情報統計
第13-15回

2020年9月19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

	16日(水) データの見える化	17日(木) 検定のこれだけは	18日(金) 分散分析と多変量解析の雰囲気	19日(土) 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　分布の仲間と、分散分析	13　補足 自習(課題、質問)
2限	2　統計の基本と用語	6　t検定	10　相関、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

- 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 n1, 不変標本分散 s1, 標本平均$\overline{x1}$

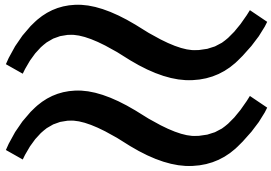
2群目：標本数 n2, 不変標本分散 s2, 標本平均$\overline{x2}$

検定統計量　$$t = \frac{\overline{x1} - \overline{x2}}{\sqrt{\frac{s1^2}{n1} + \frac{s2^2}{n2}}}$$

(近似)自由度　$$v \approx \frac{(\frac{s1^2}{n1} + \frac{s2^2}{n2})^2}{\frac{s1^4}{n1^2(n1-1)} + \frac{s2^4}{n2^2(n2-1)}}$$

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

$\circ$	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。
$\mathrm{Im},\ \mathrm{Image},\ \bullet[\bullet]$	像	写像 φ に対して、Image φ はその写像の像全体の集合（値域）を表す。写像$\varphi\colon X \to Y$に対して $\varphi[X]$とも書く。

二項関係演算

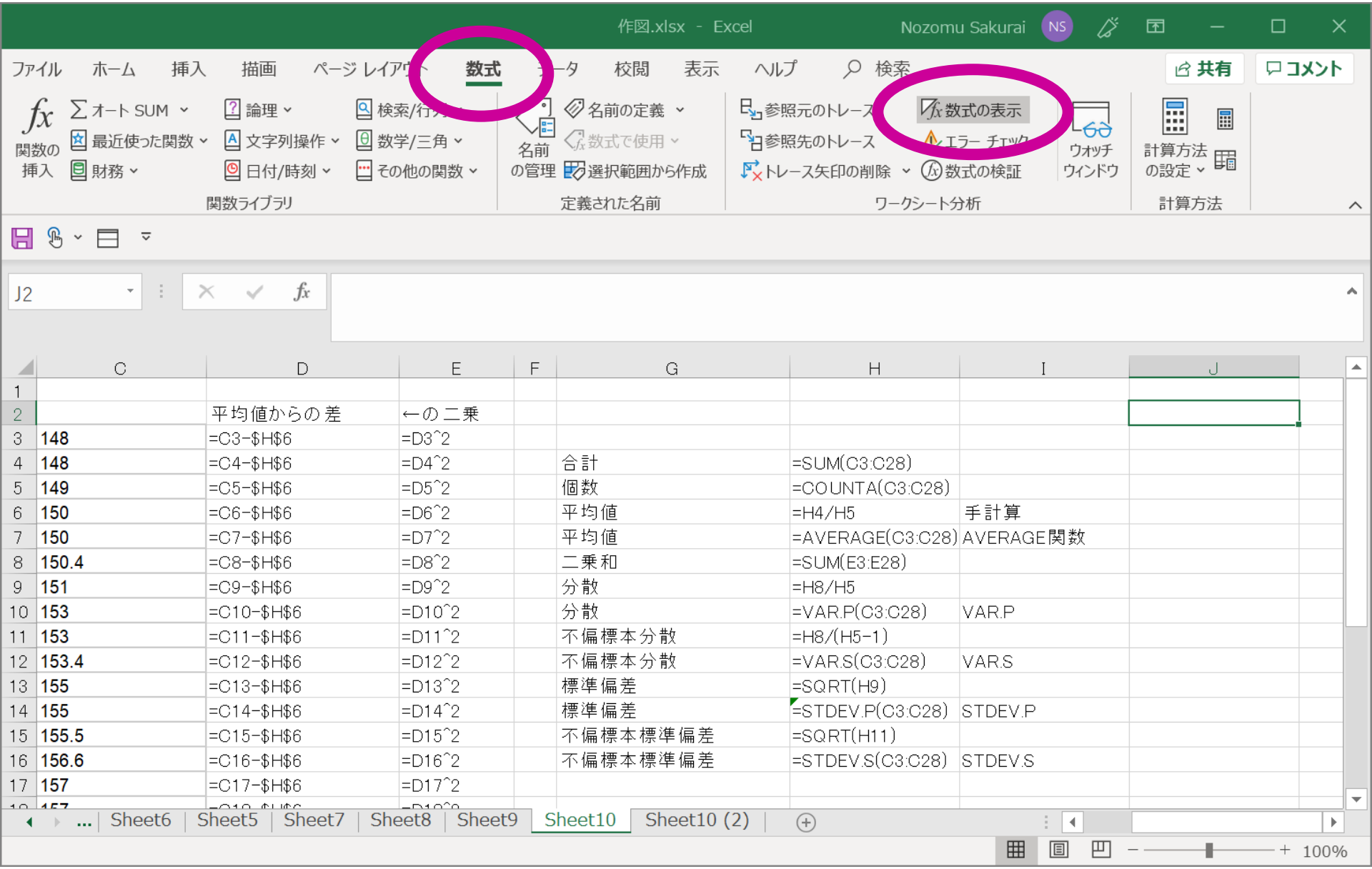
記号	意味	解説
$=$	相等	$x = y$ は x と y が等しいことを表す。
$\neq$	不一致	$x \neq y$ は x と y が等しくないことを表す。
$\fallingdotseq$, $\approx$	ほぼ等しい	「$x \fallingdotseq y$」または「$x \approx y$」は x と y がほぼ等しいことを表す。記号 $\fallingdotseq$ は日本など少数の地域でのみ通用し、$\approx$ の方が標準的である。その他にも $\sim$, $\simeq$, $\cong$ などを同様の意味で用いることもある。近似においてどのくらい違いを容認するかは文脈による。多くの場合、誤差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、そのほかにも漸近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、順序	「$x < y$」は　x と y の間に(方が「先」であることを示す

Wikipedia

Excelで数式表示

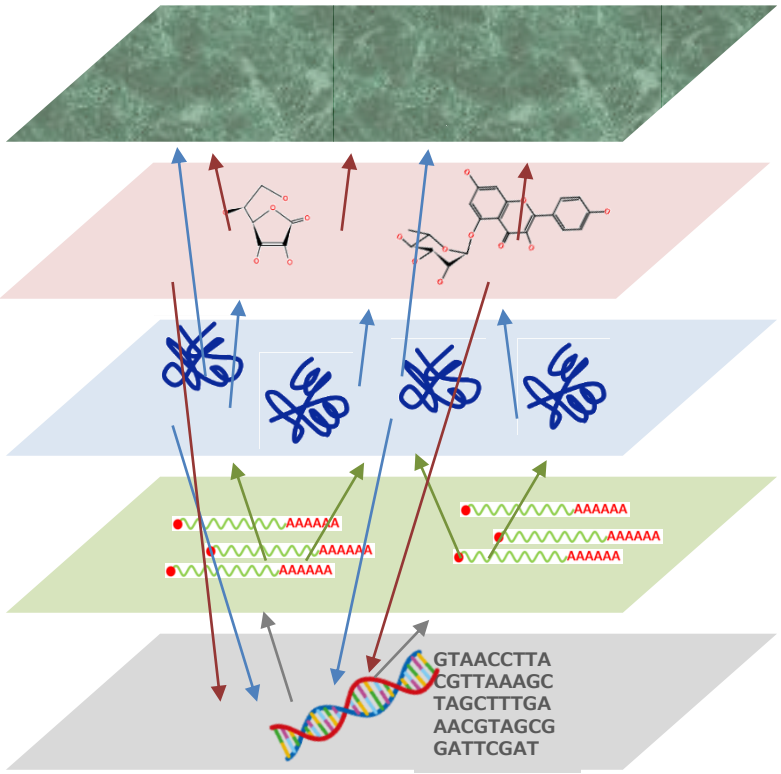


	C	D	E	F	G	H	I	J
1								
2		平均値からの差	←の二乗					
3	148	=C3-H6	=D3^2					
4	148	=C4-H6	=D4^2		合計	=SUM(C3:C28)		
5	149	=C5-H6	=D5^2		個数	=COUNTA(C3:C28)		
6	150	=C6-H6	=D6^2		平均値	=H4/H5	手計算	
7	150	=C7-H6	=D7^2		平均値	=AVERAGE(C3:C28)	AVERAGE関数	
8	150.4	=C8-H6	=D8^2		二乗和	=SUM(E3:E28)		
9	151	=C9-H6	=D9^2		分散	=H8/H5		
10	153	=C10-H6	=D10^2		分散	=VAR.P(C3:C28)	VAR.P	
11	153	=C11-H6	=D11^2		不偏標本分散	=H8/(H5-1)		
12	153.4	=C12-H6	=D12^2		不偏標本分散	=VAR.S(C3:C28)	VAR.S	
13	155	=C13-H6	=D13^2		標準偏差	=SQRT(H9)		
14	155	=C14-H6	=D14^2		標準偏差	=STDEV.P(C3:C28)	STDEV.P	
15	155.5	=C15-H6	=D15^2		不偏標本標準偏差	=SQRT(H11)		
16	156.6	=C16-H6	=D16^2		不偏標本標準偏差	=STDEV.S(C3:C28)	STDEV.S	
17	157	=C17-H6	=D17^2					

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 数万？

タンパク質 数万

転写産物 数万

ゲノム 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

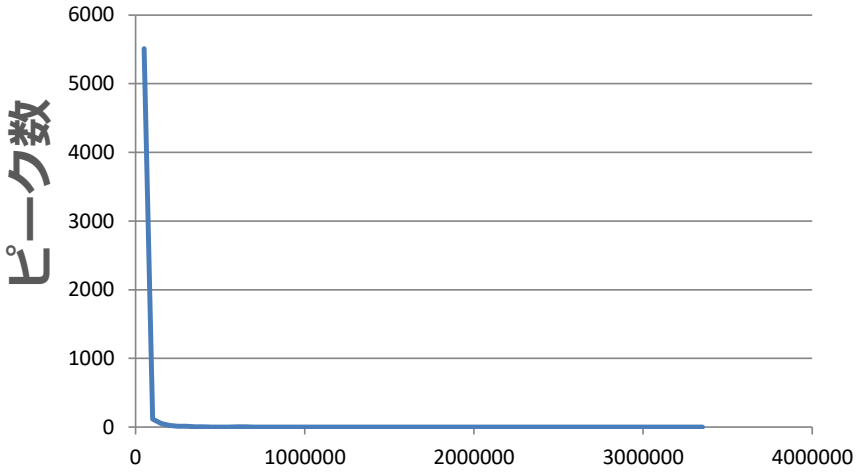
- ✓ 遺伝子発現量データ
- ✓ 質量分析での化合物検出データ

など

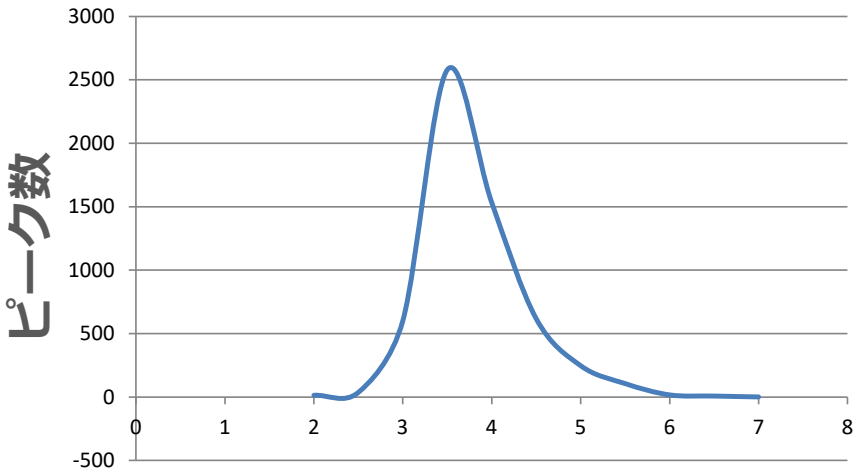
大葉（しそ）で検出された代謝物質

- 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