第3章

情報とデータサイエンス

本単元の学習内容 106
学習11 データと関係データベース 110
学習12 大量のデータの収集と整理・整形 118
学習13 重回帰分析とモデルの決定 126
学習14 主成分分析による次元削減 136
学習15 分類による予測 144
学習16 クラスタリングによる分類 152
学習17 ニューラルネットワークとその仕組み 160
学習18 テキストマイニングと画像認識 168
全体を通じた学習活動の進め方 176

第3章 情報とデータサイエンス

本単元の学習内容 【学習内容の全体像】

3 情報とデータサイエンス

ア 大量のデータの扱いとデータサイエンスが社会に果たす役割

- **1 大量のデータとデータベース**
  ビッグデータ，関係データベース
- **2 データサイエンスが社会に果たす役割**
  データの信憑性と信頼性
- **3 データの収集・整理・整形**
  データクリーニング，ワイドフォーマット，ロングフォーマット

イ データのモデリングとその表現と解釈

- **1 回帰**
  重回帰分析，交互作用項，残差分析，モデルの評価
- **2 主成分分析**
  主成分，特徴量，次元削減
- **3 分類**
  k- 近傍法，分類木，モデルの評価
- **4 クラスタリング**
  k- 平均法，デンドログラム

ウ データの分析と評価

- **1 機械学習とその評価**
  訓練データ，テストデータ，交差検証
- **2 機械学習や人工知能の応用**
  深層学習，強化学習，画像認識，テキストマイニング
- **3 これからの社会に求められること**
  コンピュータと人間との共存，データ分析の将来

全体

情報の科学的な見方・考え方を働かせて，問題を明確にし，分析方針を立て，社会の様々なデータ，情報システムや情報通信ネットワークに接続された情報機器により生産されているデータについて，整理，整形，分析などを行う。また，その結果を考察する学習活動を通して，社会や身近な生活の中でデータサイエンスに関する多様な知識や技術を用いて，人工知能による画像認識，自動翻訳など，機械学習を活用した様々な製品やサービスが開発されたり，新たな知見が生み出されたりしていることを理解するようにする。更に，不確実な事象を予測するなどの問題発見・解決を行うために，データの収集，整理，整形，モデル化，可視化，分析，評価，実行，効果検証など各過程における方法を理解し，必要な技能を身に付け，データに基づいて科学的に考えることにより問題解決に取り組む力を養う。ここで学ぶ内容は，「数学Ｂ」の(2)「統計的な推測」との関連が深いため，地域や学校の実態及び生徒の状況等に応じて教育課程を工夫するなど相互の内容の関連を図ることも考えられる。

学習目標

- 多様かつ大量のデータの存在やデータ活用の有用性，データサイエンスが社会に果たす役割について理解し，目的に応じた適切なデータの収集や整理，整形について理解し技能を身に付けるとともに，目的に応じて，適切なデータを収集し，整理し，整形する。
- データに基づく現象のモデル化やデータの処理を行い解釈・表現する方法について理解し技能を身に付けるとともに，将来の現象を予測したり，複数の現象間の関連を明らかにしたりするために適切なモデル化や処理，解釈・表現を行う。
- データ処理の結果を基にモデルを評価することの意義とその方法について理解し技能を身に付けるとともに，モデルやデータ処理の結果を評価し，モデル化や処理，解釈・表現の方法を改善する。

本単元の取扱い

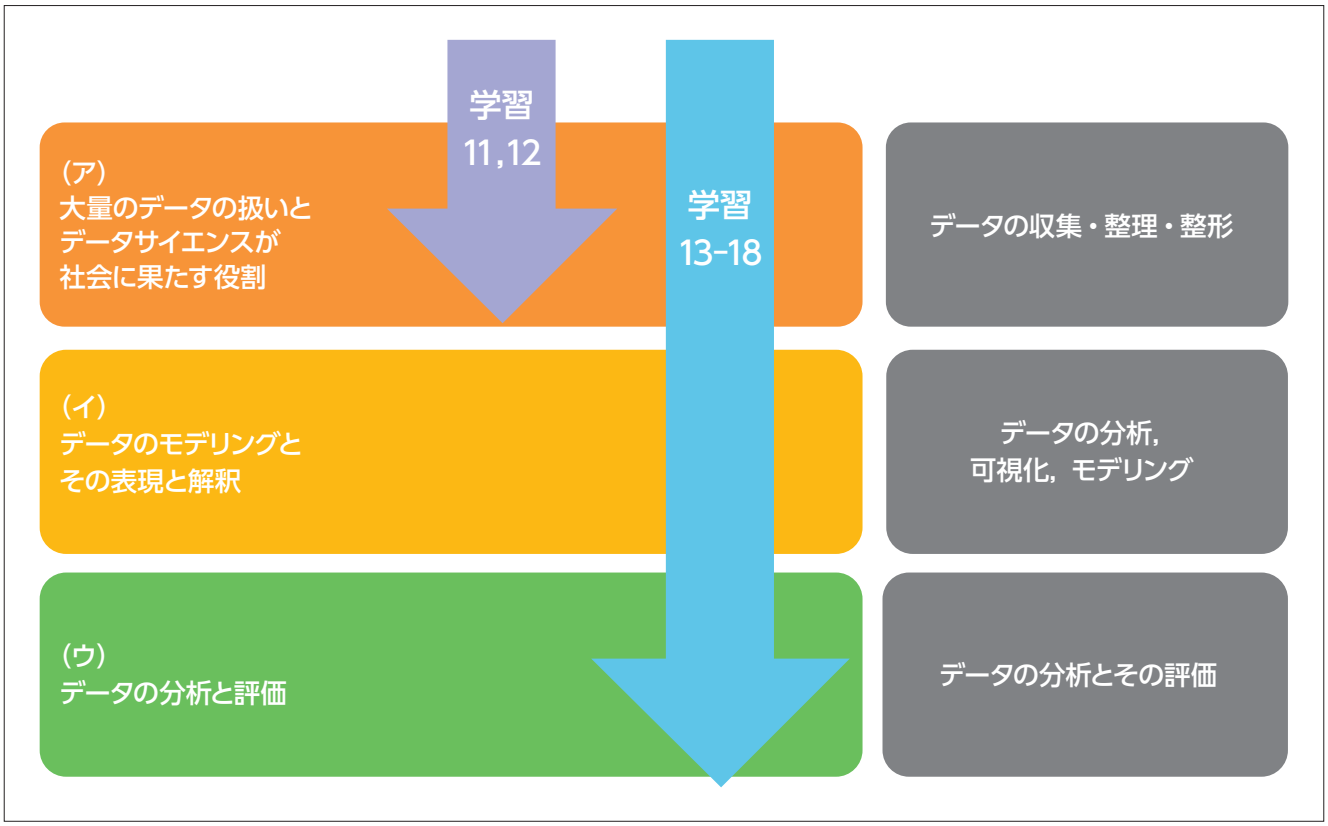
- データサイエンスによる人の生活の変化について扱う。
- 現実のデータの活用に配慮する。
- モデル化や処理，解釈・表現の方法の改善は，これらを行った結果を受けて行うようにする。

情報Ⅰの学習内容との関連

- モデル化及びプログラミングについては，(3)「コンピュータとプログラミング」，データの種類や特性及び活用については，(4)「情報通信ネットワークとデータの活用」で学習する内容と関連付けて扱う。

1 学習内容について

この章での学習は，学習指導要領解説にある(ア)(イ)(ウ)を統合したものになっている。これは，データを分析する過程が，(ア)(イ)(ウ)の一連の流れを必要とするからである。そのため，学習11と学習12に関しては，(ア)の内容を中心として書いてあるが，それ以外の学習に関しては，全ての要素を含み，データ分析手法の違いに着目して書いている。



図表1　第3章の学習とデータ分析のプロセスの関係

2 この章の学習について

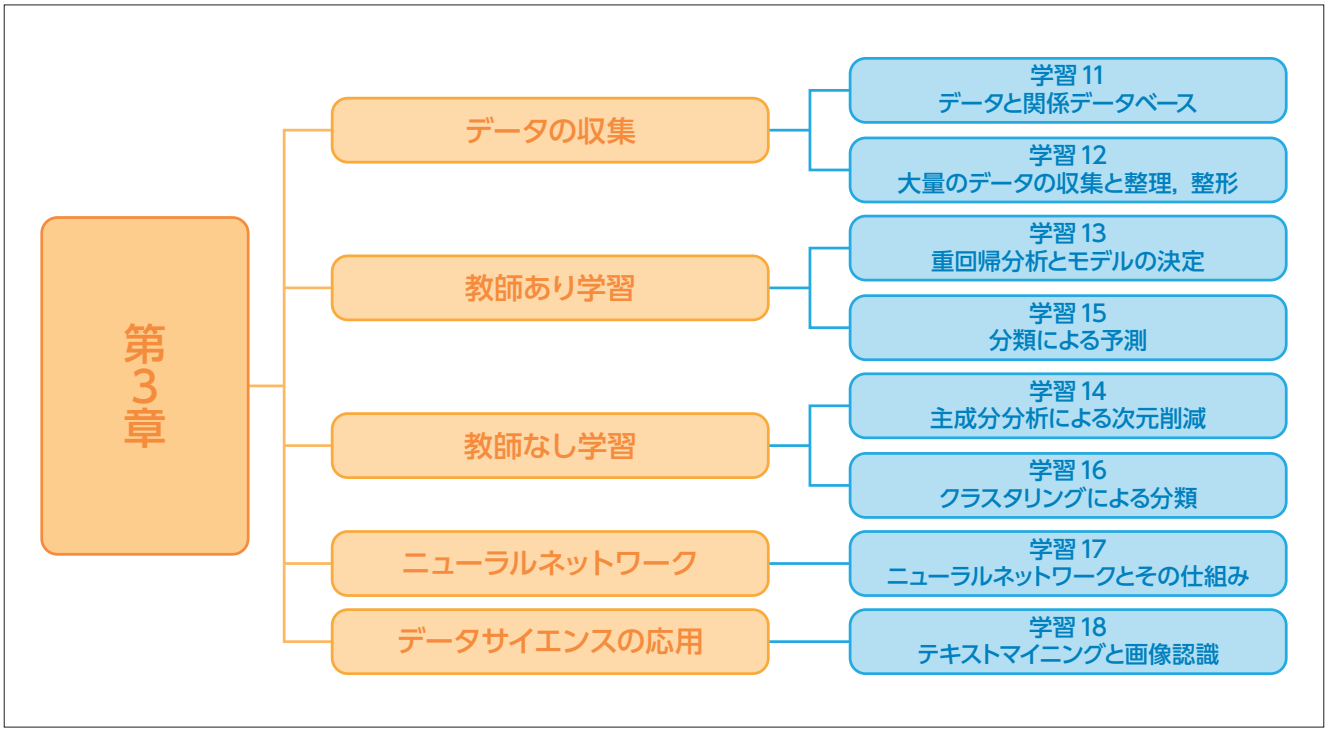
この章の学習については，Webスクレイピング，データクリーニングをはじめとして，機械学習やニューラルネットワークの基礎的な事項が学べるように配慮している。

「情報Ⅰ」でも学んだように，Webコンテンツや Webに掲載されたデータを取り込むことをWebスクレイピングといい，そのデータを分析，処理するソフトウェアに読み込ませ，分析しやすいようにするために整理，整形することをデータクリーニングという。これは，主に学習の11と12で扱う。

学習13から18までの学習では主に機械学習といわれている分野の内容を扱い，一連のデータ分析の流れを体験できるような内容としている。機械学習は，教師あり学習(supervised learning)，教師なし学習(unsupervised learning)，強化学習(reinforcement learning)の三つに大別できるが，本教材では教師あり学習と教師なし学習についてのみ扱う。学習の内容は各学習の中で解説するが，教師あり学習とは，データに対してラベルが付いているものに関して機械学習を行う手法であり，写真データに「犬」や「猫」などの正解ラベルを含んでいるものである。一方で教師なし学習とは，正解ラベルが付いていないデータに関する学習であり，それぞれ目的によって，使い分ける必要がある。どのような目的で分析するかによって，選

ぶ手法や学習の種類は異なるが，それに関しての詳細は各学習の中で解説する。

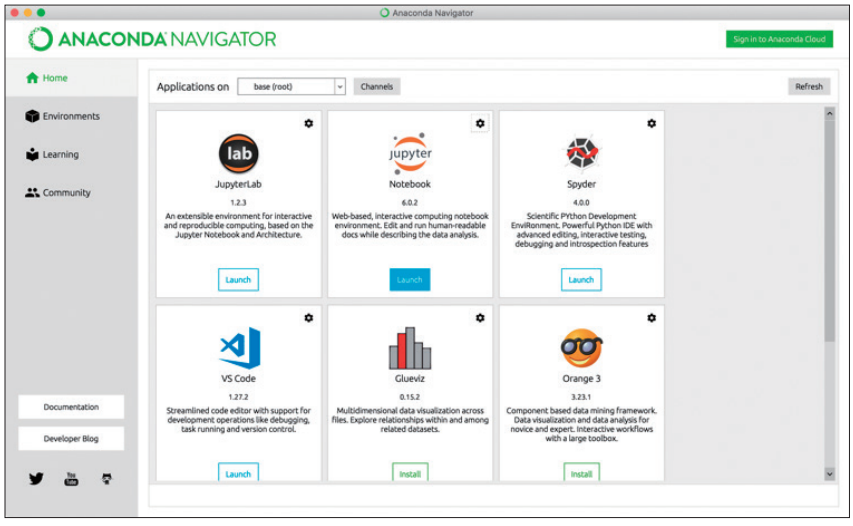
データを分析するために必要なスキルを効率よく学ぶには，学習11と12のデータの収集や整理，整形について学び，それぞれのデータ分析の目的や手法に合わせて，学習13から18までを任意に選択するとよいだろう。



図表2　データ分析と学習の分類

3 本章の学習での演習環境

本章の演習においては，表計算ソフトウェア，統計処理ソフトウェアR，プログラミング言語Pythonを想定している。Rに関しては，v.3.6以上，統合開発環境であるRStudioの利用が望ましい。Pythonに関しては，v.3.7以上で，AnacondaのJupyter Notebook，Jupyter Labo，Spyderなどの統合開発環境での利用を想定している。また一部の学習では，Google Colaboratoryも活用している。



図表3　Anaconda Navigator

学習

11 データと関係データベース

研修内容

研修の目的

- データの信憑性や信頼性について生徒に考えさせ，生徒が適切にデータを処理する力を養う授業を展開できるようになる。
- 生徒にデータ形式による違いや蓄積方法を選択することの重要性を理解させ，適切なデータ形式と蓄積方法を考えさせる授業ができるようになる。
- 生徒に関係データベース（リレーショナルデータベース：RDB）を実際に操作させ，表形式データの蓄積方法を理解させる授業ができるようになる。
- 生徒にデータベースの種類（SQL型とNoSQL型）を示し，実社会でどのように使われているか，興味を持たせる授業ができるようになる。

この学習項目で使用するプログラミング言語は Python です。

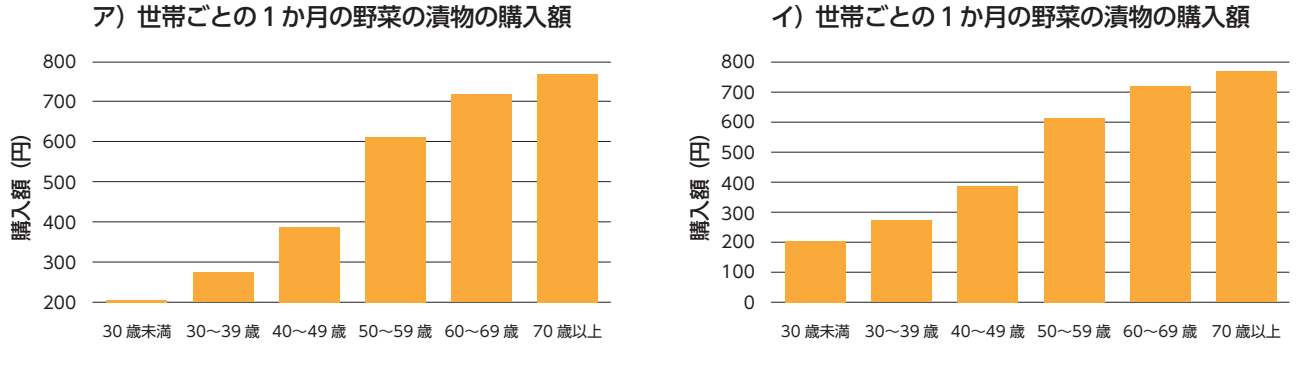
1 データの信憑性と信頼性

(1) データの信憑性

データとはデータ分析に使うもとになるものをいい，データ分析とはそのデータの傾向や性質を数量で捉えることをいう。データは様々な方法で収集することができ，集まったデータの形式も様々である。したがって，信憑性のあるデータの収集や分析をするためには注意すべき点が多くある。

この学習項目では，データ処理の方法によって同じデータでも見え方が異なることを体験するとともに，データの収集や分析の際に起こりうる基本的なデータの偏りについて確認する。

データの信憑性（credibility）とは，示されたデータが信じるに値する科学的根拠となり得るかである。Cambridge Dictionary によると credible は「able to be trusted or believed」とされている。演習 1 では，同じデータであっても目盛りの取り方や階級の幅（ビン幅）の取り方，外れ値の扱いにより異なったヒストグラムとなることを確認し，データの信憑性をデータの提供者の視点から考える図表 1。



図表1　平成26年品目別1世帯当たりの1か月間の支出より野菜の漬物の購入額　出典：全国消費実態調査のデータを利用して作成 https://www.e-stat.go.jp/

演習 1

EXERCISE

図表 1 のアとイはそれぞれ同じデータを使用して作成したものです。アとイの表現の違いを近くの人と共有し，グラフ作成の際に注意すべき点を確認しましょう。

ヒストグラムでは適切な階級の幅を設定することが重要である。分析のために同じデータでいくつかの階級の幅を試すことも有用である。適切な階級の数は計算で求める方法もいくつかある。例としてスタージェスの公式を挙げておく。

（階級の数）$= 1 + \log_2 n$ （n はデータの個数）

また，何らかの原因により，収集したデータの中で他と大きく異なる値を外れ値という。外れ値からデータ全体の重大な欠陥や特徴が見えることもあるため，外れ値には注意を払う必要がある。

(2) 母集団と調査

データの収集，読み解き（理解）の際には母集団に気を付けて標本の大きさ（サンプルサイズ）や標本抽出（サンプリング）の方法を選択する必要がある。母集団を適切に把握しないと集めたデータに偏り（バイアス）が生じることもあるため，母集団を意識してデータ収集を行い，提示されたデータを読み解く必要がある。

母集団全体を調査することを全数調査，悉皆（しっかい）調査という。日本では 5 年に一度行われる国勢調査が代表例である。母集団全体を調査できないときには有意抽出や無作為抽出を用いる。有意抽出とは母集団の様々なカテゴリから適切な割合を見ながら作為的に標本をとることである。無作為抽出では特定の属性を持ったデータに偏らないようにランダムに選んだ標本を使う。主にアンケート調査や視聴率調査にも使われている。サンプルサイズが小さすぎると信頼性は低くなる。適切なサンプルサイズは「数学 B」で学習する計算式を用いて求めることができる。国の調査などでは一般に信頼水準 95％を用いている。

$$n = 1.96 \cdot \frac{p(1-p)}{d^2}$$

（n はサンプルサイズ，p は回答率，d は許容誤差）
※信頼水準99%では1.96の代わりに2.58を使う。

これを簡単に計算できるフォームを公開している Web サイトもある。アンケートを実施するなどデータ収集の際には必要なサンプルサイズに気を付けたい。

演習 2

EXERCISE

総務省が行っている毎月の家計調査などについて，どのようにして標本を抽出しているか調べてみましょう。

(3) バイアス

バイアスには選択バイアス，生存バイアスなどがある。選択バイアスには，標本抽出の際に発生する標本バイアス（母集団の取り違えなど意識的，無意識的によらず無作為でない抽出）や膨大探索効果（膨大なデータに対して様々なモデルを当てはめると何かしらの興味深い結果が出るが，それは本当に興味深い結果か偶然のいずれかである），特定の分析結果を強調するための時間間隔をとること，分析者にとって都合の良い結果が見えたところで検定を止めることなどが含まれる。生存バイアスは，以下の例に示すように母集団全体ではなく偏ったデータ抽出をしてしまうことである。

例えば，お菓子の売れ行きが芳しくないときにそのお菓子の購入者にアンケートを採ることや，ランダムで発生させた番号の固定電話に電話をかけて聞き込み調査をするといったことである。お菓子の購入者にアンケートを採っても肝心のお菓子を買わなかった人は含まれていないし，固定電話の電話番号にかけても固定電話を持たない人は含まれていないので調査結果に偏りがあることは容易に想像できるはずだが，意外と見落としがちなバイアスである。

2 データの形式と蓄積方法

この学習項目では，表形式によるデータの保持とデータの型，関係データベースによるデータの蓄積方法について学ぶ。

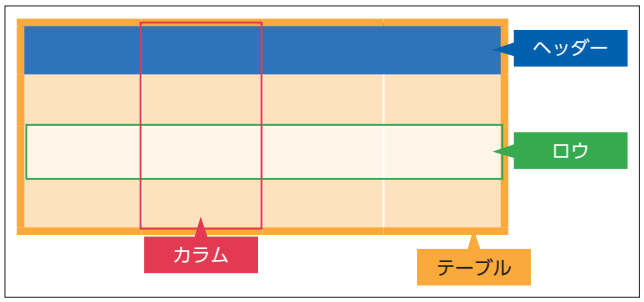
ビッグデータという言葉が一般的になるくらい身の回りにデータがあふれている。しかし，それらのデータは紙にメモ書きされたようなものから，種類ごとにまとめられたもの，観測値のように数値だけが集積されたものまである。そしてそれらのデータは往々にしてそのままでは使いづらい。例えば，地震の観測値（常にどのくらい揺れているかを測定している値）は数値の羅列であり，事前に何を表す値なのか分かっている人に説明してもらわなければ，見る人にその数値の意味が分からない。一方で，人にとってデータの意味を分かりやすいようにすると，データサイエンスでデータ処理する際には扱いづらい。例えば，通知表のように各生徒の履修科目とその成績を印刷すると，一人一人の生徒にとっては分かりやすいが，コンピュータではそれらのデータは処理しづらい。

ここで，人にもある程度データの意味を読むことができ，コンピュータでも処理しやすい形でデータを保持，蓄積する方法として表形式データがある。表形式データを蓄積する方法として関係データベース（リレーショナルデータベース，以後RDBと表記）がある。他にもデータの保持，蓄積方法はあるが，それらについてはこの学習の最後の節で扱う。

表形式データは，表計算ソフトでの扱いのように行（ロウ）と列（カラム）からなるデータ保持である図表３。この表形式のデータをRDBではテーブルと呼ぶ場所を用意して格納している。テーブルでは列名を表示するためにヘッダーを用意し，データベースの構造であるスキーマや各列の値の種類を示す型（一例を図表２に挙げている）が分かりやすいように指定する。RDBでは行のことをレコード（タプル），列のことを属性（カラム）という。RDBを操作するためにはSQL（Structured Query Language）を使う。SQLにはMySQLやSQLiteなどの無料のものから有料のものまで様々ある。また，データを処理するハードウェアが1台で完結することを想定していることが多く，複数のハードウェアで大量のデータを分散処理することは得意ではない。表形式で整理されていない大量のデータを高速に分散処理することは，本学習の最後の項「4. データベースの種類とその使い分け」で述べるNoSQLの方が得意である。NoSQLにも様々な種類があり，目的に応じて使い分けることが求められている。

データ型	例	用途
Char（固定長） （varchar可変長）	Abc	文字，文字列を表す
String（text）	Abc	文字列を表す
Int	1, 2, 3	整数を表す
Real（Double, Float）	-0.2, 3.4	実数（または浮動小数点数）を表す
Object（blob）	123abc	図，音声などのデータ

図表2　RDBで使われるデータ型の例



図表3　表形式データの構造と名称

演習 3

EXERCISE

小テストのデータを表形式に格納する際に必要なヘッダーとその型を挙げてみましょう。

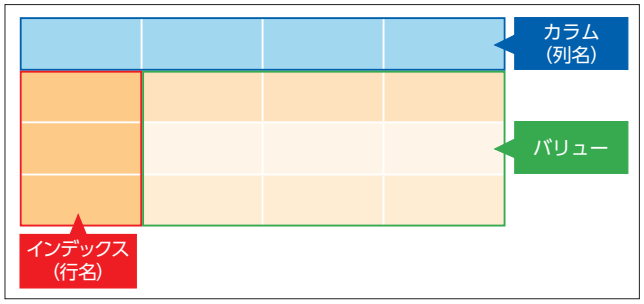
RDBは表計算ソフトや2次元配列（行列）のような形をしている図表３。ヘッダーを付けない場合もあるが，付ける際にはスキーマや格納するデータの型が分かりやすいように付ける。レコード（タプル）は表の行を表し，一つ一つのデータである。属性（カラム）は列を指し，データセットの特徴量（特定の種類）を表す。

3 データサイエンスでのデータの利用

この学習項目では，簡単なデータ操作ではなく，大量のデータをプログラミングで処理，分析するデータサイエンスでのデータ利用の基本操作や用語を確認する。

プログラミング言語では，データベースやデータをそのまま扱うこともできるが，大変面倒である。そのためRやPythonではデータフレームと呼ばれる形でデータを扱う図表4。データを扱うことに向いたこれらの言語ではRDBをSQLで直接操作するよりもずっと簡単に高速にデータを読み込み，適切な形で処理することができる。

実際にGoogle DriveとGoogle Colaboratory（以下Colabと略す）を使ってPythonでデータフレームを作成し，データの読み込みとデータ操作をしてみよう。まず，データフレームに読み込むデータTestdata.csv（別添資料）をColabからアクセスできる場所に置くか，



図表4　データフレームの基本構造と名称

Google Driveにあるファイルならば，Google Driveをマウントする※。Google Driveを別ウィンドウで開いてマイドライブの中にできているColab Notebooks（Colabのためのフォルダ）にTestdata.csvを置く。これはあるクラス39人の小テストのダミーデータである。以下のたった数行のコードで表形式のデータを読み込み，データフレームとして扱うことができる。

```
01  import pandas as pd
02  df = pd.read_csv('/content/drive/My Drive/Colab Notebooks/Testdata.csv')
03  print(df)
```

1行目はPythonでデータフレームを扱うためのパッケージの一つであるpandasを使えるようにしている。pdと名付けている。2行目のread_csv関数でデータフレームをdfとして作成している。()内の「'(引用符)」ではさまれた部分が読み込みたいデータのパス（最初にTestdata.csvを置いた場所）である。3行目はprint文を使って作成したデータフレームdfの中身を出力する。確認が不要であれば書く必要はない。

pandasのデータフレームでは，axis=0で同じ列内のデータ，axis=1で同じ行内のデータを対象に，集約計算やデータ削除の処理を行う。また，インデックスを使って行を指定することができる。これにより簡単に大量のデータから列や行の指定，削除ができる。重複データや欠損値データの抽出機能もある。欠損値の削除や補完は学習12で扱う。

演習4

EXERCISE

pandas によるデータフレームのメリットを感じてみよう

特定の行のデータは以下の一行で削除できます。print() 文を使いどの行が削除されたか表示して確認しましょう。余裕があれば重複データを見つけ（別添資料 Test_csvimportD.ipynb），PlayersNo の列で特定の番号のデータを指定して削除してみましょう。

- **特定行の削除：df = df.drop(0)**
- **全く同じ重複データの発見：print(df[df.duplicated()])**
- **PlayerNo が重複したデータの削除：df.drop_duplicates(subset='PlayerNo',keep='first',inplace=True)**

ここでは，subset = 'PlayerNo' で PlayerNo が重複したデータを抽出し，keep='first で最初の重複データを残して inplace=True で元のデータフレームを置き換えています。

※ Google Drive のマウントの方法に関しては，別添資料「プログラミング開発環境の使い方」を参照してください。

4 データベースの種類とその使い分け

この学習項目ではデータベースの種類を知り，それぞれのデータ蓄積方法の利点と欠点を考える。

データベースには大きく分けてSQL構文を使い操作を行うRDBと，RDBやRDBMS以外のNoSQLとがある。現在ではNoSQLはNot only SQLの略とされることが一般的であり，RDBとRDB以外のデータベース両方を使う手法が登場している。

大量のデータを扱う際にRDBは必ずしも効率的な方法ではなく，NoSQLでは，これらのデータをキー・バリュー型，カラム型，ドキュメント型，グラフ型などの形式で蓄積している。

解答資料に，より詳しくNoSQLを学びたい方に向けて，いくつかの無料枠や学習素材を提供しているデータベースを紹介している。

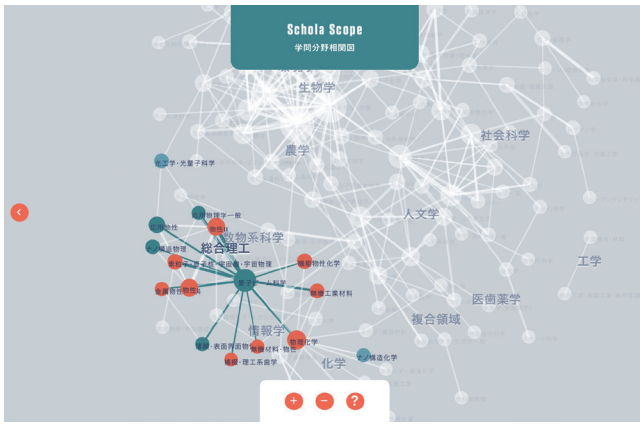
演習 5

EXERCISE

グラフ型データベースのイメージ体験

Scholar Scope を使って研究分野の関連性を探ってみましょう図表５。

興味を持った分野をクリックして関連のある分野を見つけ，更なる関連分野を探しましょう。



図表５　「Schola Scope」https://navischola.app/network/6/general-science-and-engineering/

「2.データの形式と蓄積方法」(112ページ)では，データ保持の形式として表形式を扱った。ここでは，JSON（JavaScript Object Notation)形式について基本的なことを確認しておく。JSON形式は，NoSQLでもドキュメント型で代表されるように，表形式では保持しにくいデータに対応した形式であり，国が公開しているデータベースでも提供されている形式である。

図表６のように{　と　}で囲まれた部分でひとまとまりのデータを表し，「:」で区切った左側をキー（名前)，右側をバリュー（値)という。「,」で区切って複数のデータを同じ{　}内に並べることもできる。また，入れ子構造となることや配列([と]で囲まれた部分)を要素とすることも多い。図表６では人間が理解しやすいようにインデントや改行が入っているが，これはあってもなくてもよい。コンピュータが処理する際にはこのインデントや改行はなく，全て続いた状態で保持され，読み込みも行われる。

113ページのCSVファイルの場合と同様に，Pythonではたった数行で簡単にJSON形式のデータをプログラムに読み込み，使うことができる。学習23以降では，同様にJSON形式のデータを読み込み，更に活用している。

```
{
   "名前":"たろう",
   "年齢":17,
   "履修科目":{
      "修得済":{
         "国語":["国語総合"],
         "数学":["数学１","数学 A"],
         "情報":["情報１"]
      },
      "未修得":{
         "情報":["情報２"]
      }
   }
}
```

図表６　JSON形式の例：study.json

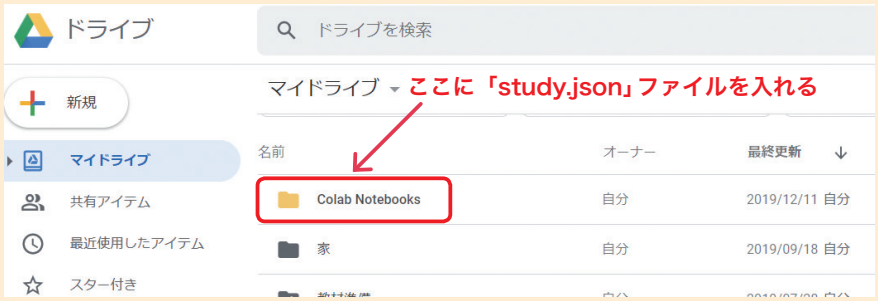
演習 6

EXERCISE

図表6に示したJSON形式のデータ「study.json」をPythonで読み込んでみましょう。

Google Colaboratoryを開いて新しくjsonload.ipynbとしてPythonファイルを作成し，Google Driveをマウントする。

別添資料にあるstudy.jsonを各自のドライブのColaboratoryのためのフォルダに配置し，マウント後に以下の４行のプログラムを実行すると，図表6のJSON形式のデータが読み込まれ，出力される。




```
01  import json
02  with open('/content/drive/My Drive/Colab Notebooks/study.json') as f :
03    jsndata = json.load(f)
04  print(jsndata)
```

```
import json
with open('/content/drive/My Drive/Colab Notebooks/study.json') as f:
  jsndata = json.load(f)
print(jsndata)
```

```
['名前': 'たろう', '年齢': 17, '履修科目': ['修得済': ['国語': ['国語総合'], '数学': ['数学1', '数学A'], '情報': ['情報1']], '未修得': ['情報': ['情報2']]]]
```

更に余裕があれば自分で書いたJSON形式のデータが正しい文法で書けているかを検証できる無料のサイトがあるので，ぜひ試してみてほしい。他にもJSON形式で表現したデータをアップロードして他の人と共有できるサイト，他の人が用意したオープンソースのデータを利用できるサイトがある。アイスクリームの講評データを集めて公表したものや，急患診療事業情報といった面白そうなデータから役立ちそうなデータまで様々あるので，ぜひ探して分析や開発に利用してほしい(参考文献参照)。

また，NoSQLで使用されるデータ保持の形式としてLOD（Linked Open Data)もある。発展的な学習の参考までに，解答資料で触れているので様々なデータ保持の形式を見てほしい。

[参考文献・参考サイト]

- 「なるほど統計学園高等部」総務省統計局 https://www.stat.go.jp/koukou/index.html
- 「政府統計の総合窓口(e-Stat)」総務省統計局 https://www.e-stat.go.jp/
- 「データサイエンスのための統計学入門」Peter Bruce，Andrew Bruce 著　黒川利明 訳　大橋真也 技術監修　オライリージャパン(2018)
- 「東京大学のデータサイエンティスト育成講座」塚本邦尊，山田典一，大澤文孝 著　中山浩太郎 監修　松尾豊 協力　マイナビ出版(2019)
- 「[第 2 版] Python 機械学習プログラミング」Sebastian Raschka，Vahid Mirjalili 著　株式会社クイープ 訳　福島真太朗 監訳　インプレス(2018)
- 「NOSQL の基礎知識」太田洋 監修　本橋信也，河野達也，鶴見利章 著　リックテレコム(2012)
- 「navischola.app Schola Scope」https://navischola.app/network/6/general-science-and-engineering/
- 「JSON.org」https://www.json.org/json-en.html
- 「LinkData.org」http://linkdata.org/work?sort=date
- 「Cambridge Dictionary」CambridgeUniversity Press 2020 https://dictionary.cambridge.org/ja/dictionary/
- 「The Linked Open Data Cloud」https://lod-cloud.net/
- 「DBpedia Jpanese」http://ja.dbpedia.org/

学習活動と展開

学習活動の目的

- データの信憑性や信頼性を考えて適切にデータ処理をできるようになる。
- 収集したデータの特性や用途から，適した蓄積方法を考える力を身に付ける。

学習活動とそれを促す問い

	問　い	学習活動
展開 1	データの信憑性や信頼性はどのように調べるか。	標本抽出の方法やバイアスについて調べる。
展開 2	表形式データをデータベースやデータフレームとしてどのように扱えばよいか。	リレーショナルデータベースや Python の pandas パッケージを使う。
展開 3	表形式でないデータに関してはどのように扱うか。	NoSQL データに触れてみる。

展開 1	
問　い	データの信憑性や信頼性はどのように調べるか。
学習活動	標本抽出の方法やバイアスについて調べる。
指導上の留意点	● データを可視化する際の留意点を理解させる。 ● 標本抽出やバイアスとデータの信頼性について関連付ける。

展開 2

問　い	表形式データをデータベースやデータフレームとしてどのように扱えばよいか。
学習活動	● リレーショナルデータベースを操作する。 ● Python の pandas パッケージを使う。
指導上の留意点	● リレーショナルデータベースの作成や操作について理解させる。 ● データフレームの構造について理解させる。

展開 3

問　い	表形式でないデータに関してはどのように扱うか。
学習活動	NoSQL データに触れてみる。
指導上の留意点	表形式でないデータに関しての格納形式や扱いについて理解させる。

まとめ

まとめ	● データの信憑性や信頼性について理解できる。 ● データベースやデータフレームなどを用いてデータの加工や変換，整理が行えるようになる。

学習

12 大量のデータの収集と整理・整形

研修内容

研修の目的

- 大量のデータを収集する方法について理解し，目的に応じたデータの収集の技能を生徒に身に付けさせる授業ができるようになる。
- データの形式の変換やデータの単位，欠損値や外れ値の処理を行う方法について理解し，目的に応じたデータを整形・整理する知識や技能を生徒に身に付けさせる授業ができるようになる。

この学習項目で使用するプログラミング言語は Python です。

1 データの収集

データを収集する際には，アンケートのような調査からの取得，計測機器からの取得，Web APIからの取得などの方法がある。国内のWebサイトだけでなく，海外のWebサイトにも目を向けると多様なデータを取得することができる。ここでは，USGS（U.S. Geological Survey，アメリカ地質調査所）のWebサイトで提供されているWeb APIを用いて，データを取得してみよう。

演習 1

EXERCISE

世界中で発生した地震について調べるために，USGS から Web API を用いて地震のデータを取得してみましょう。

一般にWeb APIを用いてデータを取得する際は，これを提供するサイトのユーザー登録とAPIを利用するためのキー（特定の文字列）の取得が必要である。ここでは，登録なしで利用できるUSGSのEarthquake CatalogのAPIを使って実習してみよう。

ブラウザのアドレスバーに，APIのエンドポイントのURL（https://earthquake.usgs.gov/fdsnws/event/1/[METHOD[?PARAMETERS]]）のMETHODにはメソッド名，PARAMETERSには必要なパラメータを与えてアクセスすることによりデータを取得できる。例えば，URLを「https://earthquake.usgs.gov/fdsnws/event/1/query?format=geojson&starttime=2019-12-01&endtime=2019-12-02」としてアクセスすることにより，2019年12月1日から2日に発生した地震のデータをGeoJSON形式で取得できる図表 1。詳細な説明はAPI Documentationのページを参照されるとよい。

```
{"type":"FeatureCollection","metadata":
{"generated":1579338433000,"url":"https://earthquake.usgs
format=geojson&starttime=2019-12-01&endtime=2019-12-02","
Earthquakes","status":200,"api":"1.8.1","count":378},"fea
{"mag":1.1799999999999999,"place":"8km NNW of Redwood Val
CA","time":1575244115740,"updated":1575408903432,"tz":-48
kes/eventpage/nc73310796","detail":"https://earthquake.us
```

図表1　USGSから取得した地震データの一部（GeoJSON形式）

（「USGS API Documentation」https://earthquake.usgs.gov/fdsnws/event/1/）

他にも，Webサイトにはオープンデータを含め，様々なデータが掲載されており，十分な価値がある。また，公式サイトの説明文や口コミサイトなどのコメントといったテキストデータも収集し分析することで，有用な情報が得られる。自動でWebサイトを巡回してデータをかき集めることをクローリングといい，そのデータを解析して必要なデータを抽出することをスクレイピングという。これらを合わせてWebスクレイピングといい，Pythonは，requestとBeautiful Soup4というパッケージ，Rはrvestというライブラリを用いる。

演習 2

EXERCISE

文部科学省から発表された内容を確認するために，新着情報のページから，項目名を抽出してみましょう。


```
<h3 class="information-date">令和2年1月6日</h3>
<ul class="news_list">
    <li>
        <div class="area_tag">
            <span class="tag genre_07">報道発表</
            span>
        </div>
        <span class="link"><a href="/
        b_menu/houdou/31/041415183_00001. htm">
        文部科学省所管特殊法人の理事長の任命について </a></span>
    </li>
```

図表2　文部科学省新着情報のページとHTML（一部抜粋）

（「文部科学省 新着情報（2020年1月6日時点）」https://www.mext.go.jp/b_menu/news/index.html）

文部科学省の新着情報のページのHTMLでは日付に，class属性がinformation-dateであるh3タグが付いている。項目は1日分ごとにclass属性がnews_listであるulタグ内に記述され，1記事ごとにaタグが付けられている。これらの特徴を基に掲載日ごとに項目名を取得するPythonによるプログラムは，次の通りである。

Webスクレイピングを用いることにより，Webサイトに掲載された様々なデータの取得が可能となるが，繰り返し文などにより多数回アクセスを試みたり，画像などのファイルサイズの大きいファイルを多数取得したりすることは，対象となるWebサーバに必要以上の負荷をかけることから，プログラムの実行の可否を検討する必要がある。また，Webスクレイピングを利用規約で禁止しているWebサイトもあるので，利用規約を確認することも必要である。

次のプログラムにより，文部科学省の新着情報のページから項目名を抽出することができる。

```
01  import requests
02  from bs4 import BeautifulSoup
03  url = 'https://www.mext.go.jp/b_menu/news/index.html'
04  r = requests.get( url )
05  soup = BeautifulSoup( r.content, 'html.parser' )
06  links = soup.find_all( 'ul', 'news_list' )
07  for l in links:
08    titles = l.find_all('a')
09    for t in titles:
10      print( t.string )
```

2 収集したデータの整理

収集したデータはそのままではプログラムで扱うことができない場合が多い。例えば，表の題名が付いていたり，不要な列が付加されていたりする。これはプログラミング言語のデータの形式と合っていないことが原因である。また，データの表記にゆれがあり，同じデータが異なったデータとして扱われてしまうことも起こりうる。更に，同じデータが重複して存在することもある。そのため，表計算ソフトなどを用いて，これらを事前に整理する必要がある。表計算ソフトのフィルター機能を用いることで，列に含まれるデータの一覧を表示することができ，該当データを抽出して修正する。このような修正をデータクリーニングという。

演習3

EXERCISE

都道府県別の人口の状況を調べるため，e-Stat で「都道府県の指標　基礎データ　人口・世帯　2020」とキーワード検索して表示されたデータを，表計算ソフトで読み込める形式でダウンロードしました。
このデータの中の 2015 年の部分をプログラミング言語で扱えるように表計算ソフトで修正しましょう。

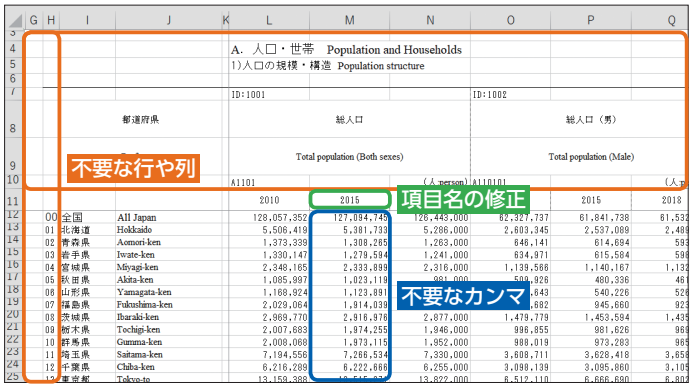
ダウンロードしたファイルは，不要な行や列が含まれるなどそのままプログラムで操作するには適していない。このファイルに対して，表計算ソフトを用いて次の操作を行う図表3, 4。

- 不要な行や列を削除する
- 項目名を修正する
- 不要なカンマが付与されないようセルの書式を「通貨」から「標準」に変更する
- プログラムでファイルを開くことができるよう「CSV UTF-8形式」で保存する

このデータは比較的修正が少なくて済むデータであったが，他に次のような修正が考えられる。

- 表記のゆれを修正する（大文字と小文字，西暦と和暦，正式名称と略称，空白の有無など）
- 不要なデータの注釈や空白文字を除去する
- 重複するデータを除去する

公開されているデータには，印刷することを前提に整形されているものも多く，このようなデータを，プログラムを用いて処理するには，このように前処理をする必要がある。



図表3　前処理が必要なデータ

	A	B	C	D	E	F
1	都道府県	総人口	出生数	死亡数	転入者数	転出者数
2	北海道	5381733	36695	60667	49407	57823
3	青森県	1308265	8621	17148	18162	24755
4	岩手県	1279594	8814	16502	18137	22430
5	宮城県	2333899	17999	23070	50024	49813
6	秋田県	1023119	5861	14794	11999	16473
7	山形県	1123891	7831	14960	13634	17663
8	福島県	1914039	14195	24205	29485	31552
9	茨城県	2916976	21700	31025	50399	58326
10	栃木県	1974255	15306	20519	34885	38607
11	群馬県	1973115	14256	21519	32038	32553
12	埼玉県	7266534	56077	62565	180451	162374
13	千葉県	6222666	47014	56079	155892	147853
14	東京都	13515271	113194	111673	456635	372404

図表4　整形後のデータ

3 データフレームを用いたデータの操作

学習11ではリレーショナルデータベースについて学習し，データフレームについても簡単に扱った。Pythonではpandasというパッケージを使うことにより，データを抽出したり，集計したりというように，まとめてデータを扱うことができるようになる。他には，リレーショナルデータベースのように二つのデータを結合したり，重複データを取り除いたりすることができる。

演習4

EXERCISE

人口減少社会といわれていますが，pandas を用いて都道府県ごとに人口 10 万人に対する人口の増減について調べてみましょう。
ここで，人口の増減は（出生数＋転入者数）－（死亡数＋転出者数）により求めるものとします。

pandasを用いてデータフレームにデータを読み込む。読み込んだデータフレームの出生数，転入者数，死亡数，転出者数から増減を計算し，新しい列として追加する。同様に人口増減率を求めて，新しい列を追加する。人口増減率が多い都道府県が分かるように，df.sort_values('増減率', ascending=False) によりデータをソートする。ascending=Falseとすることで，降順でソートすることができる図表5。

```
01 import pandas as pd
02 df = pd.read_csv("population.csv")
03 df['増減'] = (df['出生数']+df['転入者数']) - (df['死亡数']+df['転出者数'])
04 df['増減率'] = df['増減'] / df['総人口'] * 100
05 df.sort_values('増減率', ascending=False)
```

	都道府県	総人口	出生数	死亡数	転入者数	転出者数	増減	増減率
46	沖縄県	1433566	16941	11326	26384	26476	5615	0.391681
22	愛知県	7483128	65615	64060	127036	116518	1555	0.020780
12	東京都	13515271	113194	111673	456635	372404	1521	0.011254
24	滋賀県	1412916	12622	12507	27302	29403	115	0.008139

図表5　pandasにより処理された人口のデータ

日本全国での人口の増減を調べるには，列に対してsumを適用すればよい。

```
01 df['増減'].sum()
```

出力結果

```
-283642
```

4 ロングフォーマットとワイドフォーマット

表形式のデータにはワイドフォーマット(横持ち形式)とロングフォーマット(縦持ち形式)という二つの形式がある。ワイドフォーマットは図表6の商品データの例のように，一つの商品の全ての属性を，横に並べた列を使い，1行で保持する形式である。ロングフォーマットは，図表7の例のように，同じ商品であっても属性ごとに行を縦に増やして，全ての属性の情報を保持する形式である。

ワイドフォーマットは，散布図のように1行を一つの観測値として値を用いる場合や，分類やクラスタリングを行う場合には有用である。これに対し，ロングフォーマットは，属性ごとに積み上げ棒グラフを描いたり，属性ごとに折れ線グラフを描いたりする場合に有用である。

また，ワイドフォーマットでは，属性を追加した場合にそのデータを参照するシステムの修正が伴ってしまい，変更が容易ではない場合がある。更に，属性の数が多くても，全てのレコードが全ての属性についてのデータがない場合もある。このときワイドフォーマットでは，列(属性)が膨大になると表が疎(空白の項目)を多く持つ状態になってしまう。しかし，ロングフォーマットでは1行で属性とデータを組として保持するため，データが存在しない行は不要であることから，メモリを節約することにつながる。

表計算ソフトでは，ピボットテーブルを作成することにより，ロングフォーマットからワイドフォーマットに変換できる。ここでは，プログラムを用いて変換する。

商品	縦	横	高さ
A	15.4	8.6	2.6
B	20.1	14.3	7.4

図表6　ワイドフォーマットのデータ例

商品	属性	長さ
A	縦	15.4
A	横	8.6
A	高さ	2.6
B	縦	20.1
B	横	14.3
B	高さ	7.4

図表7　ロングフォーマットのデータ例

演習5

EXERCISE

演習4の変数dfを用いて，データ形式の変換に慣れましょう。変数dfはワイドフォーマットになっています。これをロングフォーマットに変換してみましょう。また，変換されたロングフォーマットのデータフレームをワイドフォーマットに変換してみましょう。

ワイドフォーマットをロングフォーマットに変換するプログラムを次に示す。

```
01  melted = df.melt( ['都道府県'] ,var_name='属性', value_name='値')
```

メソッドmeltの引数で都道府県をキーとすることとしている。これによりロングフォーマットの形式が得られる図表8。ここで，ロングフォーマットに変換するときの属性を総人口，増減，増減率に絞り込むには次のようにする。

```
01  melted2 = df.melt( ['都道府県'], ['総人口', '増減', '増減率'],
      var_name='属性', value_name='値' )
02  melted2
```

逆にロングフォーマットをワイドフォーマットに変換するプログラムは次のようになる。

```
01  table = melted.pivot_table( values='値', index='都道府県', columns='属性' )
02  table.reset_index(inplace=True)
03  table
```

メソッドpivot_tableの引数は順に値，行，列になるものをロングフォーマットの列名で与える(実行結果：図表9)。ワイドフォーマットへ変換するとき，ワイドフォーマットで表の同じ場所に集計されるべき値がロングフォーマットの複数の行に存在することもある。この場合には，デフォルトでは平均値が出力される。集計方法として合計を指定するには次のようにする。

```
01  import numpy as np
02  table2 = melted.pivot_table( values ='値', index='都道府県',
      columns='属性', aggfunc=np.sum )
03  table2
```

	都道府県	属性	値
0	北海道	総人口	5.381733e+06
1	青森県	総人口	1.308265e+06
2	岩手県	総人口	1.279594e+06
3	宮城県	総人口	2.333899e+06
4	秋田県	総人口	1.023119e+06

図表8　ロングフォーマットに変換した結果

属性	都道府県	出生数	増減	増減率	死亡数	総人口	転入者数	転出者数
0	三重県	13950.0	-6189.0	-0.340829	20139.0	1815865.0	30612.0	35188.0
1	京都府	19662.0	-5833.0	-0.223456	25495.0	2610353.0	58586.0	59224.0
2	佐賀県	7064.0	-2638.0	-0.316751	9702.0	832832.0	15900.0	18622.0
3	兵庫県	44015.0	-11376.0	-0.205536	55391.0	5534800.0	93099.0	100465.0
4	北海道	36695.0	-23972.0	-0.445433	60667.0	5381733.0	49407.0	57823.0
5	千葉県	47014.0	-9065.0	-0.145677	56079.0	6222666.0	155892.0	147853.0

図表9　ワイドフォーマットに変換した結果

5 欠損値と異常値の取扱い

欠損値は，計測機器の故障などにより値が記録されなかったり，アンケートでの無回答などの理由で値が得られなかったりして，値が欠落していることを示すものである。pandasではNaN，RではNAとして示される。演習6で扱うCSVファイルではセルが空欄になっており，これをpandasで読み込んだときの値はNaNになっている図表10。記録や回答の際に生じる以外に，ロングフォーマットのデータをワイドフォーマットに変換したときに，完全に表が埋まらないことにより欠損値となる場合がある。

欠損値を無視するのがよいのか，除外するのがよいのか，それらしい値を用いるのがよいのかを検討する必要がある。この判断によってバイアスが生じることもあるため，慎重に扱う必要がある。

演習 6

EXERCISE

そらまめ君（環境省大気汚染物質広域監視システム）からデータをダウンロードし大気汚染の状況を調べようとしています。このデータに含まれる欠損値がどの程度あるかを調べて，欠損値の処理をしましょう。

ここでは2019年12月の東京都のデータを用いる。解凍ファイルには多くのファイルが含まれるがそのうちの一つをCSV UTF-8形式で保存し直す。

次にpandasでデータを読み込み，欠損値がどの程度含まれているかを次のプログラムで確認する（実行結果：図表11）。

```
01  import pandas as pd
02  df = pd.read_csv("201912_13_13101010.csv")
03  df.isnull().sum()
```

	A	B	C	D	E	F	G	H	I
1	測定局コー	日付	時	SO2(ppm)	NO(ppm)	NO2(ppm	NOx(ppm)	CO(ppm)	Ox(pp
59	13101010	2019/12/3	10	0	0.005	0.016	0.021		0.
60	13101010	2019/12/3	11						
61	13101010	2019/12/3	12	0.001	0.003	0.012	0.015		0.
62	13101010	2019/12/3	13	0	0.003	0.011	0.014		0

	測定局コード	日付	時	SO2(ppm)	NO(ppm)	NO2(ppm)	NOx(ppm)	CO(ppm)	Ox(ppm
57	13101010	2019/12/3	10	0.000	0.005	0.016	0.021	NaN	0.0
58	13101010	2019/12/3	11	NaN	NaN	NaN	NaN	NaN	N
59	13101010	2019/12/3	12	0.001	0.003	0.012	0.015	NaN	0.0
60	13101010	2019/12/3	13	0.000	0.003	0.011	0.014	NaN	0.0

図表10　CSVファイルの空欄をpandasに読み込んだ結果
（「そらまめ君」http://soramame.taiki.go.jp/DownLoad.php）

```
測定局コード      0
日付            0
時              0
SO2(ppm)       12
NO(ppm)        12
NO2(ppm)       12
NOx(ppm)       12
CO(ppm)       744
Ox(ppm)        13
NMHC(ppmC)    744
CH4(ppmC)     744
THC(ppmC)     744
SPM(mg/m3)      5
PM2.5(ug/m3)    4
SP(mg/m3)     744
WD(16Dir)       1
WS(m/s)         1
TEMP(℃)         1
HUM(%)          1
dtype: int64
```

図表11　欠損値の個数

次に行う欠損値の処理には次のようなものがある。それぞれのデータフィールドを表示して，例えばNOxの濃度の推移を調べたいとき，NOxの濃度の平均値を調べたいとき，それぞれの場合に適切と考えられる処理について考えてみよう。

- **全ての列にデータがあるものを使う**：`df1 = df.dropna()`
- **必要な列を選び欠損値がある行を除く**：
  `df2 = df[['日付','時','NOx(ppm)']].dropna()`
- **欠損値を0として扱う**：`df3 = df.fillna(0)`
- **前の値で埋める**：`df4 = df.fillna(method='ffill')`
- **平均値で埋める**：`df5 = df.fillna(df.mean())`

データの中には，多くのデータからかけ離れた値である外れ値がある。外れ値の中でも原因を特定できるものを異常値という。外れ値は，四分位範囲などの統計量を用いたり，データ間の距離を用いたり，学習16のクラスタリングを用いたりして検出することができる。外れ値が有益なデータの可能性もあるため，その値がどのような原因や理由によって得られたかを考察することが必要である。

【参考文献・参考サイト】
●「アメリカ地質調査所 API Documentation」https://earthquake.usgs.gov/fdsnws/event/1/
●「政府統計の総合窓口（e-Stat）」https://www.e-stat.go.jp/
●「そらまめ君　環境省大気汚染物質広域監視システム」http://soramame.taiki.go.jp/DownLoad.php
●「東京大学のデータサイエンティスト育成講座」塚本邦尊，山田典一，大澤文孝 著　中山浩太郎 監修　マイナビ出版

学習活動と展開

学習活動の目的

- 大量のデータを収集する方法を理解し，その技能を身に付ける。
- 収集したデータの整理・整形の必要性を理解し，そのための技能を身に付ける。
- データを収集した後の処理について考え，そのために必要な整理・整形について判断して行うことができる。

学習活動とそれを促す問い

	問　い	学習活動
展開 1	インターネットから必要なデータを収集しよう。	インターネット上に掲載された Web ページから，Web スクレイピングによりデータを収集する（利用規約を確認してから行う）。
展開 2	収集したデータを活用しやすいように整理・整形しよう。	収集したデータをプログラムから読み込める形式に表計算ソフトなどを用いて加工する。
展開 3	収集したデータをプログラムで活用しよう。	収集したデータをプログラムから読み込み，簡単な集計を行ったり，形式を変換したりする。

展開 1	
問　い	インターネットから必要なデータを収集しよう。
▼ 学習活動	インターネット上に掲載された Web ページから，Web スクレイピングによりデータを収集する（利用規約を確認し，対象となる Web サーバに必要以上の負荷をかけないようにする）。
▼ 指導上の留意点	HTML の構造と比較しながら，収集するデータのタグを見つけられるようにする。

展開 2

問　い	収集したデータを活用しやすいように整理・整形しよう。
学習活動	収集したデータをプログラムから読み込める形式に表計算ソフトなどを用いて加工する。
指導上の留意点	プログラムからデータを読み込むときの形式を意識させて，表記のゆれをそろえたり，不要な行や列を削除したりするなど，必要な処理に気付かせるようにする。

展開 3

問　い	収集したデータをプログラムで活用しよう。
学習活動	収集したデータをプログラムから読み込み，簡単な集計を行ったり，形式を変換したりする。
指導上の留意点	プログラミング言語を用いてデータを読み込んだ後に，どのような処理を行うかを意識させる。

まとめ

まとめ	大量のデータを収集する方法を整理し，収集したデータを活用する上での注意を確認する。

学習

13 重回帰分析とモデルの決定

研修内容

重回帰分析の基本について学習し，実際のデータで重回帰モデルの構築を行う技法を習得する。データから推定された重回帰式の解釈やデータとモデルとの適合度の評価，モデル選択の基準等を学習する。

研修の目的

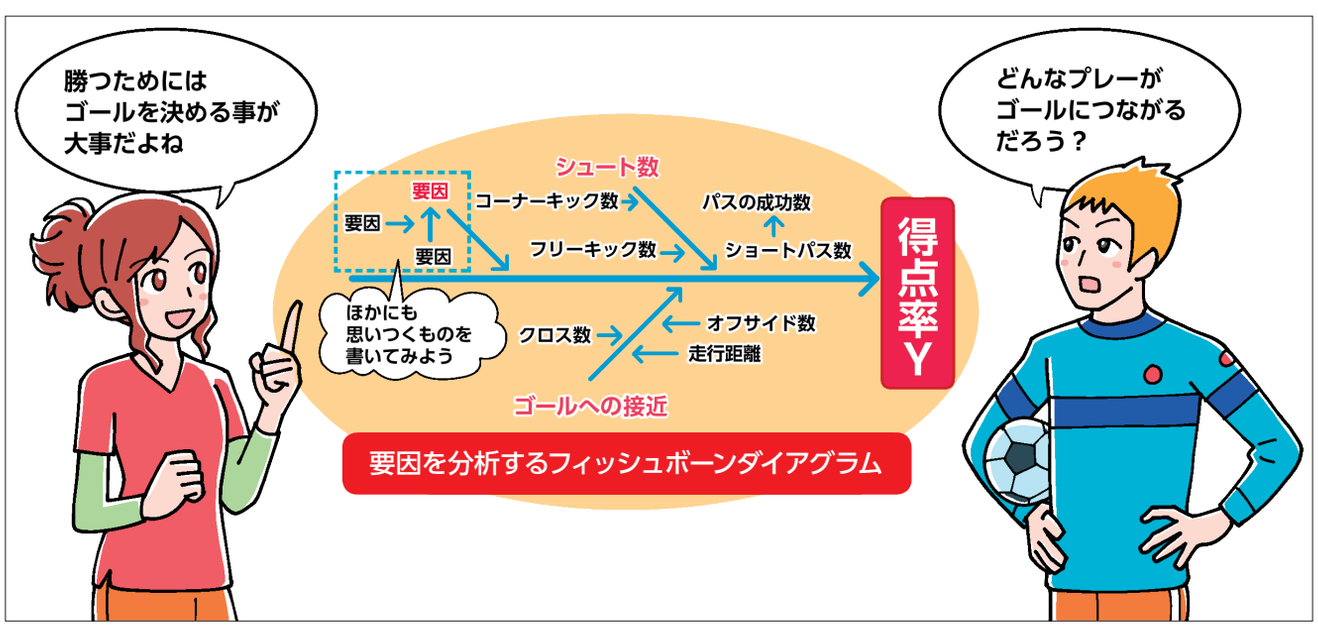
- 課題発見とデータに基づく問題解決の枠組みにおける予測と予測モデル構築の意義，要因分析と制御の位置付けを理解する。
- 量的データの予測モデルの基本である重回帰モデルの概念や用語，活用のためのデータ構造を理解する。
- 具体的なデータで表計算ソフトや統計ソフトRを使った実際の分析の方法と出力の読み方を学ぶ。
- 予測と要因分析の違いを理解し，予測精度を上げるためのモデルの改良(モデリング)の方法を理解する。

この学習項目で使用するプログラミング言語は R です。

1 課題発見とデータに基づく問題解決：予測と制御

問題解決をデータに基づいて行う場合，まず解くべき課題の発見，把握が必要である。課題発見力とは，例えば，海の環境やごみ置き場の清掃状況，地域の商店街の様子や所属するスポーツ部の成績や試験結果など，身の回りの現象に対して，現実があるべき姿(理想)の状態ではないことを知覚し，現実と理想のそれぞれの状態の明確化とそのギャップが解くべき課題であることを具体的に示す力をいう。

次に，その課題をデータと分析で解決する科学的な問題解決のフレームに落とし込むためには，理想や現実の状態を示す客観的なデータ指標Y(目的変数，ターゲット変数，教師変数，予測変数，被説明変数)を具体的に定め，その値の変化や変動に何が影響するのかをいわゆる，5W1Hやその発生に至るプロセス要因



図表1　特性要因図による要因の洗い出し

を洗い出し，論理的な構造モデルをブレーンストーミングなどで作成する必要がある。この際，ブレーンストーミングに使用する論理図には，特性要因図（フィッシュボーンダイアグラム）図表１，要因連関図，イシューツリー，ロジックツリー，ロジックモデル等がある。いずれにしても，何が何に影響を与えるかの構図（仮説）を明示することが肝要である。

データに基づく問題解決のフレームでは，問題を規定する目的変数に対して，その値の変動に影響を与えると思われる要因系の変数（図表１の矢印線の元にある要因）も具体的なデータ指標として記録される。これら要因系のデータ指標群を目的変数Ｙに対して，説明変数（予測子，要因変数）という。科学的問題解決における予測の問題では，一つもしくは複数（p個）の説明変数 $X_1,\cdots,X_p$ の値を使って，目的変数Ｙの値を規定する構造モデル（回帰モデル）をデータから推測（学習）し，その構造モデルを使って説明変数群の状況に応じたＹの値を予測（推測）する。また，各説明変数の値の変化が目的変数にどのように影響するのかという効果を推測（要因分析）したり，シミュレーションを行って最適なＹの値を探索したりと，単純にＹの値を予測するだけではなく制御する方策に関しても考察を深めることもできる。

回帰モデルを推測するためのデータは，説明変数や目的変数が個々の対象に対して，観測・記録された構造化データである。例えば，選手の勝率Ｙを上げることが目的であれば，選手が対象となり，選手のプロファイルデータを収集・整理するが，チームの勝率Ｙを上げることが目的であれば，チームが対象となり，チームのプロファイルデータを収集・整理することになる。「情報Ⅰ」教員研修用教材の学習22では，中古住宅を対象としたプロファイルデータや生徒を対象としたプロファイルデータを示している図表２。このようなデータの一般形を構造化データ（行列データ）という図表３。

近年は，スポーツの成績評価や不動産の取引（成約）価格に，予測モデルを使用したスポーツデータサイエンスや不動産データサイエンスなど，データが活用されるデータサイエンス領域が次々と生まれている。

	ID	性	身長	体重	座高	握力	上体起こし	長座体前屈	反復横跳び	シャトルラン	50m走	立ち幅
2	1	男	167.6	56.2	89.8	35	33	55	49	112	7	23
3	2	男	157.1	50.5	85.8	33	29	48	57	70	7.4	20
4	3	男	165.4	61	85.2	34	31	45	54	76	8	23
5	4	男	168	60	91.1	40	31	55	52	76	7.5	22
6	5	男	165.9	49	89.7	37	32	62	56	87	7.8	24
7	6	男	170	61.5	91.2	36	31	48	55	68	8.2	21
8	7	男	168.7	57	92.3	40	42	50	62	102	6.9	24
9	8	男	173.1	57.5	91.7	47	38	47	63	95	7.1	27
10	9	男	168.2	51.5	90.3	33	35	53	60	88	7.2	24
11	10	男	167.3	51	88.6	36	32	55	61	82	7.5	24
12	11	男	166.1	49.9	89.3	34	35	41	51	51	7.7	23
13	12	男	165.6	64.3	87.6	41	30	61	56	106	6.6	24
14	13	男	161.4	52.5	86.7	30	30	40	44	82	7.3	22
15	14	男	164.6	51.8	87.5	39	35	57	57	114	6.5	26
16	15	男	163.5	59	84.5	32	38	50	49	78	8	18
17	16	男	174.6	59	92.5	44	38	68	62	112	6.5	27
18	17	男	160.7	49.5	84.6	35	35	56	60	92	7.4	23
19	18	男	159.8	49	80.3	38	37	64	60	77	7.5	23
20	19	男	161.7	59	89.5	35	35	56	55	108	6.9	23
21	20	男	170.2	67	93.6	41	35	59	59	60	7.3	21
22	21	男	157.8	45	85.3	31	30	58	59	82	7.8	21

図表２　生徒の体力測定に関するプロファイルデータ

出典：「科学の道具箱」 https://rika-net.com/contents/cp0530/contents/index.html

	X_1	X_2	…	X_p
1	x_{11}	x_{21}	$\cdots$	x_{p1}
2	x_{12}	x_{21}	$\cdots$	x_{p2}
⋮			$\cdots$	
n	x_{1n}	x_{2n}	$\cdots$	x_{pn}

図表３　構造化データの形式（行列，矩形）

演習 1

EXERCISE

特性要因図の作成

ブレーンストーミングを行い，身の回りや社会現象で，予測したい目的変数，予測に使用する説明（要因）変数の候補となぜ，それが要因となるのかの理由を考え，特性要因図にまとめましょう。また，予測や要因分析が何に役立つのかを考えましょう。

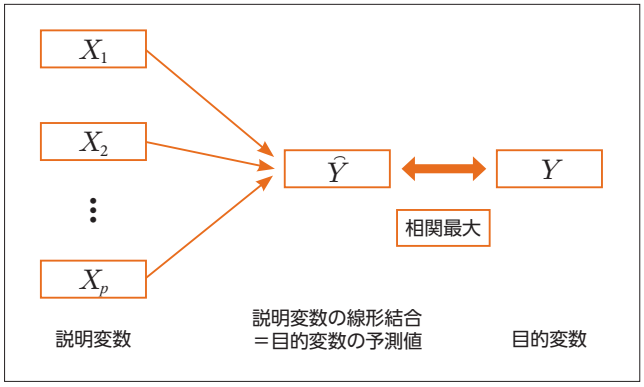
2 予測のための重回帰モデルと重回帰分析

目的変数Yを p 個の説明変数 $X_1,\cdots,X_p$ で説明する最も基本的な数学モデルが重回帰モデルである図表4。重回帰モデルでは，説明変数 $X_1,\cdots,X_p$ に対して第 i 番目の対象のデータの値 $x_{1i},\cdots,x_{pi}$ が与えられたときにその重み付きの合計を求め，それを y_i の予測値 $\hat{y}_i$（y_i ハット）とする。

$$\hat{y}_i = a + b_1 x_{1i} + \ldots + b_p x_{pi} \quad (i = 1,\cdots,n)$$

このとき，a は定数項で，$b_1, b_2,\cdots, b_p$ それぞれの説明変数に係る重みで回帰係数という。また，これらを回帰モデルのパラメータという。

特に，p=1，すなわち，説明変数を一つしか使用しないモデルを単回帰モデルといい，これは，直線の式となる。「情報Ⅰ」教員研修用教材の187ページの演習2では，高校生の体力測定データを使って，「50m走のタイム(Y)」の予測式を「立ち幅跳び(X)」を説明変数として，表計算ソフトの「散布図」上で求めている図表5。予測式は以下となる。

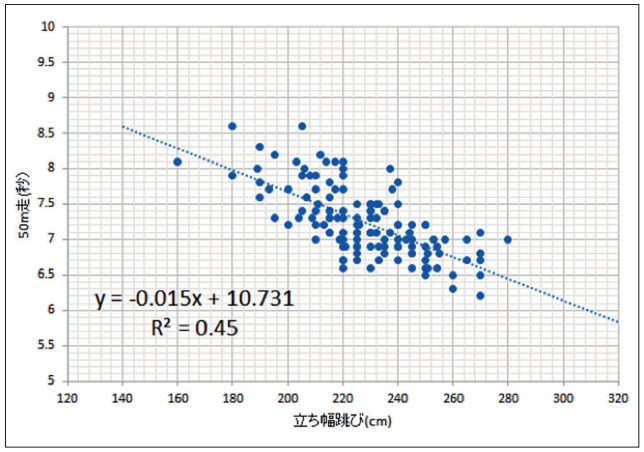


図表4　重回帰分析のモデル

50m走（秒）の予測値 ＝ − 0.015（秒／cm）× **立ち幅跳び（cm）** ＋ 10.731（秒）

回帰係数は，「立ち幅跳び」の単位cmを「50m走」の記録の単位である秒に変換する役割を担っている。この場合は，「立ち幅跳び」が1cm長いと「50m走」の予測値が0.015秒ずつ短縮されることを意味する。

当然，一つの説明変数Xで目的変数Yの変動が全て説明できるわけではない。そこで，より説明力（予測力）を上げるために，複数の説明変数を使用する必要性が出てくる。それが，重回帰モデルである。同じデータに，説明変数を追加して重回帰モデルを当てはめた結果は，以下となる。



図表5　立ち幅跳び(X)と50m走のタイム(Y)の散布図と予測式

出典：（「情報Ⅰ」教員研修用教材　より抜粋）

50m走（秒）の予測値 ＝
− 0.012（秒／cm）× **立ち幅跳び（cm）**
− 0.014（秒／m）× **ハンドボール投げ（m）**
− 0.040（秒／kg）× **握力得点（kg）**
− 0.025（秒／回）× **上体起こし（回）**
＋ 10.819（秒）

重回帰モデルの回帰係数は，Yと$X_1, X_2, \cdots, X_p$に関するデータが与えられたとき，実際のYの観測値と重回帰モデルによる予測値の差(残差)，$e_i = y_i - \hat{y}_i$の2乗和(残差平方和) $SS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ を最小にするように求められる。これを**最小二乗法**という。

最小二乗法によって求められた予測式に，$X_1, X_2, \cdots, X_p$ の各変数の平均値を代入すると，Yの平均値となる。このときの残差は0である。また，この予測値と実際の観測値との相関係数を**重相関係数**という。重相関係数は，0から1の間の値をとる。単回帰分析の場合，重相関係数はXとYの相関係数rの絶対値と等しくなる。

3 重回帰モデルの適合度

最小二乗法により当てはめられた回帰直線が実測されたデータにどの程度，適合しているのかについて，単回帰モデルの場合は，散布図上でデータ点がどの程度直線の近くに集中しているのかで視覚的に判断できる。適合とは，残差が小さいことを意味する。データ全体で，適合度に関する指標として以下がある。

◎重相関係数 R：

回帰モデルによる期待値 $\hat{y}_i$ と実測値との相関係数で，次に示す寄与率の正の平方根として求められる。

◎寄与率(決定係数) R^2：

目的変数Yの変動の何パーセントが与えられた回帰モデルで説明できたかを示す指標で，Yの平均値まわりの変動(全変動S_T)に占めるYの予測値の平均値まわりの変動(回帰による変動S_R)の割合として以下の式で求められる($0 \leqq R^2 \leqq 1$)。寄与率は，100%に近いほどモデルがデータに適合していることになる。

$R^2 = \dfrac{S_R}{S_T} = 1 - \dfrac{S_E}{S_T}$ ，ここで，S_T，S_R，S_E は以下である。

全変動 (全平方和)：$S_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$

回帰による変動 (回帰による平方和)：$S_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

残差変動 (残差平方和)：$S_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

全変動S_Tがもともとの目的変数の変動を表し，残差変動S_Eが回帰モデルで説明できない変動を表している。$S_T = S_R + S_E$が成立することから，回帰による変動S_Rは，回帰モデルで説明できたYの変動と考えることができる。

◎自由度：

S_T，S_R，S_E の各平方和には，それぞれ対応する自由度(独立する成分の数) f_T，f_R，f_Eがある。自由度は，その統計量を構成する本質的な(独立した)要素の数で，データの数nと関連する大事な数量である。もともと対象にしているのは，平均値まわりのYの変動 S_T で，自由度はデータ数から制約式の数(平均値の式) 1 を引き，$f_T = n - 1$ となる。Yの予測値の変動(回帰による変動S_R)の自由度は，回帰パラメータ数と平均制約から，$f_R = (p+1) - 1$より，$f_R = p$となり，その予測値とYの変動(残差変動S_E)の自由度は，$f_E = n - 1 - p$ となる。全変動の分解 $S_T = S_R + S_E$ と同様，自由度の分解 $f_T = f_R + f_E$ も成立する。

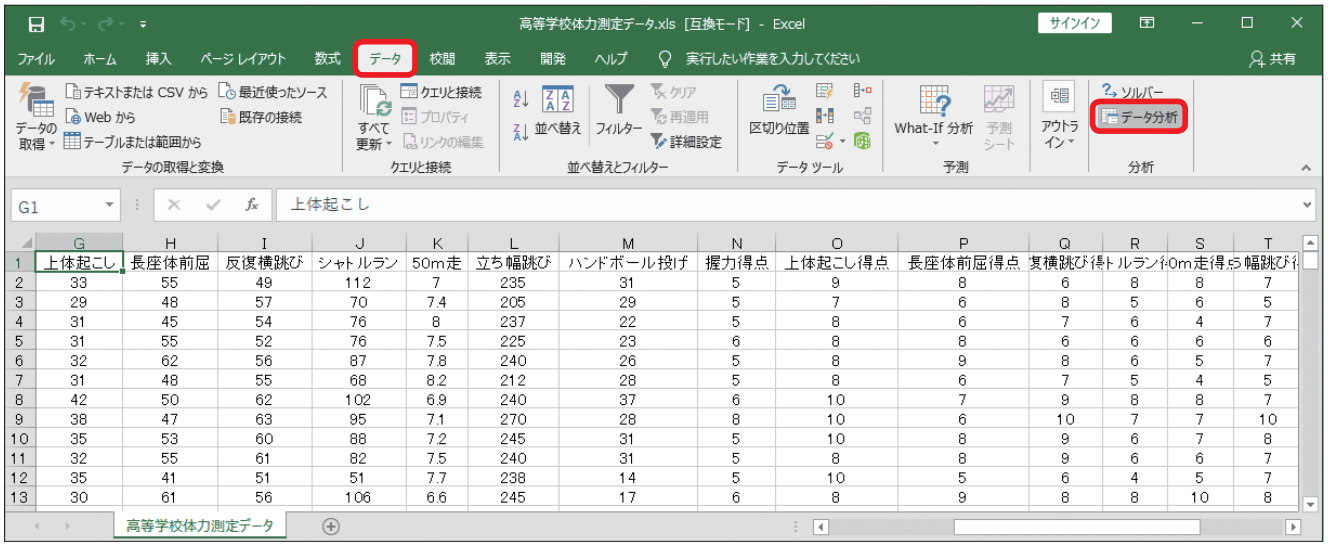
◎標準誤差：

残差の自由度を調整した標準偏差である。(S_E / f_E) の平方根で求められ，Yと同じ具体的な測定単位を持つ。例えば50m走の記録タイムの予測であれば，単位は秒である。この値が小さいほど良いモデルとなるが，単純に，残差変動S_Eが小さくなる(寄与率R^2が上がる)だけでは達成されない。残差変動の自由度f_Eが大きいことも重要となる。

4 重回帰分析のコンピュータでの実行と出力

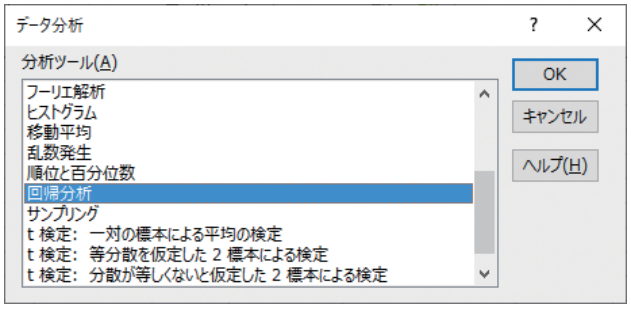
(1) 表計算ソフトでの重回帰分析の実行

重回帰分析は，Excelの「データ」メニューから「データの分析」(アドインで設定)を選択することで簡単に実行できる図表６，７，８。



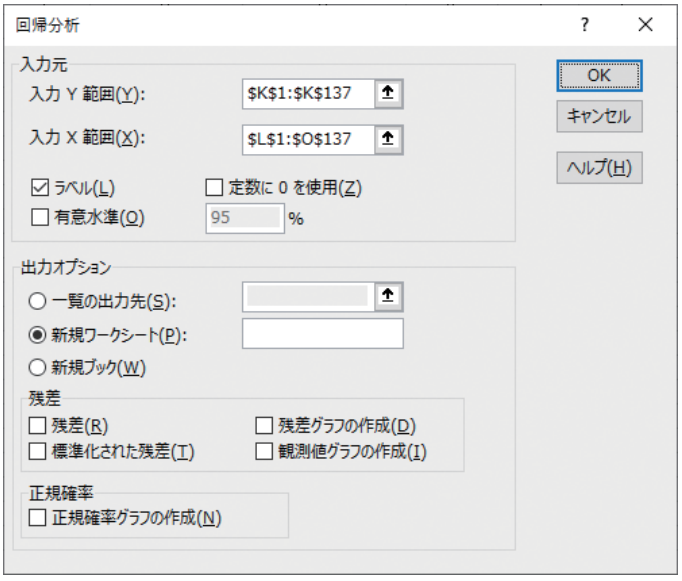
	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	上体起こし	長座体前屈	反復横跳び	シャトルラン	50m走	立ち幅跳び	ハンドボール投げ	握力得点	上体起こし得点	長座体前屈得点	反復横跳び得点	シャトルラン得点	50m走得点	立ち幅跳び得点
2	33	55	49	112	7	235	31	5	9	8	6	8	8	7
3	29	48	57	70	7.4	205	29	5	7	6	8	5	6	5
4	31	45	54	76	8	237	22	5	8	6	7	6	4	7
5	31	55	52	76	7.5	225	23	6	8	8	6	6	6	6
6	32	62	56	87	7.8	240	26	5	8	9	8	6	5	7
7	31	48	55	68	8.2	212	28	5	8	6	7	5	4	5
8	42	50	62	102	6.9	240	37	6	10	7	9	8	8	7
9	38	47	63	95	7.1	270	28	8	10	6	10	7	7	10
10	35	53	60	88	7.2	245	31	5	10	8	9	6	7	8
11	32	55	61	82	7.5	240	31	5	8	8	9	6	6	7
12	35	41	51	51	7.7	238	14	5	10	5	6	4	5	7
13	30	61	56	106	6.6	245	17	6	8	9	8	8	10	8

図表６　表計算ソフトの重回帰分析



図表７　回帰分析の選択

(図表6～8：一例としてExcelの表示を示す)



図表８　回帰分析の設定

ここでは，重回帰分析の機能を追加したExcelで体力測定のデータ(高校1年生男子)で重回帰分析を行い，その出力を示す図表９。

回帰統計	
重相関 R	0.725
重決定 R2	0.525
補正 R2	0.510
標準誤差	0.335
観測数	136

	係数	標準誤差	t	P-値
切片	10.819	0.325	33.331	0.000
立ち幅跳び	−0.012	0.002	−7.648	0.000
ハンドボール投げ	−0.014	0.006	−2.367	0.019
握力得点	−0.040	0.024	−1.677	0.096
上体起こし得点	−0.025	0.020	−1.264	0.208

図表９　重回帰分析の出力結果1

※R^2は出力では数値表示されるが，一般に，寄与率というときは%表示，決定係数というときは数値表示をする。

なお，ここで示した図表は表計算ソフトの一例としてExcelで示したが，Excelの出力結果は必要に応じて，レポート等では書き直す必要がある。

「重相関R」→「重相関係数R」

「重決定R^2」→「寄与率(決定係数) R^2」

「補正R^2」→「自由度修正済みR^2」

この出力から，予測値と実測値の相関係数(重相関係数)が0.725であること，モデルの適合度を示す寄与率R^2が52.5%であることが分かる。「立ち幅跳び」のみを説明変数とした単回帰モデルの寄与率が45%だったので図表5，説明変数を増やしたことで，モデルの適合度(Yの変動の説明力)が上がったことが分かる。

分散分析表			
	自由度	変動	分散
回帰	4	16.230	4.05
残差	131	14.689	0.11
合計	135	30.919	

図表10　重回帰分析の出力結果2

回帰係数の推定値は，「係数」の列の箇所に出力される。また，分散分析表として，先に示した変動和の分解と自由度の分解が出力される図表10。ここで，回帰の自由度はモデルの項の数(説明変数の数)を表し，モデルの複雑度に相当するが，モデルの複雑度を上げれば，残差の自由度が少なくなることも留意する必要がある。

(2) 統計ソフトRでの重回帰分析の実行

重回帰分析をRで実行する場合は，以下のコードとなる。ここでは，前項のExcelで使用した高校1年生男子の体力測定のデータを用いる。Rでデータを読み込む場合は，Excel形式をCSV形式(変数名の制限よりX50m走としてある)にしてから読み込む(2行目)。4行目が実際に推測したい回帰モデル式を入力する部分である。5行目は，分析結果を要約するコードである。これは忘れないようにする必要がある。

①コード

```
01  # データの読み込み
02  high_male <- read.csv("high_male_data.csv")
03  # 4つの説明変数による重回帰分析
04  res <- lm(X50m走 ~ 立ち幅跳び + ハンドボール投げ + 握力得点 + 上体起こし得点, data = high_male)
05  summary(res)
```

②出力結果

Rで実行した回帰分析の結果を出力させる(①の5行目)。Rの出力は基本的には英語になる。前項のExcelの出力と対応させて理解してほしい。この部分のみであれば，Excelで重回帰分析の意味を理解させた後に，Rのコードの練習としてやってみる，また，重回帰分析の大事な用語の英語を学ぶ程度の位置付けになる。後述の変数選択をする場合は，Excelではできないので，Rも知っておく必要がある。

```
Call:
lm(formula = X50m走 ~ 立ち幅跳び + ハンドボール投げ + 握力得点 + 上体起こし得点, data = high_male)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66229 -0.22824 -0.03057  0.22797  0.99194

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    10.819417   0.324609  33.331  < 2e-16 ***
立ち幅跳び     -0.012005   0.001570  -7.648  3.89e-12 ***
ハンドボール投げ -0.014393   0.006081  -2.367  0.0194 *
握力得点       -0.040185   0.023969  -1.677  0.0960 .
上体起こし得点 -0.025489   0.020163  -1.264  0.2084
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3349 on 131 degrees of freedom
Multiple R-squared:  0.5249,Adjusted R-squared:  0.5104
F-statistic: 36.18 on 4 and 131 DF,  p-value: < 2.2e-16
```

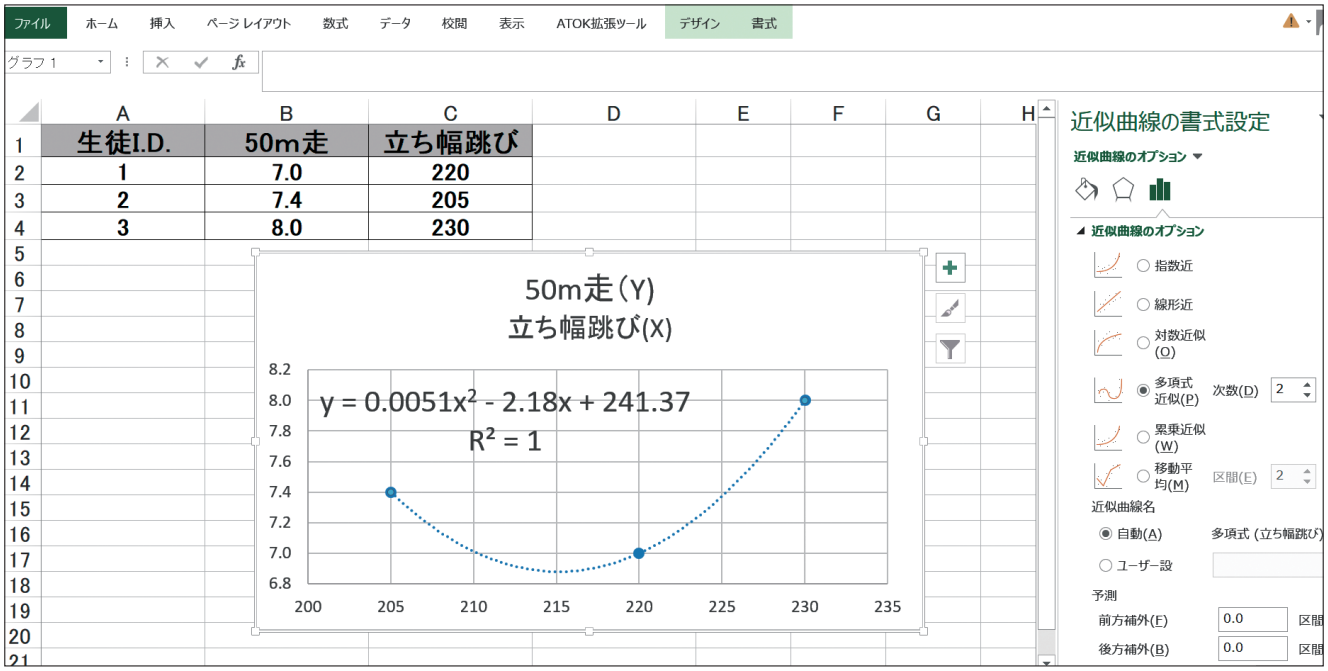
Estimate:係数の推定値
Residual standard error:
(残差)標準誤差
degrees of freedom
(残差)自由度
Multiple R-squared:
寄与率　R^2
Adjusted R-squared:
自由度修正済みR^2

5 モデル選択(モデリング)

複数の説明変数の中には，目的変数の予測に役立たないものが含まれている可能性がある。そこで，モデル式の構築にあたっては，説明変数の取捨選択が重要な課題となる。これを**変数選択**もしくは**モデル選択**(モデリング)という。その際，寄与率R^2ができるだけ大きくなることが望ましいが，説明変数を増やせば増やすほど単純に大きくなる。また，モデルに説明変数の二次項や三次項(曲線の当てはめ，非線形モデル)を追加することも説明変数を増やしたことと同じ効果となり，複雑なモデルほど寄与率R^2は大きくなる。しかし，欠点(**過剰適合**，**過学習**)も生じる。

例えば，データが2人分しかない体力測定のデータで散布図を作成し直線を当てはめた場合，必ず二つのデータ点が直線上に乗り，寄与率R^2は100%になる。3人分であれば，二次項を含め多項式でモデル化するとやはり寄与率R^2は100%となる図表11。しかし，これらのモデルでは，残差の自由度は0となり，標準誤差は無限大となり計算できない。つまり，新しいデータに対して予測力がまるでないことになる。したがって，ある程度の残差の自由度を残しつつ，寄与率R^2を上げることが望ましい。

そのため，重回帰分析では，残差の自由度を考慮した，下記のようなモデルの良し悪しを測る指標を参考にしながら，説明変数の選択(モデル選択)を注意深く行う作業が必要となる。



図表11　二次多項式モデルと寄与率R^2

◎自由度調整済み寄与率R^2

$$1-\frac{S_E/(n-p-1)}{S_T/(n-1)}$$

自由度調整済みの寄与率R^2が大きなモデルほど説明力のある良いモデルということになる。これ以外にも，モデル選択の基準として，AIC（赤池の情報量規準)などの指標がある。AICは小さいほど良いモデルとされる。

◎回帰係数の有意差の検定

体力測定の重回帰分析の出力の箇所には，各回帰係数が母集団上で0であるか否かの統計的仮説検定(帰無仮説 $H_0: b_j = 0$)の検定結果を示す有意確率(P-値)の列がある図表9。この値が，1%もしくは5%以下であれば帰無仮説は棄却され，その説明変数の値の変化が統計的に有意にYの値の変化に影響を与えるという対立仮説が採択されたことになる。この例では，「立ち幅跳び」は1%有意，「ハンドボール投げ」は5%有意

であるが，「握力」や「上体起こし」に有意差はない。つまり，有意差の出ない変数は「50m走」の値の変化に影響を与えていない可能性が高いと判断される。そこで，この二つの説明変数を外してモデルを作成し直してみる，更に新しい別の説明変数を加えて，同様な分析を繰り返すなどを行い，最適なモデルを探索する。

◎統計ソフトによる自動変数選択

Rなどの統計ソフトには，統計的な基準で変数選択を自動で行う機能がある。変数選択には，**総当法**(全ての変数の組み合わせを尽くす方法)，**変数増加法**(一つの変数からだんだんと変数を増やしていく方法)，**変数減少法**(全ての変数を採用したモデルから変数を減らしていく方法)，**変数増減法(ステップワイズ法**：変数を減らしたり増やしたりする方法)がある。変数を選択する基準が一つではなく，また，説明変数間の相関の強弱によって，どの方法を選ぶかで選択される変数，すなわち最終的に採択されるモデルの結果は異なる。何が変数として選択されたかを全くブラックボックスとし単純に予測モデルを構築したい場合は，自動変数選択の機能は便利ではあるが，説明変数が目的変数に与える効果に言及し要因分析を行う場合は，自動変数選択は安易に使用すべき手法ではない。参考までに，Rのコード(変数増減法)は以下となる。

```
01  # データの読み込み
02  high_male <- read.csv("high_male_data.csv")
03  # 全ての説明変数によるステップワイズ法（変数増減法）による重回帰分析
04  res <- lm(X50m走 ~ 立ち幅跳び + ハンドボール投げ + 握力得点 + 上体起こし得点 , data = high_male)
05  step(res)
```

ダミー変数

重回帰分析では，目的変数も説明変数も基本的には量的な変数である必要があるが，説明変数に質的な変数を用いることがある。この場合は，説明変数の質的な属性の有無を，0 と 1 の数値で対応させた変数（ダミー変数と呼ぶ）で表現し直し，ダミー変数を説明変数として重回帰分析を行う。目的変数が質的変数の場合は，ロジスティック回帰分析を行う。

演習 2

EXERCISE

「科学の道具箱」https://rika-net.com/contents/cp0530/contents/ から，体力測定のデータをダウンロードし，目的変数と説明変数を自身で決めて予測モデルを作成してみましょう。また，作成した予測モデルからどのようなことが分かるのか，説明してみましょう。

【参考文献・参考サイト】

- 「多変量解析法入門（ライブラリ新数学大系）」永田靖，棟近雅彦 著　サイエンス社(2001)
- 「図解でわかる回帰分析―複雑な統計データを解き明かす実践的予測の方法」涌井良幸，涌井貞美 著　日本実業出版社(2002)
- 「理科ネットワーク　デジタル教材　『科学の道具箱』」https://rika-net.com/contents/cp0530/start.html
- 「生徒のための統計活用～基礎編～」(生徒用，指導用) 渡辺美智子 他 著　総務省政策統括官(統計基準担当) 編　日本統計協会(2016)
- 「高校からの統計・データーサイエンス活用～上級編～」(生徒用，指導用) 渡辺美智子 他 著　総務省政策統括官(統計基準担当) 編　日本統計協会(2017)
- 「問題解決力向上のための統計学基礎―Excel によるデータサイエンススキル」迫田宇広，高橋将宜，渡辺美智子 著　日本統計協会(2014)
- 「実践ワークショップ Excel 徹底活用 統計データ分析 改訂新版」渡辺美智子，神田智弘 著　秀和システム(2008)
- 「文化情報学事典」渡辺美智子 他 編　村上征勝 監修　勉誠出版(2019)
- 「統計学Ⅲ：多変量データ解析法オフィシャルスタディノート」岩崎学，足立浩平，渡辺美智子，宿久洋，芳賀麻誉美 著　日本統計学会・日本行動計量学会 編　日本統計協会(2017)

学習活動と展開

学習活動の目的

- 課題発見とデータに基づく問題解決の枠組みにおける予測と予測モデル構築の意義，要因分析と制御の位置付けを理解する。
- 量的データの予測モデルの基本である重回帰モデルの概念や用語，活用のためのデータ構造を理解する。
- 予測と要因分析の違いを理解し，予測精度を上げるためのモデルの改良（モデリング）の方法を理解する。

学習活動とそれを促す問い

	問　い	学習活動
展開1	問題解決のための予測モデルの設計をしよう。	予測したい目的変数を決め，予測することが何に役立つのかを考える活動を行う。また，予測モデル構築のための説明変数を考えて，特性要因図などの図にまとめる活動を行う。
展開2	データを整理・整形し，実際に，重回帰分析を行ってみよう。	表計算ソフトや統計ソフトＲで重回帰分析を行い，予測式の意味を考える活動を行う。
展開3	予測モデルの改良をしてみよう。	説明変数の取捨選択を変数の有意性に着目して行い，いくつかのモデルを作成する。それらのモデルの説明力や予測力を寄与率や自由度修正済み寄与率など統計指標を適切に読み解いて考える活動を行う。

展開1

問　い

問題解決のための予測モデルの設計をしよう。

学習活動

- 問題の重要度を決める指標を目的変数とし，その変動要因を考え予測モデルを構築することが何に役立つのかを考える活動を行う。
- 予測モデル構築のための説明変数を考えて，特性要因図などの図にまとめる活動を行う。

指導上の留意点

- 生徒が興味を持てる内容で，背景や意味が分かるデータに結び付くよう指導する。
- 特性要因図など，予測のための設計図を生徒自らが説明し議論する場面を作る。

展開 2

問い

データを整理・整形し，実際に，重回帰分析を行ってみよう。

学習活動

- データを取得し，構造化されたデータ形式を確認する。
- 表計算ソフトの分析機能で，重回帰分析を実行する。
- 出力を読み取る回帰係数，寄与率，残差，標準誤差等の意味を説明できる。
- 表計算ソフトではなく，統計ソフト R を使う学習活動も考えられる。

指導上の留意点

- 個人よりもグループで相互に教え合いながら活動することを勧める。
- 統計指標については，グループ内で話し合いながら指標の有用性を納得させるとよい。統計指標の意味や使い方については，議論させることが理解につながる。

展開 3

問い

予測モデルの改良をしてみよう。

学習活動

- 推定された回帰係数の有意差検定の結果を読み取る（仮説検定の考え方と関連付ける）。
- 有意でない変数をモデルから外す意味を理解し，外したモデルで分析をやり直す。
- 元のモデルの出力（回帰係数，寄与率，残差，標準誤差等）と比較し，自由度や自由度調整済み寄与率を参照できるようにする。
- 統計ソフト R を使って，変数選択をしてみる。

指導上の留意点

- 個人よりもグループで相互に教え合いながら活動することを勧める。
- 統計指標については，グループ内で話し合いながら指標の有用性を納得させるとよい。
- 予測モデルの改良については，議論させることが理解につながる。

まとめ

まとめ

- 重回帰分析を活用し，具体的な問題解決の経験を通して，予測モデルの意義を理解させる。
- 分析結果を相互に発表させ，議論させる活動が好ましい。

学習

14 主成分分析と次元削減

研修内容

主成分分析の基本について学習し，主成分分析の意味，主成分（特徴量）の抽出と次元削減の方法について学習する。また，主成分分析の活用事例や機械学習の手法としての役割などについても学習する。

研修の目的

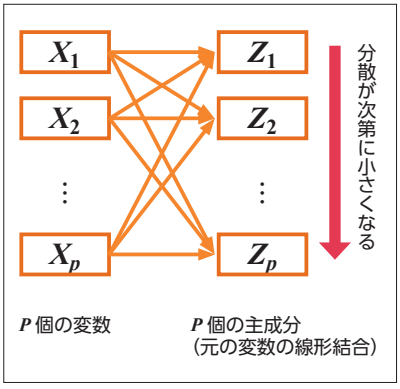
- 対象の特徴を表す多変量(高次元)のプロファイルデータから複数の変数をまとめて主成分(特徴量)を作成する方法を理解し，その意義を学習する。
- 具体的なデータで主成分の概念や実際の求め方，用語，活用(解釈)を，表計算ソフトを用いて理解する。
- Rを使った実際のデータに基づく分析の方法と出力の読み方，活用事例，次元の縮約方法を学ぶ。

この学習項目で使用するプログラミング言語は R です。

1 主成分分析とは？ ～新しい指標(特徴量)を効果的に作成する方法～

対象の特徴を一般にデータで表す際に，単一の変数だけではなく多くの変数を使うことで，対象の特徴をより詳細に分析することができる。学習13で挙げた生徒を対象とした体力測定のデータでは，体力に関する複数の種類の測定項目を使って生徒の体力の状況が調査されている。また，中古住宅のデータは，住宅の特徴を床面積や築年数など，やはり複数の変数で住宅の特徴が記述されている。しかし，変数の数が増えれば，データから個々の対象の特徴を総合的に捉えることが難しくなる。主成分分析を使えば，このような多変量(高次元)のデータに対して，変数の間の共分散や相関の強い変数同士をまとめて，個々の対象の違いを最も大きくするような主成分と呼ばれる新しい特徴量(変数)を作成することができる。その主成分(変数)軸を使って，個々の対象のポジショニングの把握や対象全体の分類を効果的に行うことができる。

具体的には，元のp個の変数$(X_1, X_2, \cdots, X_p)$から，情報を損失することなく線形結合(重み付きの合計)によって，p個の互いに独立な合成変数を主成分$(Z_1, Z_2, \cdots, Z_p)$として作成する。主成分は分散が最も大きくなるような順番で作成され，下位の主成分になるに従って，対象間で主成分の値(主成分得点)の変動(分散)が小さくなる図表１。そこで，対象の弁別に対して寄与の小さい下位の方の主成分を捨て上位の主成分のみ採用することで，元の個数pより少ない数で，対象の特徴をプロファイルすることが可能になる。このことを**次元削減(次元縮約)**といい，特に高次元の変数を扱う画像データの処理では，機械学習の一つの手法として使用されている。また，主成分分析によって，対象をうまく説明する新しい特徴量を見いだすこともできる。簡単な例で説明する。50人の生徒の成績の状況を表す5科目の試験に関する成績データがあるとする図表２。



図表１　主成分の作成

No.	国語	英語	数学	物理	化学
1	45	42	47	49	38
2	47	52	40	51	42
3	54	52	47	50	48
4	47	47	48	48	51
5	51	55	54	53	60
6	43	47	55	59	60
7	45	41	45	54	51
：	：	：	：	：	：

図表２　５科目の成績データ

5次元のデータを1次元にするために合計得点が計算され，50人の生徒を合計得点で順位付ける，ということはよく行われている。このとき，合計得点は生徒の総合能力を表す一つの特徴量と考えることができる。合計得点は各科目を同じ重み「1」で合計した変数である。

合計得点＝1×**国語**＋1×**英語**＋1×**数学**＋1×**物理**＋1×**化学**

ここで，各科目の係数を一般化して，a，b，c，d，eとして作成される合成変数を考えてみよう。

合成変数＝a×**国語**＋b×**英語**＋c×**数学**＋d×**物理**＋e×**化学**

a，b，c，d，eにいろいろな数値を与えると，単純な合計得点以外の合成変数をいろいろ作成することができる。主成分分析で作成する主成分とは，係数 a，b，c，d，eを50人の生徒の合成変数の値**(主成分得点)**が最もばらつくように，つまり，分散が最大となるように決定される一つの変数である。ばらつきが大きいことがその変数の情報量の大きさに対応する。つまり，生徒の成績のパターンの違いを最もよく表す特徴量となる。

最初に求められる主成分を第1主成分という。次に求められる第2主成分は，第1主成分と独立である(無相関)という条件の下で，分散が最大になるように求められる。このように求められる主成分同士は，互いに相関しない，情報が重複しない特徴量となる。

2 表計算ソフトで主成分を求めてみる

合成変数を作成する場合，二つの立場がある。

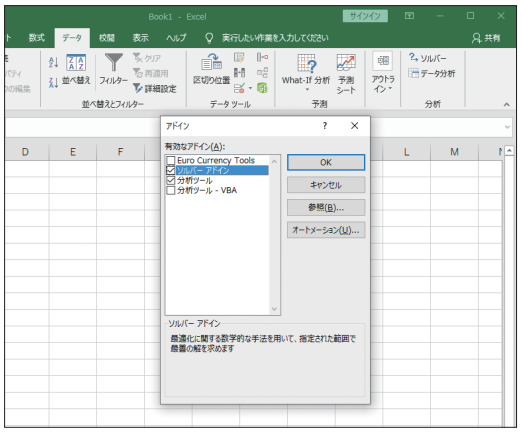
①5科目の得点のレベル(位置)をそろえる(平均を同じにする，平均からの偏差にする)。

②5科目の得点のレベル(位置)とばらつきの大きさの双方をそろえる(平均と標準偏差を同じにする，**基準化(標準化)**する，または，**偏差値**にする)。

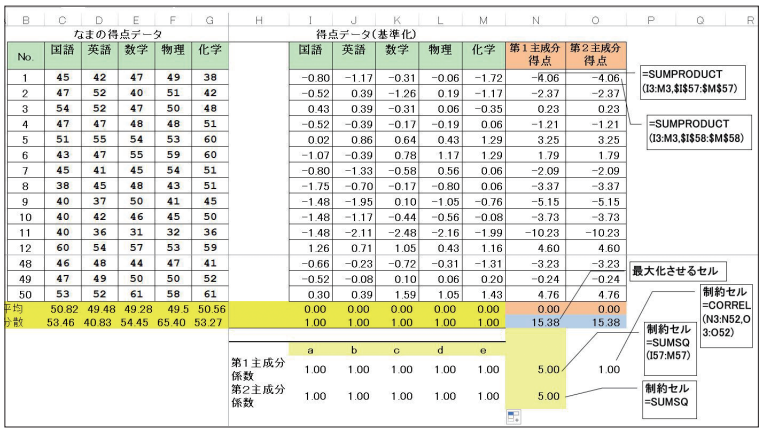
単位が異なる複数の変数を扱う場合は，②の基準化を行い，単位をそろえる必要がある。単位が同じ場合は，①と②の双方の立場で分析が可能である。①では，各変数の分散の大きさを考慮した係数が求まり，②では，各変数の分散の違いを無視した係数が求まる。

表計算ソフトを使用して分散を最大化する係数を求めるには，計算結果(主成分分析の場合は，分散)を目標値として設定し，制約条件を指定した上で変数のセル(係数のセル)を変化させ目標値を最大化させる係数を求める最適化機能を使う。ここでは，一例としてExcelで示す。Excelでの最適化機能は「ソルバー」と呼ばれる。「ソルバー」は，「ファイル」メニューの「オプション」から「アドイン」で，「ソルバーアドイン」を設定すれば，「データ」メニューから利用できる図表３(「挿入」メニューの「アドイン」から設定する場合もある)。

まず，成績データを②の立場で基準化したデータに変換し，基準化したデータを使って第1主成分，第2主成分の係数と各主成分得点を求めるためのシートを事前に準備する図表４。



図表３ ソルバーアドインの設定（一例としてExcelの設定を示す）



図表4 主成分係数をソルバーで求める

(1) 第 1 主成分を求める手順

❶それぞれの主成分の係数のセルを用意し，初期値（ここでは1）をあらかじめ入力しておく。

❷各生徒の，主成分係数によって求められる重み付き合計「主成分得点」の列を作り，SUMPRODUCT関数であらかじめ得点を全て計算し，主成分得点の平均と分散はAVERAGE関数やVARP関数で求めておく。

❸主成分の分散は係数の絶対値を大きくすればするほど大きくなる。そこで，全ての係数の2乗和を1とする制約条件の下で，分散を最大化させる必要がある。制約条件のセルをSUMSQ関数で作成しておく図表４。

❹第1主成分を求める。「データ」メニューから，「ソルバー」を選択し，「ソルバーのパラメータ」ダイアログでパラメータ設定をする図表５。

目的セル：第1主成分得点の分散を計算したセル
目標値：最大値
変数セル：第1主成分係数のセル
制約条件：係数の2乗和のセル=1
チェックボックス「非負数」はチェックしない

❺「解決」をクリックする。「ソルバーの結果」ダイアログが表示され，「OK」をクリックすると，主成分係数と得点，最大化された分散などの結果がシートに反映される図表６。

(2) 第 1 主成分の解釈

第1主成分の最大化された分散は，3.12である。

基準化された得点データの分散はそれぞれ1である。5科目の分散の合計「5」がデータ全体の分散の量と考えると，主成分1のみで，3.12 ／ 5=0.624，すなわち，全体の62.4％の情報が集約されたことになる。これを主成分の寄与率という。また，各科目がどのような重みで計算されたものかを主成分の係数の大きさから解釈すると，第1主成分は，やや理系科目に重きを置いた総合得点と考えることができる。この主成分が分散が大きくなるように数学的に導いた，50人の生徒の弁別性が最も大きくなる特徴量となる。

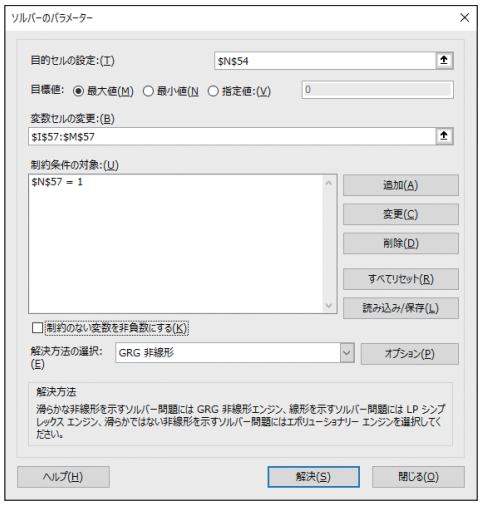
(3) 第 2 主成分を求める手順

❶第1主成分と第2主成分は独立という制約条件がある。そのため，それぞれの主成分得点の相関係数をCORREL関数で求めるセルを作っておく。

❷第2主成分係数に関しても，2乗和を1とする制約を課すため，SUMSQ関数で2乗和を計算するセルを作成する図表４。

❸ソルバーのパラメータ設定を行う図表７。

目的セル：第2主成分得点の分散を計算したセル
目標値：最大値
変数セル：第2主成分係数のセル
制約条件の追加：
第2主成分の係数の2乗和のセル=1
第1と第2主成分得点の相関係数=0
チェックボックス「非負数」はチェックしない



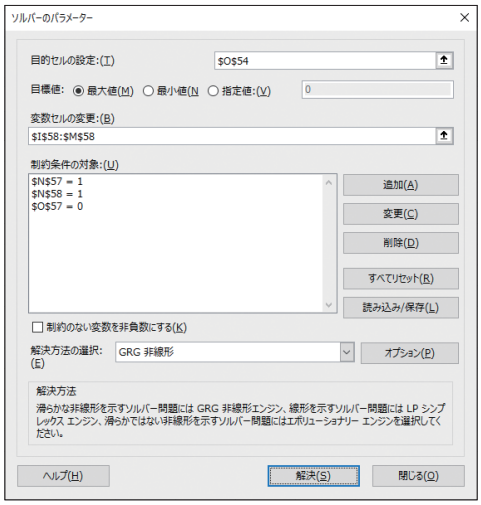
図表5　第1主成分を求めるソルバーのパラメータ設定

得点データ（基準化）

	国語	英語	数学	物理	化学	第１主成分得点	第２主成分得点
	−0.80	−1.17	−0.31	−0.06	−1.72	−1.78	−4.06
	−0.52	0.39	−1.26	0.19	−1.17	−1.13	−2.37
	0.43	0.39	−0.31	0.06	−0.35	0.01	0.23
	−0.52	−0.39	−0.17	−0.19	0.06	−0.48	−1.21
	0.02	0.86	0.64	0.43	1.29	1.53	3.25
	−1.07	−0.39	0.78	1.17	1.29	1.06	1.79
	−0.80	−1.33	−0.58	0.56	0.06	−0.81	−2.09
	−1.75	−0.70	−0.17	−0.80	0.06	−1.31	−3.37
	−1.48	−1.95	0.10	−1.05	−0.76	−2.12	−5.15
	−1.48	−1.17	−0.44	−0.56	−0.08	−1.50	−3.73
	−1.48	−2.11	−2.48	−2.16	−1.99	−4.61	−10.23
	1.26	0.71	1.05	0.43	1.16	2.01	4.60
	−0.66	−0.23	−0.72	−0.31	−1.31	−1.47	−3.23
	−0.52	−0.08	0.10	0.06	0.20	−0.03	−0.24
	0.30	0.39	1.59	1.05	1.43	2.26	4.76
	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1.00	1.00	1.00	1.00	1.00	3.12	15.38

	a	b	c	d	e	
第１主成分係数	0.33	0.41	0.50	0.47	0.49	1.00
第２主成分係数	1.00	1.00	1.00	1.00	1.00	5.00

図表6　第1主成分の出力結果



図表7　第2主成分を求めるソルバーのパラメータ設定

（図表5 ～ 7：一例としてExcelの表示結果を示す）

(4) 第 2 主成分の解釈

第2主成分の最大化された分散は，1.16となる図表8。主成分2の寄与率は，1.16 ／ 5=0.231，全体の23.1％の情報をもつ特徴量となる。主成分1と主成分2の寄与率の合計(主成分2までの累積寄与率)は，0.624+0.231＝0.855で，新しく作られた二つの特徴量で，もともとあった五つの変数の情報の85.5%をカバーしたことになる。これは，生徒の5科目の成績のパターンの特徴が二つの変数でだいたい読み取れることを意味しており，これが次元削減の意味である。また，第2主成分は，第2主成分係数から図表8，以下となる。

$$\text{第 2 主成分} = (0.68 \times \text{国語の基準化得点} + 0.50 \times \text{英語の基準化得点}) - (0.27 \times \text{数学の基準化得点} + 0.40 \times \text{物理の基準化得点} + 0.22 \times \text{化学の基準化得点})$$

これは，国語や英語に重みを置いた総合得点と，数学や理科に重みを置いた総合得点の差(対比)，すなわち，生徒の教科の興味・関心の方向性を識別する変数と考えることができる。

主成分分析の結果，第1主成分と第2主成分の累積寄与率が80%を超えており，生徒の5科目の成績は，全科目の総合的な得点と生徒の教科の興味・関心の方向性という二つの観点でほぼ特徴付けられることが分かる。また，各主成分の得点を見ることで，その二つの観点に関する生徒の数量的評価もできることになる。

得点データ(基準化)

	国語	英語	数学	物理	化学	第1主成分得点	第2主成分得点
	−0.80	−1.17	−0.31	−0.06	−1.72	−1.78	−0.64
	−0.52	0.39	−1.26	0.19	−1.17	−1.13	0.37
	0.43	0.39	−0.31	0.06	−0.35	0.01	0.63
	−0.52	−0.39	−0.17	−0.19	0.06	−0.48	−0.44
	0.02	0.86	0.64	0.43	1.29	1.53	−0.18
	−1.07	−0.39	0.78	1.17	1.29	1.06	−1.89
	−0.80	−1.33	−0.58	0.56	0.06	−0.81	−1.29
	−1.75	−0.70	−0.17	−0.80	0.06	−1.31	−1.19
	−1.48	−1.95	0.10	−1.05	−0.76	−2.12	−1.42
	−1.48	−1.17	−0.44	−0.56	−0.08	−1.50	−1.24
	−1.48	−2.11	−2.48	−2.16	−1.99	−4.61	−0.09
	1.26	0.71	1.05	0.43	1.16	2.01	0.50
	−0.66	−0.23	−0.72	−0.31	−1.31	−1.47	0.04
	−0.52	−0.08	0.10	0.06	0.20	−0.03	−0.49
	0.30	0.39	1.59	1.05	1.43	2.26	−0.77
	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1.00	1.00	1.00	1.00	1.00	3.12	1.16

	a	b	c	d	e		
第1主成分係数	0.33	0.41	0.50	0.47	0.49	1.00	0.00
第2主成分係数	0.68	0.50	−0.27	−0.40	−0.22	1.00	

図表8　第2主成分の出力結果

	国語	英語	数学	物理	化学	第1主成分得点	第2主成分得点
国語	1.000						
英語	0.685	1.000					
数学	0.301	0.499	1.000				
物理	0.185	0.388	0.808	1.000			
化学	0.369	0.440	0.767	0.750	1.000		
第1主成分得点	0.587	0.732	0.888	0.832	0.870	1.000	
第2主成分得点	0.733	0.540	−0.287	−0.433	−0.240	0.000	1.000

図表9　主成分負荷量
(図表8，9：一例としてExcelの表示結果を示す)

(5) 主成分負荷量 (主成分と元の変数との相関係数)

主成分の解釈に主成分係数を使う以外に，主成分と元の変数との相関係数(主成分負荷量または因子負荷量)を使うこともある。アドインされた「データ」メニューの「データ分析」にある「相関」で，相関行列が出力される。この場合，主成分負荷量は，相関行列の枠で囲った部分となる。相関係数は「数学Ⅰ」の「データの分析」の単元で学習しているので，生徒には分かりやすい指標である。

(6) 相関行列と共分散行列 (平均からの偏差データと基準化データ)

統計ソフトウェアの主成分分析の数学的な演算処理は，行列を用いて一度に全ての主成分に関する統計量が計算できるアルゴリズム(行列の固有値分解)を採用している。そのため，①平均からの偏差データに変換したデータに基づく主成分分析は，実際には，元の変数の分散共分散行列を固有値分解して求められている。②基準化データに変換したデータに基づく主成分分析は相関行列の固有値分解である。そのため，Rなど統計ソフトを使う場合は，分析者が，分散共分散行列か相関行列のどちらを対象にした分析を行うのかを指定しなければならないので，その意味を知っておく必要がある。また，統計ソフトウェアから出力される各主成分の分散を意味する統計量は固有値，各主成分の係数は，固有ベクトルという名称で出力されるので，用語と意味の対応付けも知っておく必要がある。

3 分析事例（野球選手の成績評価）

主成分分析の簡単な事例として，2017年の日本プロ野球の規定打席数に達した両リーグの55選手の打撃成績のデータを用いた例を紹介する。「打率」，「得点数」，「安打数」，「二塁打数」，「三塁打数」，「本塁打数」，「打点数」，「盗塁数」の8変数を使い，相関行列に基づく主成分分析を行った。その結果が右の表である図表10 図表11。

例として，第2主成分までを採択して，結果の解釈を行ってみる。第2主成分までの累積寄与率が約70%強であることから図表10，選手の元の8指標での打撃評価の変動の約70%が二つの縮約された主成分で説明できることになる。第1主成分の寄与率は約40%で図表10，主成分係数はいずれも同じ＋の符号であることから図表11，打率，安打，得点，二塁打を中心に「総合活躍度」を示した指標と解釈できる。第2主成分は寄与率が約30%で図表10，主成分係数の符号と絶対値の大きさから図表11，打点と本塁打，三塁打と盗塁の成績の対比を表した指標と考えられる。つまり，第2主成分は打撃のスタイルを意味する指標で，第2主成分得点が＋に高ければ長打力を活かすタイプの選手，－に低ければ走力を活かした選手と評価できる。

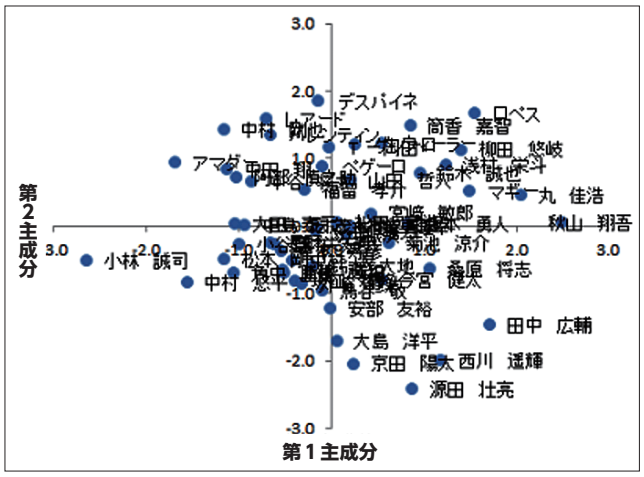
第1主成分得点と第2主成分得点で選手のポジショニングを示した散布図を作成すれば，中心円から離れた選手ほど，二つの指標（主成分）の観点で特異であることが分かる図表12。上位と下位の主成分を組み合わせて散布図に示す場合は，基準化した主成分得点を使用することもある。

主成分	固有値（分散）	寄与率	累積寄与率
1	3.335	0.417	0.417
2	2.354	0.294	0.711
3	1.055	0.132	0.843
4	0.460	0.057	0.900
5	0.381	0.048	0.948
6	0.175	0.022	0.970
7	0.141	0.018	0.987
8	0.100	0.013	1.000

図表10　主成分の分散（固有値）と寄与率

変数名	第1主成分	第2主成分
三塁打	0.153	−0.465
盗塁	0.254	−0.427
打率	0.385	−0.104
安打	0.486	−0.102
得点	0.477	0.046
二塁打	0.450	0.051
打点	0.275	0.514
本塁打	0.149	0.558

図表11　主成分負荷量

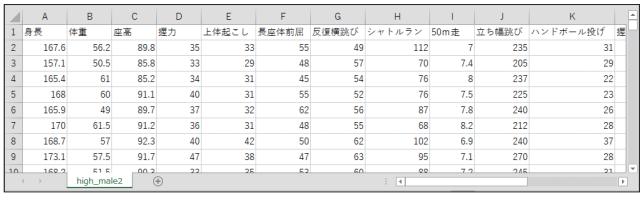


図表12　第1主成分と第2主成分得点の散布図

出典：「日本野球機構 シーズン成績2017」のデータを加工して作成
http://npb.jp/bis/2017/leagues/index_pl.html

4 Rで体力測定のデータを分析してみよう

主成分分析のRのコマンドはprcompで，主成分の英語，principal componentに因る。学習13でも紹介した体力測定のデータで，4列目の「握力」から11列目の「ハンドボール投げ」までの8変数を使って主成分分析を行う図表13。Rのコードは下記となる。ここで，共分散行列に基づく分析の場合は，コマンド内のオプションで scale = F を，相関行列のときは scale = T を入力する。体力測定の場合は，変数の単位も異なることから相関行列に基づく分析を行う。出力が英語であることから，赤字でその意味を付している。



	A	B	C	D	E	F	G	H	I	J	K
1	身長	体重	座高	握力	上体起こし	長座体前屈	反復横跳び	シャトルラン	50m走	立ち幅跳び	ハンドボール投げ
2	167.6	56.2	89.8	35	33	55	49	112	7	235	31
3	157.1	50.5	85.8	33	29	48	57	70	7.4	205	29
4	165.4	61	85.2	34	31	45	54	76	8	237	22
5	168	60	91.1	40	31	55	52	76	7.5	225	23
6	165.9	49	89.7	37	32	62	56	87	7.8	240	26
7	170	61.5	91.2	36	31	48	55	68	8.2	212	28
8	168.7	57	92.3	40	42	50	62	102	6.9	240	37
9	173.1	57.5	91.7	47	38	47	63	95	7.1	270	28

high_male2

図表13　体力測定のデータ（high_male2.csv）

①入力コード

```
# 身長，体重，座高を除き，握力からハンドボール投げまでの 8 変数を使った主成分分析
# 相関行列
# データセット ( 第 4 列から第 11 列 ) のインポート
high_male2 <- read.csv("high_male2.csv")
high_male3 <- high_male2[c(4:11)]
# 相関行列に基づく主成分分析 prcomp の実行
(res2 <- prcomp(high_male3, scale=T))
summary(res2)
pc <- res2$x
# 主成分得点のファイルへの掃き出し
write.csv(pc, file = "pca_cor.csv")
```

②出力結果

```
Rotation(nxk)=(8x8):主成分の係数
                      PC1        PC2        PC3        PC4        PC5        PC6        PC7        PC8
握力            0.3252018  0.2682176 -0.53297421  0.39993408 -0.3653663 -0.31441687  0.34209544
                                                                                    -0.17004275
上体起こし      0.3141190  0.4351668  0.42225757  0.40834395  0.4032249 -0.33321281 -0.29431157  0.08168542
長座体前屈      0.3077864  0.3745785  0.01503113 -0.75987597 -0.2411453 -0.28776668 -0.10238851  0.18941208
反復横跳び      0.3933948  0.1203619  0.05183489 -0.20404673  0.4967487  0.35638527  0.61198108 -0.19529718
シャトルラン    0.3132617 -0.4444223  0.59760197  0.01703693 -0.3900527 -0.21759749  0.17541898 -0.34157859
X50m走         -0.4057185  0.4620511  0.11729178 -0.10636452 -0.0709927  0.04215936 -0.08597965
                                                                                    -0.76329592
立ち幅跳び      0.3681042 -0.3669386 -0.40018514 -0.13933339  0.3055848 -0.10049579 -0.50594605 -0.43684157
ハンドボール投げ 0.3844997  0.1955680  0.06075045  0.15245958 -0.3852838  0.72184877 -0.34234695  0.01636705
```

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8
Standard deviation      1.968  0.9525  0.9096  0.85538  0.73271  0.68576  0.6400  0.49495  主成分の標準偏差
Proportion of Variance  0.484  0.1134  0.1034  0.09146  0.06711  0.05878  0.0512  0.03062  寄与率
Cumulative Proportion   0.484  0.5974  0.7008  0.79229  0.85940  0.91818  0.9694  1.00000  累積寄与率
```

③次元削減の基準

主成分の数をどこまで採用するのかについては，下記に基づく。

- **累積寄与率**：目安は累積寄与率が70%や80%以上になるくらいまでを採用する。
- **固有値(分散)が1以上(カイザー基準)**：固有値が1より大きい主成分を採用(相関行列を分析した場合)，寄与率が(1/変数の数)×100（%)以上になる主成分までを採用(共分散行列を分析した場合)する。
- **スクリープロット**：固有値(分散)を大きい順に左から折れ線で結んだグラフ(スクリープロット)に対して，分散の減少量が小さくなる(なだらかな減少になる)前までの主成分を採用する。

演習 1

EXERCISE

体力測定データの主成分分析結果の数値を読み取り，採用する主成分数の決定とその主成分の解釈をしてみましょう。また，「科学の道具箱」から体力測定のデータをダウンロードし，主成分分析を実行してみましょう。

【参考文献・参考サイト】

- 「主成分分析 ― 講座 情報をよむ統計学〈8〉」上田 尚一 著　朝倉書店(2003)
- 「図解入門よくわかる多変量解析の基本と仕組み」山口 和範 著　秀和システム(2004)
- 「理科ネットワーク　デジタル教材　『科学の道具箱』」https://rika-net.com/contents/cp0530/start.html
- 「日本野球機構　シーズン成績」(2017) http://npb.jp/bis/2017/stats/
- 「高校からの統計・データーサイエンス活用～上級編～」(生徒用，指導用) 渡辺美智子 他 著　総務省政策統括官(統計基準担当) 編　日本統計協会(2017)
- 「問題解決力向上のための統計学基礎―Excel によるデータサイエンススキル」迫田宇広 , 高橋将宜 , 渡辺美智子 著　日本統計協会(2014)
- 「実践ワークショップ Excel 徹底活用 統計データ分析 改訂新版」渡辺美智子，神田智弘著　秀和システム(2008)
- 「文化情報学事典」渡辺美智子 他 編　村上征勝 監修　勉誠出版(2019)
- 「統計学Ⅲ：多変量データ解析法オフィシャルスタディノート」岩崎学，足立浩平，渡辺美智子，宿久洋，芳賀麻誉美 著　日本統計学会・日本行動計量学会 編　日本統計協会(2017)

学習活動と展開

学習活動の目的

- 対象の特徴を表す多変量（高次元）のプロファイルデータから複数の変数をまとめて主成分（特徴量）を作成する方法を理解し，その意義が分かる。
- 具体的なデータで主成分の概念や実際の求め方，用語，活用（解釈）について，表計算ソフトを用いて理解する。
- 具体的なデータで表計算ソフトやＲを使った実際の分析の方法と出力の読み方，次元の縮約方法を理解する。

学習活動とそれを促す問い

	問　い	学習活動
展開1	表計算ソフトの「ソルバー」を使って，主成分分析を実行してみよう。	「ソルバー」を使用し，具体的に作業を行うことを通して，主成分が元のデータからどうやって作成されるのかを学習する。
展開2	主成分分析が使われている事例をインターネットで検索してみよう。	Google Scholar などの検索機能を使って，主成分分析が様々なテーマの実証分析に広く活用されていることを学習する。
展開3	Ｒのコードやフリーの主成分分析実行環境（SAS University Edition, HAD など）を使って，実際のデータで主成分分析をやってみよう。	実際のデータで分析を実行し，出力の解釈の方法を学ぶ。

展開 1

問　い

表計算ソフトの「ソルバー」を使って，主成分分析を実行してみよう。

学習活動

- 複数の量的データで特徴付けられる対象に対して，情報を集約した主成分の作り方を学ぶ。
- ５教科の成績データから，情報の損失を最小限に抑えて次数（次元）を削減するため，分散を最大化する係数を求める「ソルバー」の機能を体験する。
- 主成分の寄与率，累積寄与率，係数，負荷量，主成分得点などの意味と主成分の解釈を学ぶ。

指導上の留意点

- 平均からの偏差，基準化，主成分への変換など，変数変換すること，できること，それぞれの目的と有用性に気付くように指導する。
- グループ学習で進める。

展開 2

問い

主成分分析が使われている事例をインターネットで検索してみよう。

学習活動

- 自身が興味のあるテーマやキーワード（ex. スポーツ，生徒指導，サッカー，リーダーシップなど）と主成分分析をかけて，検索する。
- 関連する論文の中で，分析対象は何か，変数は何か，主成分分析を適用した目的は何か，何が分かったのかなどを読み取る。
- 主成分の寄与率，累積寄与率，係数，負荷量，主成分得点などの結果数値がどのように活用されているのかを調べる。

指導上の留意点

主成分分析の適用は，かなり広範囲で多様なジャンルに及んでいる。市場調査や実験データの解釈にも使われている。先入観等で決め付けることなく，自由にテーマやキーワードを選択させ，グループで結果を共有させるよう指導する。

展開 3

問い

R のコード，もしくはフリーの主成分分析の実行環境（SAS University Edition， HAD など）を使って，実際のデータで主成分分析をやってみよう。

学習活動

- データを取得し，主成分分析を実際に行う。
- 出力された主成分の寄与率，累積寄与率，係数，負荷量，主成分得点などの結果数値を読み取り，分析結果をまとめ，発表する。

指導上の留意点

現実のデータ分析に際し，主成分の解釈は難しい。最初から教えるのではなく，グループで議論させ相互に理解が深まる場面を作るように指導する。

まとめ

まとめ

- 主成分分析が広範な領域で活用されていること，複数の変数をまとめることで対象の特徴を効率的に捉えることができること，次元削減の考え方などを理解させる。
- 分析結果を相互に発表させ，議論させる活動が好ましい。