統計学入門

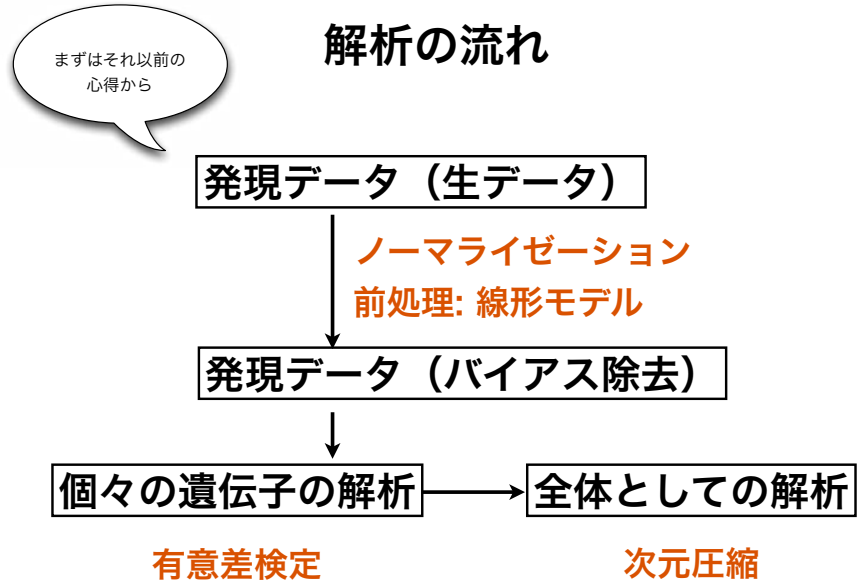
慶應義塾大学　先端生命科学研究所
佐藤昌直

遺伝子発現解析における統計の役割

統計の役割（一般論）

- 仮説検定
  - 推定
- 予測（モデル構築）

多くの遺伝子発現解析



私の担当（統計解析）で重視しているポイント

- 研究全体における統計の役割、実験と統計との連携を意識する
- 遺伝子発現解析の基礎的な概念を解説する

これらをこれから学習していくためには

- 測定、統計とは何かを見直す
- 汎用される統計の仕組みを知る
- 教科書を読めるように統計用語・表記に慣れる
- 道具を準備する - R

A workflow (that you might imagine)

Planning an experiment

↓

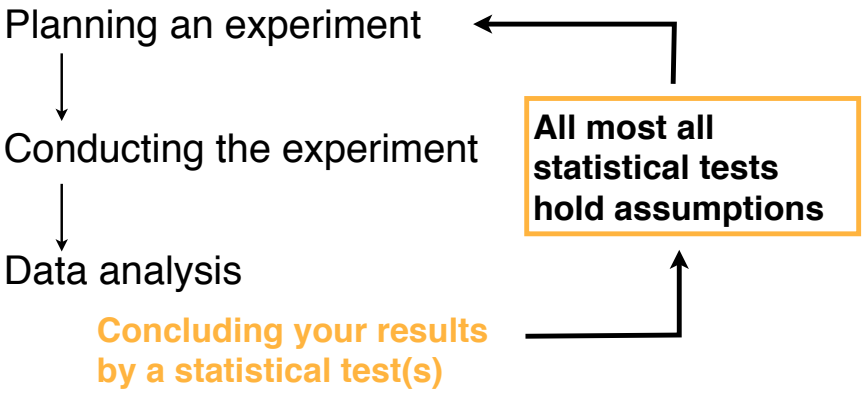
Conducting the experiment

↓

Data analysis

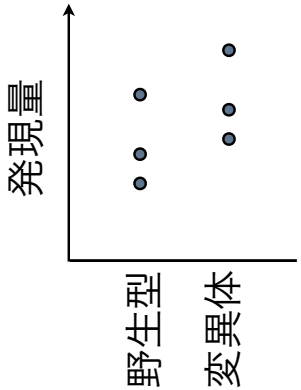
Concluding your results by a statistical test(s)

Reality: You have to design your experiments BEFORE you obtain results



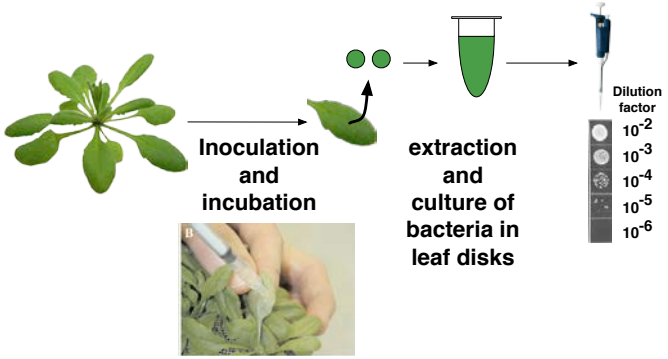
データのばらつきと
実験デザイン・統計学的観点

測定データはバラつく



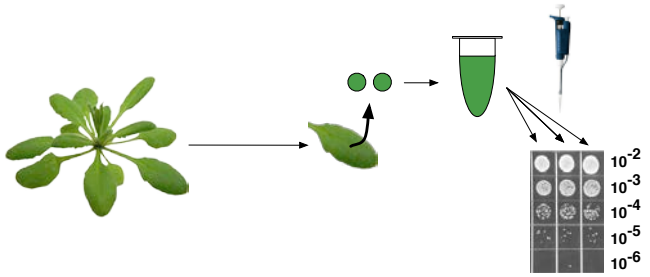
- **実験（測定）を反復する**
- 何を**「真」**と考えるか
- 論文として発表できるデータには**再現性**が必要

例: バクテリア増殖定量

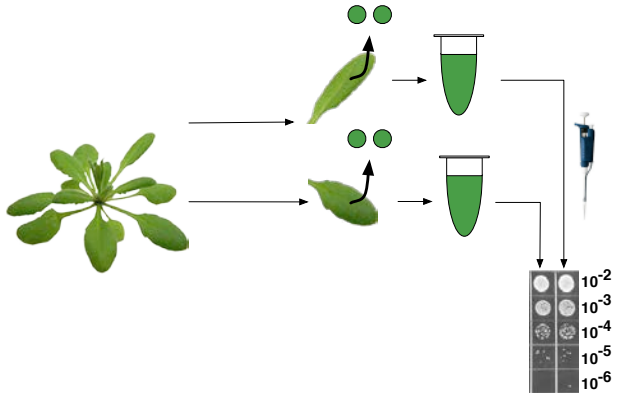


Katagiri, Thilmony R, and He S (2002) The Arabidopsis Thaliana-Pseudomonas Syringae Interaction. The Arabidopsis Book.

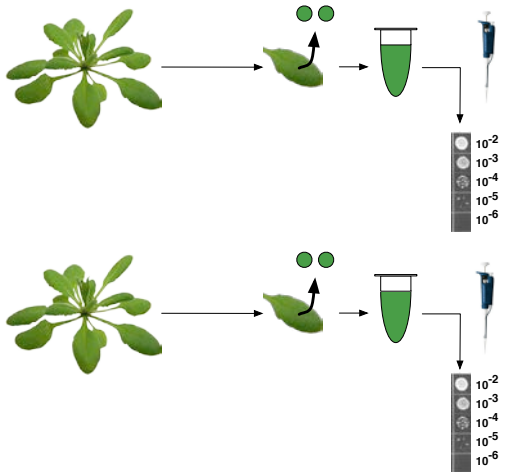
反復？



反復？



反復？



我々が1データポイントから得ているもの

- 生物学的にばらつきの中のある1点
- 測定技術のばらつきの中のある1点

我々が1データポイントの測定で得ているもの

- 生物学的にばらつきの中のある1点
- 測定技術のばらつきの中のある1点

測定における2要因

- Precision - 精度
- Accuracy - 正確度

測定における2要因

- Precision - 精度

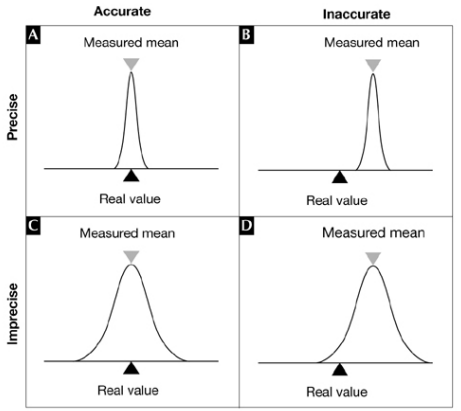
ある1測定を繰り返した際のばらつきの尺度

測定における2要因

- Accuracy - 正確度

ある測定値が「真の値」にどれだけ近いかの尺度

測定における2要因



Real value: 真の値

Measured mean: 測定値から得られた平均

Harm van Bakel & Frank C.P. Holstege (2004) *EMBO reports* (2004) 5, 964 - 969

我々が1データポイントから得ているもの

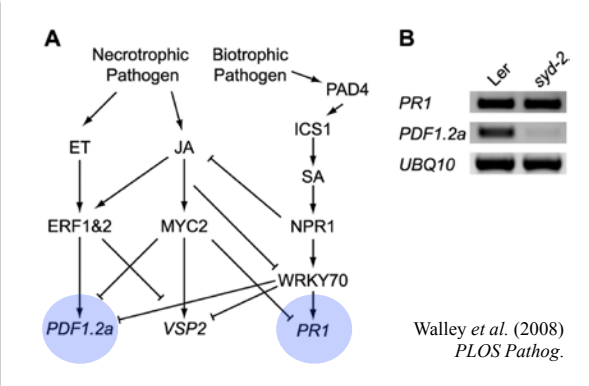
- **生物学的にばらつきの中のある1点**
- **測定技術のばらつきの中のある1点**

定量的測定が可能且つ要求される時代の再現性のあるデータとは何か？

- 何が再現されうるか？再現されたとするか？
  - いつ行っても再現できる?
  - どこで行っても再現できる?
  - 誰が行っても再現できる?

“マーカー遺伝子”測定

- 何が再現されうるか？再現されたとするか？

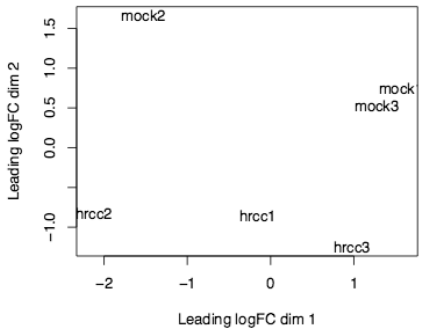


Walley *et al.* (2008) *PLOS Pathog.*

明瞭な違いを示す遺伝子:
明瞭な再現性

“トランスクリプトーム”測定

- 何が再現されうるか？再現されたとするか？



網羅的測定:
再現性の再定義

Chen *et al.* (2015) edgeR User’s Guide page 63

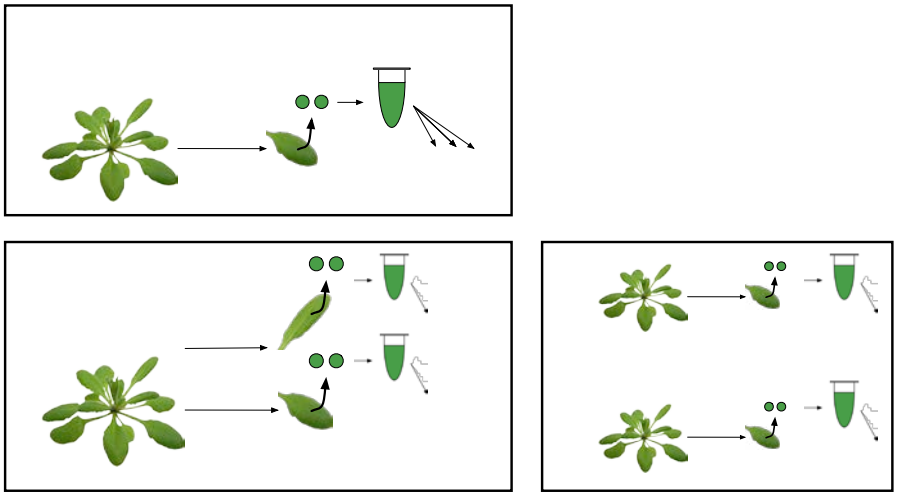
定量的測定が可能且つ要求される時代の再現性のあるデータとは何か？

- 何が再現されうるか？再現されたとするか？

- いつ行っても再現できる?
- どこで行っても再現できる?
- 誰が行っても再現できる?

バラつきの
定量と**割当て**

何を知るための実験か？
再現性のあるデータとは何か？
どのように反復を行うのが適切か？



統計学を使って我々ができることを
考えてみましょう

我々にできる事

少数の測定値から
「母集団」を推定すること

我々の実験対象

- ある遺伝子型の生物の
- ある環境での　　＋ 制御不能な実験要因
- ある遺伝子の発現量 ＋ 生化学反応のノイズ

我々にできる事

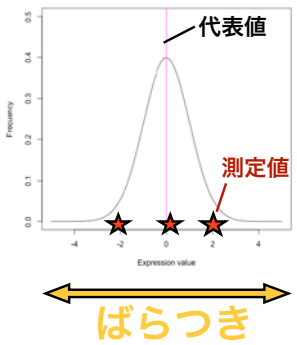
少数の測定値から
「母集団」を推定すること

生体サンプルを繰り返し取る:
biological replicates

同一サンプルを繰り返し測る:
technical replicates

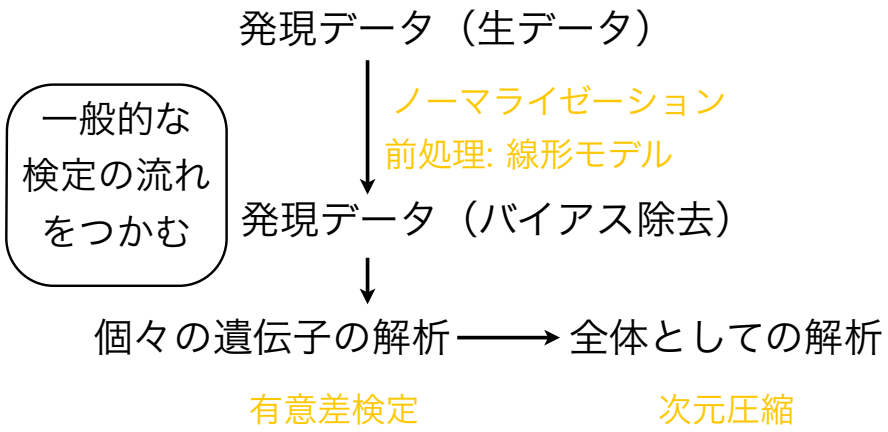
母集団を推定する統計量

1. (真の値に近い)代表値
2. ばらつきの範囲



仮説検定 - t検定を例に

解析の流れ



この講義の目標

- t検定の背景知識を得る - 勉強のとっかかりを作る
  - t統計量
  - t分布
  - 自由度
  - p値

統計における検定の手続き

1. 仮説を立てる
2. 統計量を求める
3. 求めた統計量を確率分布に照らし合わせる
4. 判定: 求めた確率が棄却限界値より大きいか、小さいか

統計における検定の手続き　t検定

1. 仮説を立てる
   - 2つのサンプル間で遺伝子発現量（平均値）の違いがある？
2. 統計量を求める
   - 平均、標準誤差、自由度からt統計量を求める
3. 求めた統計量を確率分布に照らし合わせる
   - t分布からp値を求める
4. 判定: 求めた確率が棄却限界値より大きいか、小さいか
   - 有意差の判定

1. 仮説を立てる:

帰無仮説

- 最終的に棄却される仮定:

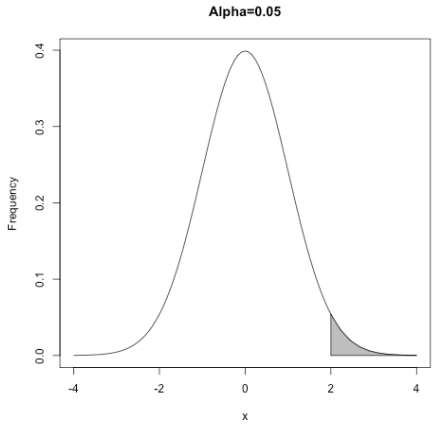
  「AとBに差がある」かを検定する場合は「AとBには差がない」と仮定する

2. 統計量を求める:

- **統計量: データから導いた具体的な数値**

  ↔ 母数: 未知の数値

3. 確率分布を求める:



4. 判定: 帰無仮説が棄却されるか？

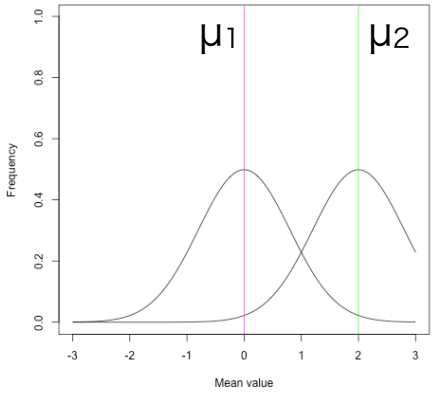
- 最終的に棄却される仮定:

  「AとBに差がある」かを検定する場合は「AとBには差がない」という仮定

t検定:

2サンプルの平均の検定

- 平均値 = μ_1, μ_2
- データは正規分布



統計量その1

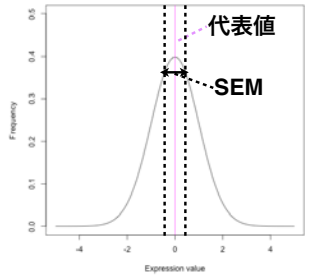
平均値: 相加平均。すべてのデータを足して、データ数で割って得られる値

$$\bar{x} = \frac{\sum_{n=1}^{i} x_i}{n}$$

統計量その2：平均値も推定値

（平均）標準誤差

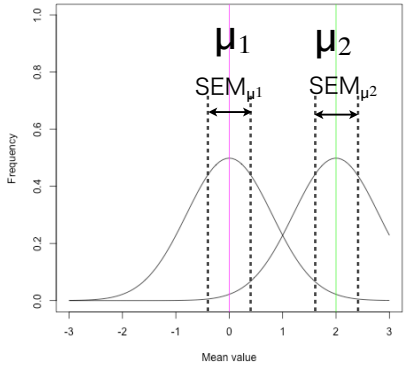
$$SEM = \frac{s}{\sqrt{n}}$$



統計量その3: 平均の差とその誤差

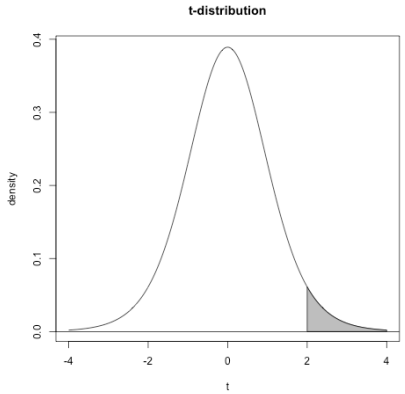
- t統計量

$$t = \frac{|\hat{\mu}_1 - \hat{\mu}_2|}{SEM_{|\hat{\mu}_1 - \hat{\mu}_2|}}$$



確率分布-t分布

- 得られたt統計量がどのくらいの確立で起きうるか
- t分布の確率分布を標本のt統計量と自由度を使って参照



自由度とは？

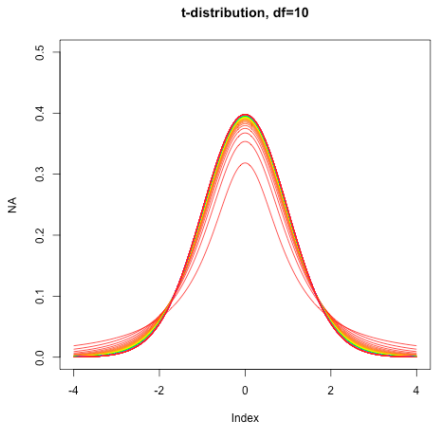
統計量を求めるのに使うことができる**「独立」**な標本数

我々の測定、検定では：

- 母分散が未知
- よって、確率密度は自由度によって変化

例) 3つの観察で得られた平均値と100観察から得られた平均値はどちらが確からしいか

1から100までの自由度でのt分布

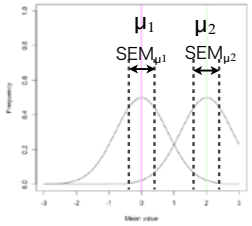


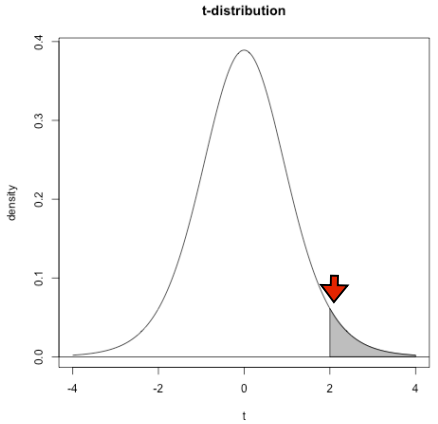
p値とは：

- 標本に基づいた統計量が帰無仮説の下、起きうる確率
- 多くの場合、0.05が危険率

確率分布-t分布

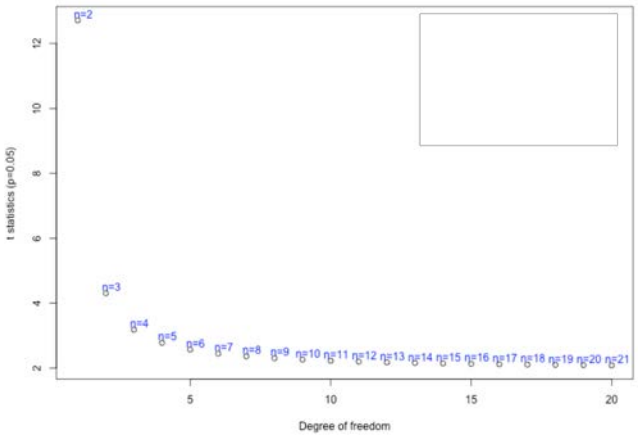
$$t = \frac{|\hat{\mu}_1 - \hat{\mu}_2|}{SEM_{|\hat{\mu}_1 - \hat{\mu}_2|}}$$





t分布表を参照する→Rで求めましょう

補足:
t統計量
自由度
反復
p値
検出力



多重検定の補正

p値とは：

- 標本に基づいた統計量が帰無仮説の下、起きうる確率

- 多くの場合、**0.05**が危険率

p値とは：

- 標本に基づいた統計量が帰無仮説の下、起きうる確率

- 多くの場合、**0.05**が危険率　　**= 100回に5回起きる**

多重検定の補正

- p = 0.05の検定を100回*繰り返すと、**5回はランダムに間違い**

*NGS解析では数万回以上繰り返すことになります

多重検定の補正

Bonferroniタイプ

False discovery rate (FDR):
- Benjamini-Hochberg
- Storey

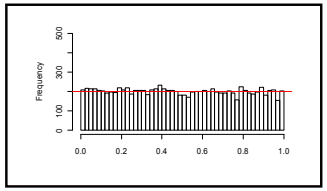
Bonferroniタイプの多重検定の補正

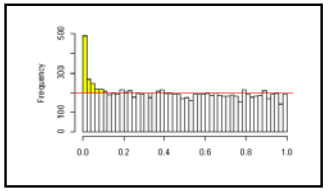
危険率を検定数で調整

危険率 = α/k

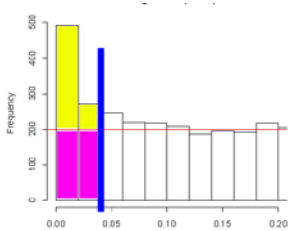
α: 元の危険率、
k: 検定数

False Discovery Rate (FDR)





False Discovery Rate (FDR)



q値:
補正されたp値。そのq値以下の検定のうち、どのくらいの割合でfalse positiveが含まれているか。

復習／発展学習

- 検定の手順
  - 統計量
  - 自由度
  - p値
- 統計解析の結果は確率 → 多重検定の補正:
  ランダムな危険率以下の検定をどう補正するか？
- Storeyの方法によるq値の求め方
- 多重検定の補正における仮定: 時系列データにFDRは使ってよいか？

分散分析・線形モデル:
多変数データを系統立てて解析する
- 実験デザインと統計の連携

解析の流れ

発現データ（生データ）

↓ ノーマライゼーション
↓ 前処理: 線形モデル ⇦

発現データ（バイアス除去）

↓

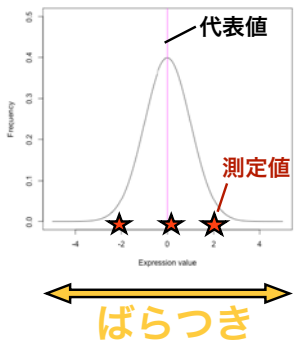
個々の遺伝子の解析 ⟶ 全体としての解析

有意差検定　　次元圧縮

- 線形モデルの概念を掴む
- 実験デザインがどう統計に影響するかを考えるきっかけとする

リマインド:
母集団を推定する統計量

1. (真の値に近い)代表値
2. ばらつきの範囲



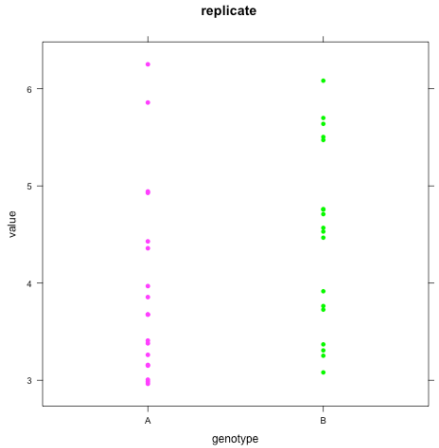
t検定: 平均値の検定

$$x_i = \bar{x} + (x_i - \bar{x})$$

偏差: 平均値からのばらつき

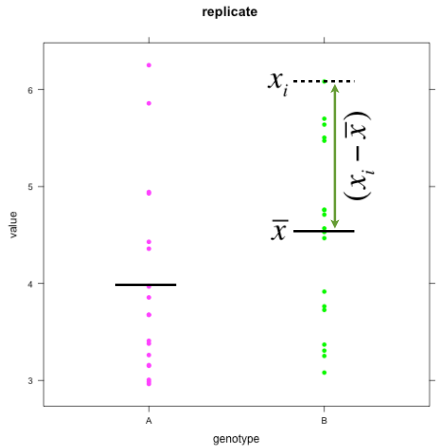
- genotype A, Bについて

  6検体ずつ3回反復して計測



- genotype: A, B
- replicate: 1, 2, 3
- value:

  計18個/ genotype



$$x_i = \bar{x} + (x_i - \bar{x})$$

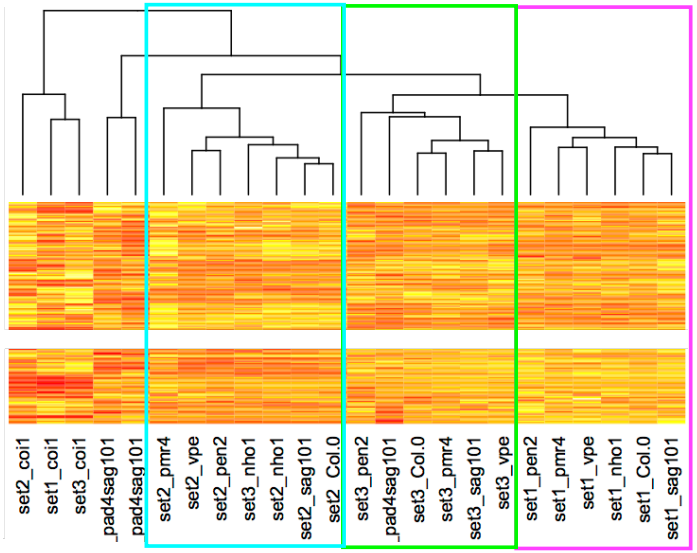
線形モデルの枠組みで考えてみる

$$x_i = \bar{x} + (x_i - \bar{x})$$

$$x_i = \bar{x} + \varepsilon_i$$

残差 (観察値-推定値):
想定要因では説明できない
データの変動

考慮するのは1要因で良いか？



観察値を複数要因の影響によるものとして分解

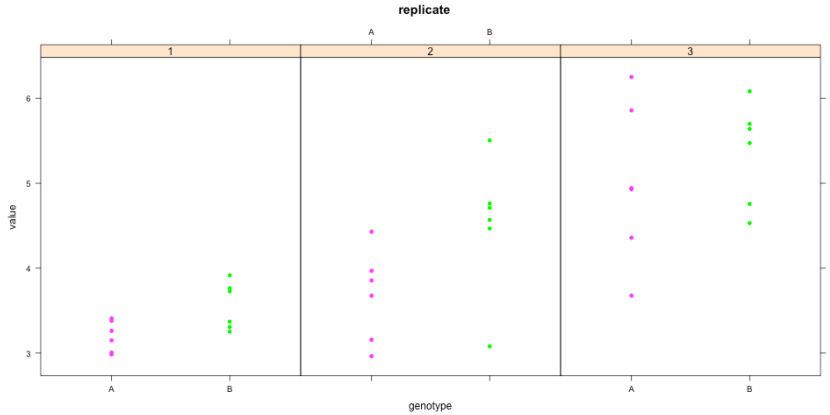
$$x_i = \bar{x} + (x_i - \bar{x})$$

$$x_i = \bar{x} + \varepsilon_i$$

genotypeとreplicateの影響を同時に考えられないか？

$$O_{ij} = x_i + y_j + \varepsilon_{ij}$$

例: 2遺伝子型の測定を3反復したデータ

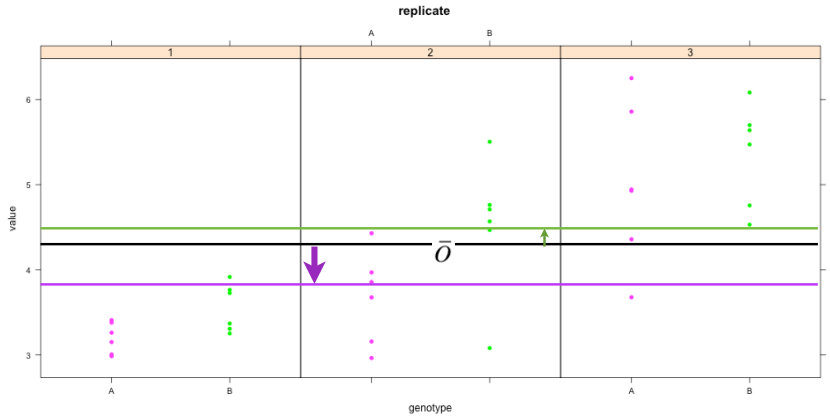


線形モデルの仕組み

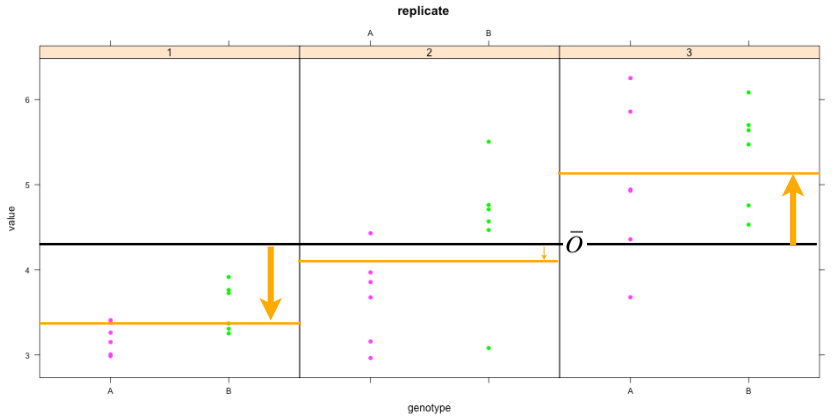
$$O_{ij} = x_i + y_j + \varepsilon_{ij}$$

$$O_{ij} = \bar{O} + (\bar{x}_{i\bullet} - \bar{O}) + (\bar{y}_{\bullet j} - \bar{O}) + \varepsilon_{ij}$$

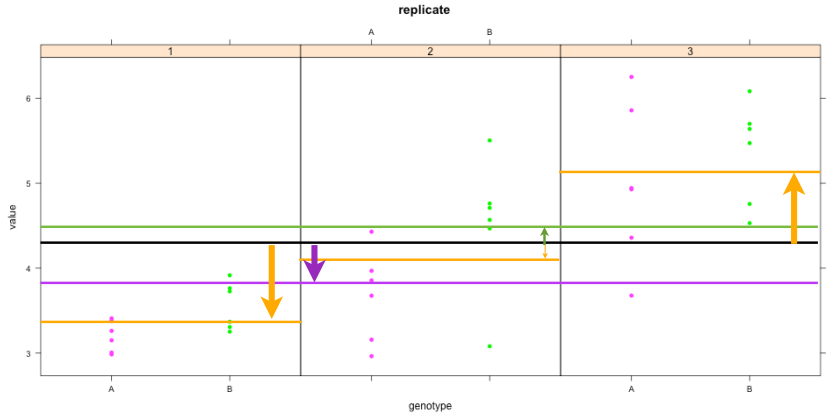
$(\bar{x}_{i\bullet} - \bar{O})$ 遺伝子型による変動



$(\bar{y}_{\bullet j} - \bar{O})$ 反復ごとの変動



各計測値は $O_{ij} = \bar{O} + (\bar{x}_{i\bullet} - \bar{O}) + (\bar{y}_{\bullet j} - \bar{O}) + \varepsilon_{ij}$ と表せる



分散分析・線形モデルの枠組み

$$O_{ij} = x_i + y_j + \varepsilon_{ij}$$

$$O_{ij} = \bar{O} + (\bar{x}_{i\bullet} - \bar{O}) + (\bar{y}_{\bullet j} - \bar{O}) + \varepsilon_{ij}$$

↓ 教科書・論文風に書くと

$$O_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

応答変数　　説明変数

線形モデルとは

応答変数　～　説明変数1 + 説明変数2 + ... + 誤差

と観察値を説明する（かもしれない）
変数でそれらの関係性を書き下すこと

- 実際には: Rでlmなどの関数を使う

実験デザインの重要性

- -omicsデータは”**batch effect**”という体系的なバイアスが多くの場合、混入する。
  例: 実験時期、餌

Nature Reviews Genetics (2010) 11, 733-

- 線形モデルで推定・除去

実験デザインの重要性

- 線形モデルで推定・除去

$$O_{ij} = \alpha_i + \beta_j + \varepsilon_{ij}$$

α_i : 遺伝子型／処理など注目している効果の要因

β_j : 反復（実験日時）／実験者などバイアス要因

- α_i の推定値、標準誤差のみを使う

再現性のあるデータとは何か？

- 自分自身で再現できる
- いつ行っても再現できる
- どこで行っても再現できる
- 誰が行っても再現できる

実験デザインの重要性

- 線形モデルで推定・除去

$$O_{ij} = \alpha_i + \beta_j + \varepsilon_{ij}$$

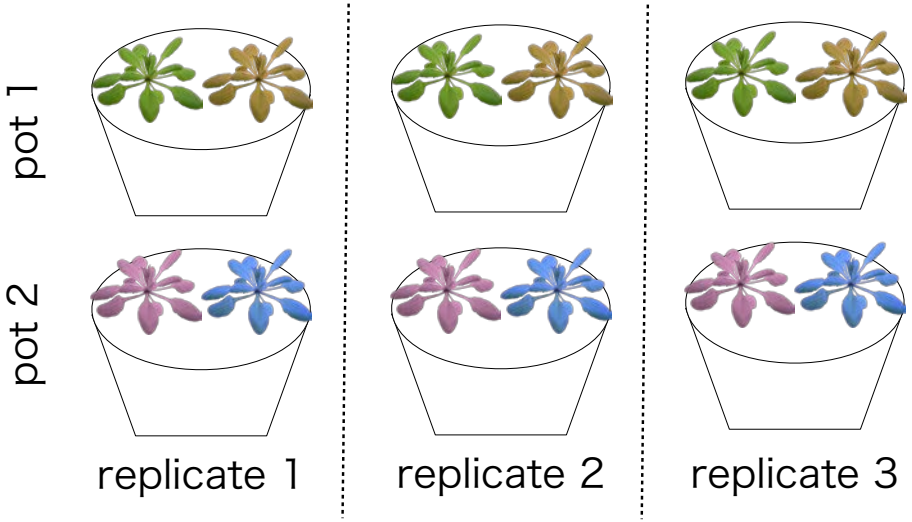
α_i : 遺伝子型／処理など注目している効果の要因

β_j : 反復（実験日時）／実験者などバイアス要因

- α_i の推定値、標準誤差のみを使う

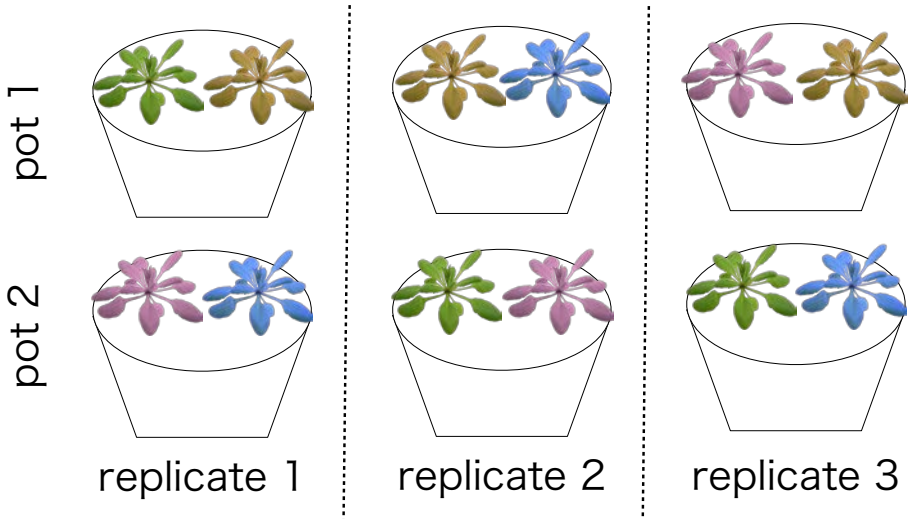
実験デザインの重要性:

genotype+replicate+potモデルを当てはめるには？



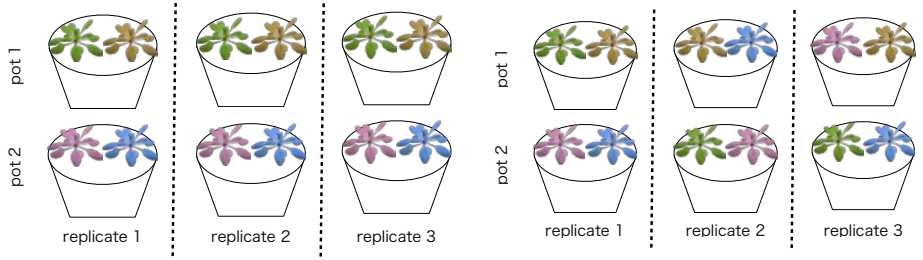
実験デザインの重要性:

genotype+replicate+potモデルを当てはめるには？



実験デザインの重要性:

genotype+replicate+potモデルを当てはめるには？



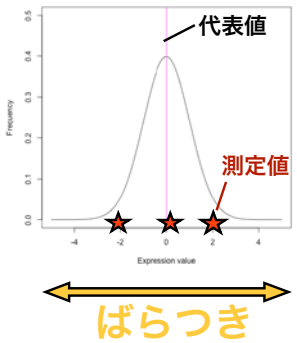
↑

genotypeとpotが独立ではない
(切り分けられない)

リマインド:

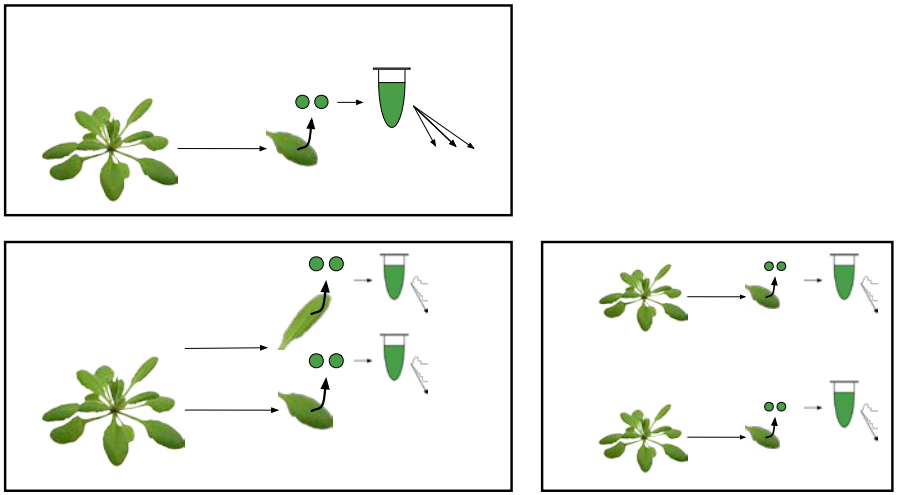
母集団を推定する統計量

1. (真の値に近い)代表値
2. ばらつきの範囲



何を知るための実験か？

求めたい代表値は何の代表値か、扱うばらつきは何に由来するか？
ある実験デザインで求めうる代表値・ばらつきは何を表すか？



実験デザインの重要性

- 要因効果を推定するための実験デザイン
  - 各実験要因を適切に反復させた実験デザイン
- 実験デザインとモデル
  - 要因: データ取得**「前」**に想定しておくもの
  - データの変動を説明しない要因を解析時に減らすことは可能。実験デザイン時に計画しなかった要因を増せない。

まとめ

- 計測データセットに影響を与える要因が一つではない場合、分散分析・線形モデルの枠組みが有効
- 理屈は難しいかもしれないが、Rで簡単に実行できるので実験デザインと連動したモデルを立てることが重要

復習／発展学習

- 回帰（最小二乗法）
- 実験計画法
- 交互作用
- Bioconductor: limma、edgeRパッケージ