

私が重視しているポイント

統計学入門

北海道大学 農学研究院
数理・データサイエンス
教育研究センター
佐藤昌直

- 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

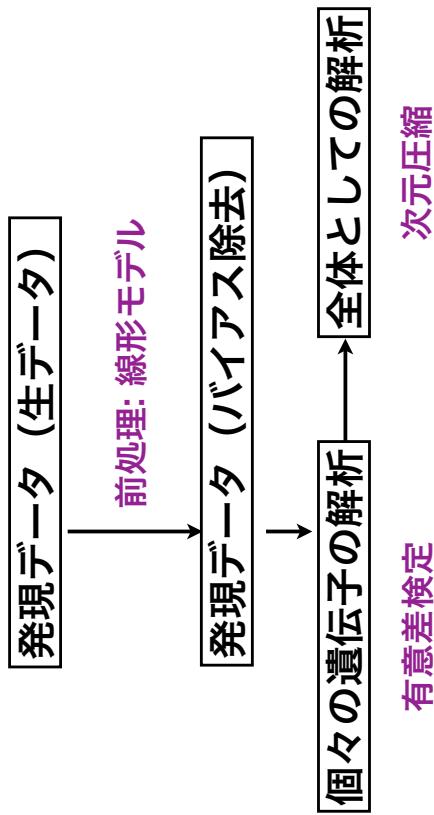
基本的な統計の用途

これらを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

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

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

ねらい

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

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

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？

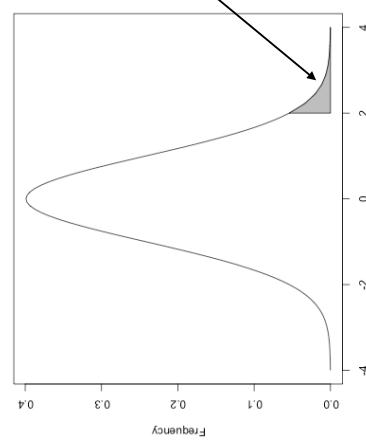
例2. 野生型と変異体Aの遺伝子発現プロファイル間の相関係数は0.35だった。これらは有意に相關していると考えられるか？

3. 確率分布と照らし合わせる

**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



統計量

統計における検定の手続き

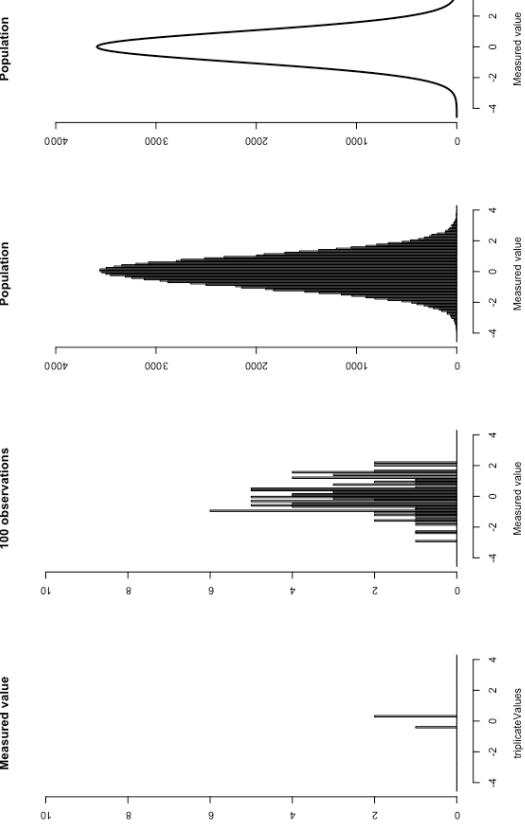
1. 仮説を立てる

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

4. 判定：帰無仮説が棄却されるか？

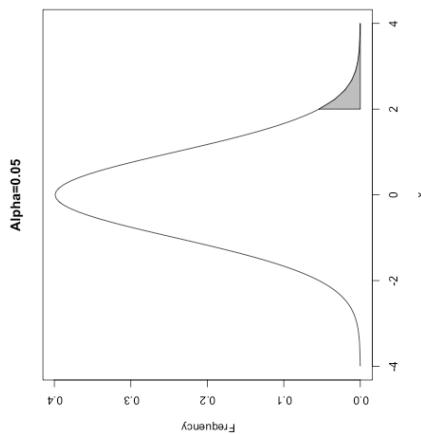


確率分布？面積？

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



統計的検定の手続き

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

2. 統計量を求める：

ポイント

**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

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
ヘ (ハット) は
推定を表す

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

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

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でか？

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

(ほぼ全ての検定方法に
前提がある)

ばらつき: 分散 / 偏差

分散:

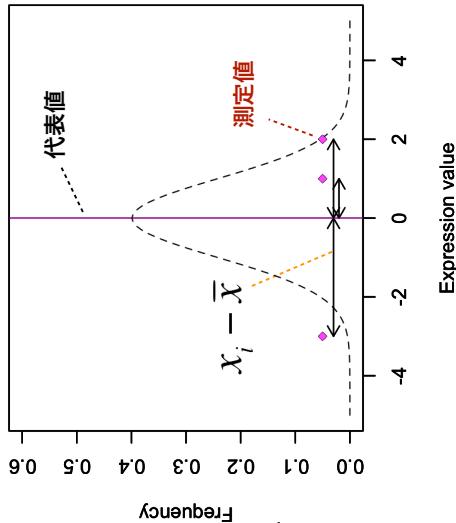
$$\sum_{i=1}^n (x_i - \bar{x})^2$$

$$\frac{1}{n-1}$$

標準偏差:

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

$$\frac{1}{n-1}$$

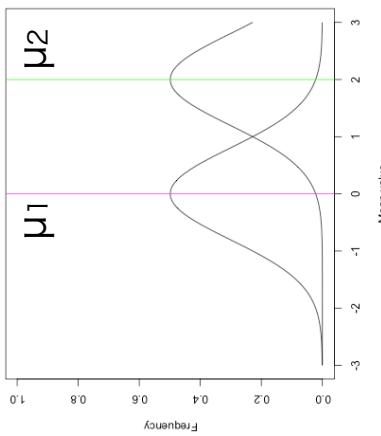


ポイント

検定:

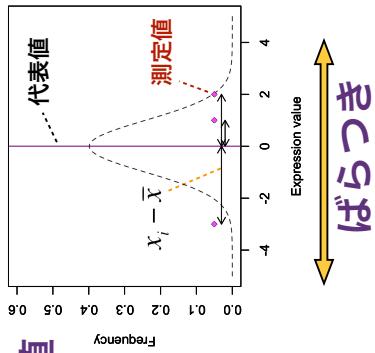
2サンプルの平均の検定

- 平均値 = μ_1, μ_2
- データは正規分布



母集団を推定する統計量

1. (真の値に近い) 代表値



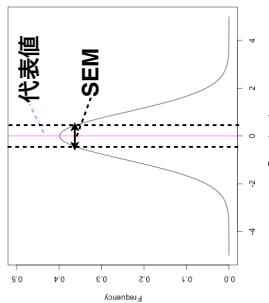
2. ばらつきの範囲

統計量その2：

平均値もあくまで推定値

(平均) 標準誤差：「統計量」の偏差

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



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

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

નોંધ

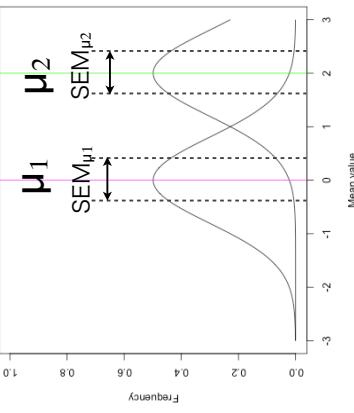
statistical
mind

統計量その3:

平均の差とその誤差

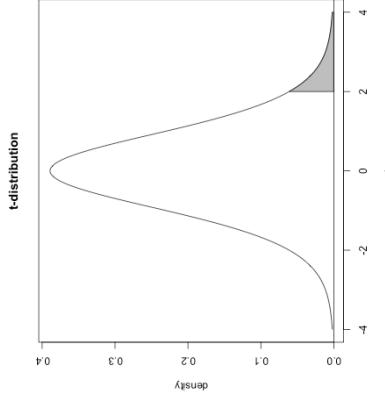
統計量

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



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

statistical mind

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

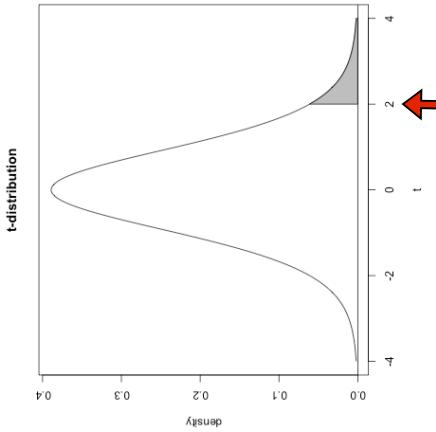
→正規分布ではなく、t分布

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

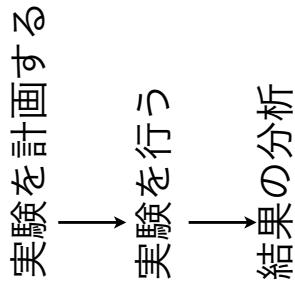
例) 3つの観察で得られた平均値と10観察から
得られた平均値はどちらが確かしいか

p値とは：

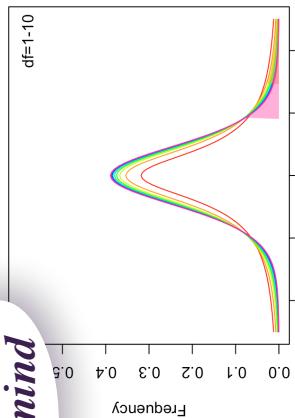
- 標本に基づいた統計量が帰無仮説の下、起きうる確率
- 多くの場合、0.05が危険率



研究の手順（危険な例）



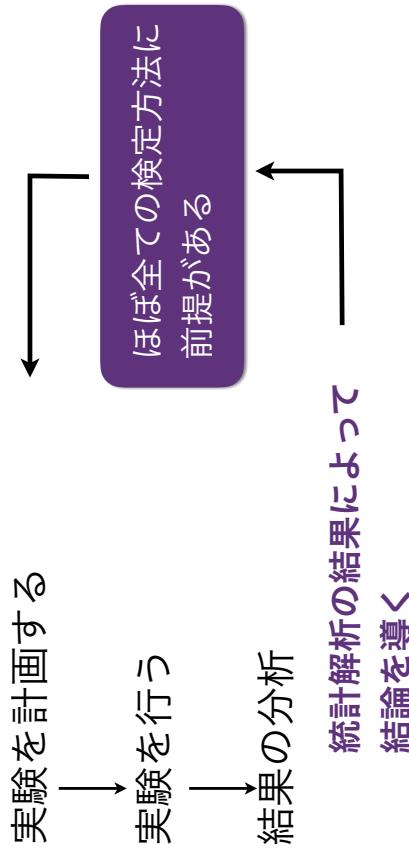
ポイント



統計解析の結果によつて
結論を導く

ポイント

現実には：実験デザインはデータを
取得する「前」に練つてある必要がある



ほぼ全ての検定方法に
前提がある

ポイント

ex. t 検定: 正規分布、等分散

どの確率分布を想定する？

連続値：正規分布、ガンマ分布（非負）

離散値（カウントデータ）：

ポアソン分布（平均=分散= λ ）

負の二項分布（ λ がガンマ分布）

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

多重検定の補正

+ 統計検定における重要な思考

*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

多重検定の補正の必要性

- $\alpha = 0.05$ 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

多重検定の補正

1. Bonferroniタイプ

危険率を検定数で調整

2. False discovery rate (FDR):

- Benjamini-Hochberg [R:p.adjust]
- Storey [R:qvalue]

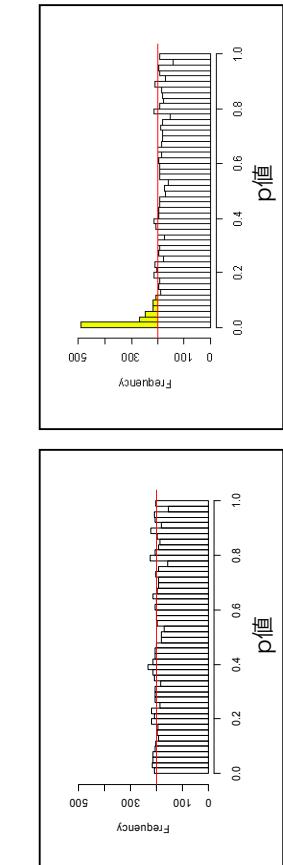
Bonferroniタイプの多重検定の補正

危険率 = α / k

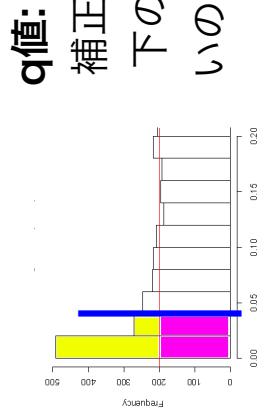
α : 元の危険率、
k: 検定数

False Discovery Rate (FDR)

False Discovery Rate (FDR)



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生じてしまう各p値の頻度は？



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。

p値、q値の違い

$$\begin{aligned} p\text{値の視点: } & \mathbf{FP}/(\mathbf{TN}+\mathbf{FP}) \\ q\text{値の視点: } & \mathbf{TP}/(\mathbf{TP}+\mathbf{FP}) \end{aligned}$$

Statistical test

positive negative

実験結果	True positive	False negative
+	True positive	False negative
-	False positive	True negative

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

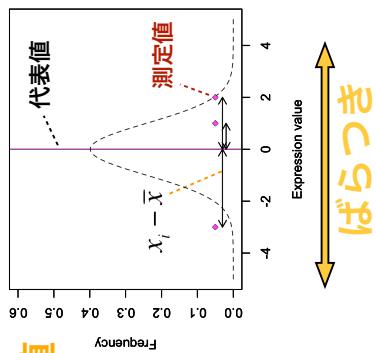
ポイント

母集団を推定する統計量

データのばらつきと 実験デザイン・統計学的観点

1. (真の値に近い)代表値

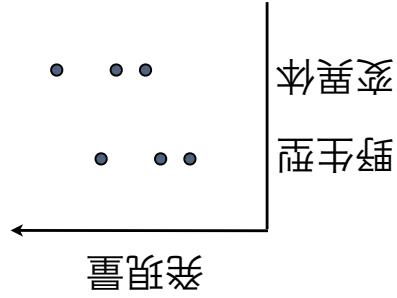
2. ばらつきの範囲



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

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

我々にできる事

少數の測定値（標本）から
「母集団」を推定すること

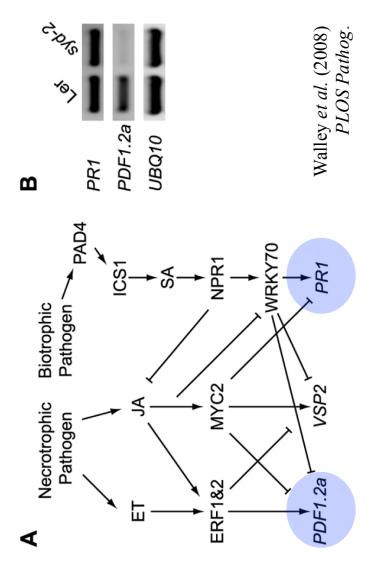
生体サンプルを繰り返し測る:
biological replicates

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？

非NGS測定：“マーカー遺伝子”測定

- 何が再現されるか？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

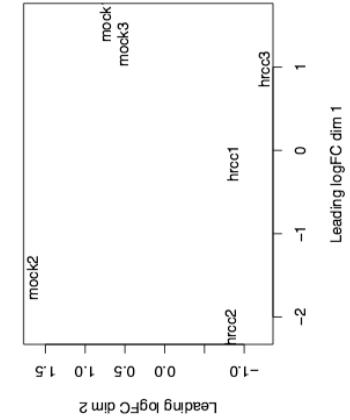
- 何が再現されるか？再現されたとするか？

網羅的測定：
**再現性の
再定義**

ポイント

“トランスクリプトーム”測定

- 何が再現されるか？再現されたとするか？



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？

- 何が再現されるか？再現されたとすると何が？
 - 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

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

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

解析の流れ

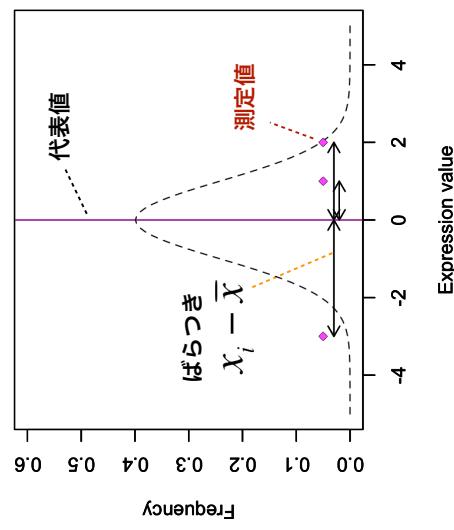


有意差検定

次元圧縮

リマインド:

母集団を推定する統計量



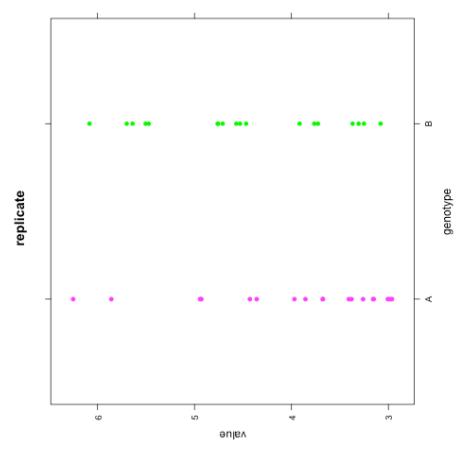
t検定: 平均値の検定

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

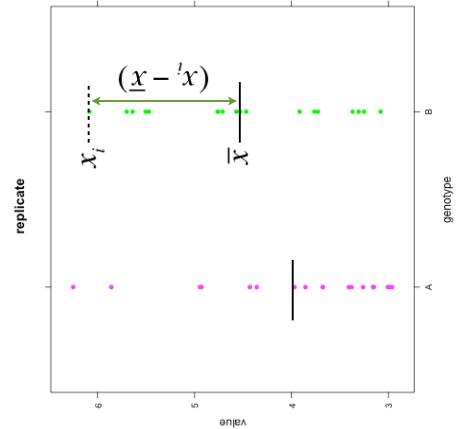
偏差: 平均値からのばらつき

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



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



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

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

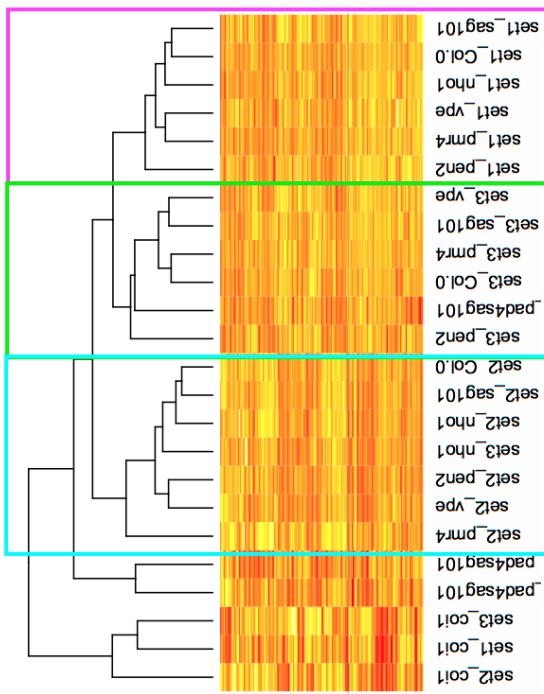
考慮するのは1要因で良いか？

線形モデルの枠組みで考えてみる

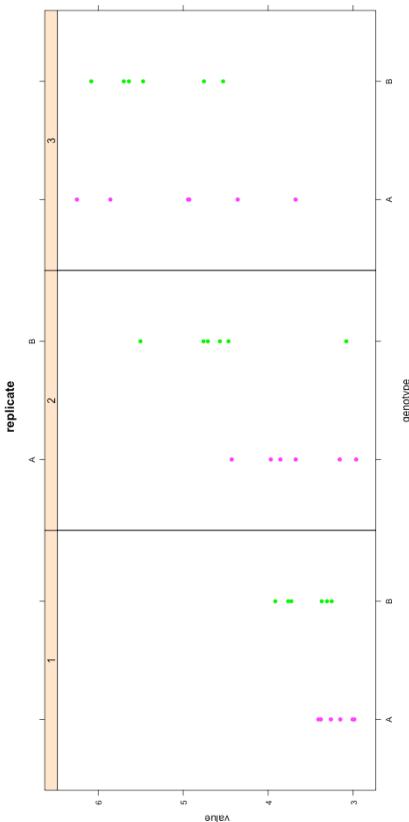
$$x_i = \bar{x} + (x_i - \bar{x})$$

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

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



例：2遺伝子型の測定を3回複したデータ



観察値を複数要因の
影響に起因するものとして分解

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

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

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

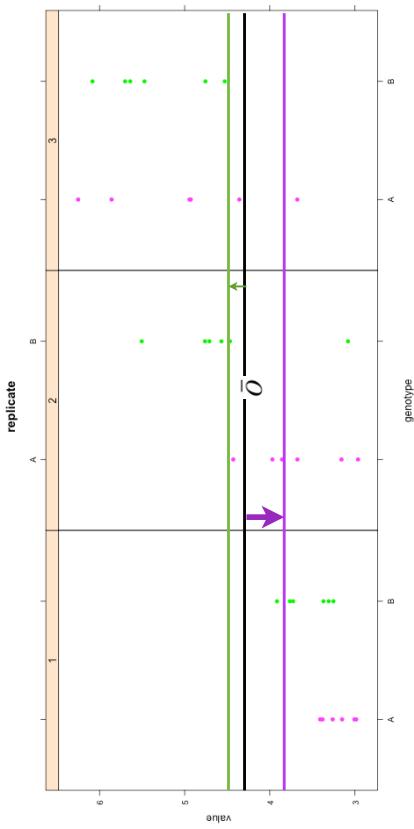
$$O_{ij} = x_i + y_j + \varepsilon_{ij}$$

線形モデルの仕組み

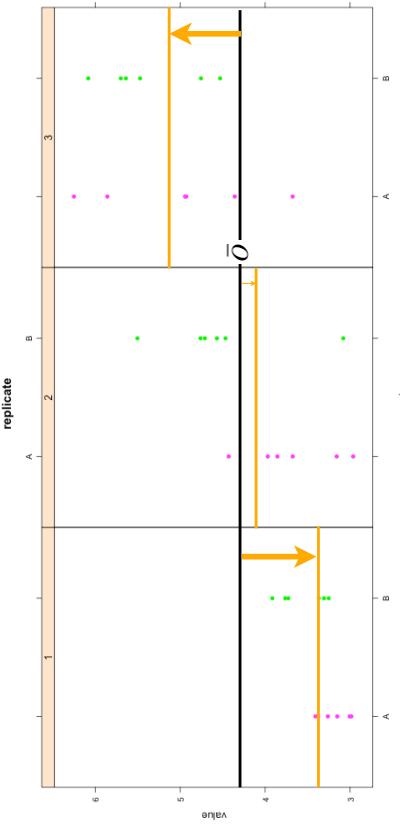
$$O_{ij} = X_i + Y_j + \varepsilon_{ij}$$

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

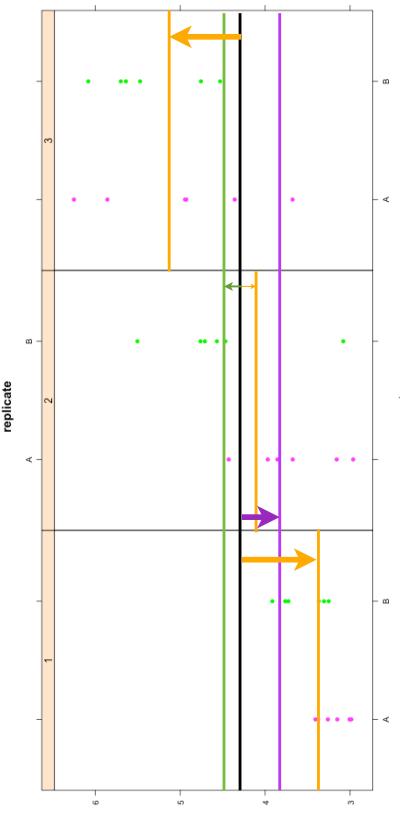
$(\bar{x}_{i\bullet} - \bar{O})$ 遺伝子型による変動



$(\bar{y}_{\bullet j} - \bar{O})$ 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

$$\begin{aligned} O_{ij} &= x_i + y_j + \varepsilon_{ij} \\ O_{ij} &= \bar{O} + (\bar{x}_i - \bar{O}) + (\bar{y}_j - \bar{O}) + \varepsilon_{ij} \\ &\quad \downarrow \text{教科書・論文での書き方} \\ O_{ij} &= \mu + \alpha_i + \beta_j + \varepsilon_{ij} \\ &\quad \swarrow \quad \downarrow \end{aligned}$$

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

$O_{ij} = \bar{O} + (\bar{x}_i - \bar{O}) + (\bar{y}_j - \bar{O}) + \varepsilon_{ij}$

と観察値を説明する（かもしけない）
変数でそれらの関係性を書き下すこと

- 実際には: Rでlm, glmなどの関数を使う

ポイント

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

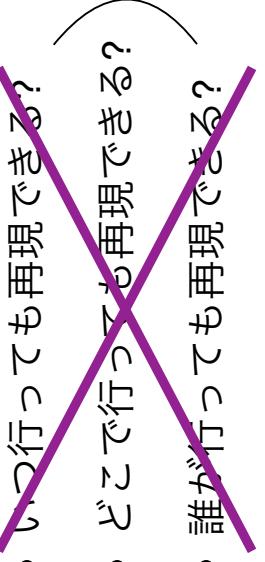
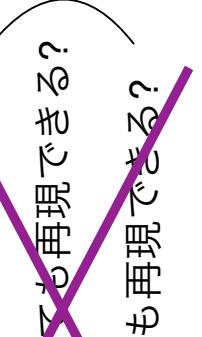
Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Strainic, Benjamin Langmead, W. Evan Johnson, Donald Ceman, Keith Baggerly and Rafael A. Irizarry

Nature Reviews Genetics (2010) 11, 733-

- 線形モデルで推定・除去

- 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

定量的測定が可能な且つ要求される時代の 再現性のあるデータとは何か？

- 何が再現されるか？再現されたとすると何か？

 - いつ行っても再現できる？

 - どこで行つても再現できる？

 - 誰が行っても再現できる？

- 体系的な
バラつきの
定量と
説明変数への
割当て**

R (edgeR) での実装

```
> x <- read.delim("TableOfCounts.txt", row.names="Symbol")
> group <- factor(c(1,1,2,2))
> y <- DGEList(counts=x, group=group)
> y <- calcNormFactors(y)
> design <- model.matrix(~group)
> y <- estimatedDisp(y,design)
> fit <- glmFit(y,design)
> lrt <- glmLRT(fit,coef=2)
> topTags(lrt)
```

Chen, et al., edgeR User's Guide (December 26, 2017)

ポイント

Rを用いた線形モデルにおける

実験デザイン指定: factor, model.matrix

```
> x <- read.delim("TableOfCounts.txt", row.names="Symbol")
> group <- factor(c(1,1,2,2))
> y <- DGEList(counts=x, group=group)
> y <- calcNormFactors(y)
> design <- model.matrix(~group)
> y <- estimatedDisp(y,design)
> fit <- glmFit(y,design)
> lrt <- glmLRT(fit,coef=2)
> topTags(lrt)
```

Chen, et al., edgeR User's Guide (December 26, 2017)

ポイント

model.matrixで生成される出力

```
group      <- factor(rep("M", 3), rep("H", 3)))
replicates <- factor(c(1:3, 1:3))
model.matrix(~group+replicates)

(Intercept) groupM replicates2 replicates3
1           1       1       0       0
2           2       1       1       0
3           3       1       1       1
4           4       1       0       0
5           5       1       1       0
6           6       1       0       1
attr(,"assign")
[1] 0 1 2 2
attr(,"contrasts")
attr(,"contrasts")$group
[1] "contr.treatment"
```

0と1の行列
contrasts

`model.matrix`で生成される出力

線形モデルとmodel.matrixの関係

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

↓
 i, j を書き下すと

$$O_{M,1} \sim \mu + \alpha_M + \beta_1 + \varepsilon_{M,1}$$

$$O_{M,2} \sim \mu + \alpha_M + \beta_2 + \varepsilon_{M,2}$$

$$O_{M,3} \sim \mu + \alpha_M + \beta_3 + \varepsilon_{M,3}$$

$$O_{H,1} \sim \mu + \alpha_H + \beta_1 + \varepsilon_{H,1}$$

$$O_{H,2} \sim \mu + \alpha_H + \beta_2 + \varepsilon_{H,2}$$

$$O_{H,3} \sim \mu + \alpha_H + \beta_3 + \varepsilon_{H,3}$$

ポイント

ポイント

線形モデルとmodel.matrixの関係

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

$$O_{M,1} \sim \mu + \alpha_M + \beta_1 + \varepsilon_{M,1}$$

$$O_{M,1} \sim \mu \times 1 + \alpha_M \times 1 + \alpha_H + \beta_1 + \beta_2 \times 0 + \beta_3 \times 0 + \varepsilon_{M,1}$$

$$O_{M,1} \sim \mu \times 1 + \alpha_M \times 1 + \beta_2 \times 0 + \beta_3 \times 0 + \varepsilon_{M,1}$$

0 0 1 1

ポイント

推定する係数の数: 6

$$\mu, \alpha_M, \alpha_H, \beta_1, \beta_2, \beta_3$$

推定したい係数の数よりも
観察数が多くなくてはなら
ない)
contrasts: 1番目の水準の
係数を0として残りと比較
→係数の数を削減

$$\sigma_{H,3} \sim \mu + \alpha_H + \beta_3 + \varepsilon_{H,3}$$

57

model.matrixまとめ

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

(Intercept) groupM replicates2 replicates3

1	1	0	0
2	1	1	0
3	1	1	1
4	1	0	0
5	1	0	1
6	1	0	0

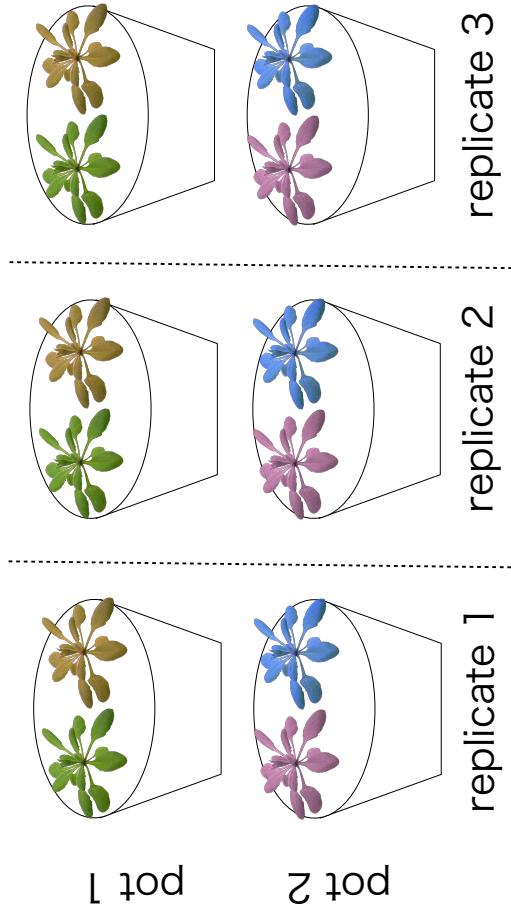
ポイント

- 0と1の意味
- (この場合の) contrastsの概念、 $\mu=\text{replicate1}$ の処理Hの係数
- 観察数、実験デザインとの関連

ポイント

実験デザイン的重要性:

genotype+replicate+potモデルを当てはめるには？



ポイント

実験デザイン的重要性

• 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
(発展学習: 無作為化)

• 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。一方、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

(少しだけ) 線形モデル→一般化線形モデル

【予測】

実現象に即し、データにあてはまるモデル

どの確率分布を想定する？

連続値：正規分布 [R:lm]

離散値（カウントデータ）：

負の二項分布 [R:glmFit, glm.nb]

- 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ないが、実行はRで簡単にできる。理解に努める努力と実験デザインと運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）：contrast、切片
- 実験計画法
- 交互作用
- Bioconductor: limma、edgeR/ペッケージ