情報統計
第11回

2024年8月1日　神奈川工科大学

櫻井 望
公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

補足

- カイ二乗検定の計算
  - ✓ どこで使われる？
  - ✓ 期待度数などの計算
- 主成分分析
  - ✓ 第n軸が全く同じ分散になったら？

	ビール好き	ビールあんまり
男性	69	36
女性	21	24

	A型	B型	AB型	O型
男性				
女性				

	ゲーム好き	ゲームしない
朝食食べる		
朝食食べない		

	ペット飼ってる	ペットいない
独身		
既婚		

	治った	治らなかった
薬剤投与		
コントロール		

分割表
数字の大小で表せないものを扱う

期待度数

(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	ビール好き	合計	割合
男性	69	36	105	0.7
女性	21	24	45	0.3
合計	90	60	150	
割合	0.6	0.4		

(2) (1)の割合から、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	ビール好きではない
男性	=E10*C13	42
女性	27	18

自由度

自由度	=(COUNTA(C3:D3)-1) * (COUNTA(B4:B5)-1)

観測データ

	ビール好き	ビール好きではない
男性	69	36
女性	21	24

研究紹介

主成分分析、散布図などの例

情報統計
第12回

2024年8月1日　神奈川工科大学

櫻井 望
公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

補足

- 数学記号
- ログ変換

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$\overline{x1}$

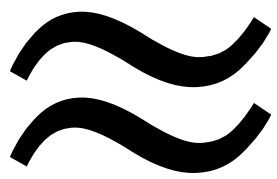
2群目：標本数 n2, 不変標本分散 s2, 標本平均$\overline{x2}$

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。

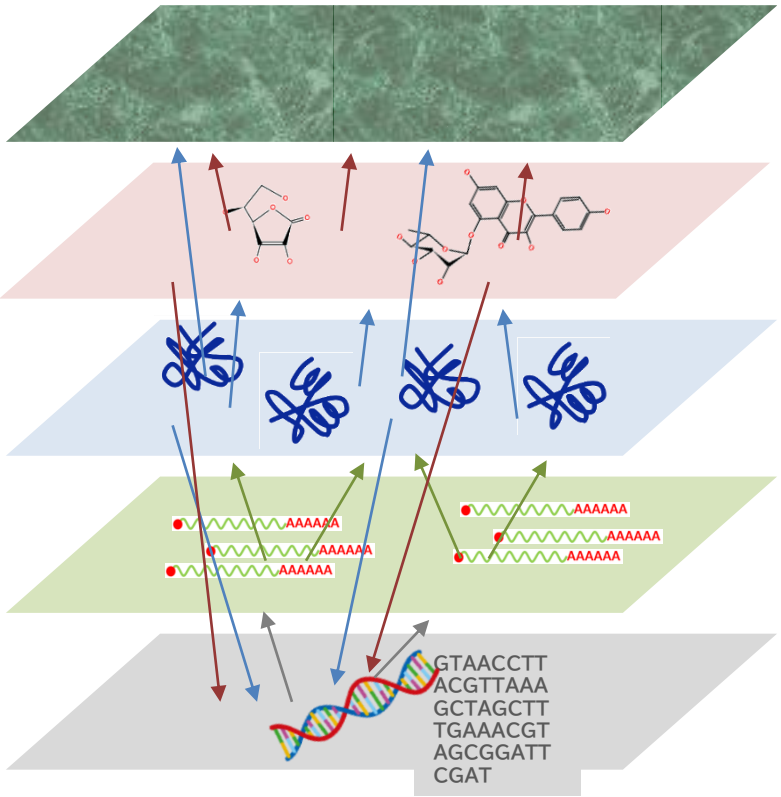
順序構造

記号	意味	解説
$<, >$	大小関係、順序	「$x < y$」は　x と y の間に何方が「先」であることを示す。

補足

- 数学記号
- ログ変換

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

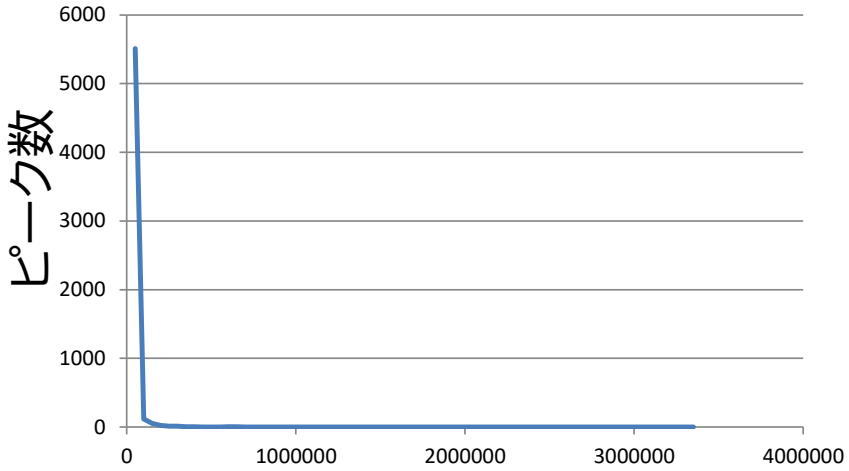
- ✓ 遺伝子発現量データ
- ✓ 質量分析での化合物検出データ

など

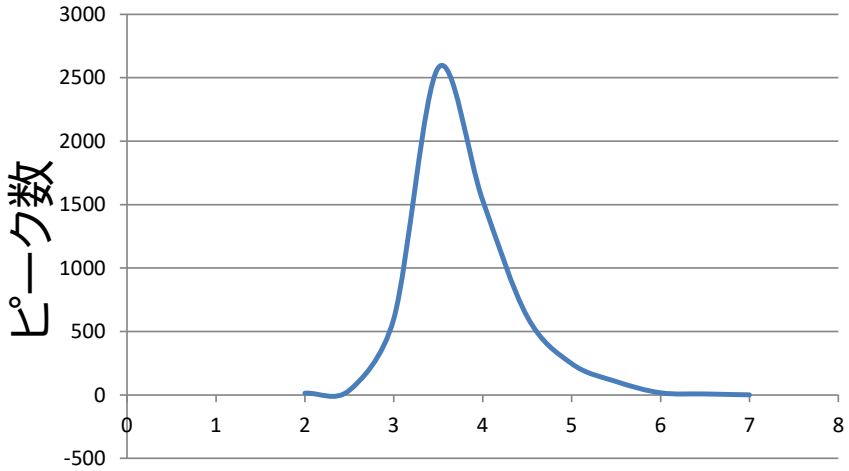
大葉(しそ)で検出された代謝物質

- 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

情報統計
第13回

2024年8月1日　神奈川工科大学

櫻井 望
公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

自習

課題の準備

情報統計
第14回

2024年8月1日　神奈川工科大学

櫻井 望
公益財団法人かずさDNA研究所
先端研究開発部 シーズ開拓研究室
藻類代謝エンジニアリングチーム

発表会

お疲れさまでした