

情報統計 第13-15回

2019年9月23日 神奈川工科大学



櫻井 望

国立遺伝学研究所
生命情報・DDBJセンター

スケジュール

	17日（火） データの見え る化	18日（水） 検定のこれだけ は	19日（木） 多変量解析の雰 囲気	23日（月） データ準備 発表会
1限	1 ガイダンス、 PC環境準備、 データの見え る化	5 区間推定、 分布とその使い 方	9 相関	13 自習（課題、 質問）
2限	2 統計の基本 と用語	6 t検定	10 主成分分析	14 自習（課題、 質問）
3限	3 プログラミ ングの基礎	7 検定で注意 すること	11 他の多変量 解析	15 発表会
4限	4 自習（課題 検討、復習）	8 自習（課題 検討、復習）	12 自習（課題 検討、復習）	

補足

- 等分散性の検定（F検定）
- 分散分析（F分布を使う）
- ログ変換
- 主成分分析の例

F検定

等分散性の検定

1群目：標本数 n_1 , 不変標本分散 v^2_1

2群目：標本数 n_2 , 不変標本分散 v^2_2

検定統計量：
$$F = \frac{v^2_a}{v^2_b}$$

※ v^2_a , v^2_b は、 v^2_1 , v^2_2 のいずれか、分散の大きい方を分子にする。数値は1以上になる

自由度： $n_1 - 1$, $n_2 - 1$

※分子と分母に対応させて、二つ与える

帰無仮説：2群の分散は等しい

F分布を扱うExcel関数：F.DIST, F.DIST.RTなど

例) 身長データの場合

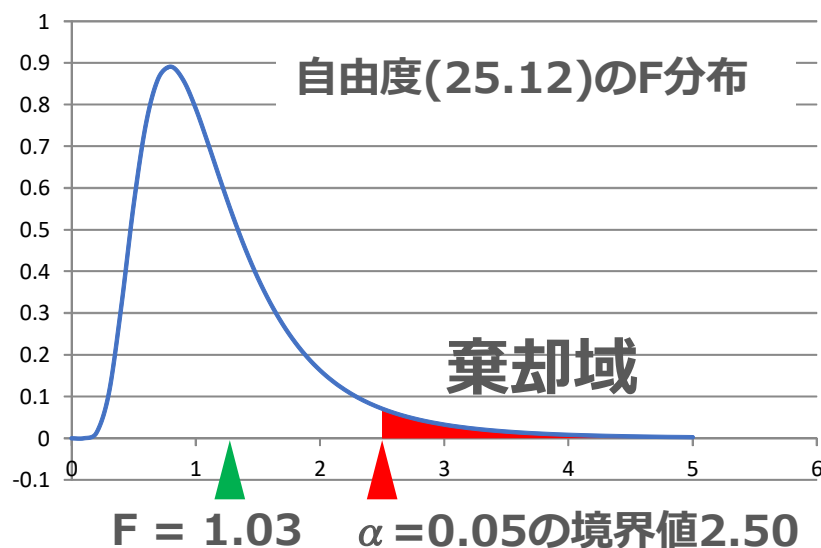
女性 : $n_1 = 26, v_1^2 = 23.63$

男性 : $n_2 = 13, v_2^2 = 23.02$

有意水準 : 0.05とする

$$F = 23.63 \text{ (女性)} / 23.02 \text{ (男性)} = 1.03$$

自由度(25, 12)のF分布から、F.DIST.RT関数を使って求めた右側確率pは、0.50



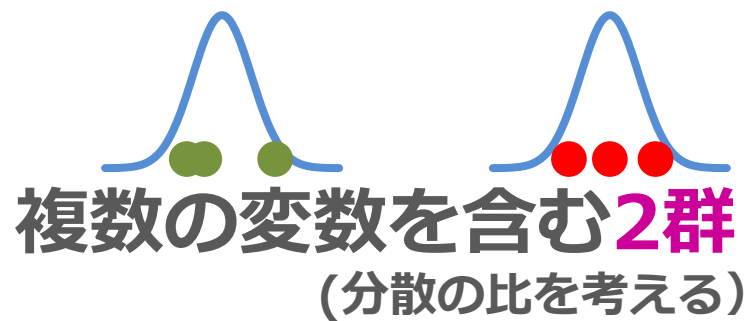
F値が棄却域の境界値より内側
($1.03 < 2.50, p=0.50 > \alpha$)
なので、帰無仮説は棄却できず、
「2群の分散に差があるとは言えない」と結論づけられた。

留意すべきこと

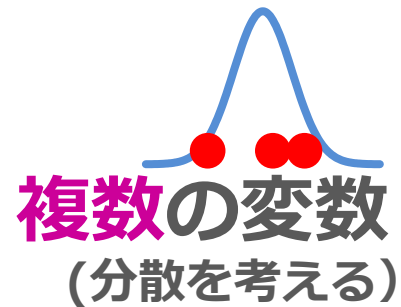
F検定で「分散に差がある」という結論を得たのち、2群の平均値に差があるかどうかをt検定すると、「**検定の多重性**」の問題にあたってしまう。

近年では、等分散かどうかに関係なく適用できるウェルチの検定を最初から行うことが望ましいという考えも出てきている。

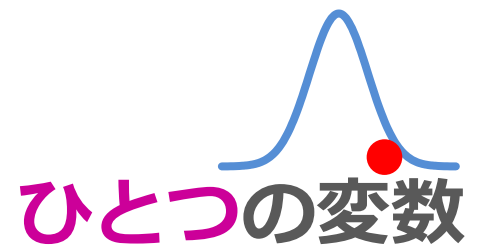
F分布



カイ二乗分布

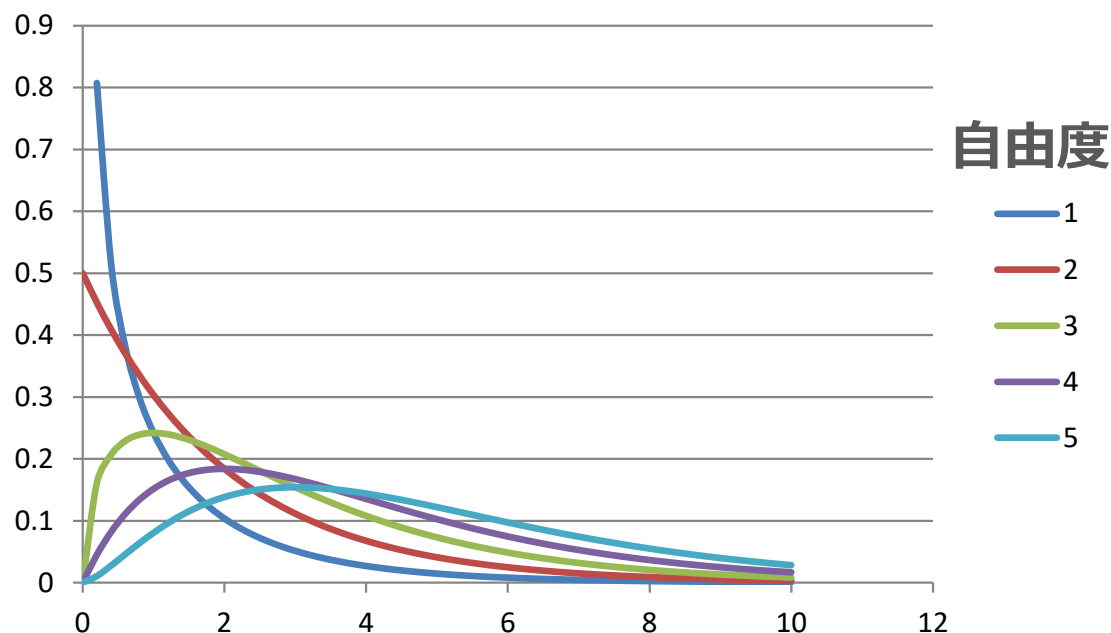
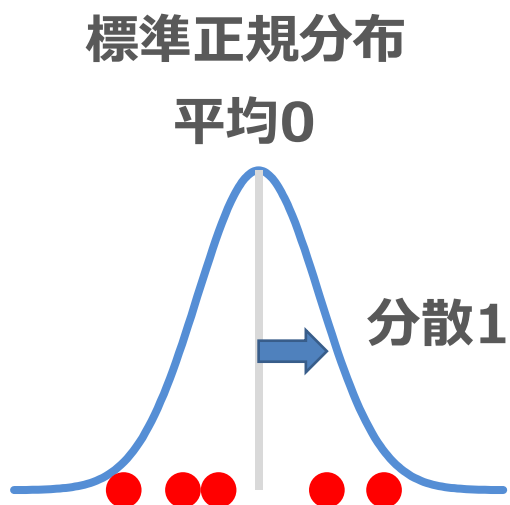


標準正規分布



カイ二乗分布

標準正規分布に従った**独立した**変数がいくつ
つかあるとき、その**二乗和**が従う分布



カイ二乗分布の性質

正規分布 $N(\mu, \sigma^2)$ に従った k 個の変数 x_i について、
偏差（平均からの差）の平方和と分散の比は、自由度 k のカイ二乗分布に従う

$$\chi^2 = \frac{\sum_{i=1}^k (x_i - \mu)^2}{\sigma^2} = \sum_{i=1}^k \left(\frac{x_i - \mu}{\sigma} \right)^2$$

カイ二乗検定

	ビール 好き	ビール あんまり
男性	23	12
女性	7	8

二つのカテゴリに関連があるかを調べたい

帰無仮説：

二つのカテゴリは独立である（関連がない）

有意水準：0.05

カイ二乗検定の手順

(1) 観測データから、カテゴリーごとに割合を出す

	ビール好き	ビールあんまり	合計
男性	69	36	105 70%
女性	21	24	45 30%
合計	90 60%	60 40%	150 100%

(2) 割合から、カテゴリーが独立な場合の度数（期待度数）を出す

	ビール好き	ビールあんまり	合計
男性	63	42	105 70%
女性	27	18	45 30%
合計	90 60%	60 40%	150 100%

カイ二乗検定の手順

(3) 観測度数と期待度数の差を出す

	ビール好き	ビールあんまり
男性	6	-6
女性	-6	6

(4) その二乗を出す

	ビール好き	ビールあんまり
男性	36	36
女性	36	36

(5) 期待度数で割る

	ビール好き	ビールあんまり
男性	$36/63 = 0.57$	$36/42 = 0.86$
女性	$36/27 = 1.33$	$36/18 = 2$

(6) その和を求める

$$\chi = 0.57 + 0.86 + 1.33 + 2 = 4.76$$

このように求めた値 χ は、カイ二乗分布に近似できる。

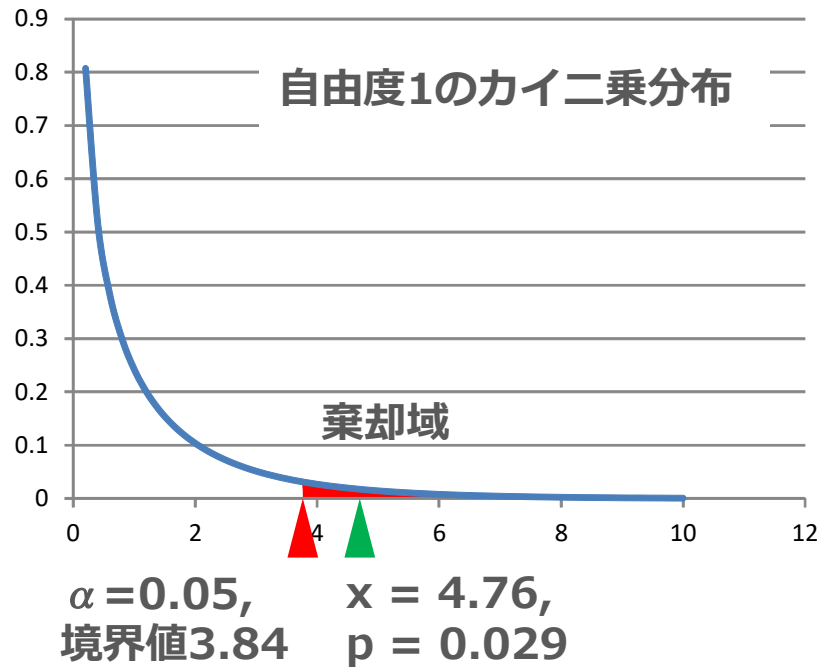
自由度は、各カテゴリ（性別、ビールの好み）の要素数をそれぞれ n_1 , n_2 とすると、 $(n_1-1)*(n_2-1)$ 。

この例の場合では、 $(2-1)*(2-1) = 1$

カイ二乗検定の手順

(7) 結論

xの値が棄却域の境界値の外側 ($3.84 < 4.76$, $p=0.029 < \alpha$) なので、帰無仮説は棄却され、「二つのカテゴリは独立ではない」と判断された。



よって、この母集団においては、「性別とビールの好みとの間に何かしらの関連性がある」と結論づけられた。

カイ二乗分布を扱うExcelの関数：
CHISQ.DIST, CHISQ.DIST.RT, CHISQ.INV.RTなど

カイ二乗検定の留意点

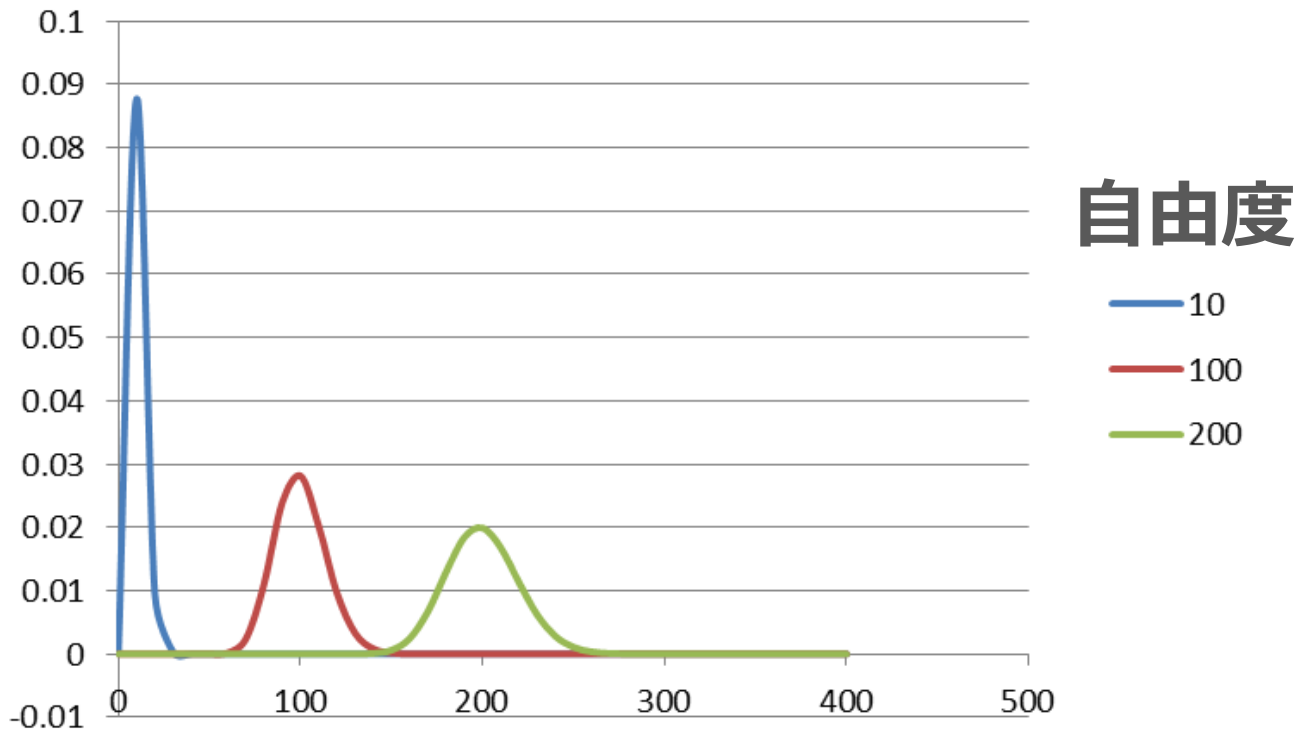
観測数が少ないとカイ二乗分布への近似ができないので、その場合はフィッシャーの正確確率検定を行う。

目安：

期待度数が5未満のセルが、全セルの20%以上で存在する場合、近似が不正確と考えられる
(コクラン・ルール)

期待度数が1未満のセルがあってはならない

カイ二乗分布の性質 その2



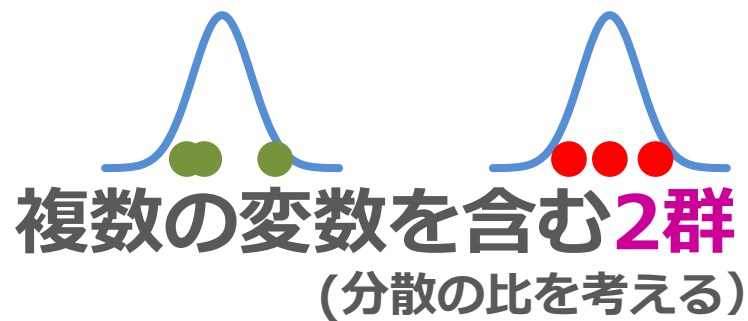
自由度 k が大きくなると、

平均値： k

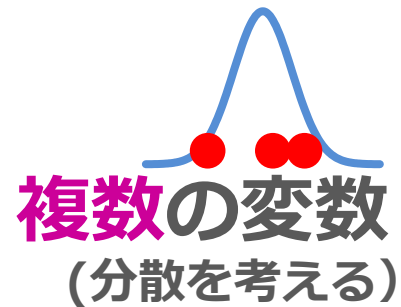
分散： $2k$

の正規分布に近づいてゆく

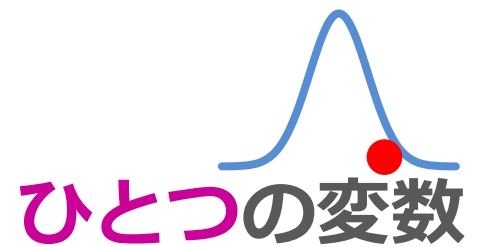
F分布



カイ二乗分布



標準正規分布



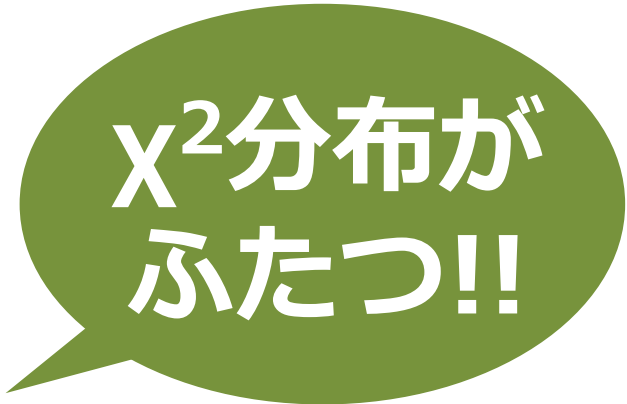
F分布とカイ二乗分布の関係

自由度 k_1 のカイ二乗分布 χ^2_1

自由度 k_2 のカイ二乗分布 χ^2_2

があるとき、次の値 F は、自由度 (k_1, k_2) のF分布に従う

$$F = \frac{\chi^2_1 / k_1}{\chi^2_2 / k_2}$$



χ^2 分布が
ふたつ!!

F 分布の活用

正規分布 $N(\mu_1, \sigma^2_1)$ に従った母集団から得た標本、
標本数： n_1 、不偏標本分散： v^2_1

正規分布 $N(\mu_2, \sigma^2_2)$ に従った母集団から得た標本、
標本数： n_2 、不偏標本分散： v^2_2

があるとき、

$$F = \frac{\chi^2_1/k_1}{\chi^2_2/k_2} = \frac{v^2_1/\sigma^2_1}{v^2_2/\sigma^2_2}$$

二つの母集団の分散 σ^2_1 と σ^2_2 が等しいと仮定できる場合は、

$$F = \frac{v^2_1}{v^2_2}$$



これをF検定で利用している！

F 分布の活用

カイ二乗分布の性質

$$\chi^2 = \frac{\sum_{i=1}^k (x_i - \mu)^2}{\sigma^2} \quad \text{自由度} k$$

この式を変形すると、

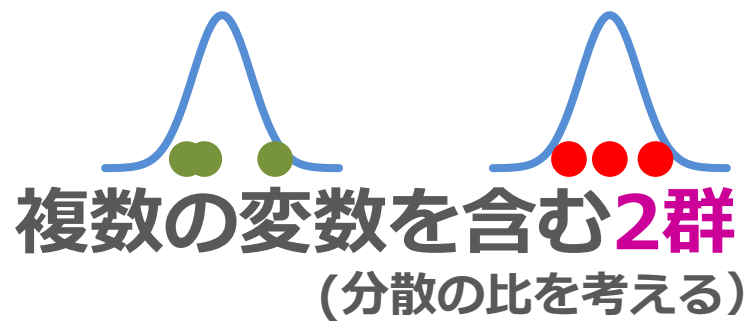
不偏標本分散 v^2 になっている！

$$\chi^2 = \frac{k \times \frac{\sum_{i=1}^k (x_i - \mu)^2}{k}}{\sigma^2} = \frac{k \times v^2}{\sigma^2}$$

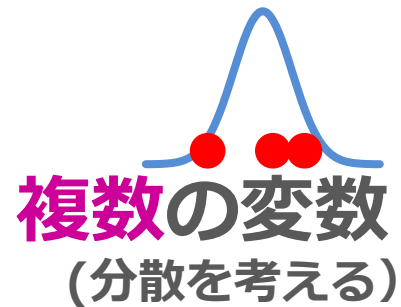
したがって、

$$\frac{\chi^2}{k} = \frac{k \times v^2}{\sigma^2} \times \frac{1}{k} = \frac{v^2}{\sigma^2}$$

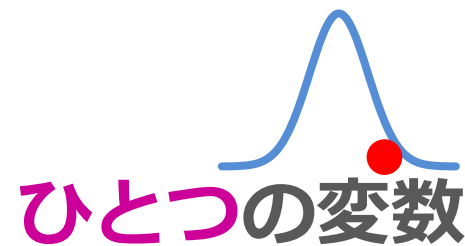
F分布



カイ二乗分布



標準正規分布



by 櫻井

補足

- 等分散性の検定（F検定）
- 分散分析（F分布を使う）
- ログ変換
- 主成分分析の例

分散分析

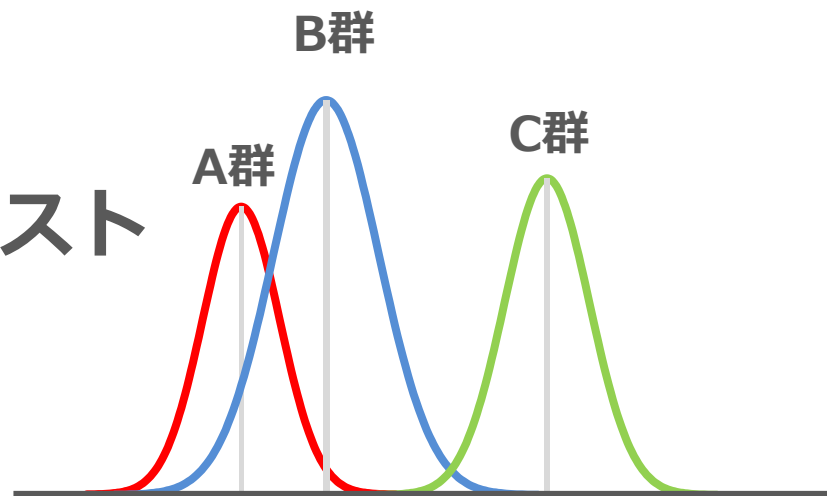
Analysis of Variance

ANOVA

- ✓ 3つ以上の群があるとき、
- ✓ 群の母平均に差があるかどうかを、
- ✓ 分散（F分布）を使って、

検定する方法

例）1組、2組、3組で、テストの平均点に差があるか？



帰無仮説：

A群、B群、C群の母平均は等しい

対立仮説：

**A群、B群、C群の母平均の中に、
異なる値がある**

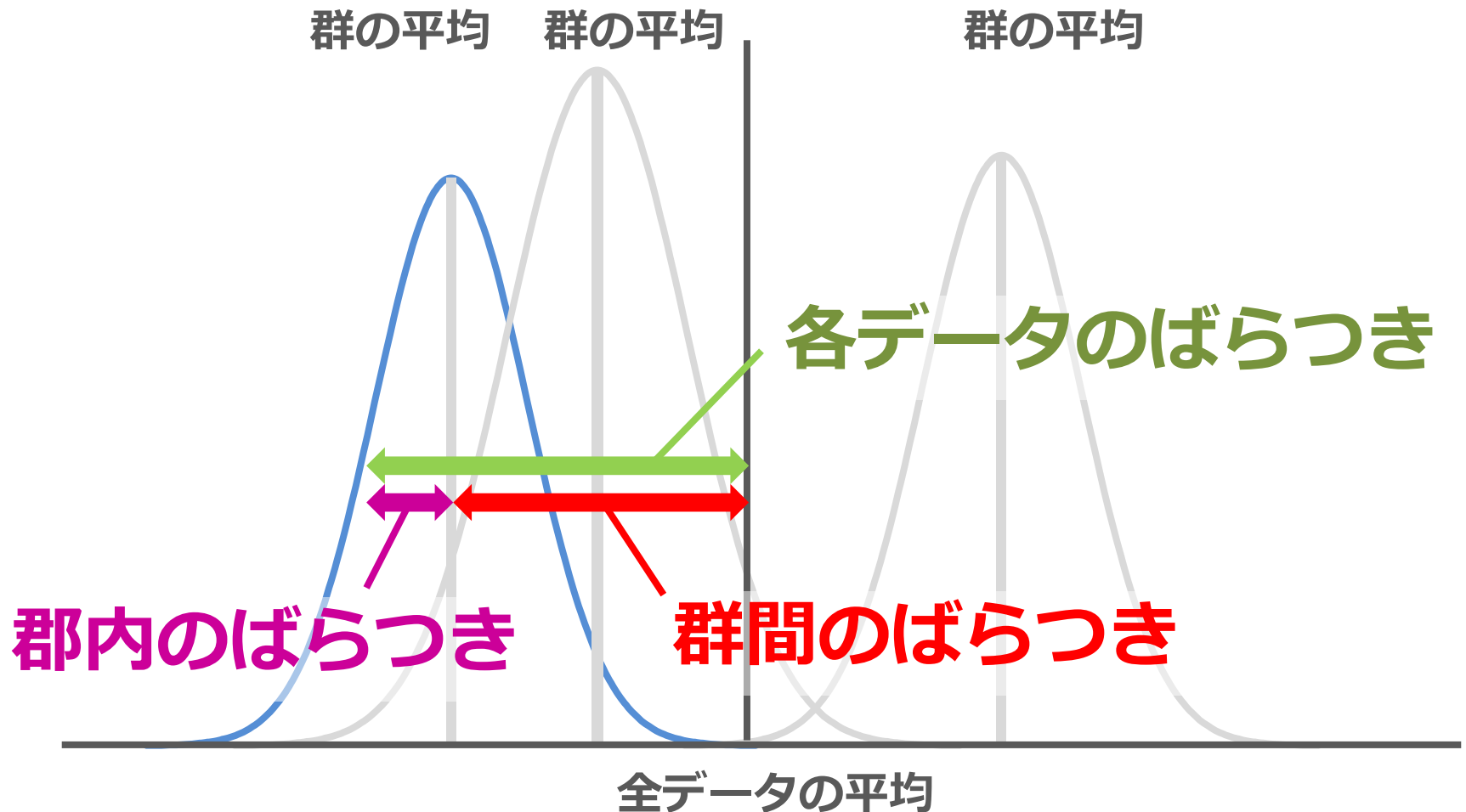


どれが異なっているかまではわからない！

帰無仮説が棄却されたときは、解釈に注意が必要

分散分析のイメージ

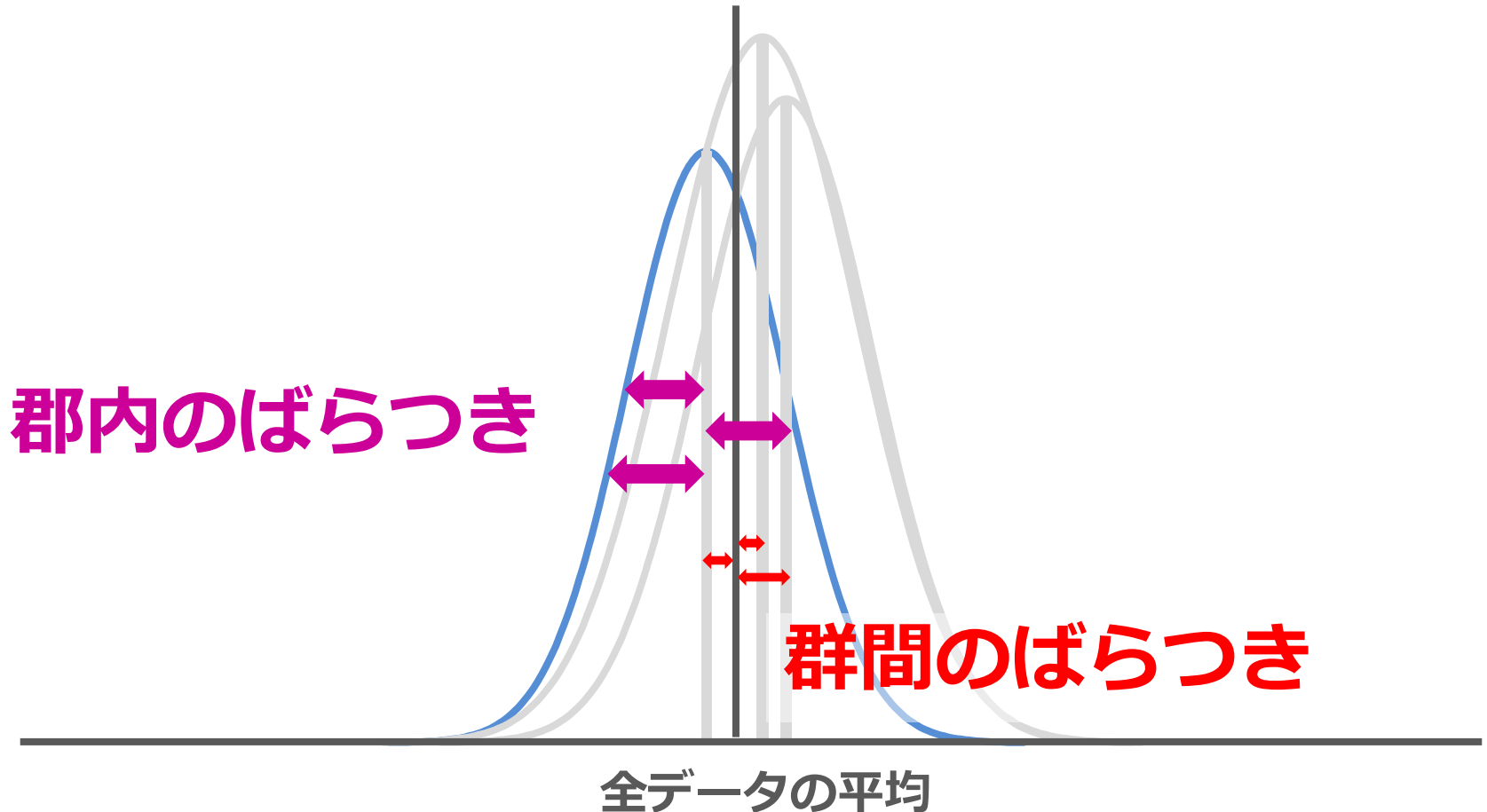
データのばらつきを、**群間**のばらつきと、**偶然により起こる群内**のばらつきに分けて考える



分散分析のイメージ

群の平均に差がなければ、

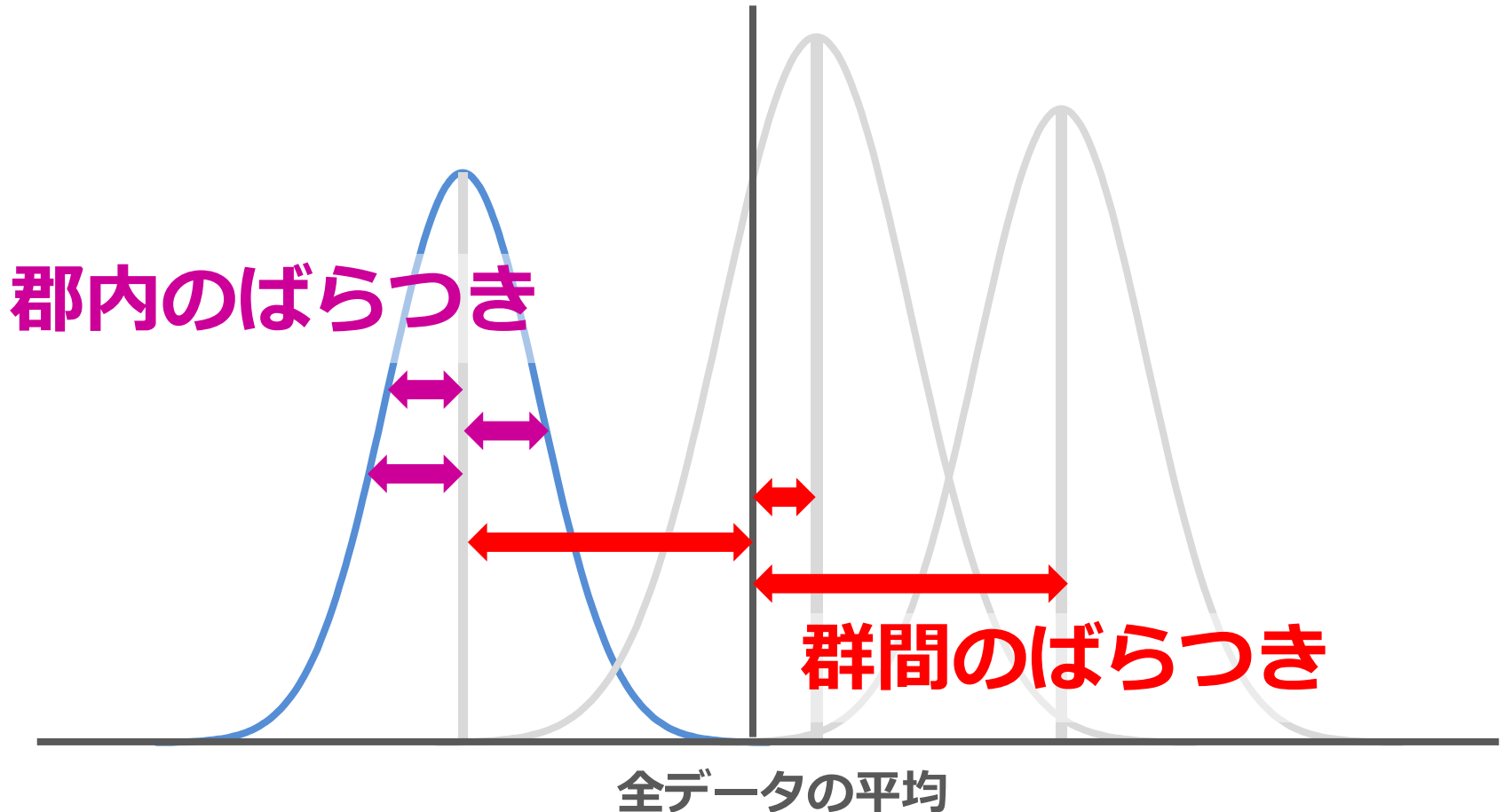
群内のばらつき > **群間**のばらつき



分散分析のイメージ

群の平均に差があるほど、

群内のばらつき < **群間**のばらつき



分散分析の手順

分散分析表を穴埋めしてゆく

要因	平方和 S	自由度 df	不偏標本分散 V ²	F値
群間 (因子)	S(群)	df(群) =群の数-1	V ² (群) =S(群)/df(群)	V ² (群)/V ² (残差)
群内 (残差)	S(残差)	df(残差) =全データ数-群 の数	V ² (残差) =S(残差)/df(残差)	
全体	S(全体)	df(全体)		

分散分析の手順

例) A～Dの異なる生育環境で育てた植物の、ある成分の含量

A群	341	347	328	329	352
B群	305	317	342	322	319
C群	342	313	350	323	
D群	331	327	303	314	

エクセルファイル:190923_anova.xlsx

以下の基本情報を計算する

- ①群ごとのデータ数
- ②全データの個数
- ③群の平均値
- ④全データの平均値

以下の差（ずれ）を計算する

- ⑤全データについて、全体の平均からの差
- ⑥各群の平均について、全体の平均からの差
- ⑦郡内の各データについて、群平均からの差

差（ずれ）の二乗を計算する

- ⑧全データについて、全体の平均からの差の二乗
- ⑨各群の平均について、全体の平均からの差の二乗
群のデータ数を乗じる
- ⑩郡内の各データについて、群平均からの差の二乗

二乗和を計算する

- ⑪ 全データについての全体の平均からの差の二乗和
- ⑫ 各群の平均についての全体の平均からの差の二乗和
- ⑬ 群内の各データについての群平均からの差の二乗和

分散分析表を埋める

⑭ 二乗和

⑪ = ⑫ + ⑬ となっているはず

⑮ 自由度

全体：② 全データ数 - 1

群間：群の個数 - 1

群内：全体の自由度 - 群間の自由度

⑯ 不偏標本分散（群間、群内について）

二乗和 / 自由度

⑰ F値

不偏標本分散の比（群間/群内）

用語

要因：
データに影響を与えるもの

因子：
要因の中で特に母平均の差に影響すると思われたため、解析の対象とするもの

残差：
偶然によって生じたばらつき

p値、 α のF境界値を計算する

⑮⑯で求めたF値と自由度から、F.DIST.RT関数を使って、p値を計算する

⑰有意水準 α に対応するF境界値を、F.INV.RT関数を使って計算する

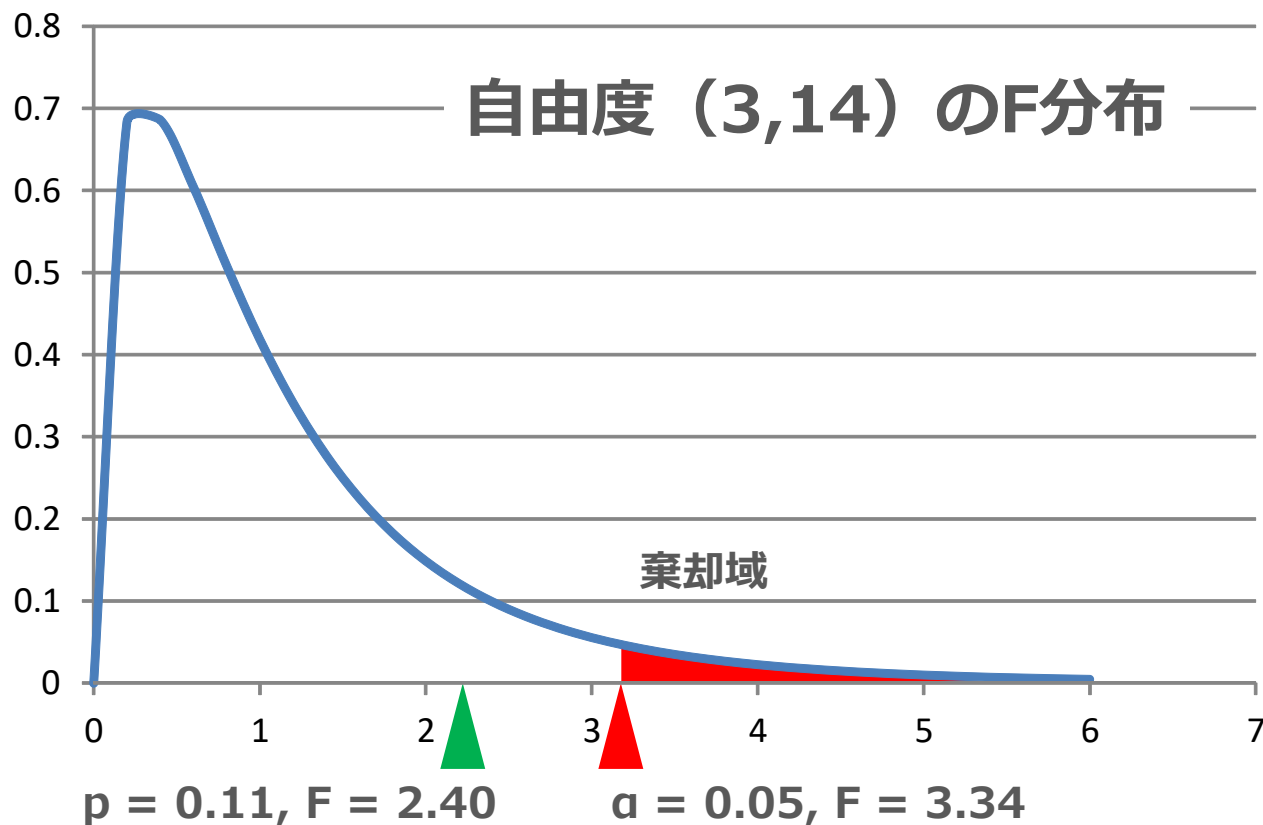
⑱F.DIST関数を用いて当該自由度のF分布を描く

p値の大きさ、 α に対応する境界値の大きさなどから、検定統計量が棄却域に入ったかどうかを判断する

結論づけをする

結論

p値は0.11となり、有意水準0.05で帰無仮説は棄却されなかった。したがって、「A～Dの生育方法によって成分の平均値に差があるとは言えない」と結論付けられた。



分散分析の種類



今回やった
もの

一元配置の分散分析 one-way ANOVA

一つの因子からなるデータを分析する方法

二元配置の分散分析 two-way ANOVA

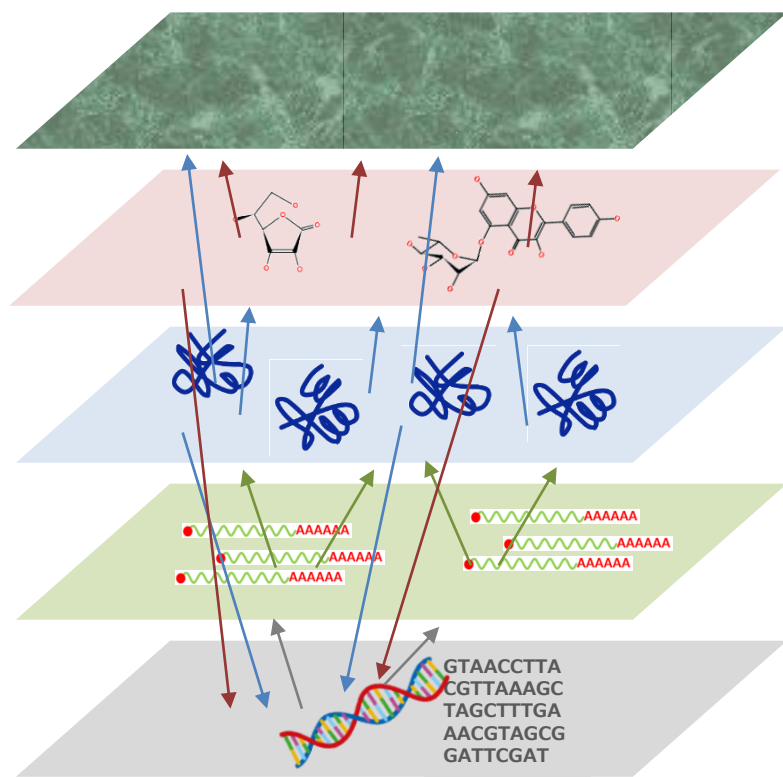
二つの因子からなるデータを分析する方法。例) 薬剤の種類と投与量など。二つの要因が組み合わさる交互作用(相乗効果)を確認することもできる

多元配置の分散分析

補足

- 等分散性の検定（F検定）
- 分散分析（F分布を使う）
- ログ変換
- 主成分分析の例

生物の遺伝子情報の流れとオミクス



表現型

代謝成分

タンパク質

転写産物

ゲノム

?

数万?

数万

数万

数万

オミクス

それぞれの要素を一斉に検出
しようとする技術・学問

一見、正規分布のように見えないデータでも、ログスケール（対数）にすることで、正規分布に近い分布になることがある

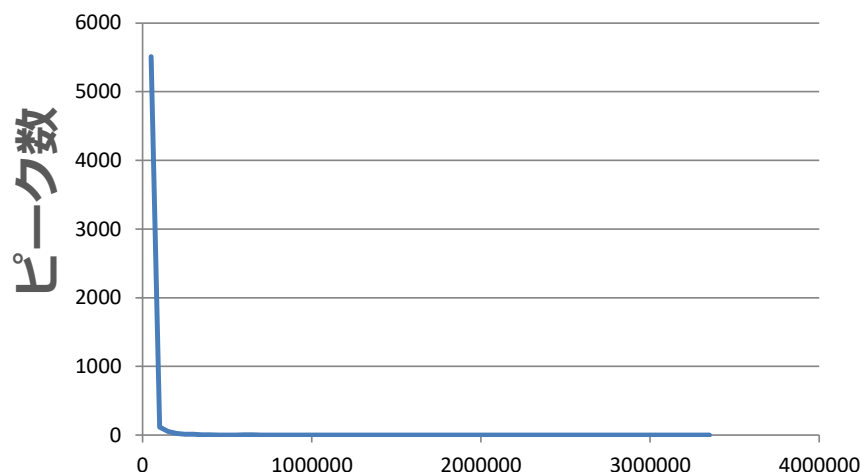
- ✓ 遺伝子発現量データ
- ✓ 質量分析での化合物検出データ

など

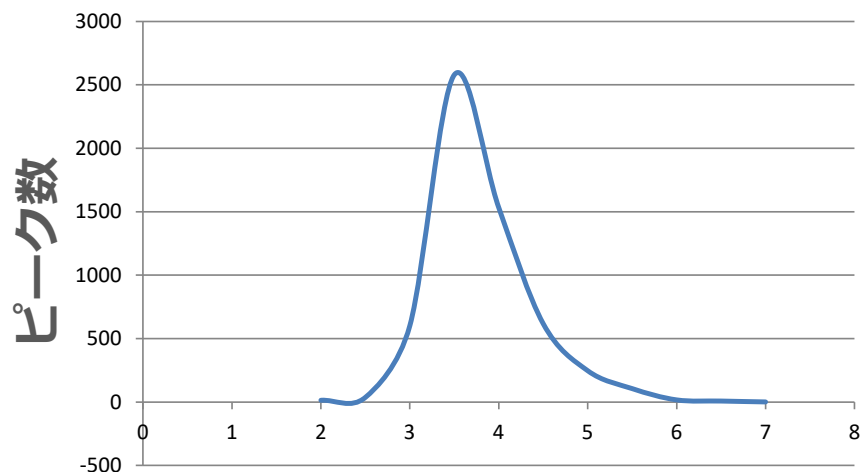
大葉（しそ）で検出された代謝物質

- 液体クロマトグラフィー-質量分析
- ESIポジティブモード

計5760ピーク



検出値
(リニアスケール)



log10変換後
(ログスケール)

Excel関数: LOGなど

ログスケールにするメリット

シグナル強度によるばらつき（分散）の変化を打ち消すことができる

例）強度10のピークの10%のばらつきは1の差なのに対し、強度1000のピークでは、同じ10%のばらつきで100の差になる。

logに変換すると、どんな強度でも同じ数値幅のばらつきにすることができる（等分散）



データの分布をExcelで描いて判断

補足

- 等分散性の検定（F検定）
- 分散分析（F分布を使う）
- ログ変換
- 主成分分析の例

自習

課題検討

発表会