情報統計
第9-11回

2019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

	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）		

相関

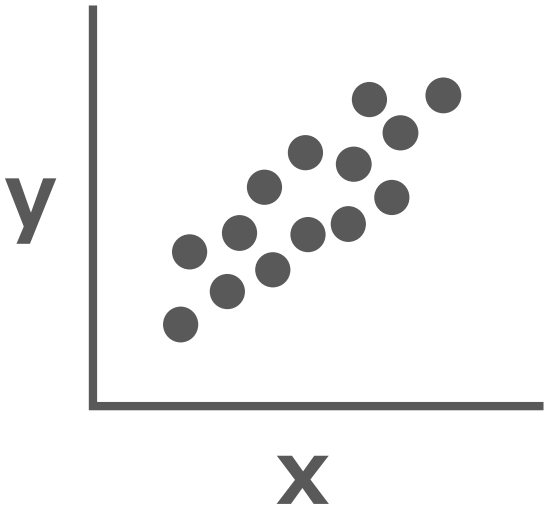
学習目標

相関のあるなしを評価できるようになる

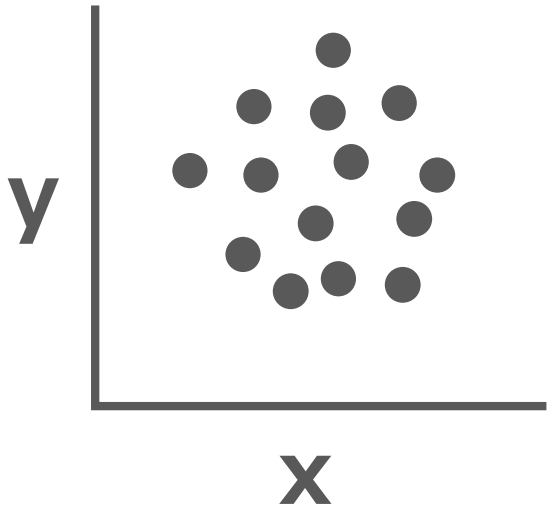
相関関係と因果関係の違いが分かる

散布図

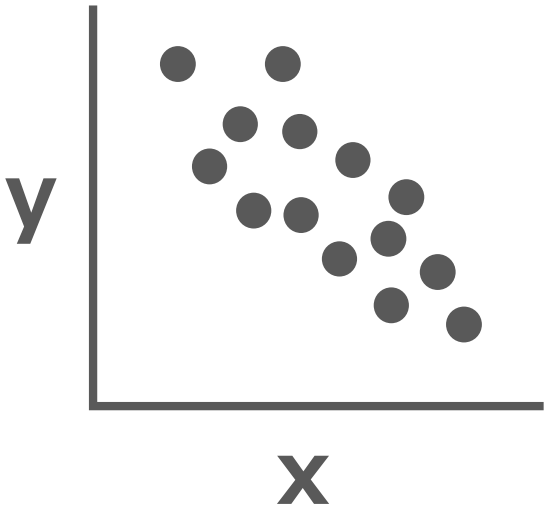
二つの変数の間の関係性を見える化する手法



正の相関がある

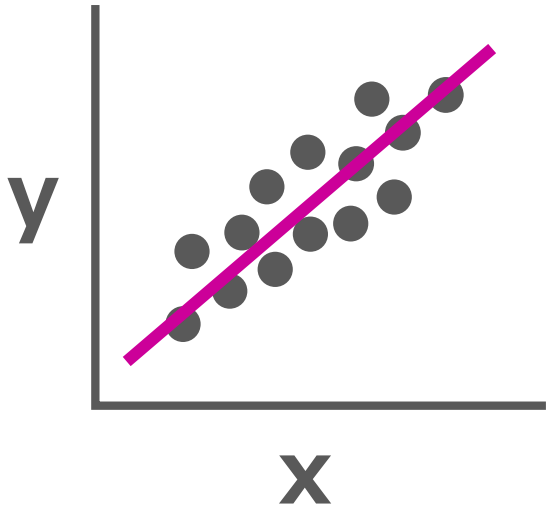


相関がない



負の相関がある

散布図の回帰曲線



エクセルのグラフ上でプロットを右クリックし、挿入できる

相関係数

- 二つの変数の間の関係性の強さを数値化したもの
- -1～1の間の値をとる

0.7～1.0：強い正の相関
0.4～0.7：中程度の正の相関
0.2～0.4：弱い正の相関

-1.0～-0.7：強い負の相関
-0.7～-0.4：中程度の負の相関
-0.4～-0.2：弱い負の相関

-0.2～0.2：相関がない

- ExcelではPEARSON関数で計算できる

相関関係を見てみる

都道府県別の統計

https://todo-ran.com/

都道府県別統計とランキングで見る県民性

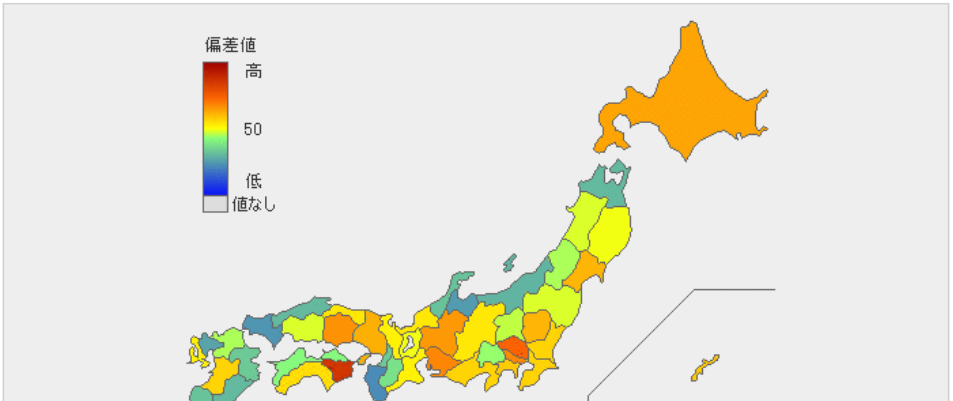
都道府県別統計を比較した都道府県ランキング。**1339** ランキング掲載中

odomon@gmail.com

トップ

最新ランキング

2019年参議院比例代表：NHKから国民を守る党得票率 [2019年 第一位 徳島県]



当サイト一番人気

都道府県ベスト&ワースト

各都道府県の1位と47位だけを一覧表にまとめました。県民性が一目で分かります。

都道府県比較

東京vs大阪、埼玉vs千葉vs神奈川など任意の都道府県の似たとこ、似ていないところを一

記事を探す

- 検索から探す（googleサイト内検索）

Google カスタム検索

サイト内検索

- カテゴリから探す
  政治・経済などカテゴリ別全記事表示
- 新着から探す
  新しい順に全記事表示

データを集めてみる

例）
神奈川県の高いランクのうち、
「しゅうまい消費量」と
「最低賃金」や「農業就業人口」との相関

- サイトでデータをコピー
- エクセルに貼り付け
- エクセルで加工（県の列で並び替え）
- 散布図を描く
- PEARSON関数で相関係数を計算する

相関係数を手で計算する

ピアソンの積率相関係数

$$r = \frac{s_{xy}}{s_x s_y}$$

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

sxy: xとyの**共分散**
sx: xの標準偏差
sy: yの標準偏差
n: xとyのペアの数

無相関の検定

帰無仮説：
母集団の相関係数は0（無相関）である

分布： *t*分布

検定統計量： $t = \dfrac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$

自由度： n-2

※|r|はrの絶対値
エクセルではABS関数
で計算できる

そのほかの相関係数

- スピアマンの順位相関係数
- コサイン相関係数

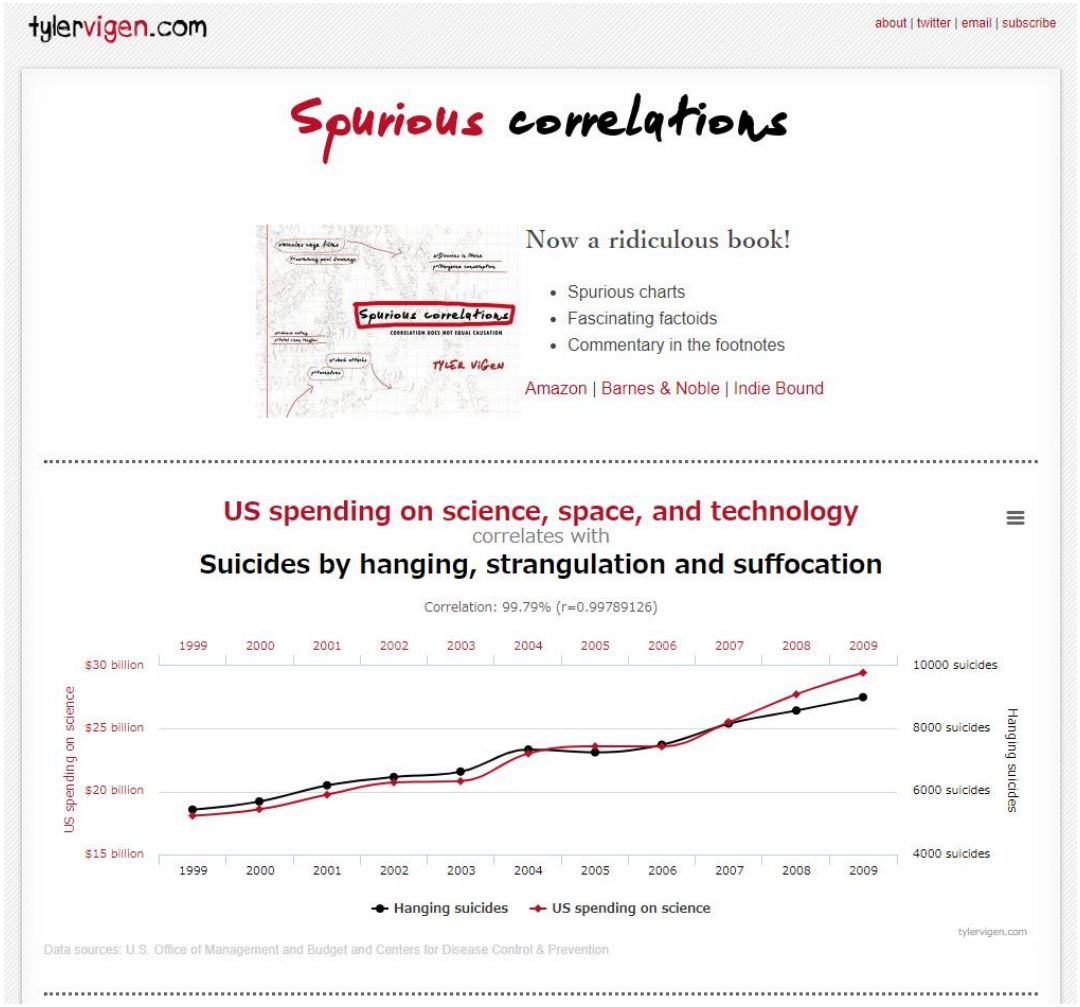
相関と因果

相関関係：
二つの事柄に関連性がある

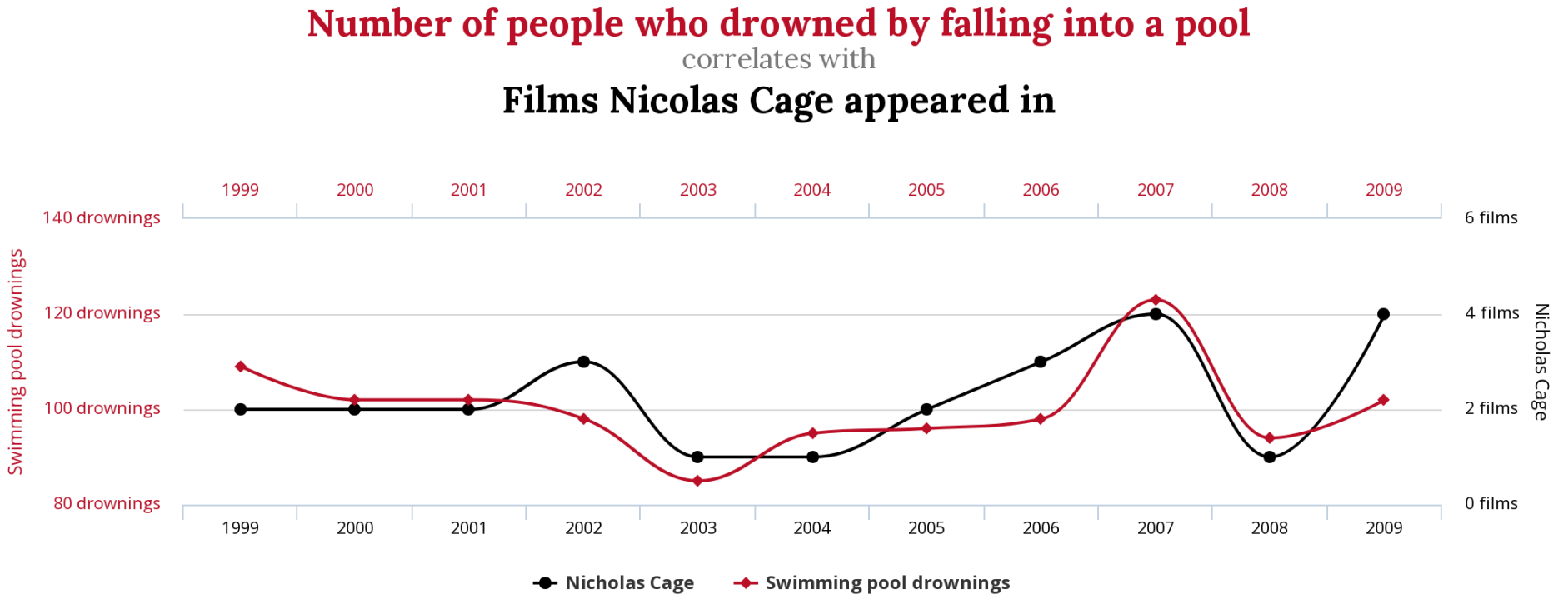
因果関係：
二つの事柄が、原因と結果の関係である

疑似相関

https://www.tylervigen.com/spurious-correlations



ニコラス・ケイジの映画出演本数と、プールでおぼれた人の数には、高い相関があるが、、、？



中室牧子, 津川友介著、ダイヤモンド社2017年

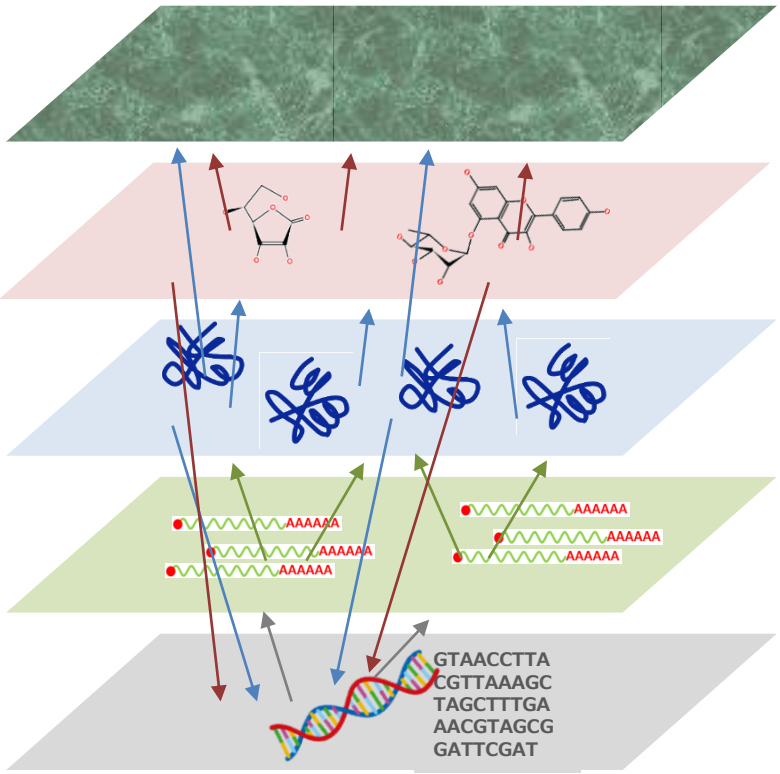
多変量解析

多変量データの例

- 大規模な疫学研究データ
- 生物等のオミクスデータ

など

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

多変量解析の目的

- データを要約して解釈しやすくする
- データに含まれる潜在的な因子を見つける
- 状況を判別したり、分類したりする
- 状況を予測する

さまざまな多変量解析

- 似ているものをグルーピングする

  クラスター解析

- データを要約する

  主成分分析

- 判別、分類、予測

  判別分析、PLS、PLS-DA、
  重回帰分析

など

主成分分析

学習目標

主成分分析について

- 概念を理解する
- 結果の解釈の仕方を理解する
- （Rによる計算をする）

主成分分析で扱うデータ

組織ごとの生体試料など

		対象				
		1	2	3	…	n
変数	X_1	X_{11}	X_{21}	X_{31}		X_{n1}
	X_2	X_{12}	X_{22}	X_{32}		X_{n2}
	X_3	X_{13}	X_{23}	X_{33}		X_{n3}
	…					
	X_m	X_{1m}	X_{2m}	X_{3m}		X_{nm}

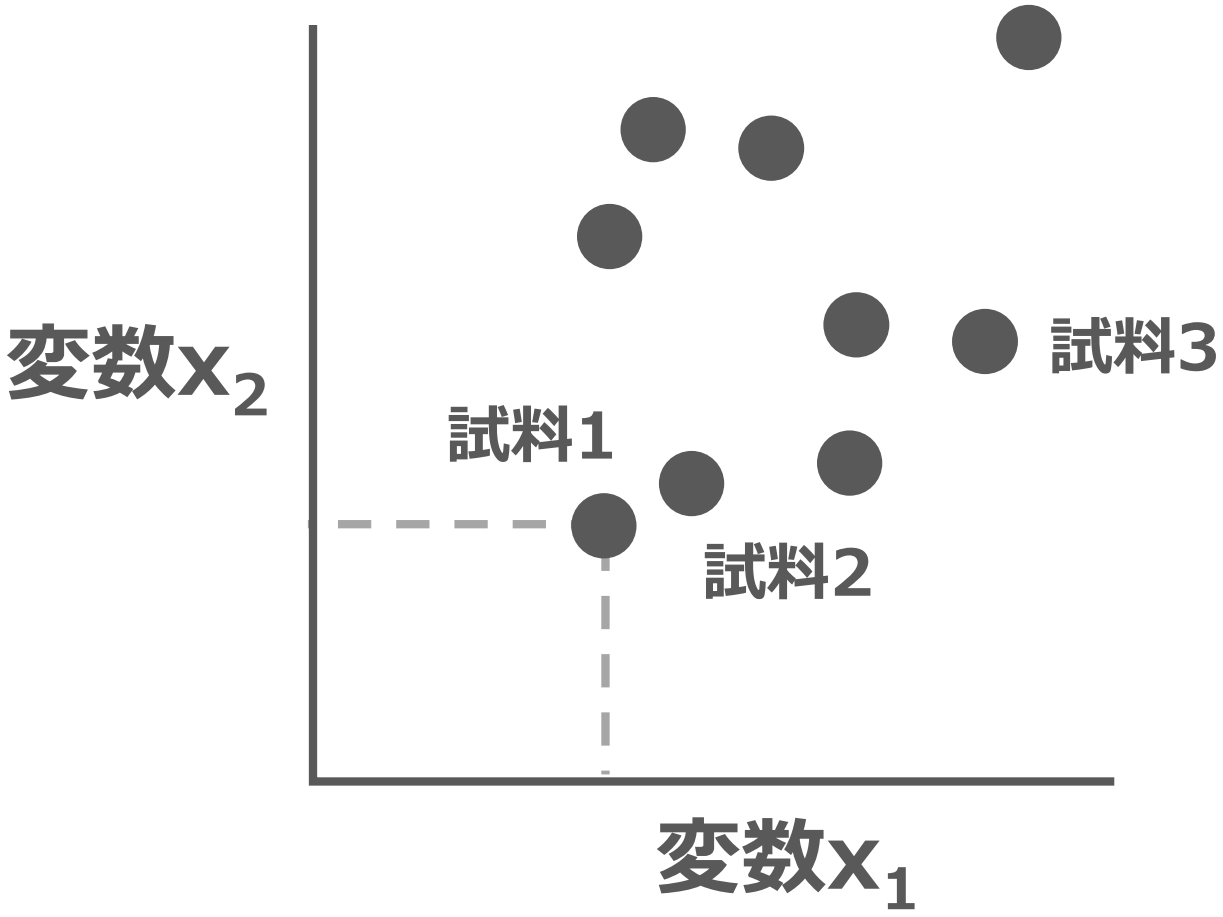
遺伝子など

説明変数，観測変数

遺伝子発現量など

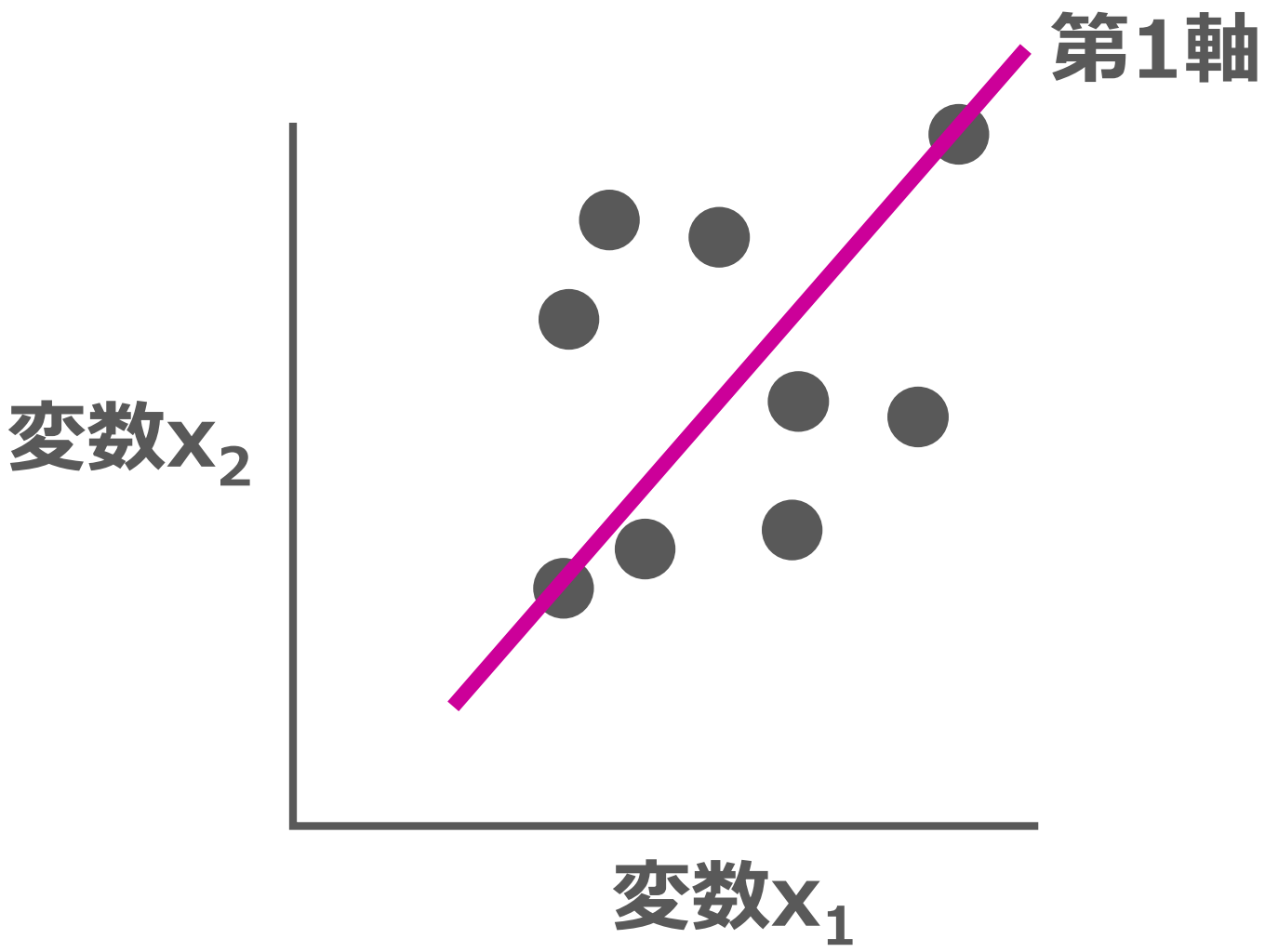
主成分分析のイメージ

①例えば変数が2個しかないとき、2次元の散布図に、試料ごとに変数をプロットできる



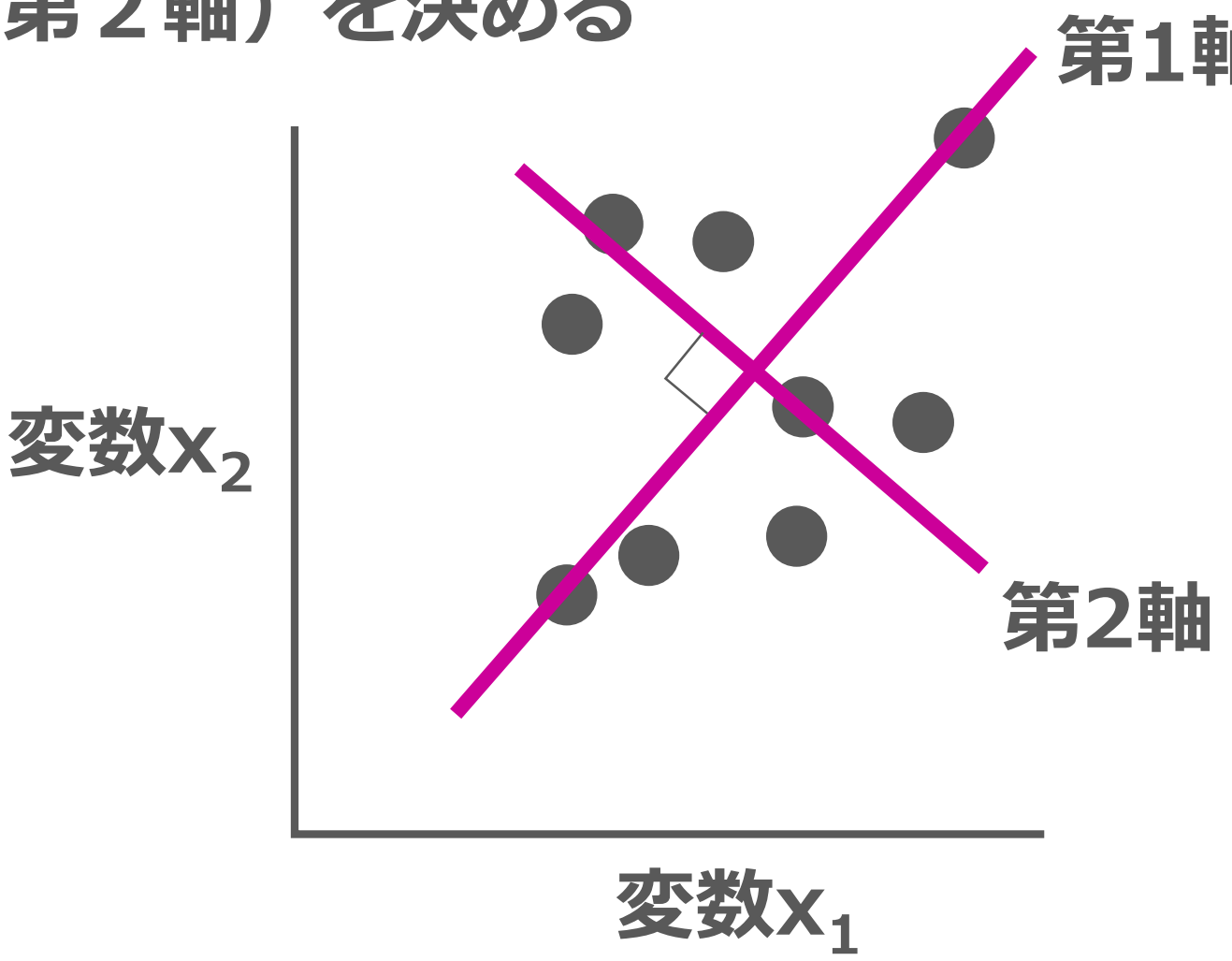
主成分分析のイメージ

②一番分散の大きい軸（第１軸）決める



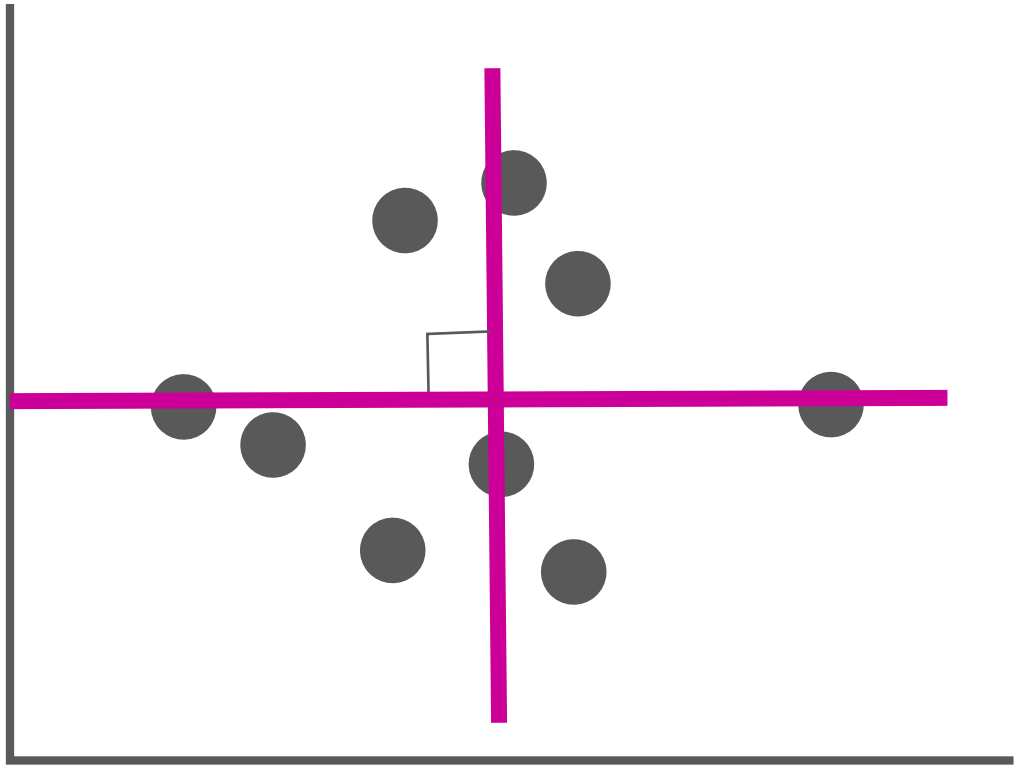
主成分分析のイメージ

③第1軸に直角に交わり、次に分散が大きい軸（第２軸）を決める



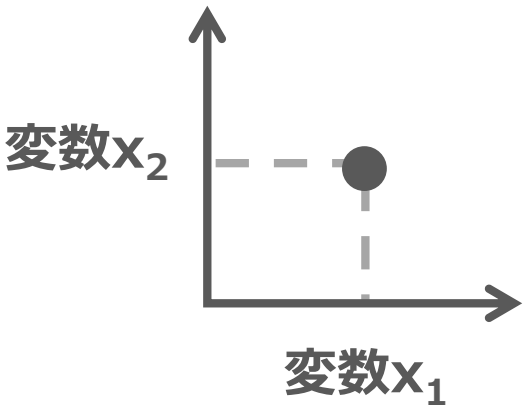
主成分分析のイメージ

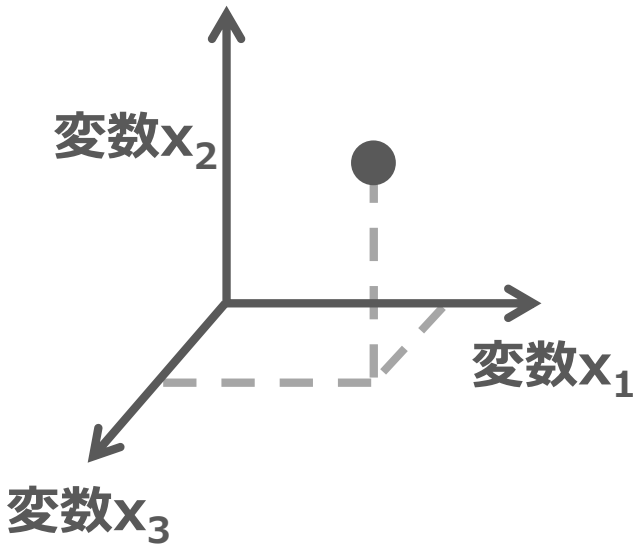
④第1軸がx軸、第2軸がy軸になるように、図を回転させた新たな図を作る



主成分分析のイメージ

m個の変数の値をm次元の図にプロットし、同様の計算を行うことが可能





描けないが計算上は可能

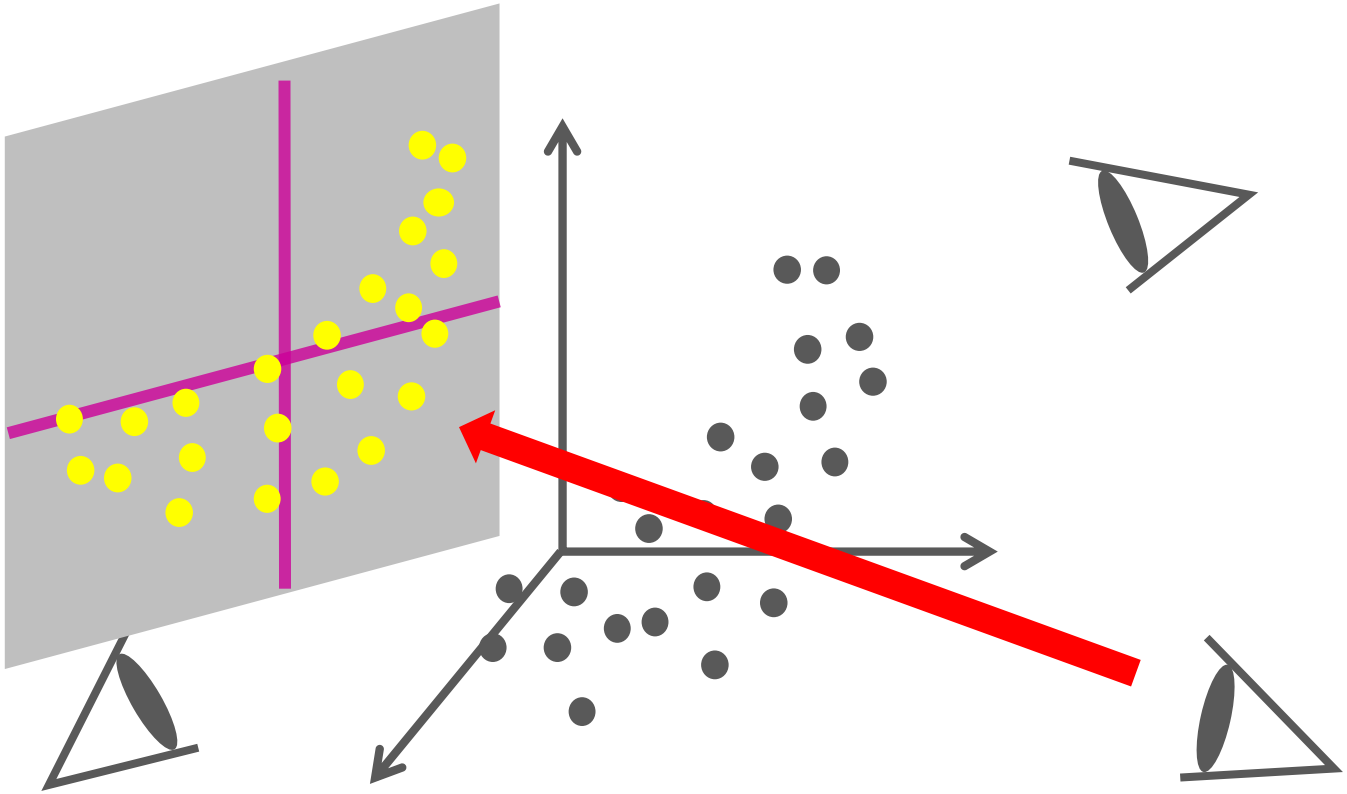
変数2個
2次元

変数3個
3次元

変数m個
m次元

主成分分析のイメージ

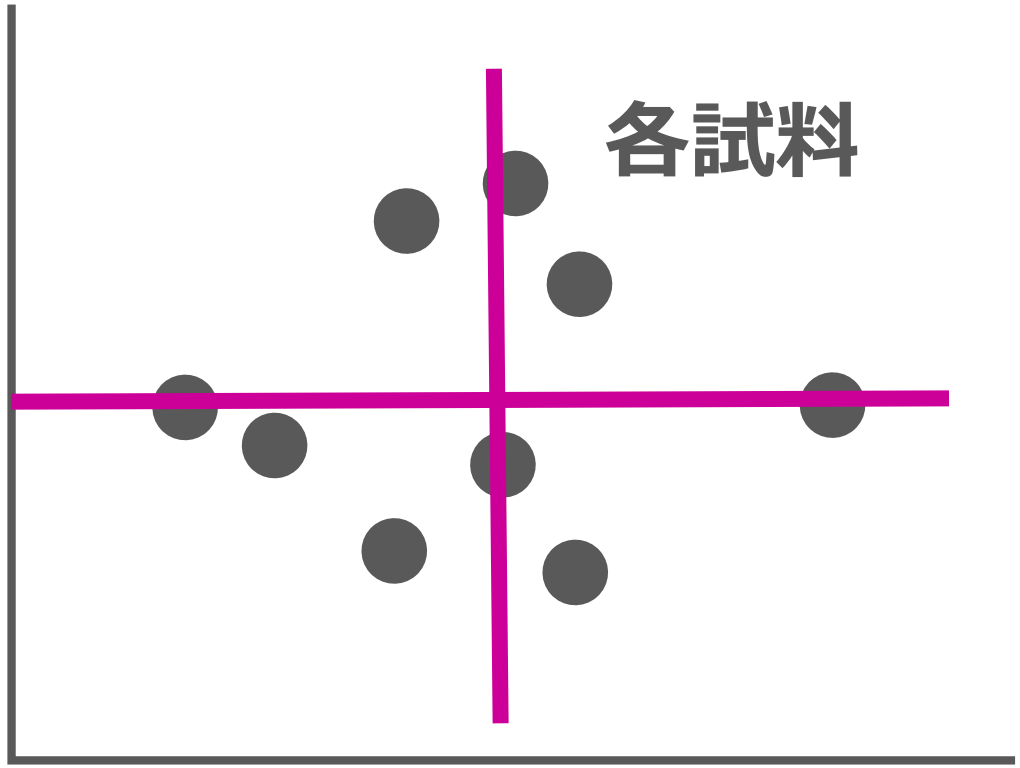
試料間の違い（特徴）が一番はっきりと見える方向から見た図が描ける



スコアプロット

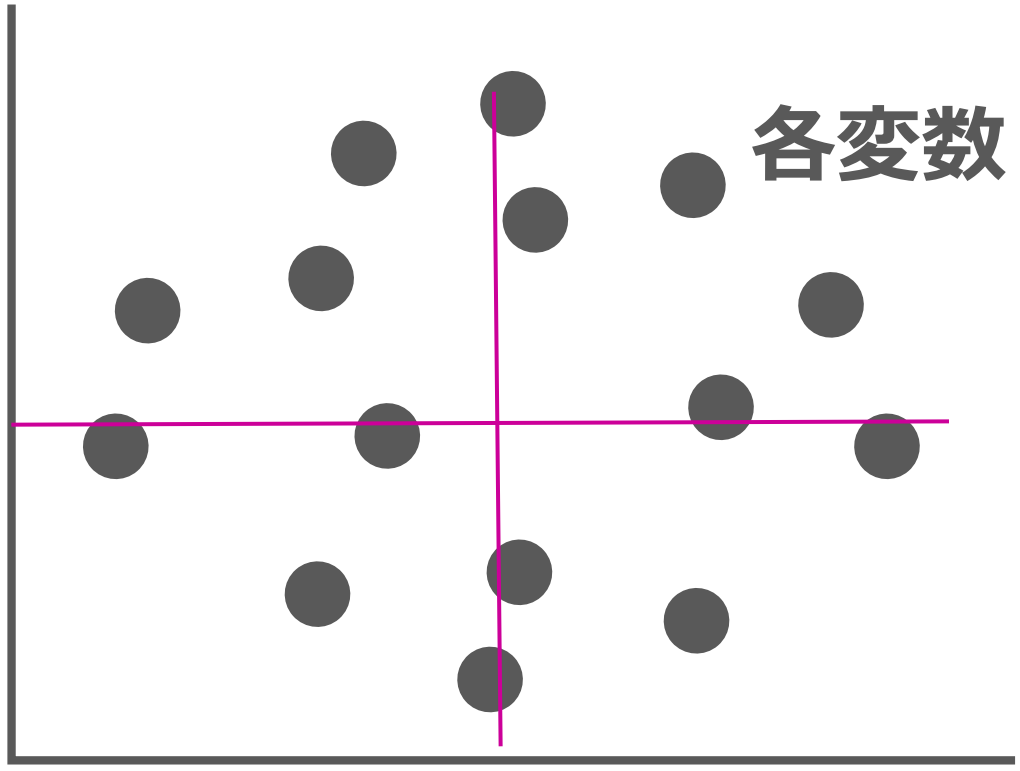
主成分軸に各試料を投影しなおした図

軸に示した%は**寄与率**と呼び、全体の分散のうち各主成分軸が説明する分散の比率を表す。第１主成分の寄与率が最も大きい。



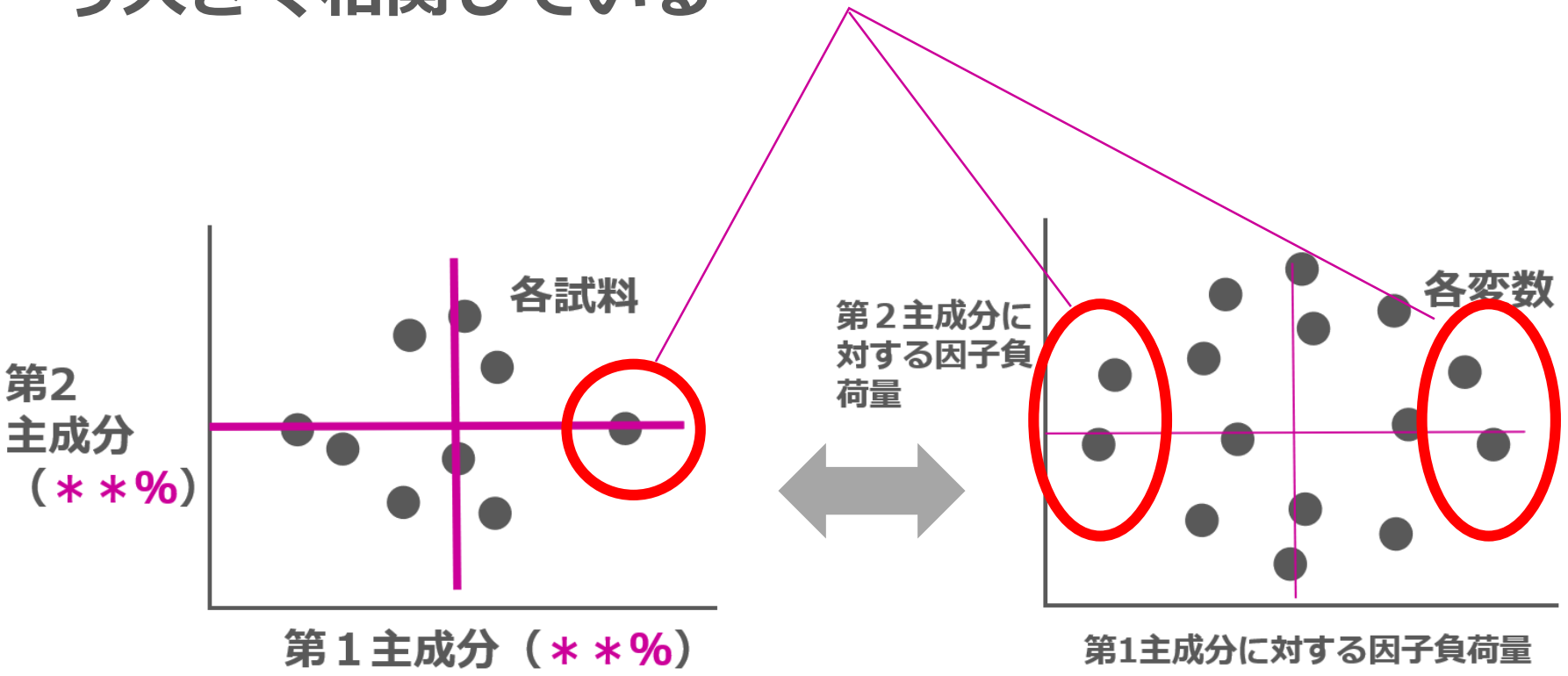
ローディングプロット

ローディングは、因子負荷量とも呼ばれ、各試料の主成分スコアと、変数の間の相関係数に相当する。
（厳密には、数値の前処理の条件などいくつか制約がある）



二つの図をセットで見る

この試料と他の試料との違いは、これらの変数がより大きく相関している



スコアプロット　　ローディングプロット

そのほかの多変量解析

さまざまな多変量解析

- 似ているものをグルーピングする

  クラスター解析

- データを要約する

  主成分分析

- 判別、分類、予測

  判別分析、PLS、PLS-DA、
  重回帰分析

など

PLS

Partial Least Squares

部分最小二乗

PLS-DA

Partial Least Squares-Discriminant Analysis

部分最小二乗-判別分析

PLS、PLS-DAで扱うデータ

目的変数が存在する

組織ごとの生体試料など

説明変数との関連を調べたい試料の分類や、試料の特徴量など
例）別途測定した、生理活性データなど

目的変数

		対象				
		1	2	3 …		n
変数	y_1	y_{11}	y_{21}	y_{31}		y_{n1}
	y_2	y_{12}	y_{22}	y_{32}		y_{n2}
	…					
	y_p	y_{1p}	y_{2p}	y_{3p}		y_{np}
変数	x_1	x_{11}	x_{21}	x_{31}		x_{n1}
	x_2	x_{12}	x_{22}	x_{32}		x_{n2}
	x_3	x_{13}	x_{23}	x_{33}		x_{n3}
	…					
	x_m	x_{1m}	x_{2m}	x_{3m}		x_{nm}

遺伝子など
説明変数, 観測変数

遺伝子発現量など

PLS、PLS-DAで得られる結果

- PCAと類似したスコアプロットとローディングプロットが得られる
- 目的変数（ｙ）を説明変数（x）で説明するためのモデルが構築される
- 目的変数を説明する変数重要度（VIP）が計算される

情報統計
第12回

2019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

自習

課題準備

おさらい

やったこと

- 統計的手法
- 記述統計
  - ✓平均値等の計算
  - ✓相関係数、回帰式
- 推測統計
  - ✓ 推定、仮説検定
- 多変量解析

- エクセル関数
- プログラミング
- Phython, R

統計って？

集団の状況を
数値で表したもの

目的：集団の〇〇を知りたい

統計学

- データを集める
- 解析する
- 解釈する

ための方法論

結果：集団の〇〇がわかった！

結論を言う

重要！

統計的結論から、設定した目的に
対する結論を導くことが最も重要。

発表会の
テンプレート

表紙　1枚

- タイトル
- 名前
- 報告日など

背景と目的　１～枚

- 何に疑問を持ち、どんな目的のためにこの課題を行ったか？
- その疑問に至った背景

方法のページ 1～枚

- どんなデータ、どんな統計的手法を使って実施したか。

だれもが追試、検証できるように

結果のページ　1～枚

- どんな結果が得られたか
- そこから言えることは何か

結果に基づいて得られた情報について述べる

考察のページ　1～枚

- 結果を総合して、目的に対してどんな結論が得られたか

最初に掲げた疑問に対する答えや、得られた結果の価値について述べる

（将来展望のページ　1～枚）
もしあれば

- 今後こんなデータを集めれば…
- 今後こんな統計的手法を適用すれば…

もっとこんなことがわかるだろう、など

未来に対する夢を述べる

よいスライドの作り方

田中佐代子著、講談社2013年

自習

課題準備