統計学入門

北海道大学　農学研究院
佐藤昌直

これらをこれから学習していくためには

- 汎用される統計の仕組みを知る
- 測定、実験計画を見直す
- 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

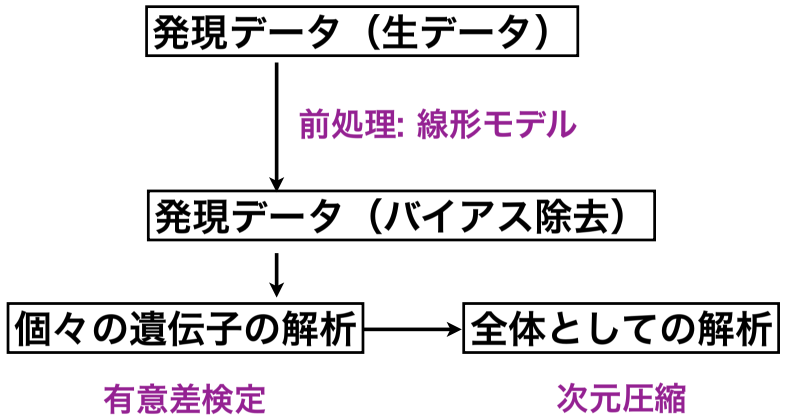
私が重視しているポイント

- 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
- 遺伝子発現解析に必要な**統計の基礎概念**を解説する
- ***“statistical mind”***を養う

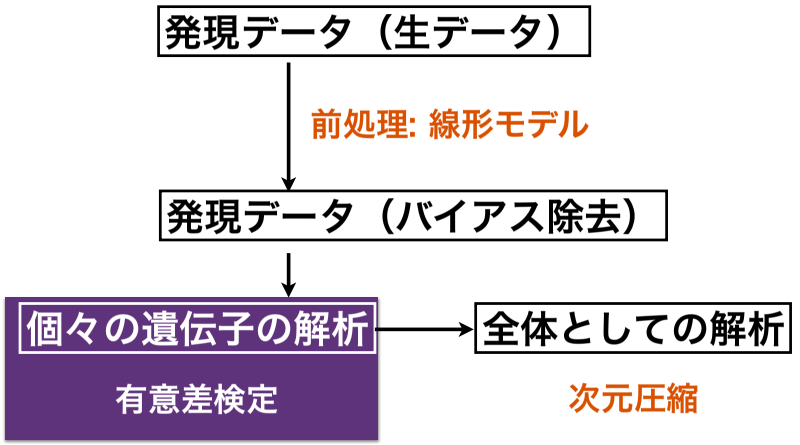
基本的な統計の用途

- 仮説検定
- 予測 (モデル構築)

遺伝子発現解析における統計の役割



遺伝子発現解析における統計の役割



仮説検定 - *t* 検定を例に

ねらい

*t*検定から検定の背景知識を得る:

- 検定の流れを知る
- 勉強のとっかかりを作る

用語の意味の整理

- 統計量、確率分布、自由度、*p*値

宿題

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: 求めた確率と棄却限界値との比較

ポイント

2. 統計量を求める:

統計量: データから導いた
<u>具体的</u>な数値

↔ **母数**: 未知の数値

我々ができること: 少数の測定値（**標本**）から
「母集団」を推定すること

ポイント

1. 仮説を立てる:

帰無仮説

statistical mind

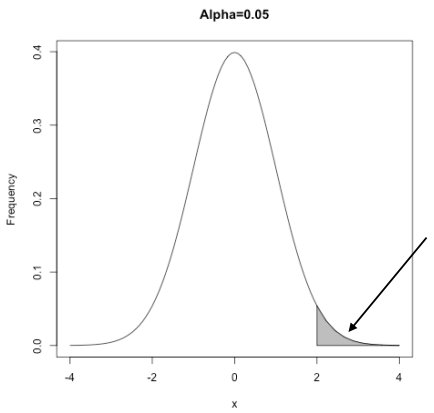
最終的に棄却される仮定:

「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

例1. 野生型と変異体Aの遺伝子xの発現量に違いがあるか？

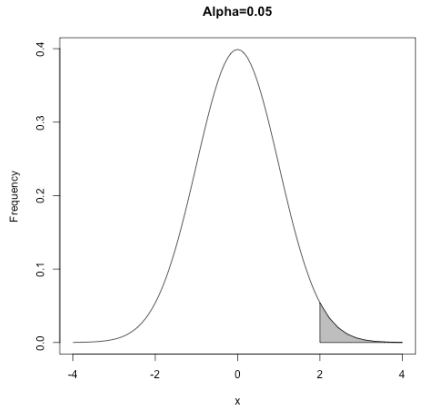
例2. 野生型と変異体Aの遺伝子発現プロファイル間の相関係数は0.35だった。これらは有意に相関していると考えられるか？

3. 確率分布と照らし合わせる



棄却限界値によって
規定される面積
(通例: 全体の5%)

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

代表値

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

-（バー）は平均を表す
^（ハット）は推定を表す

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

中央値: データを小さいものから順に並べたときに中央にくる値

ポイント

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

我々ができること: 少数の測定値（**標本**）から「**母集団**」を推定すること

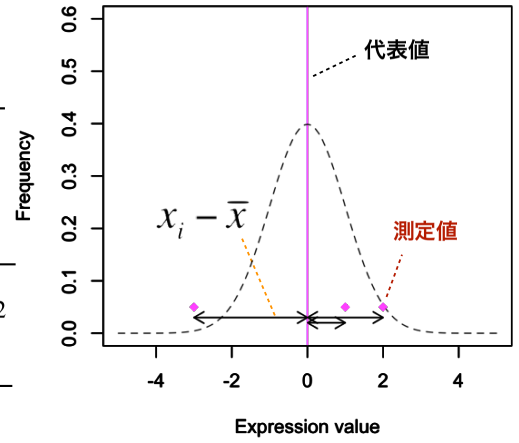
ばらつき: 分散／偏差

分散:

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

標準偏差:

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$



n -1?

なぜ、平均を求める時と分散を求める時では分母が変わるのか？

自由度: 統計量を求めるのに使うことができる**「独立」**な標本数

統計的検定の手続き　　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: 求めた確率と**棄却限界値**との比較	有意差の判定

母集団を推定する統計量

(標本)平均:

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

-（バー）は平均を表す

標準偏差: $\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

^（ハット）は推定を表す

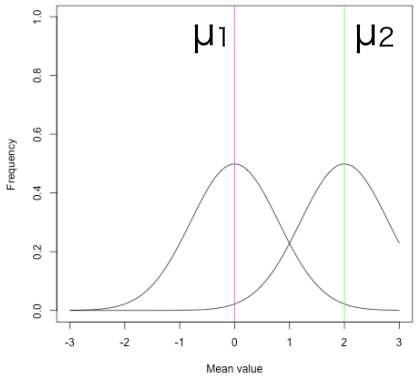
ポイント

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

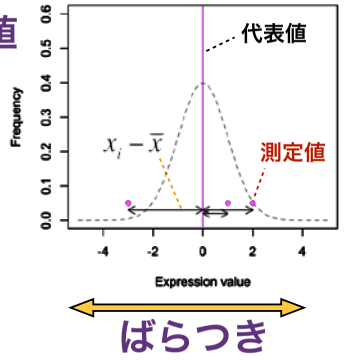
ほぼ全ての検定方法に前提がある



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**



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

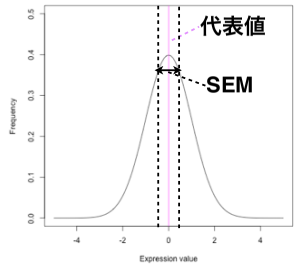
$$\bar{x} = \hat{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

統計量その2：
平均値も**あくまで推定値**

statistical mind

(平均) 標準誤差:
「統計量」の偏差

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

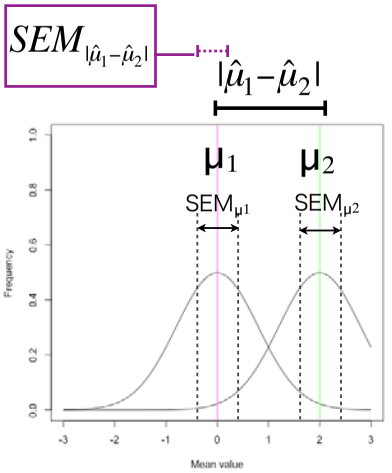


統計量その3:
平均の差とその誤差

statistical mind

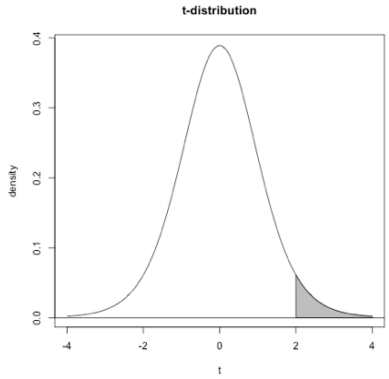
t統計量

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



確率分布-*t* 分布

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



【おさらい】自由度: 統計量を求めるのに使うことができる独立な標本数

データの分布、仮説検定に即した確率分布を使う

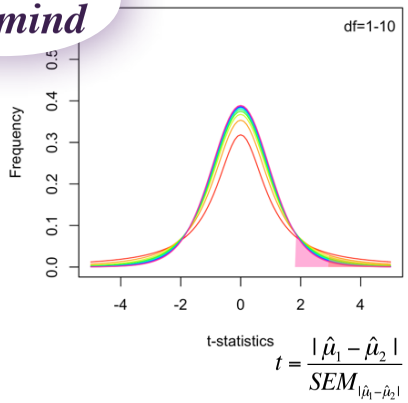
ポイント

我々の測定では

- 母分散が**未知**
- したがって確率密度は**自由度**によって変化

→正規分布ではなく、t分布

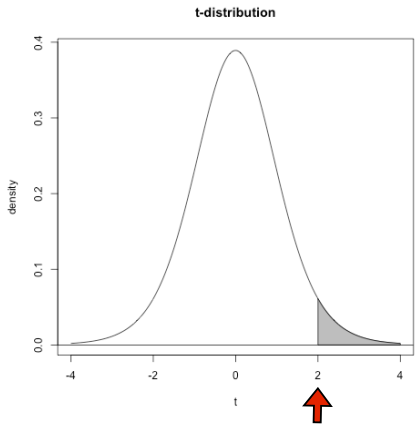
statistical mind



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

*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



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

研究の手順（危険な例）

実験を計画する

↓

実験を行う

↓

結果の分析

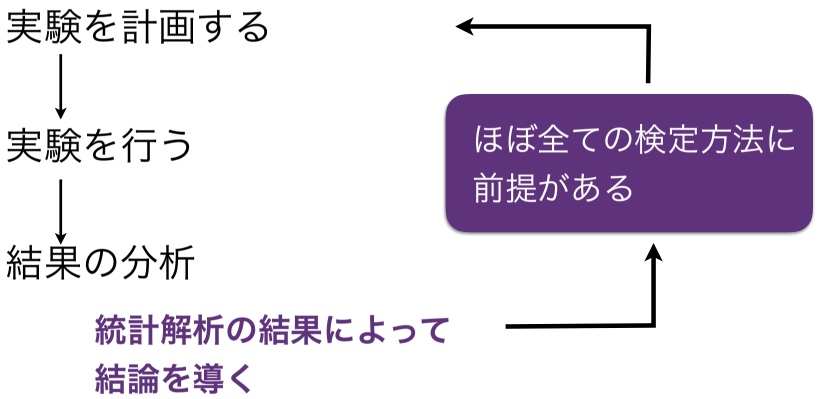
統計解析の結果によって
結論を導く

多重検定の補正

+ 統計検定における重要な思考

ポイント

現実には: 実験デザインはデータを
取得する「前」に練ってある必要がある



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

1. **Bonferroniタイプ**
2. **False discovery rate (FDR):**
   - Benjamini-Hochberg
   - Storey

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

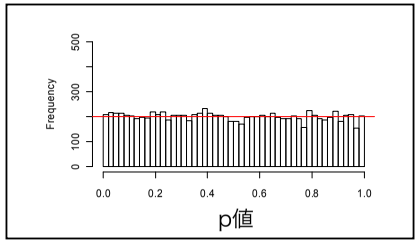
Bonferroniタイプの多重検定の補正

危険率を検定数で調整

危険率 = α/k

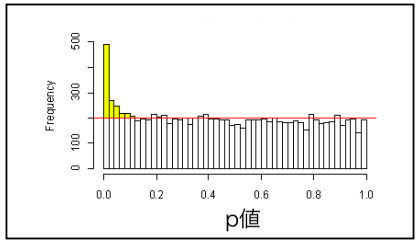
α: 元の危険率、
k: 検定数

False Discovery Rate (FDR)



帰無仮説

全ての範囲のp値が
同等の頻度で観察される
←どのp値を選んでも
ランダムに選ぶのと同じ



観察

あるp値（閾値）以下のp
値は有意な検定結果である
→では、ランダムに生じて
しまう各p値の頻度は？

ポイント

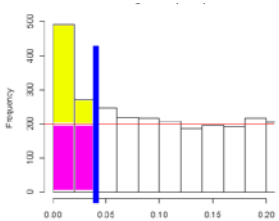
*p*値、*q*値の違い

*p*値の視点: **FP/(TN+FP)**

*q*値の視点: **TP/(TP+FP)**

		Statistical test: positive	Statistical test: negative
Real	+	True positive	False negative
Real	-	False positive	True negative

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
  - 確率分布
  - 自由度
  - p値

  } → 実験デザインの見直し、解析方法理解の基礎

- 統計解析の結果は確率に判断して得られたもの、トランスクリプトーム解析ではそれを多数行う
  → 多重検定の補正

- 検定方法、多重検定の補正における仮定
  例) 時系列データの比較にFDRは使えない

データのばらつきと
実験デザイン・統計学的観点

我々の実験対象の例

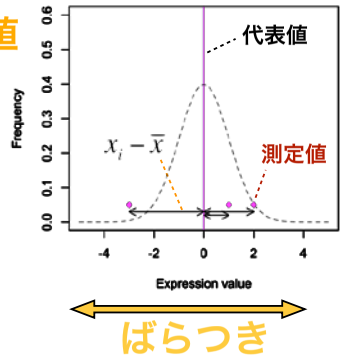
- ある遺伝子型の生物の
- ある環境での　　＋ 制御不能な実験要因
- ある遺伝子の発現量 ＋ 生化学反応のノイズ

我々に実行できる事

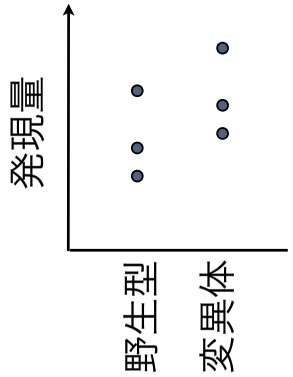
少数の測定値（**標本**）から
「**母集団**」を推定すること

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

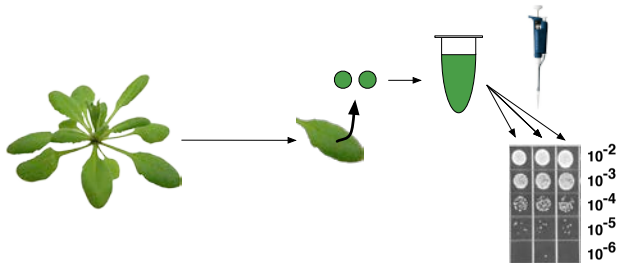


測定データはバラつく

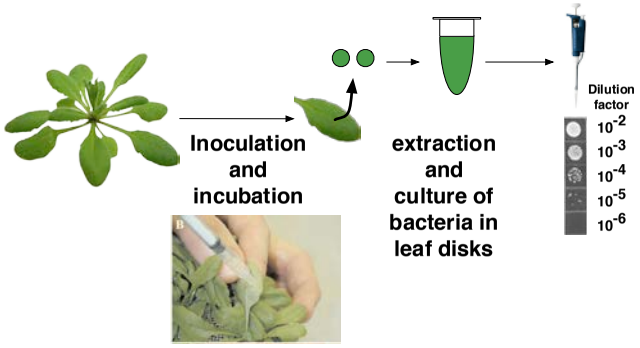


- **実験（測定）を反復する**
- 何を**「真」**と考えるか
- 論文として発表できるデータには**再現性**が必要

反復？

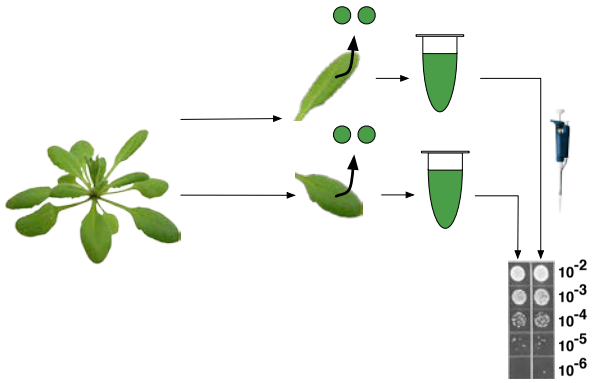


例: バクテリア増殖定量

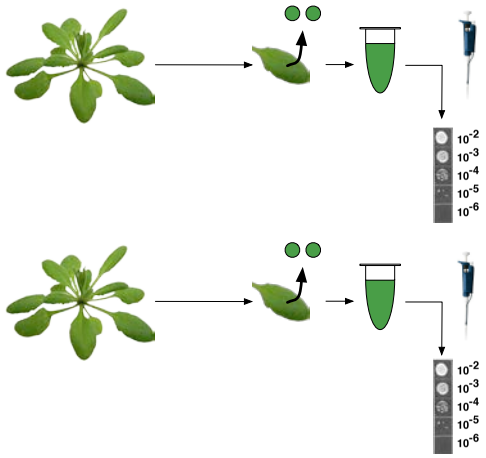


Katagiri, Thilmony R, and He S (2002) The Arabidopsis Thaliana-Pseudomonas Syringae Interaction. The Arabidopsis Book.

反復？



反復？



1データポイント:
潜在的に複数のばらつきを含む

- 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

1データポイント:
潜在的に複数のばらつきを含む

- 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

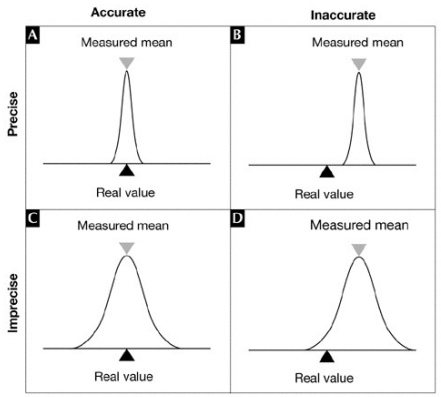
- Precision - 精度
- Accuracy - 正確度

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

ポイント



ある測定を無限に繰り返した際の測定値ヒストグラム

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

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

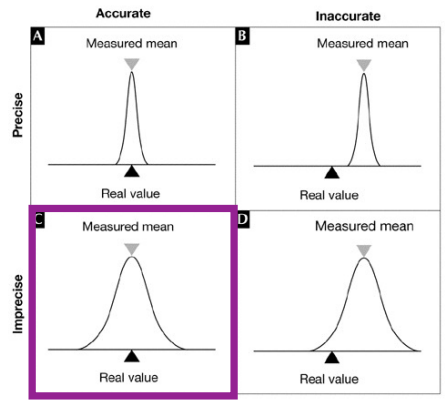
どちらが繰り返し測定することにより改善しうるか？

- **Precision - 精度**
  ある1測定を繰り返した際のばらつきの尺度
- **Accuracy - 正確度**

  ある測定値が「真の値」にどれだけ近いかの尺度

ポイント

Technical replicates: 測定値の改善



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

我々にできる事

少数の測定値（**標本**）から
「**母集団**」を推定すること

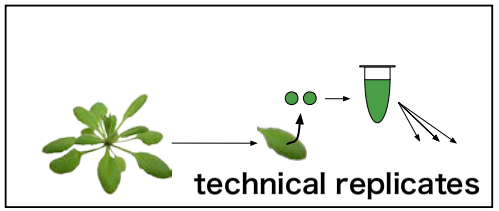
生体サンプルを繰り返し取る:
biological replicates

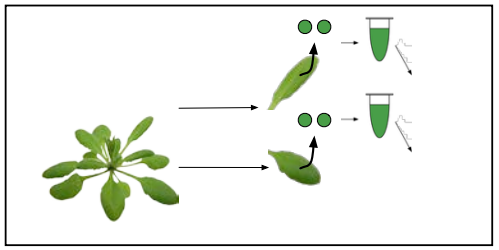
同一サンプルを繰り返し測る:
technical replicates

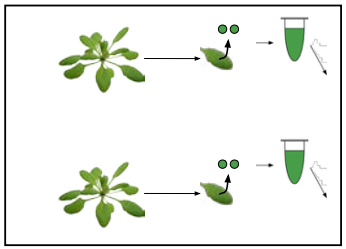
我々が1データポイントから得ているもの

- 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

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？







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?

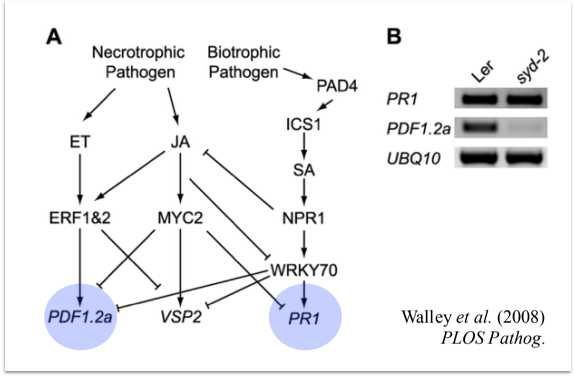
定量的測定が可能且つ要求される時代の再現性のあるデータとは何か？

- 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?~~

バラつきの定量と説明変数への割当て

“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？



明瞭な違いを示す遺伝子:
明瞭な再現性

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）

↓ **前処理: 線形モデル**

発現データ（バイアス除去）

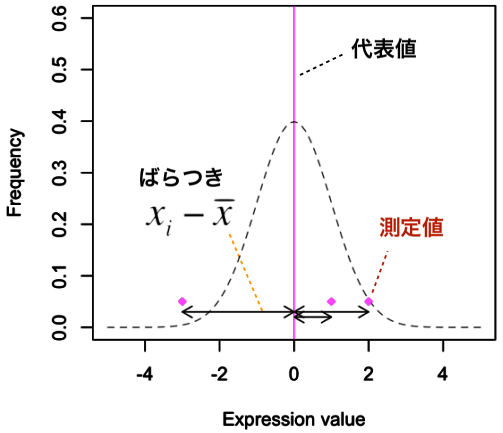
↓

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

リマインド:

母集団を推定する統計量



目標

- 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

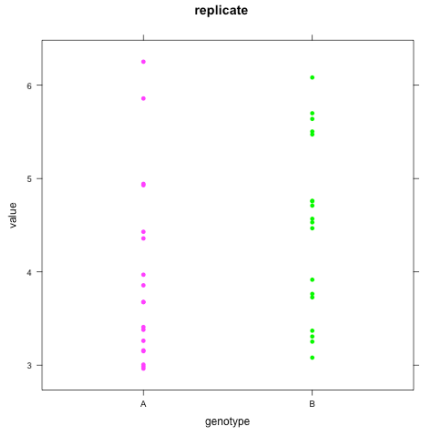
t検定: 平均値の検定

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

偏差: 平均値からのばらつき

あるRT-qPCR実験

- 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:

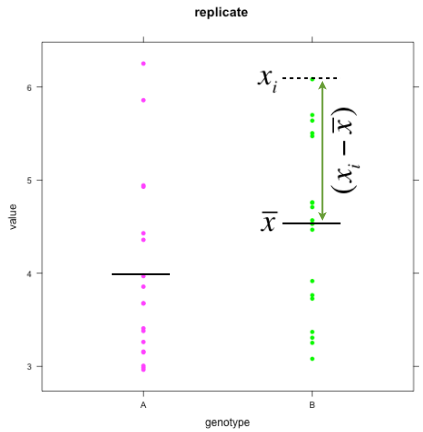
  計18個/ genotype

線形モデルの枠組みで考えてみる

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

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

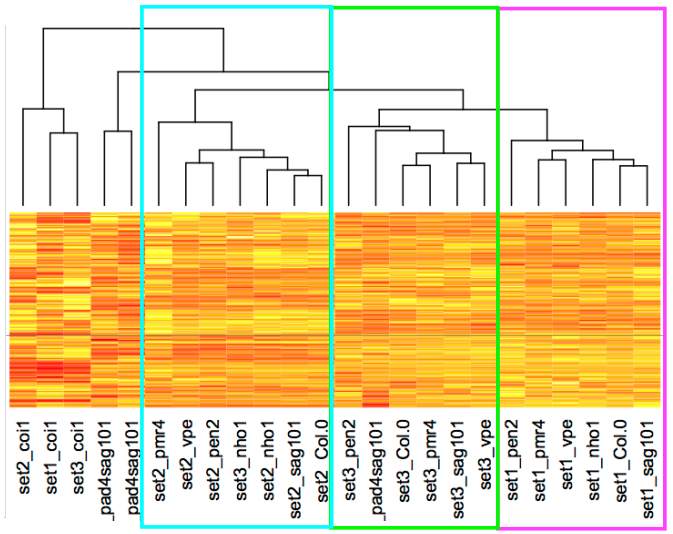
残差 (観察値-推定値):
想定要因では説明できない
データの変動



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

考慮するのは1要因で良いか？

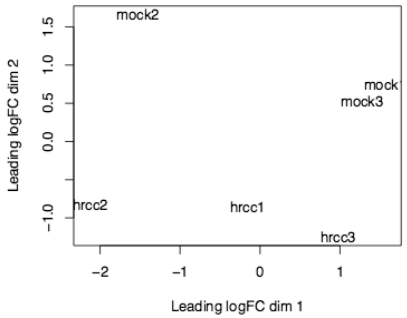
ポイント



ポイント

“トランスクリプトーム”測定

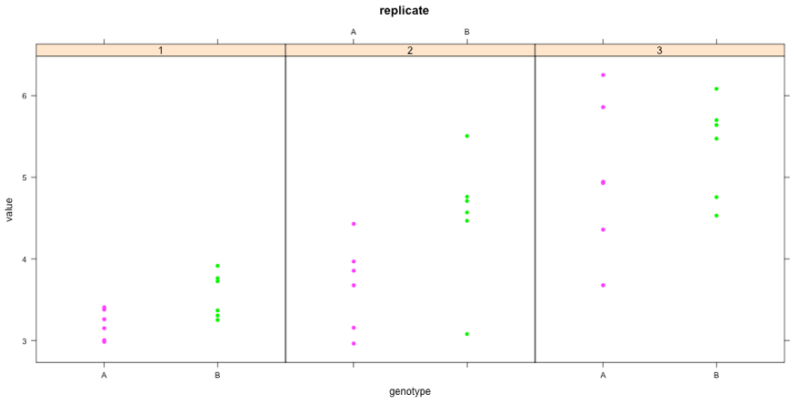
- 何が再現されうるか？再現されたとするか？



網羅的測定:
再現性の
再定義

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

例: 2遺伝子型の測定を3反復したデータ



観察値を複数要因の影響に起因するものとして分解

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

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

↓

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

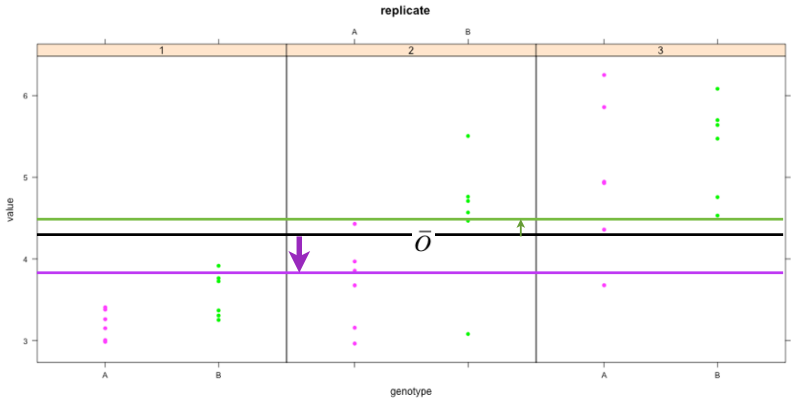
***genotype*と*replicate*の影響を同時に考えられないか？**

線形モデルの仕組み

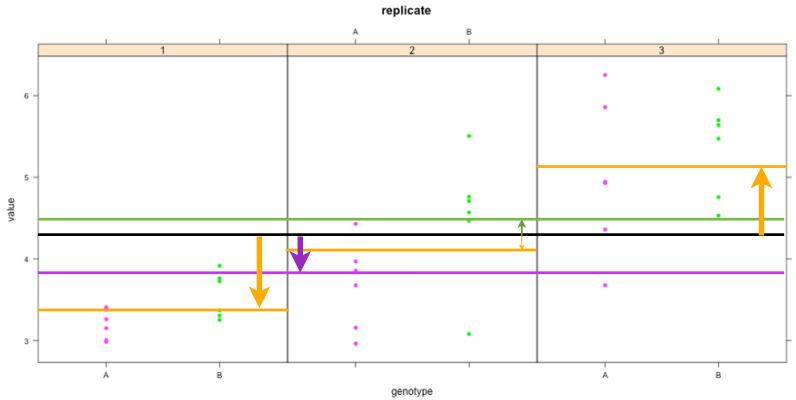
$$O_{ij} = x_i + y_j + \varepsilon_{ij}$$

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

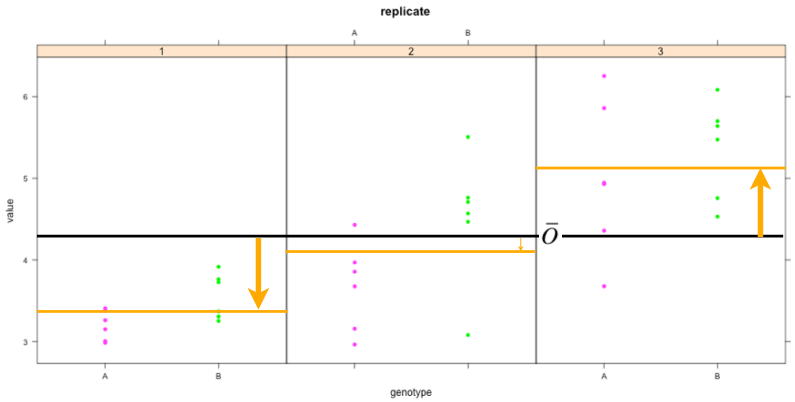
$(\bar{x}_{i\bullet} - \bar{O})$遺伝子型による変動



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



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



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・論文での書き方

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

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Rの関数を使う

ポイント

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

- -omicsデータは“**batch effect**”と呼ばれる体系的なバイアスが混入する。
  例: 実験時期、実験者、餌


OPINION

Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly and Rafael A. Irizarry

Nature Reviews Genetics (2010) 11, 733-

- 線形モデルで推定・除去

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？

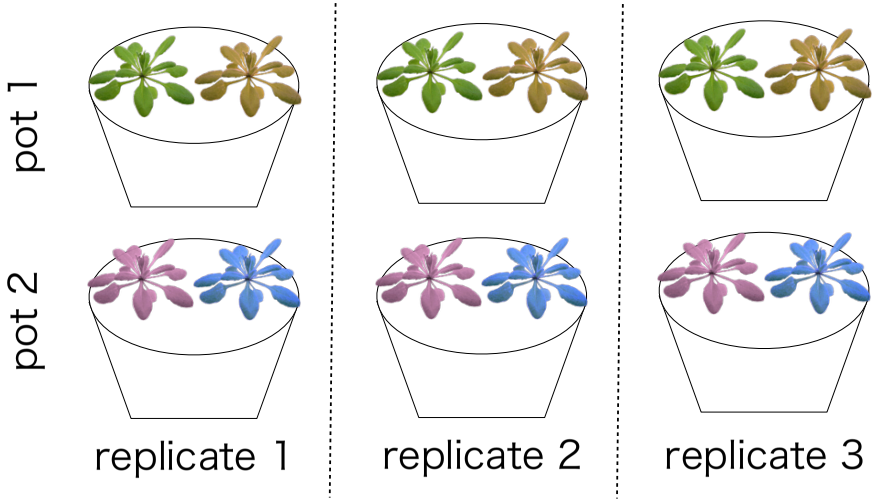
- 何が再現されうるか？再現されたとするか？

~~- 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?~~

バラつきの定量と説明変数への割当て

実験デザインの重要性:

genotype+replicate+potモデルを当てはめるには？



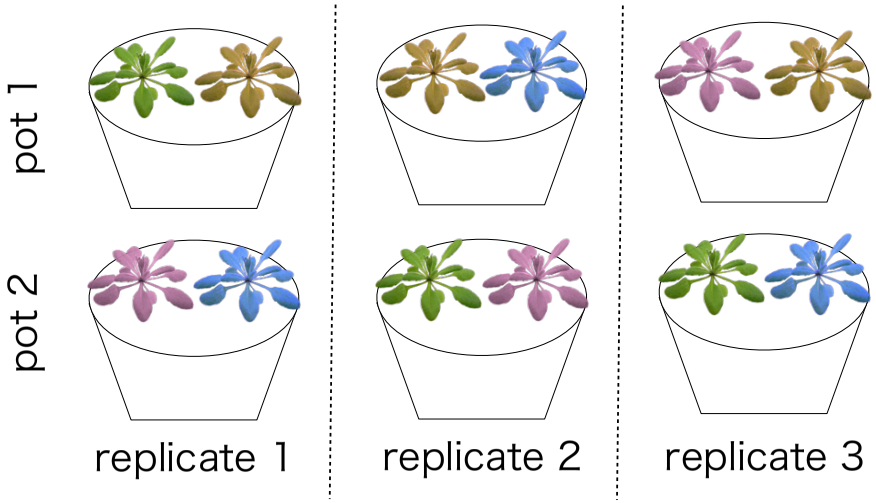
ポイント

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。

実験デザインの重要性:

genotype+replicate+potモデルを当てはめるには？



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
- 理論を理解するのは難しいかもしれないが、実行はRで簡単に行える。理解に努める努力と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