統計学入門

北海道大学大学院農学研究院
（兼）数理・データサイエンス
教育研究センター
佐藤昌直

この“統計学入門”のスコープ

- ゲノムワイドな測定データを扱うポイントを掴み
- 統計的な視点で考えられるようになる

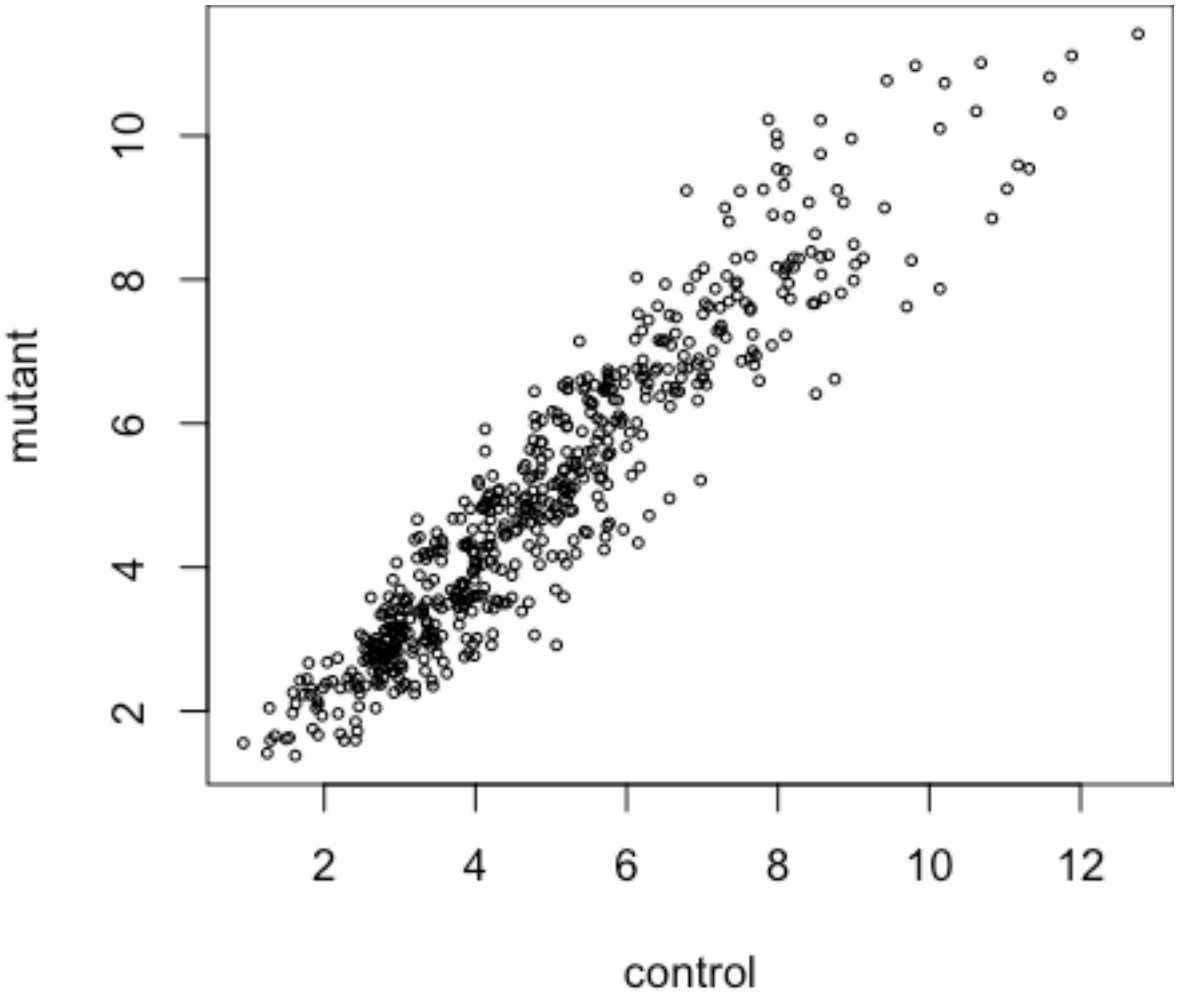
この“統計学入門”のスコープ

- 多パラメーターの分析
  （多変量解析・可視化）
- 基本的な統計の概念・知識

+ 実践・自習に向けて

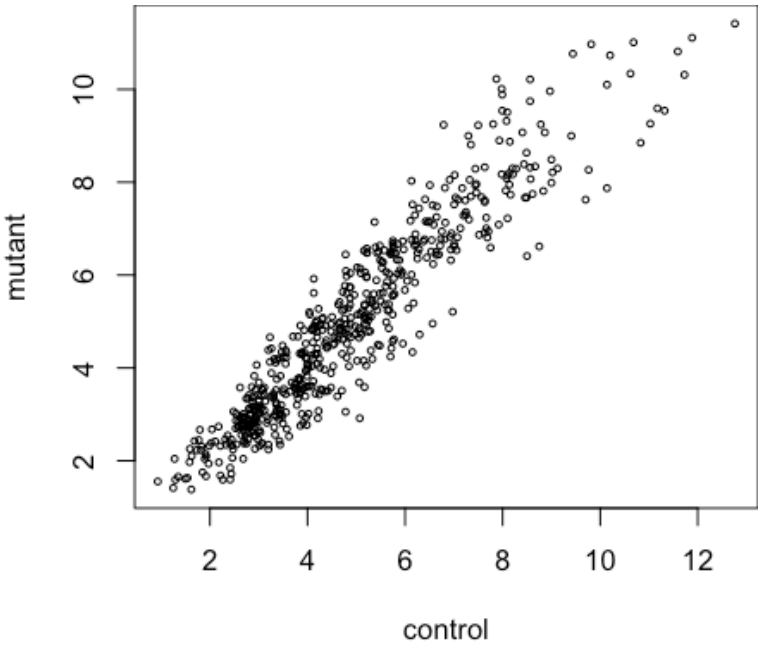
- 専門用語・概念を使えるようになる/頭に入れる
  → 生成AIとの協働
- 数式を読む
  → 教科書を読めるように
    → 言語化されていない解釈を導く

RNA-seq: ゲノムワイドな
トランスクリプトームデータ



問い

- 発現変動遺伝子群（differentially expressed genes, DEG)
- システム解析
- ….



解析方針の整理

- 実験データの質　**再現性/ばらつきの統計量**
- データ全体の吟味　**多変量解析**
- パラメーターごとの評価　**検定、多重検定の補正**

ポイント

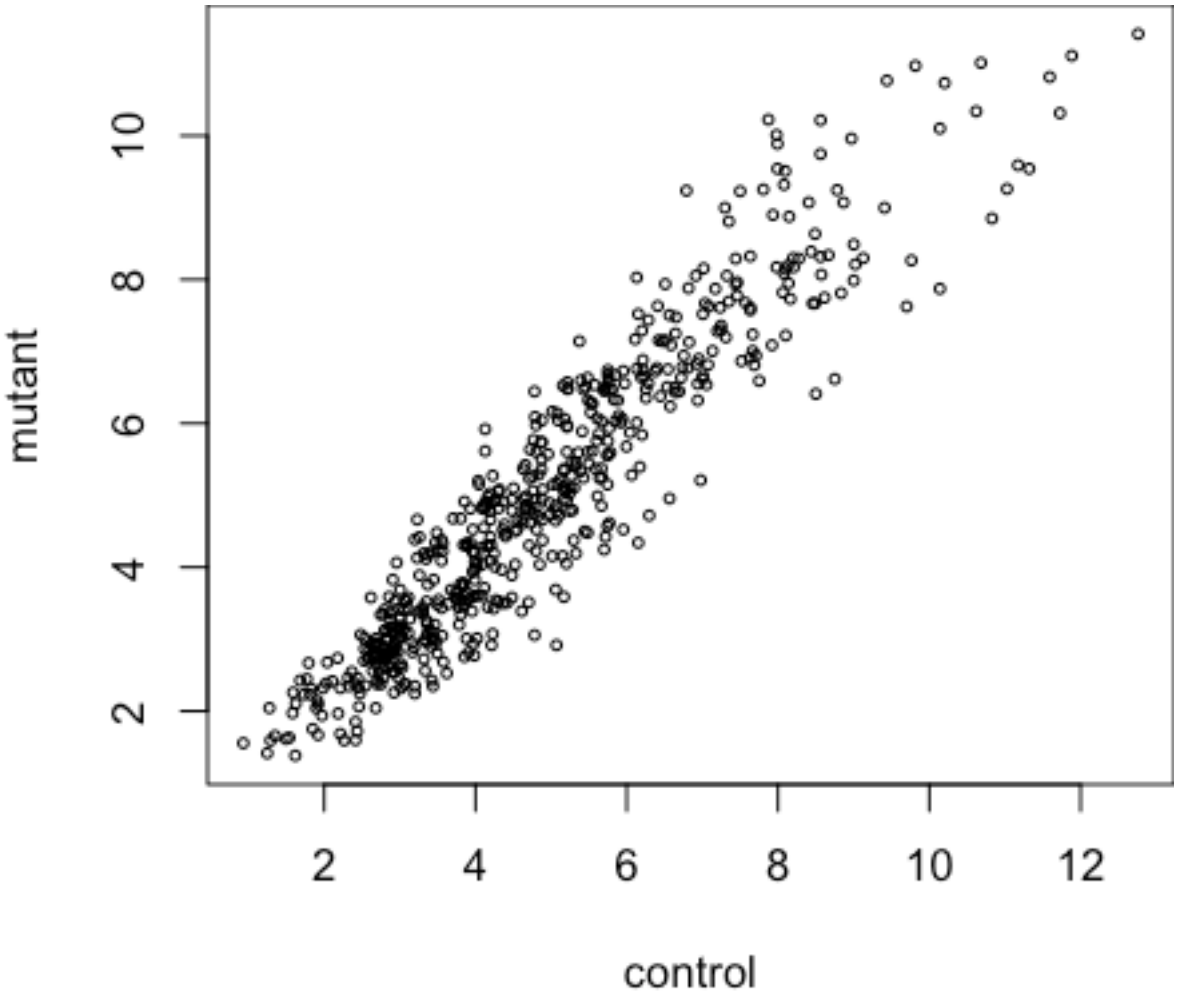
用語の整理: 統計量

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

再現性・ばらつきの統計量: 相関係数



再現性・ばらつきの統計量: 相関係数

$$\frac{(a_1-\bar{a})(b_1-\bar{b})+(a_2-\bar{a})(b_2-\bar{b})+(a_3-\bar{a})(b_3-\bar{b})}{\sqrt{(a_1-\bar{a})^2+(a_2-\bar{a})^2+(a_3-\bar{a})^2}\sqrt{(b_1-\bar{b})^2+(b_2-\bar{b})^2+(b_3-\bar{b})^2}}$$

教科書・論文での表記：数列・配列

$$\frac{(a_1-\bar{a})(b_1-\bar{b})+(a_2-\bar{a})(b_2-\bar{b})+(a_3-\bar{a})(b_3-\bar{b})}{\sqrt{(a_1-\bar{a})^2+(a_2-\bar{a})^2+(a_3-\bar{a})^2}\sqrt{(b_1-\bar{b})^2+(b_2-\bar{b})^2+(b_3-\bar{b})^2}}$$

$$a_1-\bar{a}$$

サンプルAの

I番目の要素

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

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

-（バー）は平均を表す
^（ハット）は推定を表す

→ 内山　R入門「基本統計量の計算」

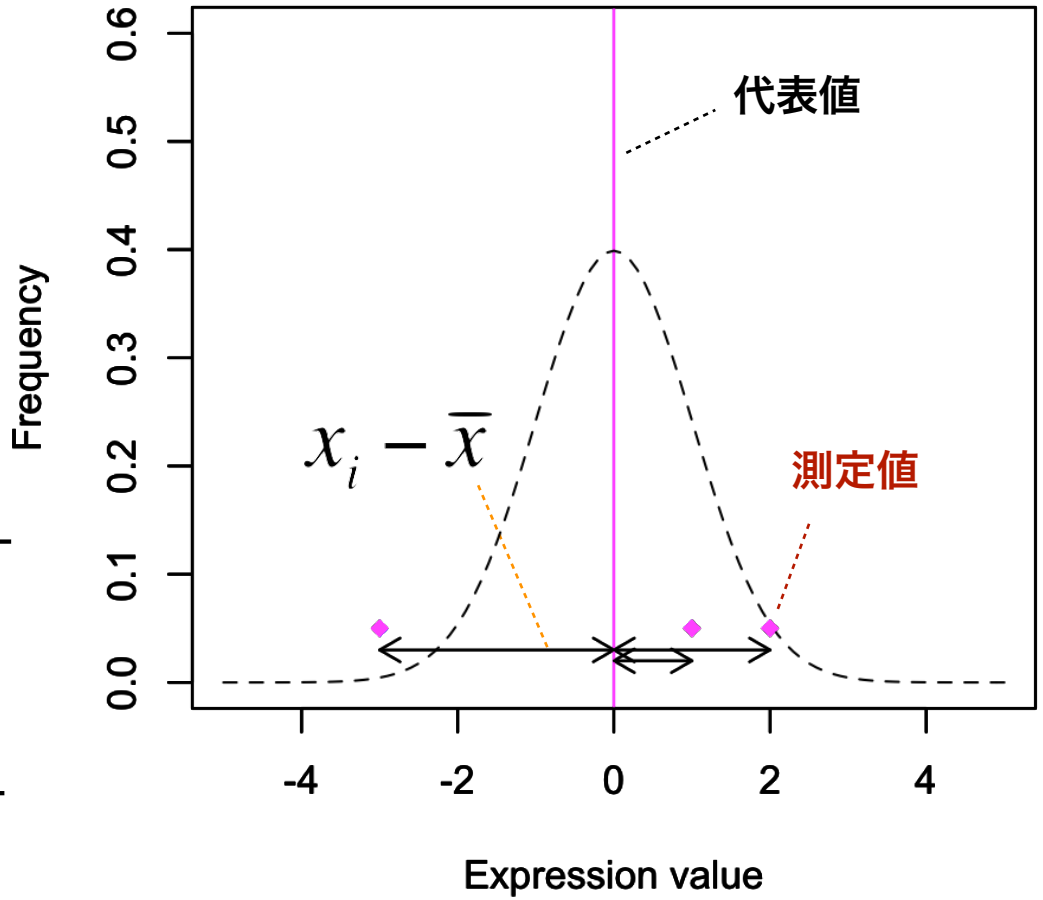
ばらつき: **分散／偏差**

分散:

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

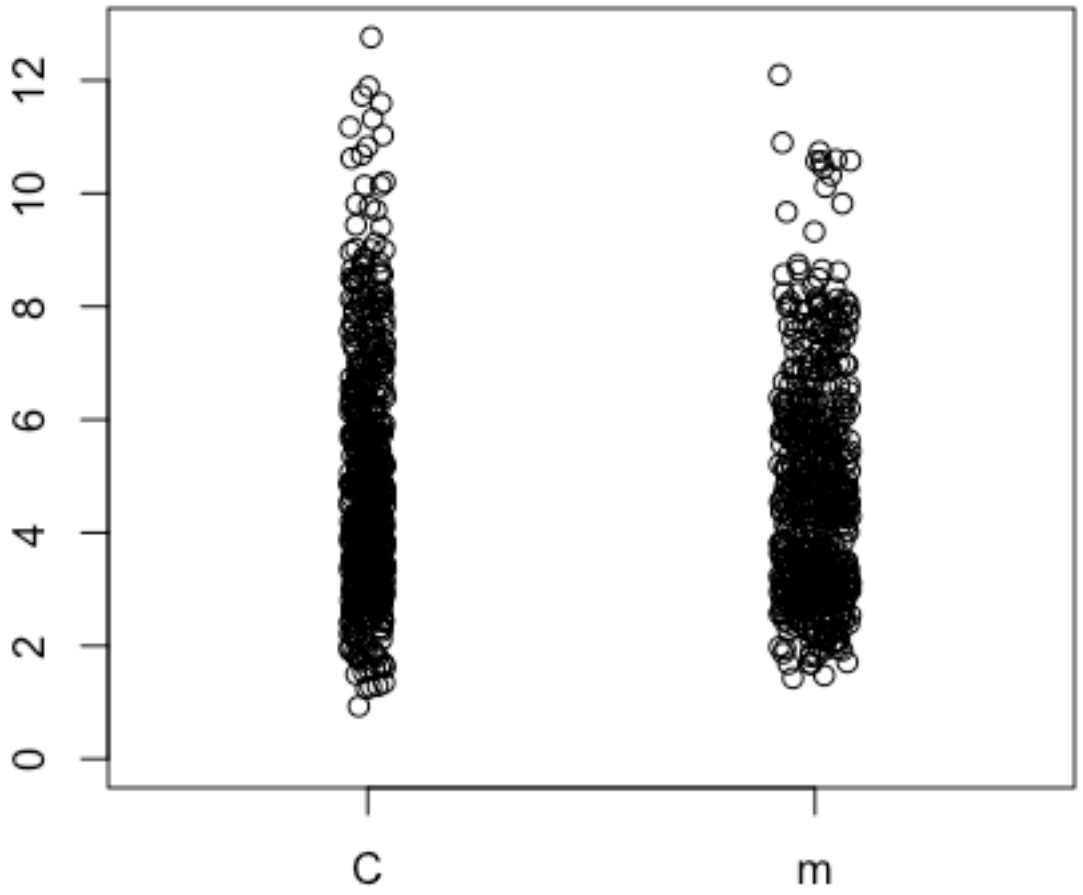
標準偏差:

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



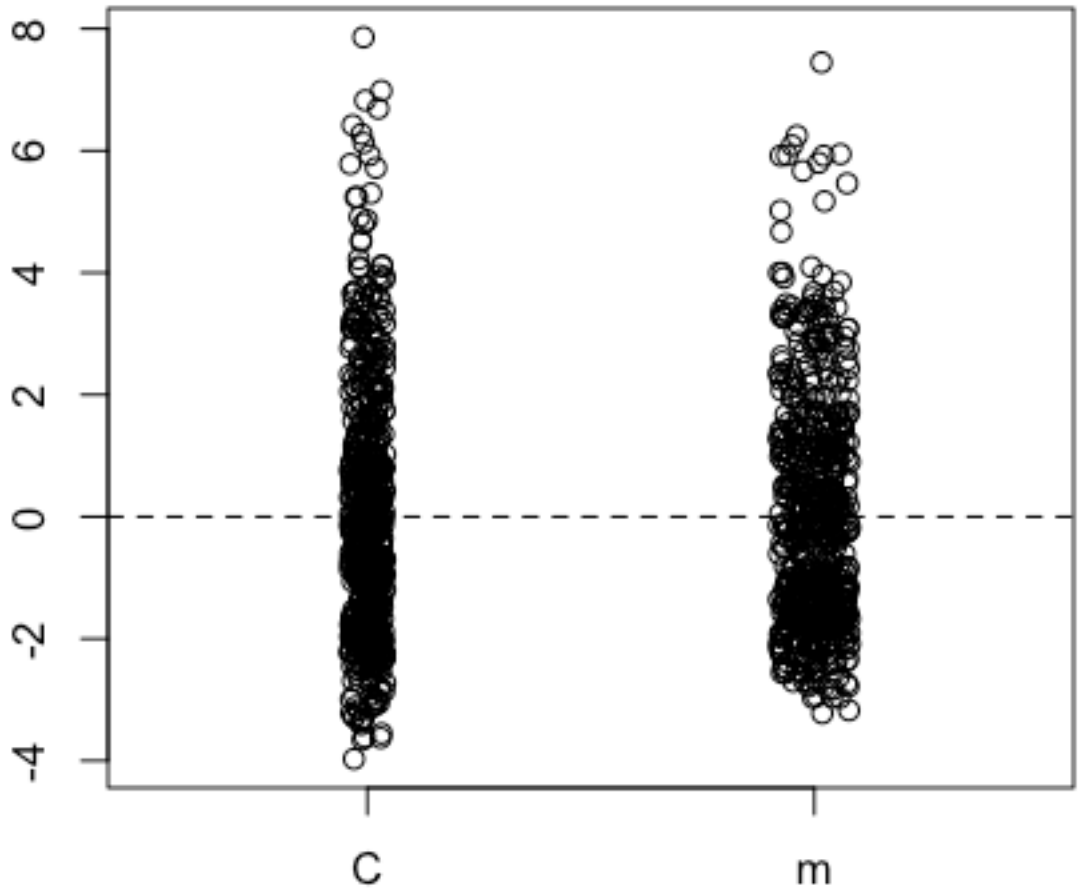
→ 内山　R入門「基本統計量の計算」

RNA-seq: normalizationしたデータ



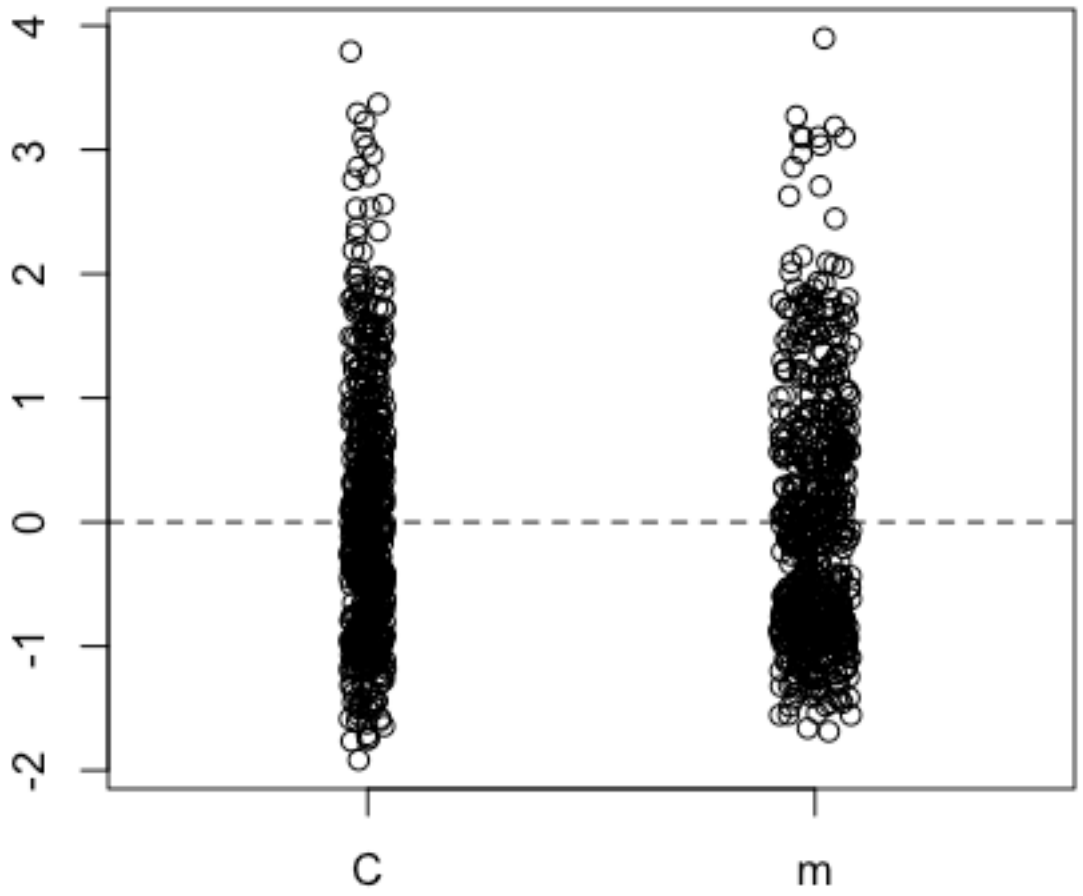
RNA-seq: センタリングしたデータ

$$a_i - \bar{a}$$



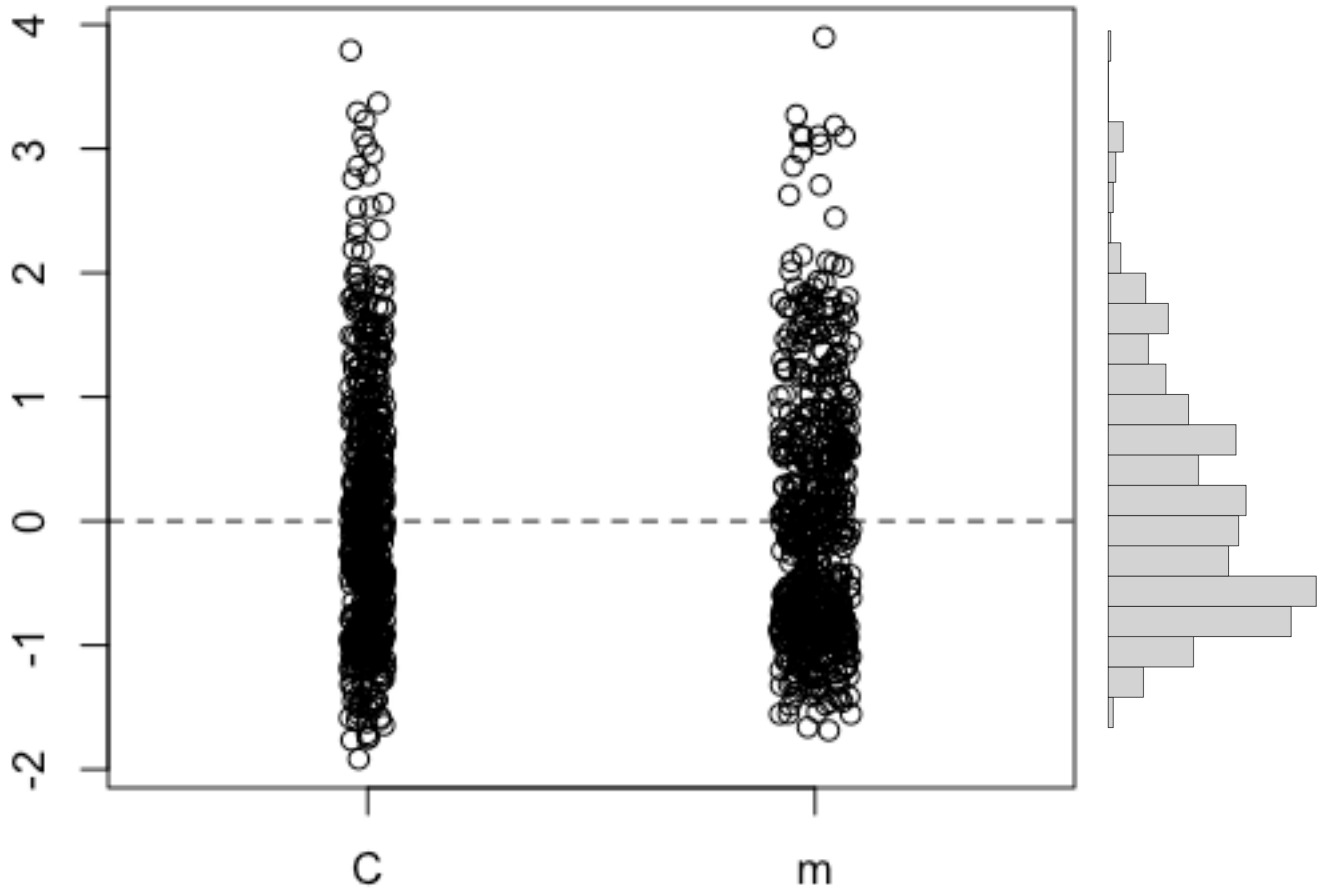
RNA-seq: センタリング、スケーリングしたデータ

$$a_i - \overline{a} / \text{S.D.}(a)$$



平均値を使うと言うことは:

正規分布を仮定している



データの分布を仮定した/しない手法

“パラメトリック/ノンパラメトック”解析

- （ピアソン）相関係数
- スピアマンの相関係数

ピアソンの相関係数
（パラメトリック）

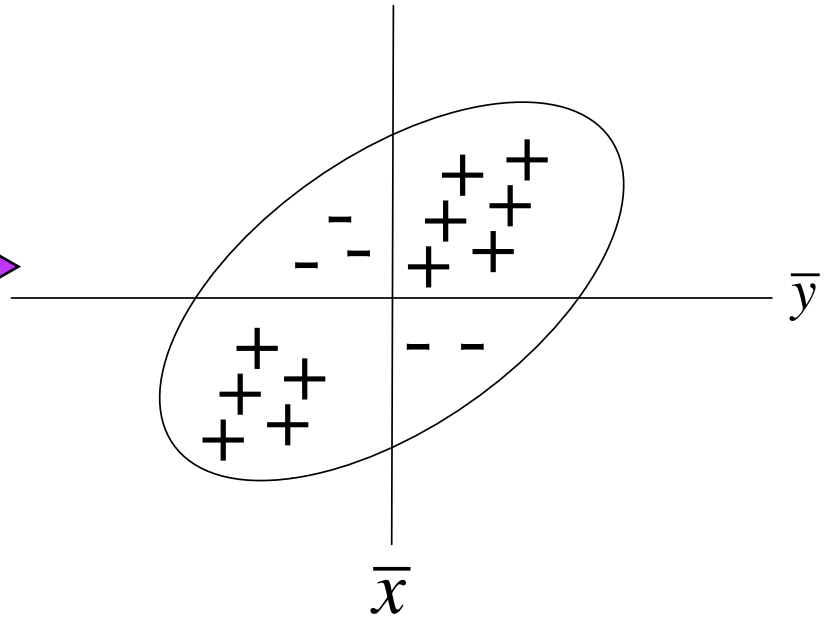
$$r_S = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

(-1≤ r ≤ 1)

x, yは正規分布している仮定

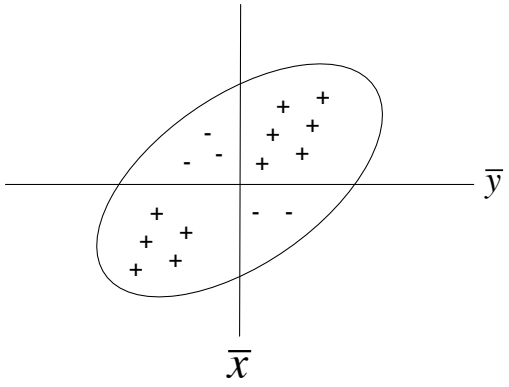
ピアソン相関係数の分子：
共分散

$$r_S = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$



ピアソン相関係数の分母：
標本標準偏差

$$r_S = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{S.D.x \times S.D.y}$$



Scaling function

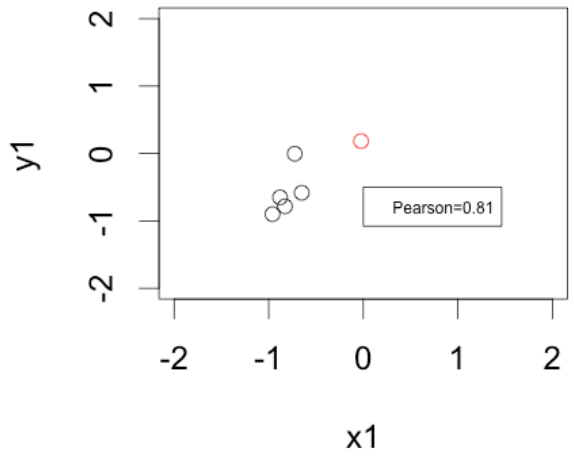
スピアマンの相関係数
(ノンパラメトリック)

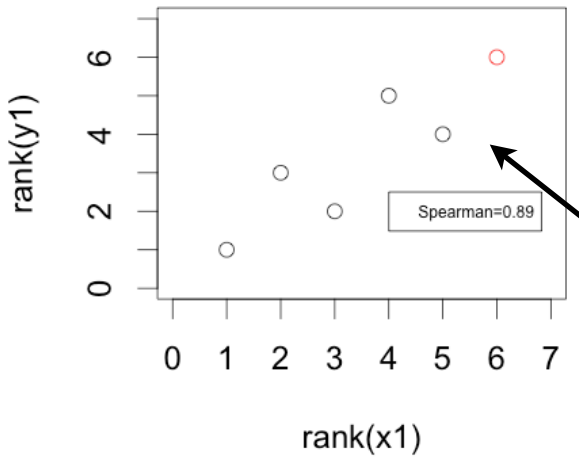
$$r_S = \frac{S_{r_x r_y}}{\sqrt{S_{r_x r_x} S_{r_y r_y}}} = \frac{\sum r_{x_i} r_{y_i} - \frac{(\sum r_{x_i})(\sum r_{y_i})}{n}}{\sqrt{\{\sum r_{x_i}^2 - \frac{(\sum r_{x_i})^2}{n}\}\{\sum r_{y_i}^2 - \frac{(\sum r_{y_i})^2}{n}\}}}$$

$(-1 \leq r \leq 1)$

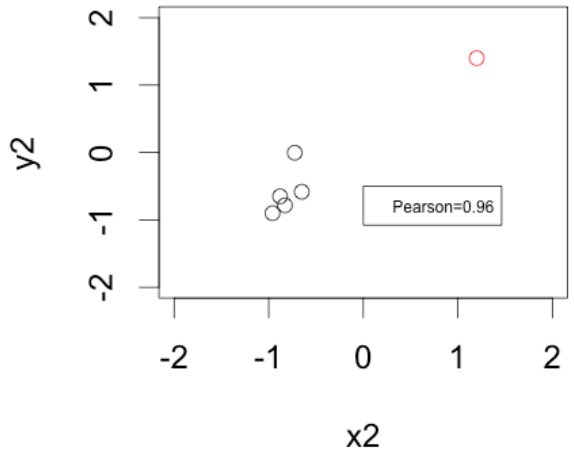
x, yの分布を仮定しない

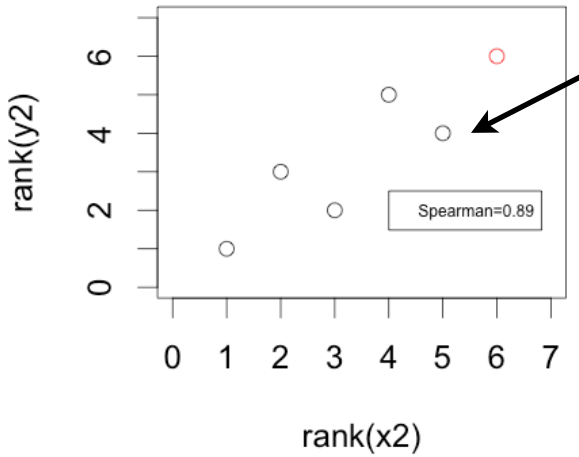
相関係数の比較: ピアソン vs スピアマン





ピアソンの相関係数は定量的な類似性を表現しているが、分布の影響を受ける





実習：再現性を評価するために

1. 反復データの散布図を描画する
2. 個々のデータの分布を確認する
3. 分布をもとに相関係数を求める

得られた相関係数にどのような生物学的意味を見出せるか？

同一処理データ間の相関係数の大きさ
と
異なる処理データの相関係数の大きさ
を
総当たりで比較してみる

相関係数行列

	set1_Col.0	set1_rps2.101C	set2_Col.0	set2_rps2.101C	set3_Col.0	set3_rps2.101C
set1_Col.0	1	0.897	0.941	0.882	0.943	0.864
set1_rps2.101C	0.897	1	0.798	0.961	0.832	0.955
set2_Col.0	0.941	0.798	1	0.783	0.955	0.772
set2_rps2.101C	0.882	0.961	0.783	1	0.805	0.929
set3_Col.0	0.943	0.832	0.955	0.805	1	0.855
set3_rps2.101C	0.864	0.955	0.772	0.929	0.855	1

距離: 各サンプルが近い/遠い

相関係数

1: ２つのデータに強い相関

0: ２つのデータに相関なし

-1: ２つのデータが逆向きで相関

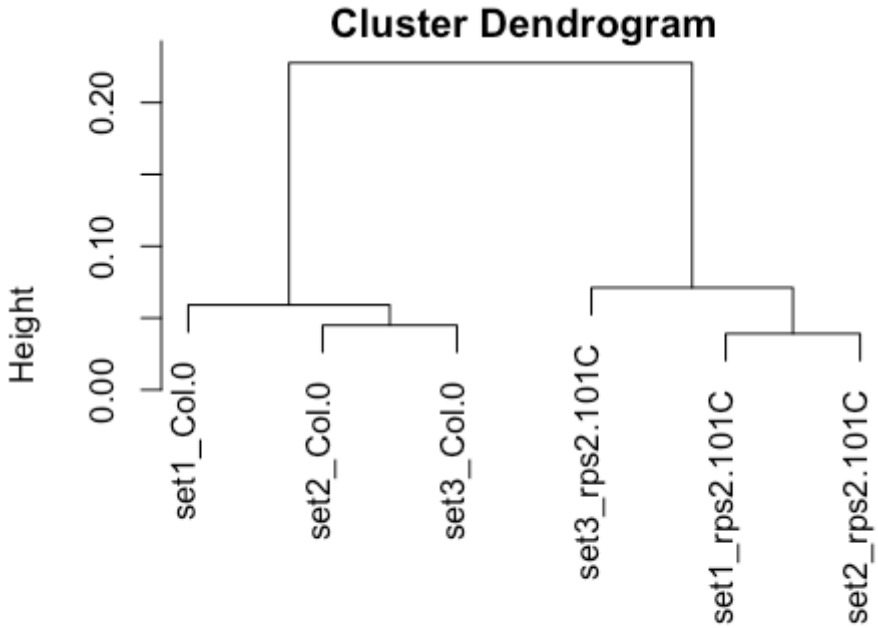
つまり、相関が強い（似ている）サンプルを
近くに配置するには
1 - 相関係数

距離行列

	set1_Col.0	set1_rps2.101C	set2_Col.0	set2_rps2.101C	set3_Col.0	set3_rps2.101C
set1_Col.0						
set1_rps2.101C	0.103					
set2_Col.0	0.059	0.202				
set2_rps2.101C	0.118	0.039	0.217			
set3_Col.0	0.057	0.168	0.045	0.195		
set3_rps2.101C	0.136	0.045	0.228	0.071	0.145	

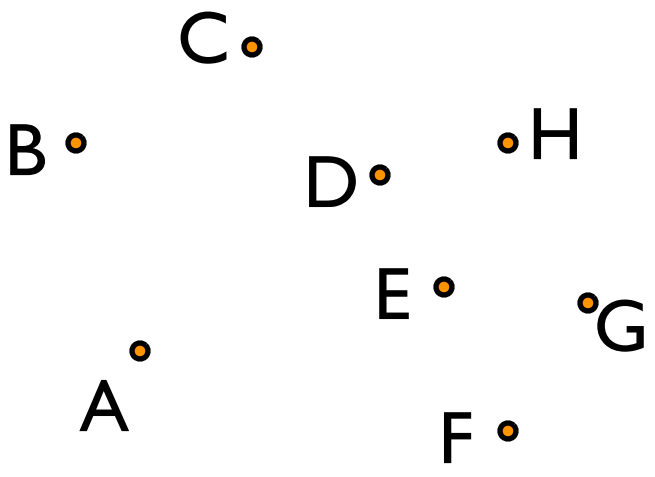
距離行列の可視化: デンドログラム

距離に基づいた
グルーピング
↓
クラスタリング
Clustering

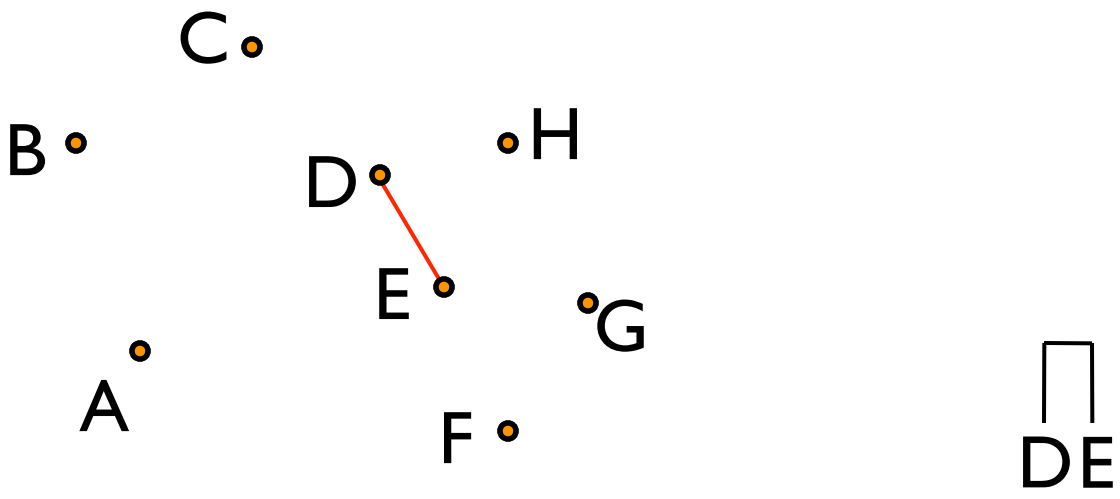


階層クラスタリング

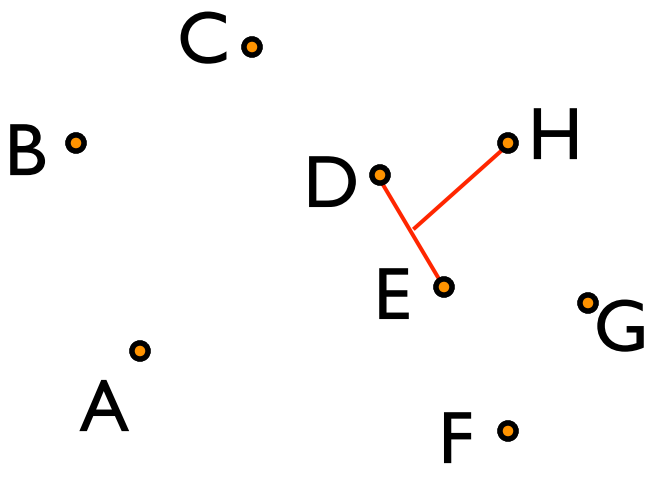
Agglomerative hierarchical clustering



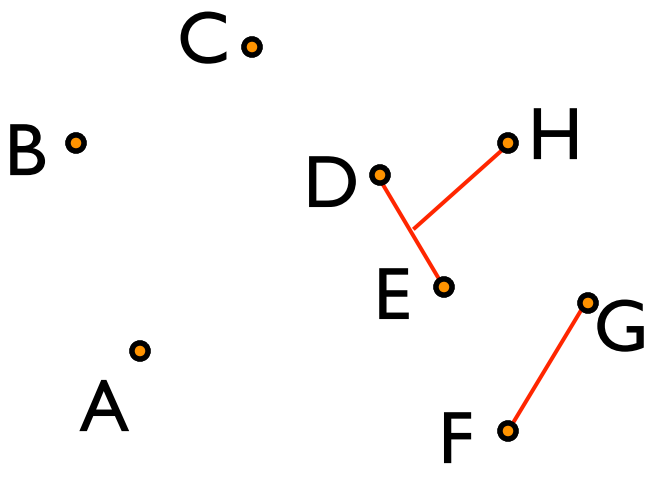
Agglomerative hierarchical clustering

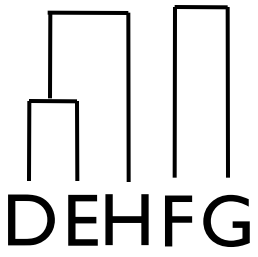


Agglomerative hierarchical clustering

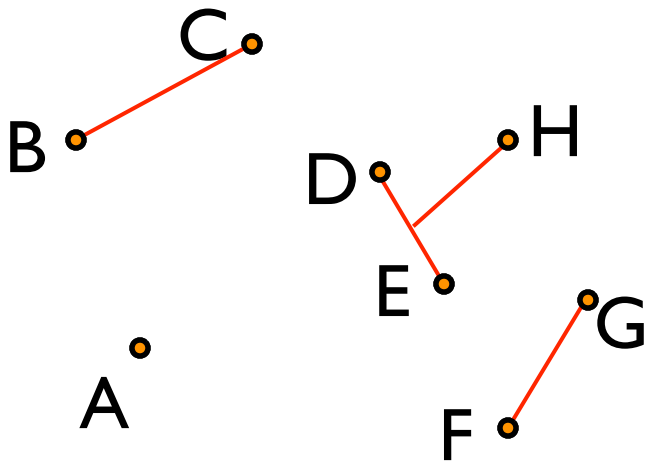


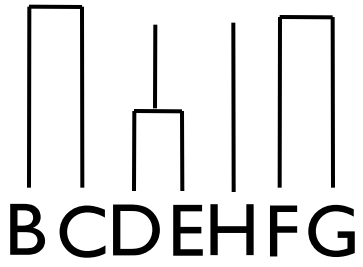
Agglomerative hierarchical clustering



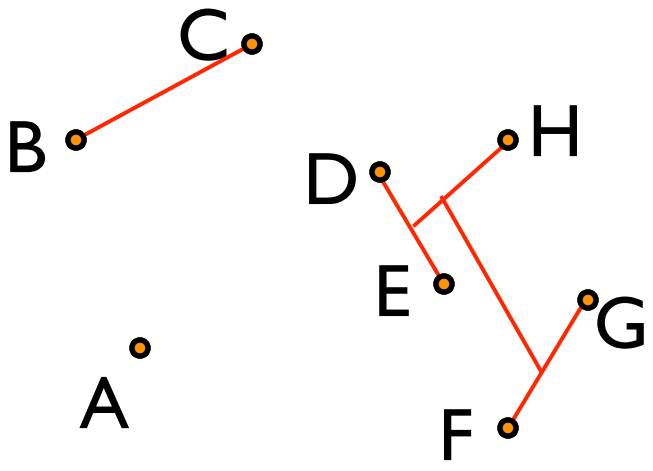


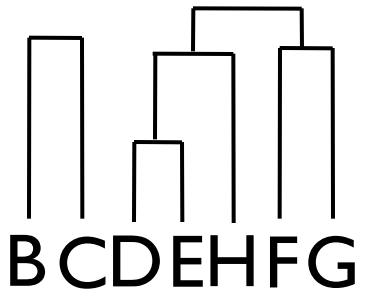
Agglomerative hierarchical clustering



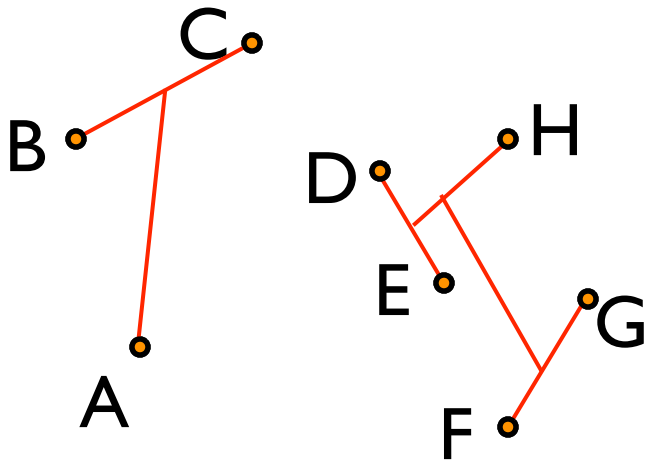


Agglomerative hierarchical clustering

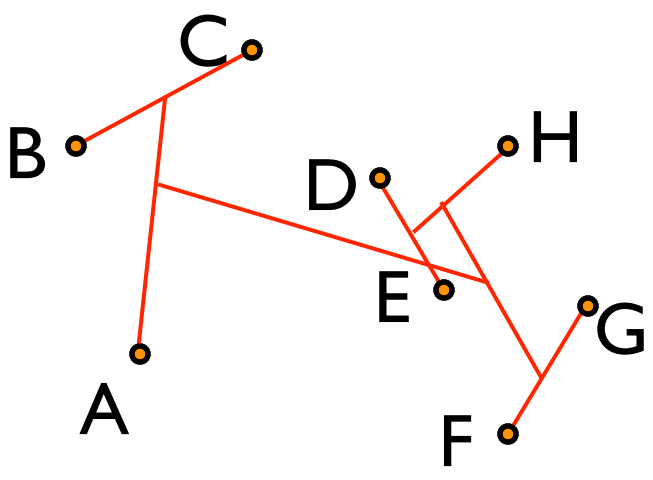


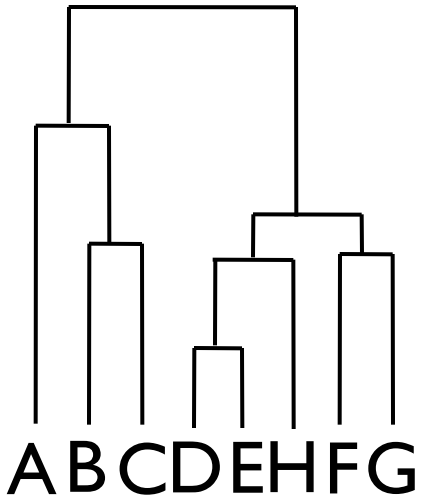


Agglomerative hierarchical clustering

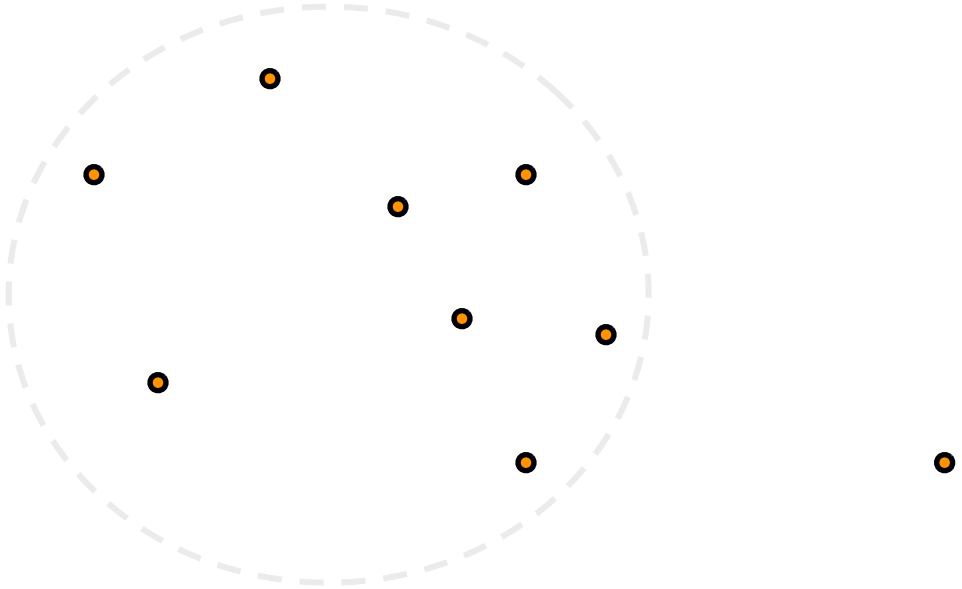


Agglomerative hierarchical clustering

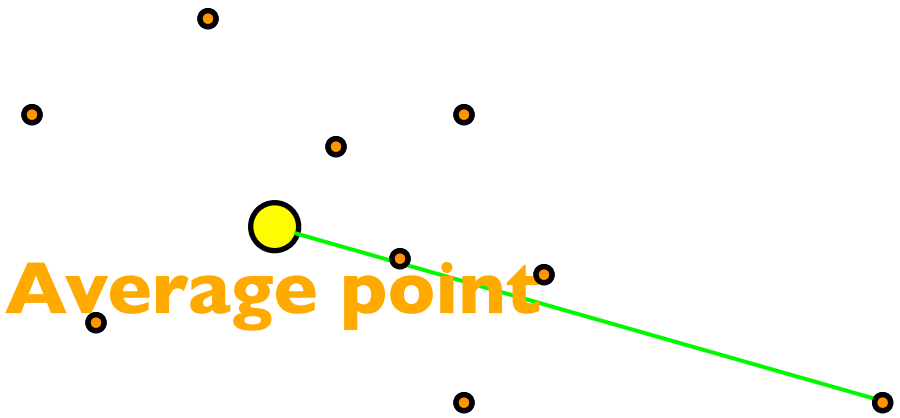




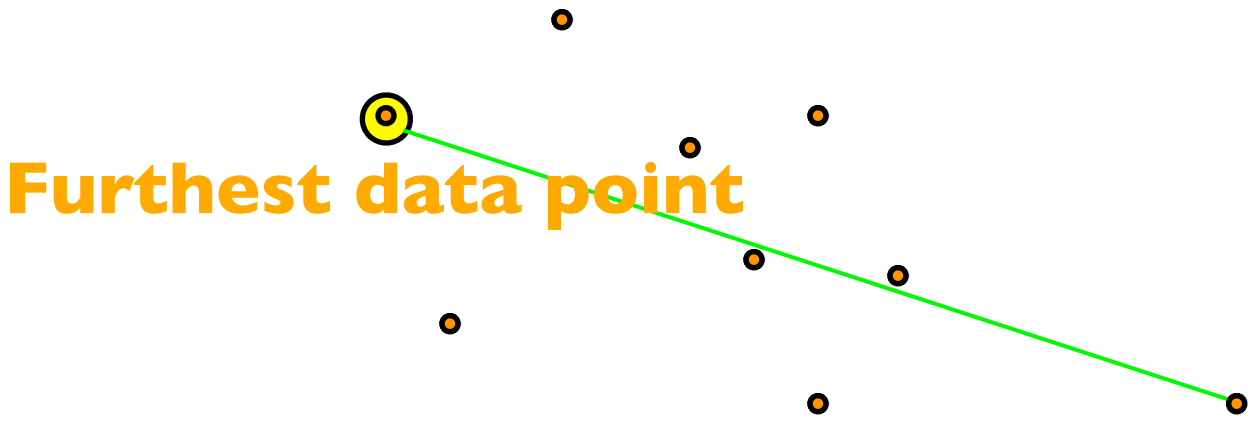
クラスター定義手法



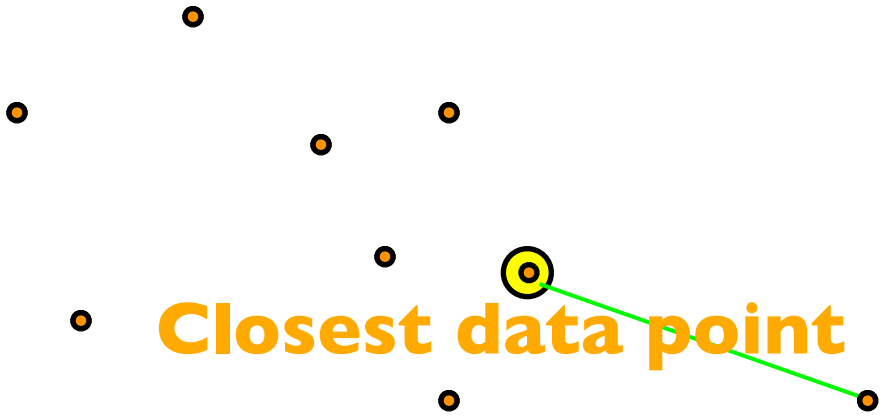
Average linkage



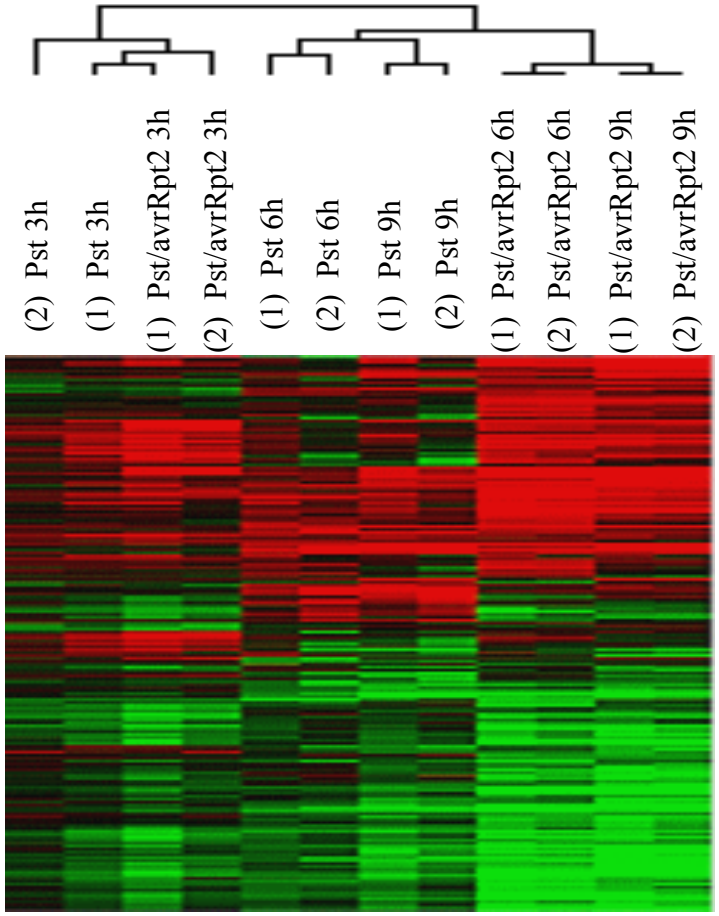
Complete linkage



Single linkage



階層クラスタリングを使った生物学的な分析



Tao *et al.* (2003) *Plant Cell*

- 3hでは処理の違いでクラスターにならない
- Pst処理では6h・9h間の違いが、Pst/avrRpt2でのそれよりも大きい

（現象の性質の類推、
解析シナリオ構想、
DEG解析する処理間の決定）

主成分分析

Principal component analysis

主成分分析とは？

モチベーション:

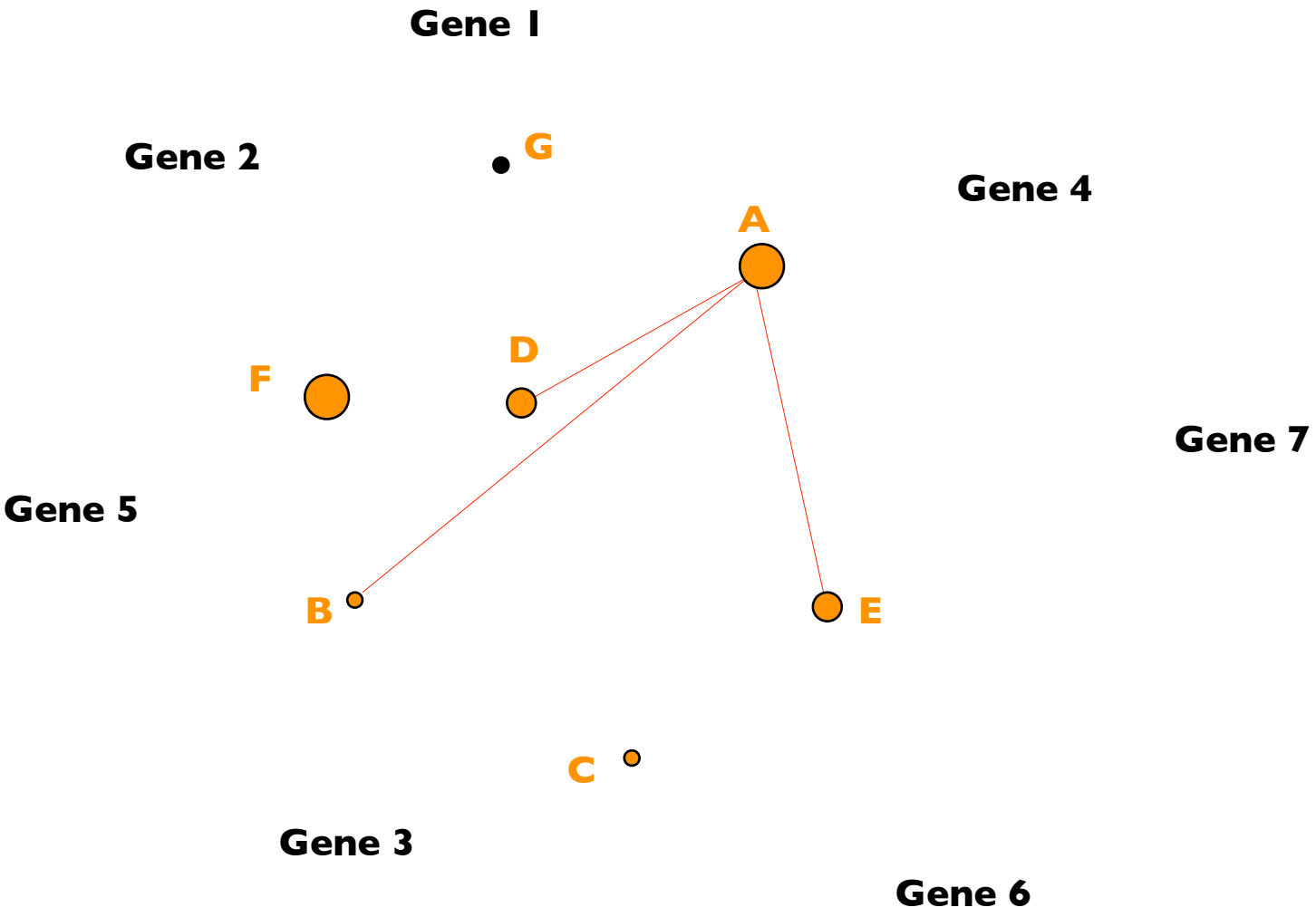
多次元データ（多数の遺伝子もしくは多数のサンプル）に含まれる特徴を

- 大きなものから抽出して**新たな軸**を作り
- **情報量の大きな低次元でデータを可視化する**

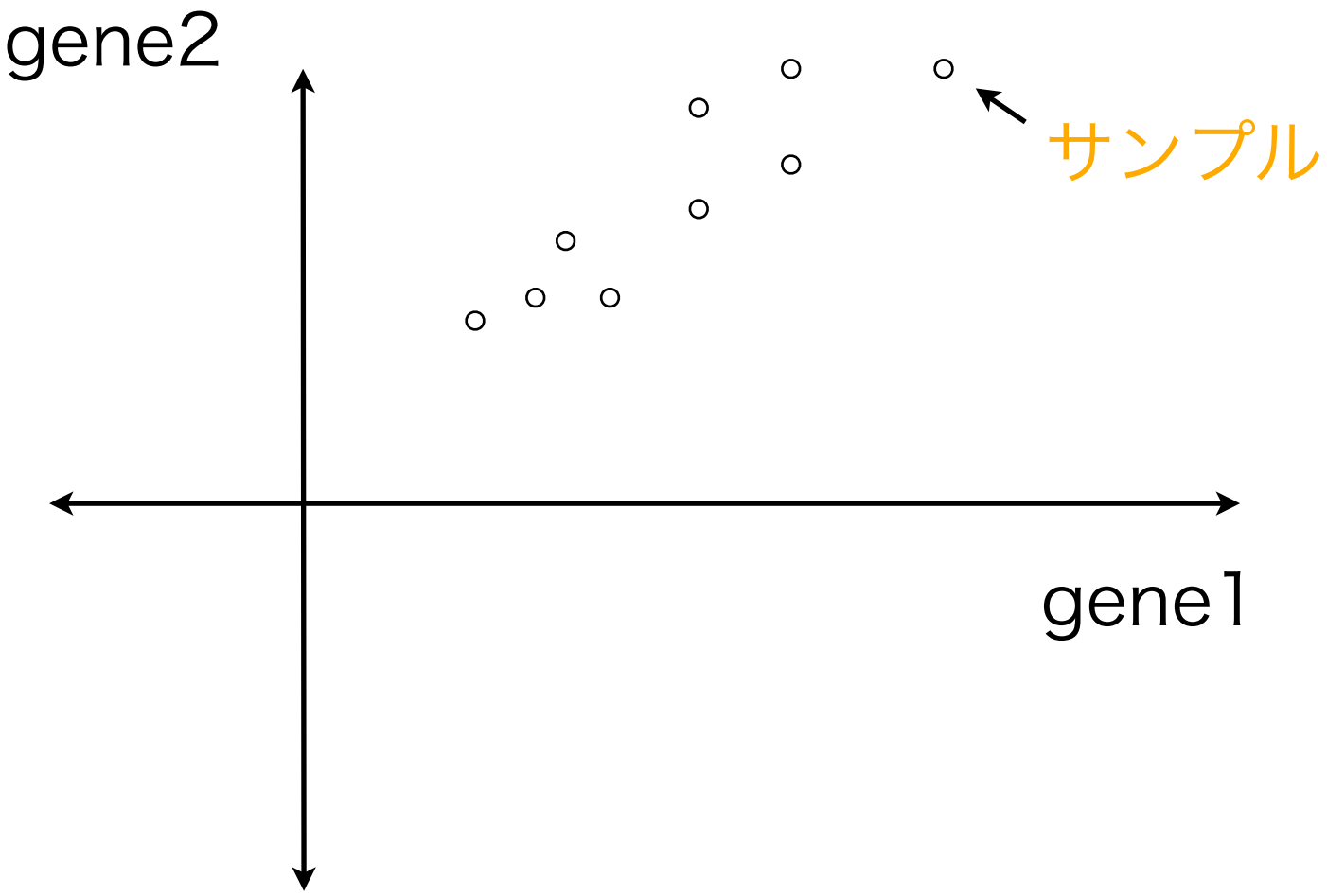
→ **人間が新たな解釈を与える**

階層クラスタリング:

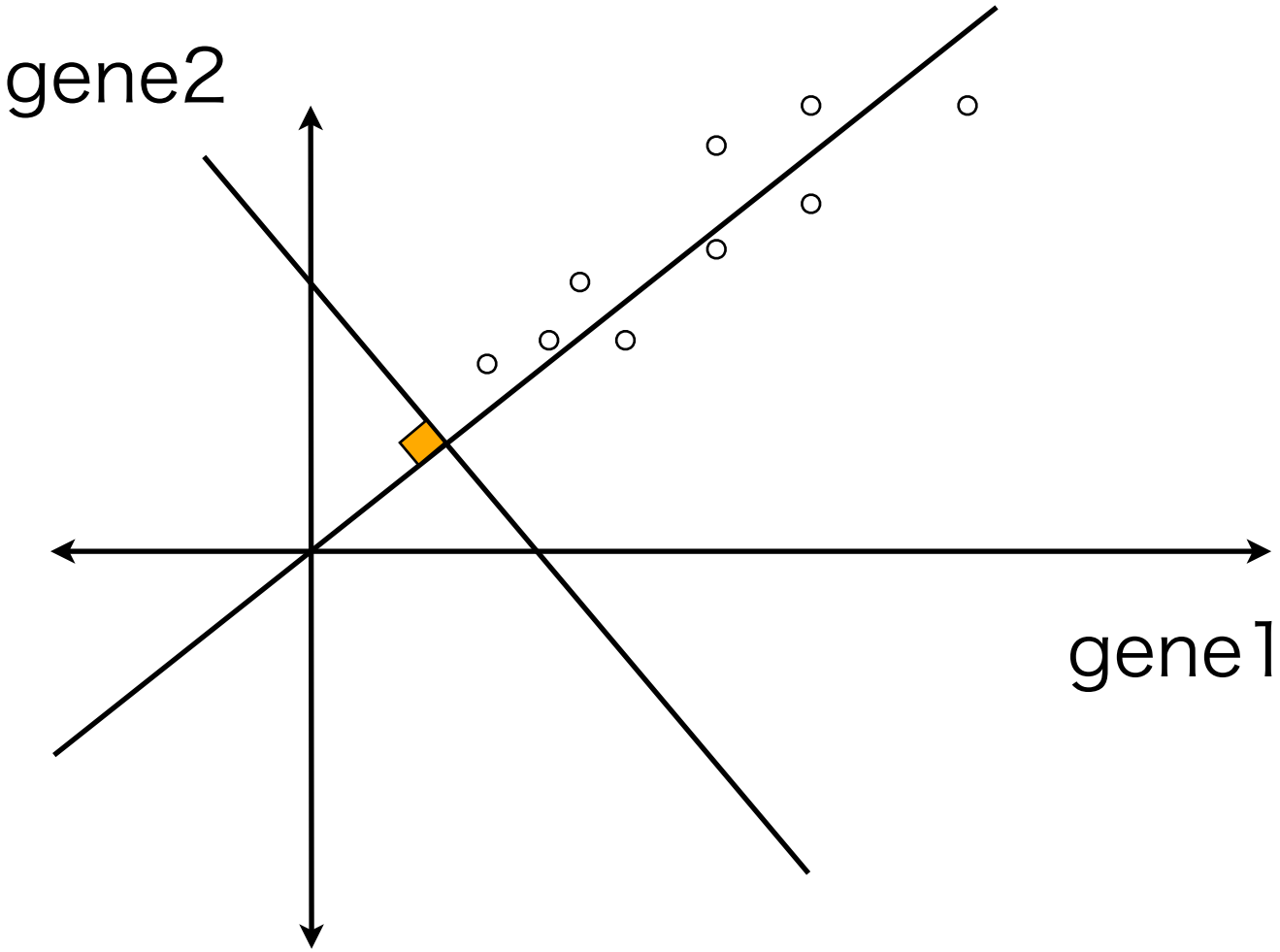
プロファイル間の類似性は空間での**1つの距離**によって決まる



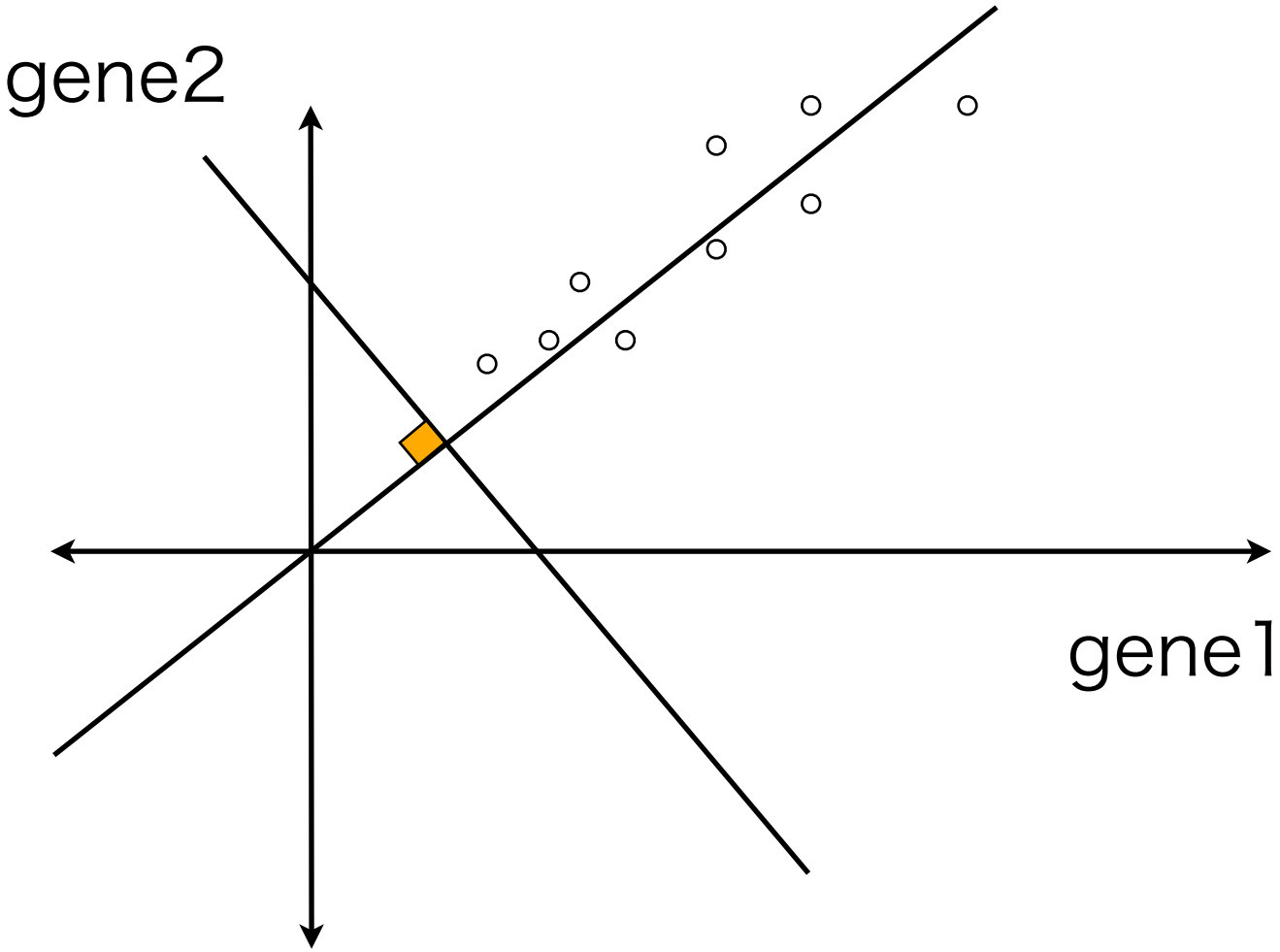
PCAは何をするのか？



PCAは何をするのか？



PCAは何をするのか？



PCAの概略(2次元)

1. 各サンプル（1..n）の観察値(x_n, y_n)を

$$u_n = a_1 x_n + b_1 y_n$$
$$v_n = a_2 x_n + b_2 y_n$$

足し算→線形

とおく

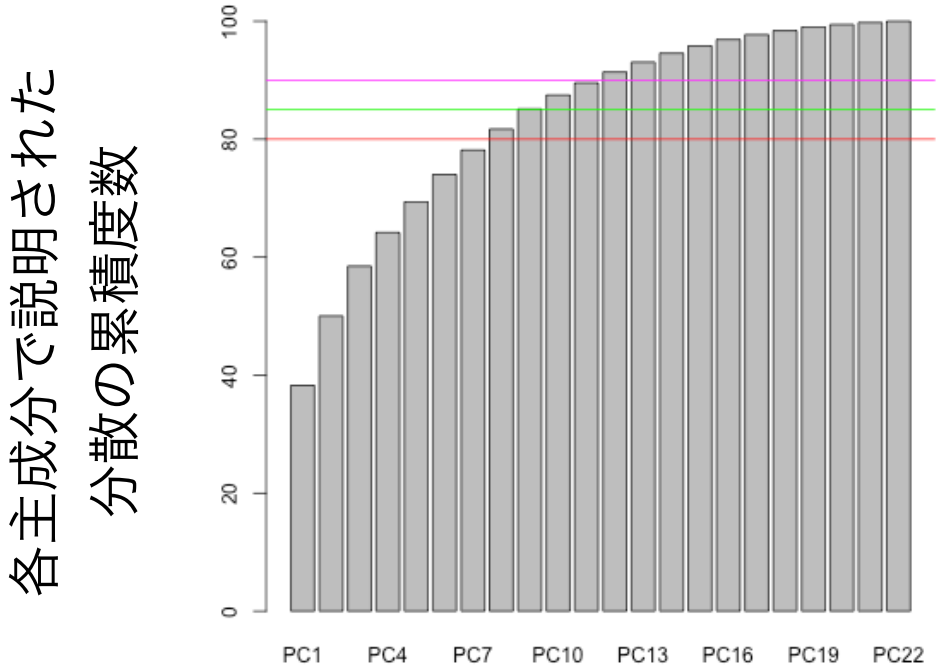
2. $a^2 + b^2 = 1$, uとvの相関係数0という制約の下でこれを解いて a_n, b_n を求める。

PCAで得られる重要な統計量

- 寄与率
- 因子負荷量
- 主成分得点

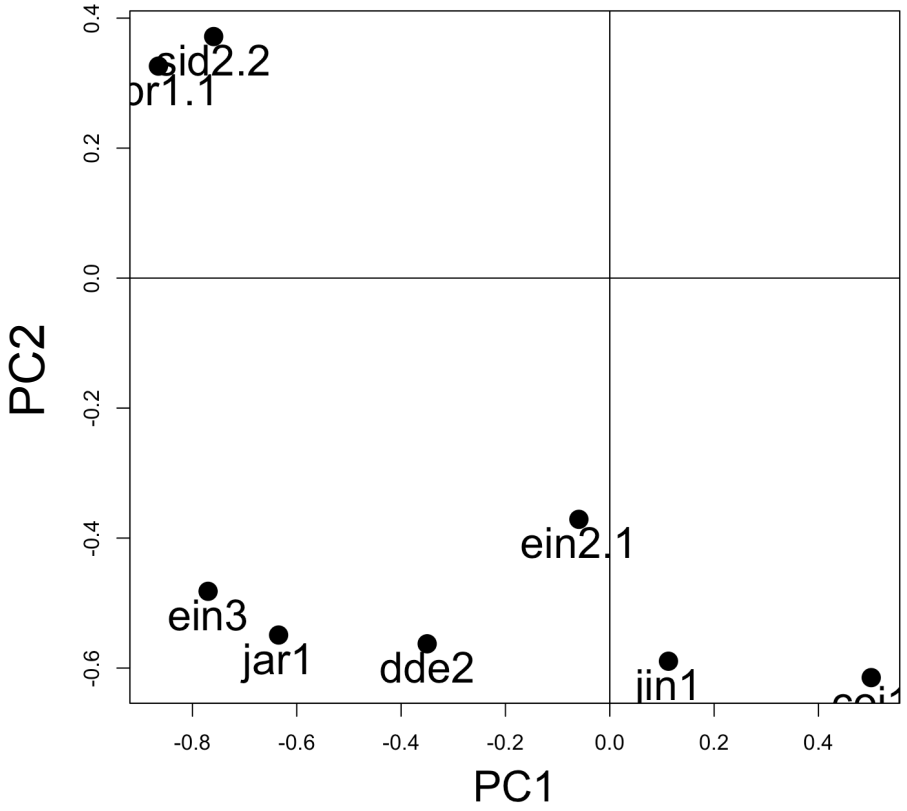
寄与率

- 各主成分が説明する分散の割合



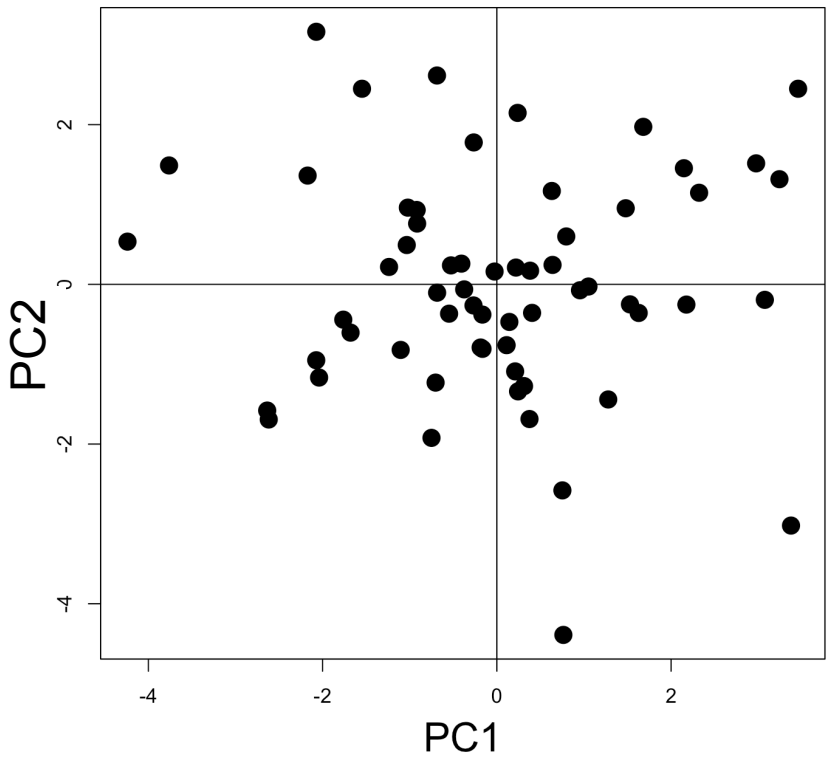
負荷量 loadings

- 得られた主成分と元データのパラメーターの相関
- 各パラメーターがもとのデータの情報をどれだけ有するか



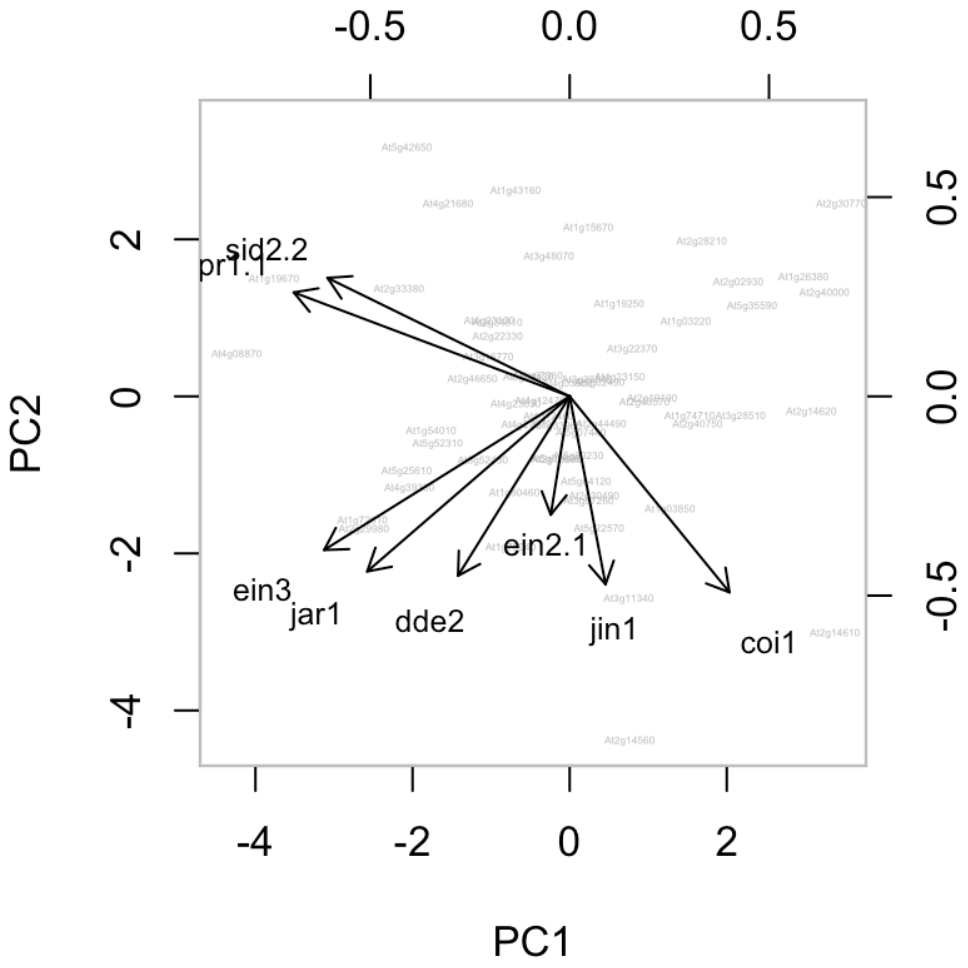
主成分得点 scores

- 各パラメーター（遺伝子）の値を各主成分について標準化した統計量



標準化: 平均0, SD=1

biplot:
因子負荷量と
主成分得点を
同時に可視化



主成分分析(まとめ)

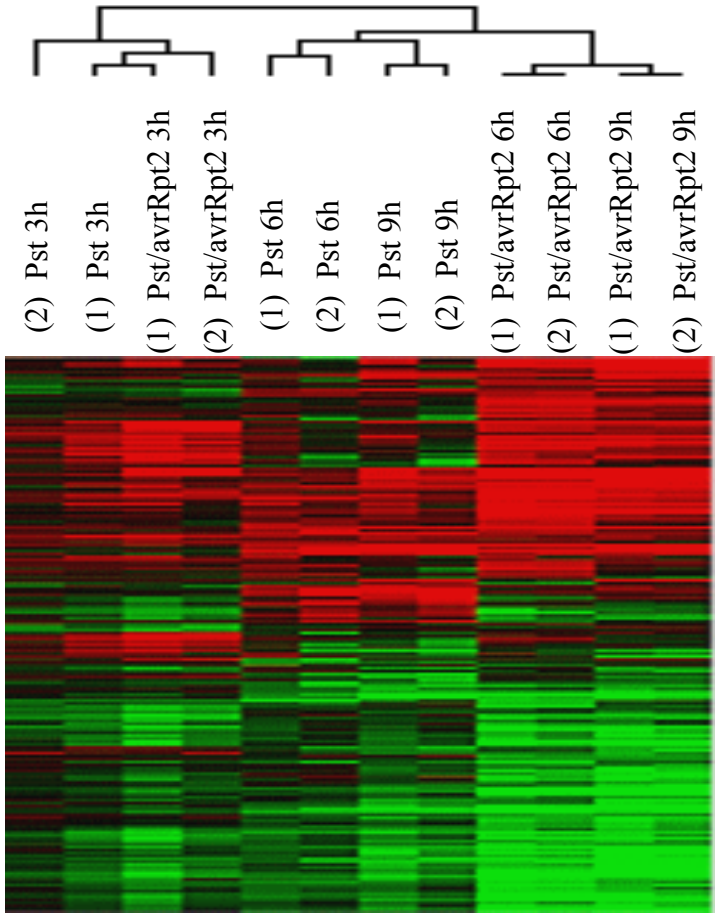
- 主成分分析は**データの分散を説明する新たな軸を計算する**方法
  - **寄与率**
  - **因子負荷量**
  - **主成分得点**

まとめ(I)

- 解析するデータの性質を明らかにする
  - 統計量を使って調べたい性質を要約する（ここでは相関係数）
  - 統計量は適切に選ぶ
  - 可視化する

パラメーターごとの評価

データ全体から部分・詳細へ



Tao *et al.* (2003) *Plant Cell*

（現象の性質の類推・
解析シナリオ構想、
DEG解析する処理間の決定）

仮説構築

- Pst/avrRpt2処理はPst処理よりもロバストか？ → 相関係数はより高いか？
- 抵抗性と相関した遺伝子群は？
  → Pst/avrRpt2 vs Pst処理間のDEG

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:

- 検定の基本的な流れ
- 検定のポイント

用語の意味の整理

- 統計量、確率分布、自由度、*p*値

Rでの統計を正しく使うために:

- エラーが出る/出ない、ではなく、Rを正しく使う

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

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.
遺伝子Aと遺伝子Bの発現プロファイルの相関係数は0.51だった。これら2遺伝子は有意に共発現しているか？

ポイント

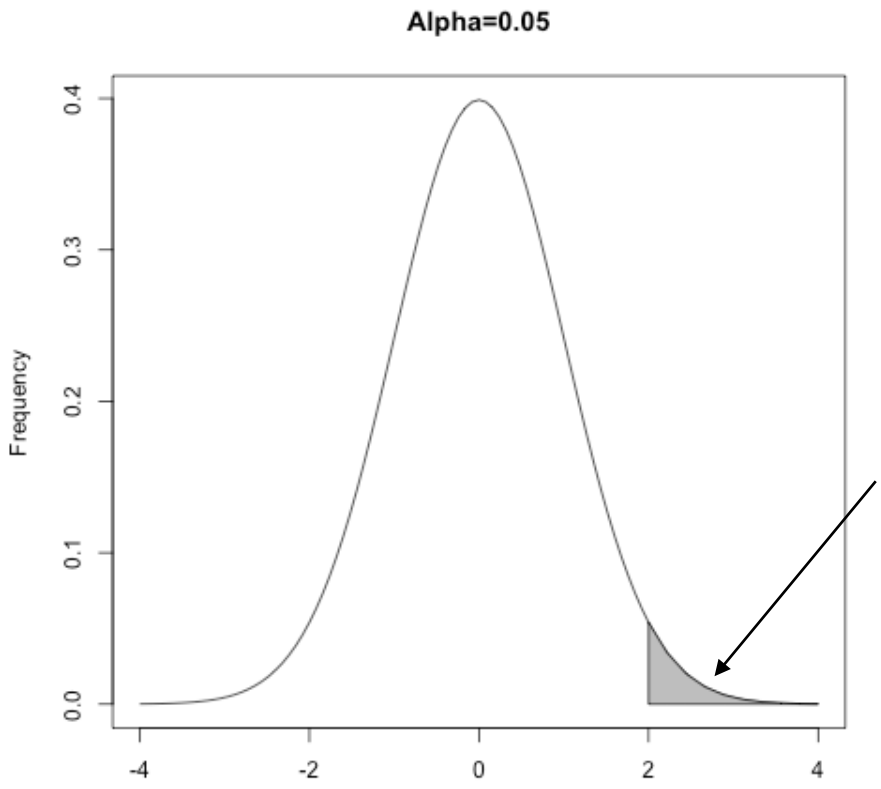
2. 統計量を求める:

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

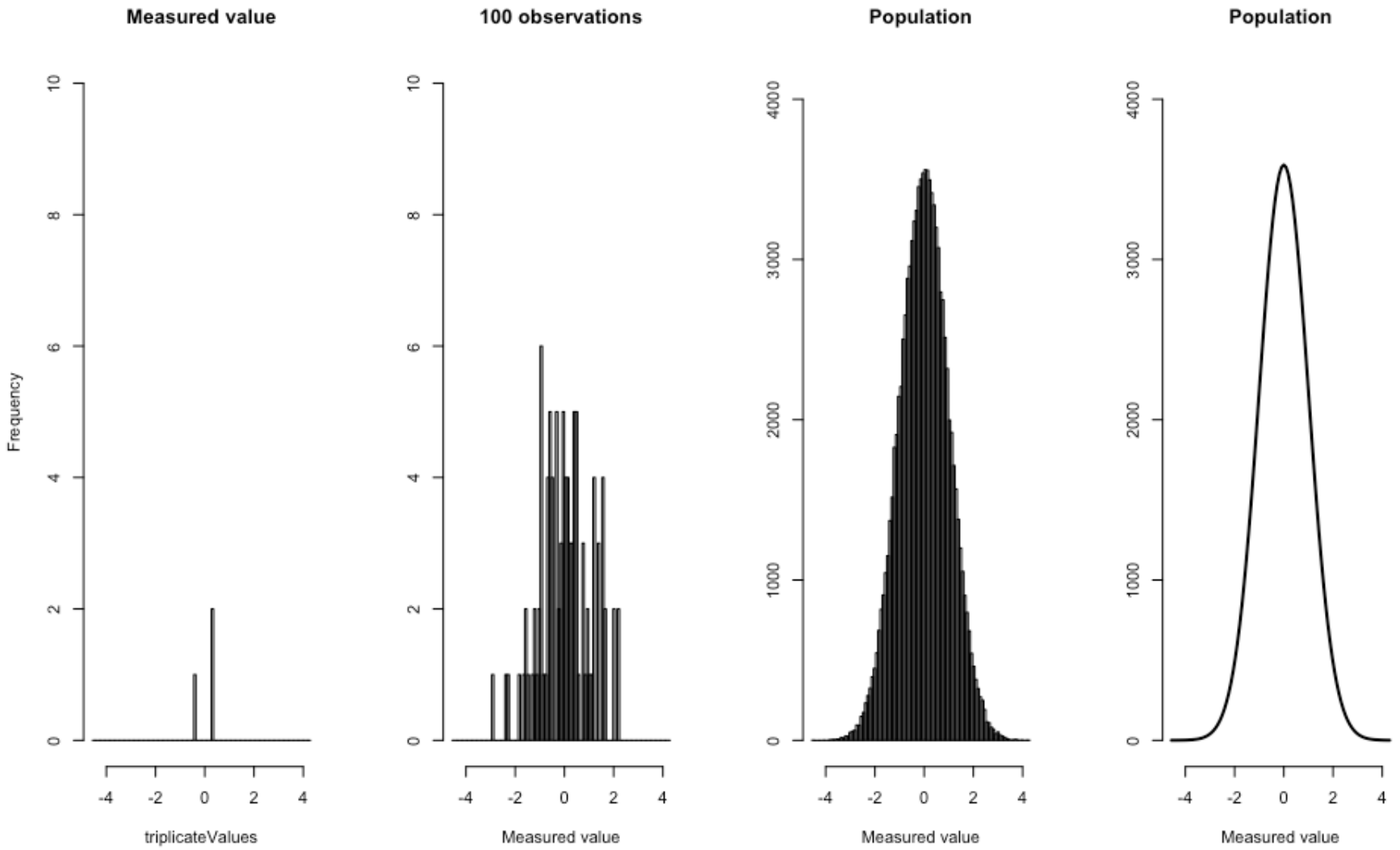
3. 確率分布と照らし合わせる



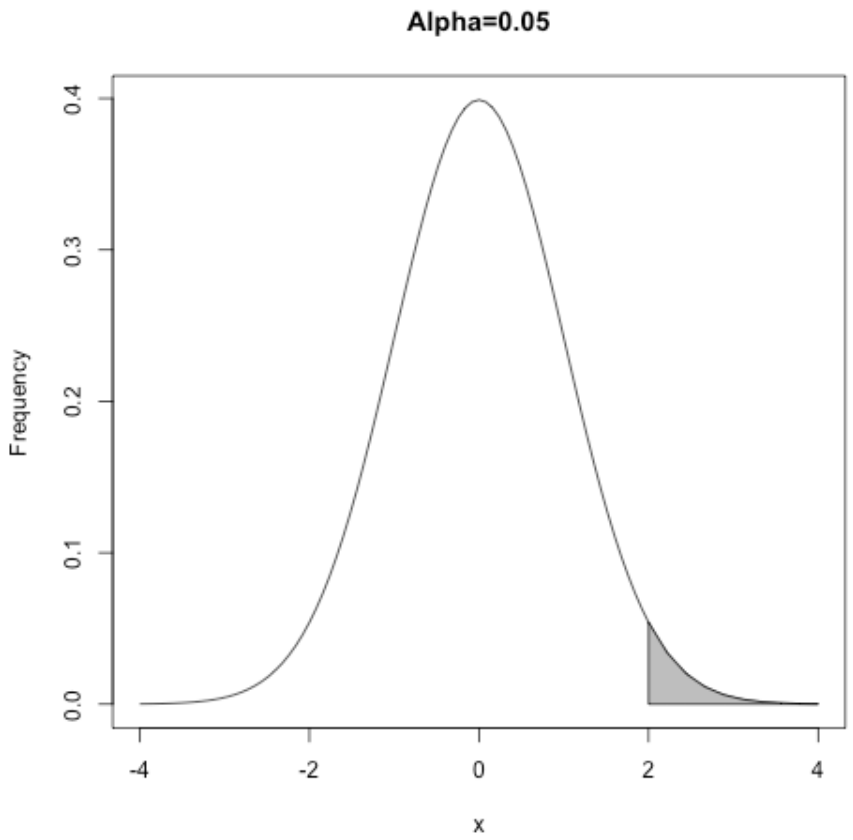
棄却限界値によって規定される面積
(通例: 全体の5%)

統計量

確率分布? 面積？


Measured value
100 observations
Population
Population
Frequency
0
2
4
6
8
10
1000
2000
3000
4000
-4
-2
0
2
4
triplicateValues
Measured value
Measured value
Measured value

4. 判定: 帰無仮説が棄却されるか？



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	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

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

代表値

-（バー）は
平均を表す
^（ハット）は
推定を表す

べてのデータを足して、データ数で
る値

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

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。データの分布に依存しない。

→ 内山　R入門「基本統計量の計算」

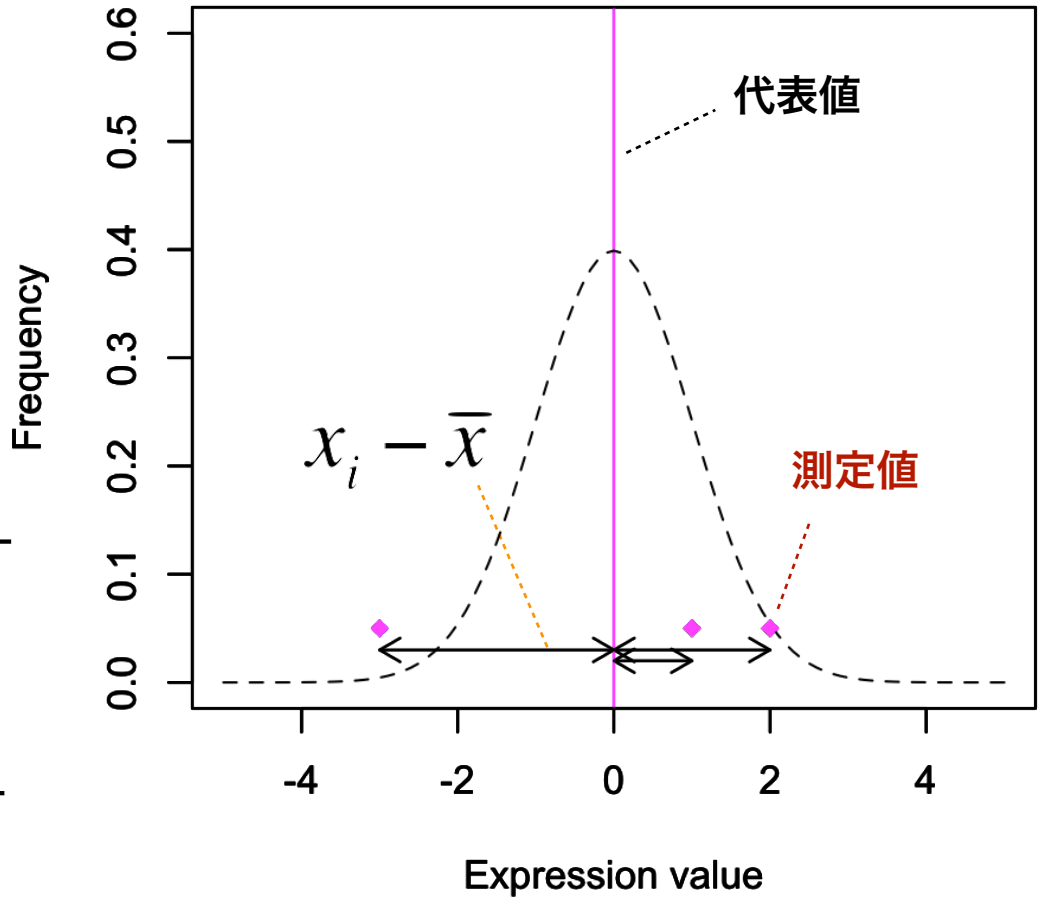
ばらつき: **分散／偏差**

分散:

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

標準偏差:

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



→ 内山　R入門「基本統計量の計算」

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

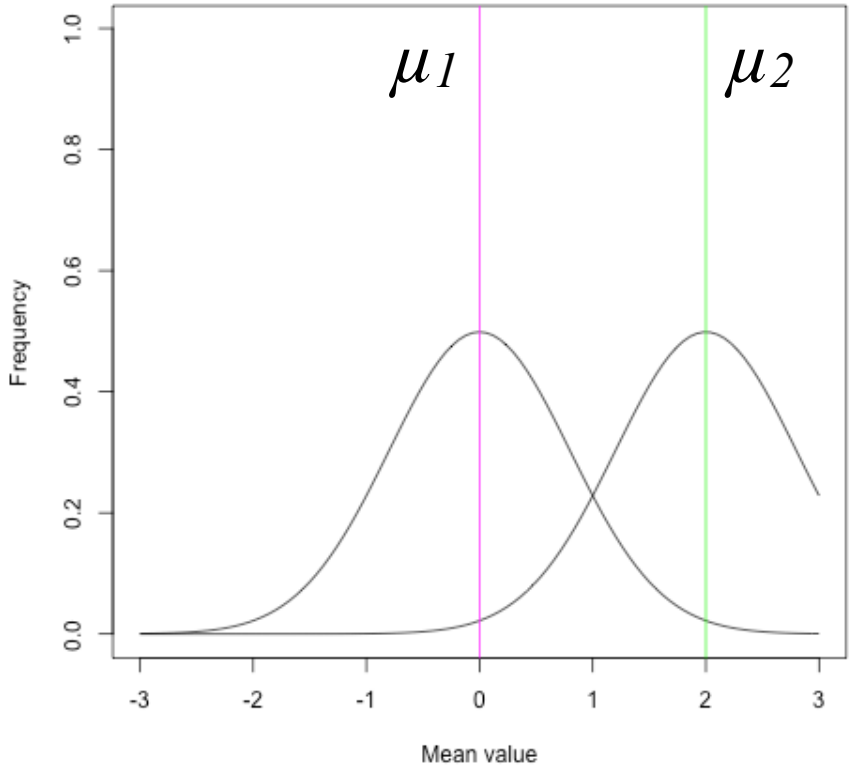
ポイント

t 検定:

2サンプルの平均の検定

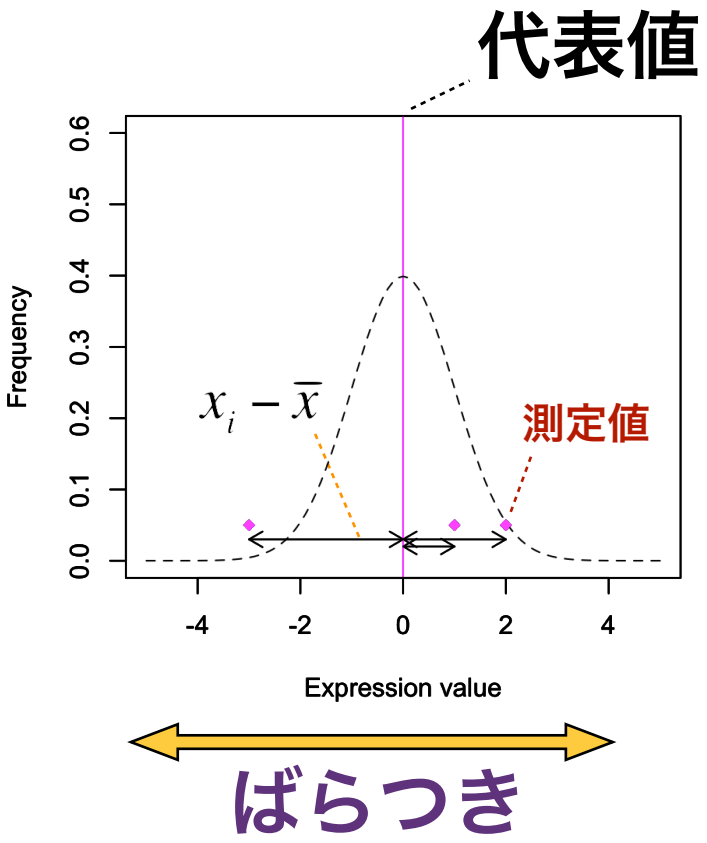
- 平均値 = μ_1, μ_2
- データは正規分布

ほぼ全ての検定方法に前提がある



t 検定で用いる統計量

1. 代表値: 平均値
2. ばらつきの範囲:
   平均標準誤差
3. 自由度



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

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

統計量その2：

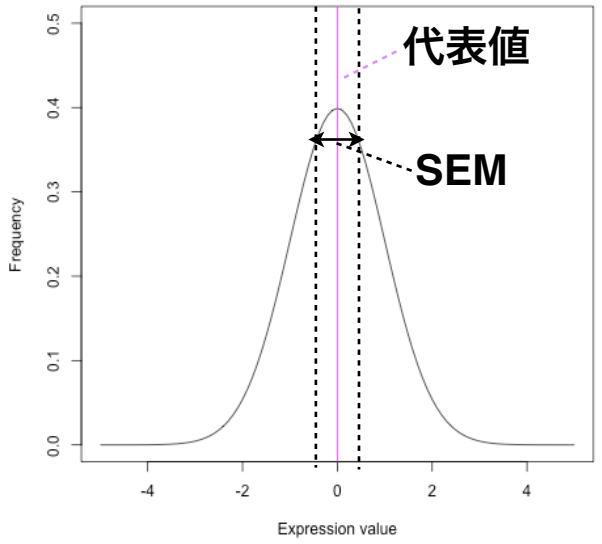
statistical mind

平均値も**あくまで推定値**

（平均）標準誤差:

「統計量」の偏差

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



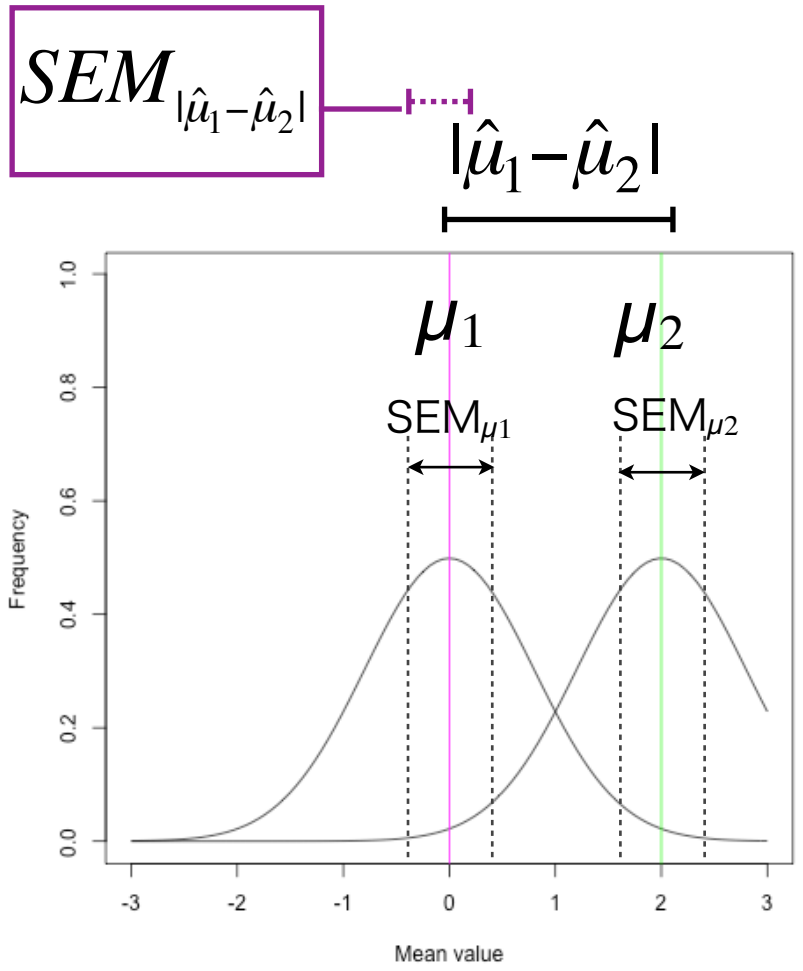
s: standard deviation 標準偏差

statistical mind

統計量その3:
平均の差とその誤差

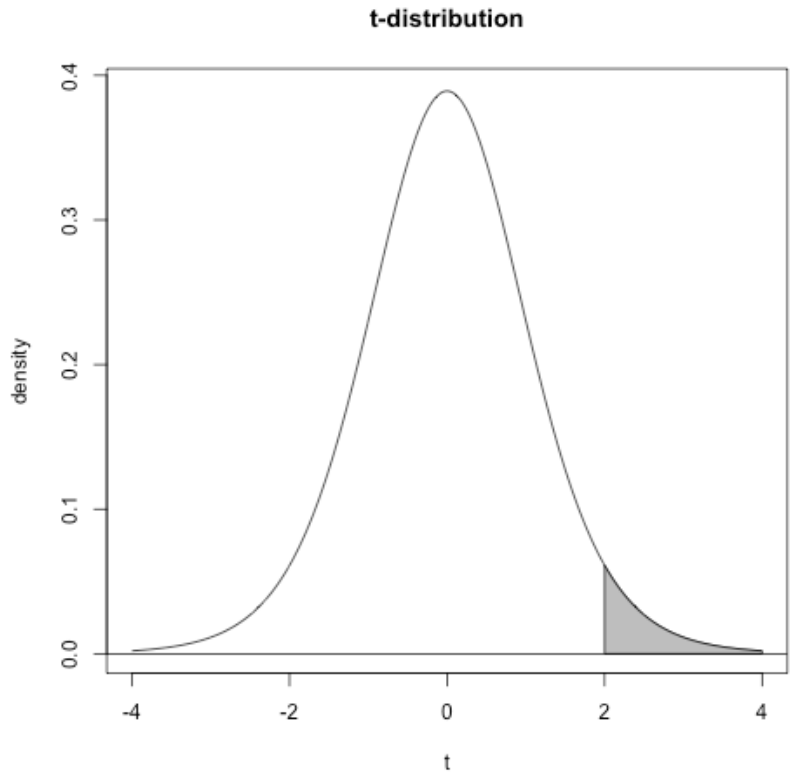
t統計量

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



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

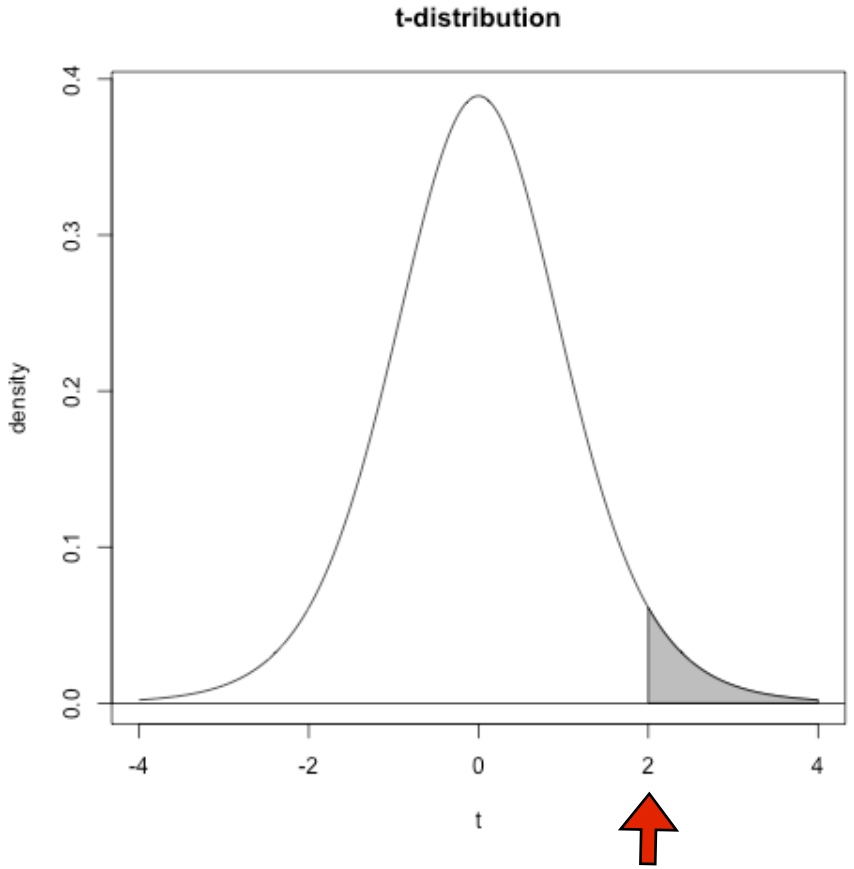


【おさらい】自由度: 統計量を求めるのに使うことができる独立な標本数

*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

- 汎用される閾値（危険率）：0.05



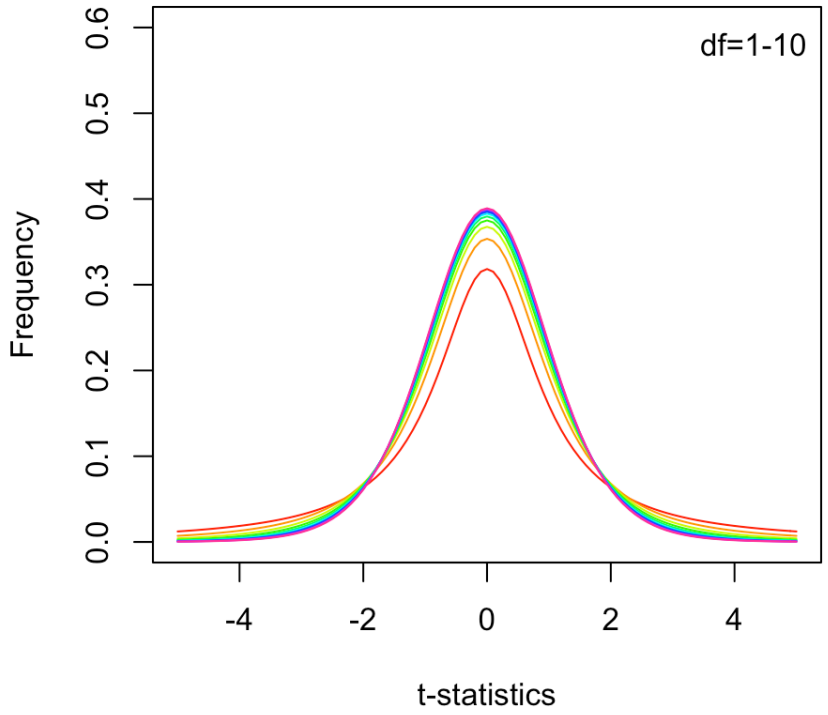
ポイント

データの分布、仮説検定に即した確率分布を使う

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

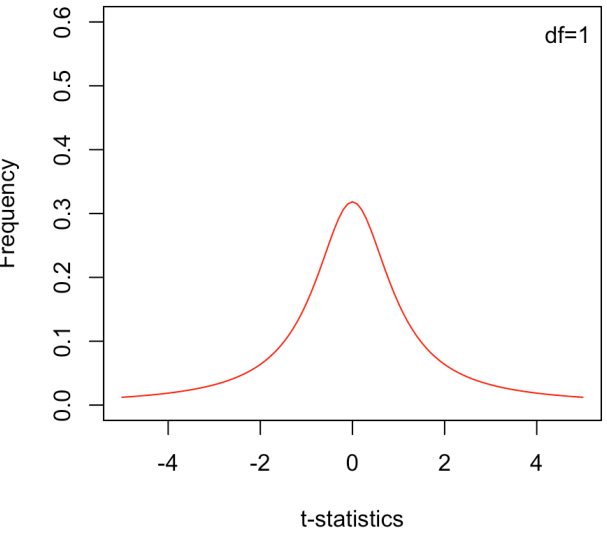
→正規分布ではなく、t分布

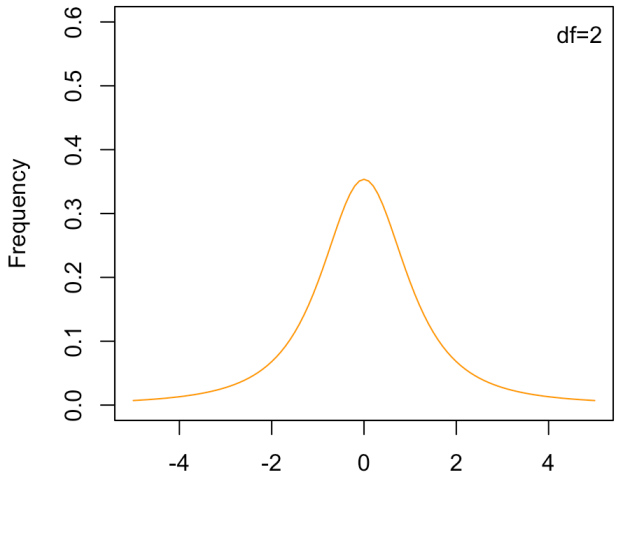


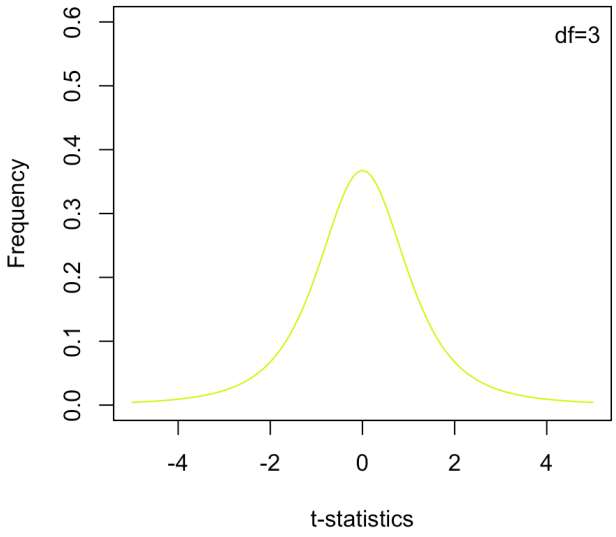
$$t = \frac{|\hat{\mu}_1 - \hat{\mu}_2|}{SEM_{|\hat{\mu}_1 - \hat{\mu}_2|}}$$

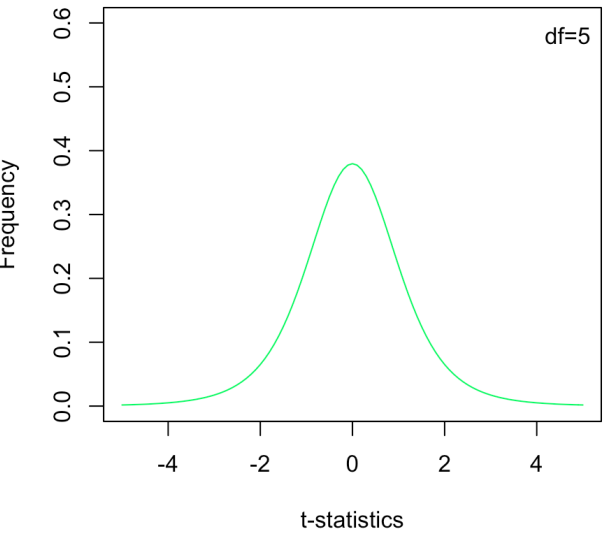
ポイント

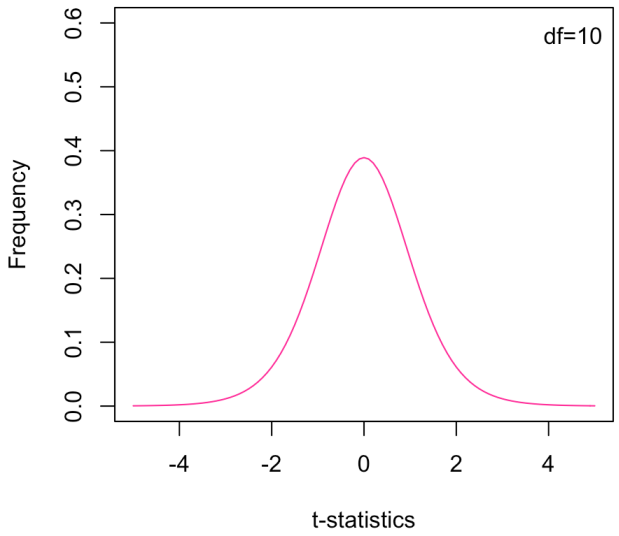
t 分布と自由度の関係

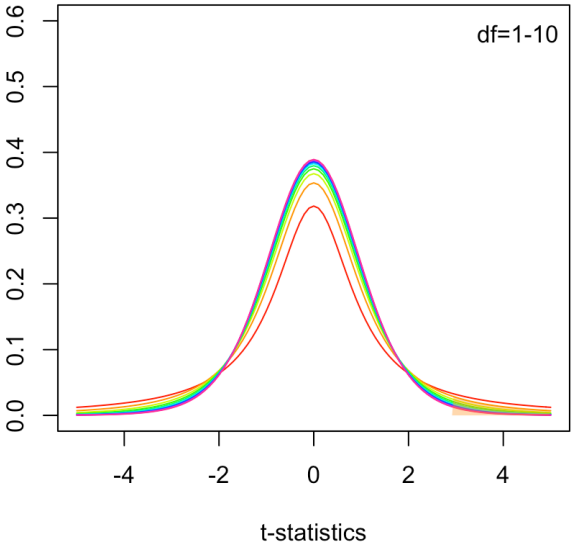






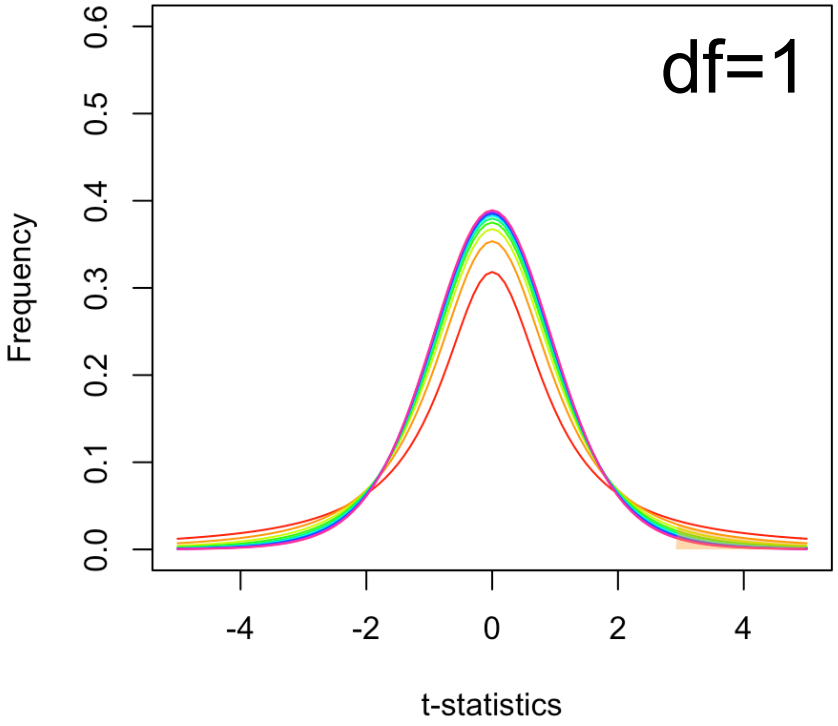


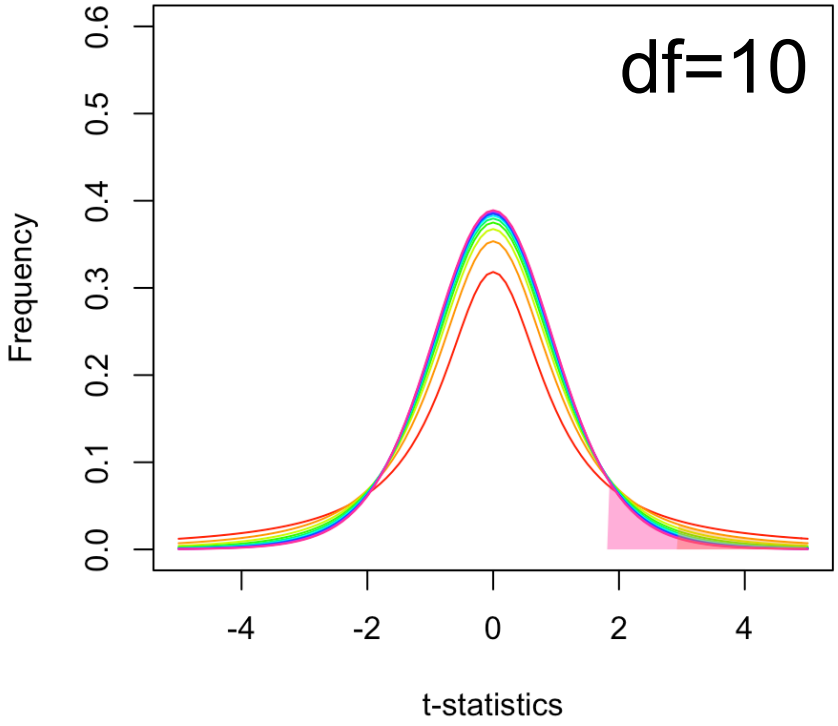




ポイント

t検定における自由度の違い: 検出力の違い





statistical mind

仮説検定を研究に導入する心構え

研究における手続き

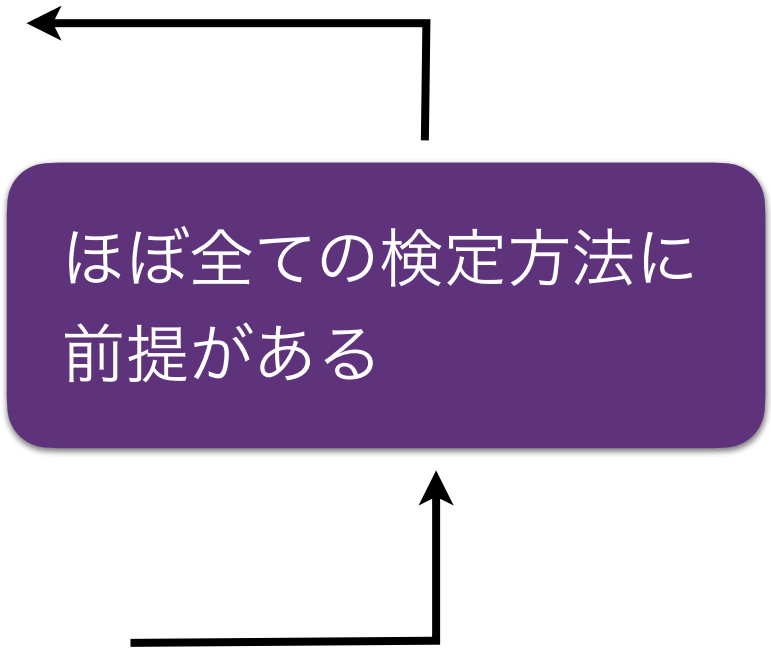
実験を計画する
↓
実験を行う
↓
結果
↓
仮説の検定

検定の結果によって
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」に練ってある必要がある

実験を計画する
↓
実験を行う
↓
結果
↓
仮説の検定



検定の結果によって
結論を導く

ポイント

ほぼ全ての検定方法に
前提がある

ex. *t* 検定: 正規分布

どの確率分布を想定すべきデータ？

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)

パラメトリック？　ノンパラメトリック？

ここまでのまとめ

検定

- 帰無仮説
- 統計量
  - 「平均値も推定値」
- 確率分布

検定の前提

- データの「型」、「分布」

多重検定の補正

+ 統計検定における重要な概念

*p*値とは：

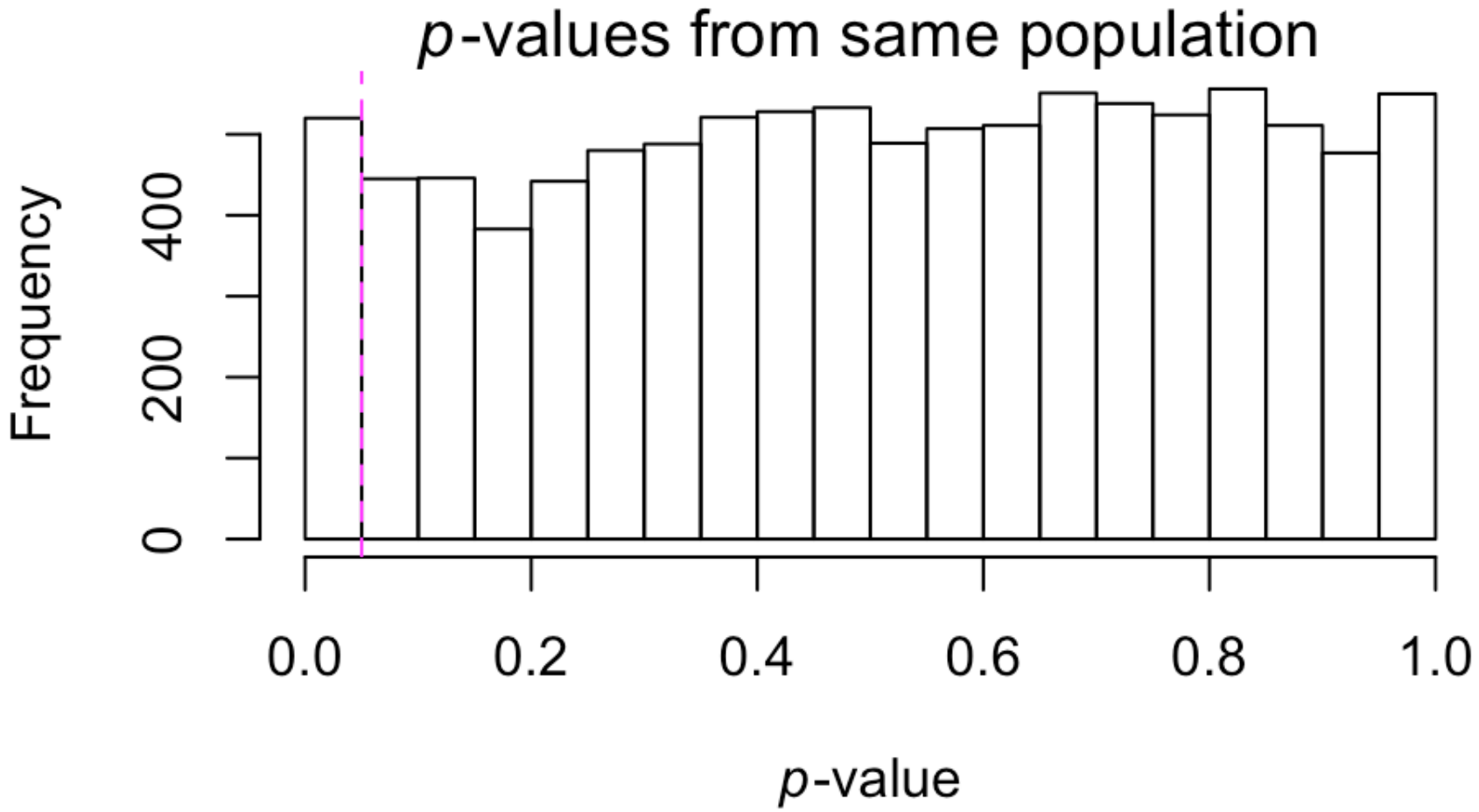
- 標本に基づいた統計量が　　帰無仮説の下、起きうる確率

- 汎用される危険率（閾値）：

0.05 = 100回に5回起きる

演習：同一平均値集団間のt検定の繰り返しをシミュレートしてみる

同一平均値集団間の*t*検定でも
# $p < 0.05$が得られる



多重検定の補正の必要性

- 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**

多重検定の補正

1. **Bonferroniタイプ**

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

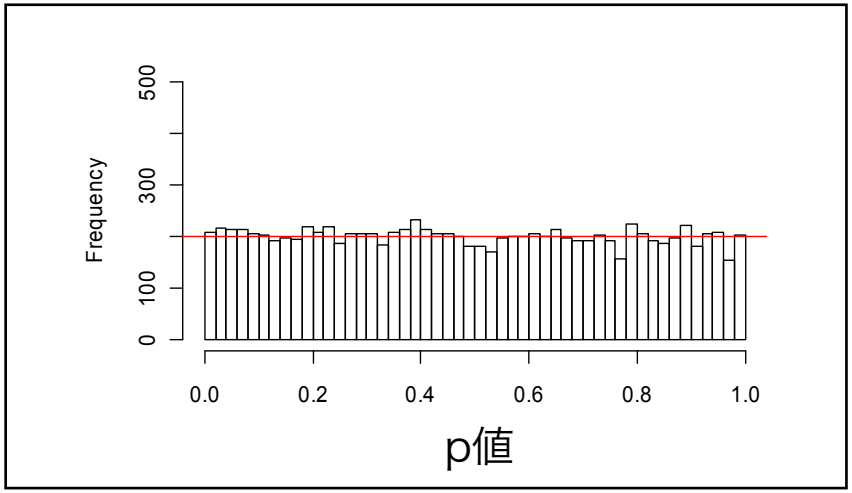
危険率を検定数で調整

危険率 ＝ α /k

α: 元の危険率、

k: 検定数

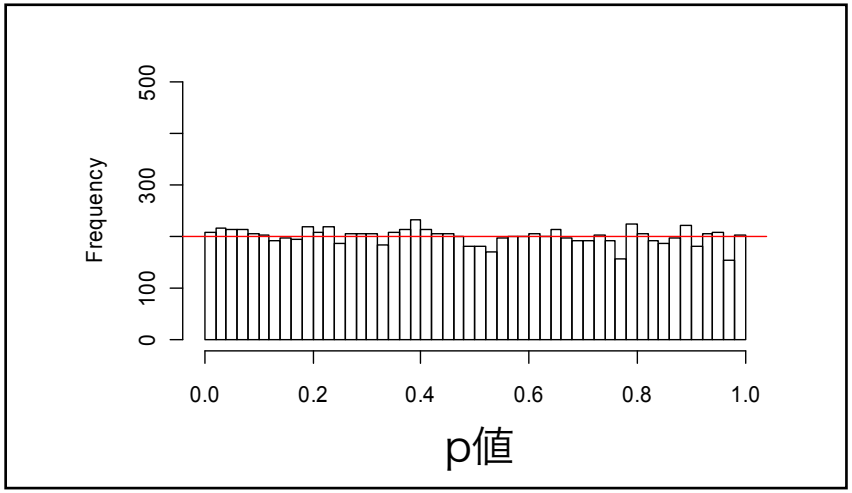
False Discovery Rate (FDR)



帰無仮説

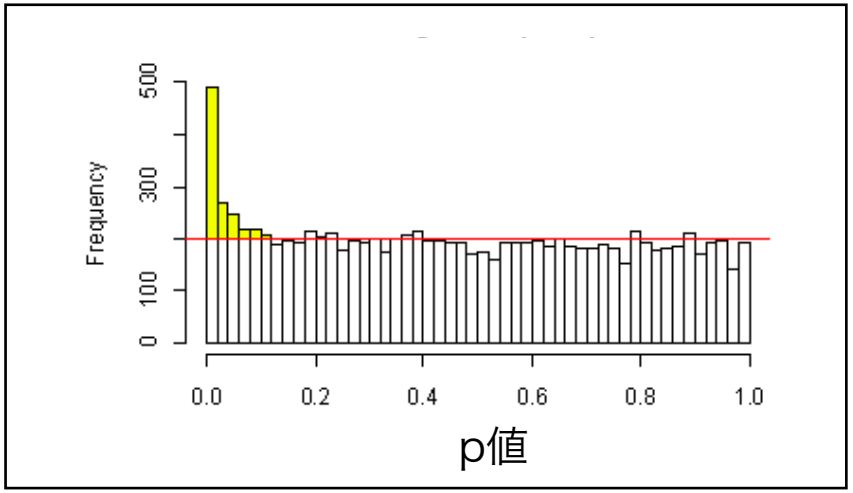
全ての範囲のp値が
同等の頻度で観察される
←どのp値を選んでも
ランダムに選ぶのと同じ

False Discovery Rate (FDR)



帰無仮説

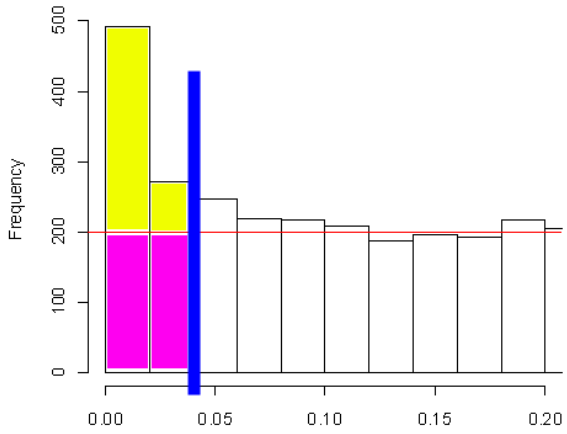
全ての範囲のp値が
同等の頻度で観察される
←どのp値を選んでも
ランダムに選ぶのと同じ



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？

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。

ポイント

*p*値、*q*値の違い

*p*値の視点: **FP/(TN+FP)**

*q*値の視点: **FP/(TP+FP)**

		検定	
		+	-
真の答え	+	True positive	False negative
	-	False positive	True negative

復習／発展学習

- *p*値、*q*値とは？
- 検定結果は確率
  - トランスクリプトーム解析では多数繰り返す
    → 多重検定の補正
- 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

→Bonferroni法等の多重検定の補正方法を
押さえておく

まとめ(2)

- 検定（解析）する仮説を設定し、検定する統計量を特定する
- それに従って検定手法を決定する
- ゲノムワイドな（多変量の）検定では、p値の取り扱いに注意。補正が必要なケースがほとんど。

パラメーター評価後の解析

解析の再検討は必要？

- 実験データの質　**再現性/ばらつきの統計量**
- データ全体の吟味　**多変量解析**
- パラメーターごとの評価　**検定、多重検定の補正**

↓ フィルタリング

- サブセットでの評価

解析の再検討は必要？

- 例: DEGのみでの相関係数

  測定したごく一部しかDEGではない。
  つまり、処理間の相関係数はほとんどの発現変動していない遺伝子群で計算されている。
- 自身が知りたい、論文で主張したい内容はどのデータ・どの統計量で表現すべきか

まとめ

- ゲノムワイドな測定データを扱うポイントを掴み
- 統計的な視点で考えられるようになる

視て

- 多変量の可視化
- 統計量での要約

分析する

- 仮説構築
- 検定