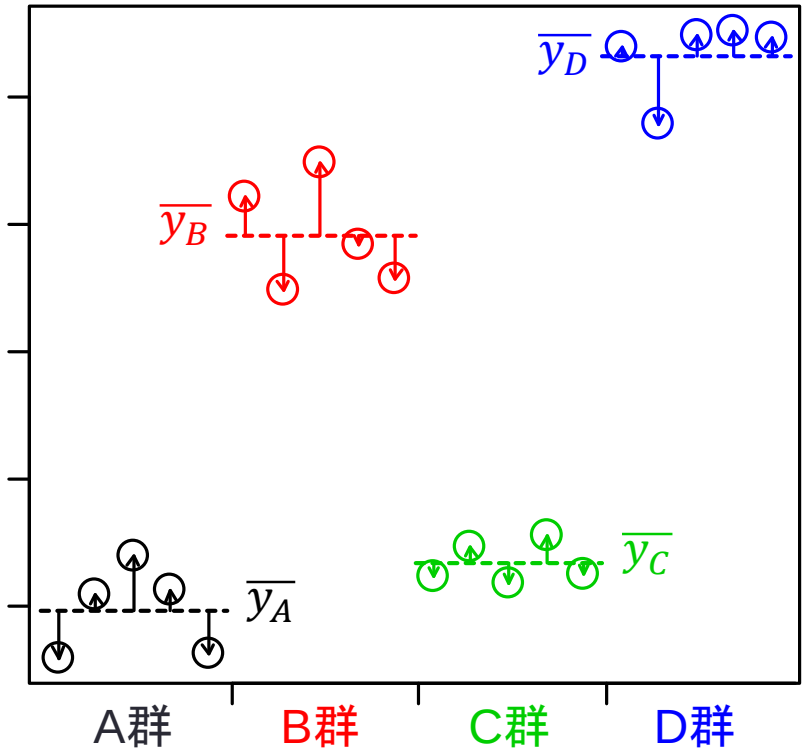
- Kruskal-Wallis検定（対応なしの場合）　kruskal.test()
- Friedman検定（対応ありの場合）　friedman.test()

がある

線形モデルへのあてはめ

i群j番目の観測値をy_{ij}と表記

$$\text{観測値}y_{ij} = \text{各処理の平均値}\overline{y_i} + \text{誤差}_{ij}$$



こいつが処理ごとに
大きくばらついていて

こっちのばらつきが
十分に小さければ

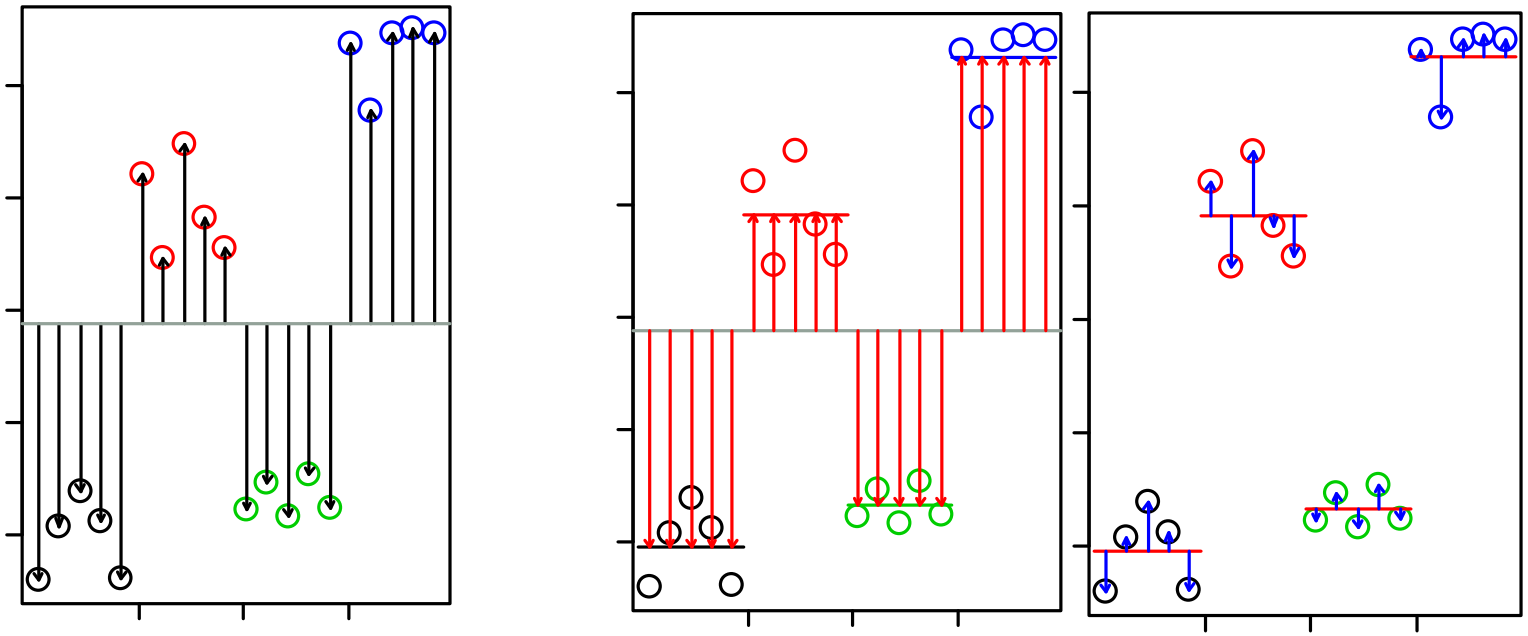
処理の効果があると言って
良さそう

ばらつきを分割する

観測値のばらつき　群間のばらつき　群内のばらつき＝誤差

$$SS_Y = SS_{Group} + SS_{Error}$$

$$\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij}-\bar{y})^2 \quad \sum_{i=1}^{p} n_i(\bar{y}_i-\bar{y})^2 \quad \sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2$$



分散を分析する

平方和SSはデータ数が多いほど大きくなる指標なので分散を用いて比較

（不偏）分散： $s^2 = \frac{\text{平方和}}{\text{自由度}} = \frac{\sum(y_i - \bar{y})^2}{n-1}$

p: 処理の種類数
n: 処理内の繰り返し

	平方和 SS	自由度 df	平均平方 $MS = \frac{SS}{df}$
処理	SS_G	p-1	$SS_G/(p-1)$
残差	SS_E	np-p	$SS_E/(np-p)$
全体	SS_Y	np-1	$SS_Y/(np-1) = s^2$

*F*比（*F*値）の計算

処理の平均平方（群間の分散：MS_G）と
残差の平均平方（郡内の分散：MS_E）の比をとったものがF比

	平方和 SS	自由度 df	平均平方MS（＝SS/df）	*F*
処理	SS_G	p-1	MS_G	$\frac{MS_G}{MS_E}$
残差	SS_E	np-p	MS_E	

帰無仮説（処理間での成長率に差はない＝処理間分散はゼロ）
のもとでF比は自由度p-1,np-pの*F*分布の従う

分散分析表（ANOVA Table）

- 分散分析の結果は下記のような分散分析表に表す

	平方和 SS	自由度 df	平均平方MS（＝SS/df）	F
処理	0.654	3	0.218	218
残差	0.017	16	0.001	

$$P(F_{3,16} \geqq 218) = 3.49 \times 10^{-13}$$

Rで下の式を入れれば計算されます
1-pf(218,3,16)

よって、帰無仮説は棄却され、
「処理間で成長率には違いがないとはいえない（≒違いがある）」

Rで分散分析

lm()とanova()を用いる　　※aov()関数でもできます

```
> model<- lm(growth~trt,data2.2)
> anova(model)
Analysis of Variance Table

Response: growth
          Df  Sum Sq Mean Sq F value    Pr(>F)
trt        3 0.65422 0.21807   205.8 5.473e-13 ***
Residuals 16 0.01695 0.00106
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.
```

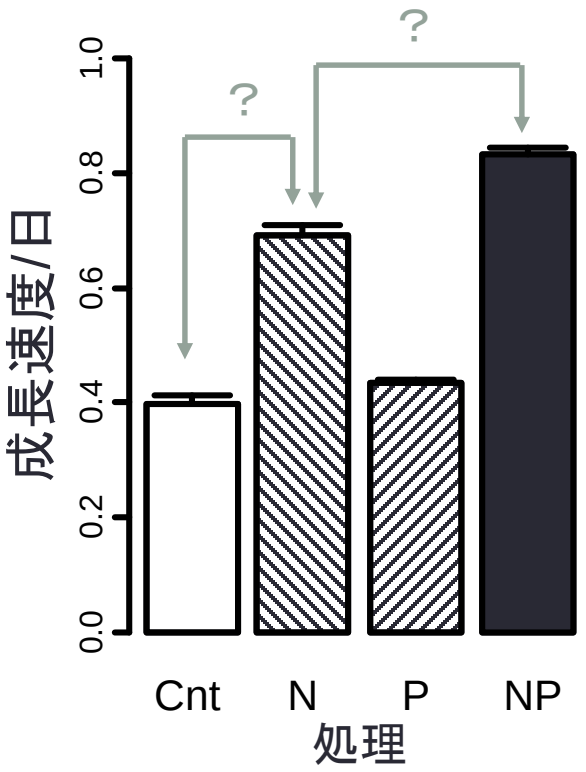
#lm()関数を用いてモデル式を建てる
#anova()で分散分析

応答変数
説明変数
自由度　平方和　平均平方　F比　P値(F比に基づく)
残差

→「処理間で成長率には違いが見られた($F_{3,16}$=205.8, P<0.001)」

群間での多重比較

「処理間で成長率には違いが見られる」はわかったが、
「どの処理とどの処理の間で具体的に差があるか」はわからない



そこで2群ずつ検定してやる

2群の差の検定→*t*検定

4群間の比較は $_4C_2=6$ 通り

6回*t*検定すればすべての組み合わせでの差がわかる

多重検定の落とし穴

統計学的有意性：帰無仮説のもとでデータより極端な値が得られる確率が十分に低い（α=0.05）

α＝0.05　の基準で検定した場合、

20回に1回は差がないものを差があると判断してしまう

（Type-I Error・第一種の過誤）

一連の仮説の中で、同じような検定を何度も繰り返す（多重検定）と、どこかでType-I Errorを起こす確率が高い！

一連の検定のいずれでもType-I Errorを起こさないようにα（もしくはP値）を修正する必要がある

→Family-wise Type-I Errorを調節する

Family-wise Type-I Errorを考慮した多重比較

- TukeyのHSD法

  統計の詳細は省略　　`TukeyHSD(aov(Y~X,data))`

- Bonferroni法

  n回多重比較するなら、P値をn倍（αを1/n）しちゃえという発想

  `pairwise.t.test(Y,X,”bonferroni”)`

- Sequential Bonferroni（Holm）法

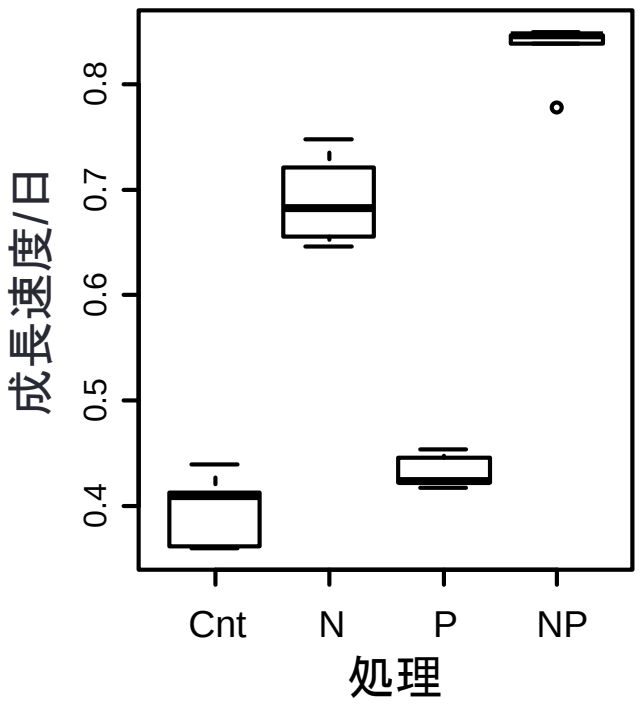
  Bonferroniは厳しすぎて差があるものをないと言ってしまう危険（Type-II Error）が高いので、差が大きいものから検定を行い、補正を厳しいものから徐々に緩める方法

  `pairwise.t.test(Y,X,”holm”)`

分散分析まとめ

目的：Xのカテゴリー間でのYの差の検定

原理：カテゴリー間分散とカテゴリー内分散（誤差）に分割して
その比をみる



箱ひげ図を書いてみる
（または棒グラフ）
`plot(Y~X)`

モデル式を当てはめる
`model<- lm(Y~X)`

分散分析からXの
効果を判断
`anova(model)`

必要なら多重比較
`TukeyHSD()`
`pairwise.t.test()`

線形回帰 (単回帰 simple linear regression)

- 変数Y(従属変数・応答変数)の変数X(独立変数・説明変数)に対する線形な関係を解析

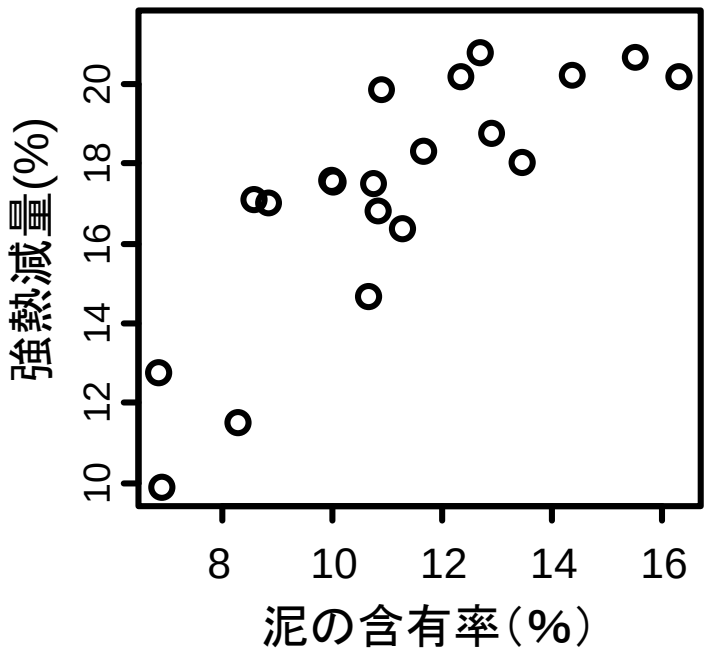


図1. 強熱減量と土壌中の泥の含有率の関係

- Xが高いほどYが高い(低い)といった単調増加/単調減少の関係を見たい場合の解析
- Yに対してXが与える影響を見る(逆の因果関係はダメ)
- 普通、左のような散布図を書いてから解析する

RではplotY~X)で実行

左の散布図だと、有機物を多く含む泥の含有率が高い土壌ほど、強熱減量が高くなるという因果関係を想定

実験では、実験者がコントロールできる要因をx軸(温度設定など)、測定対象をy軸(成長率など)におく

回帰の前提条件

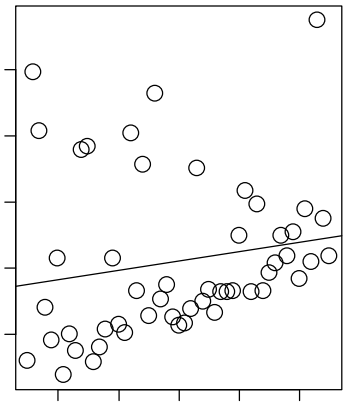
やはり母集団の正規分布を仮定しているので、以下の条件を満たす必要あり

正規性 回帰からのばらつきが**正規分布に従う**こと

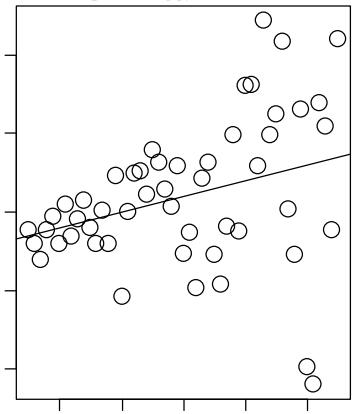
等分散性 個々のデータで回帰からの**分散が等しい**こと

独立性 個々のデータは**互いに独立**であること

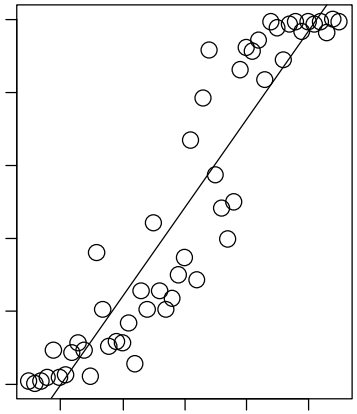
正規性×



等分散性×



等分散性×



（%で示す指標など）上限・下限にぶつかっているデータ
→逆正弦変換

$$Y = \sin^{-1}\sqrt{y}$$

```
asin(sqrt(y))
```

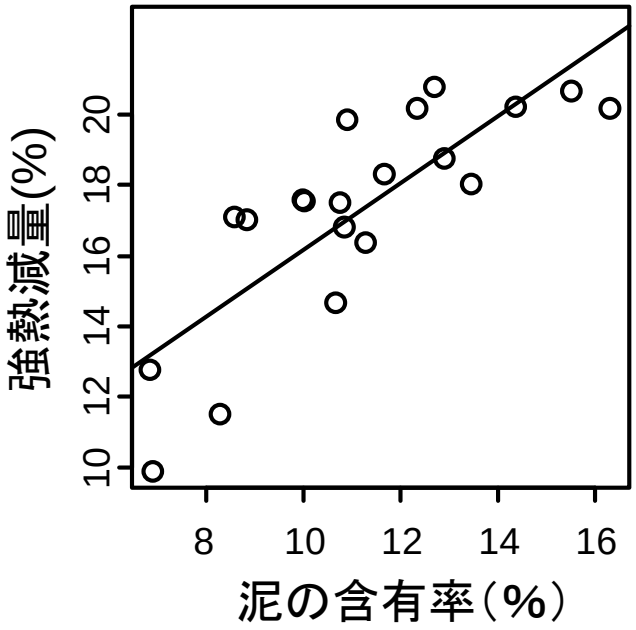
ただし、死亡率など整数/整数のデータは一般化線形モデル（次々回？）であてはめること

モデルへの当てはめ

- 地点iにおける強熱減量の泥含有率に対する関係を式に表すと

$$\text{強熱減量}_i = \text{切片} + \text{傾き} \times \text{泥の含有率}_i + \text{誤差}_i$$

（一般的な式にすると：$y_i = \beta_0 + \beta_1 \times x_i + \varepsilon_i$）

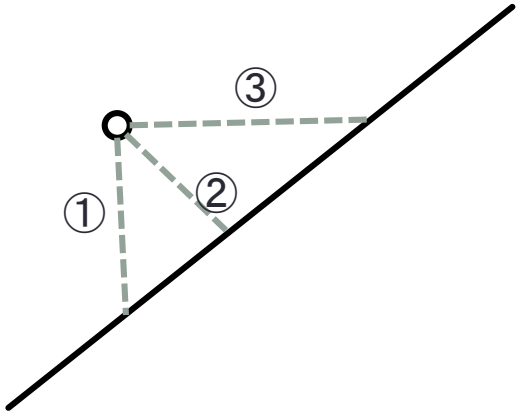


回帰分析は、回帰直線の
切片（β_0）と傾き（β_1）を
誤差が最小になるように推定

回帰における誤差

- 回帰直線からのデータのずれ（誤差）が最小になるように推定

誤差ってどこ？



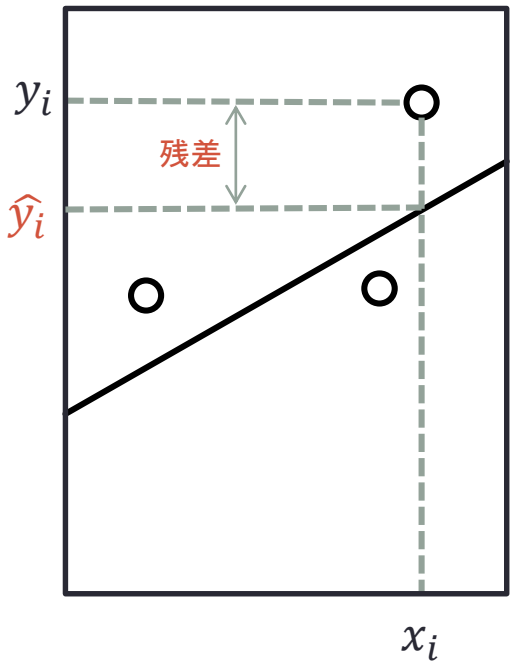
① y軸方向のずれ
② 回帰直線との最短距離
③ x軸方向のずれ

⇒正解は①
どの誤差を最小化するかで、回帰直線は異なる

回帰はxに対するyの反応を見たい解析なので、
x側は実験者が設定しているので誤差は生じない、
y側には反応の誤差が含まれる　と想定

最小二乗法～誤差を最小化

- 地点i のデータにおける回帰直線からの誤差（残差）の二乗は、



$$(\text{観測値}_i - \text{回帰の期待値}_i)^2$$
$$= (y_i - \hat{y}_i)^2 = \left(y_i - (\beta_0 + \beta_1 \times x_i)\right)^2$$

これを各点に対し計算し、合計すると

$$\text{残差}^2\text{合計} = \sum (y_i - \hat{y}_i)^2$$
$$= \sum \left(y_i - (\beta_0 + \beta_1 \times x_i)\right)^2$$

分散分析のSS_E(誤差平方和)と同じもの

これを最小にする切片β_0と傾きβ_1を推定（最小二乗法）

Rによる推定

- lm()関数を用いる

```
> model<- lm(kyounetu~mad,data2.1)
> summary(model)

Call:
lm(formula = kyounetu ~ mad, data = data2.1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3671 -1.1481 -0.1322  1.5003  2.8080

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.6897     1.8724   3.573  0.00217 **
mad           0.9463     0.1634   5.792 1.73e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.85 on 18 degrees of freedom
Multiple R-squared:  0.6508,    Adjusted R-squared:  0.6314
F-statistic: 33.55 on 1 and 18 DF,  p-value: 1.729e-05
```

lm(応答変数~説明変数）で表す
summaryで結果表示

推定したモデル式

残差の情報

係数表
これが推定結果

モデルの当てはまり
F検定の結果など

係数表（coefficient table）

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.6897     1.8724   3.573  0.00217 **
mad           0.9463     0.1634   5.792 1.73e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

推定値

推定値の標準誤差

t値

P値（t値に基づく）

切片（β_0）

madに対する傾き（β_1）

P<0.01の記号

P<0.001の記号

この表からわかること

$$\text{強熱減量}_i = 6.6897 \pm 1.8724 + 0.9463 \pm 0.1634 \times \text{泥の含有率}_i$$

P=0.002　　P<0.001

に推定された

帰無仮説：係数の値は0である

決定係数 r^2

この辺の値はANOVAの部分で説明

```
Residual standard error: 1.85 on 18 degrees of freedom
Multiple R-squared:  0.6508,    Adjusted R-squared:  0.6314
F-statistic: 33.55 on 1 and 18 DF,  p-value: 1.729e-05
```

$r^2 = 0.6508$

回帰でどれくらいのばらつきが説明できたか
（0～1の値）

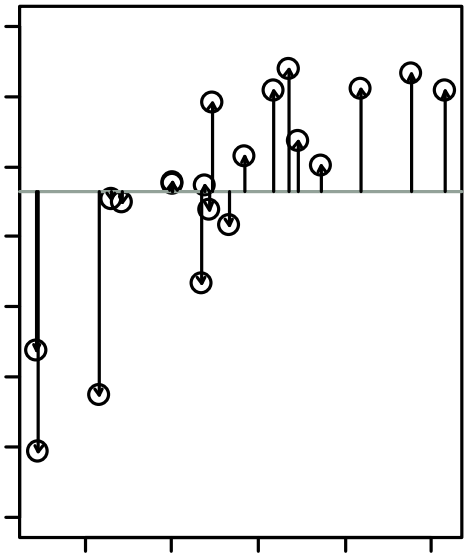
$$r^2 = \frac{\text{回帰で説明できたばらつき}}{y\text{全体の持つばらつき}} = \frac{\mathrm{SS}_{\mathrm{Regression}}}{\mathrm{SS}_{\mathrm{Y}}}$$

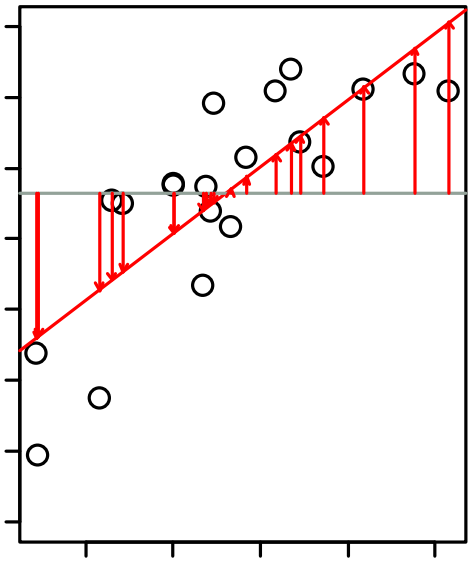
回帰におけるばらつきの分割

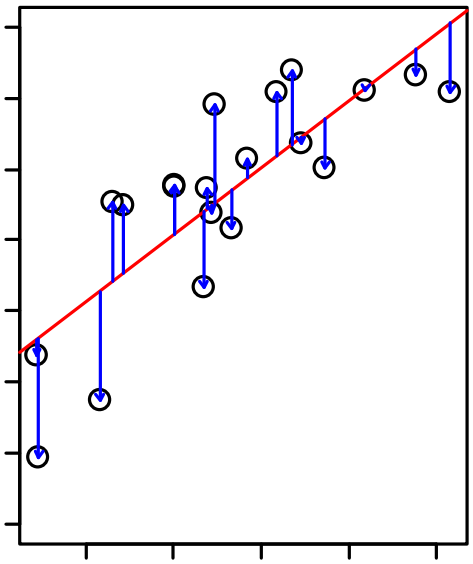
観測値のばらつき　　回帰で説明できるばらつき　　説明できないばらつき＝誤差

$$SS_Y = SS_{Regression} + SS_{Error}$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$







回帰→分散分析

回帰の場合も同様に、分散分析による検定を行う

	平方和 SS	自由度 df	平均平方MS（＝SS/df）	F
回帰モデル	SS_R	1	MS_R	$\frac{MS_R}{MS_E}$
残差	SS_E	n-2	MS_E	
全体	SS_Y	n-1		

回帰の時の自由度は分子は1,分母はn-2になる

分散分析表（ANOVA Table）

- 回帰の結果の分散分析表

	平方和 SS	自由度 df	平均平方MS（＝SS/df）	*F*
mad	114.79	1	114.79	33.55
Residuals	61.58	18	3.42	

$P(F_{1,18} \geqq 33.55) = 0.0000173$

Rで下の式を入れれば計算されます
1-pf(33.55,1,18)

よって、帰無仮説は棄却され、
「強熱減量は泥の比率によって説明される」という対立仮説を採用できる

Rによる検定（分散分析）

- anova関数を用いる

```
> model<- lm(kyounetu~mad,data2.1)# lm(応答変数~説明変数）で表す
> #summary(model)                 # summary()で係数表を表示
> anova(model)                    # anova()で分散分析表を表示
Analysis of Variance Table

応答変数
Response: kyounetu
          自由度   平方和    平均平方     F値      P値（F値に基づく）
           Df  Sum Sq Mean Sq F value    Pr(>F)
説明変数
mad         1 114.789 114.789   33.55 1.729e-05 ***
残差
Residuals 18  61.585   3.421
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

→「強熱減量は泥の含有率によって異なった（$F_{1,18}$=33.55, P<0.001）」

回帰係数or相関係数or決定係数？

回帰係数 $\hat{\beta}$ YがXの変化に応じてどの程度変化するか $[-\infty \sim \infty]$

相関係数 r YとXが互いの変化に応じてどの程度変化するか $[-1 \sim 1]$ Rではcor(x,y)

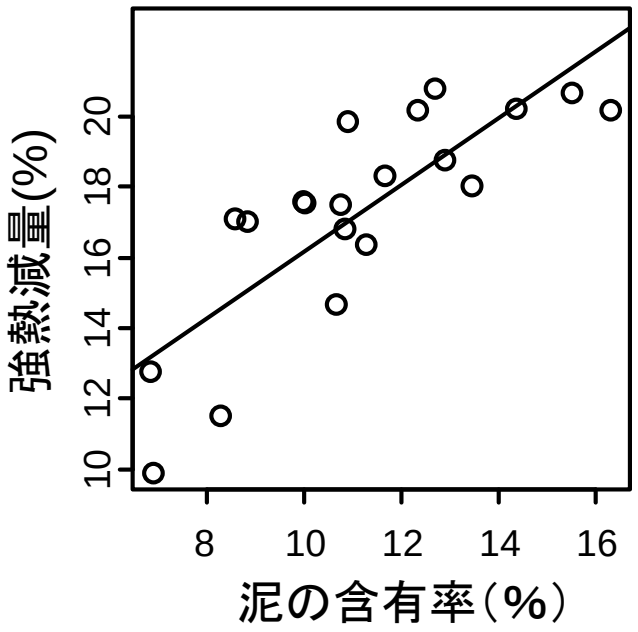
決定係数 r^2 Yのばら付きがXによってどの程度説明できたか $[0 \sim 1]$

$$SS_X = \sum(x - \bar{x})^2;\ SS_Y = \sum(y - \bar{y})^2;\ SS_{XY} = \sum(x - \bar{x})(y - \bar{y})$$

$$\hat{\beta} = \frac{SS_{XY}}{SS_X} \qquad r = \frac{SS_{XY}}{\sqrt{SS_X \times SSY}} \qquad r^2 = \frac{\hat{\beta}^2 \times SS_X}{SS_Y} = \frac{{SS_{XY}}^2}{SS_X \times SSY}$$

回帰まとめ

目的：YのXに対する直線的な関係の推定（＋分散分析による検定）
原理：誤差が最も小さくなる直線を当てはめる（推定）
　　回帰で説明できたばらつきと説明できない誤差を比較



散布図を書いてみる　`plot(Y~X)`

↓

モデル式を当てはめる　`model<- lm(Y~X)`

↓

係数表で回帰直線の推定結果・r^2を見る　`summary(model)`

↓

分散分析による検定結果を見る　`anova(model)`

実は分散分析も回帰もほとんど同じ

	分散分析	回帰
説明変数	カテゴリカル変数	連続変数
モデル式	Y＝各カテゴリーの平均値＋誤差	Y=切片＋傾き×説明変数＋誤差
検定	$F=\frac{\text{カテゴリー間誤差}}{\text{カテゴリー内誤差}}$	$F=\frac{\text{回帰で説明される誤差}}{\text{説明されない誤差}}$

$y_{ij} = \beta_0 + \sum \beta_i \times x_{ij} + \varepsilon_{ij}$ の形で記述できるモデル

↓

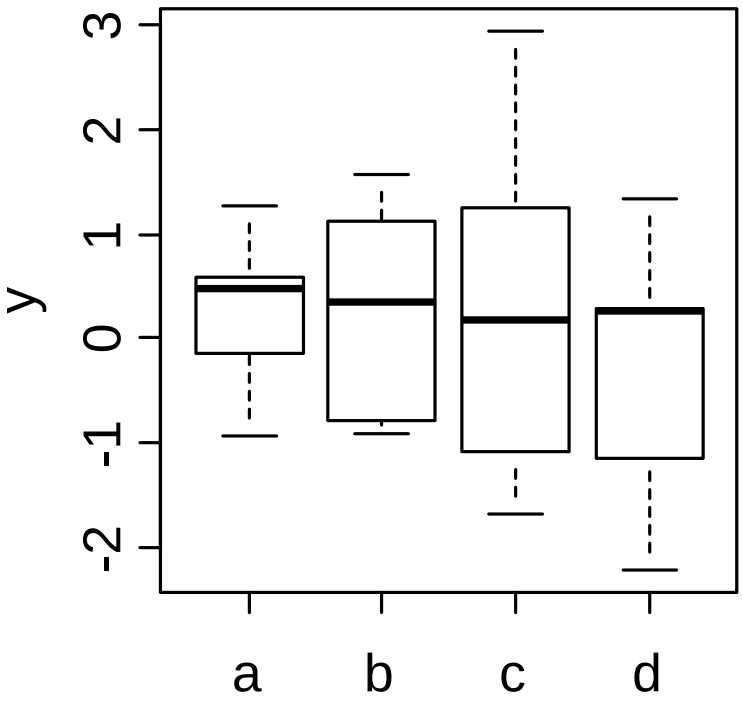
一般線形モデル（General Linear Model）

回帰・分散分析の表現

- plot関数による作図
- 回帰直線の追加
- barplot関数による作図
- エラーバーの追加
- 分散分析表

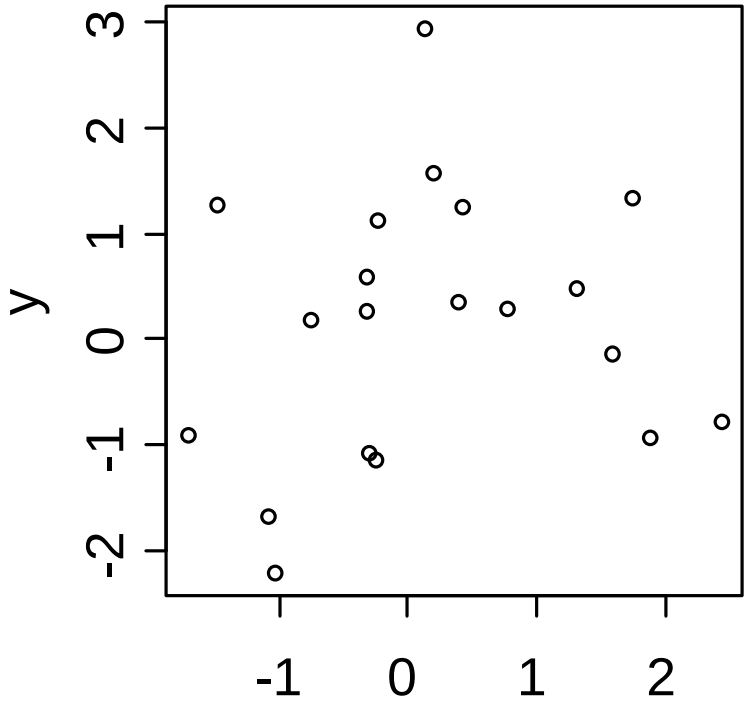
とりあえずplot

- plot関数は渡された変数の種類によって、作図内容を変える



X1：カテゴリカル変数

plot(y~x1)　箱ひげ図



X2：連続変数

plot(y~x2)　散布図

散布図

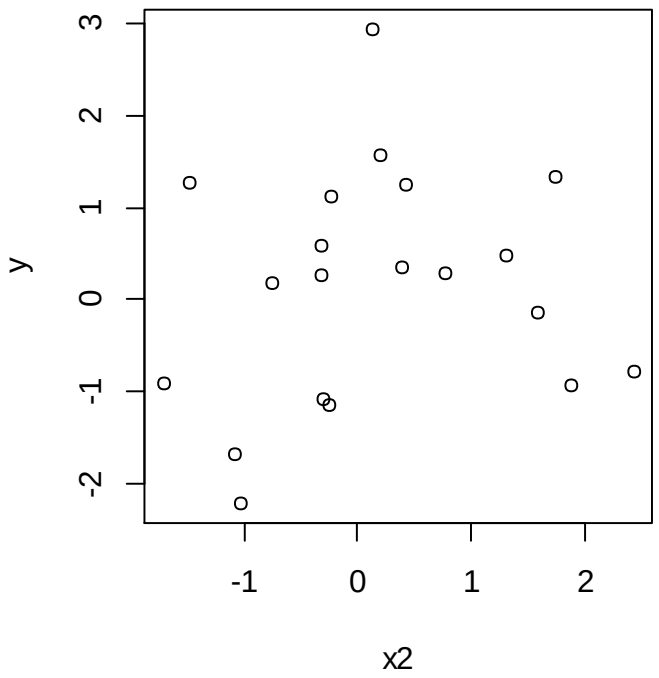
R-tipsのホームページなんかが参考になるかも

- plotのオプション設定（よく使うものの例）

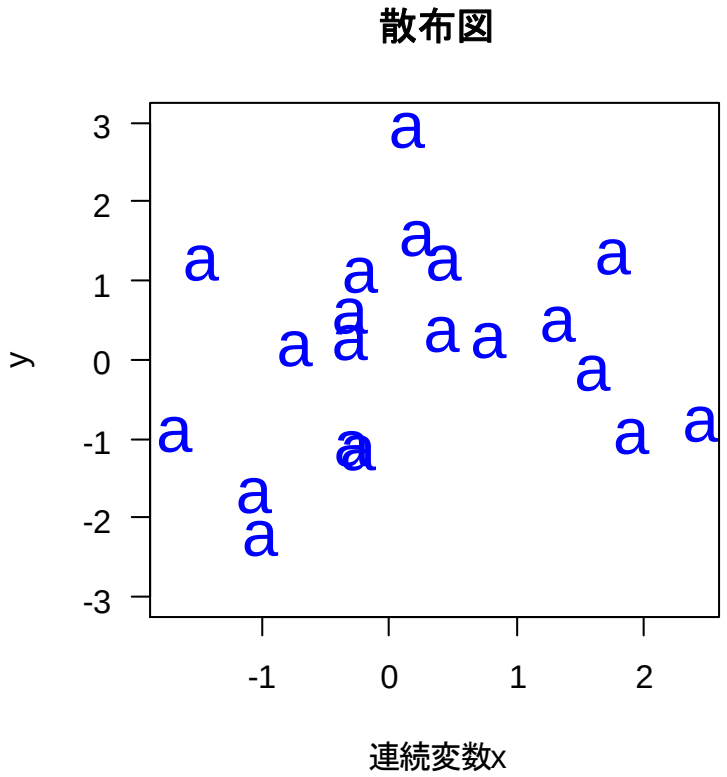
引数	内容	例
main	グラフタイトルを設定	main=“成長率9月”
ylab,xlab	軸の名前を設定	xlab=“処理”
ylim,xlim	軸の範囲を設定	ylim=c(0,1)
log	軸を対数に	log=“x”
col	点の色を変更	col=“red”
pch	点の種類を変更	pch=“a”
cex	プロットの点のサイズを変更	cex=2
las	軸の文字の方向を変更	las=2

plot例

```
plot(y~x2)
```




```
plot(y~x2, main="散布図", xlab="連続変数x", ylim=c(-3,3), col="blue", pch="a", cex=2, las=1)
```

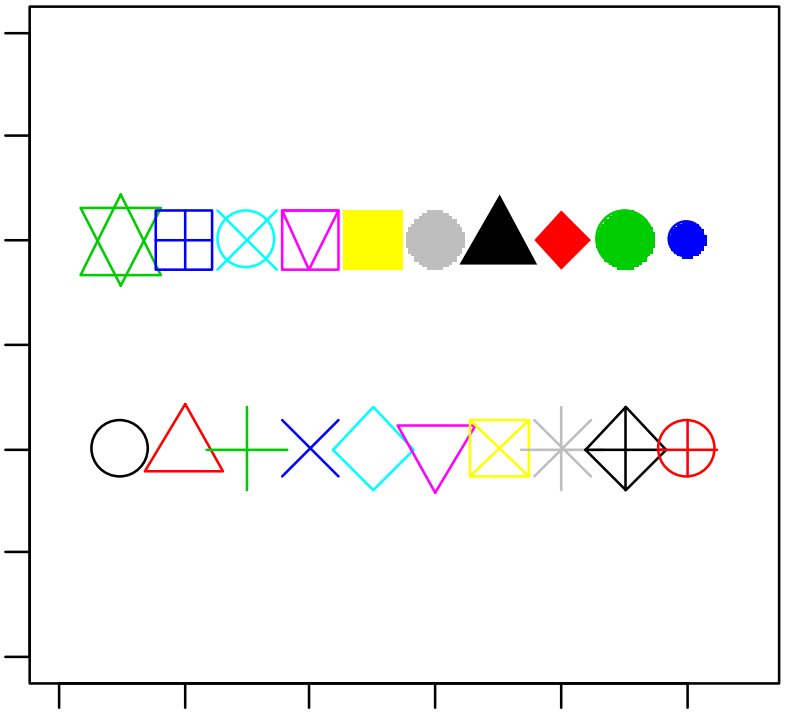


ただし、日本語はパワポでグループ解除すると文字化けする
Font familyで設定可能？？（普段は英語で出力してパワポで修正するほうが楽かも）

col,pchの応用

- colやpchは数字で指定できる

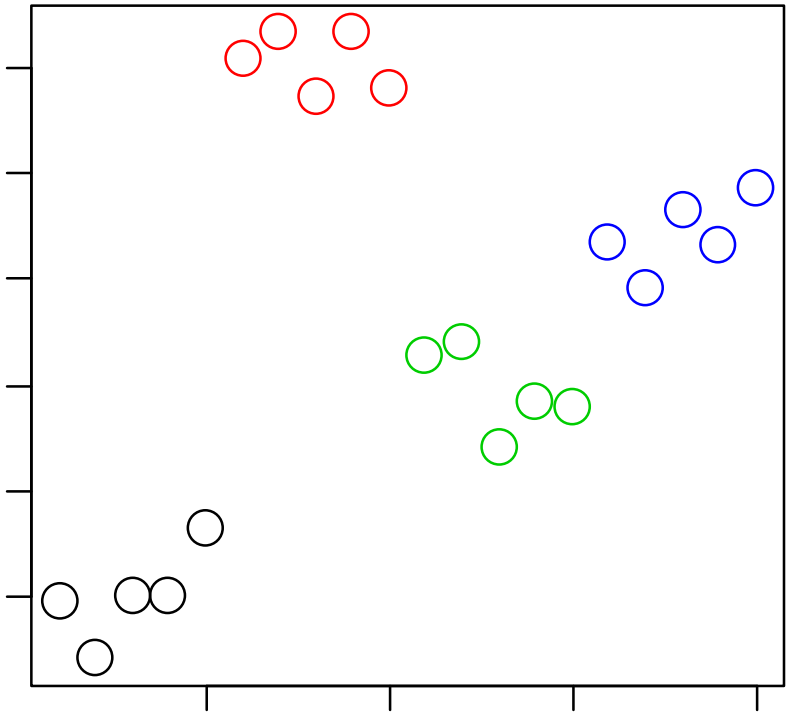
  col=gray(0.8) とかで黒(0)～白(1)も指定可能



上段(11－20); 下段(1－10)

- ベクトルでの指定もできる

→処理ごとに色や形を変えられる



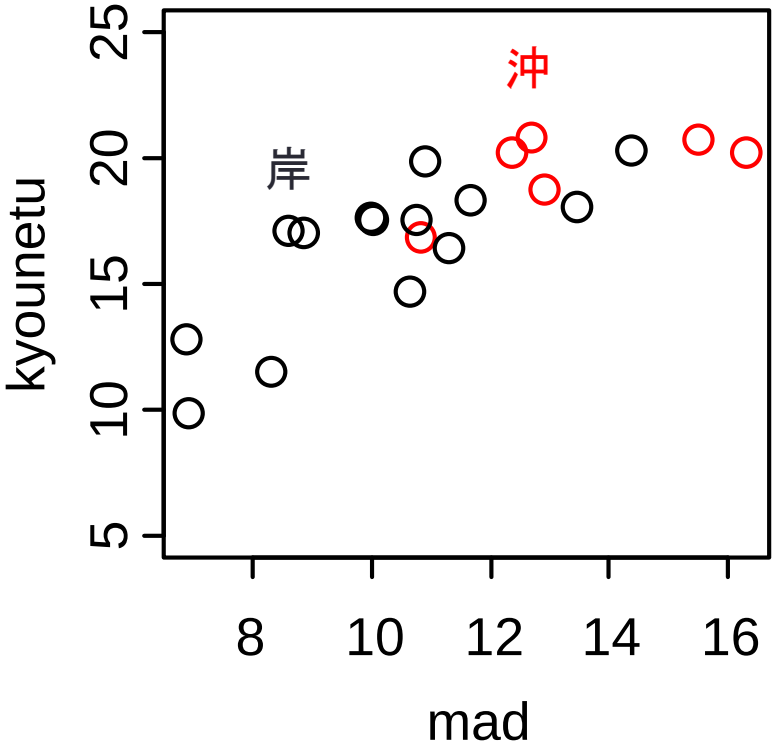
```
trt<- rep(1:4,rep(5,4))
plot(y~x,col=trt)
```

練習問題

- 沖と岸で色分けされた、散布図（Y軸：kyounetu、X軸：mad）をY軸5-25の範囲で作図してみよう(data2.1)

- pchは指定なし、もしくはpch=16あたりが見やすい
- cex=2くらいが見やすい
- oki・kishiといったカテゴリー変数は、引数colの中では、アルファベット順に1・2といった数字で認識される

  oki・kishi→2・1→赤・黒

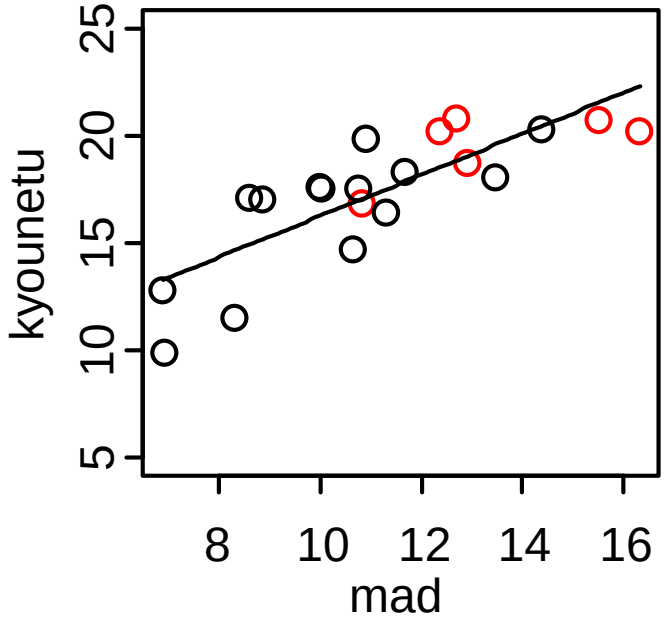


回帰線を引く

lm()で推定された回帰係数を使って、回帰直線を引く

$$\text{強熱減量}_i = 6.6897 + 0.9463 \times \text{泥の含有率}_i$$

```
> model<- lm(kyounetu~mad,data2.1)
> summary(model)
>
>
> abline(model)
> curve(6.6897+0.9463*x,
                col="red",add=T)
```

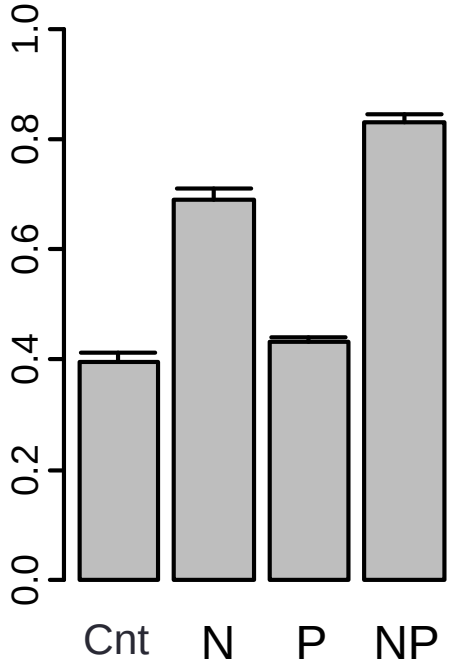


abline()：指定された切片と傾きの直線を引く。lmで推定したモデルを指定すると回帰線を引いてくれる。

curve()：xの関数であれば曲線でも引ける。add=Tで既存の図に追加。

barplot()による棒グラフの描画

- 放り込んだらエラーバー付きで書いてくれる･･･わけではない
- エクセルで書く場合と同様の手順を踏む



手順

- カテゴリーごとの平均を計算する
- カテゴリーごとのSDを計算する
- カテゴリーごとのnを集計する
- カテゴリーごとのSEを計算する

- 平均を元に棒グラフを書く
- SEをもとにエラーバーを追加する

tapply関数の利用

trt_id	trt	growth
1	cnt	0.359607
1	cnt	0.408672
1	cnt	0.439557
1	cnt	0.411698
1	cnt	0.362153
2	N	0.720722
2	N	0.64689
2	N	0.747944
2	N	0.68312
2	N	0.656175
3	P	0.422572
3	P	0.446498
3	P	0.417347
3	P	0.454348
3	P	0.424048
4	NP	0.838112
4	NP	0.777603
4	NP	0.847251
4	NP	0.850416
4	NP	0.845936

処理ごとに平均とSEを計算したい！

- カテゴリーごとに同じ計算をあてはめたい場合に有効な関数

※類似した関数で、行列を行方向もしくは列方向に関数をあてはめて集計するapply関数もある

```
tapply(data2.2$growth,data2.2$trt,mean)
```

growthをtrtごとにmeanせよの意

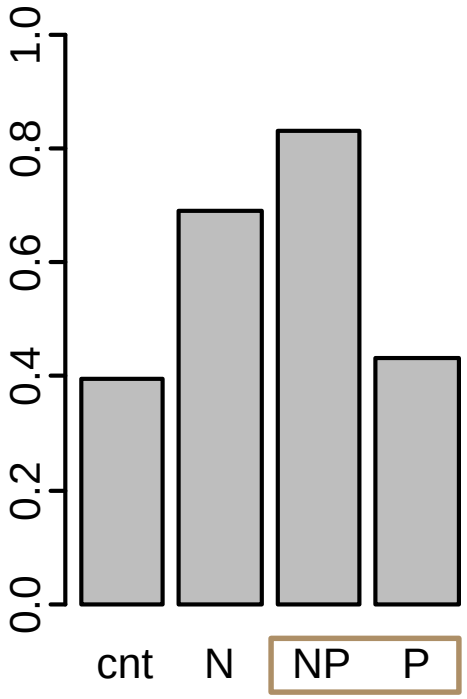
```
      cnt         N        NP         P
0.3963376 0.6909703 0.8318639 0.4329628
```

これを使って、カテゴリごとのSD[関数sd()]およびn[関数length()]を集計し、SEも計算できる

barplot()

- 平均値を使って、棒グラフを書いてみる

```
growth_mean<- tapply(data2.2$growth,data2.2$trt,mean)
barplot(growth_mean,ylim=c(0,1))
```



NPとP入れ替えたい！

根本的な対処方法

- カテゴリーの順序を定義し直して再度集計

```
data2.2$trt<- factor(data2.2$trt,
                     levels=c("cnt","N","P","NP"))
```

その場しのぎの対処方法

Cnt=1C,N=2N,P=3P,NP=4NP
など振り直したもの

- trtの名前を付け直して再集計

```
growth_mean2<- tapply(data2.2$growth,
                      data2.2$trt2,mean)
```

- growth_meanの順番を入れ替えて描画

```
barplot(growth_mean[c(1,2,4,3)])
```

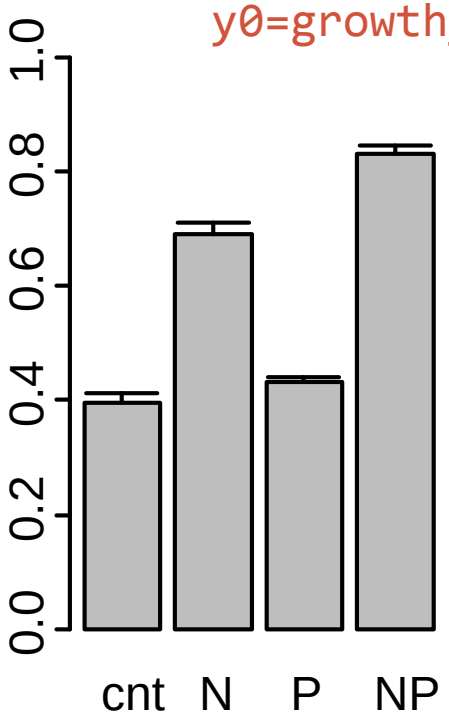
- パワポで修正する

作業量多くないならこれでも

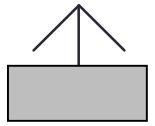
arrows()でエラーバーを追加する

- arrows()は矢印を書く関数。始点と終点の座標を指定する必要。

```
growth_se<-tapply(data2.2$growth,data2.2$trt,sd)/
                          sqrt(tapply(data2.2$growth,data2.2$trt,length))
arrows(x0=-0.5+1.2*c(1:4),
       y0=growth_mean, y1=growth_mean+growth_se, angle=90)
```




```
# x0=x1の時、x1は省略可
#angleは矢印の開きを調節
#angle=45だとこんなの→
```



X座標の `-0.5+1.2*c(1:4)` って？

barplotのデフォルトで、棒の間隔0.2、棒の幅1.0に設定されている
なので、棒の中心のx座標は0.7から1.2刻みで
`c(0.7, 1.9, 3.1, 4.3, ･･･)` になっている

論文中での表現・棒グラフ

- エラーバーを書いた場合、それがSD（標準偏差）なのかSE（標準誤差）なのかを図の脚注に明記する必要がある
- 有意差を示す＊やアルファベットをふった場合も説明が必要

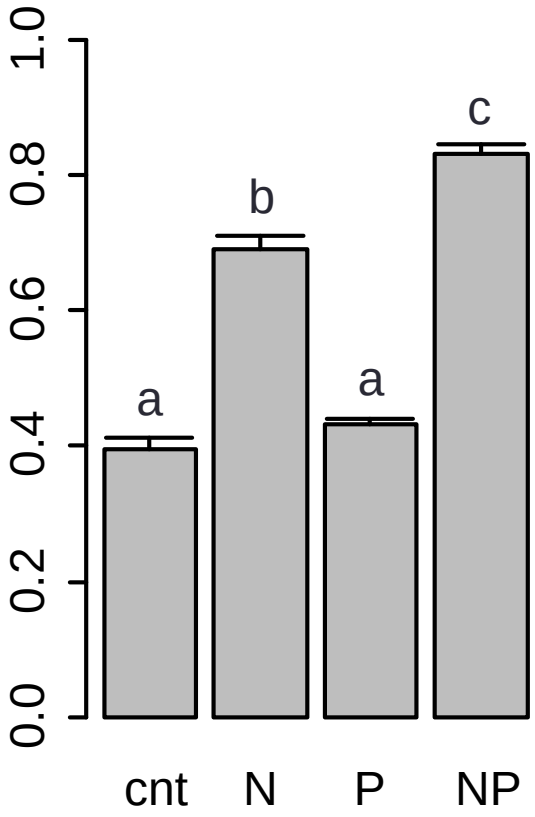


図1. 9月の各処理における植物プランクトンの一日あたりの成長率

平均＋SEを示す。
平均および標準誤差を示す。
棒の高さは平均値。エラーバーは標準誤差。

異なるアルファベットの処理間では有意差が見られることを示す（P<0.05）。

論文中での表現･分散分析表

- 悪い例

表1. 9月のANOVA解析の結果。

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
trt	3	0.65422	0.21807	205.8	5.473e-13 ***
Residuals	16	0.01695	0.00106		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

- 改善例

表1. 9月の植物プランクトンの成長率を目的変数とした分散分析表。

	自由度	平方和	平均平方	*F*	P
処理	3	0.654	0.218	205.8	<0.001
残差	16	0.017	0.001		

論文中での表現･方法

以下にいくつかの例を列挙

- ～を目的変数[応答変数･従属変数]、～を説明変数[独立変数]として、（一元配置）分散分析[単回帰分析]を行った。
- 分散分析で有意差が見られた場合、各処理間での差に関して、TukeyHSD法による多重比較を行った。
- ～とした分散分析を行った後、各処理間での差に関して、*t*検定を行った。多重比較の際のP値は、Bonferroni法による補正を行った。

- すべての統計解析はR3.0.2 (R Core Team 2013)によって行った。

citation()で確認できます

論文中での表現･結果

以下にいくつかの例を列挙

- 泥含有率が高い地点ほど強熱減量も高い傾向が見られた（強熱減量 $= 6.89 + 0.96 \times$泥含有率, $r^2 = 0.65, F_{1,18} = 33.55, \mathrm{P} < 0.001,$ 図2）。
- ～が多くなるにつれ～が少なくなる傾向が見られた（表1, 図2）。
  #図表をみれば統計量や推定値、P値がわかる場合
- ～は処理によって異なり（表１）、処理Aは他の処理の２倍近く高い値を示した（図３）。
- 砂の含有率と泥の含有率の間には負の相関が見られた（r=-078）。
  #ただの因果のない関係性の記述なら相関でよいかも（方法でもそのように書く）
- 「図1は～と～の関係である。」より、「～と～の間には～な関係が見られた（図1）。」が好ましい（前者は図の脚注に書くべき内容）。

次回予告

- R編：データの型
  - 連続変数とカテゴリカル変数
  - パッケージのインストール
- 統計編：一般線形モデル
  - 二元配置分散分析
  - 交互作用
  - 重回帰分析
  - 共分散分析
- 表現編：グラフ表現いろいろ（？）