

# 情報統計 第13-15回

2021年9月14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

	13日(月) データの見える化	14日(火) 検定のこれだけは	15日(水) 分散分析と多変量解析の雰囲気	16日(木) データ準備 発表会
1限				13 補足 自習(課題、質問)
2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 分布の仲間と、 分散分析	14 自習(課題、質問)
3限	2 統計の基本と 用語	6 t検定	10 相関、主成分 分析	15 発表会
4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	
5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

- 数学記号
- ログ変換
- 主成分分析の例

# 2群のt検定（独立2群）

等分散が仮定できない場合 ウェルチの方法

1群目：標本数  $n_1$ , 不変標本分散  $s_1^2$ , 標本平均  $\bar{x}_1$

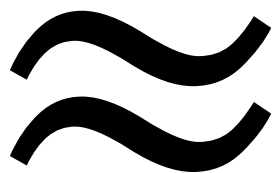
2群目：標本数  $n_2$ , 不変標本分散  $s_2^2$ , 標本平均  $\bar{x}_2$

検定統計量  $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

(近似)自由度  $v \approx \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}}$

帰無仮説：2群の母集団の平均値は等しい

で、同様に検定できます 参考まで



ほぼ等しい

数学記号

○	合成写像	「 $f \circ g$ 」は写像 $g$ と写像 $f$ の合成を表す。すなわち $(f \circ g)(x) = f(g(x))$ である。
Im, Image, • [•]	像	写像 $\varphi$ に対して、Image $\varphi$ はその写像の像全体の集合（値域）を表す。写像 $\varphi: X \rightarrow Y$ に対して $\varphi[X]$ とも書く。

二項関係演算

記号	意味	解説
=	相等	$x = y$ は $x$ と $y$ が等しいことを表す。
≠	不一致	$x \neq y$ は $x$ と $y$ が等しくないことを表す。
≐, ≈	ほぼ等しい	「 $x \doteq y$ 」または「 $x \approx y$ 」は $x$ と $y$ がほぼ等しいことを表す。記号 $\doteq$ は日本など少数の地域でのみ通用し、 $\approx$ の方が標準的である。その他にも $\sim, \simeq, \cong$ などを同様の意味で用いることもある。近似においてどのくらい違いを容認するかは文脈による。多くの場合、 <a href="#">誤差</a> 解析的な意味で用いられ、ある誤差の見積もりの下で両者が等しいことを示すが、そのほかにも <a href="#">漸近</a> 解析においては漸近的に等しいという意味で用いられる。

順序構造

記号	意味	解説
< . >	大小関係、 <a href="#">順序</a>	「 $x < y$ 」は $x$ と $y$ の間に1方が「先」であることを示す

# Excelで数式表示

作図.xlsx - Excel Nozomu Sakurai NS

ファイル ホーム 挿入 描画 ページレイアウト **数式** データ 校閲 表示 ヘルプ 検索 共有 コメント

fx  $\Sigma$  オートSUM  $\downarrow$  関数の挿入 最近使った関数  $\downarrow$  財務  $\downarrow$  関数ライブラリ 論理  $\downarrow$  文字列操作  $\downarrow$  日付/時刻  $\downarrow$  数学/三角  $\downarrow$  その他の関数  $\downarrow$  検索/行/列  $\downarrow$  数学/三角  $\downarrow$  その他の関数  $\downarrow$  名前 名前の定義  $\downarrow$  数式で使用する  $\downarrow$  選択範囲から作成 定義された名前 参照元のトレース 参照先のトレース トレース矢印の削除  $\downarrow$  数式の表示 エラー チェック 数式の検証 ワークシート分析 ウォッチ ウィンドウ 計算方法の設定  $\downarrow$  計算方法

J2

	C	D	E	F	G	H	I	J
1								
2		平均値からの差	←の二乗					
3	148	=C3-\$H\$6	=D3^2			=SUM(C3:C28)		
4	148	=C4-\$H\$6	=D4^2		合計	=COUNTA(C3:C28)		
5	149	=C5-\$H\$6	=D5^2		個数	=H4/H5	手計算	
6	150	=C6-\$H\$6	=D6^2		平均値	=AVERAGE(C3:C28)	AVERAGE関数	
7	150	=C7-\$H\$6	=D7^2		平均値	=SUM(E3:E28)		
8	150.4	=C8-\$H\$6	=D8^2		二乗和	=H8/H5		
9	151	=C9-\$H\$6	=D9^2		分散	=VAR.P(C3:C28)	VAR.P	
10	153	=C10-\$H\$6	=D10^2		分散	=H8/(H5-1)		
11	153	=C11-\$H\$6	=D11^2		不偏標本分散	=VAR.S(C3:C28)	VAR.S	
12	153.4	=C12-\$H\$6	=D12^2		不偏標本分散	=SQRT(H9)		
13	155	=C13-\$H\$6	=D13^2		標準偏差	=STDEV.P(C3:C28)	STDEV.P	
14	155	=C14-\$H\$6	=D14^2		標準偏差	=SQRT(H11)		
15	155.5	=C15-\$H\$6	=D15^2		不偏標本標準偏差	=STDEV.S(C3:C28)	STDEV.S	
16	156.6	=C16-\$H\$6	=D16^2		不偏標本標準偏差			
17	157	=C17-\$H\$6	=D17^2					
18	157	=C18-\$H\$6	=D18^2					

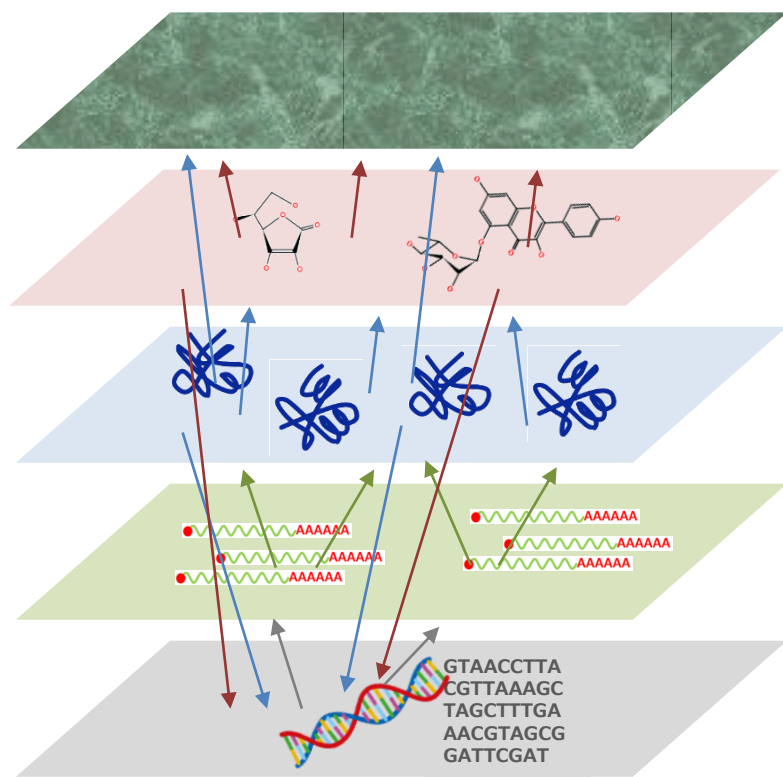
Sheet6 Sheet5 Sheet7 Sheet8 Sheet9 Sheet10 Sheet10 (2) 100%

# 補足

- 数学記号
- 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

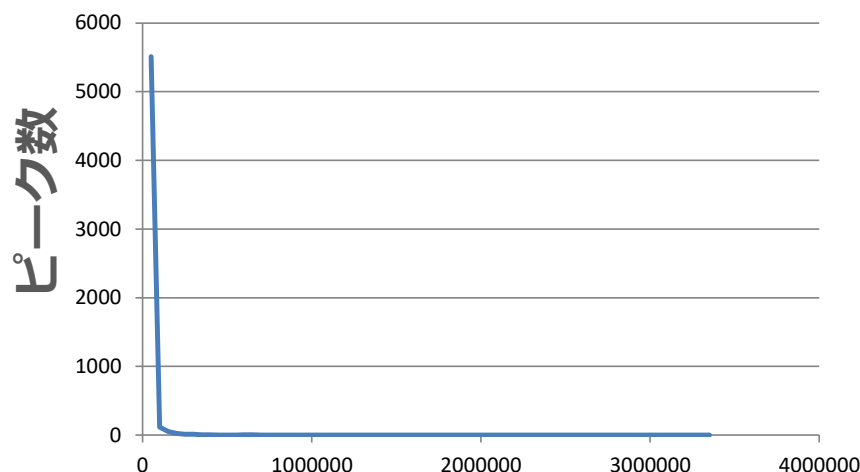
- ✓ 遺伝子発現量データ
- ✓ 質量分析での化合物検出データ

など

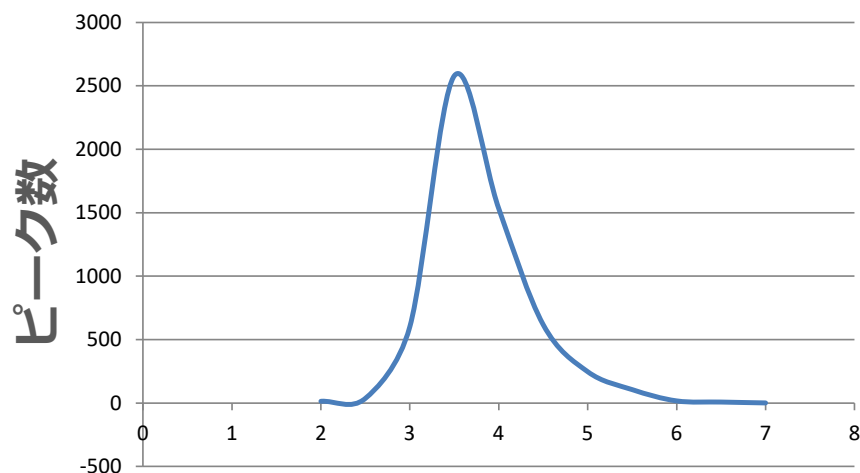
# 大葉（しそ）で検出された代謝物質

- 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

- 数学記号
- ログ変換
- 主成分分析の例

自習

課題検討

発表会