多変量解析とは何か

単変量解析

- 平均　　1,2,3,4,5の平均値は、(1+2+3+4+5)/5=3

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- 分散　　1,2,3,4,5の(不偏)分散は2.5

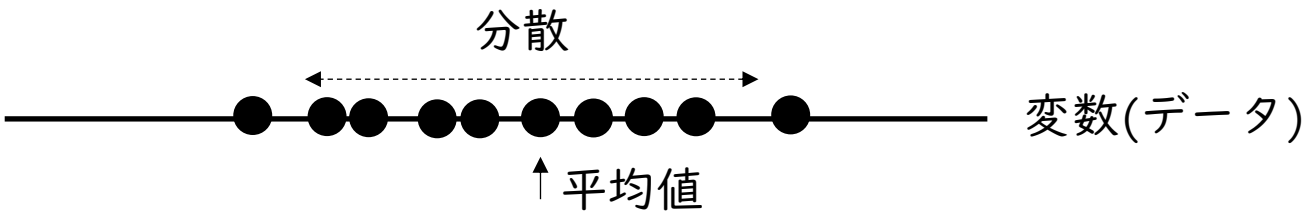
(不偏)分散　$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$

- 標準偏差　　1,2,3,4,5の標準偏差は1.58

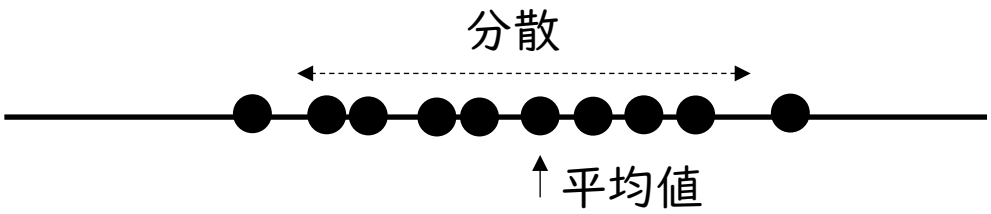
(不偏)標準偏差　$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$

単変量解析と多変量解析の違い

- 単変量解析
  - 1つの変数、または変数が複数ある場合でも、
    それぞれの変数の平均、分散などを考える

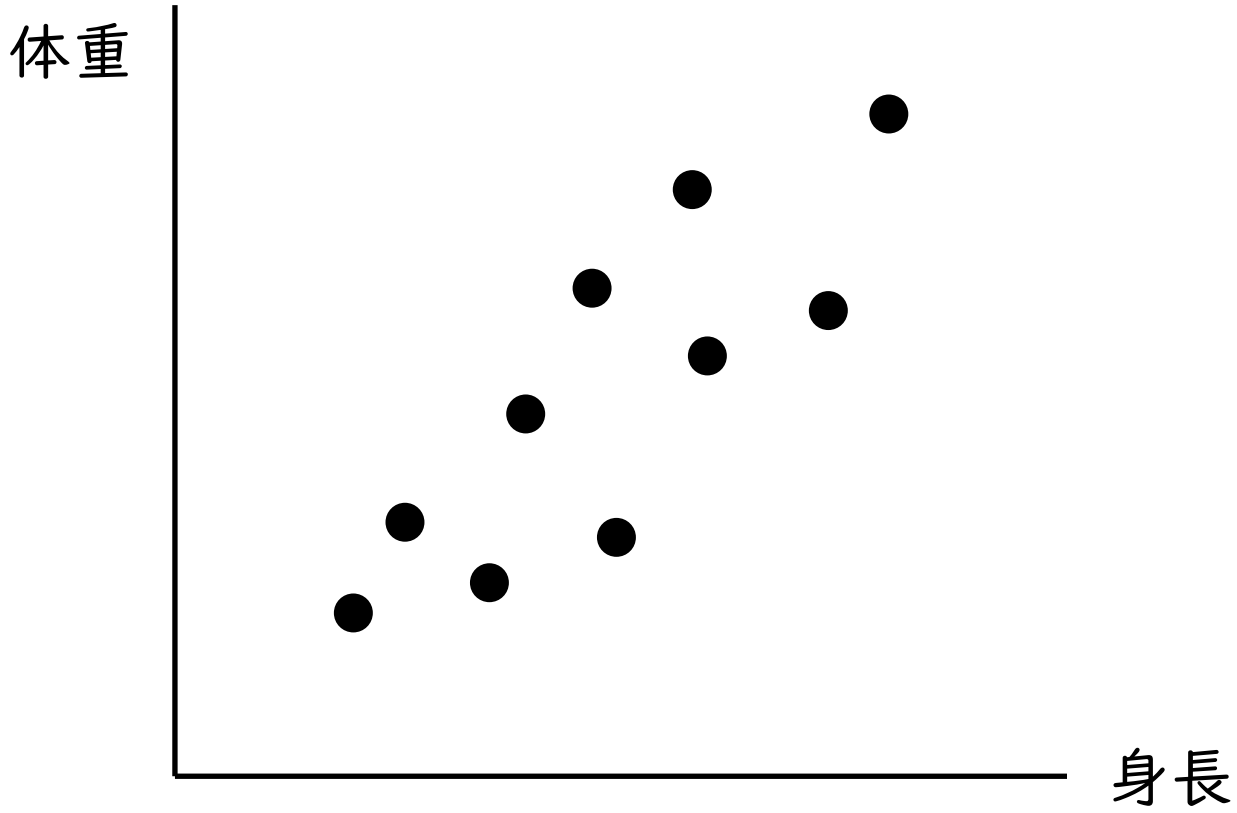


- 多変量解析
  - 多変量解析は、複数の変数をまとめた1つの変数である合成変数の統計(平均や分散など)を考える
  - 主成分分析は、複数の変数をまとめた1つの変数である主成分スコアの分散を考える



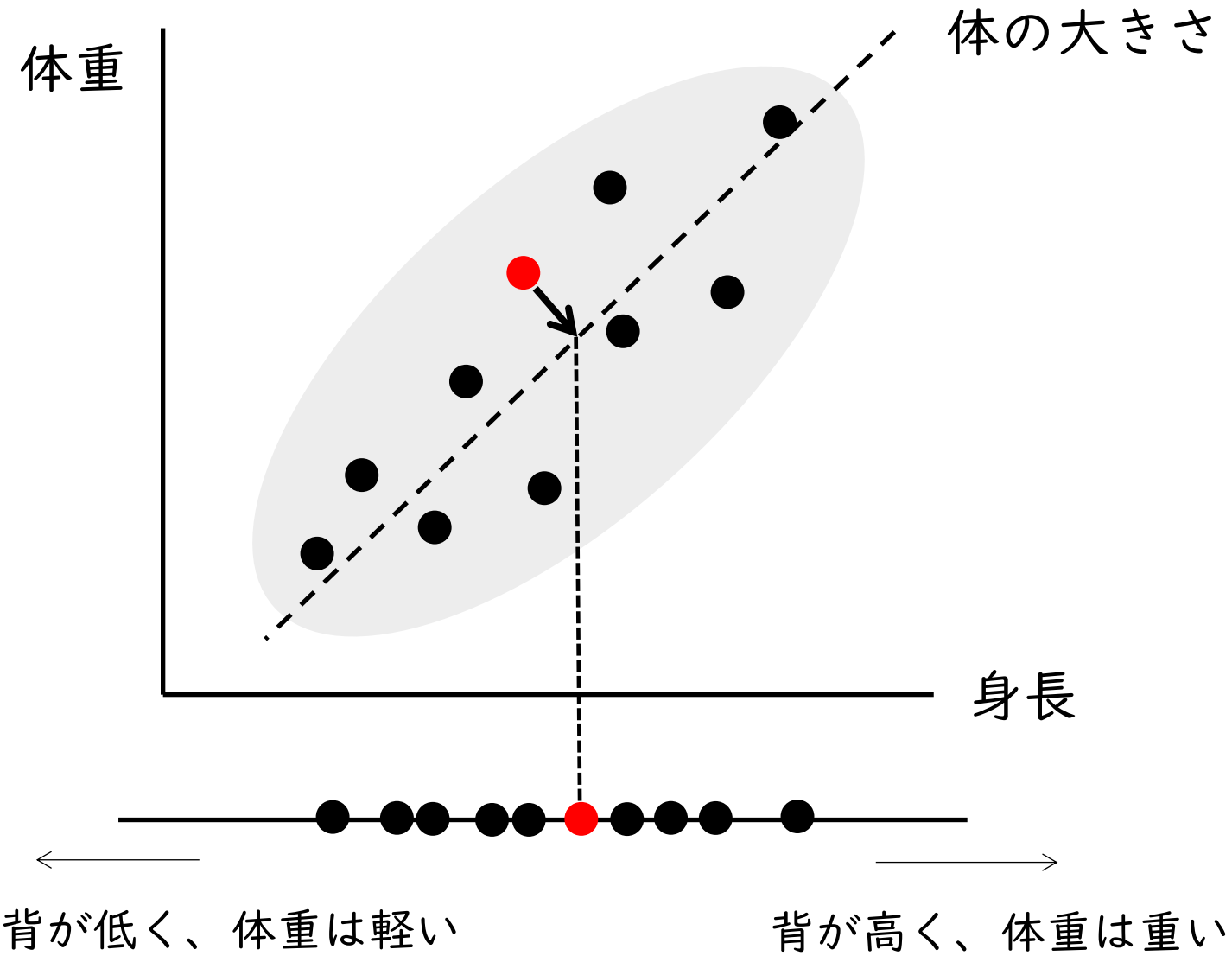
合成変数
主成分スコア

合成変数とは？　身長と体重から新たな変数を作る

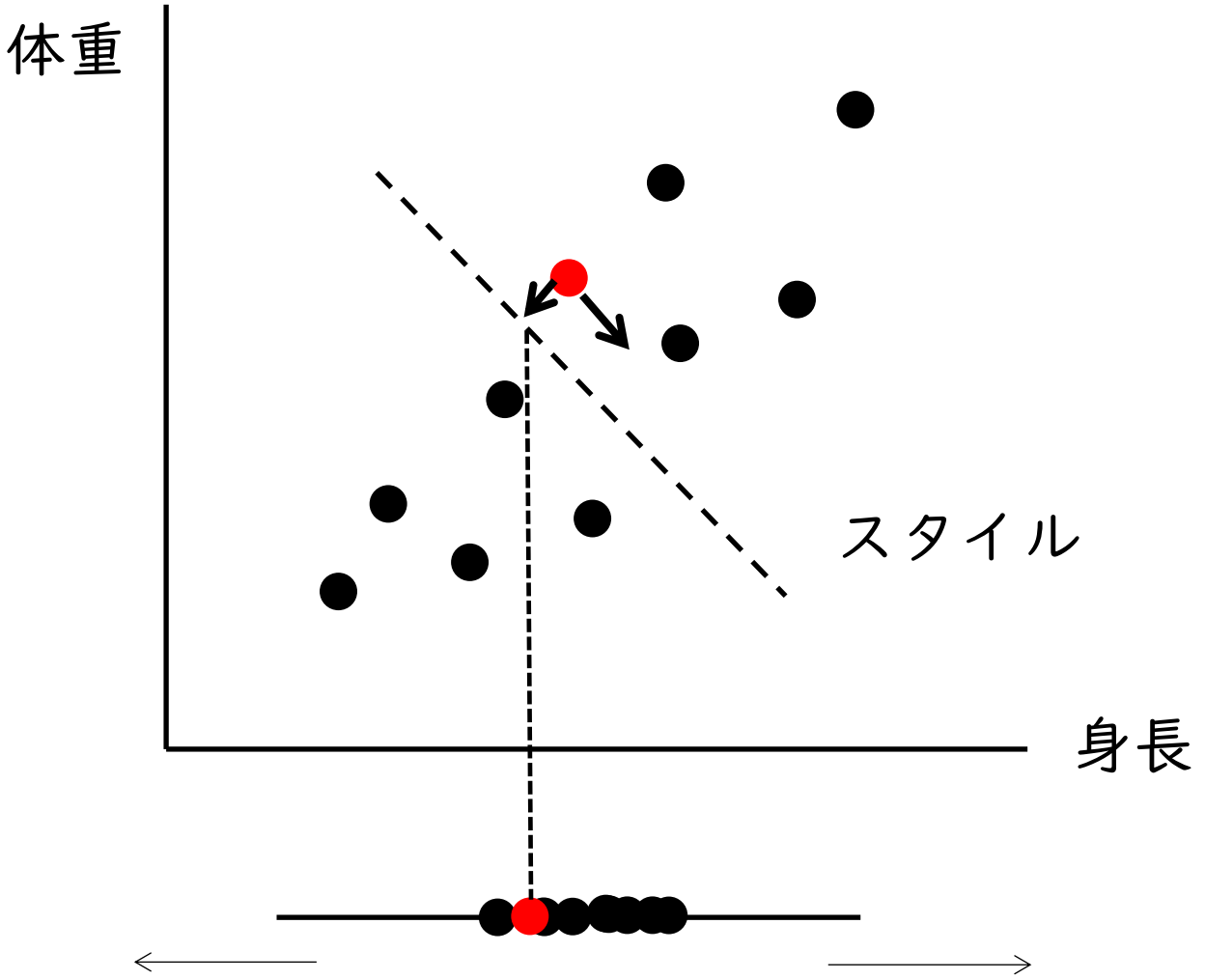


身長と体重の2つの変数から、新しい変数を作るとすれば、
どのようなものが考えられるでしょうか？

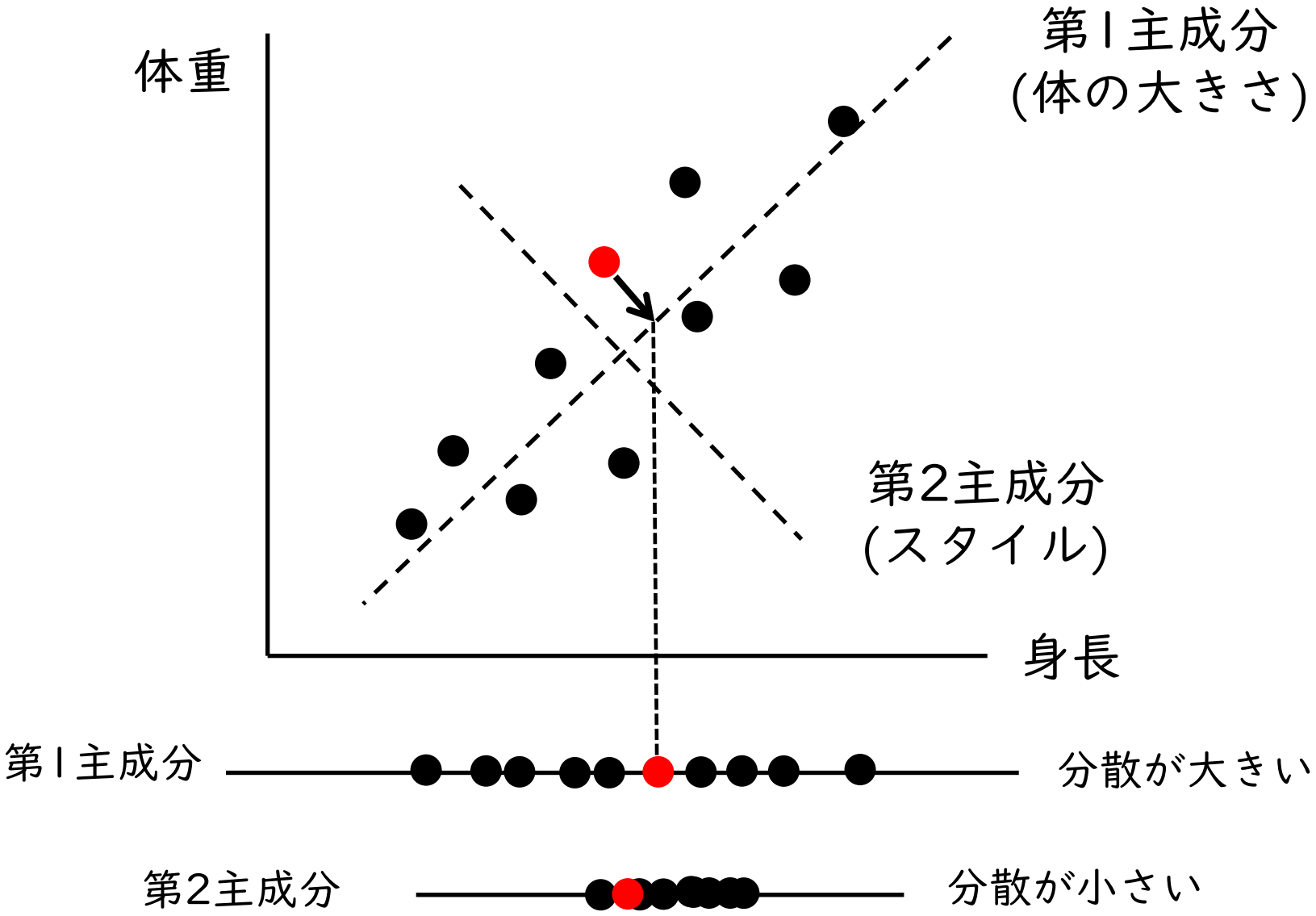
元の変数(身長、体重)から新たな変数(体の大きさ)を作る



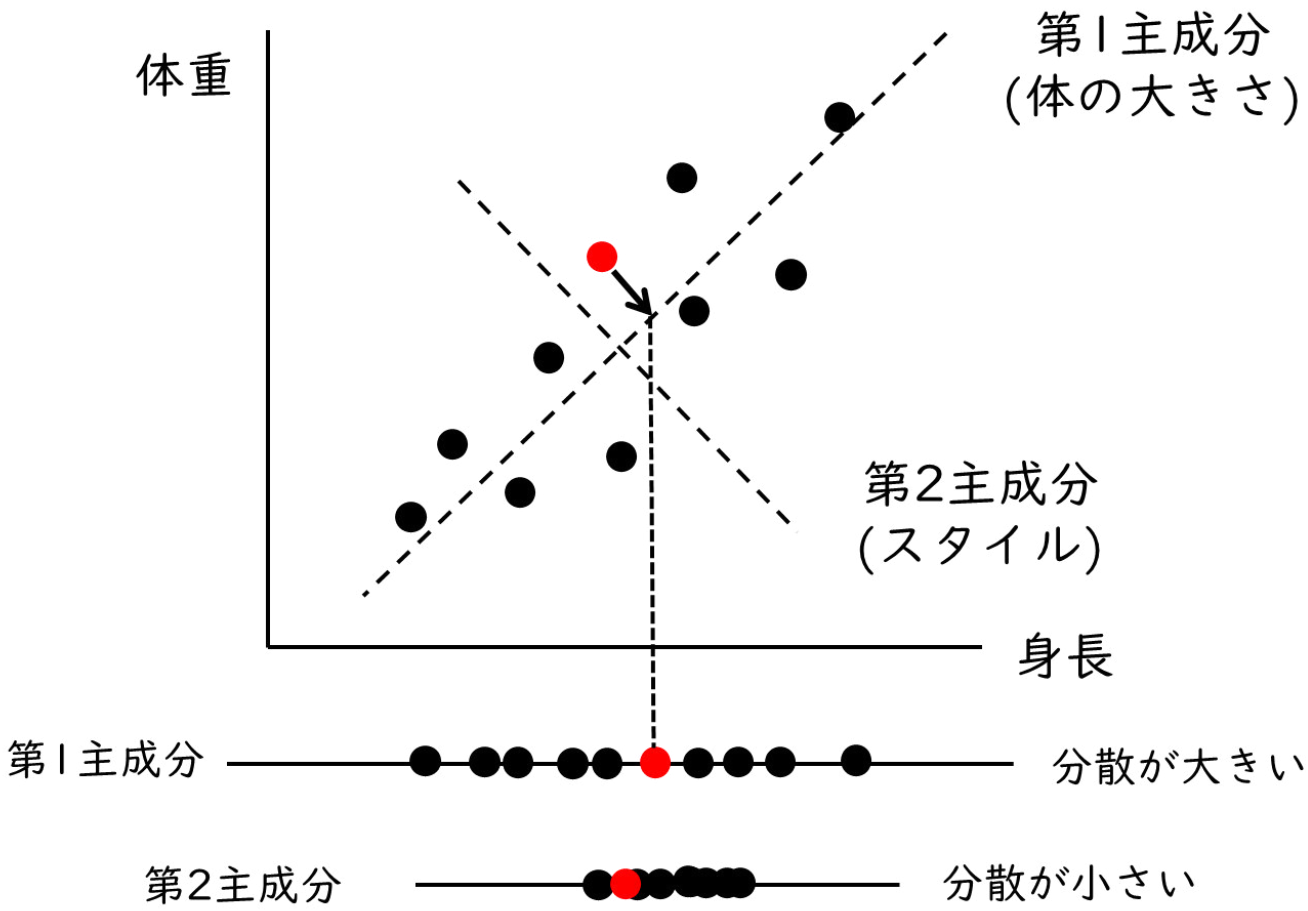
元の変数(身長、体重)から新たな変数(スタイル)を作る



主成分分析との関係

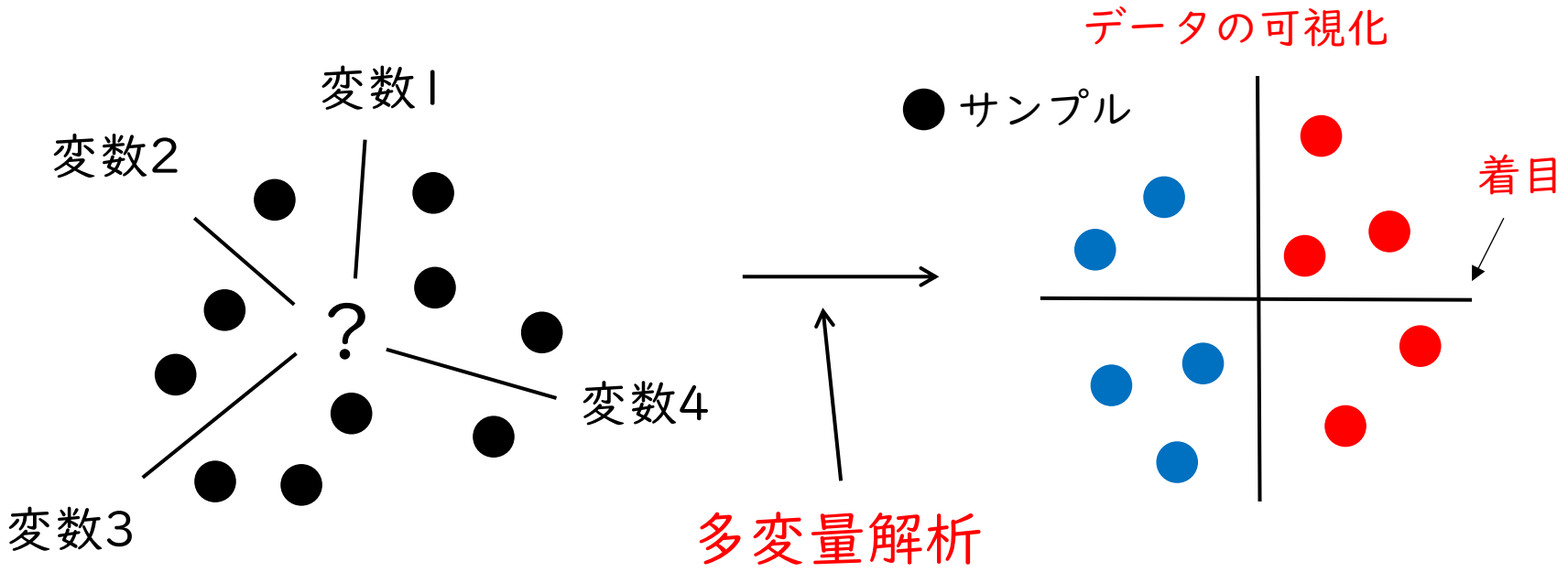


身長・体重と実際のデータとの違い



実際のデータでは変数の組み合わせが多く、身長と体重を
体の大きさで代表するというように、直感的に考えることは難しい

多変量解析を用いたデータの可視化



視覚的に表現できない
多次元データ

視覚的に表現できる
2次元データ

データを可視化し、群や時系列の情報など
表現型を表す（主）成分を見つける

主成分係数を用いた重要な変数の選び方

身長と体重の例で言えば、

$$\begin{matrix} \text{体の大きさ} \\ \text{(合成変数)} \end{matrix} \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix} = \begin{matrix} \text{身長} \\ \downarrow \end{matrix} \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} w_1 + \begin{matrix} \text{体重} \\ \downarrow \end{matrix} \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} w_2$$

実際のデータでは、

$$\begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix} = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} w_1 + \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} w_2 + \cdots + \begin{bmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{bmatrix} w_p$$

w：重み・主成分係数・固有ベクトル

例えば係数w_2の値が大きければ、

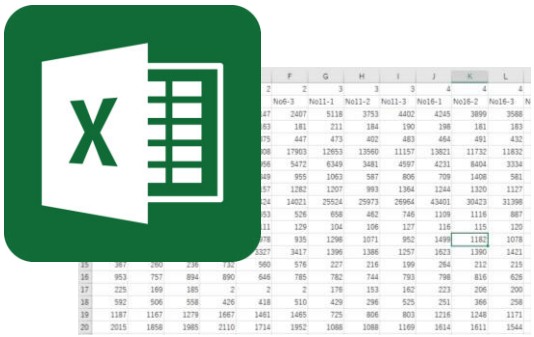
第1(主)成分と元のデータの2番目の変数との関連が高い

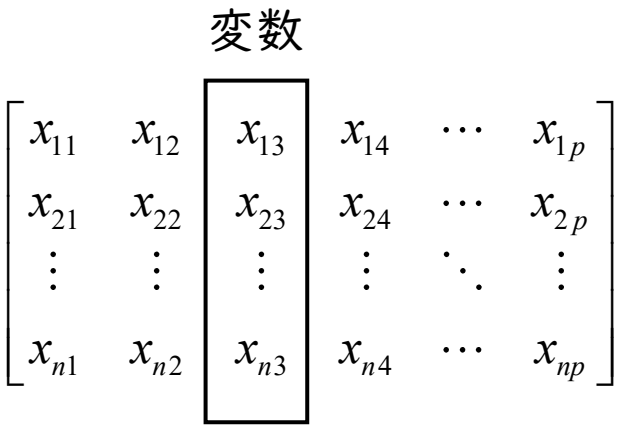
データを可視化し、着目した(主)成分と関連の高い変数を選ぶ

多変量解析

- 多変量解析とは
  - 統計学において、複数の独立変数(説明変数)からなる多変量データを統計的に扱う手法。主成分分析、因子分析、クラスター分析などがある。

“https://ja.wikipedia.org/wiki/多変量解析”より一部改変





観測された変数の数が1つではなく複数ある

(測定項目、検査項目など)

主成分分析

- 主成分分析とは
  - 主成分分析は1901年にカール・ピアソンによって導入された。ピアソンとは独立に1930年代にハロルド・ホテリングよっても導入され、主成分分析 (principal component analysis、PCA) と呼ばれるようになった。

“https://ja.wikipedia.org/wiki/主成分分析”より一部改変

1890年の米国国勢調査のデータ処理で初めて使用されたタビュレーティングマシン

1937年から1942年にかけて開発され、世界最初のコンピューターとも言われるABCマシン(の復元)

1957年にNASAで使用されるIBM704(大型コンピュータ)

1968年に統計ソフトウェアのSPSS Iがリリース

データの集計 → 連立方程式の計算 → 多変量解析

Wikipediaより一部改変

多変量解析といっても色々な方法がある
どの方法を用いるべきか？

- 分野によって、主に用いられる統計解析手法は異なる
  - 歴史的な背景で決まっていることが多い
    - その分野の論文を見ると、概ね決まって同じ解析手法が用いられている
  - その分野で最も良く用いられている方法を学ぶのが効率的

- 本セミナーでは主成分分析とPLSを中心に説明
  - 化学・生物(ケモメトリックス)分野での多変量解析としては、主成分分析とPartial Least Squares(PLS)が最も良く用いられている

- その他の手法
  - 心理学におけるパーソナリティの特性論的研究など、心理尺度の研究手法としては因子分析が主に用いられている

“https://ja.wikipedia.org/wiki/因子分析”を一部改変

ケモメトリックス(計量化学)は、化学・生物分野での多変量解析

- ケモメトリックス、計量化学とは
  - 数理科学、統計学、機械学習、パターン認識、データマイニングなどの手法により、（広義の）化学分野における諸問題を解決しようとする分野である
  - 計量化学では実験で得られた（多くの場合）大量のデータに対し、次元の圧縮・視覚化・回帰・判別・分類などを行うことによって、実験結果の解釈に重要な情報を提供することを目的とする
  - 1960年代後半から1970年代前半にかけて、コワルスキ 、ウォルドらの活躍によってであり、ケモメトリックス という用語は、ウォルドによって初めて用いられた
    - SPSS 1がリリースされたのと同時期であることから、この時代から実際のデータの解析として、多変量解析が一般に広く使われるようになったのでは？

“https://ja.wikipedia.org/wiki/計量化学” より一部改変

メタボロミクス

- メタボロミクス、メタボローム解析とは
  - メタボロミクスとは、アミノ酸、有機酸、脂肪酸、糖など、分子量が約1000以下の低分子の代謝物(メタボライト)を網羅的に解析すること
- サンプル
  - 動物の臓器、微生物の培養サンプル、人の血液、尿など様々な種類のサンプルを、質量分析装置で測定を行い、メタボロームデータを取得する
- 応用範囲
  - 薬剤の作用機序の解明や、疾患バイオマーカーの探索、食品の機能性研究など様々な研究に利用されている
- 多変量解析の利用
  - 主成分分析やPLSを用いたデータの可視化とローディングを用いて重要な代謝物を選ぶ

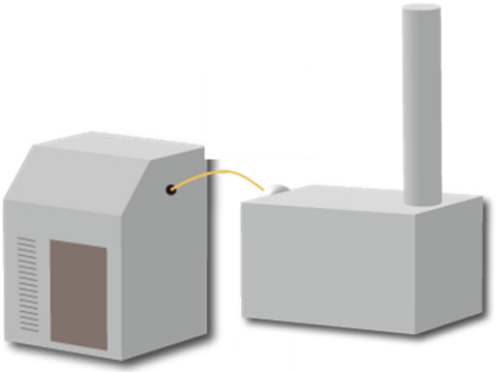
メタボロミクスの研究例

絶食マウス肝臓のメタボロームデータ

通常飼育
(n=5)

12時間絶食
(n=5)

肝臓サンプル



キャピラリー電気泳動
飛行時間型質量分析を用いた
メタボローム解析

10サンプル

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & x_{24} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \cdots & x_{np} \end{bmatrix}$$

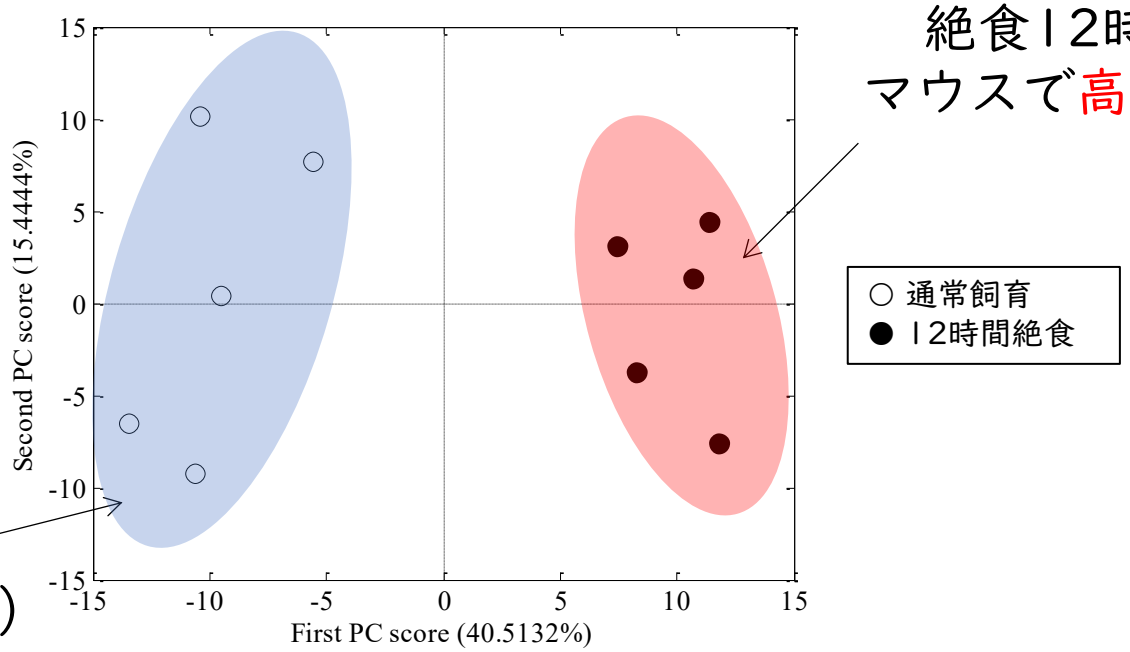
282変数(代謝物質)

主成分分析

Yamamoto et al., "Statistical hypothesis testing of factor loading in principal component analysis and its application to metabolite set enrichment analysis" BMC Bioinformatics, (2014) 15(1):51.

主成分分析を用いたデータの可視化の例

絶食12時間
マウスで高値(+)



通常飼育
マウスで低値(−)

身長・体重での合成変数「体の大きさ」と同様に合成変数を計算

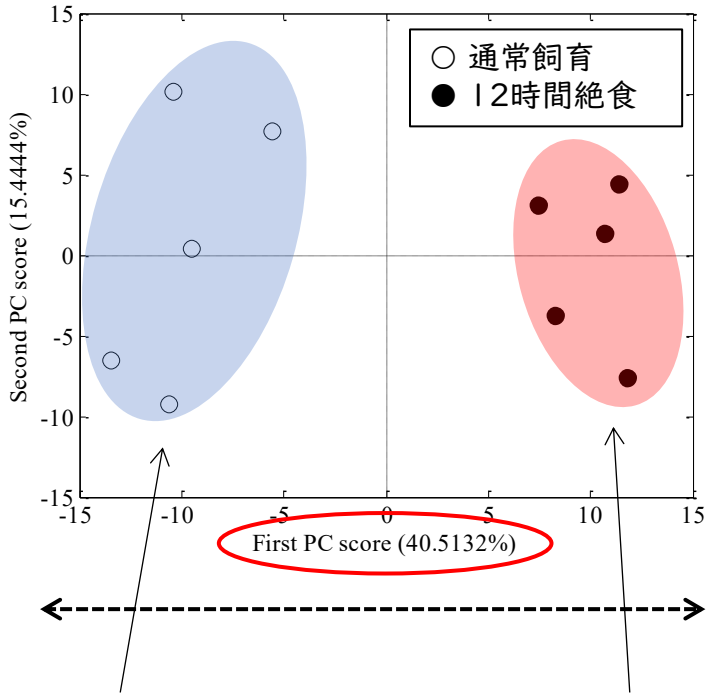
↓

実際のデータでは変数の組み合わせが多く、身長と体重を体の大きさで代表するというように、直感的に考えることは難しい

データを可視化した後に着目する主成分スコア(合成変数)を決める

絶食マウスでの主成分分析の解析例

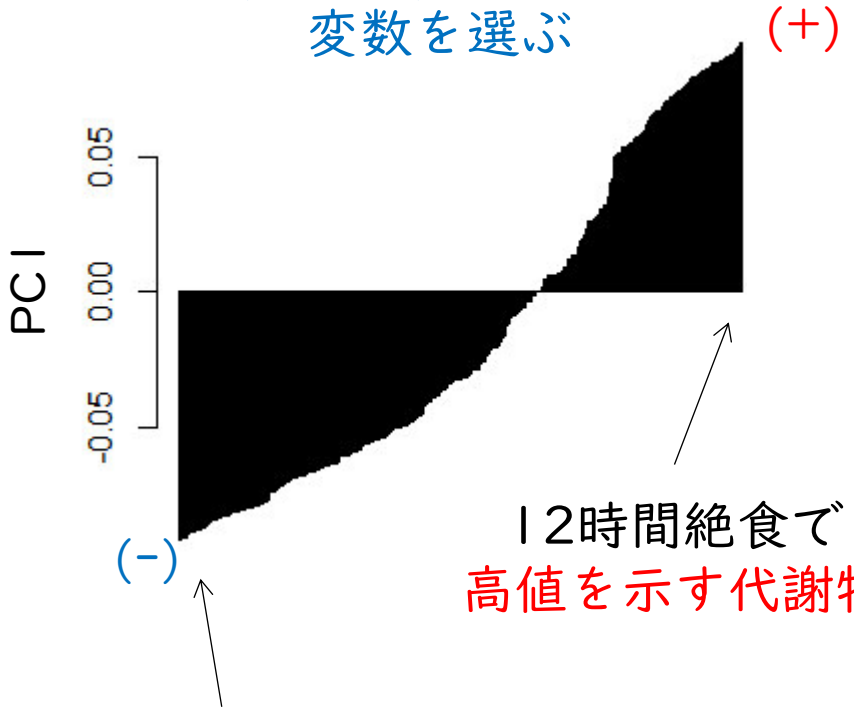
データの可視化



通常飼育
マウスで低値(−)

絶食12時間
マウスで高値(+)

重み(主成分)係数を用いて
変数を選ぶ



12時間絶食で
高値を示す代謝物

12時間絶食で
低値を示す代謝物

合成変数に対する主成分係数wから変数を選ぶ

⇒ 実際は、主成分負荷量を用いて重要な変数を選ぶ