

# 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

# 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

そのためには

- 測定、実験計画を見直せるように
  - 仕組みを知る
  - 試す - R
- 統計用語・表記に慣れる

# 基本的な統計の用途

- 仮説検定
- 予測 (モデル構築)

# 仮説検定 - $t$ 検定を例に

# ねらい

## **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

# 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？



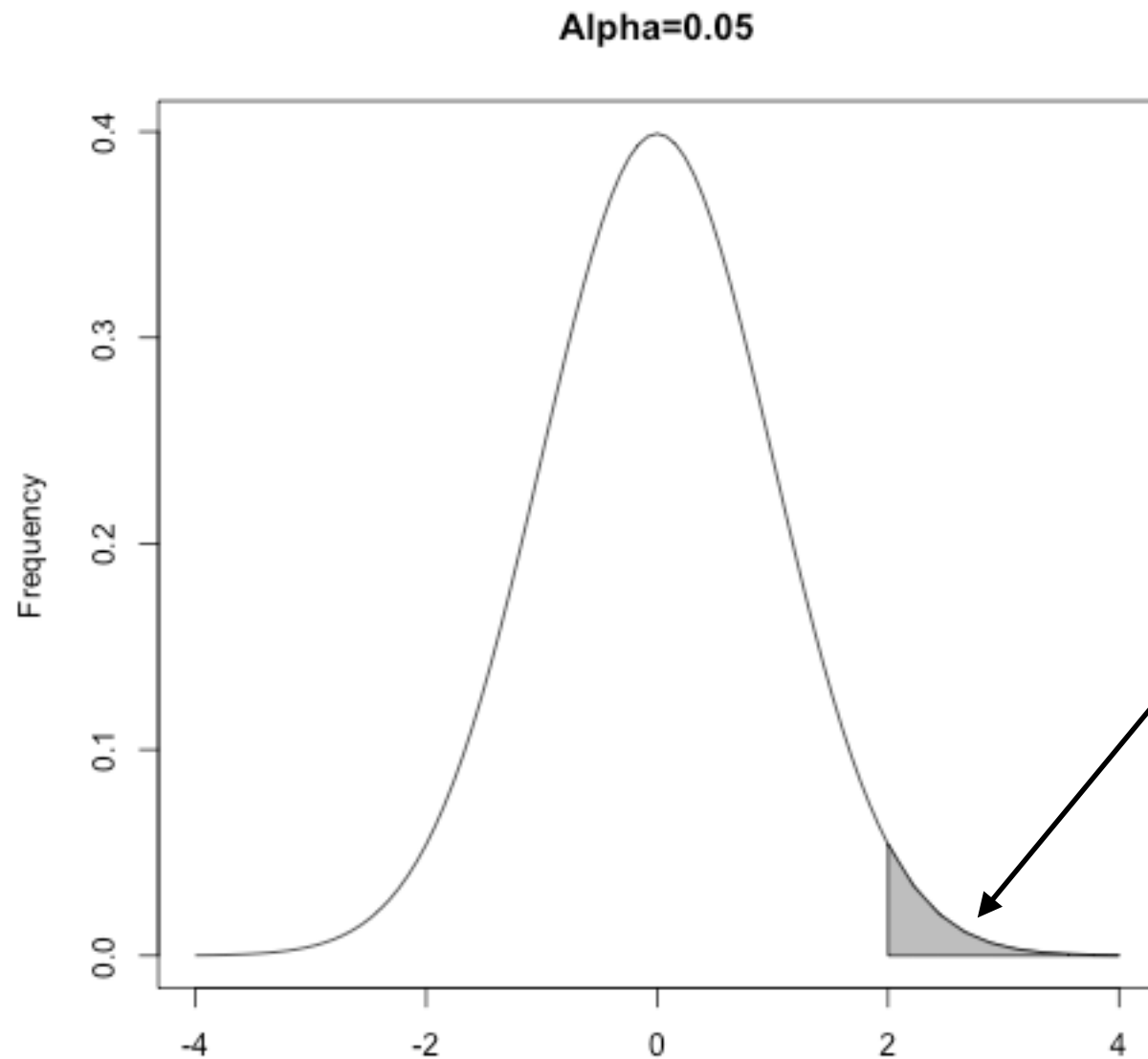
## 2. 統計量を求める:

**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

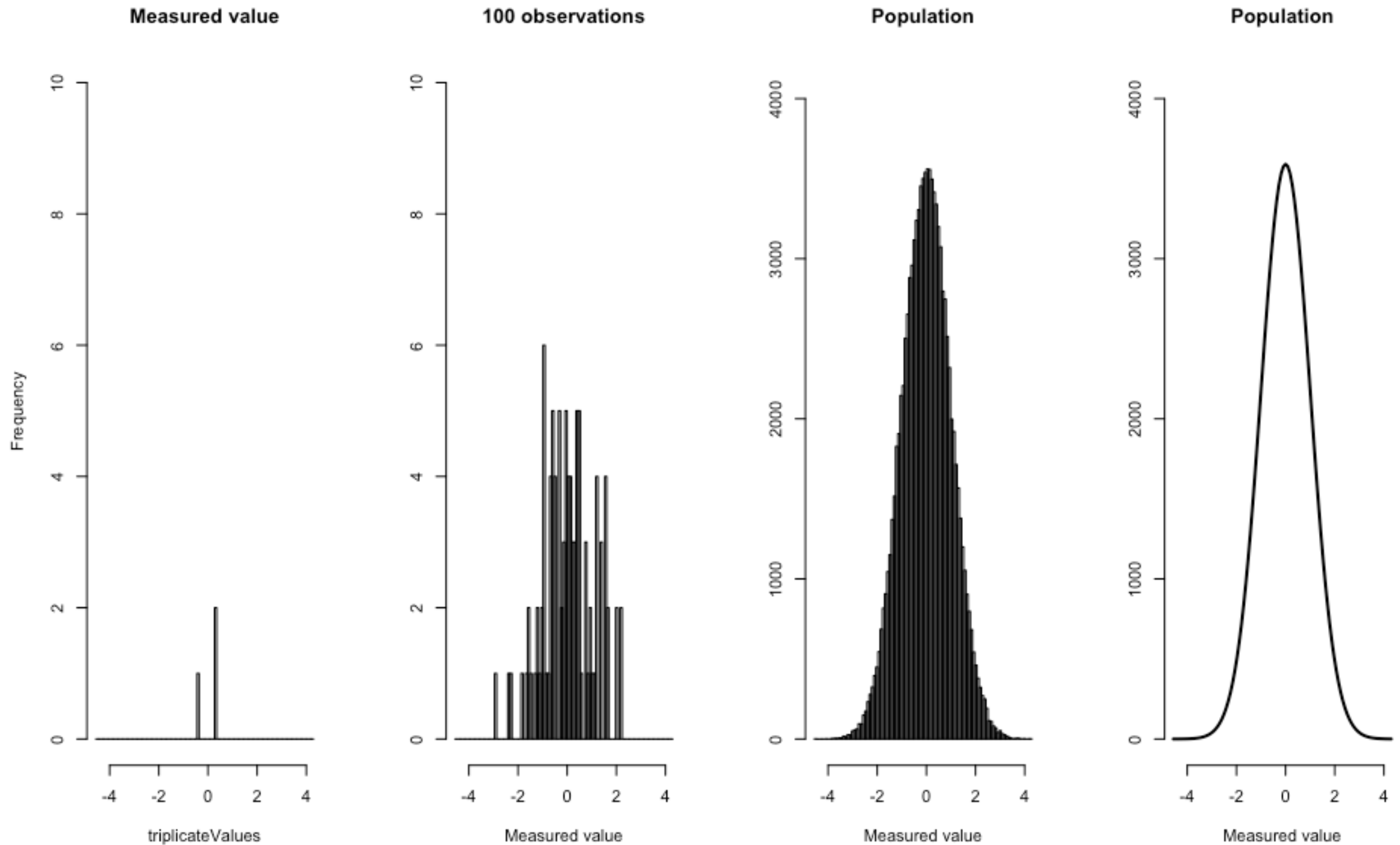
### 3. 確率分布と照らし合わせる



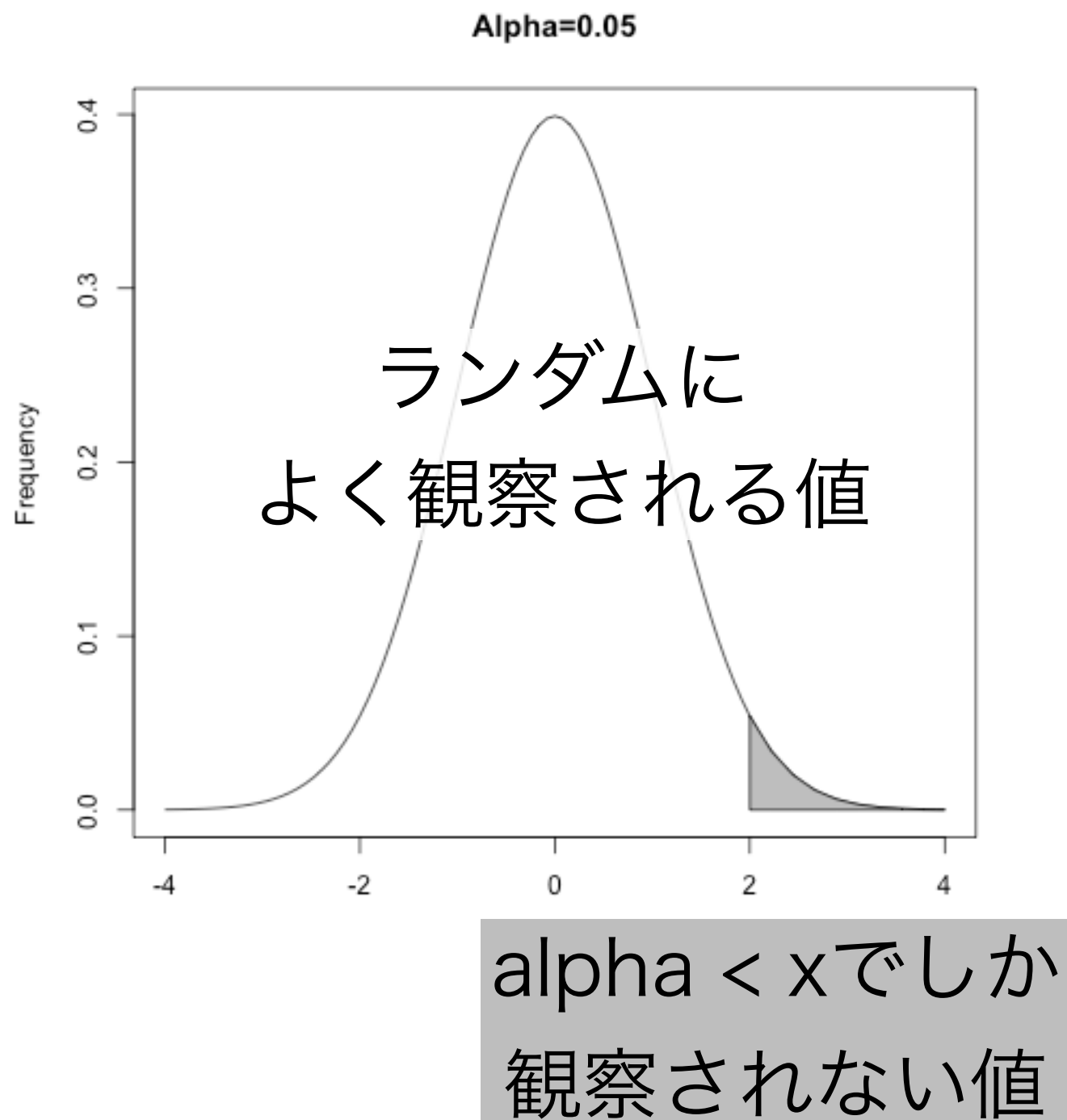
棄却限界値によって  
規定される面積  
(通例: 全体の5%)

統計量

# 確率分布？面積？



## 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

## 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

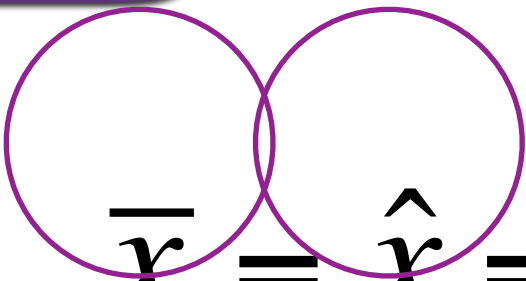
**母数** : 未知の数値

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

すべてのデータを足して、データ数で  
割る値


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

**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」

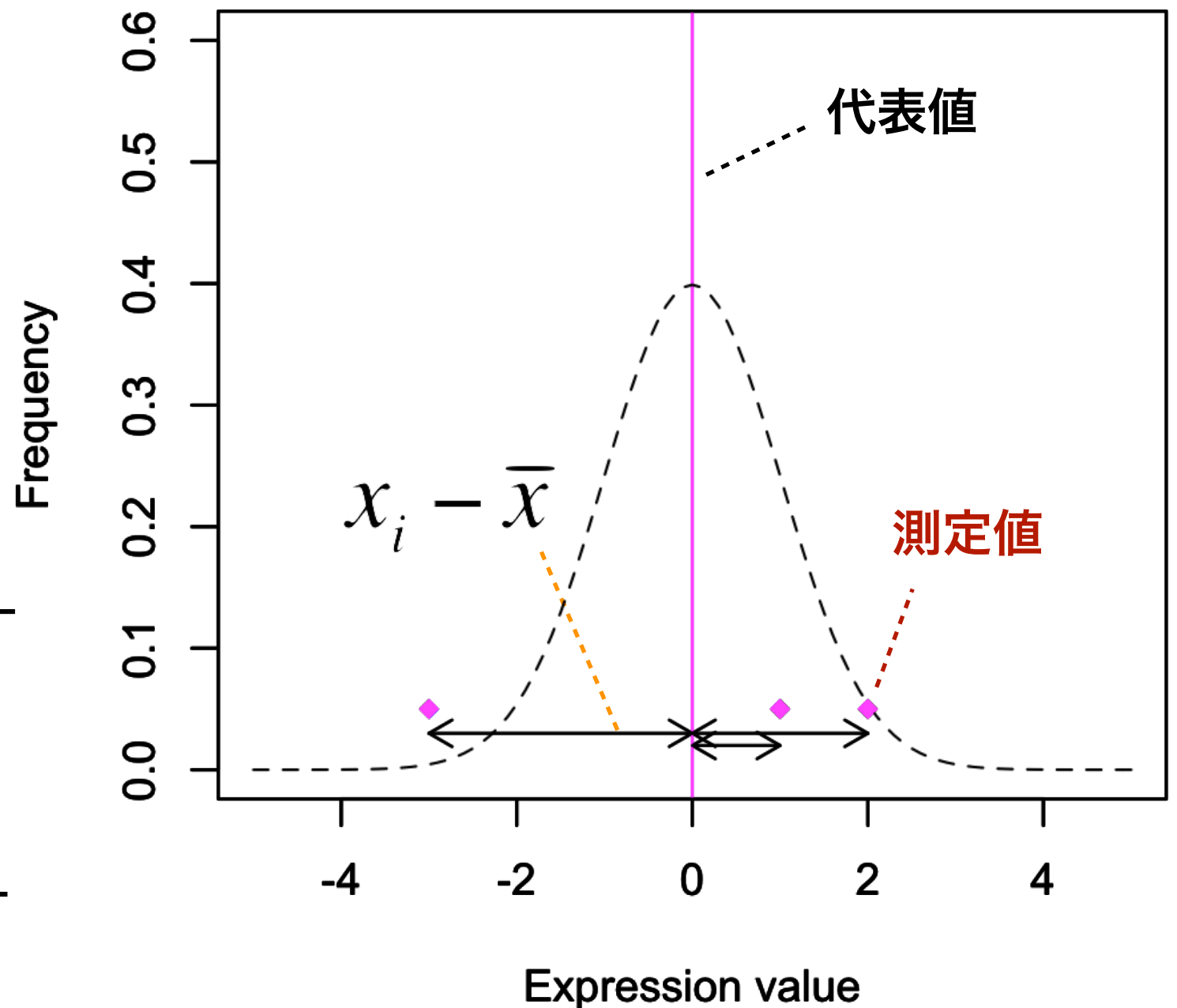
# ばらつき：分散／偏差

分散:

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

標準偏差:

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





# $n-1$ ?

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？

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

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

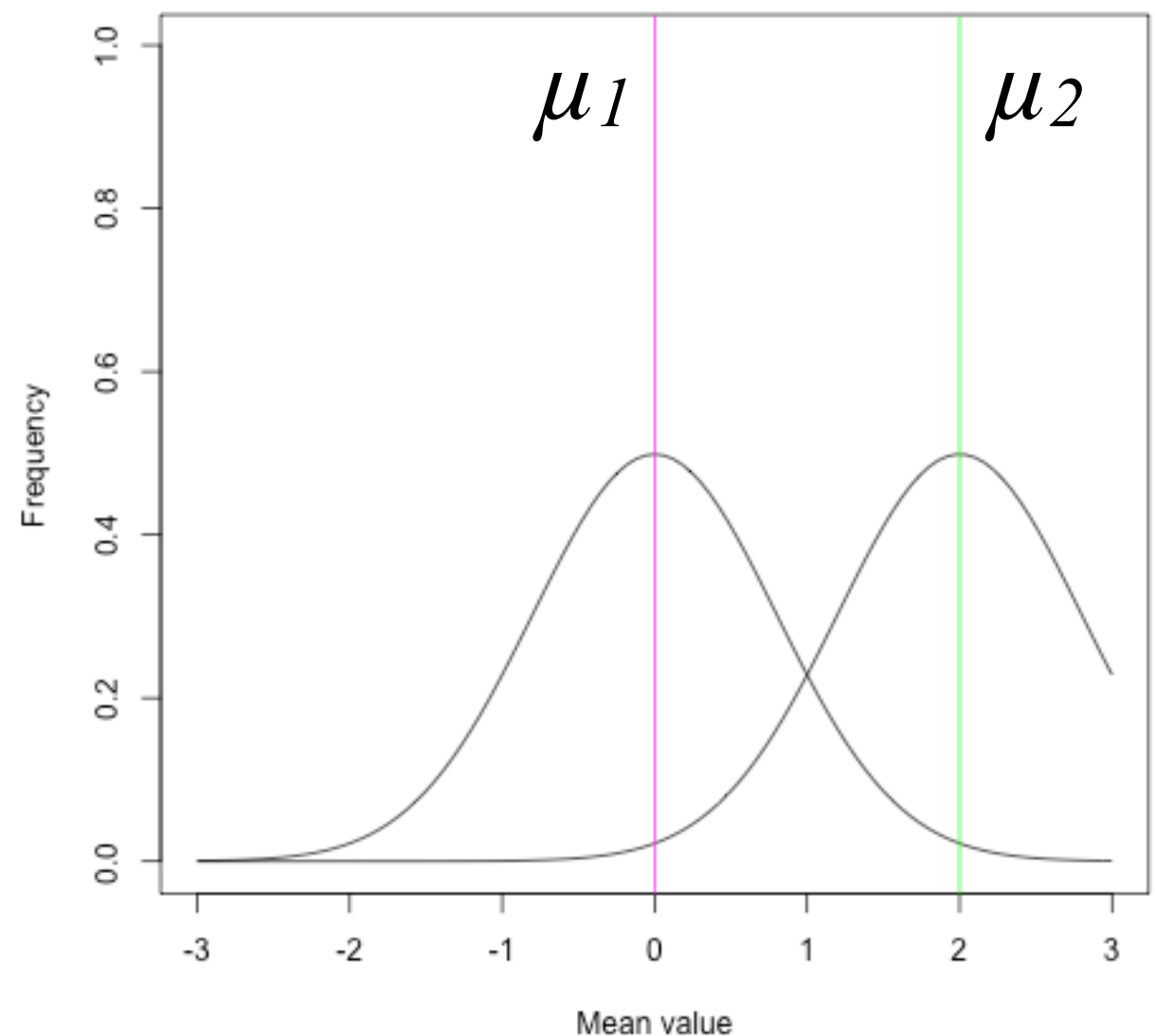
自由度: 統計量を求めるのに使うことができる「独立」な標本数

# $t$ 検定:

## 2サンプルの平均の検定

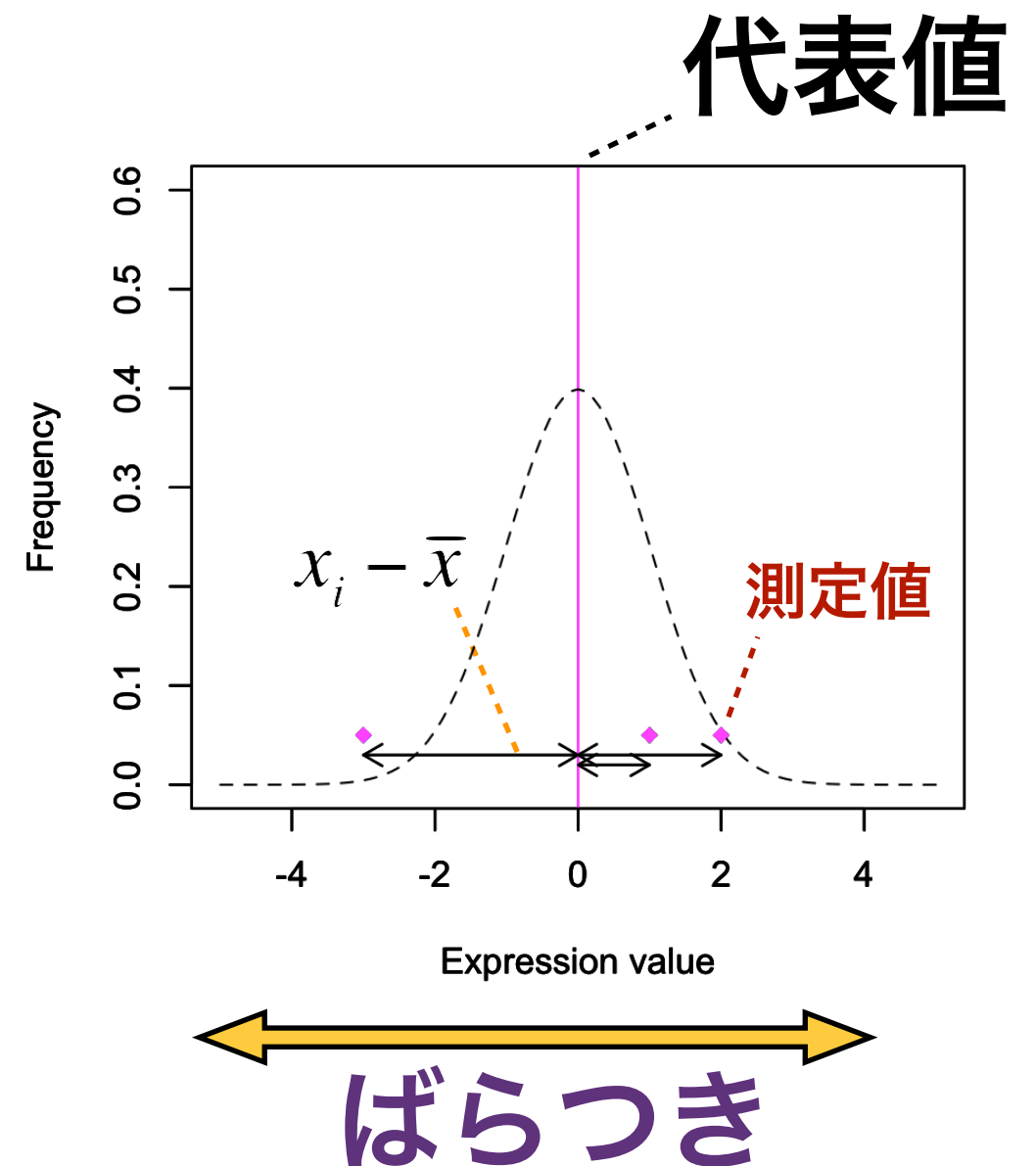
- 平均値 =  $\mu_1, \mu_2$
- データは正規分布

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

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

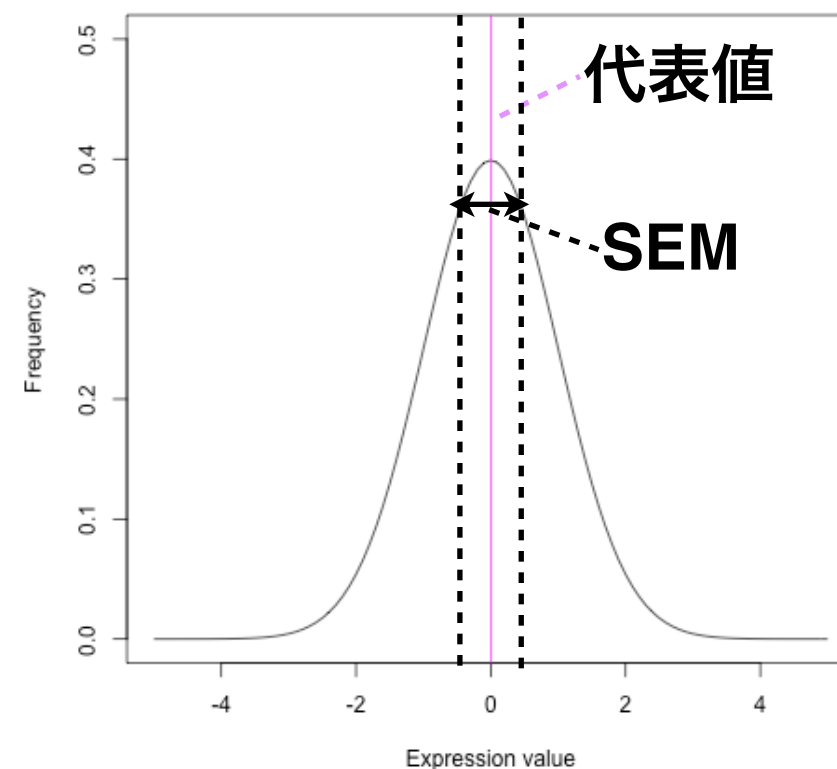
# 統計量その2:

*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

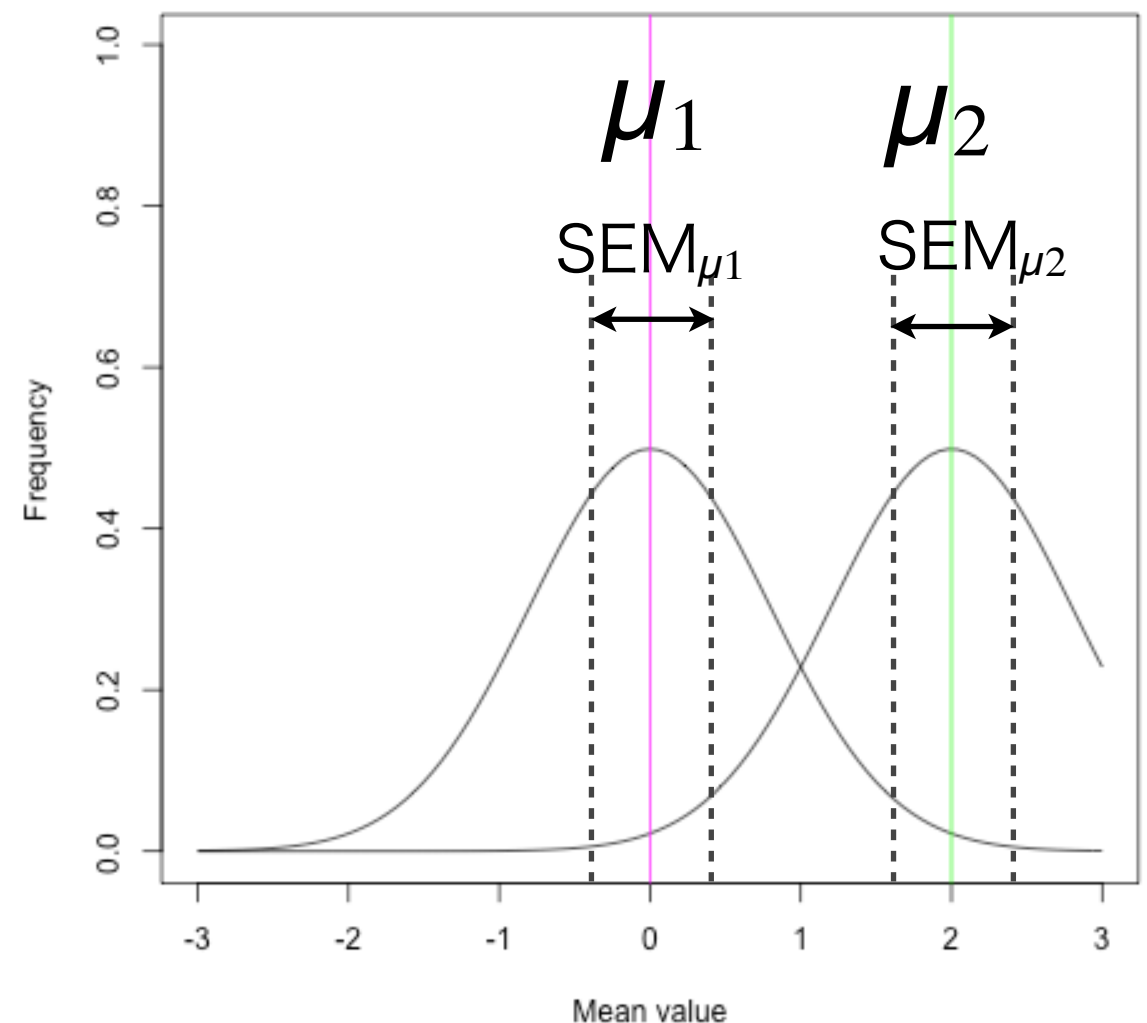
*statistical  
mind*

# 統計量その3: 平均の差とその誤差

t統計量

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

$$SEM_{|\hat{\mu}_1 - \hat{\mu}_2|} \quad | \quad |\hat{\mu}_1 - \hat{\mu}_2|$$



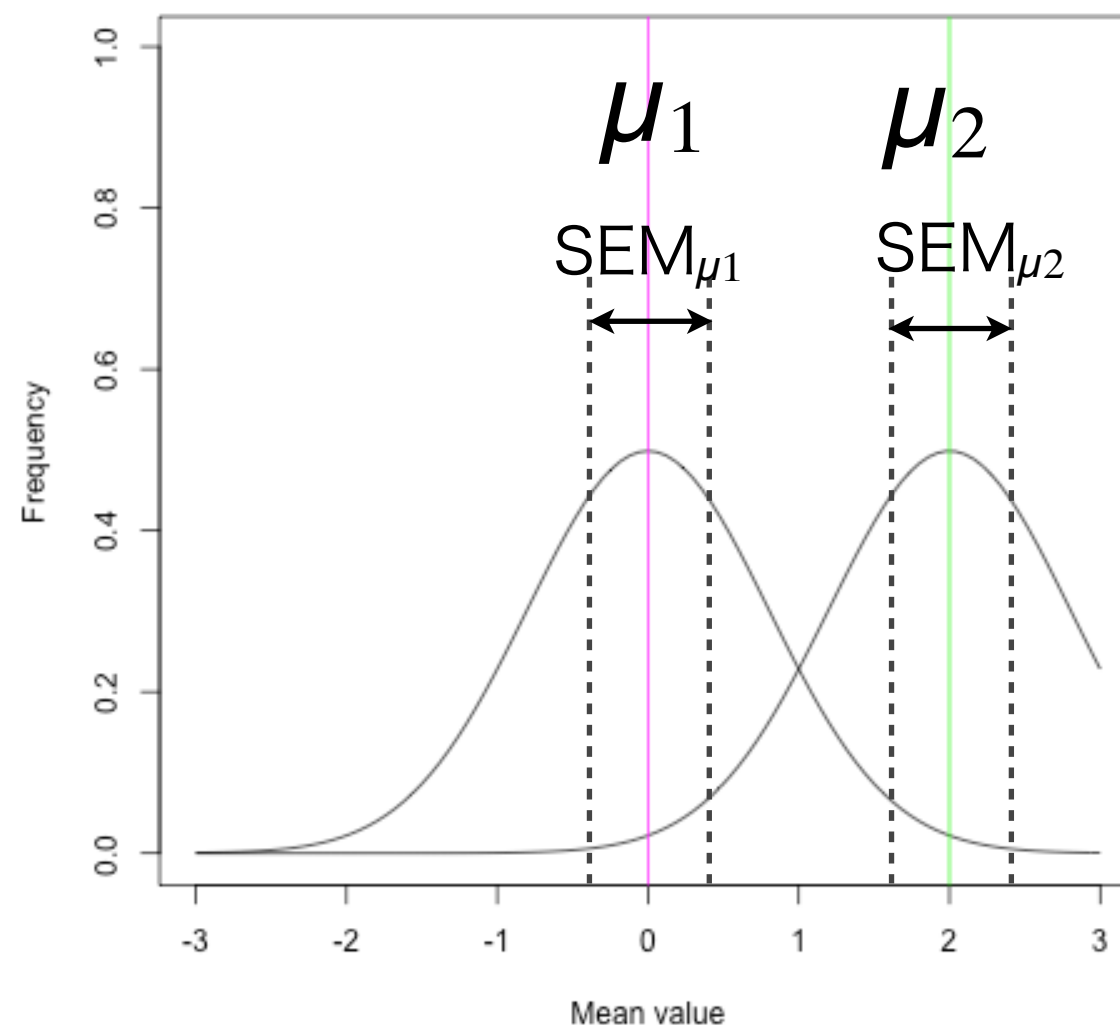
# 統計量その3: 平均の差とその誤差

$$SEM_{\hat{\mu}_1 - \hat{\mu}_2}$$

$$SEM_{|\hat{\mu}_1 - \hat{\mu}_2|}$$

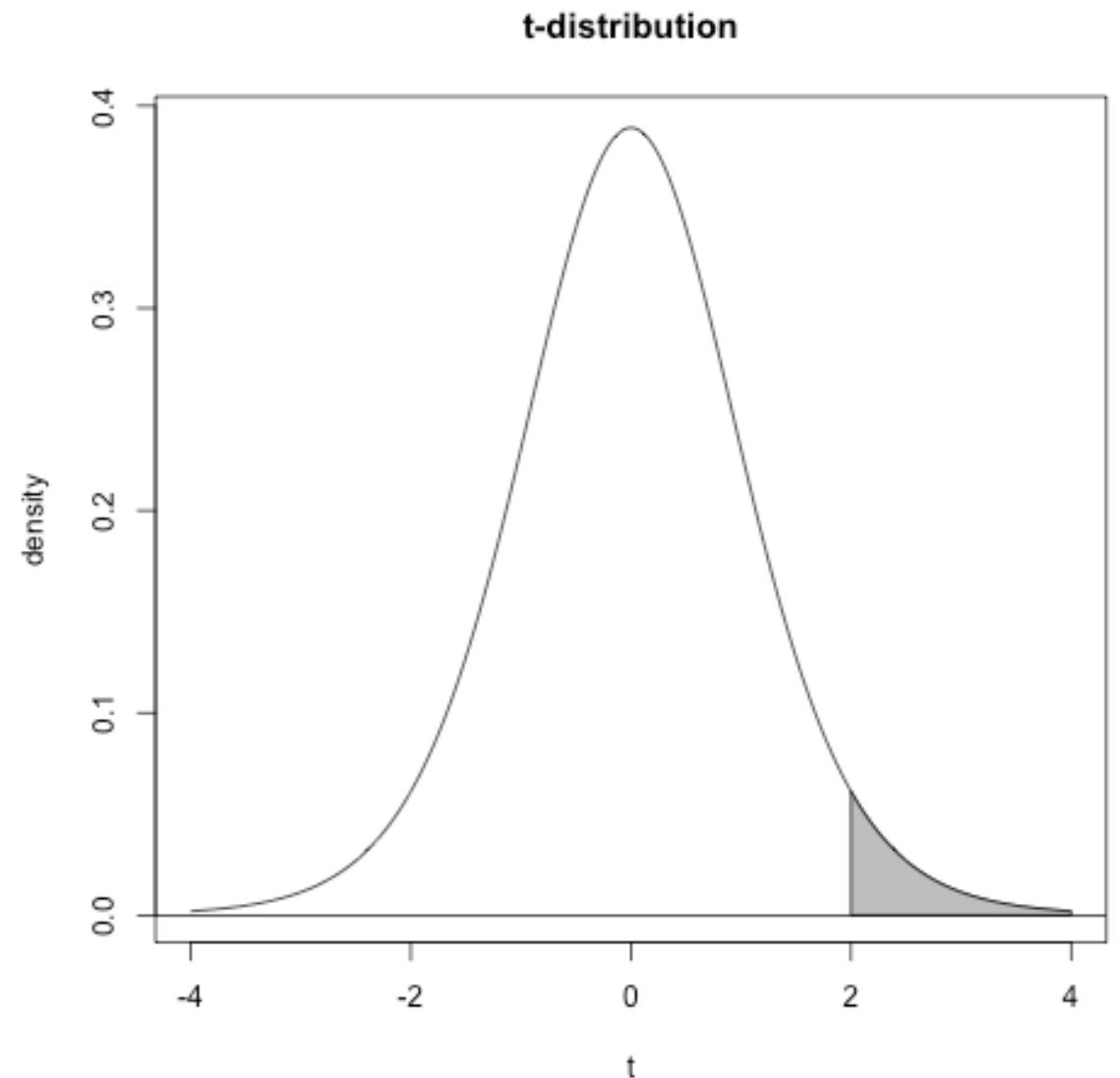
$$= \sqrt{SE_{\hat{\mu}_1}^2 + SE_{\hat{\mu}_2}^2}$$

$$= \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$



# 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

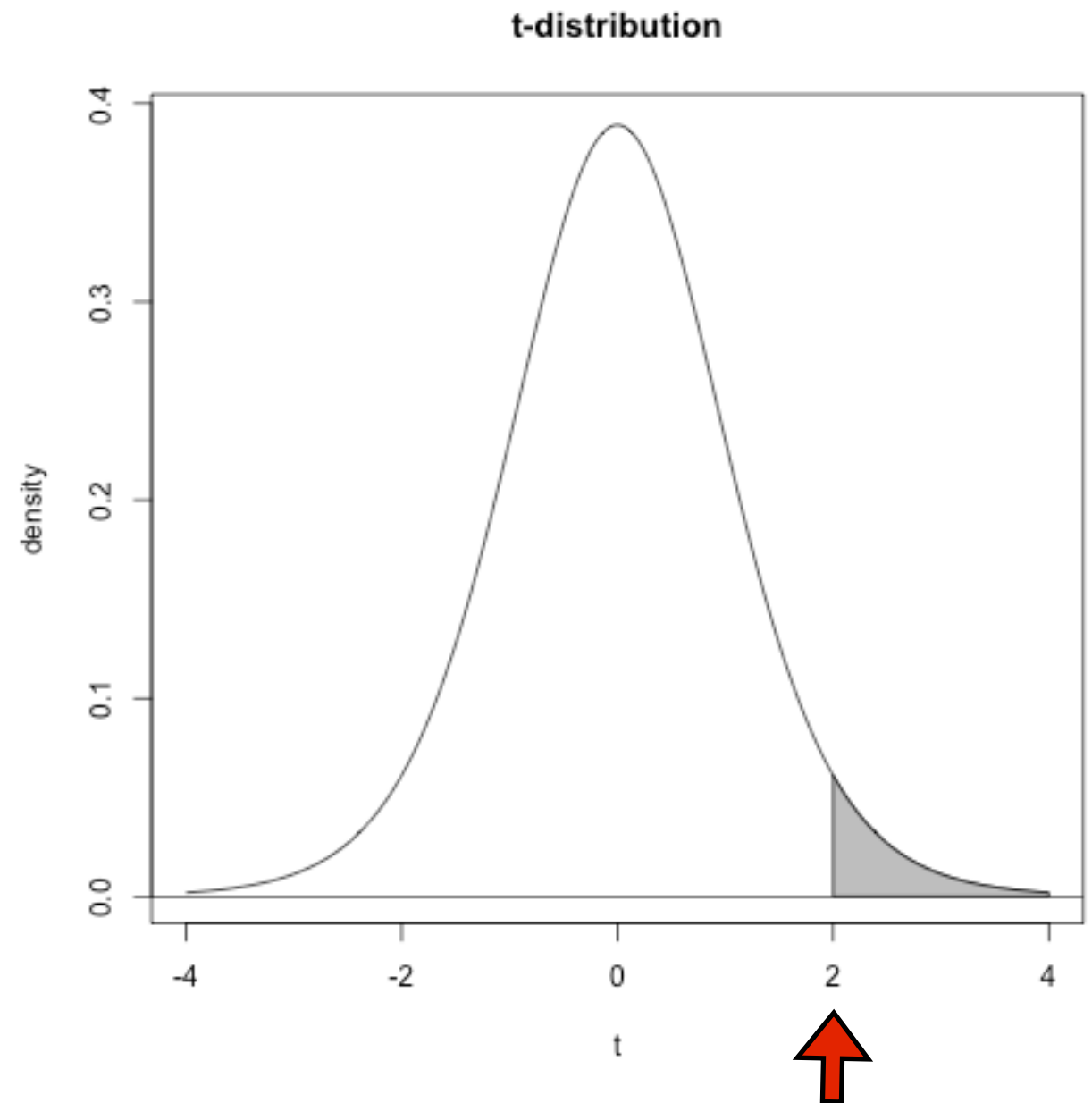


【おさらい】 自由度: 統計量を求めるのに使うことができる独立な標本数



$p$ 値とは：

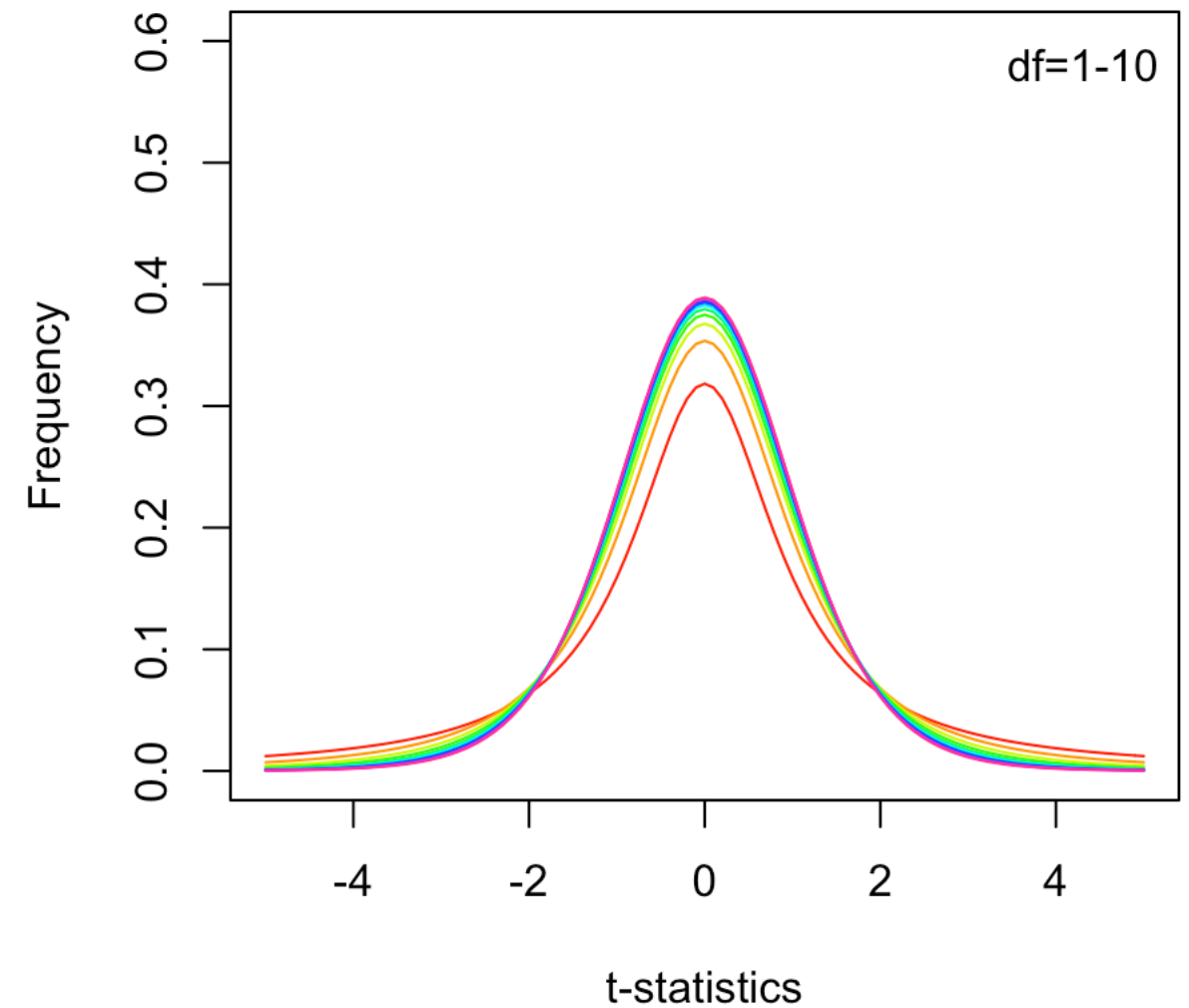
- 標本に基づいた統計量が帰無仮説の下、起きうる確率
- 汎用される閾値（危険率）：0.05



# データの分布、仮説検定に即した確率分布を使う

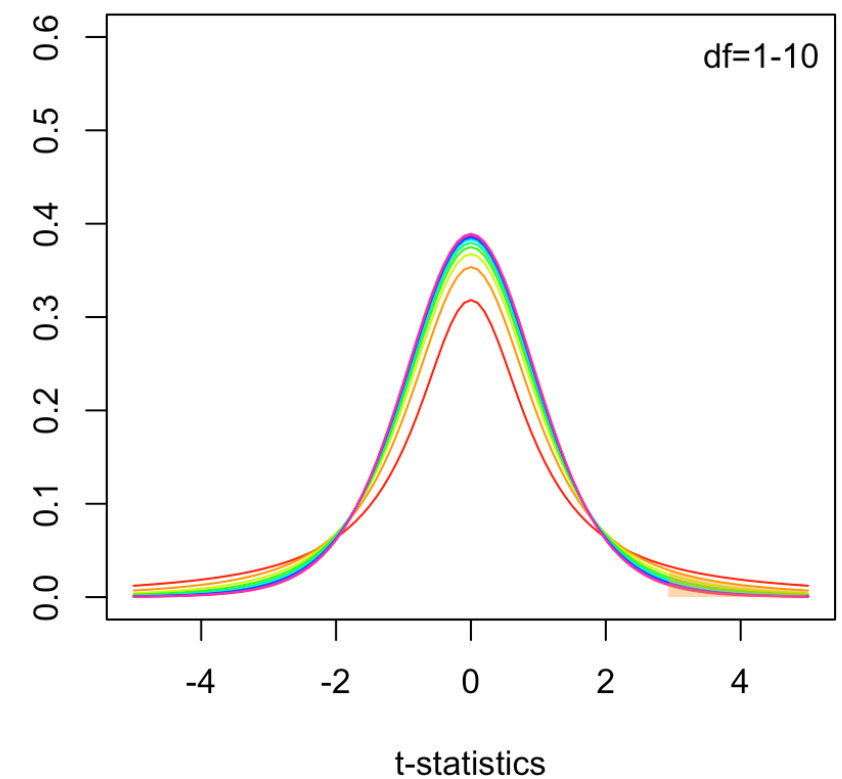
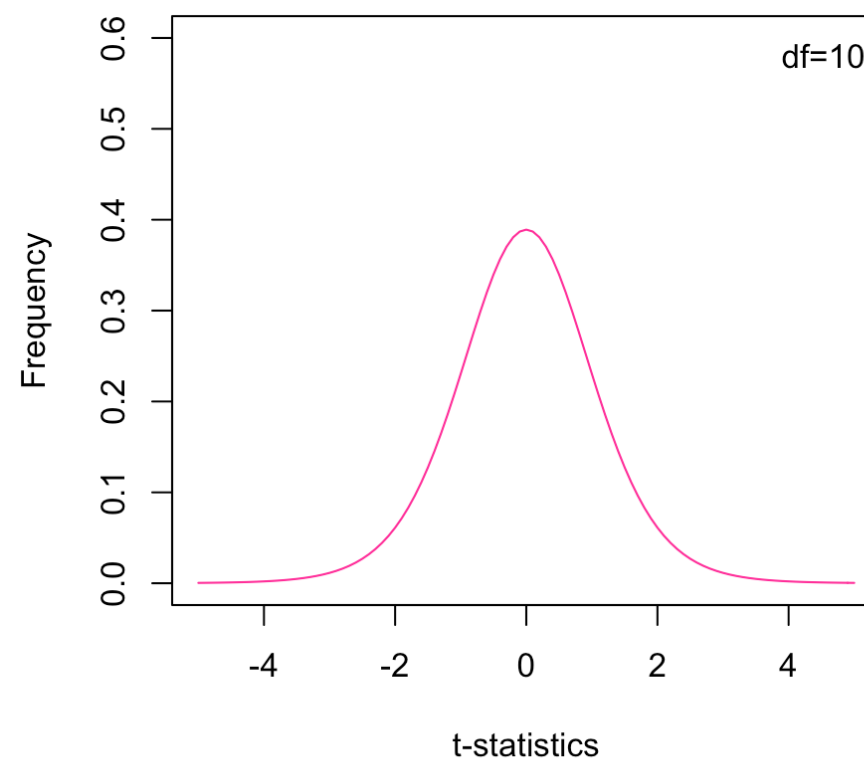
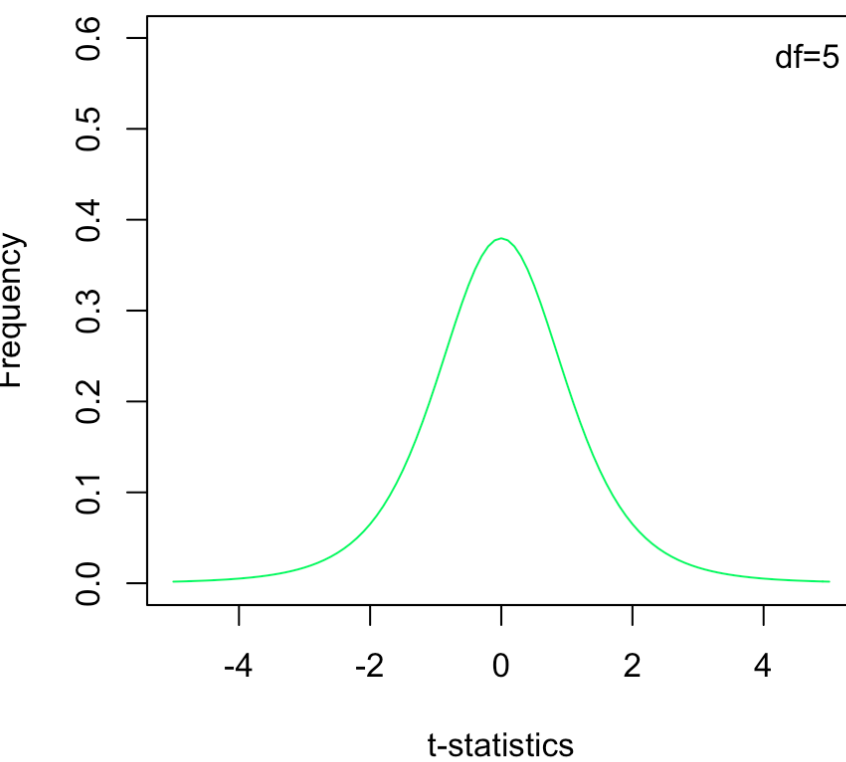
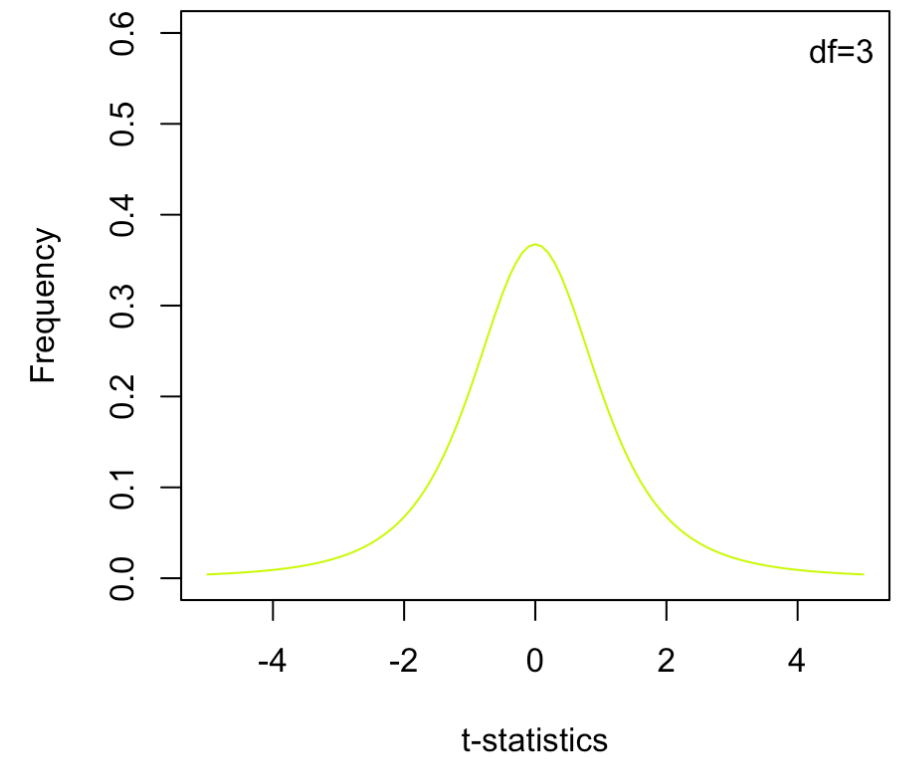
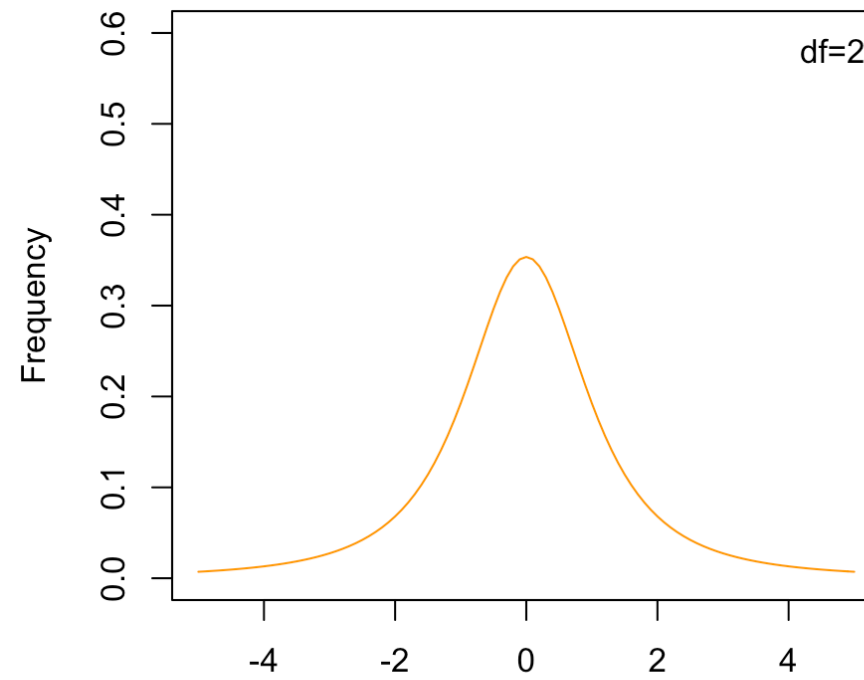
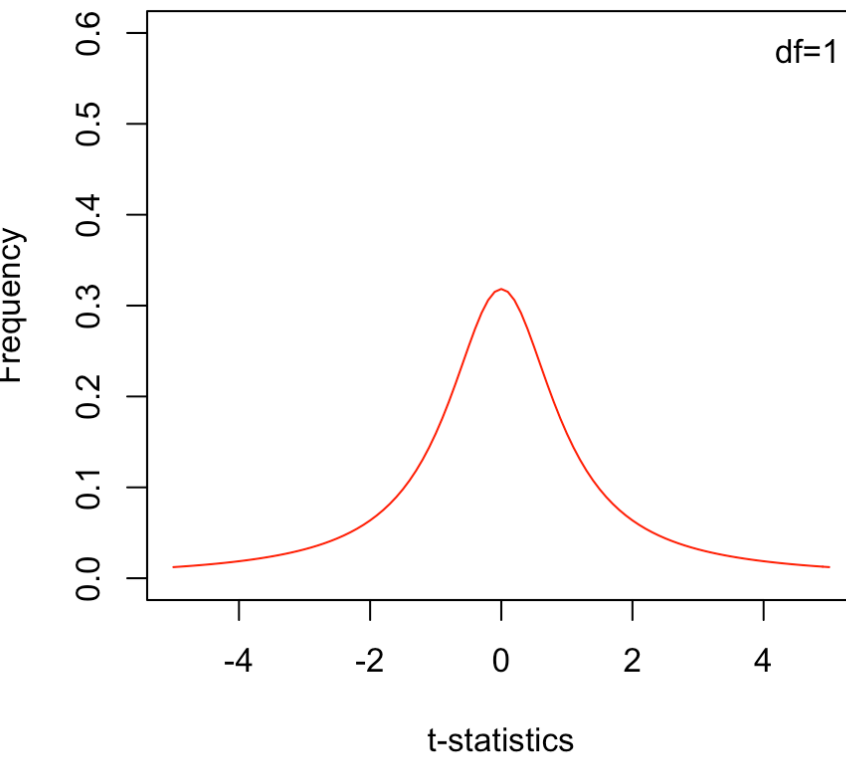
我々の測定では

- ・ 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

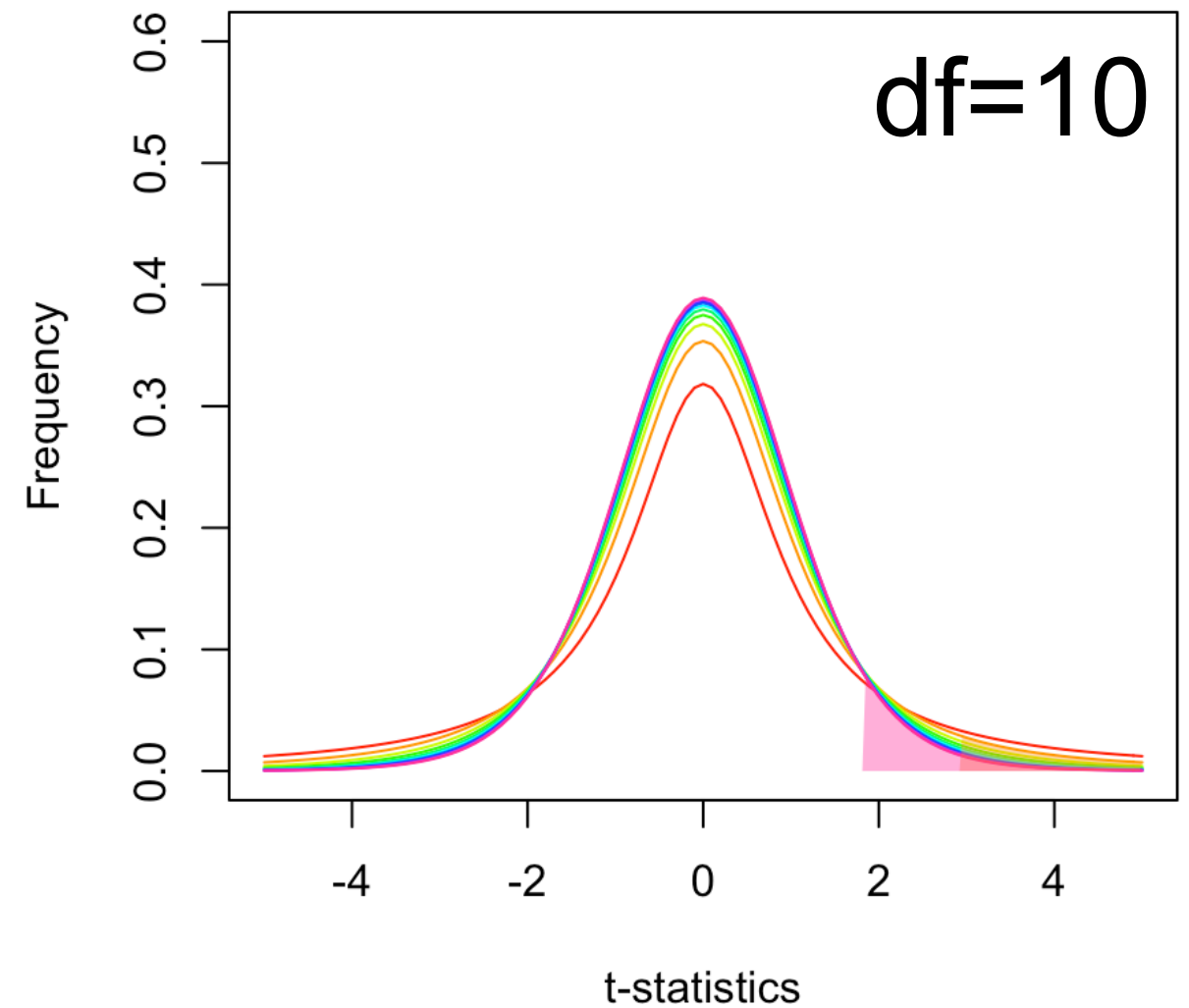
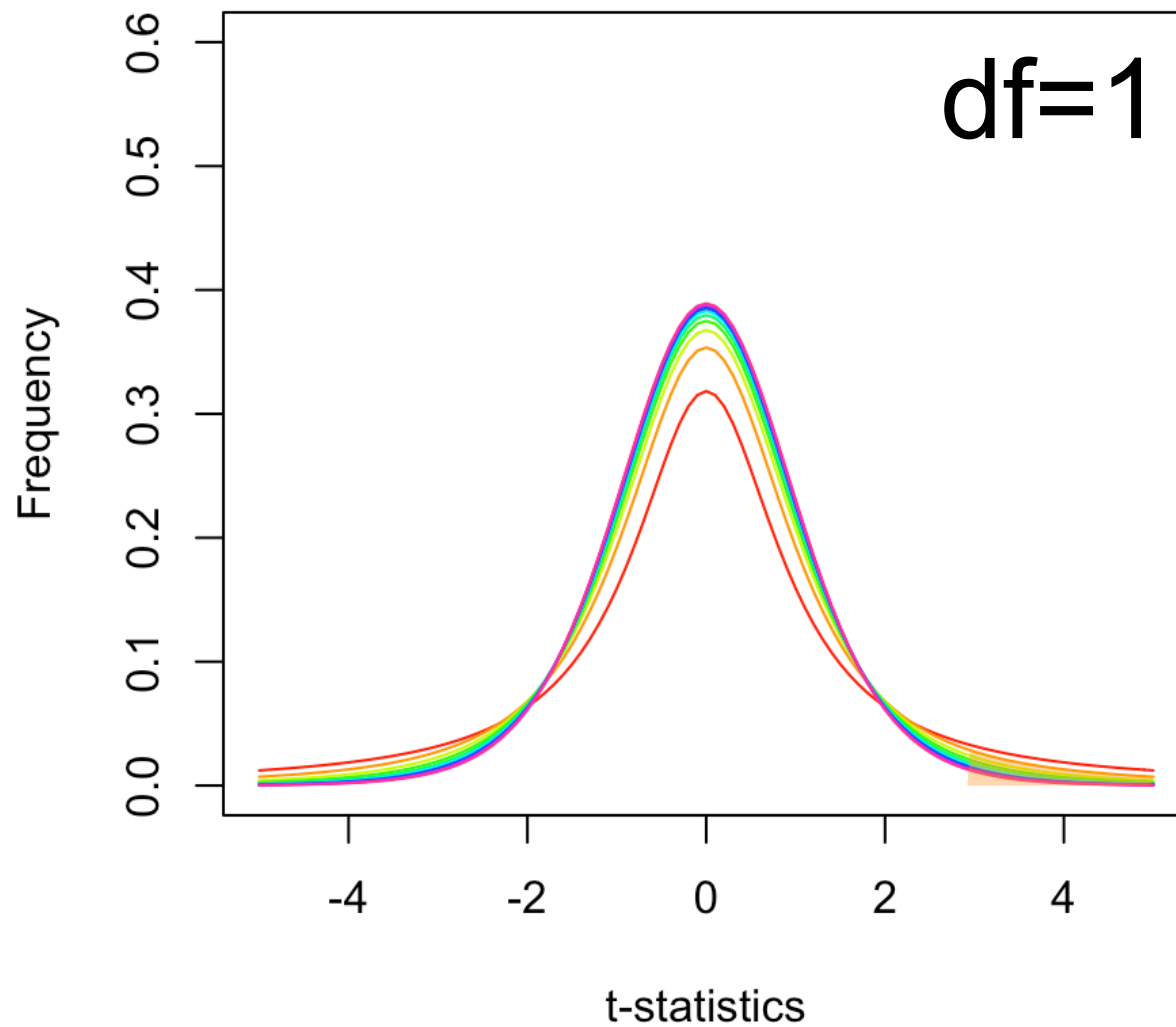


$$t = \frac{|\hat{\mu}_1 - \hat{\mu}_2|}{SEM_{|\hat{\mu}_1 - \hat{\mu}_2|}}$$

# $t$ 分布と自由度の関係



# $t$ 検定における自由度の違い: 検出力の違い



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

*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

# 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

ほぼ全ての検定方法に  
前提がある

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= $\lambda$ ）

**負の二項分布** ( $\lambda$ がガンマ分布)



# 二項分布：離散値の確率分布の基本

$$P(k) = \underbrace{{}_n C_k}_{\text{組み合わせ}} \underbrace{p^k}_{\text{確率}} \underbrace{(1-p)^{n-k}}_{\text{確率}}$$

$n$ 回の試行で事象が	事象が	事象が
$k$ 回観察される	観察される	観察されない
組み合わせ	確率	確率

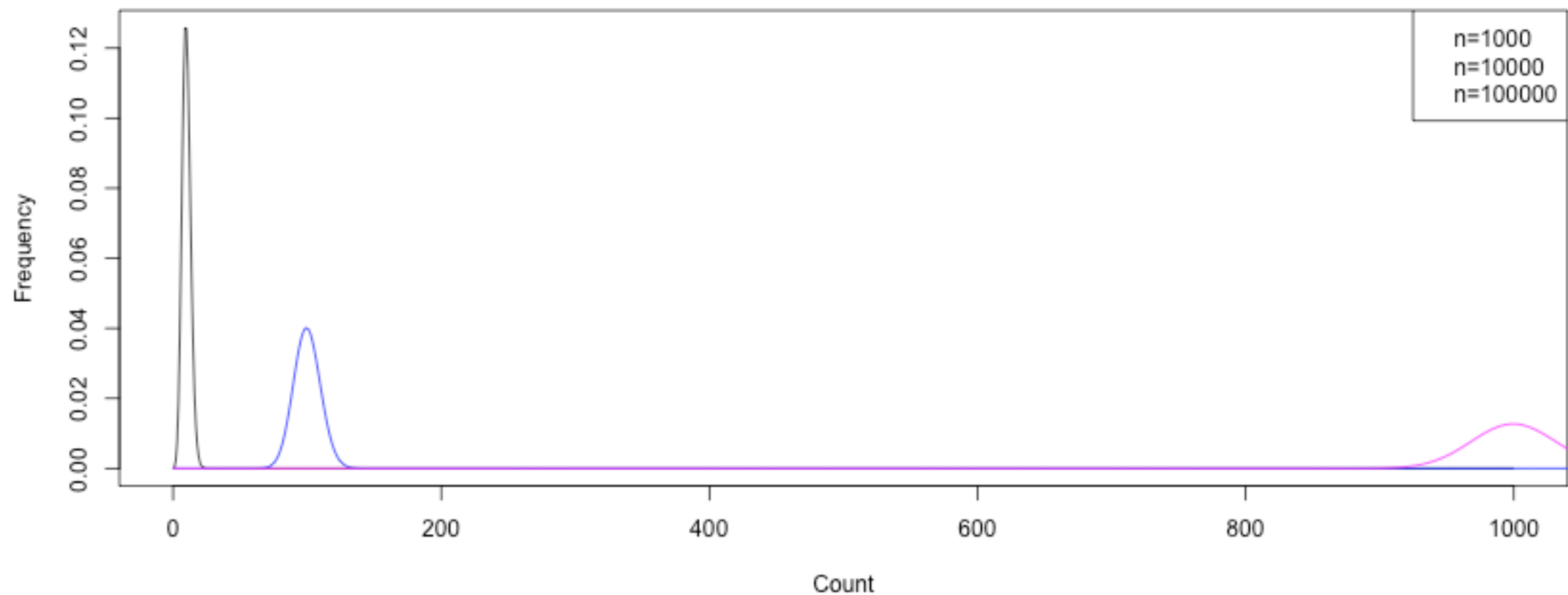
## 二項分布:

総リード数中のある遺伝子にマップ  
されたリード数として考えると

$$P(k) = \underbrace{{}_n C_k}_{\text{組み合わせ}} \underbrace{p^k}_{\text{観察される確率}} \underbrace{(1-p)^{n-k}}_{\text{観察されない確率}}$$

$n$ 個のリードから	その遺伝子	その遺伝子
$k$ 回カウントされる	リードが	リードが
組み合わせ	観察される	観察されない
	確率	確率

# 二項分布の限界

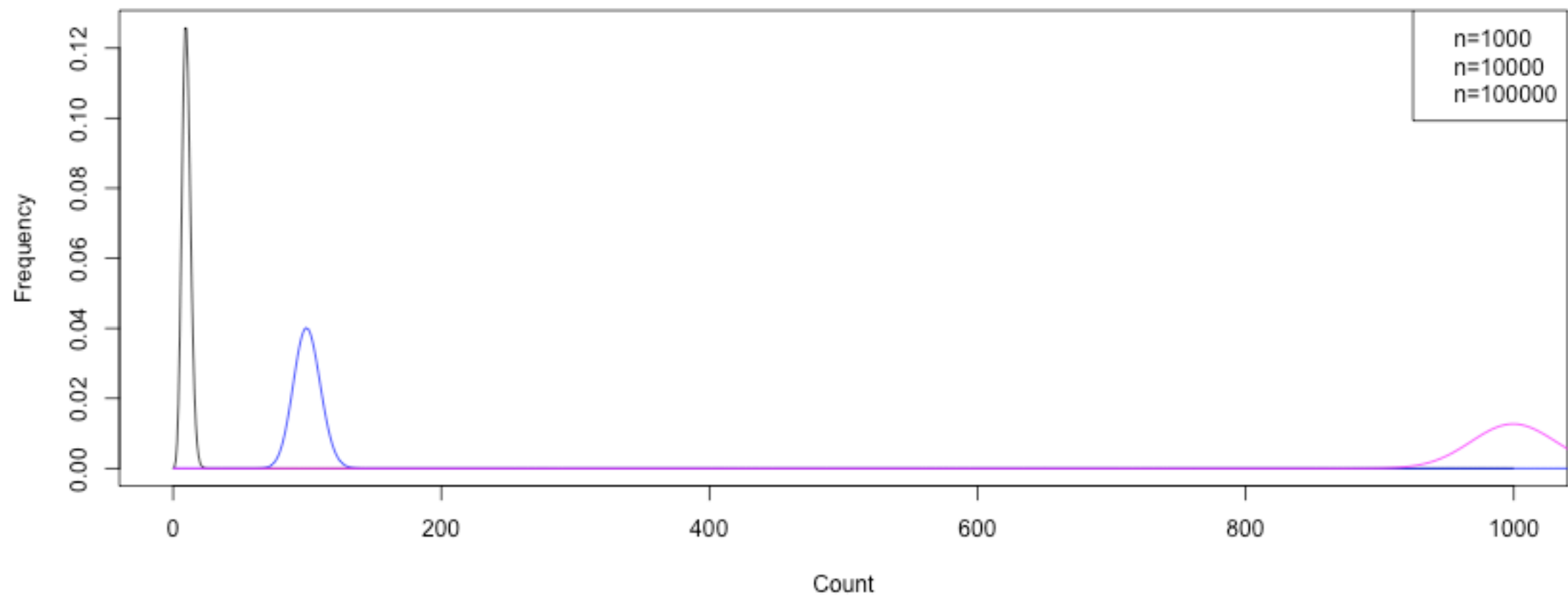


$$P(k) = \underbrace{n}_{0.01} \underbrace{C_k}_{0.99} \underbrace{p^k (1-p)^{n-k}}_{0.99}$$

リード数が増えれば  
増えるほど

- 観察確率が下がっていく
- 計算上の問題

# ポアソン分布の登場

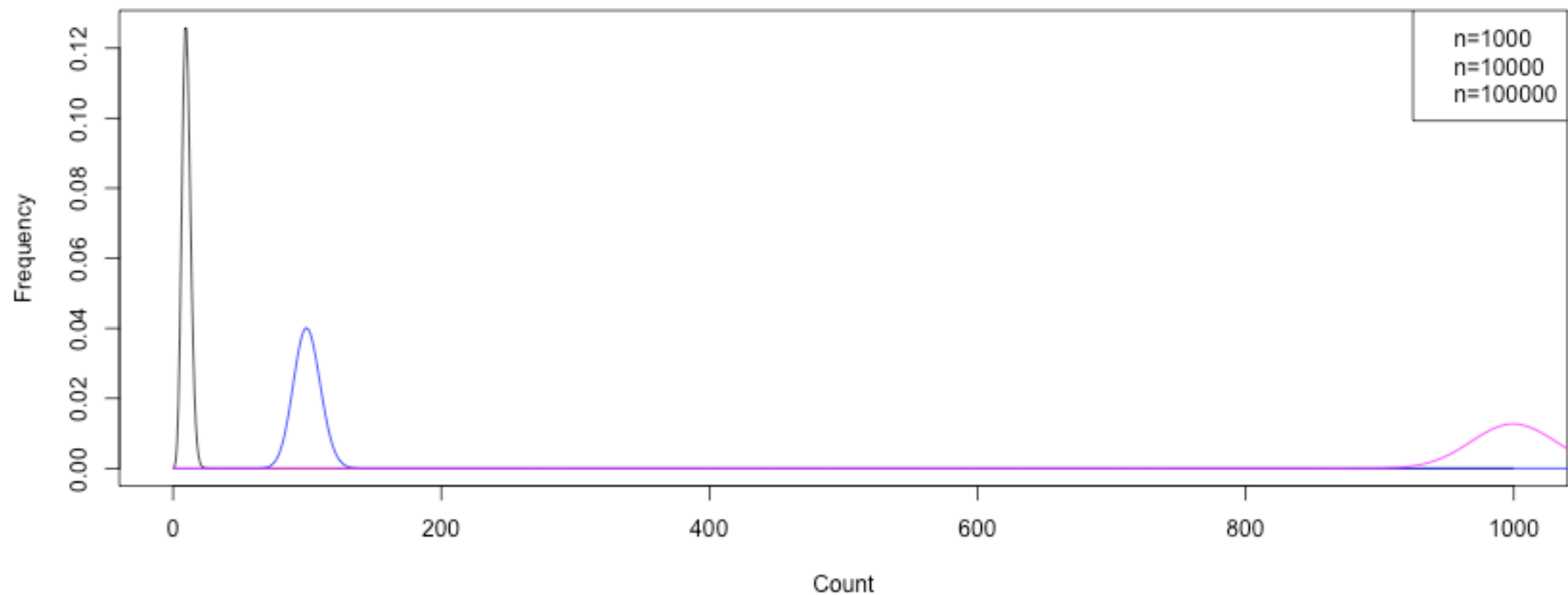


$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

ある遺伝子が $\lambda$ 回カウントされる事象が  
 $k$ 回観察される確率

平均:  $\lambda$ , 分散:  $\lambda$

# ポアソン分布の問題



$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

平均:  $\lambda$ , 分散:  $\lambda$

実データでは「平均=分散」  
が当てはまらない

**過分散 overdispersion**

# 負の二項分布の登場

$$P(k) = \underline{{}_k+r-1 C_k} p^k (1-p)^r$$

$r$  = 失敗回数  
(ノイズ)

$$P(k) = \underline{{}_n C_k} \underline{p^k} \underline{(1-p)^{n-k}}$$

$n$ 回の試行で事象が  $k$ 回観察される  
組み合わせ

事象が

観察される確率

事象が

観察されない確率

ほぼ全ての検定方法に  
前提がある

**データが取る分布が不明？**

→データから推定する

**リサンプリング**：観測（測定）データを  
母集団とした標本抽出

# リサンプリングの活用場面

## 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

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？



# 相関係数の帰無仮説の分布が不明？

→データから推定する

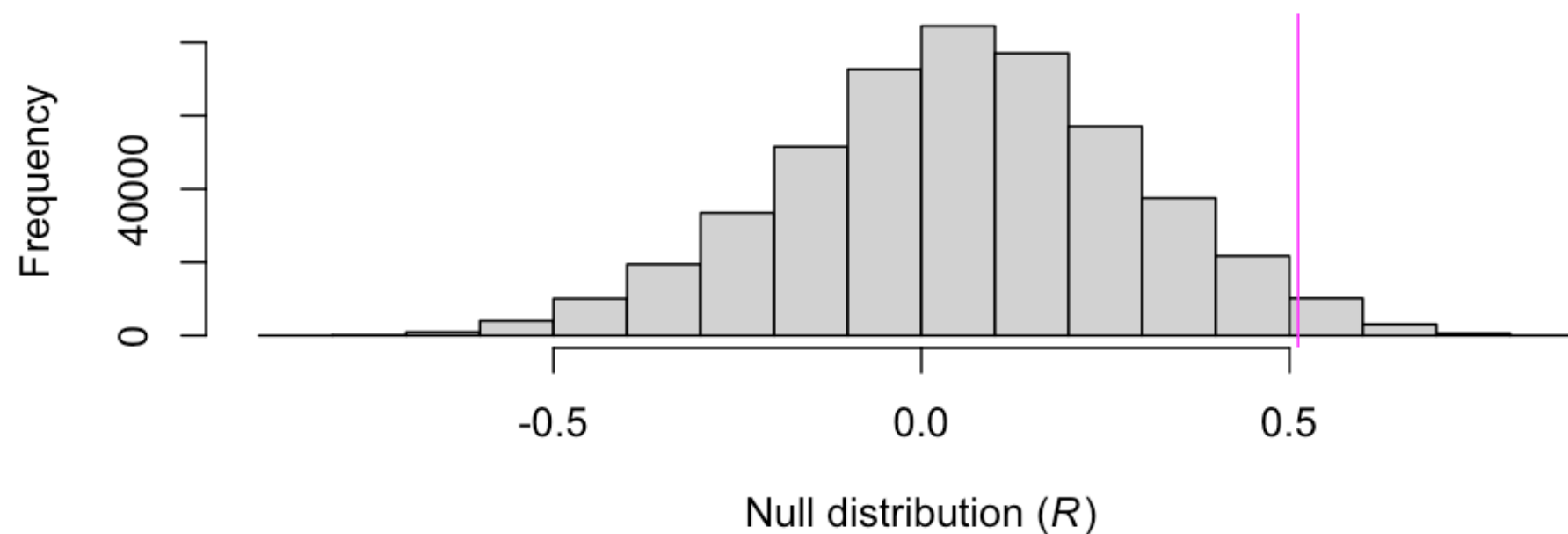
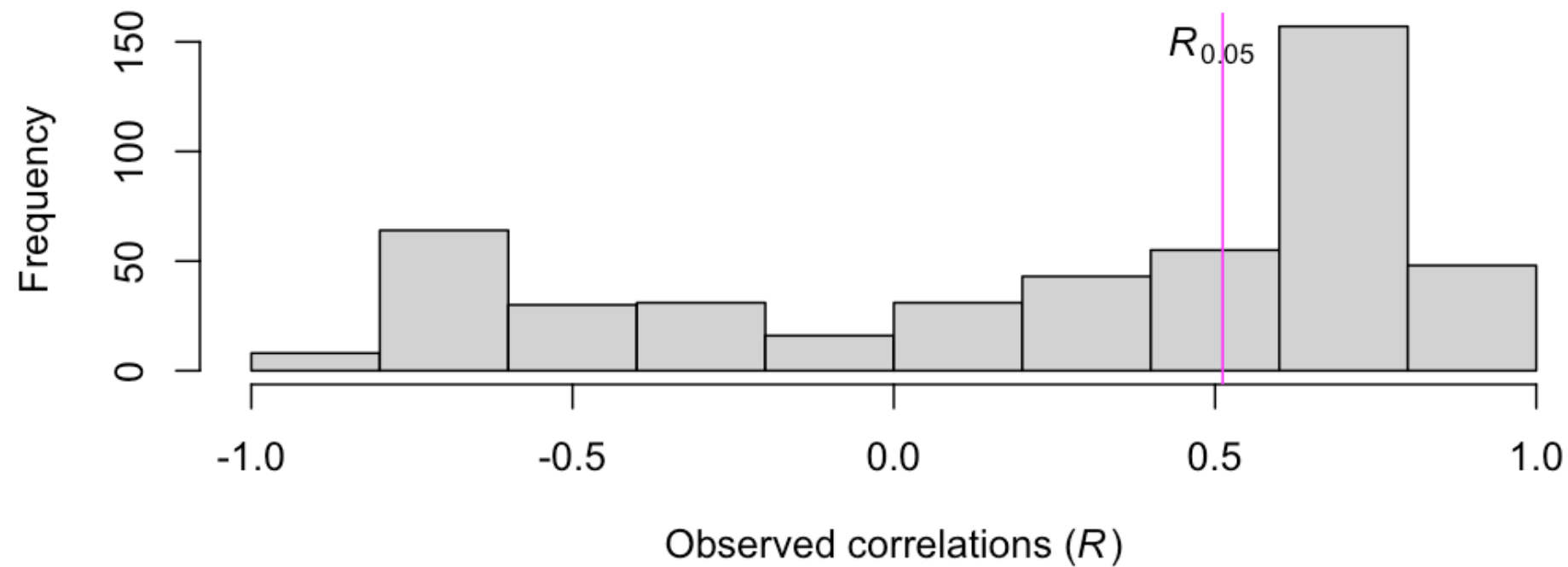
帰無仮説：遺伝子Aと遺伝子Bの  
発現プロファイルは相関していない



データをランダムに並べ替え、  
「**相関しない**」発現プロファイルを作り、  
相関係数を計算する  
(発現データを**無作為抽出**する)

# 相関係数の帰無仮説の分布が不明？

→ データから推定する



# 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：

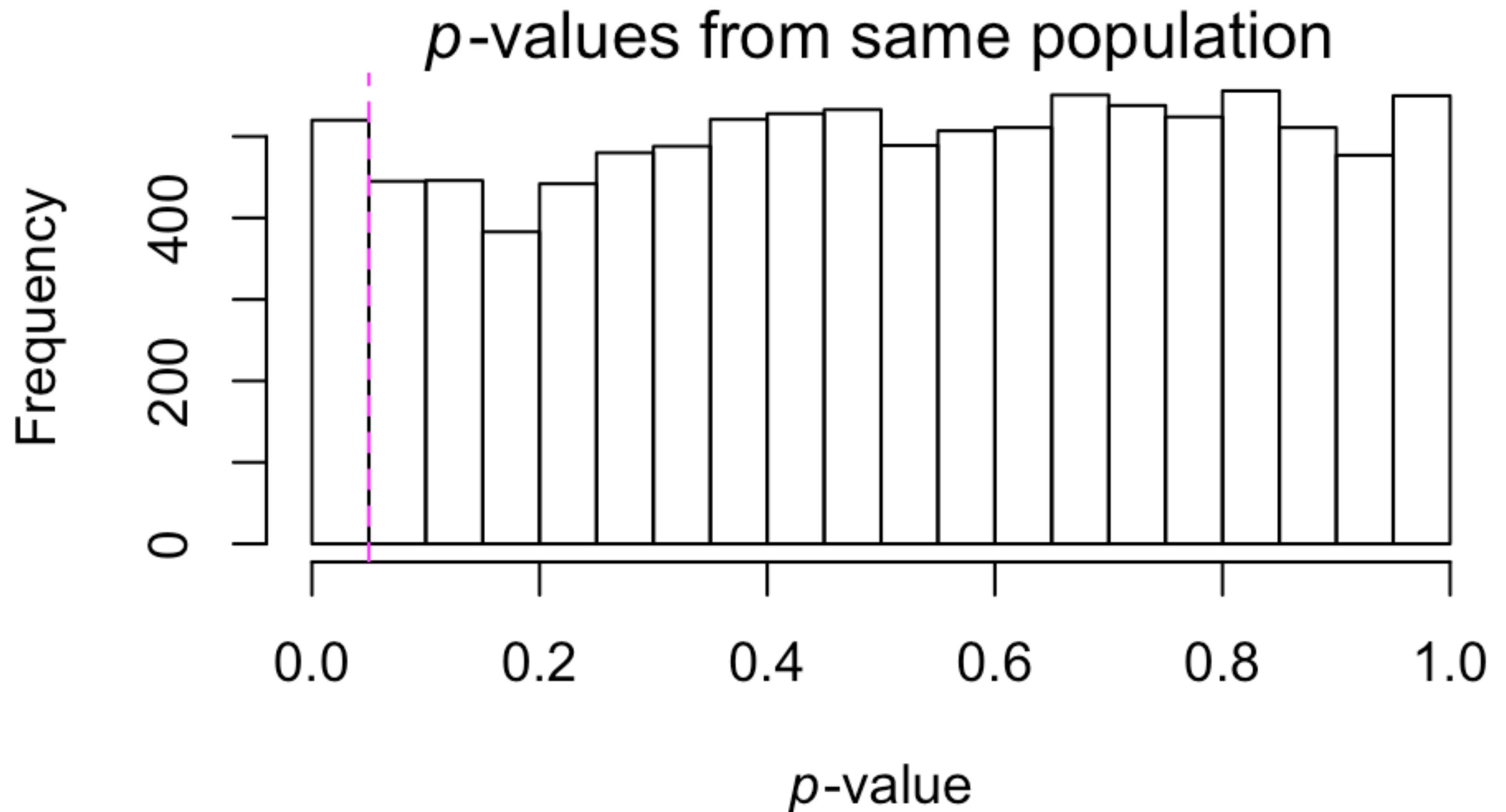
- 標本に基づいた統計量が  
帰無仮説の下、起きうる確率
- 汎用される危険率（閾値）：

**0.05 = 100回に5回起きる**

同一平均値集団間のt検定の繰り返しを実験してみましょう

→ nibb-unix/gitc202309-unix/ex3 復習問題3 統計学入門 2.

# 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

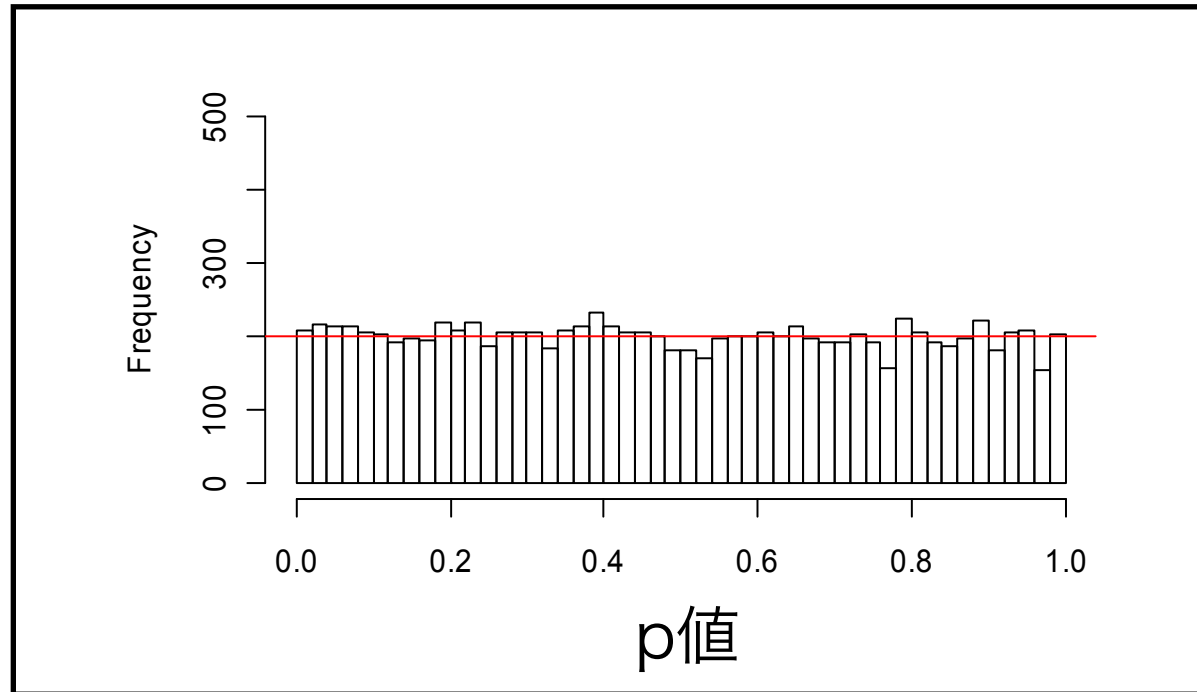
危険率を検定数で調整

$$\text{危険率} = \alpha / k$$

$\alpha$ : 元の危険率、

$k$ : 検定数

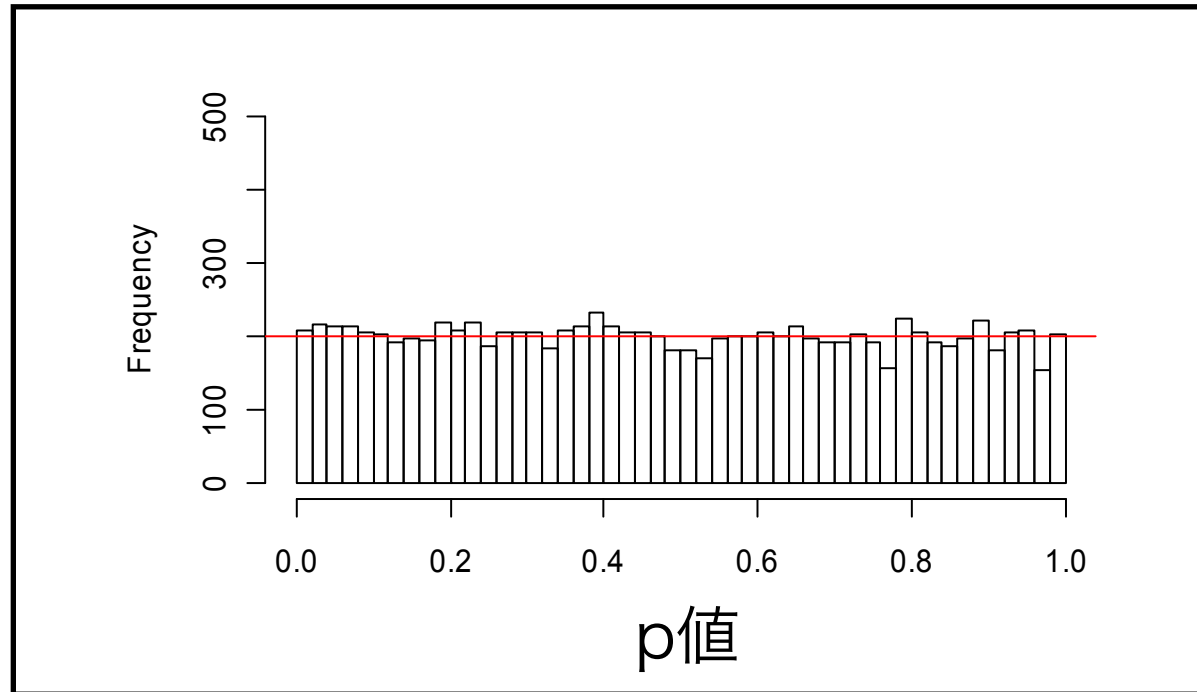
# False Discovery Rate (FDR)



帰無仮説

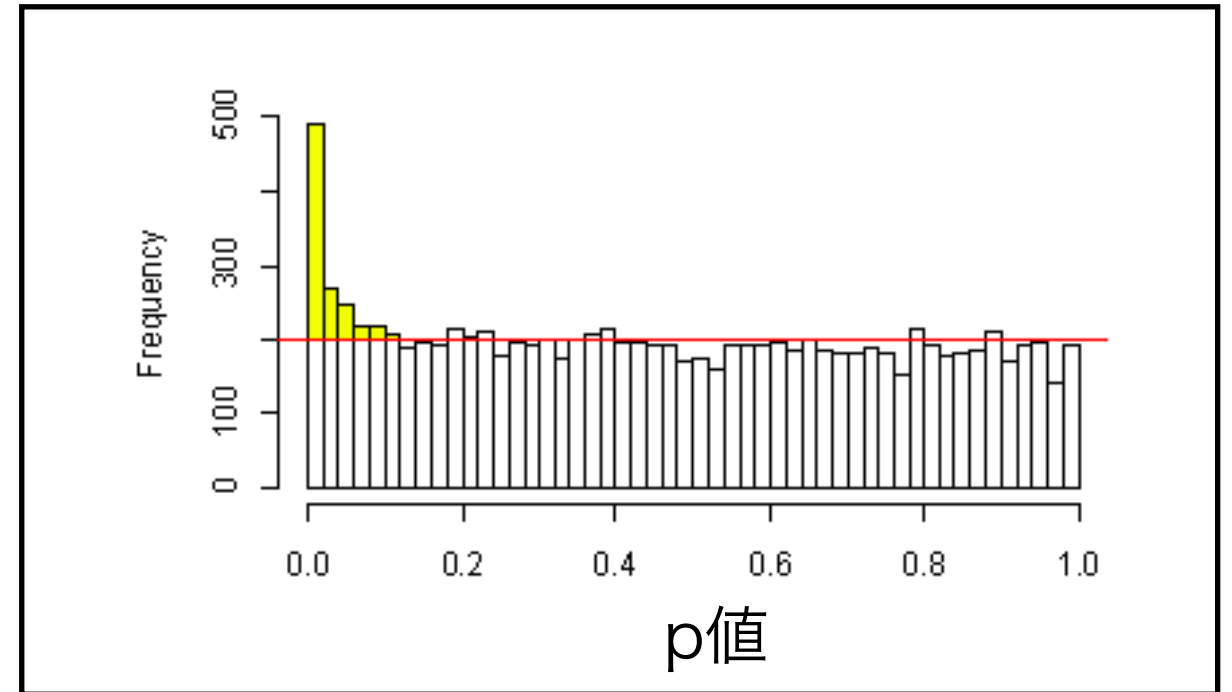
全ての範囲のp値が  
同等の頻度で観察される  
←どのp値を選んでも  
ランダムに選ぶのと同じ

# False Discovery Rate (FDR)



帰無仮説

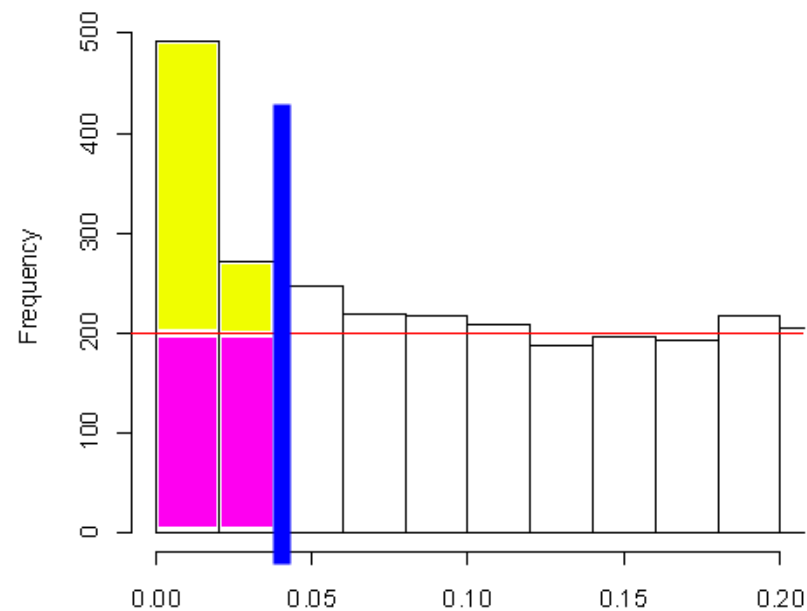
全ての範囲のp値が  
同等の頻度で観察される  
←どのp値を選んでも  
ランダムに選ぶのと同じ



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。

# $p$ 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**

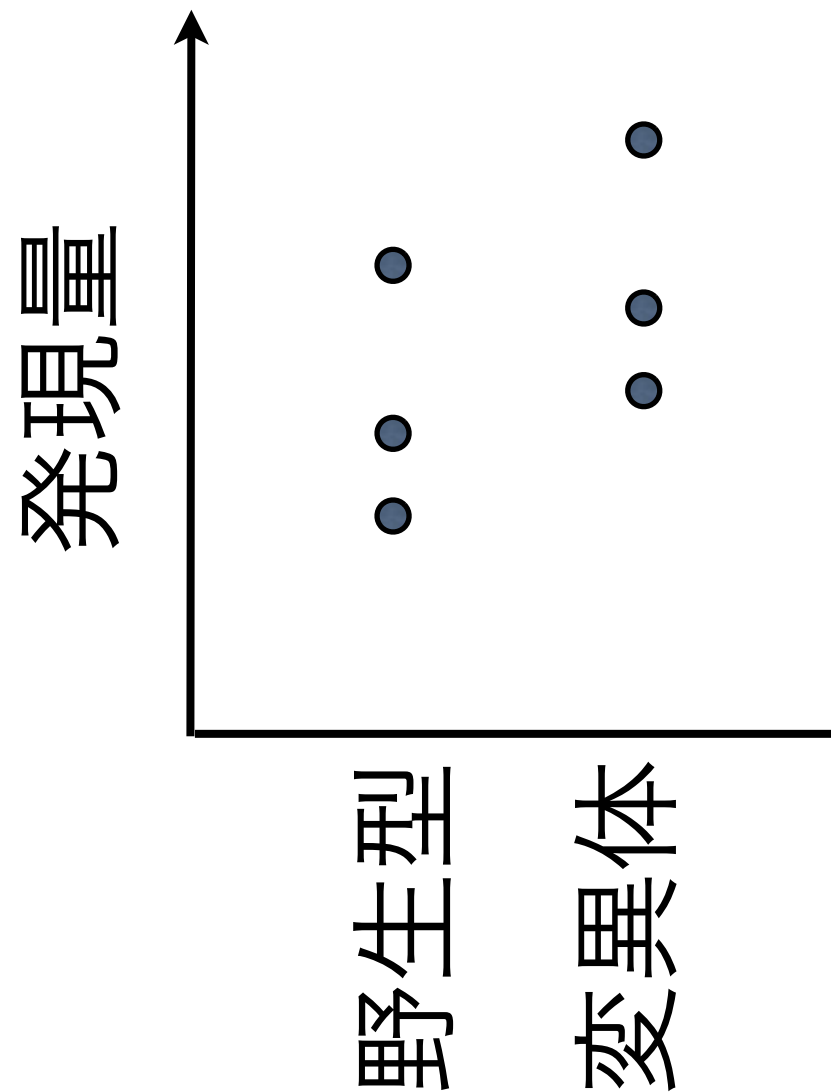
**True negative**

# 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

# 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



# 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**

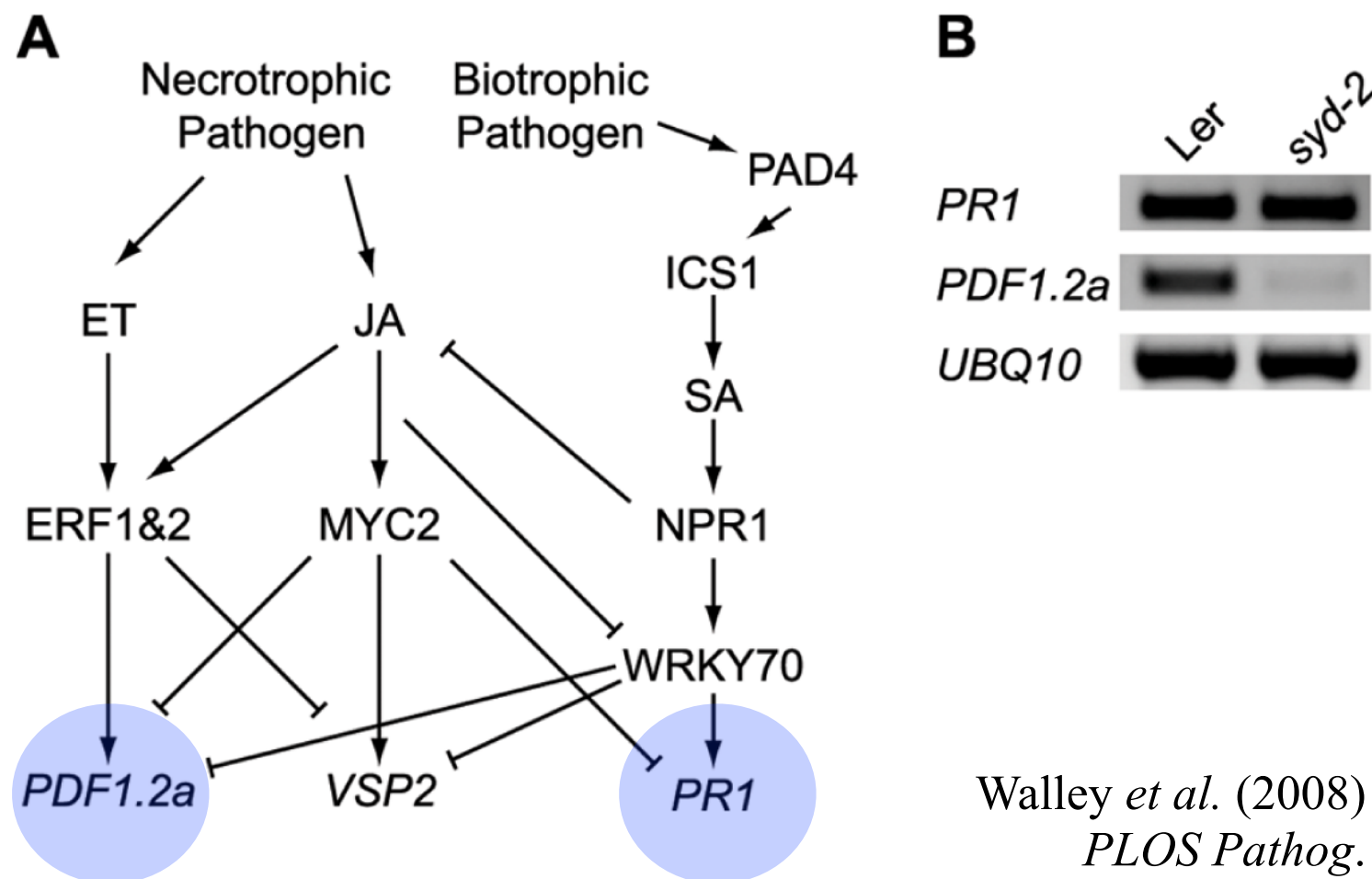
同一サンプルを繰り返し測る:  
**technical replicates**

# 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？

# 非NGS測定：“マーカー遺伝子”測定

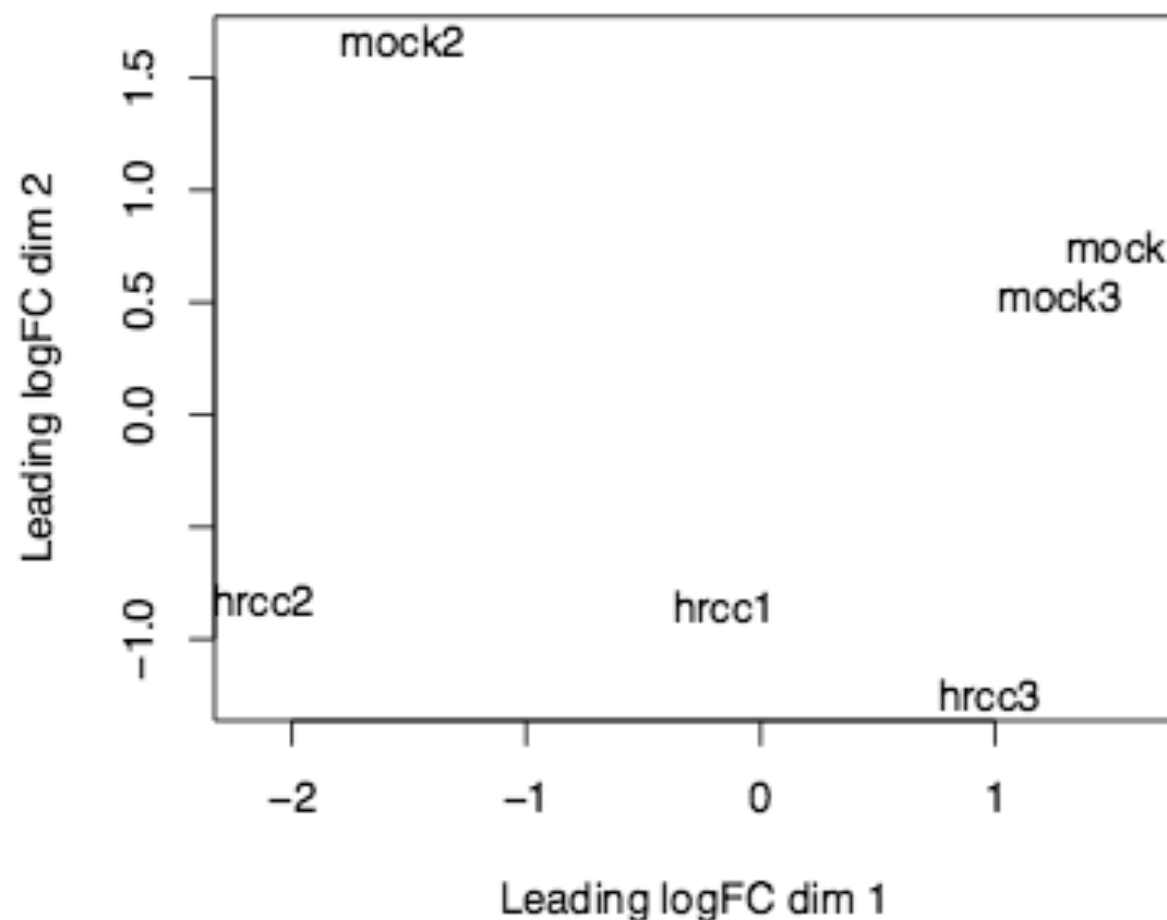
- 何が再現されうるか？再現されたとするか？



明瞭な違いを  
示す遺伝子:  
明瞭な再現性

# “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

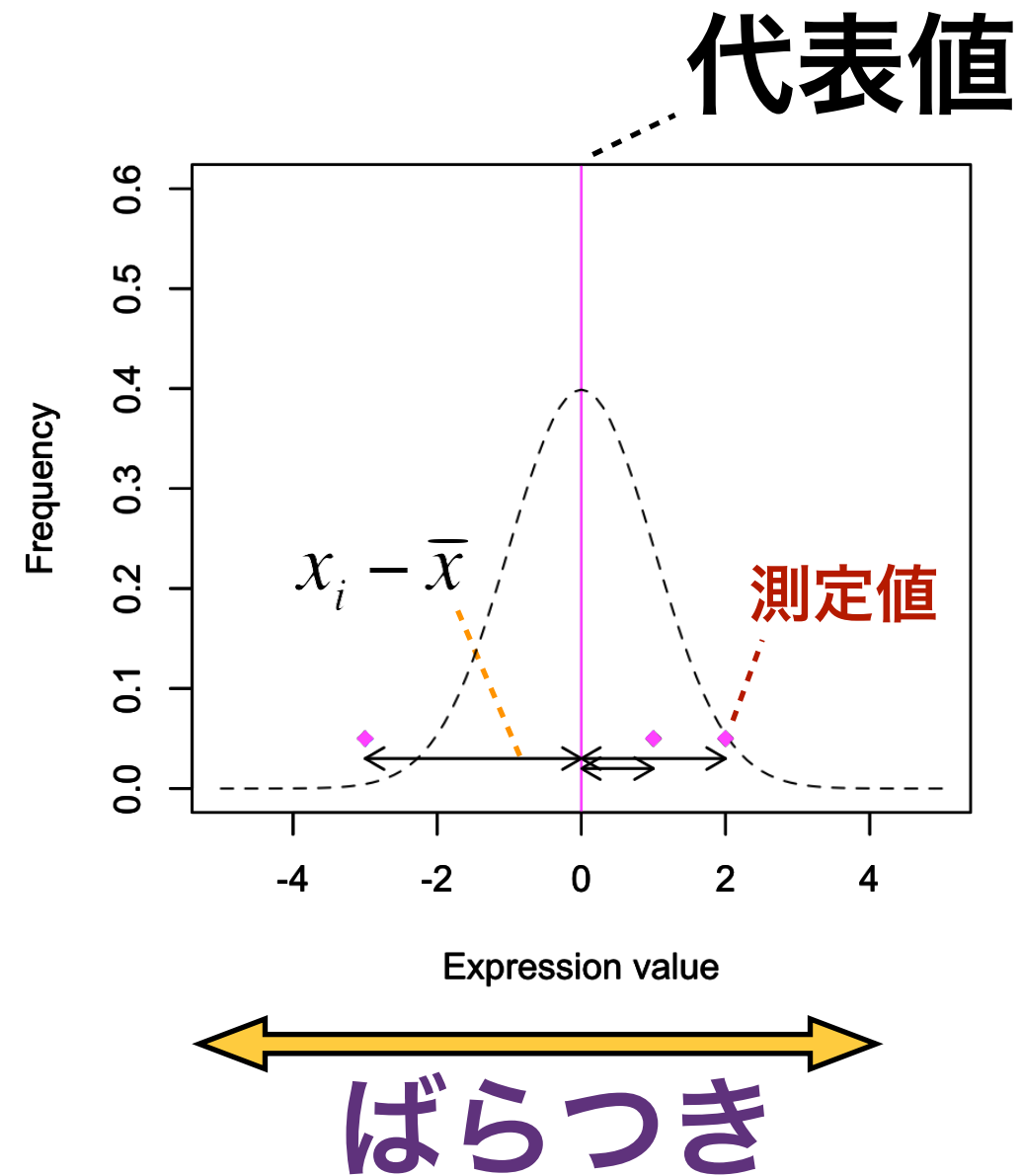
# 定量的測定が可能且つ要求される時代の 再現性のあるデータとは何か？

- 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？~~

バラツキの  
定量と  
説明変数への  
割当て

ここまでの統計量はサンプルという  
一要因のみを考慮



# 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

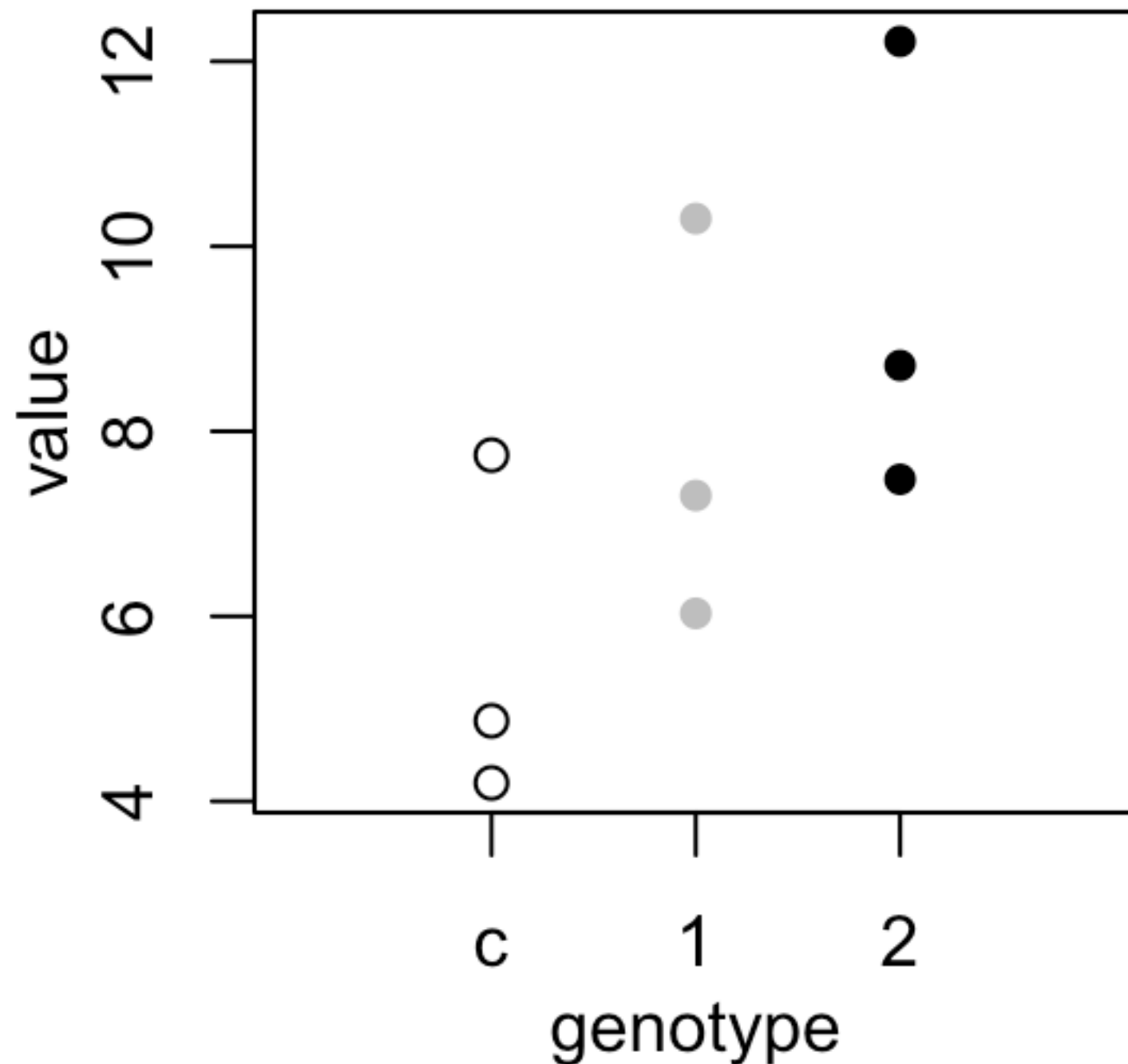
次元圧縮



# 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をつかむ

# あるRT-qPCR実験: 生データ

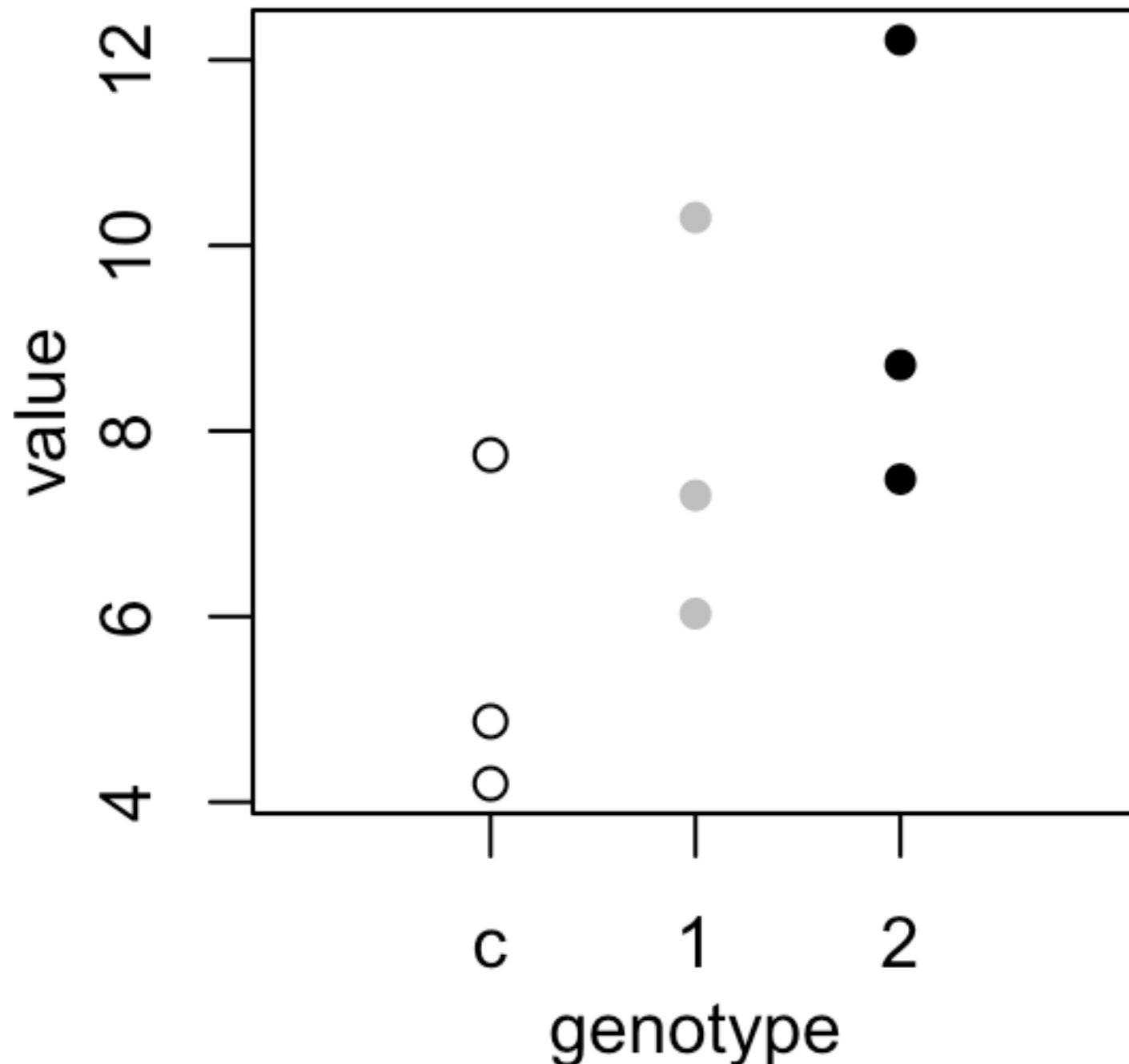


**genotype**

- control
- strain1
- strain2

**replicate: 1, 2, 3**

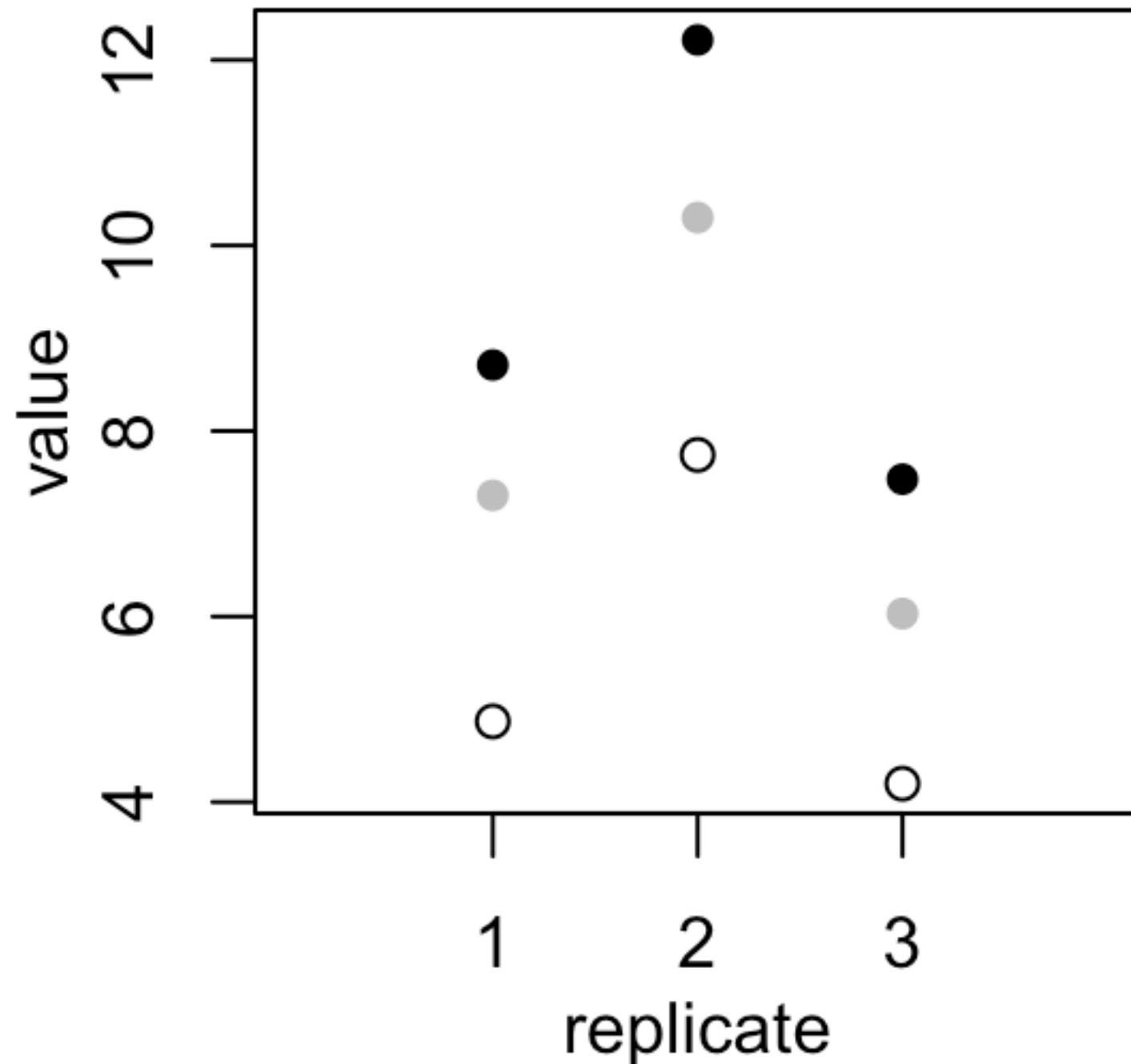
## あるRT-qPCR実験: $t$ 検定結果



### p-values

- control vs strain1  
= 0.2456
- control vs strain2  
= 0.1011

# 生データをreplicateについて可視化



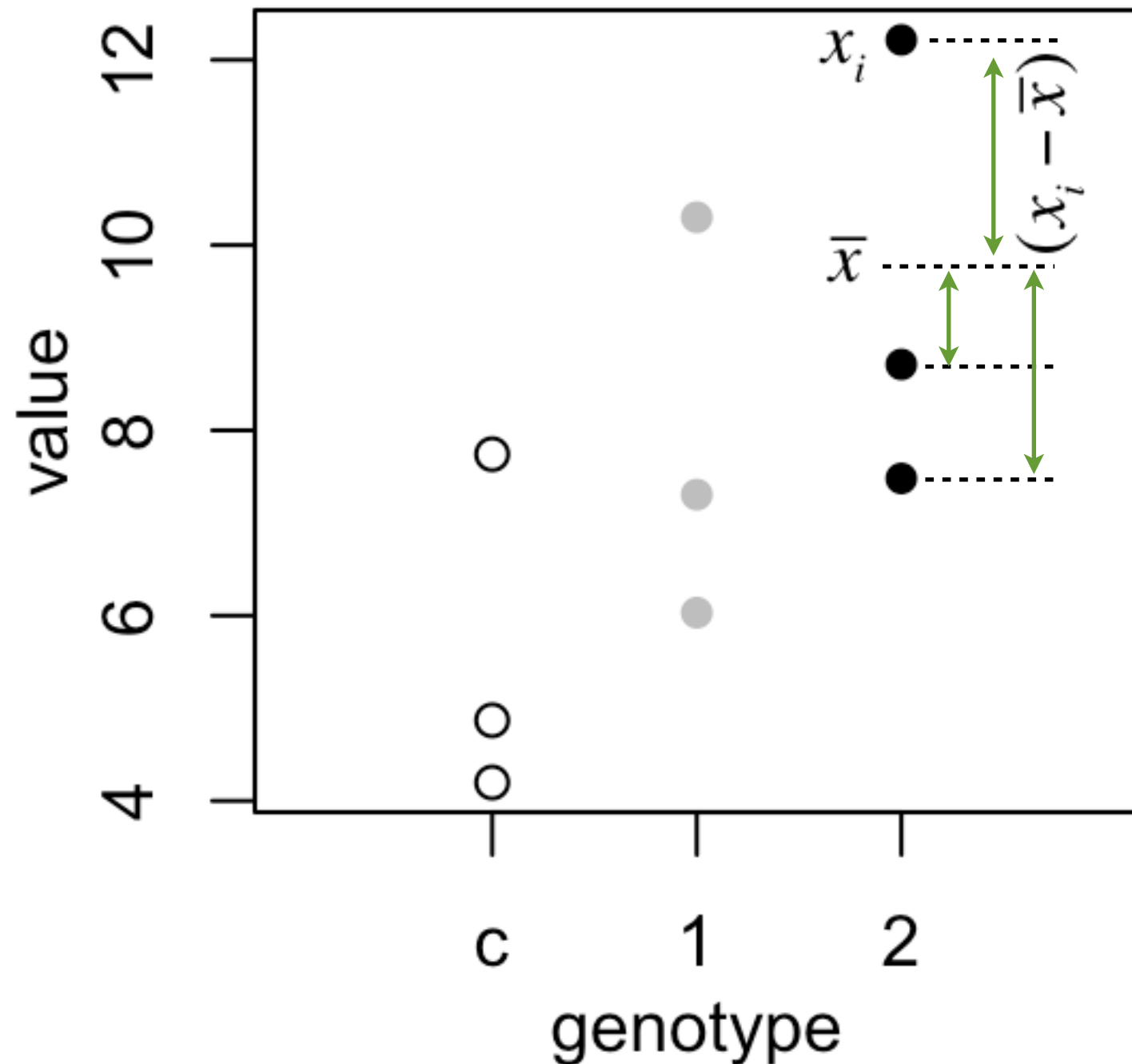
**genotype**

- control
- strain1
- strain2

**replicate: 1, 2, 3**

**検定から推定（予測・モデル構築）へ：  
線形モデルへの転換**

# 線形モデルで考えてみる：モデル式



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

# 線形モデルで考えてみる：モデル式

$$x_1 = \bar{x} + (x_1 - \bar{x})$$

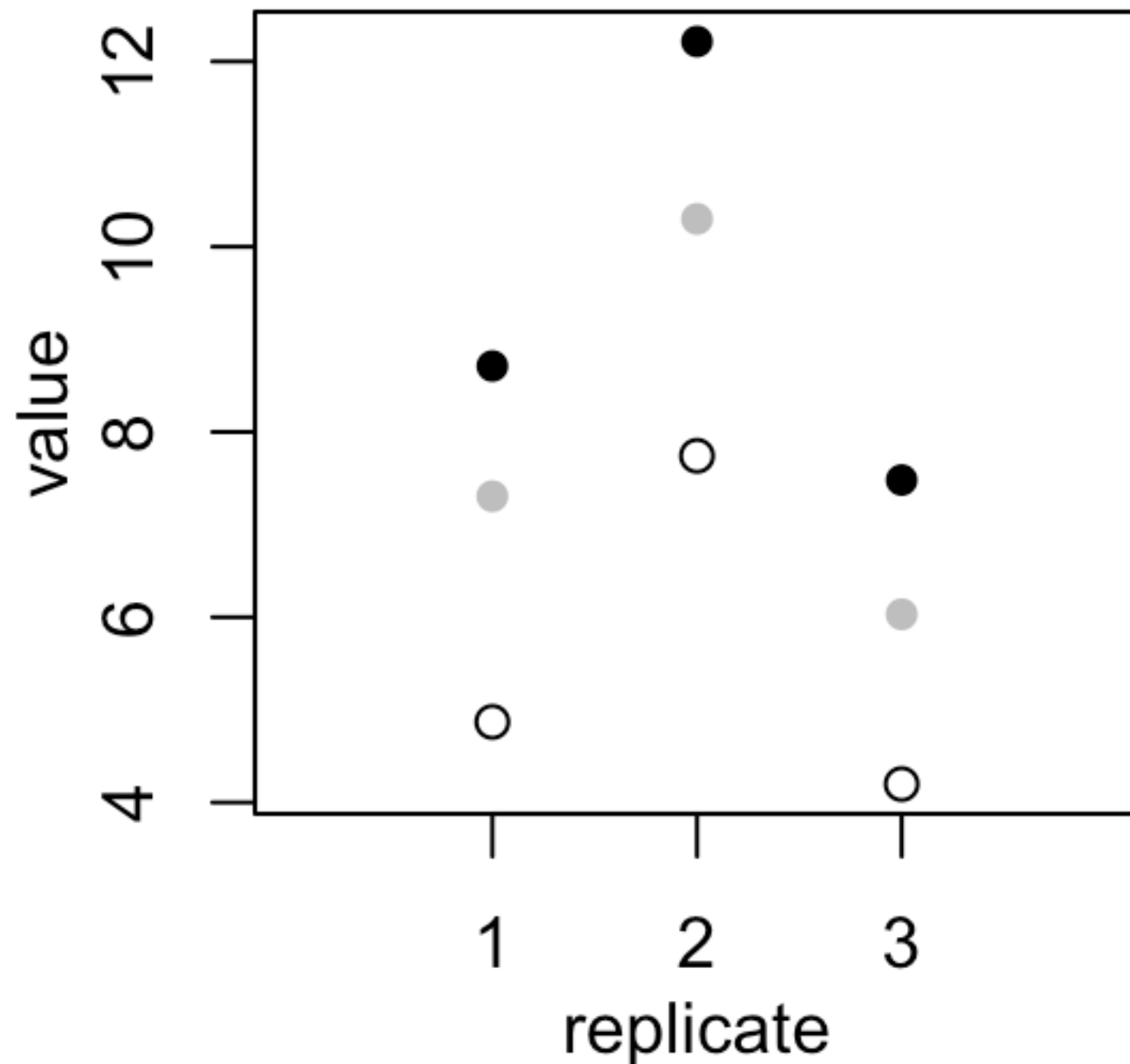
$$x_1 = \bar{x} + \underline{\varepsilon_1}$$

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

# Replicateによる影響を考慮すべき



**genotype**

- control
- strain1
- strain2

**replicate: 1, 2, 3**



# 観察値を複数要因の 影響に起因するものとして分解

$$x_1 = \bar{x} + (x_1 - \bar{x})$$

$$x_1 = \bar{x} + \varepsilon_1$$

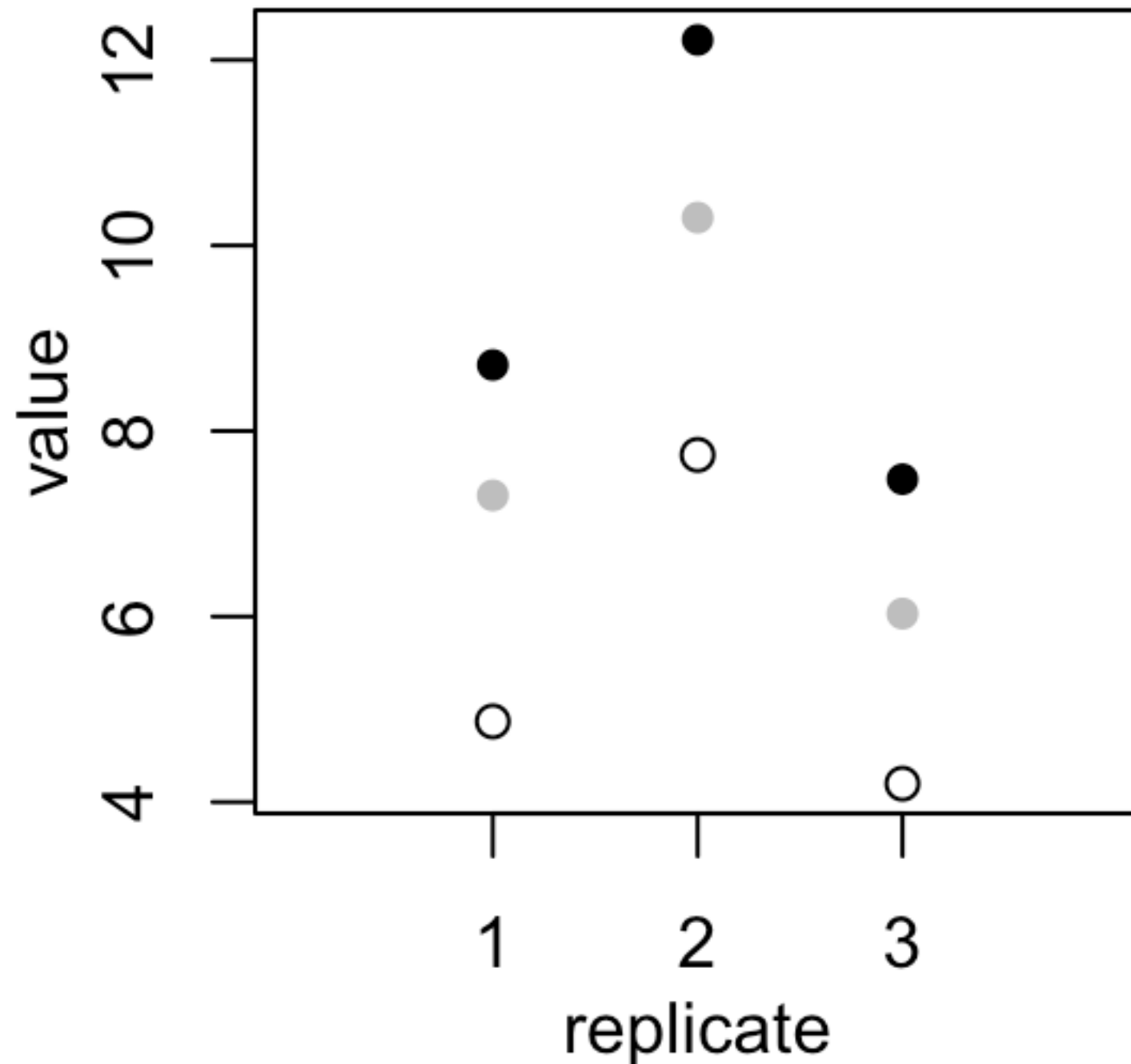


$$O_{ij} = x_i + y_j + \varepsilon_{ij}$$

*genotype*と*replicate*の

影響を同時に  
考えられないか？

# 生データをreplicateについて可視化

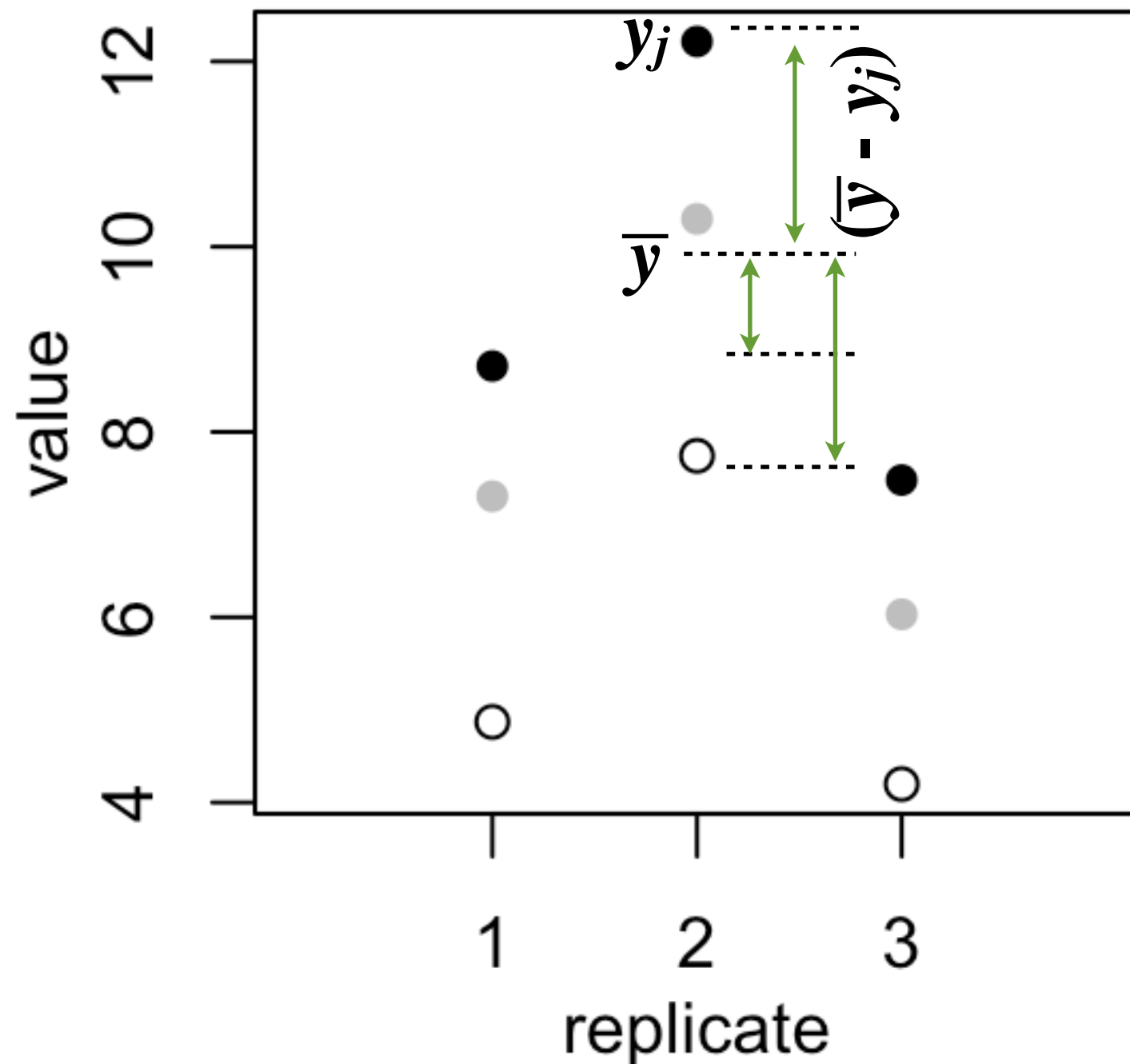


**genotype**

- control
- strain1
- strain2

**replicate: 1, 2, 3**

# replicateの影響も推定



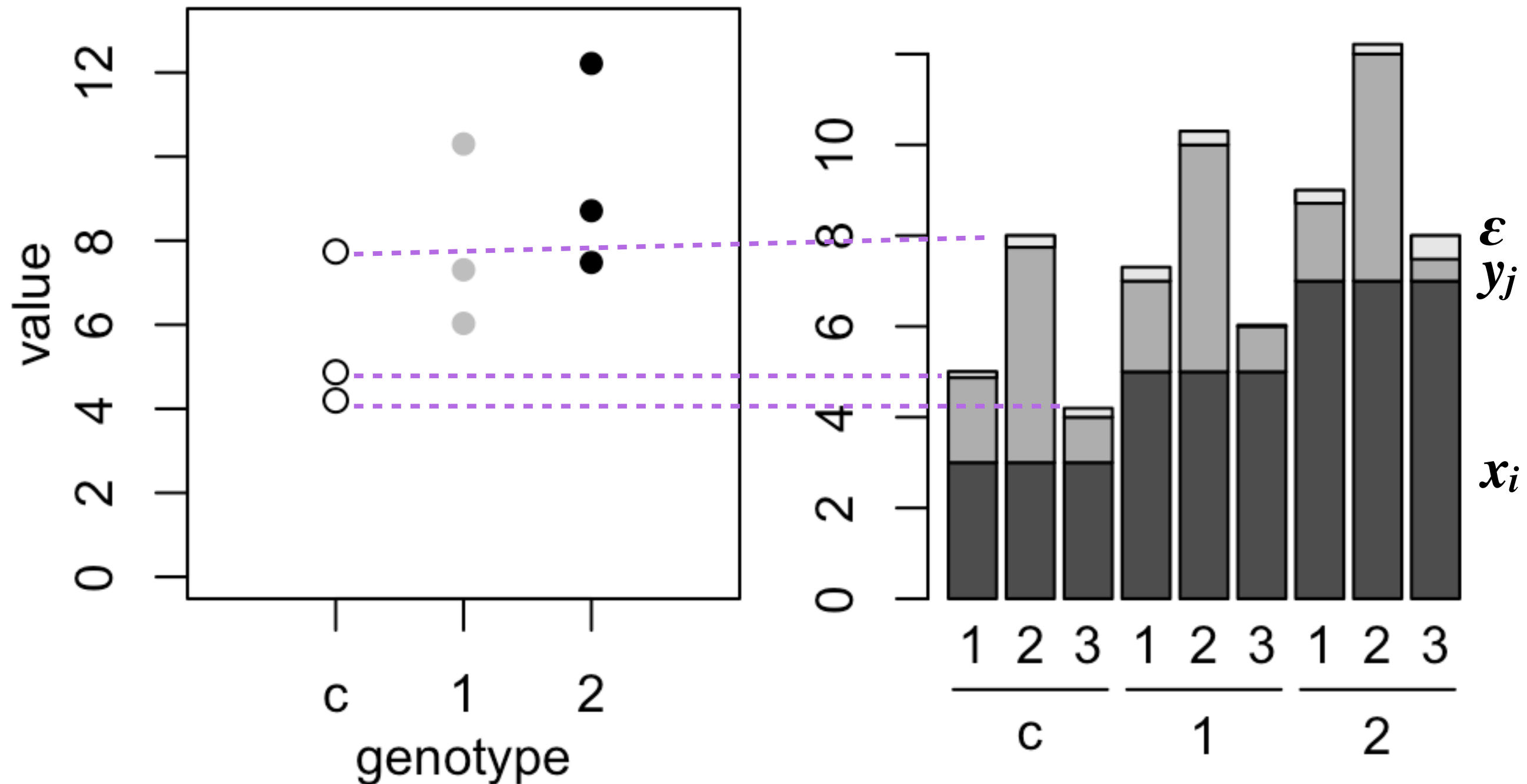
## genotype

- control
- strain1
- strain2

**replicate: 1, 2, 3**

genotype, replicateの影響を

同時に推定する:  $O_{ij} = x_i + y_j + \varepsilon_{ij}$



# 分散分析・線形モデルの枠組み

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

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  $\sim$  説明変数1 + 説明変数2 + .... + 残差

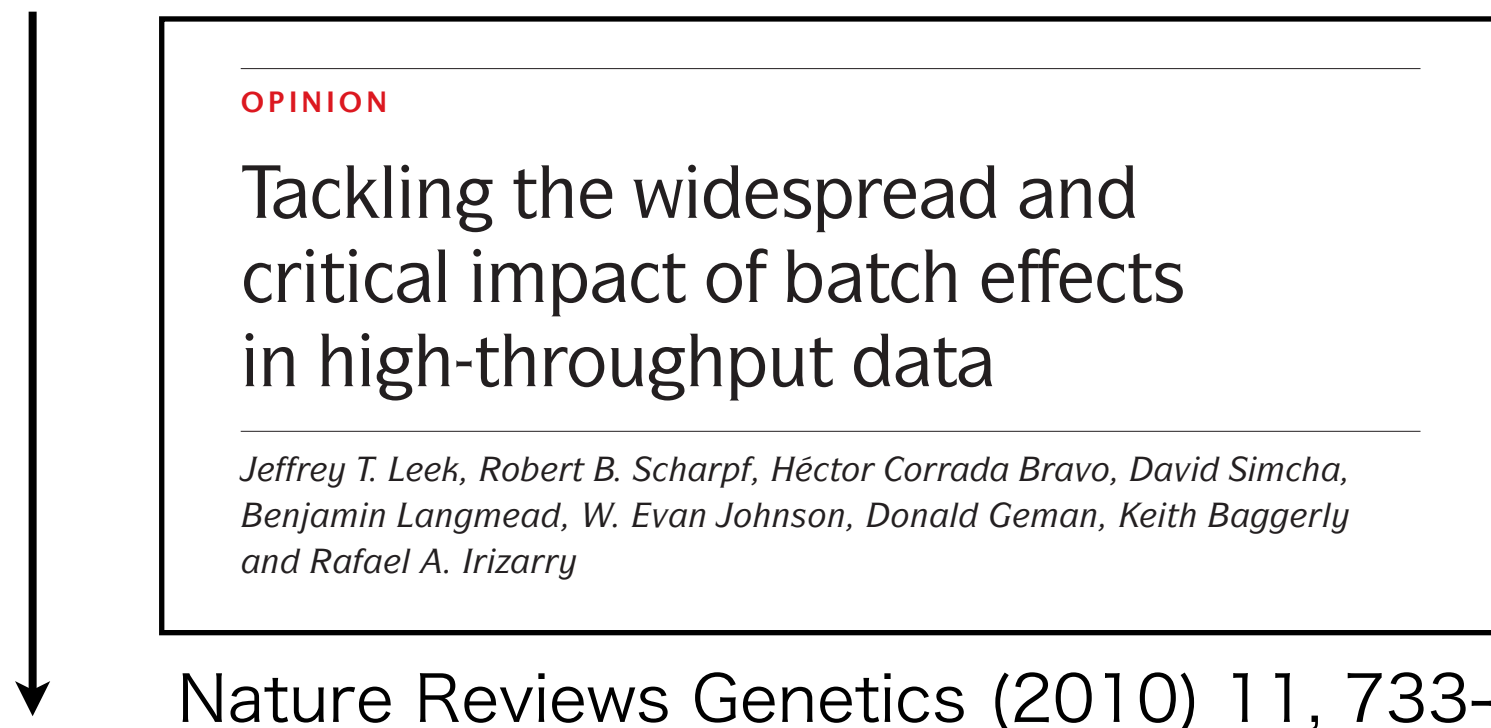
と観察値を説明する（かもしれない）  
**変数の足し算**で応答変数への貢献度を  
推定する

- R: lm, glm, glmFitなどの関数を使う  
→ 内山 R入門「モデル式と線型モデル」

# 線形モデル・実験デザインの重要性

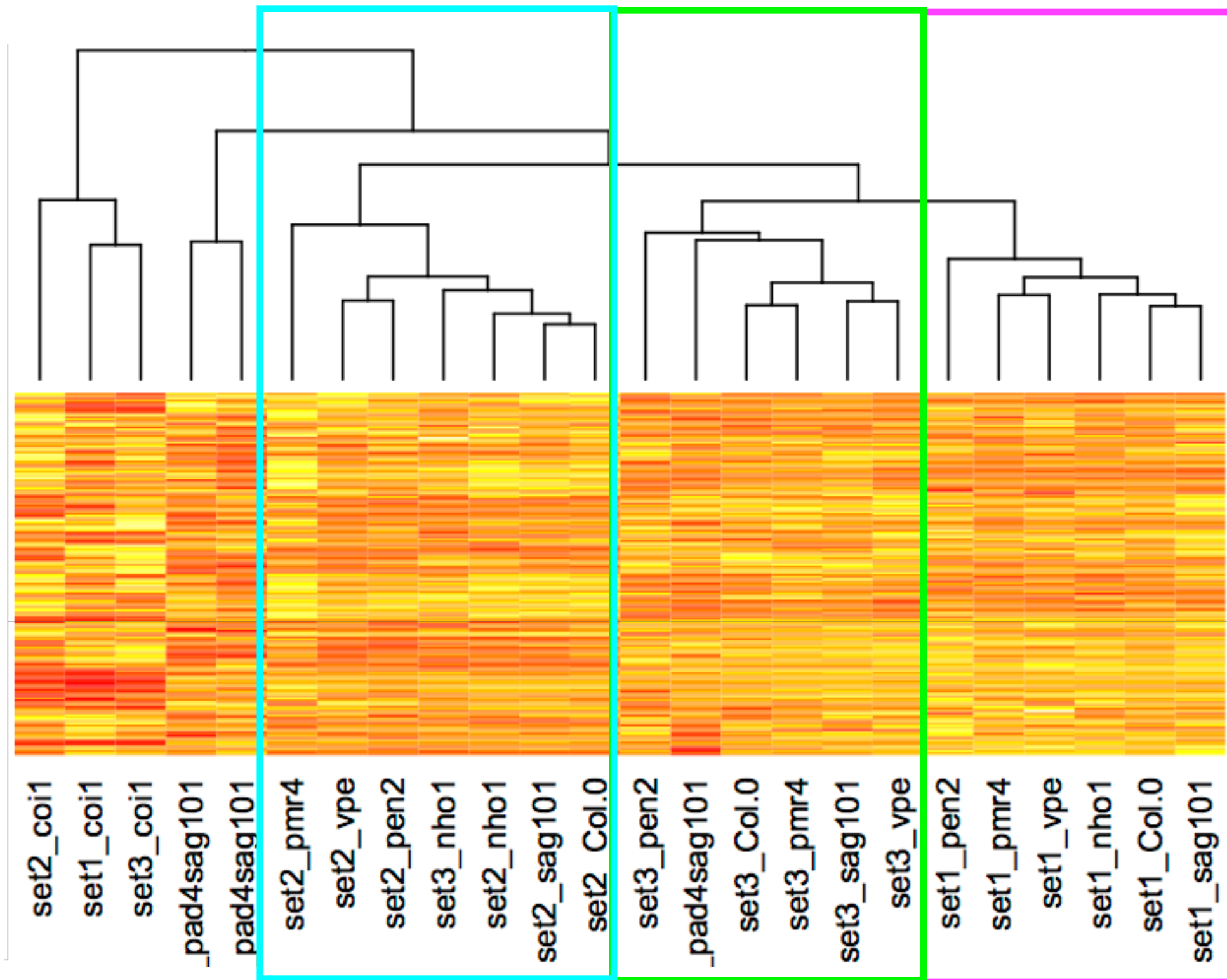
- -omicsデータは“**batch effect**”と呼ばれる体系的なバイアスが混入する。

例: 実験時期、実験者、餌



- 線形モデルで推定・除去

# batch effect の トランスクリプトームへの影響





# 線形モデル・実験デザインの重要性

- 線形モデルで推定・除去

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

$\alpha_i$  : 遺伝子型 / 処理など注目している効果の要因

$\beta_j$  : 反復（実験日時） / 実験者などバイアス要因

- $\alpha_i$  の推定値、標準誤差のみを使う

# 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？

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

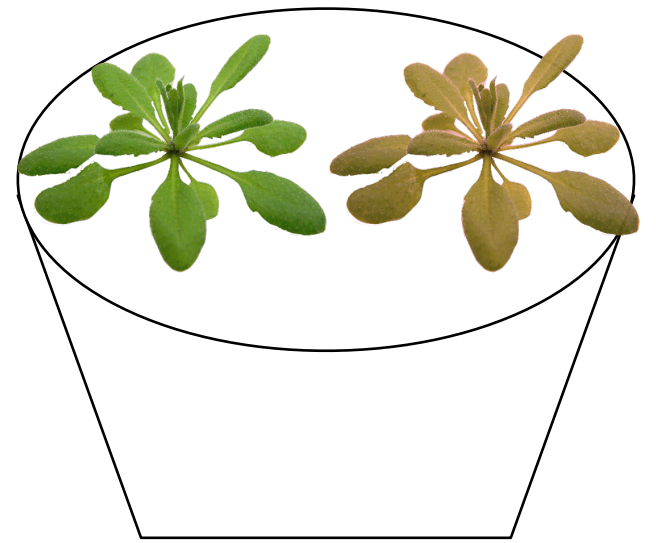
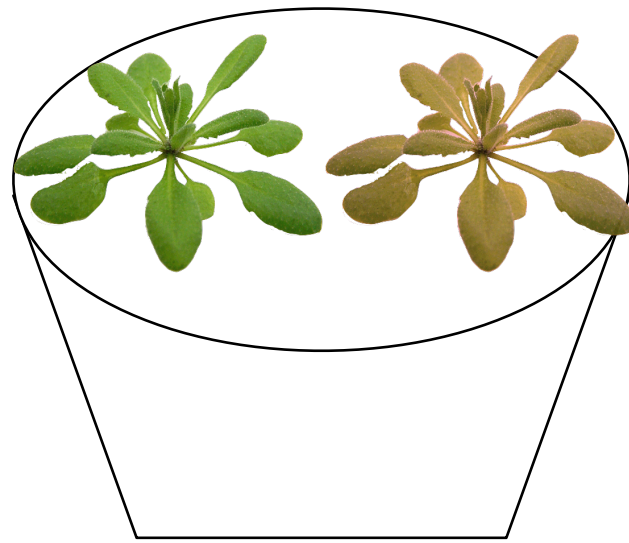
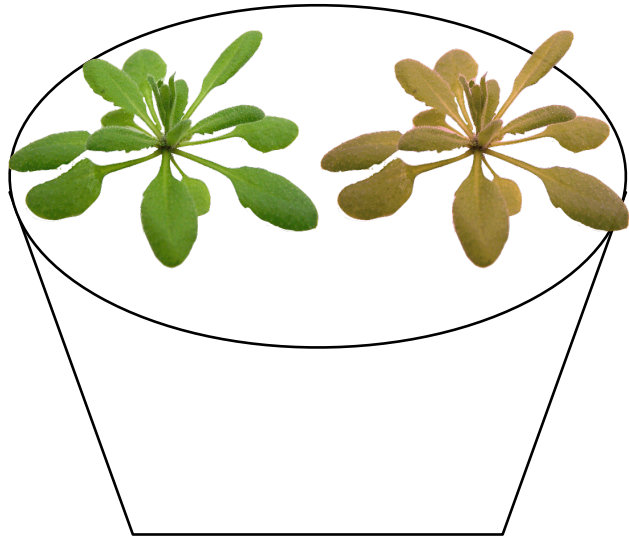
体系的な  
バラつきを  
説明変数として  
割当て

- 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？

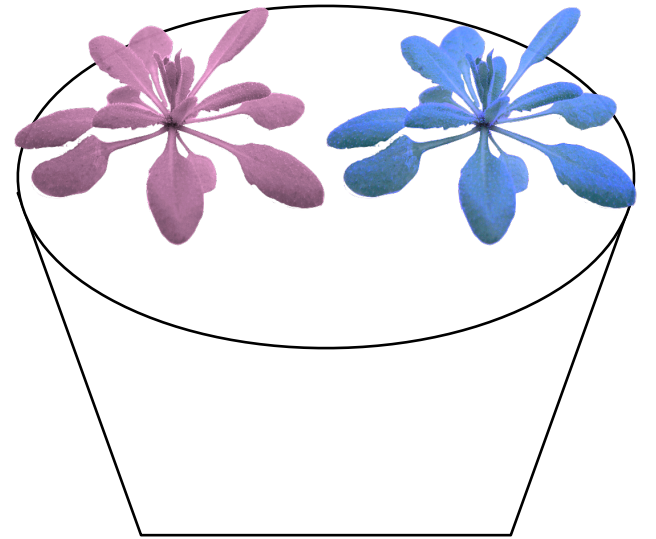
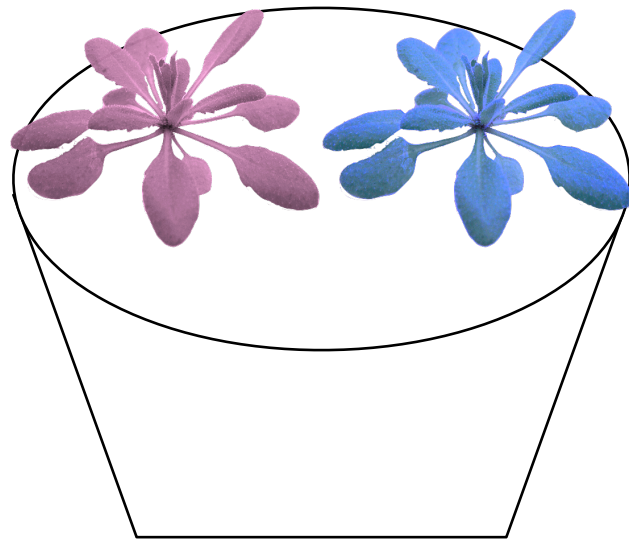
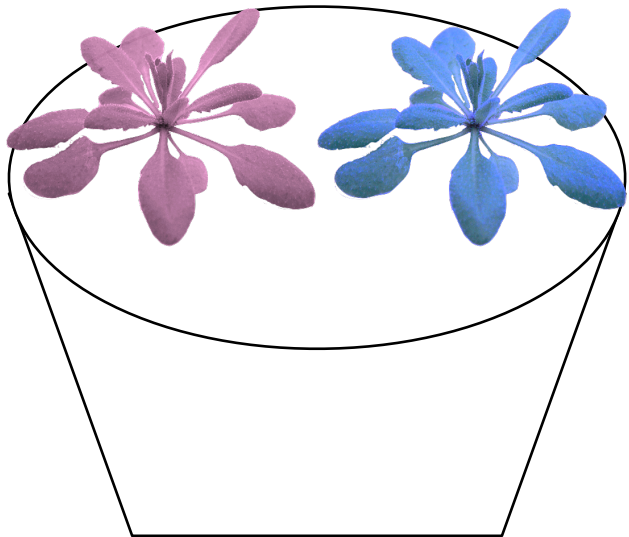
実験デザインの重要性:

**genotype+replicate+pot**モデルを当てはめるには？

pot 1



pot 2



replicate 1

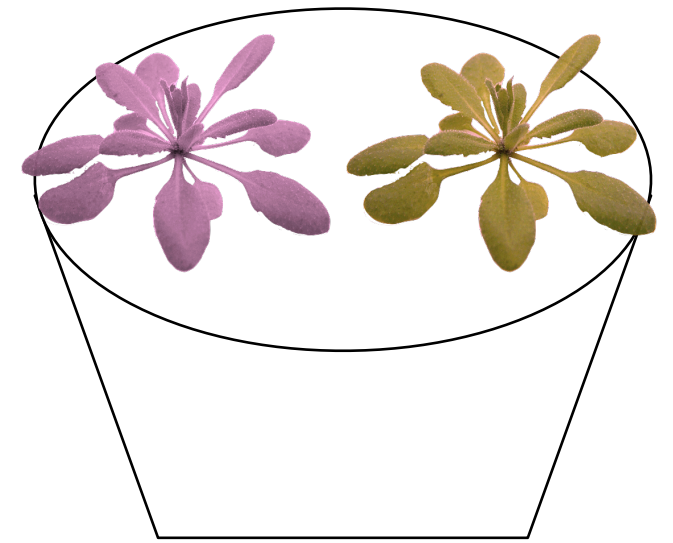
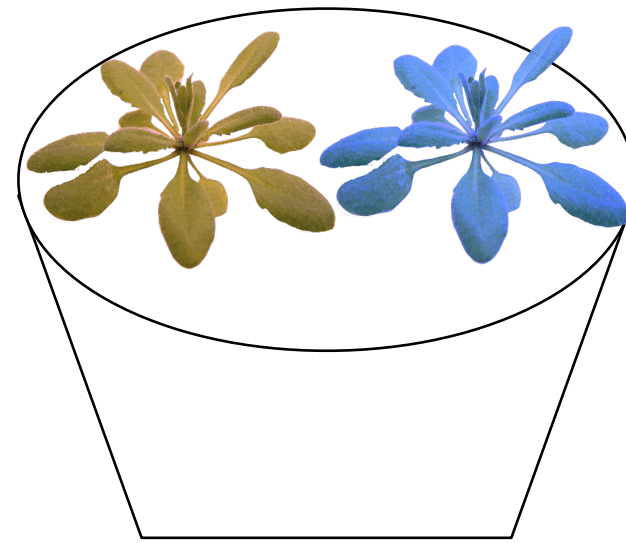
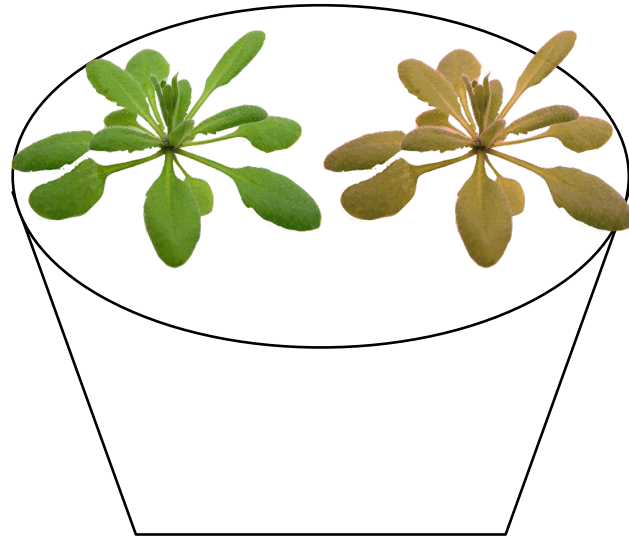
replicate 2

replicate 3

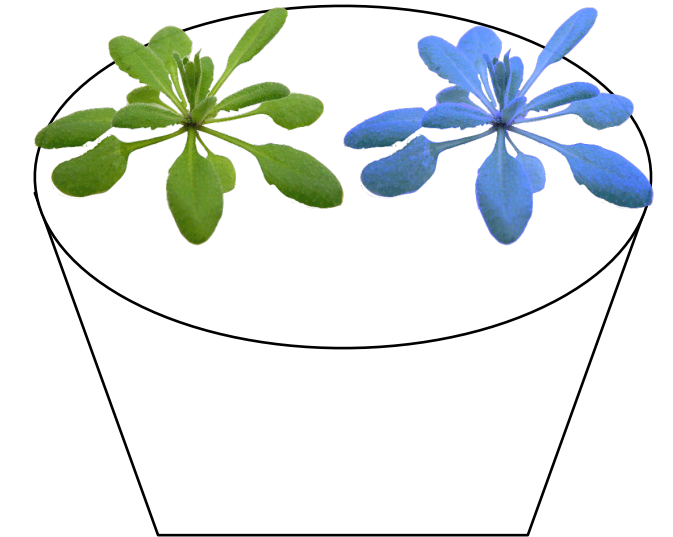
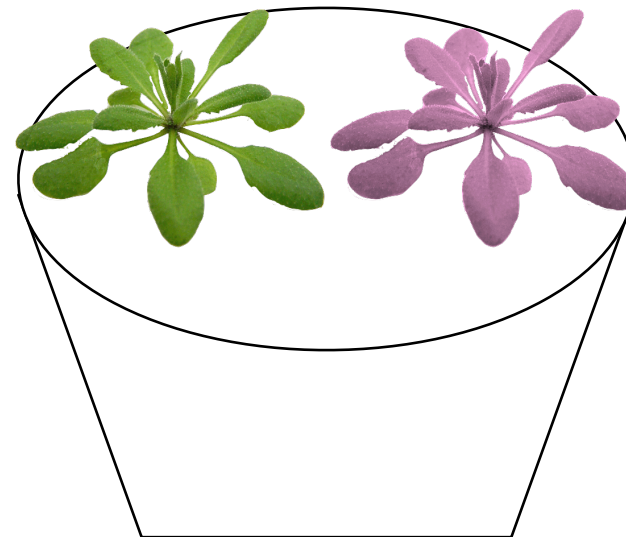
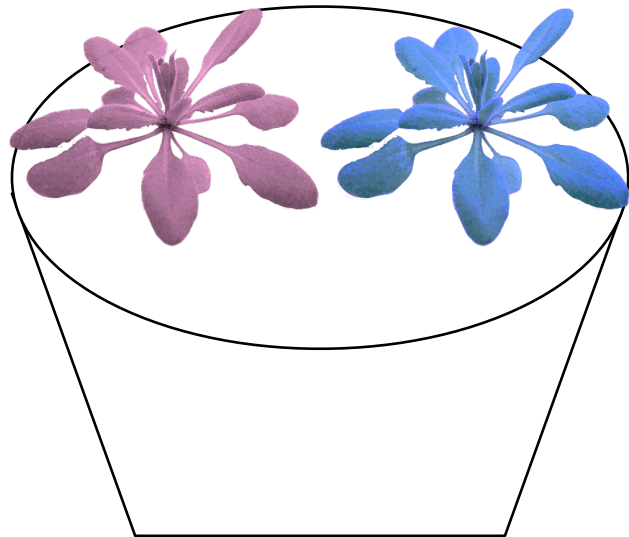
実験デザインの重要性:

**genotype+replicate+pot**モデルを当てはめるには？

pot 1



pot 2



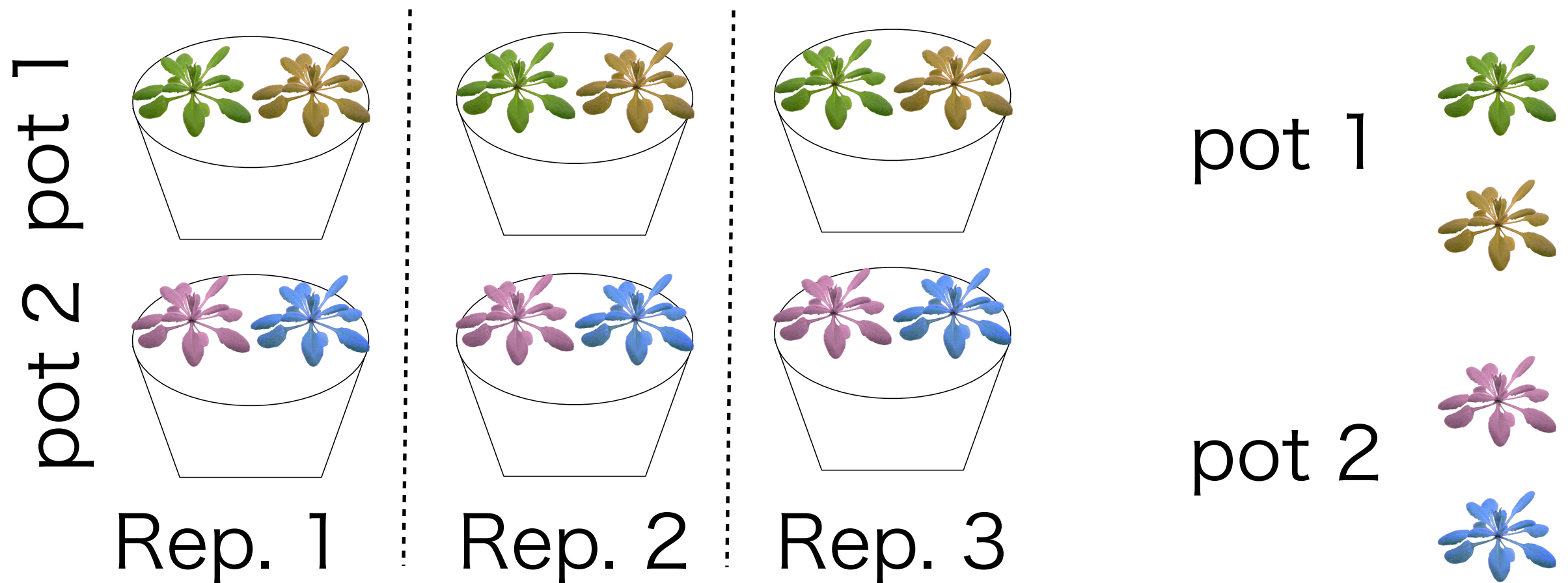
replicate 1

replicate 2

replicate 3

線形モデル・実験デザインの重要性  
→ 無作為化 (randomization)

✖ 無作為化されていない実験デザイン



# 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  
(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。

# (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：残差が正規分布 [R:lm]

離散値（カウントデータ）：

**負の二項分布** [R:glmFit, glm.nb]



# 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）：切片、**contrast**  
➡ R入門 モデル式と線型モデル
- 実験計画法
- 交互作用 ex.replicate特異的なgenotypeの影響
- Bioconductor: limma、edgeRマニュアル