基礎情報教育: データ科学入門
第 1 章 社会と教育における変化

T. MIYAGUCHI

Naruto Universality of Education

データ科学・講義 (木曜) および実習 (火曜)

担当者 (数学コース・宮口)

アウトライン

第 1 章: 社会と教育における変化

第 2 章: データ科学とは?

第 3 章: データ分析の基礎

第 4 章: 可視化

第 5 章: データ分析実習

授業中の質問

講義 (木曜) は対面で実施します. スライドに細かい図等があるため, スライドの **pdf** を手元で見ながら受講してください. また Moodle のチャットを通した質問も受け付けます (質問には加点有).

配布物

- 講義スライドと各種データを Moodle にアップしてあります.

成績について

① **[講義 (木曜)]** 小テスト **(moodle)**

授業中に実施

② **[実習 (火曜)]** 授業後に課題提出 **(moodle)**

授業の翌日 (水曜) **18:00** まで

に提出すること.

講義全体のアウトライン

- 第 1 章: 社会と教育における変化
  - 仮説駆動とデータ駆動
  - データに基づく思考や判断
- 第 2 章: データ科学とは?
  - 統計学・計算機科学とデータ科学
  - AI・機械学習・倫理
- 第 3 章: データ分析の基礎
  - データとは?
  - 代表値・散らばりの指標・関係性の指標
- 第 4 章: 可視化
  - 可視化の必要性
  - 量の表現・割合の表現・分布の表現・関係性の表現・系列の表現
- 第 5 章: データ分析実習
  - 問を立てよう・統計量と可視化
  - 機械学習に挑戦・データ分析の実践
- 第 6 章: 様々な話題

授業の背景

講義・実習で使用するデータは次のページに置いてある:

moodle 基礎情報・データ科学

この章では次の 3 点に着目して授業の背景について考えていく.

(1) 統計教育と情報教育の充実

(2) 近年のコンピュータ技術の発展に伴なって, 人間が思考したり何かを発見するプロセスにも変化が現われている. 特に

- 仮説駆動
- データ駆動

という 2 つの思考様式を通して, 近年の傾向を理解していこう.

(3) データに基づいて考えたり判断することが必要になってきている.

統計教育の充実 (日本)

統計教育

小学校: 算数

- 円グラフ, 帯グラフ
- 代表値 (平均値・中央値・最頻値)
- 度数分布表とグラフ

中学校: 数学

- 累積度数
- 四分位範囲・箱ひげ図

高校: 数 I

- 分散・標準偏差・散布図・相関係数

中学校数学 指導要領 $H29.7$

ヒストグラムや相対度数などを手作業で作成したり求めたりすることは, その必要性と意味を理解するために有効であるが, 作業の効率化を図り, 処理した結果を基にデータの傾向を読み取ることを中心とする学習においては, コンピュータなどを積極的に利用するようにする.

高等学校数学 指導要領 $H30.7$

多くのデータを取り扱う場合や実験においては, コンピュータなどの情報機器を積極的に用いるようにすることが大切である.

情報教育の充実 (日本)

情報教育

小学校

- **Scrach** 等を用いたプログラミング教育の必修化

中学校: 技術・家庭 (技術分野)

- プログラミングに関する内容を充実

高校: 情報科

- 情報 **I** の必修化 (全ての生徒がプログラミングやシミュレーションなどを学ぶ).
- データ分析や **AI** の分野で最近注目されているプログラミング言語 **Python** の採用 (本講義で使用するグラフのほとんどは **Python** で作成したものである).

NEXT 統計教育と情報教育の充実が図られている理由は?

社会でおきている変化

Internet of Things (IoT)
モノのインターネット

人と人がつながるインターネット **(Facebook, Twitter, Line...)** に対し，自動車や家電のような「モノ」をインターネットにつなげて活用することを，**IoT** と呼ぶ.

例 1 エアコンをインターネットにつなげれば，遠隔操作して帰宅時間に合わせて空調を整えることは **IoT** の一例です **(スマートハウス)**。

例 2 車をインターネットに接続することで，混んでいる道路の状況を正確に把握できるようになり、利用者にとっての利便性が向上します。

Society 5.0

- **Society 1.0** 狩猟
- **Society 2.0** 農耕
- **Society 3.0** 工業
- **Society 4.0** 情報
- **Society 5.0**

  仮想空間と現実空間を高度に融合させたシステムにより、経済発展と社会的課題の解決を両立する新たな未来社会.
  鍵となる技術として，**IoT**，ビッグデータ，人工知能 **(AI)**，ロボット.

日本の学校教育の現状

右図は PISA(2015) における ICT 活用調査の結果の 1 つです.

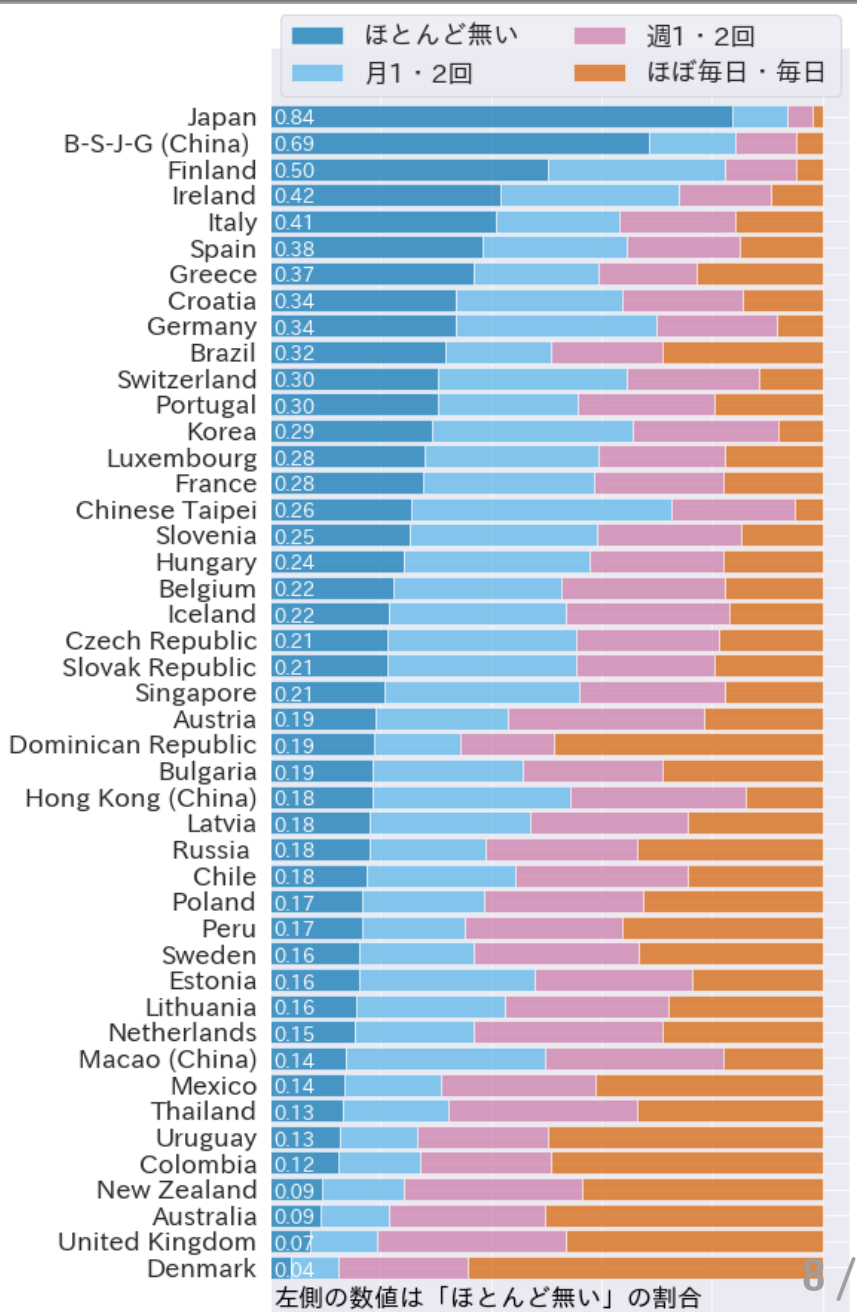
調査の質問は「コンピュータを使って宿題をしますか?」日本の生徒の 84% は「ほとんどない」

Take Home Message

コンピュータを使えば良いというわけではありませんが, 日本の大学生はコンピュータがとても苦手です.

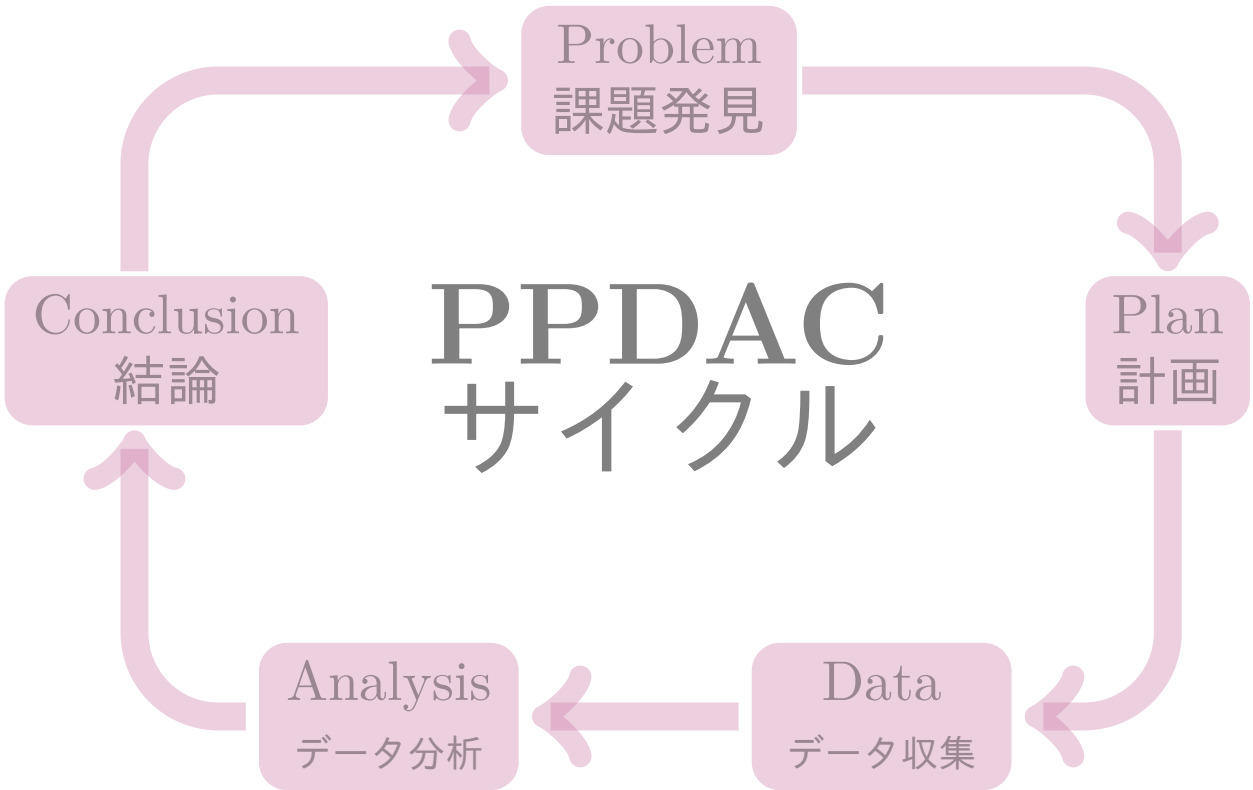
http://www.oecd.org/pisa/data/2015database/

右のグラフを作成する際に, 無回答を除き, 有効回答のみで 100% になるようにした.



PPDAC サイクル

データ分析活動の一般的な流れは **PPDAC** サイクルで表される (下図の時計周りのサイクル).



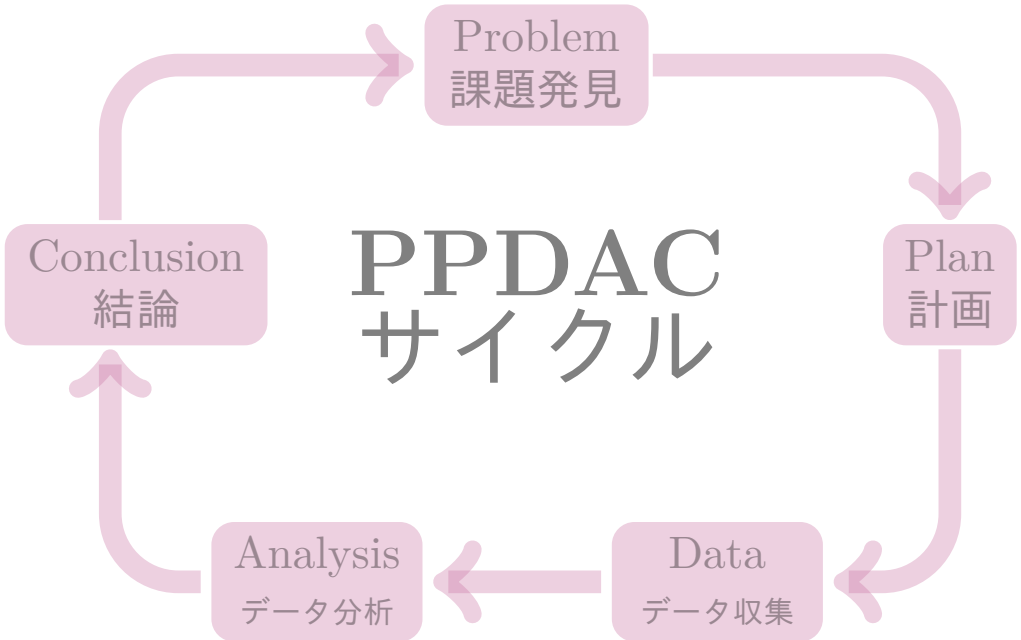
次節では,「問を立てること」について詳しく見てみよう.

Take Home Message

「結論」といっても課題が完全に解決することはまれである. 多くの場合, 部分的な解決にとどまり, さらなる課題が見えてくることが多い.

すなわち「収集したデータやその分析」から「新たな課題」(問や疑問) を見出すことが非常に重要である.

補足: PPDAC サイクルを回すには?



興味ある実データを得る方法として，「相互アンケート」がある．

相互アンケート

① クラスをグループ分け

② グループ毎にアンケートを作成

③ 他グループのアンケートに答える

④ アンケート結果の分析

⑤ 発表

PPDAC サイクルが回る条件

- 実データ
- 興味・好奇心を感じるデータ
- 比較的複雑なデータ (多変量データ)

補足: 架空のデータと実データ

架空のデータ

- 概念の理解には有効
- 準備が楽
- 授業の結論をコントロールできる (ただし, それが良いことなのかどうかは慎重に考えるべき)

実データ

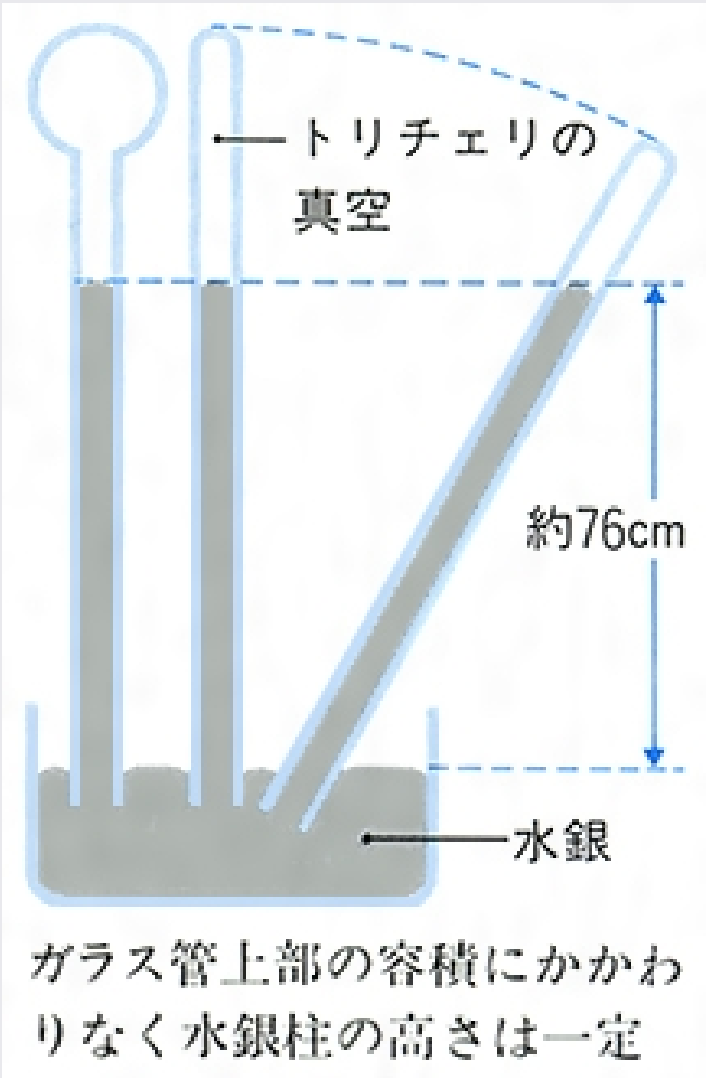
- 興味や関心を引き出すことができる
- 準備が大変
- 授業の結論が曖昧になりがち

Point

実データを授業で用いるのは大変ですが, 好奇心や探究心の育成の必要条件だと思います.

トリチェリの実験 1

トリチェリの実験 (中学理科)



出典: 小学館 日本大百科全書

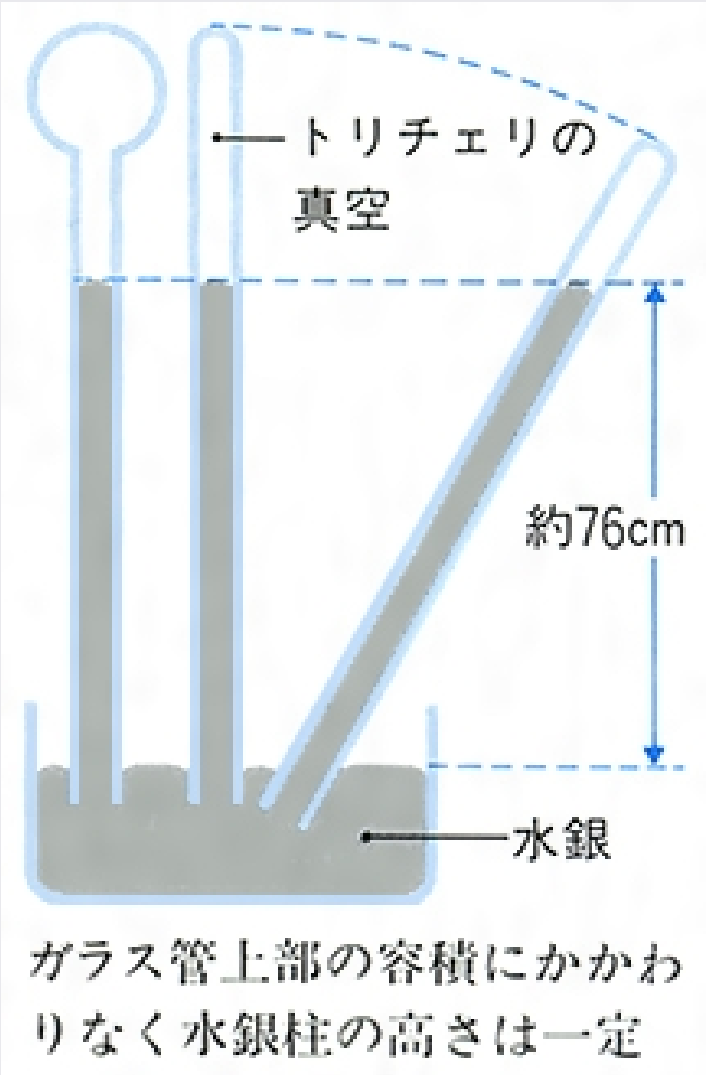
片方が閉じた長いガラス管に水銀を満たし，同じく水銀を入れた容器に，ガラス管の開口部を差し込む．

すると，水銀柱の上端は高さ約 $76\,\mathrm{cm}$ の所まで下がり静止する．ガラス管の上部には真空ができている．

この事実は 1643 年 トリチェリによって発見された．

トリチェリの実験 2

トリチェリの実験 (中学理科)



出典: 小学館 日本大百科全書

トリチェリがこのことを発見した時代 (17 世紀), 大気や真空については良く分かっていなかったが, トリチェリは水銀がガラス管の中を下がって来ないのは,

大気には重さがあり, それが容器の水銀表面を押して, ガラス管内に水銀を押し上げているから

という推論をした (もちろん今では正しいことが分かっている).

パスカルの実験 1

水銀の密度は約 $13.6\mathrm{g/cm^3}$，水銀柱の高さは約 $76\,\mathrm{cm}$ である．

トリチェリの仮説が正しいとすると

$1\mathrm{cm}^2$ $(1\,\mathrm{cm} \times 1\,\mathrm{cm})$ 当りに約

$$13.6\ \mathrm{g/cm^3} \times 76\ \mathrm{cm} \fallingdotseq 1\ \mathrm{kg/cm^2}$$

の重さ (大気の重さ) がかかっていることになる．例えば，手の平 $(5\mathrm{cm} \times 20\mathrm{cm} = 100\mathrm{cm}^2)$ 程度の面積には約 $100\mathrm{kg}$ の重さがかかっている計算だ．

空気が何なのか分かっていなかった当時，これはとても受け入れがたい仮説だったはずだ．

しかし，この仮説が正しいと信じてさらに実験をしたのがパスカルである (「人間は考える葦である」と言った人)．

パスカルの実験 2

問題

トリチェリの仮説

大気には重さがあり，それが容器の水銀表面を押して，試験管内に水銀を押し上げている

が正しいことを示すにはどのような実験をすれば良いだろうか?

もし，水銀柱が下がらないのが大気の重さが原因ならば，高い所で同じ実験をすれば，水銀柱の高さは下がるはずだ，とパスカルは考えた．

そこでピュイ・ド・ドームという高い山で，トリチェリの実験をやってみた (実際に山に登って実験したのはパスカルの義兄のペリエ)．実験結果は，パスカルの予想通りだった．

パスカルのデータ

表: パスカルのデータ

麓からの高度 [m]	水銀柱の高さ [m]
0	71.0
13	70.9
52	70.4
290	67.5
970	62.6

江沢 洋 『だれが原子をみたか』より抜粋.

右寄せにして数値を書くと値の大小が分かり易い (ヒストグラムと同じ原理)

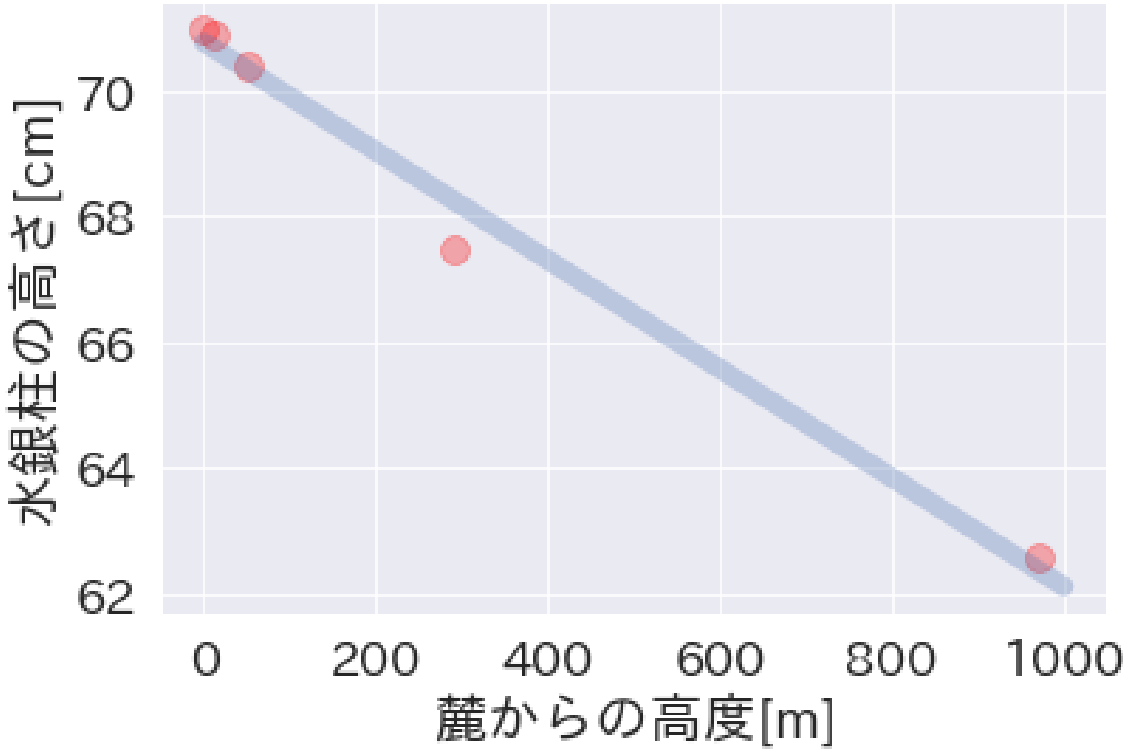


図: パスカルのデータをプロットした. 直線は最小二乗法で求めた回帰直線.

表のメリット: 詳細な数値が分かる

グラフのメリット: 関数関係が明確になる (ここでは直線的な関係が見える)

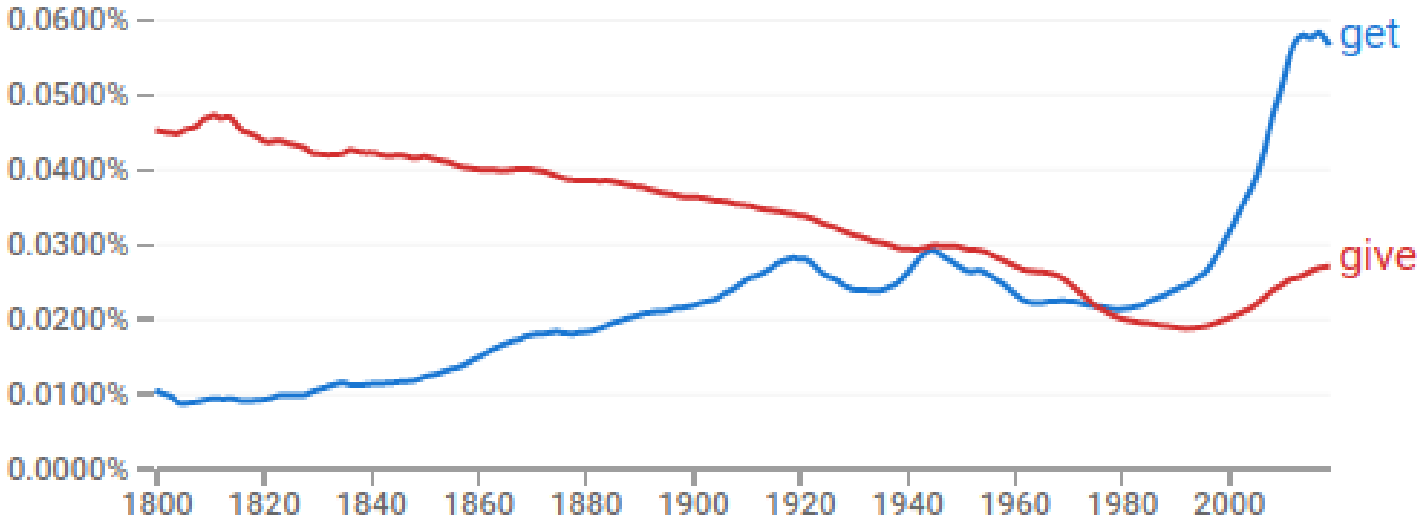
Google Ngram Viewer: get と give

Google Ngram Viewer を用いると **1800** 年から **2019** 年までの言葉の (書籍における) 使用頻度を調べることができる. 例えば,

get (得る) **give (与える)**

の使用頻度を比較すると, **get** が増加傾向にあり, **give** が減少傾向にあることが分かる (人間は利己的になってきている?):

縦軸の値はそれぞれの年に出版された書籍に含まれる全単語の中に占める検索ワードの割合



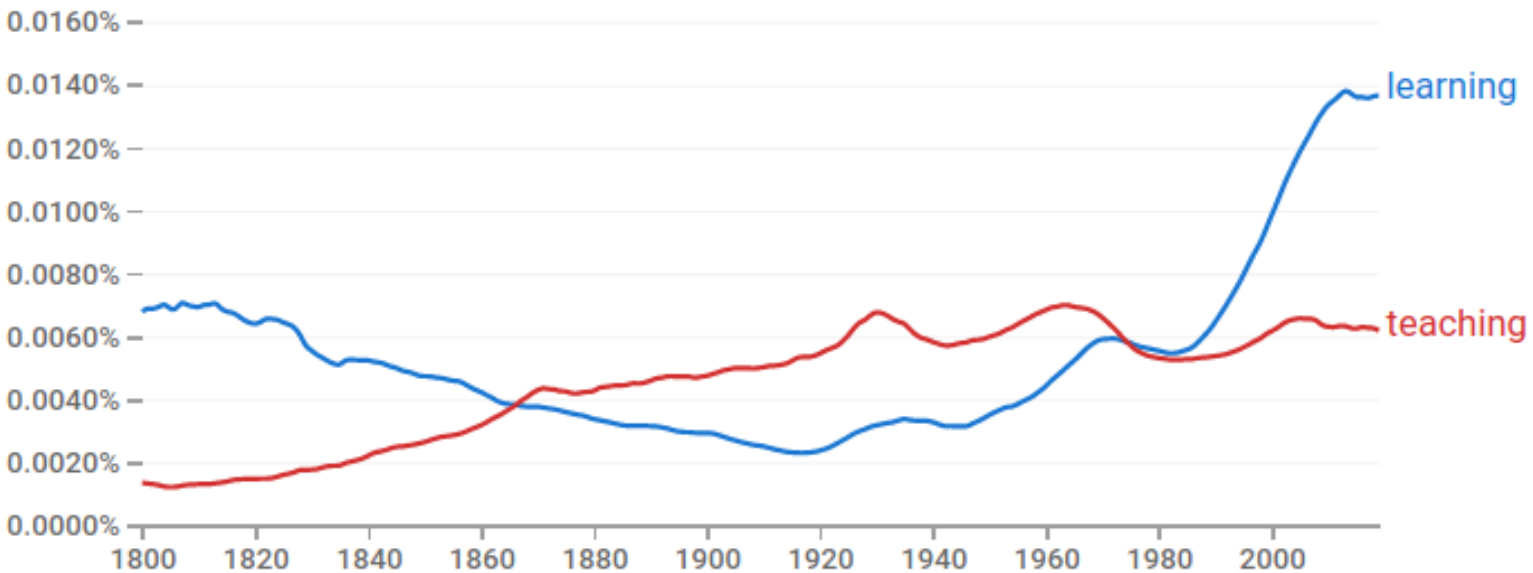
Greenfield は, 都市部に住む人口が増加したこととの関連を指摘している **(Greenfield 2013, The Changing Psychology of Culture From 1800 Through 2000).**

Google Ngram Viewer: Learning と Teaching

博士過程の学生 (小学校の元先生) の解答例を紹介する:

learning (学ぶ) **teaching (教える)**

の使用頻度を比較すると, **learning** が増加傾向にあり, **teaching** が減少傾向にあることが分かる:



この結果からどのようなことが予想できるか? (詰め込みから自発的学びへの転換?)

Google Ngram Viewer: 問を立てること

Google Ngram Viewer の例から分かることは,

データが先にある
もしくは
容易に入手できる

ということである. もちろん, 常にそうとは限らないけれど, そういうことが多くなっているようだ. すなわち,

データ駆動型

の活動が重要になってきている.

問題 (実習)

Google Ngram Viewer を用いて, いくつかの言葉の使用頻度を比較せよ.

その結果からどのような洞察が得られるか?

考えてみると分かるように, このような漠然とした問に答えることは容易ではない. このような活動を行うには,

- 様々な現象 (自然現象・社会現象) に対する探求心・好奇心
- 興味深い「問」を立てることができる創造性

を育てることが重要である.

NBA 選手のデータ

問

授業のページ

moodle 基礎情報・データ科学
「NBA 選手のデータセット」

から 「NBA 選手のデータセット 1」 をダウンロードしよう．このファイルは csv ファイルであり，Excel で開くことができる．

このファイルには，1950 年から 2017 年までの期間に活躍した NBA 選手の身長・体重・生年・出身大学などのデータである．このデータから問を立ててみよ．

csv ファイルは，データを保存・記録する際によく用いられるファイルフォーマット．

上のデータは以下のサイトからダウンロードした:

https://www.kaggle.com/drgilermo/nba-players-stats

www.kaggle.com では，さまざまなデータ分析のコンペティションが行なわれている (成績が良ければ賞金ももらえる).

NBA 選手のデータセット 1

このファイルには次のような情報が含まれている:

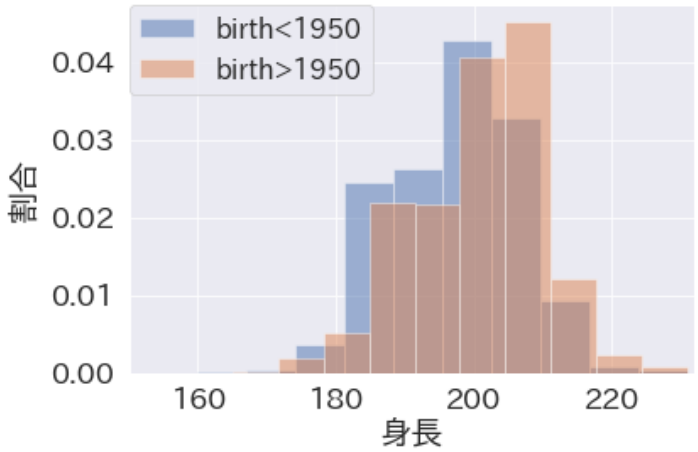
選手名	身長	体重	出身大学	生年	出生地 (市)	出生地 (州)
C. Armstrong	180	77	**Indiana University**	1918		
Cliff Barker	188	83	**Univ of Kentucky**	1921	**Yorktown**	**Indiana**
L. Barnhorst	193	86	**Univ of Notre Dame**	1924		
Ed Bartels	196	88	**N.Carolina State Univ**	1925		
Ralph Beard	178	79	**Univ of Kentucky**	1927	**Hardinsburg**	**Kentucky**
Gene Berce	180	79	**Marquette Univ**	1926		
⋮						

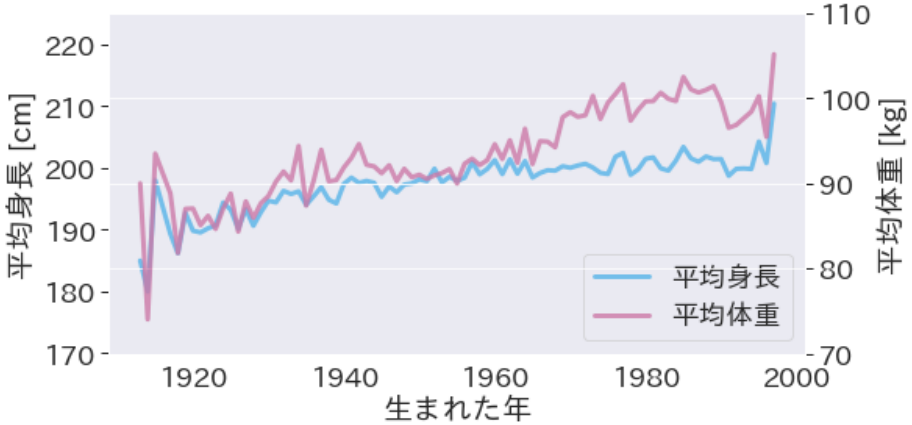
合計 3921 名の **NBA** 選手のデータである. 現実のデータによくあることだが, 値が欠損している箇所がある.

NBA 選手のデータ: 分析例 1

問

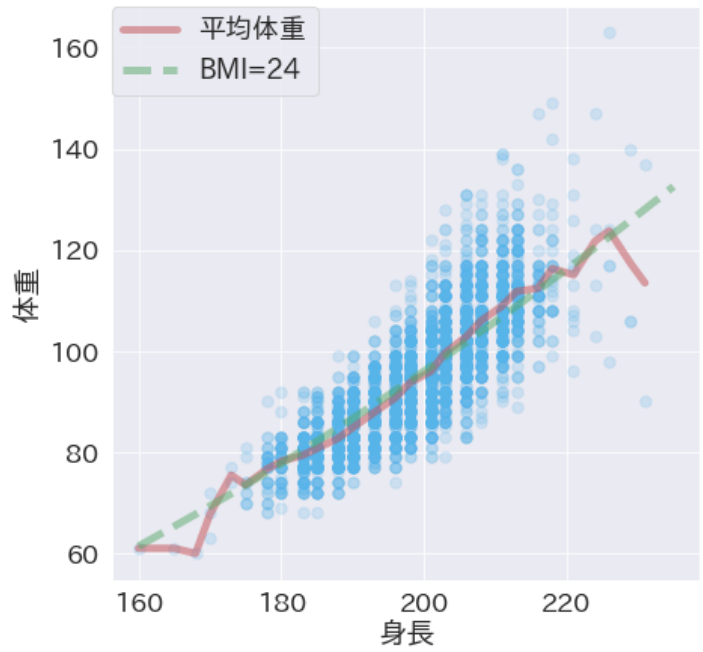
平均身長に経年変化があるか?





問

NBA 選手の BMI はどれくらい?



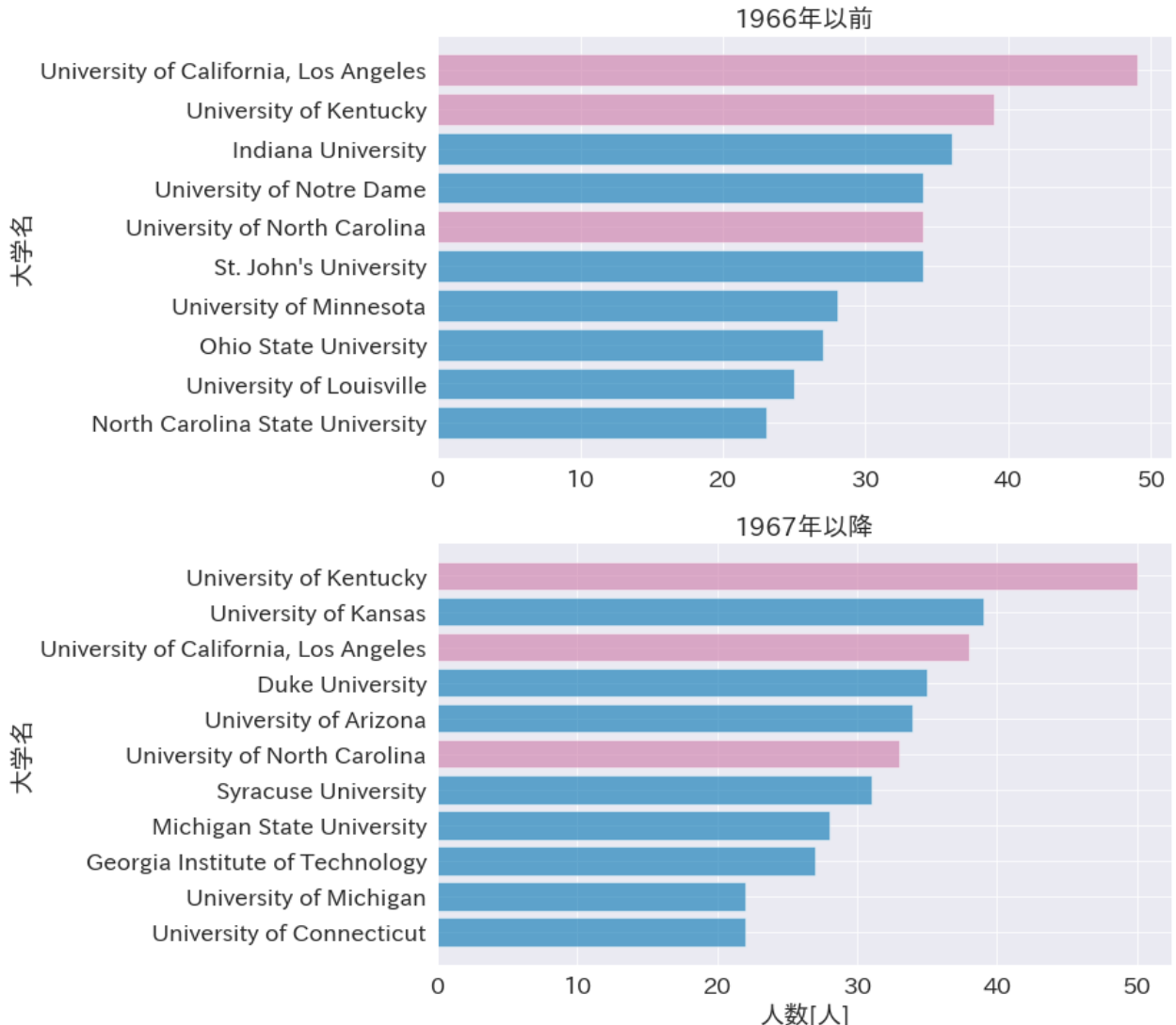
$$\text{体重 [kg]} = \text{BMI} \times (\text{身長 [m]})^2$$

(そもそも) なぜ 3 乗ではなく 2 乗?

NBA 選手のデータ: 分析例 2

問

出身大学はどこが多い? 時代の変化はある?



「数値」だけがデータではない. 「文字列」もデータになりうる.

Take Home Message

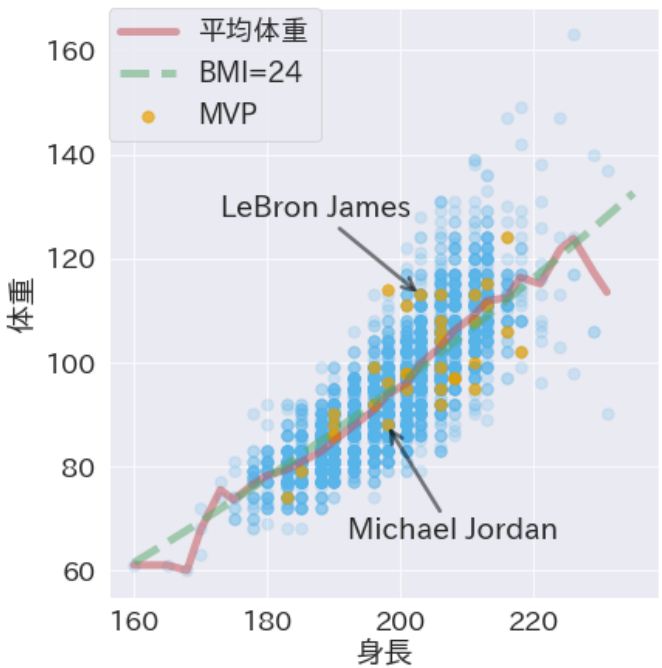
比較的単純なデータ (今回の場合 4 つの変量だけを使った) であっても, 非常に多くの見方が可能である

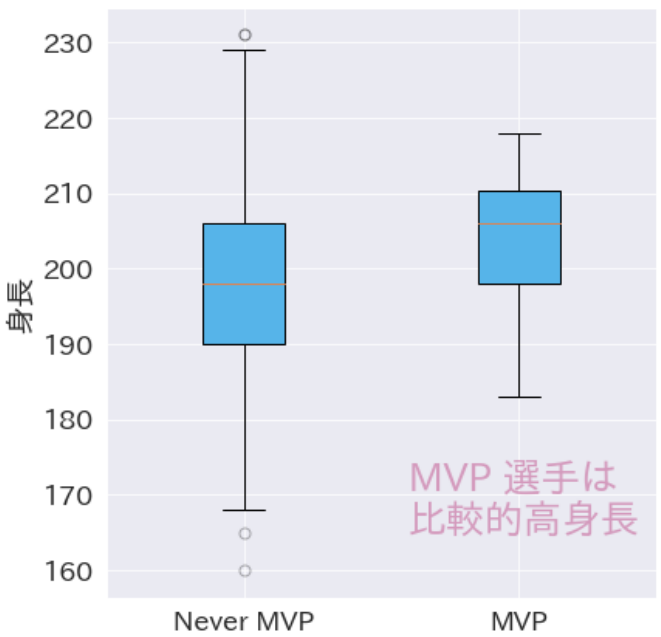
NBA 選手のデータ: 分析例 3

問

名選手の BMI は高い? 低い?

NBA 選手のデータだけで, この問に答えるのは難しい. そこで, 年間 MVP の情報を追加して考察しよう.





Take Home Message

別のソースから得たデータと組合わせることで, 新たな発見につながることがある.

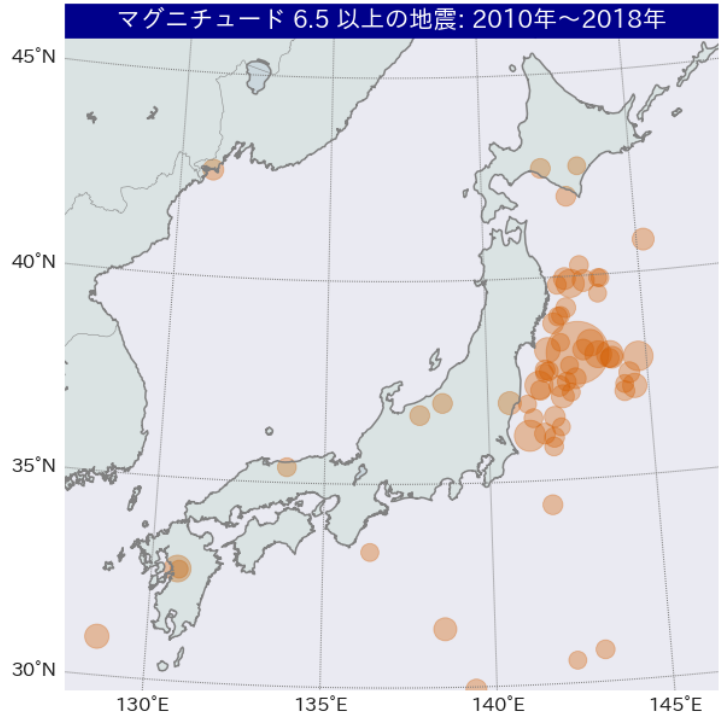
まとめ: 発見プロセスにおける仮説駆動とデータ駆動

仮説駆動

- データは比較的少なく簡単
- データにアクセスできるのは科学者や技術者など一部の人に限られる.
- 現在でも発見をするための重要な思考様式である.

データ駆動

- データは比較的複雑
- 公開されているデータ (オープンデータ) の場合, 誰でもアクセスできる. あた, 多くの組織でデータの蓄積が進んでいるが, データの整理や分析はそれほど進んでいない.



↑気象庁の HP のデータから作成. 1919 年から最近まで日本付近で発生した地震のデータ (発生日時, 震源の緯度・経度・深度, マグニチュード) を誰でも入手できる.

地震の予測方法を見つけたい!!

という子供が出てくるかも. わくわくしませんか?

まとめ: 探究心・創造性

探究心 現象に関心・好奇心を持ち, データ (数値) から洞察を得ようとすること

創造性 データから興味深い「問」を立てられること (そのためには, 文理融合・分野横断・異分野協力が鍵. 「文系だから関係無い」は通用しない)

Take Home Message

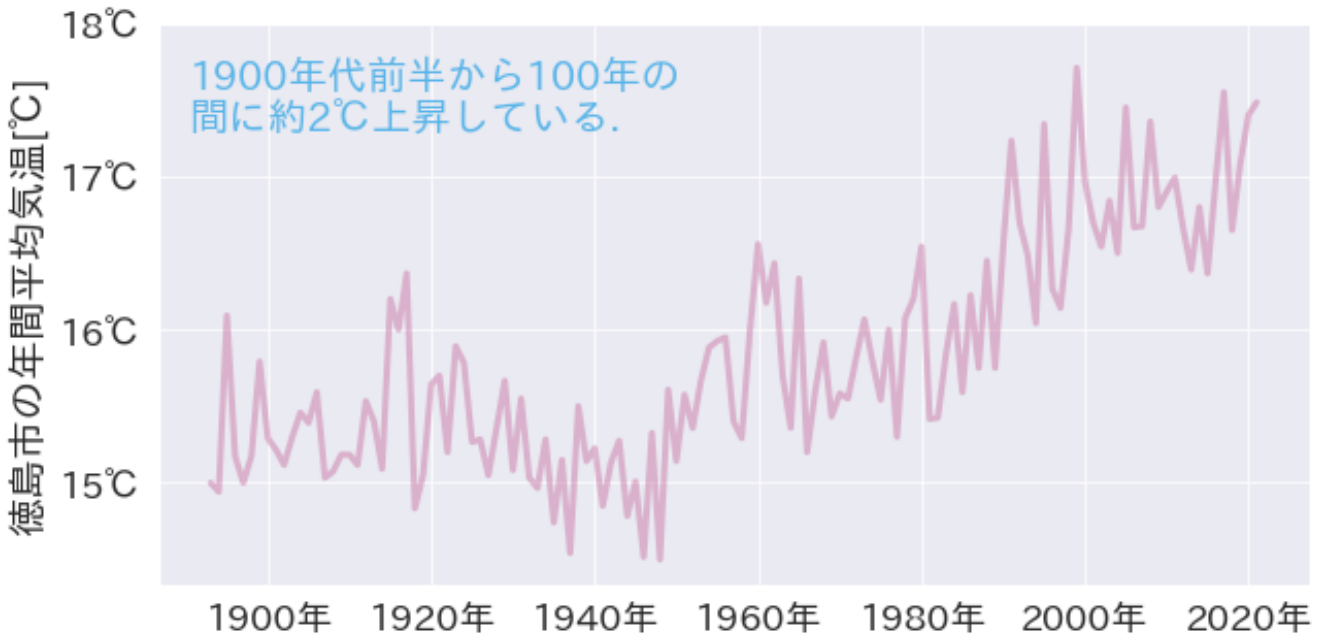
(1) 子供のこのような資質を延ばすことが, 学校教育に求められています. このような資質は専門的な研究分野だけでなく, 多くの業種で必要になってきています.

(2) データに関する探究心や創造性を支えるのは, 統計学やコンピュータに関する知識や技術です (第 2 章).

気温は上昇しているか?

問: 気温は上昇しているのか?

地球温暖化が問題になっているが, 本当に気温は上昇しているのだろうか? 上昇しているとするとどの程度上昇しているのか?



データソース: 気象庁

世間で言われているから, 「なんとなくそうなのかな」と思っていることが沢山ある.

そのようなことを実際に手を動かして調べてみることが大事になってきている.

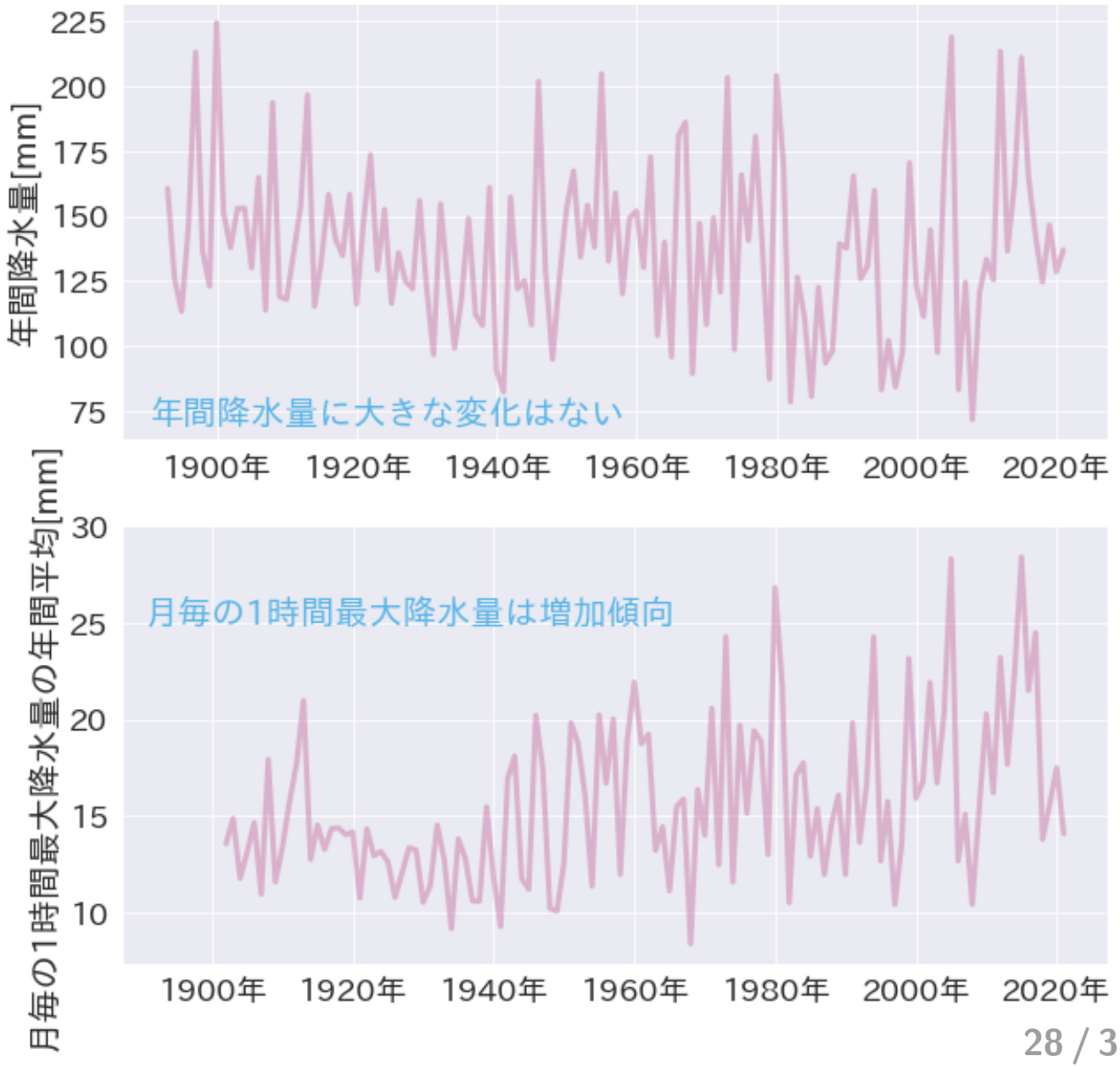
降水量は増加しているか?

問: 降水量は増加しているか?

近年, 水害が各地で生じているが, 降水量は増加しているのだろうか?

降水量自体には増加傾向は見えないが, 1 時間降水量の最大値は増加している (集中豪雨が増えている).

データソース: 気象庁



データに基づく教育

校務系データ，授業・学習系データに基づいた学習指導・生徒指導、学級・学校経営の質の向上に関するモデル事例（渋谷区の取り組み）

https://www.mext.go.jp/content/1387543_02.pdf より引用

Take Home Message

学校教育でもデータ重視の方向へ進んでいる．そのためこれからの学校教員はデータの基礎的扱いに習熟している必要がある（将来的には AI が導入される可能性も高いので，AI に関する基礎的素養も必要）．

教育は人間を対象とする営みなので，データで全てが解決されるわけではないけれど，経験だけに頼らず，データを活用しようとすることは重要です．

都会より地方の方が...1

問: 「都会より地方の方が高齢化が進んでいる」は本当か?

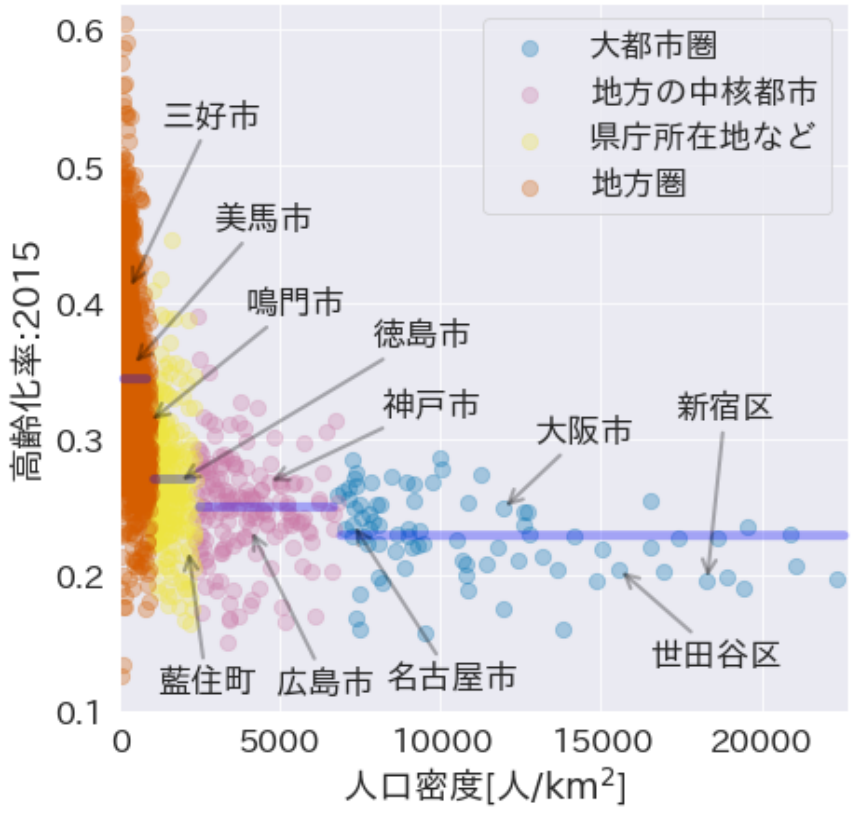
yes か no か, 予想してみよう.

$$\text{高齢化率} = \frac{\text{65 歳以上人口}}{\text{総人口}}$$

$$\text{人口密度} = \frac{\text{総人口}}{\text{可住地面積}}$$

人口密度を 4 つの階級に分け, 階級毎に高齢化率の平均値を計算した (青い線). 都市部ほど高齢化率が低い傾向が見られる.

また地方の中にも高齢化率が低い所もあるから, 地方はどこも高齢化が進んでいるというわけではない.



データソース: 教育用標準データセット

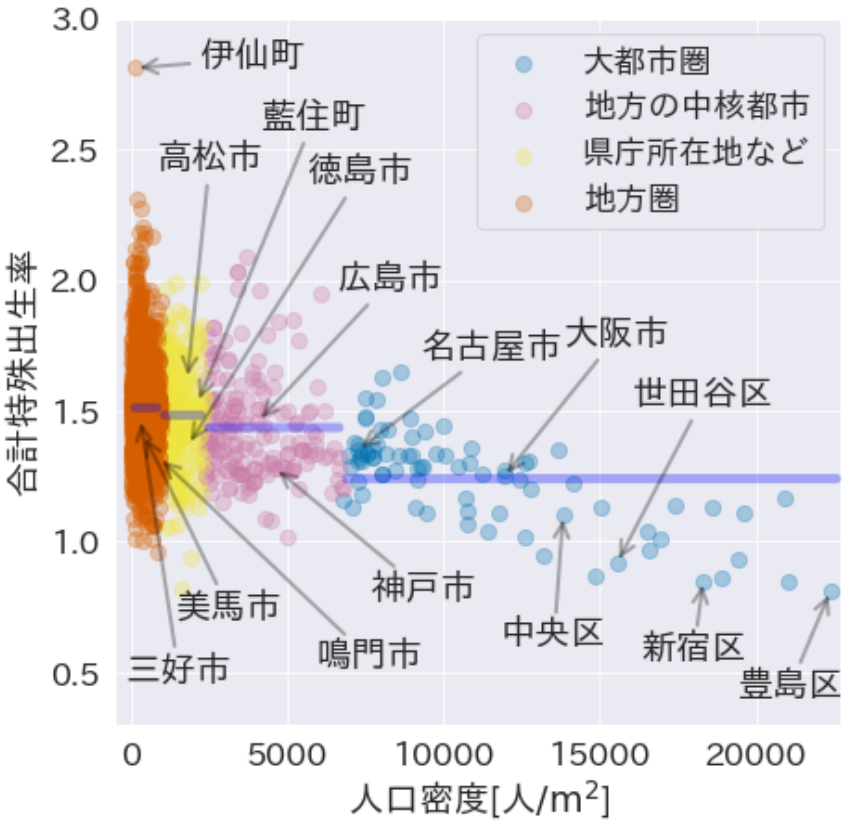
都会より地方の方が...2

問: 「都会より地方の方が高齢化が進んでいるのは, 地方は出生率が低いからである」は本当か?

yes か no か, 予想してみよう.

大都市圏が一番出生率が低い. それ以外の地域では大きな差異は見られないが, 地方ほど高い傾向が見られる.

地方の高齢化率が都会より高いのは, 出生率が低いからではなく, (特に若い世代の) 人口流出が主要な要因である.



データソース: RESAS

日本人は寛容か?

問: 「日本は多神教だから寛容だ」は本当か?

このような考えは昔から多くあったようだが, 本当に正しいのだろうか? **yes** か **no** か, 予想してみよう.

[以下の文章は「不寛容論 (森本あんり著)」の引用である]

『現代日本の宗教事情 (国内編 I)』では, 編者の堀江宗正が「世界価値調査」のデータを用いて日本と他国を比較し, その「惨憺たる」結果を示している. 指標に選ばれているのは中国, インド, アメリカ, ブラジル, パキスタンで, それぞれ無宗教, 多神教, 一神教など, 多様な宗教情勢を抱えた国々である.

日本は (中略)「他宗教の信者を信頼する」人の割合では中国に次いで下から 2 番目, 「他宗教の信者も道徳的」と考える人の割合が最低である.

「他宗教の信者と隣人になりたくない」と答える人は 6 つの国の中でいちばん多く、「移民や外国人労働者と隣人になりたくない」はインドに次いで多い。これらの数字は、宗教的にきわめて不寛容な日本の現実を浮かび上がらせている。

データに基づく思考や判断

データが身近になり入手しやすくなったことで，さまざまな主張をデータを用いて検証できるようになった．

何らかの主張をする際に，その主張が本当に正しいのかデータを分析して検証することが必要である．そのためには自分自身の考えや経験を批判的に見ることが重要である．

有名な本から引用しておこう：

Take Home Message

情報を批判的に見ることも大事だけれど，自分自身を批判的に見ることも大事 (『ファクトフルネス』より引用)

問

普段「なんとなくそうかな」と思っていることで，データ分析したいテーマ (問) を 3 つ挙げよ．

1 章全体のまとめ

学校教育で児童生徒の次のような資質を延ばすことが重要である：

- 探究心 (様々な現象に関心・好奇心を持つこと)
- 創造性 (様々な問を立てること)
- 批判的精神 (自分自身の考えを批判的に見ること)

また，これからの学校教員には，データに基く教育が求められる．

基礎情報教育: データ科学入門
第 2 章 データ科学とは?

T. MIYAGUCHI

Naruto Universality of Education

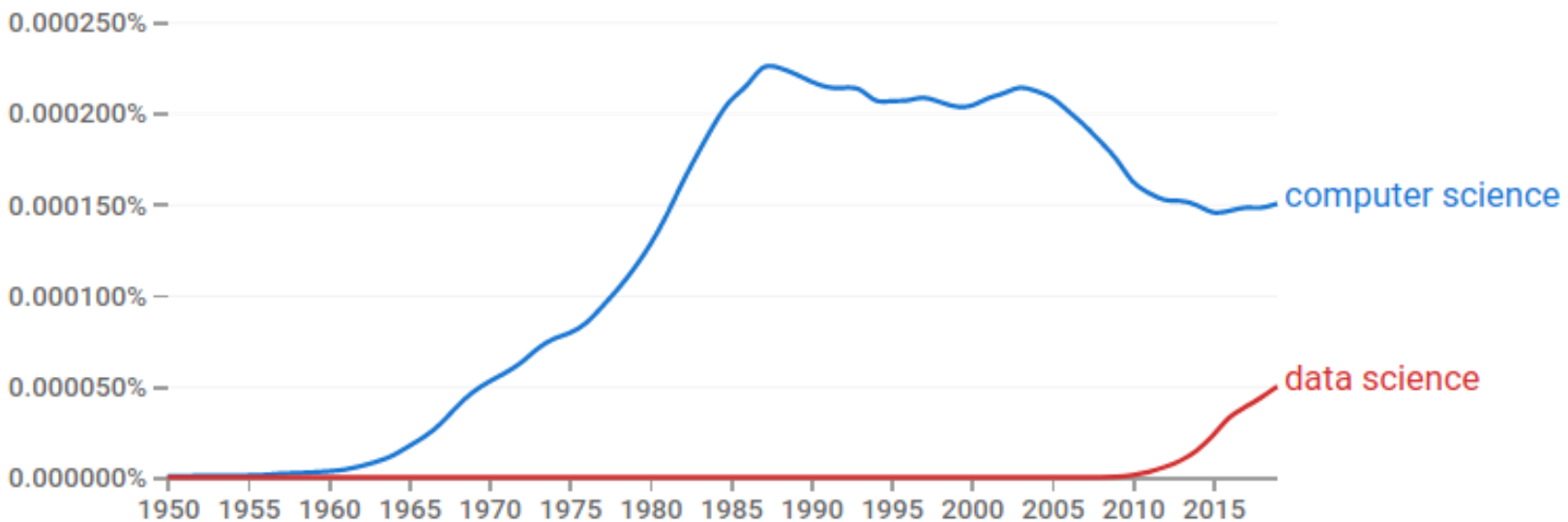
講義全体のアウトライン

- 第 1 章: 社会と教育における変化
  - 仮説駆動とデータ駆動
  - データに基づく思考や判断
- 第 2 章: データ科学とは?
  - 統計学・計算機科学とデータ科学
  - AI・機械学習・倫理
- 第 3 章: データ分析の基礎
  - データとは?
  - 代表値・散らばりの指標・関係性の指標
- 第 4 章: 可視化
  - 可視化の必要性
  - 量の表現・割合の表現・分布の表現・関係性の表現・系列の表現
- 第 5 章: データ分析実習
  - 問を立てよう・統計量と可視化
  - 機械学習に挑戦・データ分析の実践
- 第 6 章: 様々な話題

データ科学 (データサイエンス) とは

大量のデータを入手・分析することが可能になったことで，「データ科学 (データサイエンス)」 という分野が注目されている (下図).

しかし，一般的に認められた定義はなく，批判も多いようである．ここでは代表的な批判を 2 つ紹介し，それに対する個人的な見解を述べることで，「データ科学」とは何かを考えてみたい.



データ科学 (データサイエンス) への批判

批判

「データを扱わない科学」は存在しないので，「データ科学」は「科学」そのものでは?

回答: 「データ科学」というのは「自然科学」や「社会科学」などと同じような専門的な学術分野を表すものではない．実際，ビジネスやスポーツなどの分野の方が，データ科学という言葉がより頻繁に利用されているようだ．つまり，「データ科学」は「科学」より圧倒的に広いのだ．

しかし，「データ科学」が専門的な学術分野を表すものではないとしたら一体何なのだろうか?

データ科学 (データサイエンス) への批判 2

批判

データ分析 (data analysis) と同じでは?

回答: データ科学とデータ分析の違いは:

> **"(Data science) puts a human face on the data analysis process" (Blei and Smyth 2017, *Science and data science* から引用)**

Take Home Message

すなわち, 方法論 (統計学・計算機科学) を中心とする立場から, 好奇心や探究心, 創造性など人間的側面を中心とする立場への転換 (の決意表明) と捉えるべき. そういう意味では

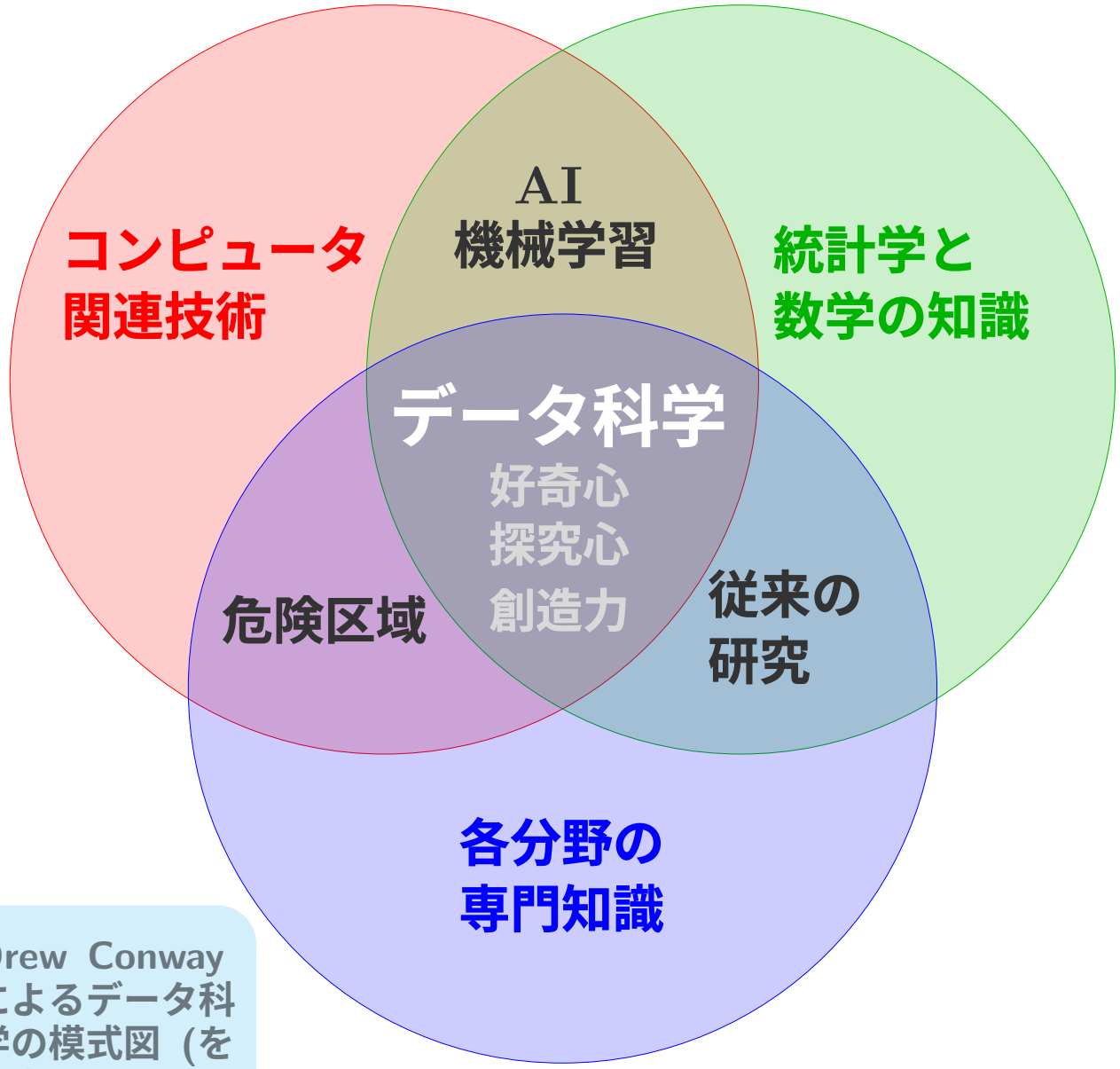
「科学的な好奇心を持ってデータに向き合うこと」

が「データ科学」と言えるかもしれない.

データ科学 (データサイエンス) とは?

データやその背後にある現象に対する好奇心や探究心を中心とするデータ科学は, コンピュータ関連技術 (プログラミングなど) と統計学や数学の知識に支えられている.

各分野の専門知識が必要であることから, 異分野間の協力が鍵を握っている.



Drew Conway によるデータ科学の模式図 (を改変したもの)

統計学とデータ科学の相違点

統計学 データから規則性や不規則性を見いだす数学的手法についての研究分野

統計学では統計量や統計的推測を通して

データから情報を抽出する方法

が興味の中心 (あくまで, そういう傾向があるということです).

データ科学 データ科学では

データやその背後にある現象

が関心の中心.

しかし, 統計学の基礎を理解せずに, データ分析の結果のみを見ていると, 思わぬ考え違いをすることもある. 例えば次の問題を考えてみよう.

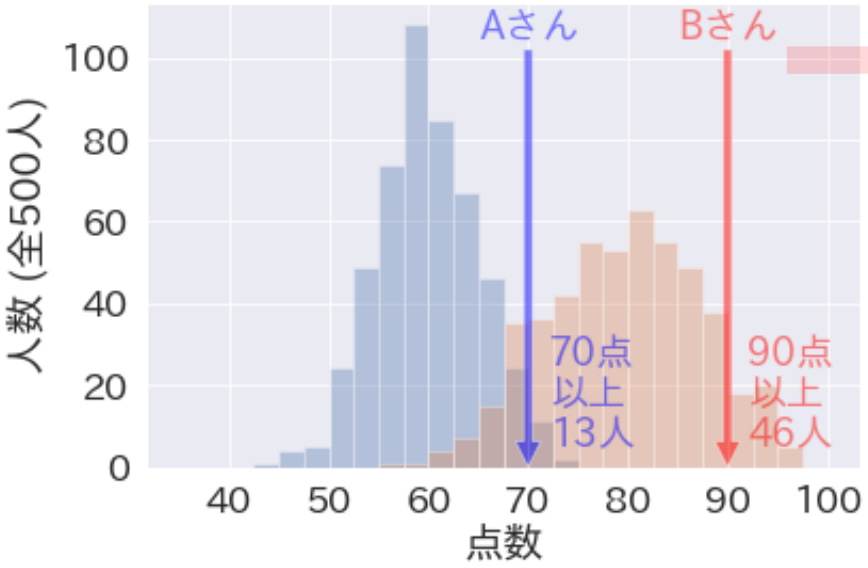
問: 成績が良いのはどっち?

A さん: 平均点が 60 点の試験で 70 点

B さん: 平均点が 80 点の試験で 90 点

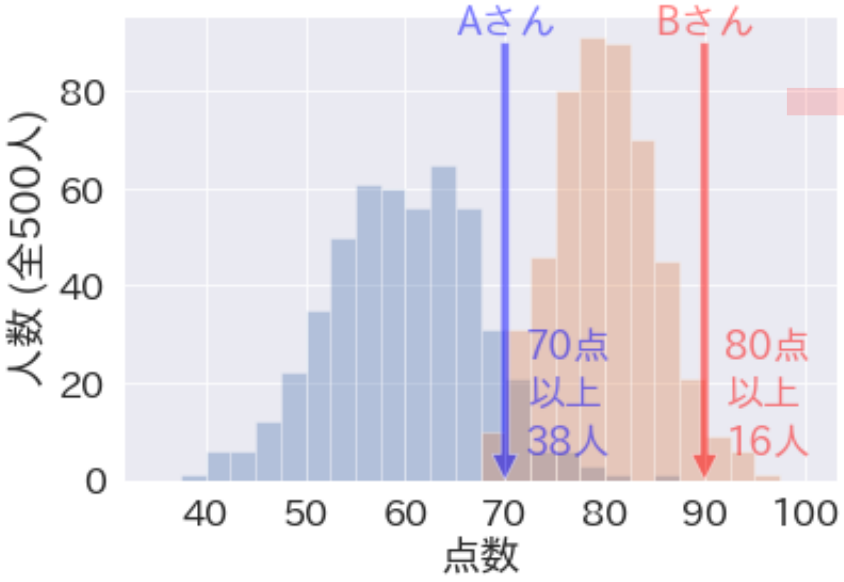
答: 何とも言えない (理由は次のページ). A さん もしくは B さんと答えてしまった人は, 成績評価を正確にできない可能性がある.

試験の点数の評価



A さんの方が成績が良い (と判断できそう).

B さんの方が成績が良い (と判断できそう).

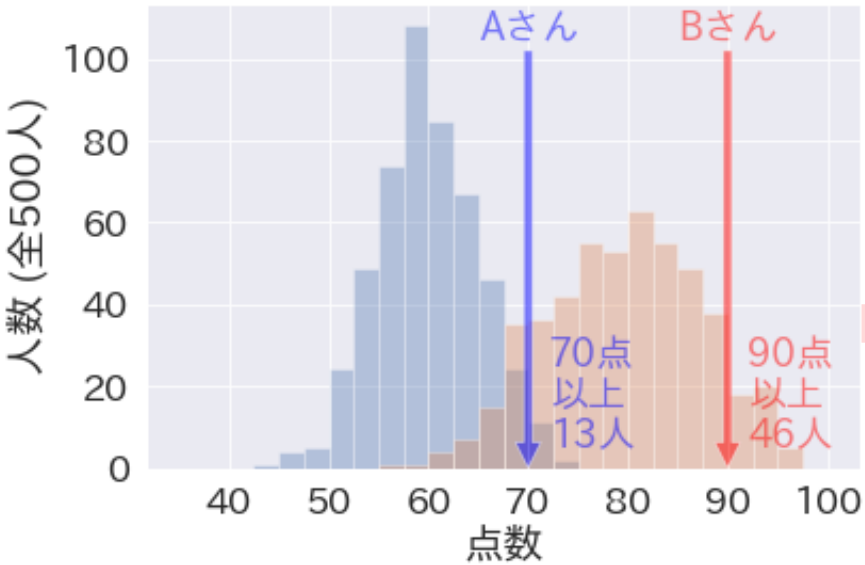


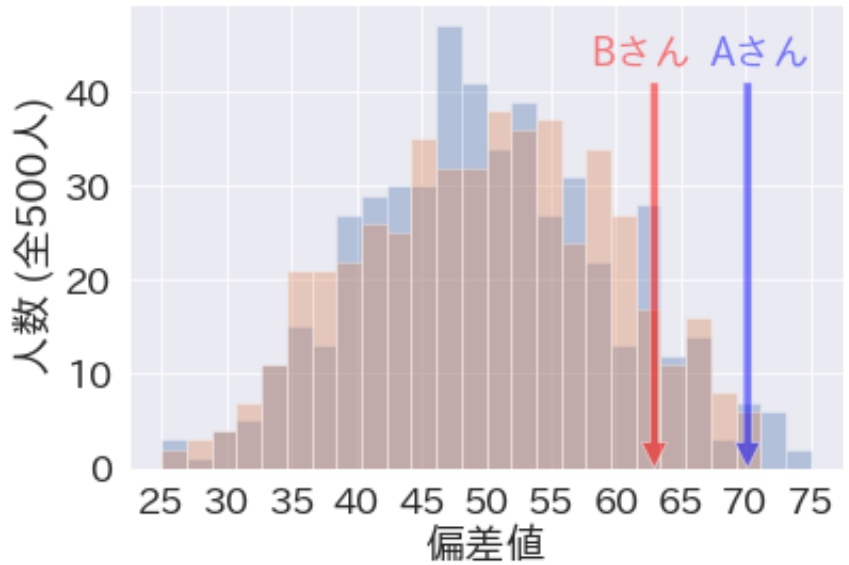
平均点の周りの点数のバラツキを考慮しないと判断できない.
→ バラツキを考慮するため, 偏差値を導入する.

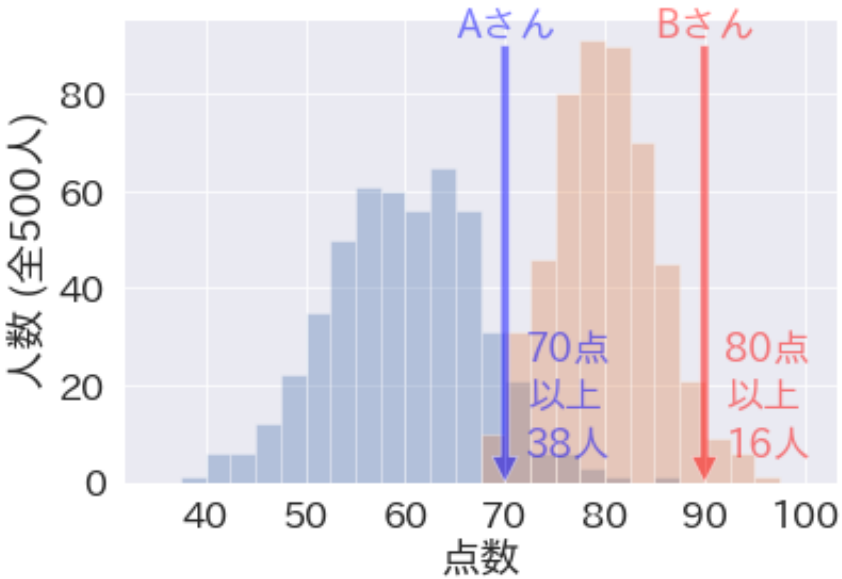
偏差値

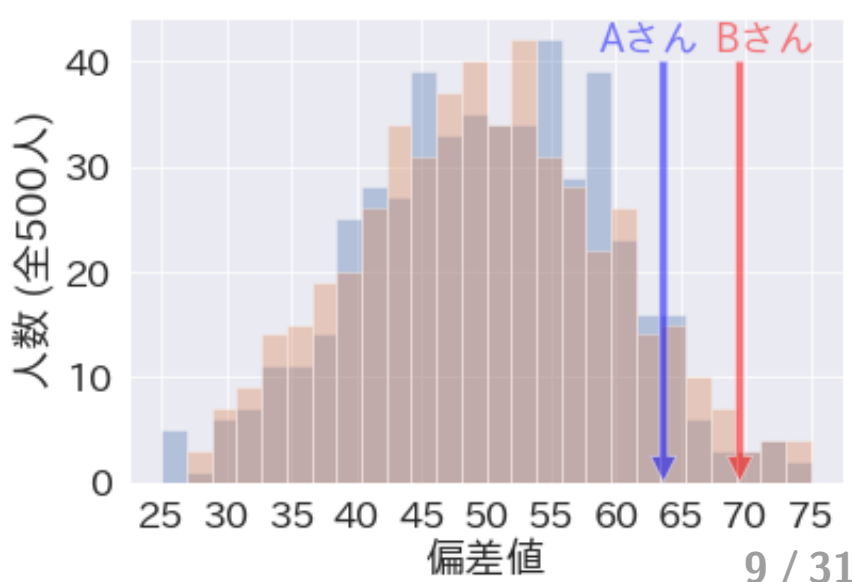
$$(\textbf{偏差値}) = \frac{(\textbf{点数}) - (\textbf{平均点})}{(\textbf{標準偏差})} \times 10 + 50$$

偏差値


Aさん
Bさん
70点
以上
13人
90点
以上
46人
人数 (全500人)
点数


Bさん
Aさん
人数 (全500人)
偏差値


Aさん
Bさん
70点
以上
38人
80点
以上
16人
人数 (全500人)
点数


Aさん
Bさん
人数 (全500人)
偏差値

危険区域

偏差値は順位付けの便利な手法だが，次の点に注意する必要がある:

(1) 同じ統計分布に従っているかどうか (どの試験の点数分布も正規分布に近いか)

(2) 同じ (もしくはほぼ同等とみなせる) 母集団から無作為抽出しているかどうか

これらが満たされていない場合，得られた順位は信頼できない可能性がある.

Take Home Message

データ分析の結果を元に適切に判断・意思決定するには，統計学の知識が必要である (統計学の知識なしに手法だけを用いるのは危険なことがある).

ランキング 1

順位付け (ranking) は判断や意思決定の重要な例であり，数学や統計が応用されている．例えば，

- Google 検索した際に表示される順番
- Amazon で商品を薦める順序
- FIFA のサッカーのランキング
- TOEIC などで利用されている項目反応理論

などは，数学的な手法でランキングを計算している．これらは，

「真の順位」をデータからうまく推測する方法

と言うことができる．

Take Home Message

でも「真の順位」なんて本当にあるのだろうか?

あったとしても，本当に正確に推測できているんだろうか?

という意識は重要．便利だけれど，鵜呑みしてはいけません．

ランキング 2

FIFA のサッカーのランキングを考えよう.

サッカーの国際試合を総当たりで実施すればランキングは容易につけられるが, そうすると各チームは 200 試合以上試合しなくてはならない. かといって, 勝率で決めるのも問題がある (弱いチームとばかり対戦すれば勝率が上がる).

したがって, 総当たりや勝率以外の方法でランキングを付けるうまい方法が必要だ (対戦してもいない国より格下だと言われて納得できるか?).

問

FIFA ランキングがどのような方法で行われているか調べてみよ.

計算機科学とデータ科学の相違点

計算機科学 情報と計算の理論，およびそのコンピュータ上への実装と応用に関する研究分野.

計算機科学ではコンピュータ上で計算を実現する

アルゴリズムやその性能

などの方法論が興味の中心 (あくまで，そういう傾向があるということです).

データ科学 一方，データ科学では

データやその背後にある現象

が関心の中心.

しかし，計算機科学が提供するアルゴリズムやソフトウェア，プログラミング言語などの技術の習得は極めて重要である．特に，

Take Home Message

コンピュータ関連の技術の習得が，データ分析をしようとする意欲を高める.

その結果，データやその背後にある現象に対する興味や関心を高めることがある.

可視化

2020 年のセンター試験に出題された箱ひげ図

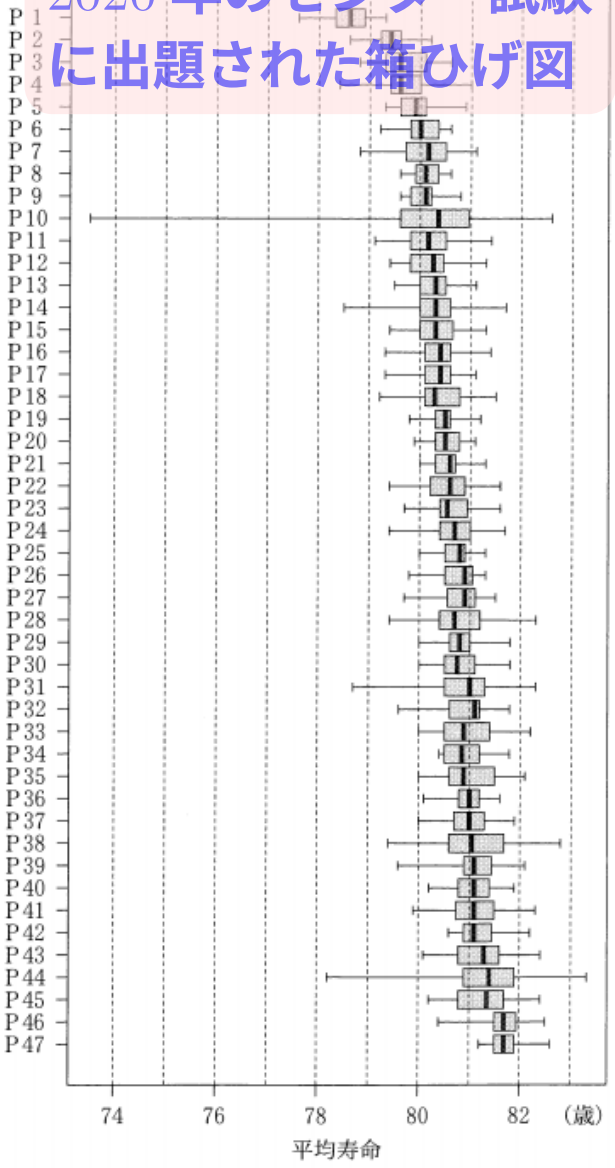


図 1　男の市区町村別平均寿命の箱ひげ図
(出典：厚生労働省の Web ページにより作成)

左の図について

- 縦軸は都道府県名 (記号で置き換えられているため具体的な都道府県名は分からない)
- 横軸は市区町村ごとの「男性の平均寿命」

箱ひげ図について

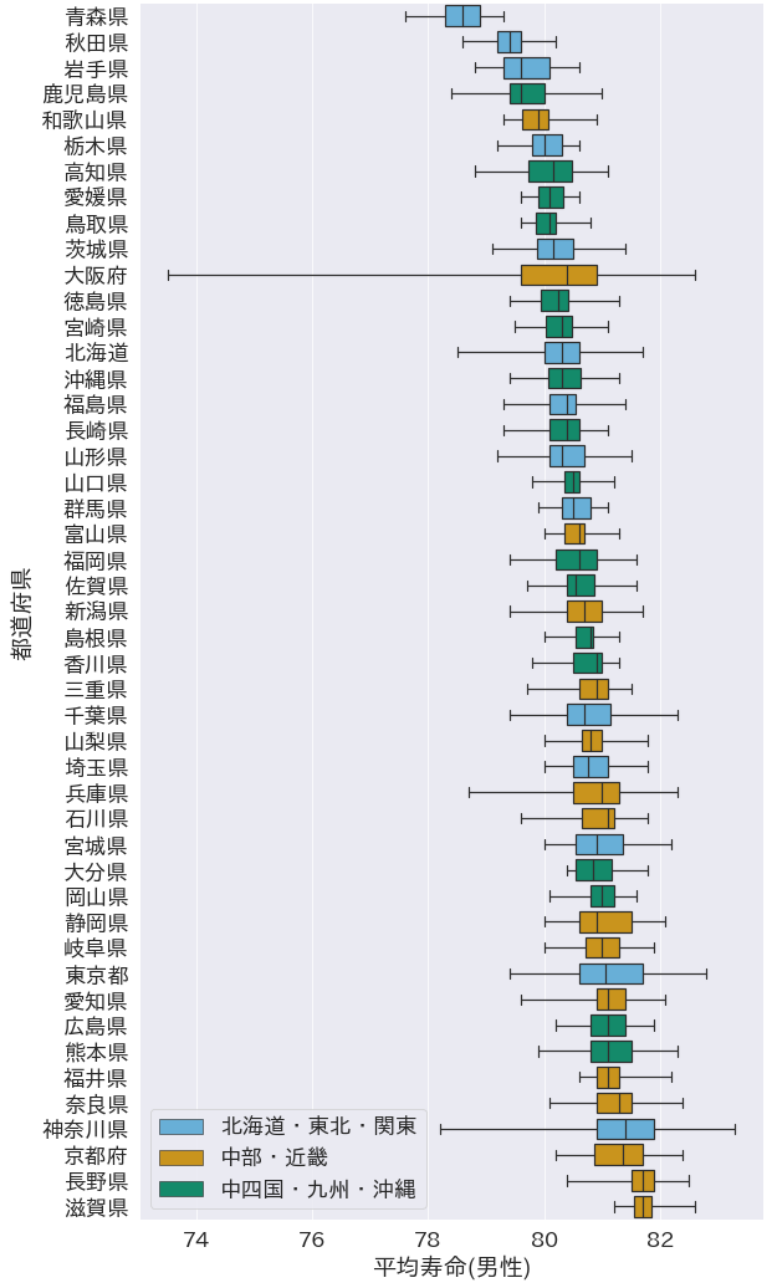
- 箱の中の黒い縦線の値は中央値
- 箱の左右の端の値は (第 1 および第 3) 四分位数
- 箱から延びている線の両端が最小値と最大値

問題

下のページから「平均寿命のデータ」をダウンロードし，センター試験と同様の箱ひげ図を作成せよ (どのような方法でも可).

http://mygch.g2.xrea.com/DS/ds.html

Python による可視化



←プログラミング言語 Python を用いて，センター試験と同じ図を出力した (Python は高校で必修化される授業「情報 I」で主として用いられる言語となることが予想される).

ちなみに，中学校指導要領 (H29.7) には

コンピュータなどの情報手段を用いるなどしてデータを整理し箱ひげ図で表すこと.

とある (統計量の意味を理解することの重要性はよく強調されるが，このような可視化技術は軽視される傾向がある).

データの前処理 1

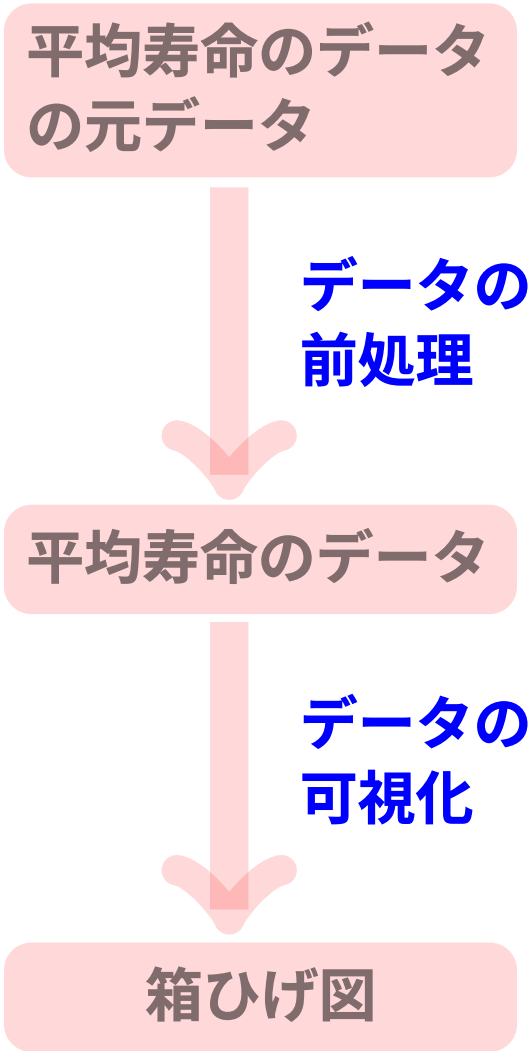
平均寿命の元にしたデータは e-stat (政府統計の総合窓口) からダウンロードした:

同じファイルを HP にも置いてある:

http://mygch.g2.xrea.com/DS/ds.html

「平均寿命のデータの元データ」

p これを元にして,「平均寿命のデータ」を作成した. このような処理をデータの**前処理**という.



データの前処理 2

問題

実際に「平均寿命のデータの元データ」ダウンロードして，どのようにして前処理を行うかを考えよ．

(次のスライドで説明するように) Python による前処理プログラムの行数は約 25 行であり，比較的複雑な処理が必要となる．

一方，「平均寿命のデータ」 があれば，それをグラフにするには 10 行以下で，処理も比較的単純である．

Take Home Message

データ分析の作業の中で「データの前処理」にはかなり時間がかかる．プログラミングの技術が不可欠な作業である．

Python のプログラム作成には，次のサイトを参考にした．

https://ameblo.jp/ken-pc-works/entry-12568760871.html

データの前処理 3

前処理のプログラム

```
import pandas as pd
import csv

csvfile = open('center-exam2020.csv')
cols=['p_code',' 都道府県','c_code','c_code2',' 市区町村',' 平均寿命（男性）',' 平均寿命（女性）']
df = pd.DataFrame(index=[], columns=cols)
cnt = 0
for row in csv.reader(csvfile):

    cnt = cnt + 1
    if (row[0] == ' 平成 27 年'): cnt = 1 # 各自治体の先頭行
    elif (cnt == 2): # 市町村コードと自治体名を取る
        code = row[0]
        pre  = row[1].split("  ")[0]
        city = row[1].split("  ")[1:]
    elif (cnt == 6): m_ls = row[6]  # 男性の平均余命 (male life span)
    elif (cnt == 28): f_ls = row[6] # 女性の平均余命 (female life span)
    elif (cnt == 49 and m_ls != '…'): # データフレームへ追加
        lst = [code[1:3], pre, code[3:], code[4:], ''.join(city), float(m_ls), float(f_ls)]
        rec = pd.Series(lst, index=cols)
        df = df.append(rec, ignore_index=True)

df = df[df.c_code2 != '00'] # 全国平均・県平均 (000) と区がある市 (.00) を削除
df.drop('c_code2', axis=1).to_csv('center-exam2020_mod.csv', index=False)
```

可視化のプログラム

```
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(9, 18))
grouped = df.groupby(' 都道府県')[' 都道府県', ' 平均寿命（男性）'] # 県名でグループ化
ordered = grouped.mean().sort_values(' 平均寿命（男性）') # 男性の平均でソート
kwargs = {'order':ordered.index,'whis':(0, 100),'showmeans':False}
sns.boxplot(y=' 都道府県', x=' 平均寿命（男性）', data=df, **kwargs)
```

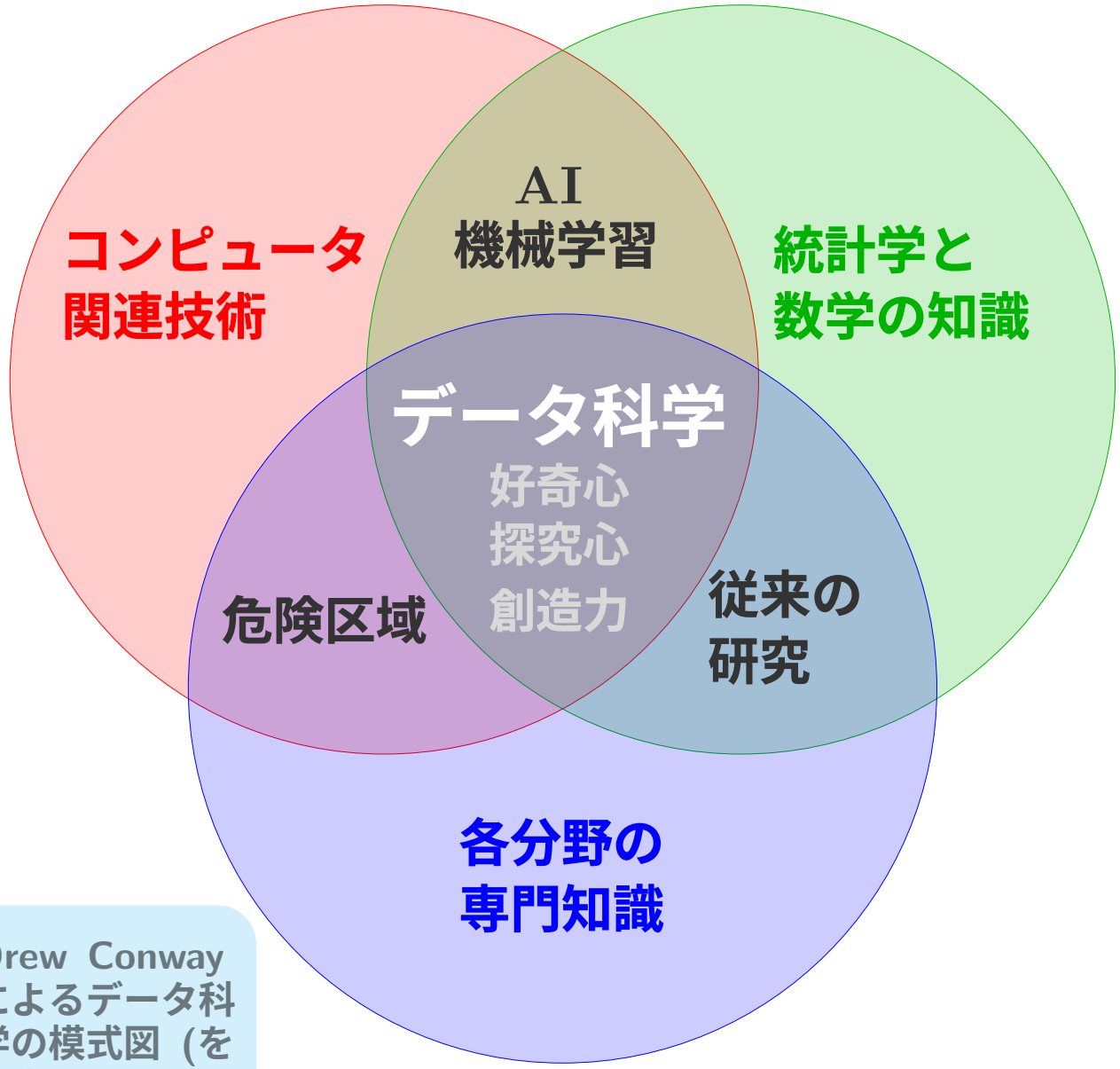
可視化のプログラムより，前処理のプログラムの方が複雑で長い．

一般にデータの前処理は非常に手間がかかる作業であり，プログラミングの技術が不可欠であることが多い．

データ科学 (データサイエンス) とは?

データやその背後にある現象に対する好奇心や探究心を中心とするデータ科学は, コンピュータ関連技術 (プログラミングなど) と統計学や数学の知識に支えられている.

各分野の専門知識が必要であることから, 異分野間の協力が鍵を握っている.



Drew Conway によるデータ科学の模式図 (を改変したもの)

データ駆動型○○とは 1

様々なデータ駆動型○○

- データ駆動型社会
- データ駆動型ビジネス
- データ駆動型農業
- データ駆動型開発
- データ駆動型研究

これまで，「長年の経験」による意思決定や判断が，様々な分野でなされてきた．

これに対して，データ (とその分析結果) に基づいた意思決定や判断をすることを差して，「データ駆動型○○」と呼ぶ．

問

どのような「データ駆動型○○」があるか調べてみよ．そのうちのひとつに注目して，どのような活動が行われているか調べよ．

データ駆動型○○を実現するためには，人工知能 (Artificial Intelligence) の利用が多くの場合不可欠である．

データ駆動型○○とは 2

エビデンスに基づく政策形成

2016 年にオバマ政権下で新しい法律 「エビデンスに基づく政策のための評議会設置法」が制定されました. この評議会の使命のうち主要なものは,

- **政策効果の因果関係がデータ分析により解明される仕組みを作る**
- **政府が持つ詳細な行政データを研究者に利用させ分析させる体制を整える.**

「データ分析の力 因果関係に迫る思考法 (伊藤公一郎著)」より (一部改変した上で) 引用

人工知能 (Artificial Intelligence)

人工知能とは

人工的につくられた人間のような知能、ないしはそれをつくる技術 (松尾豊,「人工知能は人間を超えるか」より引用)

AI には過去 3 回のブームがあったとされる (歴史の詳細は略します).

現在は第三次 AI ブームと呼ばれているが, その中心にある技術が機械学習である. 機械学習の中には, (テレビなどのメディアでも注目されるようになった) 深層学習 (Deep Learning) も含まれる.

AI の利用例: 需要予測

小売業における商品発注業務

問題点

- 発注量の不足は、顧客離れの大きな要因に
- 発注量の過剰は、売れなかった際に多くの在庫が発生

過去のデータを参照し需要分析を行っても、天候の急変やイベントなど外的要因が複雑に絡み合い、ベテラン担当者の知識と経験に頼る部分が多い。また, 人材不足のため、ベテラン担当者の知見を若手人材へ継承していくことも難しい。

そこで, イトーヨーカドーは AI を活用して需要を予測し、発注業務の精度向上と時間短縮を目指す試みが進められた.

この結果, 発注業務にかかる時間が平均 35% 短縮し, 欠品率は 27% 減少したそうである.

詳細はリンクを参照

機械学習

機械学習とは, データを理解するために数理モデルを構築すること.

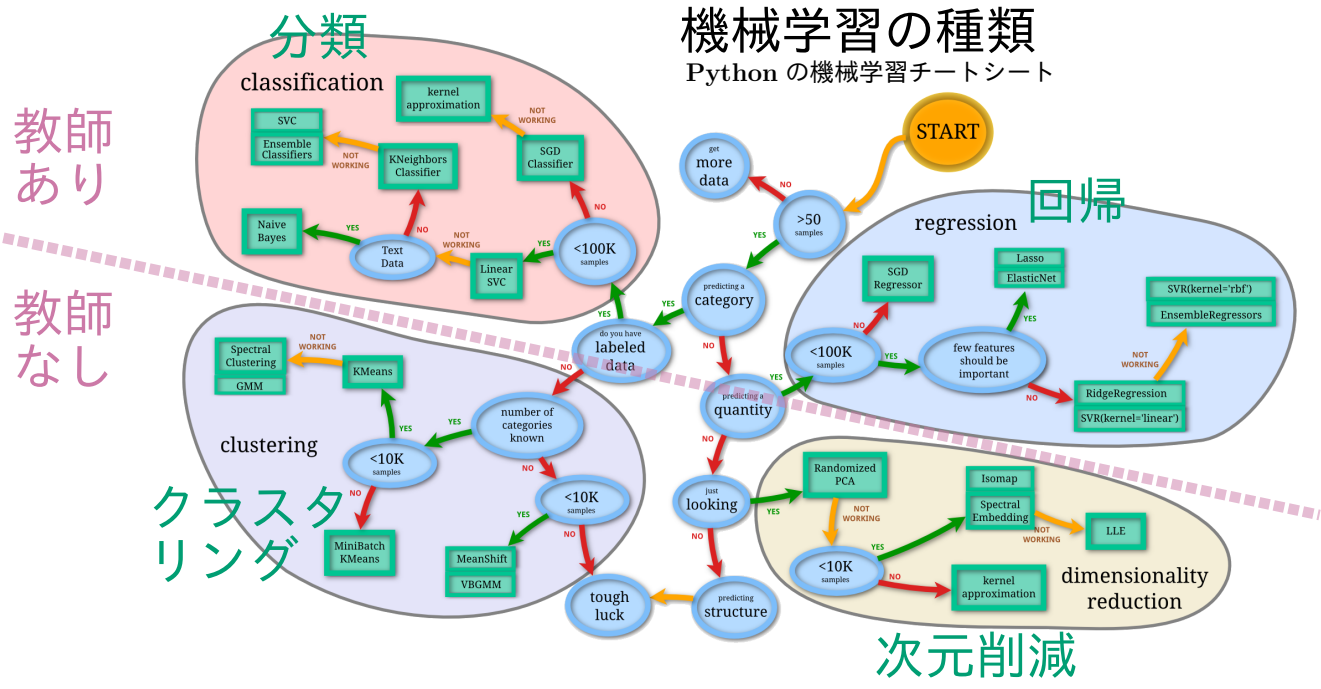
ここで「学習」とは, 数理モデルのパラメータを観測データに適応するために調節することを差す. この調整をコンピュータ (すなわち「機械」) が行うので, 「機械が学習する」と呼ばれている (VanderPlas, *Python Data Science Handbook* **より引用; 多少意訳している**).

例: 線形回帰の場合, 数理モデルは一次関数

$$y = ax + b$$

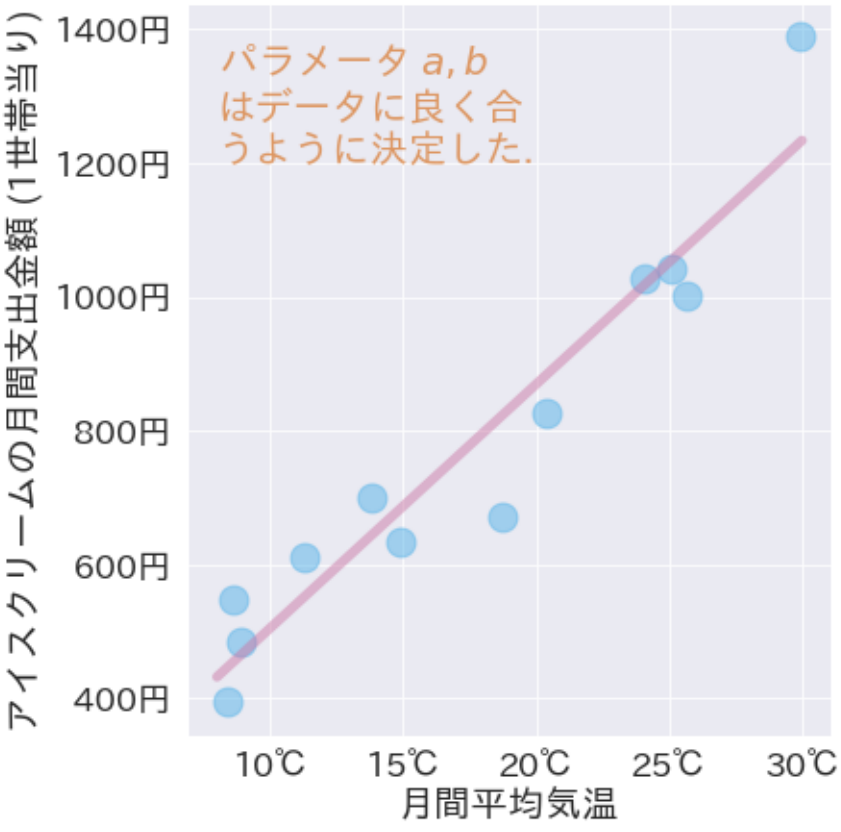
である. パラメータは傾き a と切片 b であり, データに基いて決定される.

機械学習は様々な手法に対する総称である (右図に参照).

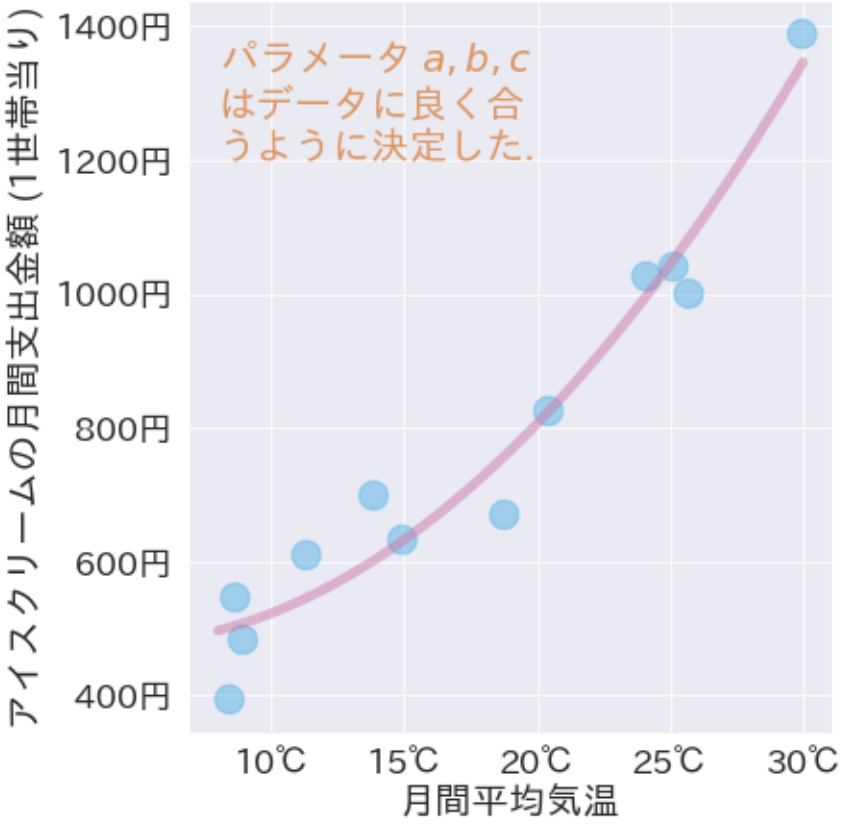


回帰とは

データを一次関数 $y = ax + b$ でモデル化



データを二次関数 $y = ax^2 + bx + c$ でモデル化



アイスクリームのデータは E-STAT, 月間平均気温は気象庁からダウンロードした. どちらも 2020 年の徳島市についてのデータ.

分類とクラスタリング

「分類」と「クラスタリング」は良く似たデータ分析手法である. どちらも, データを複数のグループに分ける手法である. 例として

例① 動物の写真から種ごとにグループ分け

例② 顧客のデータから購買特性 (どのような品を購入しそうかなど) ごとにグループ分け

などがある.

「分類」と「クラスタリング」の本質的な違いは, 「分類」が「教師あり」であるのに対し, 「クラスタリング」は「教師なし」である点である.

問

「分類」と「クラスタリング」の具体例を調べてみよう.

「分類」の簡単な例を実習の時間に扱う予定である.

データ科学の倫理的課題

Take Home Message

データ科学は直面している倫理的な課題は，次の 2 つの間のバランスをいかに取るかということである:

- **社会の安全や利益**
- **個人やマイノリティーの自由およびプライバシー**

(Kellerher and Tierne 2018, Data Science)

この節では，具体的な例を通して，データ科学や AI の発展により，どのような問題が発生しているのかを考えてみる．具体例は次の書籍を参考にした:

『おそろしいビッグデータ 超類型化 AI 社会のリスク』(山本龍彦著)

妊娠予測とベビー用品広告

シナリオ 1

小売業の A 社は大量の顧客データを解析して，「無香料スキンローション，特定のサプリメント，大きめのバッグなどの商品を同時期に購入した，ある年齢層の女性は妊娠している可能性が高い」というパターンを発見した．

A 社は巨大な顧客データベースからこのようなパターンに一致する女性を見つけ，彼女たにに対してのみベビー用品のクーポン券を送った．

問 (5 分)

シナリオ 1 について，あなたの意見を述べよ．

A 社にとって適切な広告をそれを必要とする人に届けることができる．妊娠した女性にとっても，必要な品に関する広告が届くので効率的である．これは「社会の利益」と言っても良いかもしれない．

しかし，「妊娠」は非常にプライベートな出来事である．もしかすると家族や親しい友人にも打ち明けていない段階で，A 社はその事実を「知っていた」かもしれない．家に届いたクーポンを，その事実を知らされていない家族が見るかもしれない．

信用力スコアと社会的排除 1

シナリオ 2

B 社は求職者の職務遂行能力を予測する AI を開発した．C さんは大学卒業後，非正規雇用を数年経験し，その頃クレジットカードの支払いを滞納したことがあった．

その後 C さんは正規雇用の職を求めて就職活動を始めた．しかし B 社の AI は非正規雇用期間や滞納歴があると職務遂行能力が低いと予測するようで，C さんは B 社の AI を導入している全ての企業で不採用となった．仕方なく C さんは低賃金の非正規雇用を転々としたが，そしてそのことがさらに AI の評価を下げた．

A さんは B 社の AI を導入している全ての組織から排除され続け，次第に自らが劣った存在であると感じるようになった．

①中国のある企業は個人の信用力を査定するサービスを始めており，融資や住宅の賃貸，ビザ取得や裁判などで利用されているという．② 日本ではソフトバンクが新卒採用のエントリーシート評価に AI を利用しているそうだ．(山本 龍彦，「前掲書」) ③アメリカのある都市では学校教師の評価に AI が利用され，評価が低い判断され解雇になった人もあるという (Cathy O'Neil, *Weapons of Math Destruction*)．

問 (5 分)

シナリオ 2 について，あなたの意見を述べよ．

信用力スコアと社会的排除 2

信用力スコアを用いると，評価を効率化することができる．また，良い人材を確保したり，融資や住宅の賃貸においては滞納する可能性が低い人を見つけることができる．

しかし，問題点も多い．

(1) AI による評価は C さん個人の評価ではない．C さんと同じような履歴 (非正規雇用歴や滞納歴) を持っている人達に対する統計的評価である．そうした評価は多くの場合正確かもしれないが，不正確なケースも当然ありうる．そのような不正確性を含む評価を就職や裁判など，人生を左右する局面で使用するのは危険である．

(2) 悪循環 (ネガティブ・フィードバック) が発生することも多い．C さんのケースでは，就職できなかったことで非正規雇用を続けるしかなく，経済的に困窮した (再びクレジットカードの支払いを滞納したかもしれない)．そのことで，さらに AI の評価を下げるという悪循環に陥いった．このような悪循環により AI は既に存在する差別を助長することがある．

個人の政治的信条の予測と選択的ニュース配信

シナリオ 3

SNS を提供する D 社は利用者の政治的信条を予測する AI を開発し，利用者の信条に合致したニュースや投稿を「あなたへのおすすめ」として配信していた．

保守的な考え方を持つ E さんには E さんが好むであろう保守派の言論が配信され，リベラルな政治的言論は排除された．そのため E さんがリベラルな言論に触れる機械は著しく減少し，保守派傾向はますます強まった．一方，リベラルな考え方を持つ F さんにも同様のことが起き，F さんのリベラルな傾向が強まった．

友人であった E さんと F さんは今では，相手が自分とは全く相容れない「他者」であると感じている．

問 (5 分)

シナリオ 3 について，あなたの意見を述べよ．

(1)「他者」の意見に触れる機会の減少は「他者」への敵意や恐れを生み，社会の政治的分断を助長するという指摘もある．

(2) 選挙をコントロールする危険性．フェイスブックが 2010 年 11 月に行った実験で，一部の利用者に投票を促進する情報 (投票所の場所や友人の投票行動) を配信することで，実際に投票率を上昇させることに成功したという (0.39%)．

基礎情報教育: データ科学入門
第 3 章 データ分析の基礎

T. MIYAGUCHI

Naruto Universality of Education

講義全体のアウトライン

- 第 1 章: 社会と教育における変化
  - 仮説駆動とデータ駆動
  - データに基づく思考や判断
- 第 2 章: データ科学とは?
  - 統計学・計算機科学とデータ科学
  - AI・機械学習・倫理
- 第 3 章: データ分析の基礎
  - データとは?
  - 代表値・散らばりの指標・関係性の指標
- 第 4 章: 可視化
  - 可視化の必要性
  - 量の表現・割合の表現・分布の表現・関係性の表現・系列の表現
- 第 5 章: データ分析実習
  - 問を立てよう・統計量と可視化
  - 機械学習に挑戦・データ分析の実践
- 第 6 章: 様々な話題

データの種類

量的データと質的データ

量的データ 身長や体重など，数値で表されるデータ

質的データ 性別や職業など，調査対象の性質を表すデータ

質的データは，数値化して上で扱うことが多い．例えば

$$\begin{aligned} \text{教員} &\Longrightarrow 0 \\ \text{会社員} &\Longrightarrow 1 \\ \text{自営業} &\Longrightarrow 2 \end{aligned}$$

ただし，質的データの数値は量的データと同じように扱うことに意味がないことが多い．例えば，

$$\text{会社員} - \text{教員} = \text{自営業} - \text{会社員}$$

となるが，この等式が成立することに意味はない．

データセットの種類

一変量データ

生徒	数学の点数
生徒 A	80
生徒 B	90
生徒 C	70
生徒 D	80
生徒 E	60
生徒 F	100

多変量データ

各個体の特徴 (変量)

生徒	数学の点数	英語の点数
生徒 A	80	80
生徒 B	90	90
生徒 C	70	70
生徒 D	80	80
生徒 E	60	60
生徒 F	100	100

各個体の名前 (index)

Take Home Message

ほとんどのデータは表形式 (relational database と呼ばれる) で表すことができる (数学の集合論を用いて証明されている). 元となるデータから前処理を経て, まずこのような表形式のデータにまとめることがデータ分析のスタート地点.

一次データと二次データ・メタデータ

一次データと二次データ

何らかの目的のために新たに収集されたデータを**一次データ** と呼びます．一方，誰かが他の目的にすでに収集してあったデータを再利用する場合**二次データ**と呼びます．

メタデータ

001	89
002	76
003	52
004	96
005	79

メタデータ

確率・統計学 期末試験
2021 年 8 月 2 日実施

学籍番号	点数
001	89
002	76
003	52
004	96
005	79

もし二次利用されることを想定してデータを公開するなら，様々なメタデータを付加することが必要です(個人で扱う場合でも，後から見返して分かるように，最低限のメタデータは付加するべきでしょう)．

どのようなデータが活用されているか 1

シャアリングエコノミー

「シェアリング・エコノミー」とは，典型的には個人が保有する遊休資産 (スキルのような無形のものも含む) の貸出しを仲介するサービスであり，貸主は遊休資産の活用による収入、借主は所有することなく利用ができるというメリットがある (総務省のサイトより抜粋)。

Uber や Airbb が有名です．これらの産業が最近発展しているのは，データを AI によって，「貸主」と「借主」のマッチングが容易できるようになったことが挙げられます．

外食産業

回転寿司店をチェーン展開するスシ●ーは，全ての寿司皿に IC タグを取り付け，レーンに流れる寿司の鮮度や売上状況を管理している．

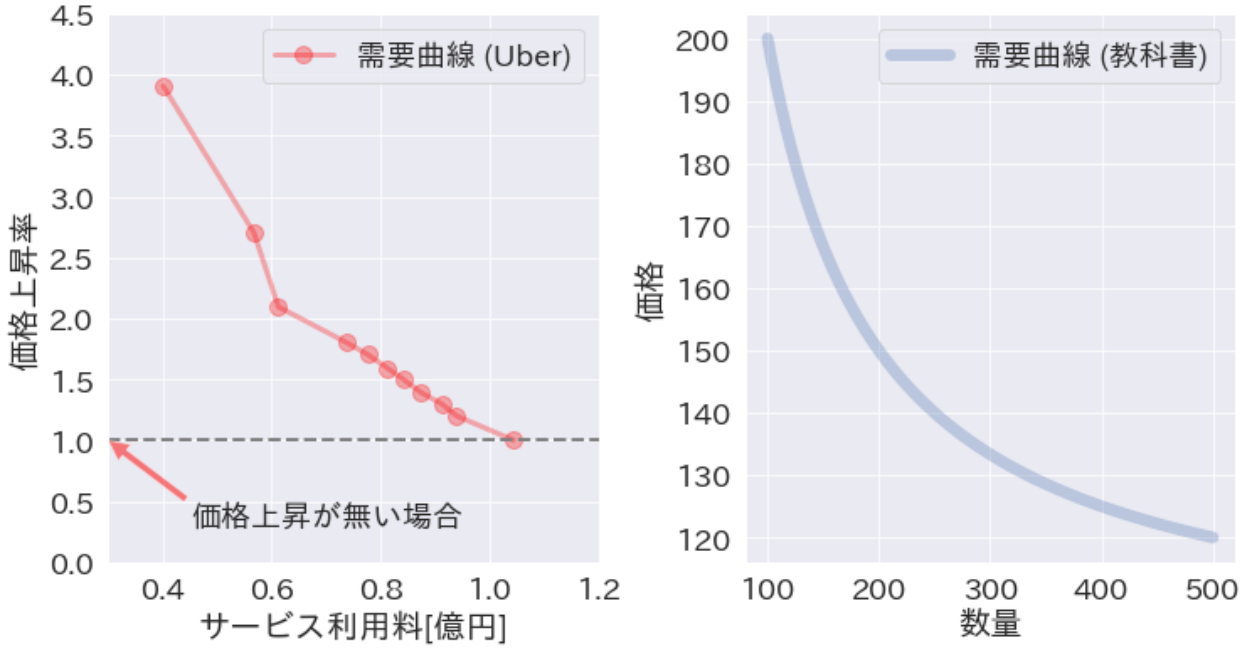
どの店で，いつどの寿司がレーンに流され，いつ食べられたのか，どのテーブルでいつどの商品が注文されたのか，などのデータを毎年 10 億件以上蓄積することで，需要を予測し，レーンに流す寿司の量をコントロールしている．この結果，廃棄量を 1/4 におさえることに成功しているという

どのようなデータが活用されているか 2

Uber

Uber は路上に出ているドライバーの数に比べて利用者数が多い場合には, 利用価格を上げている.

このような利用価格の設定で重要なのが需要曲線です (中学校で習います). 実際の利用データから, この需要曲線を求めたのが左の図です.



論文 Cohen et al (2016), *"Using big data to estimate consumer surplus: The case of uber"* のデータを基に作成.
「データ分析の力 因果関係に迫る思考法 (伊藤公一郎著)」 も参照.

教科書などでは架空の需要曲線 (右の図) が紹介されている.

どのようなデータが活用されているか 3

Web マーケティング

インターネット技術の発達により，Web を用いたマーケティングが広く行われるようになってきた.

オバマ元大統領は Web ページのメッセージ (1. Sign Up; 2. Sign Up Now; 3. Learn More; 4. Join Us Now) や写真 (下図. 検証した案には，この他に 3 つの動画もあった) をどのように構成すれば，最も効果的かを，Web ページの訪問者 (約 31 万人) のデータから分析し，約 72 億円の追加的選挙支援金を得たそうです.

学校教育における多変量データの例

問: 学校教育で多変量データを最初に扱うのはいつ?

答: 小学校

次の表は小学校 6 年生の算数の教科書に出ている表である.

6年1組

番号	記録(m)	番号	記録(m)	番号	記録(m)
①	14	⑪	19	㉑	32
②	24	⑫	25	㉒	28
③	29	⑬	40	㉓	29
④	14	⑭	33	㉔	18
⑤	38	⑮	23	㉕	17
⑥	22	⑯	37	㉖	20
⑦	33	⑰	26	㉗	21
⑧	24	⑱	24	㉘	29
⑨	36	⑲	23		
⑩	40	⑳	32		

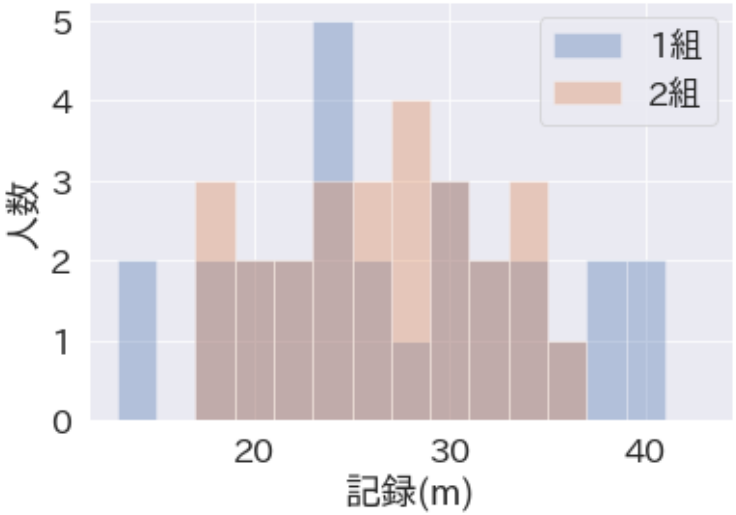
6年2組

番号	記録(m)	番号	記録(m)	番号	記録(m)
①	23	⑪	26	㉑	27
②	18	⑫	30	㉒	29
③	31	⑬	24	㉓	33
④	35	⑭	21	㉔	17
⑤	22	⑮	26	㉕	26
⑥	28	⑯	20	㉖	23
⑦	27	⑰	30		
⑧	19	⑱	18		
⑨	34	⑲	32		
⑩	33	⑳	28		

このデータは次のような表にまとめることができる.

名前	記録 (m)	組
児童 1	14	1
児童 2	24	1
⋮	⋮	⋮
児童 27	21	1
児童 28	28	1
児童 29	23	2
児童 30	18	2
⋮	⋮	⋮
児童 53	26	2
児童 54	23	2

学校教育における多変量データの例 (発展 1)



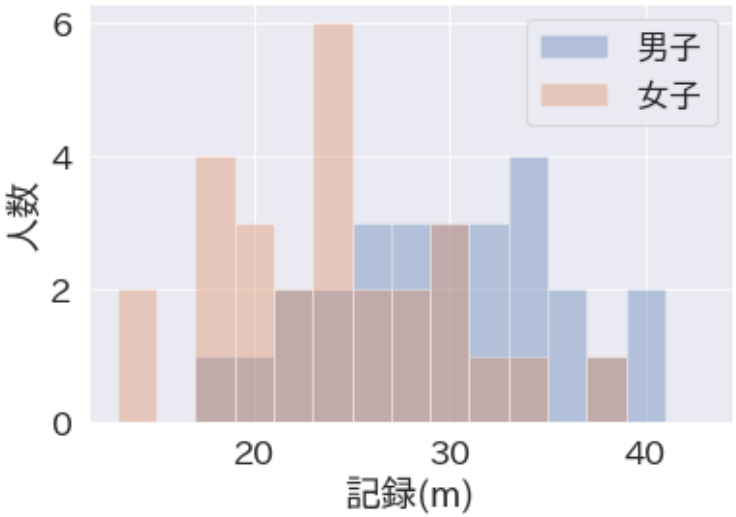


表: 小学校のデータに性別の情報を追加 (3 変量データ)

名前	記録 (m)	組	性別
児童 1	14	1	女子
児童 2	24	1	女子
⋮	⋮	⋮	⋮
児童 27	21	1	男子
児童 28	28	1	女子
児童 29	23	2	男子
児童 30	18	2	女子
⋮	⋮	⋮	⋮
児童 53	26	2	女子
児童 54	23	2	男子

問: 他にどのようなヒストグラムがかけますか?

変量が増えるとヒストグラムの種類も増える.

学校教育における多変量データの例 (発展 2)

表: 小学校のデータに性別の情報を追加 (3 変量データ)

名前	記録 (m)	組	性別
児童 1	14	1	女子
児童 2	24	1	女子
⋮	⋮	⋮	⋮
児童 27	21	1	男子
児童 28	28	1	女子
児童 29	23	2	男子
児童 30	18	2	女子
⋮	⋮	⋮	⋮
児童 53	26	2	女子
児童 54	23	2	男子

上の表 (relational database) から, ピボットテーブルを作成することができる.

例えば, 次の表がピボットテーブルの例である:

表: 25m 以上投げた人の人数

性別 \ 組	1 組	2 組
女子	4	6
男子	10	10

(注) このように表内部が度数である場合, ピボットテーブルは「クロス集計表」と呼ばれる.

問: ピボットテーブルのような概念を学校教育で最初に扱うのはいつ?

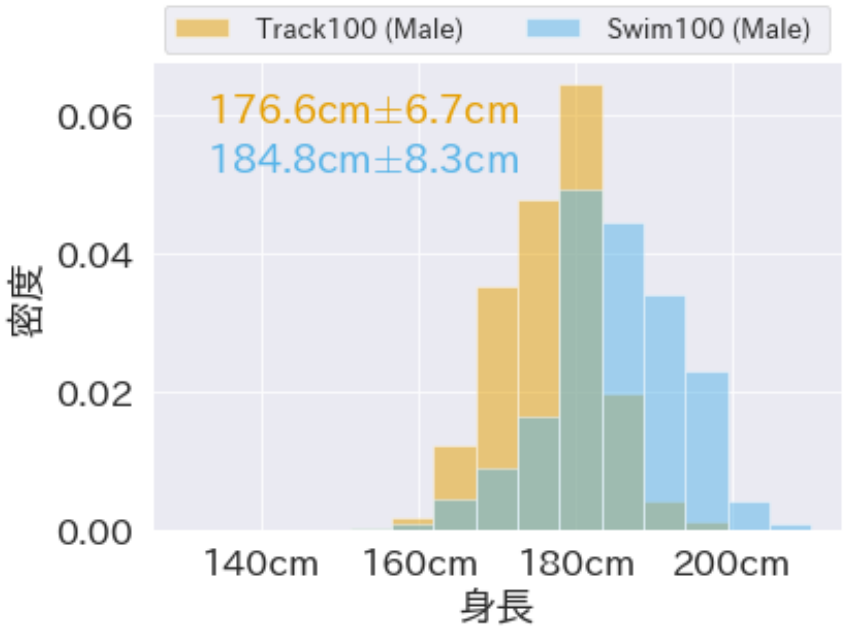
小学校 4 年生: 「データを二つの観点から分類整理する方法を知ること (指導要領 H29.9)」

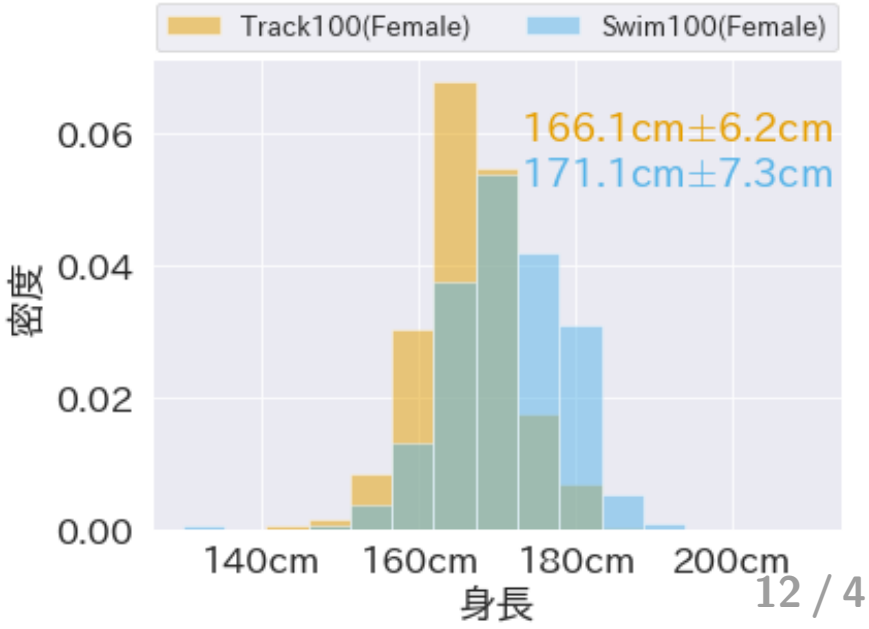
要約統計量

右の図はオリンピックの陸上 100m 短距離と水泳 100m 自由形に出場した選手の身長分布を表している (上は男性, 下は女性; このような図をヒストグラムという).

一見して水泳の方が身長が高いことが分かるが, 例えば, 「男女差が大きいのはどちらの競技か?」を図がら読み取ることは難しい. このように精密な比較をする場合には平均値や標準偏差などんの要約統計量が便利である.

データソース: Kaggle





平均値

平均値は次のように定義される:

平均値 (小学校)

$$(\text{平均値}) = \frac{(\text{データの中の全ての値の総和})}{(\text{データの数})}$$

元になるデータがあれば上のように計算すれば良いが, 集計後の表 (度数分布表) しか入手できないケースが多い. そのような場合は, 度数分布表から平均値を求める:

表から計算する場合の平均値 (中学数学)

$$(\text{平均値}) \fallingdotseq \frac{\sum[(\text{階級値}) \times (\text{度数})]}{\sum(\text{度数})}$$

問

この式の近似 $\fallingdotseq$ は等号 $=$ であることもある. どういうとき, 等号になるのだろうか?

表から計算する平均値

23	10	8	22	9	13	13	15
17	10	17	20	21	4	9	21
19	6	11	23	27	25	8	25
15	15	15	18	20	14	8	18
18	11	11	19	18	8	8	11
26	9	18	15	21	12	8	13
18	4	12	17	20	18	11	19
15	6	11	24	23	16	13	15
22	8	8	17	15	7	13	9
12	12	11	18	16	15	9	19

階級 (以上)		階級 (未満)	階級値	度数
2	–	5	3.5	2
5	–	8	6.5	3
8	–	11	9.5	15
11	–	14	12.5	16
14	–	17	15.5	12
17	–	20	18.5	16
20	–	23	21.5	8
23	–	26	24.5	6
26	–	29	27.5	2
計				80

$$(\text{平均値}) = \frac{23+10+8+22++9+13+13+15+\cdots+12+12+11+18+16+15+9+19}{80}$$

$$= \frac{4+4+6+6+7+8+8+8+8+8+\cdots+23+23+23+24+25+25+26+27}{80}$$

$$= \frac{(4+4)+(6+6+7)+(8+8+8+8+8+\cdots)+\cdots+(23+23+23+24+25+25)+(26+27)}{80}$$

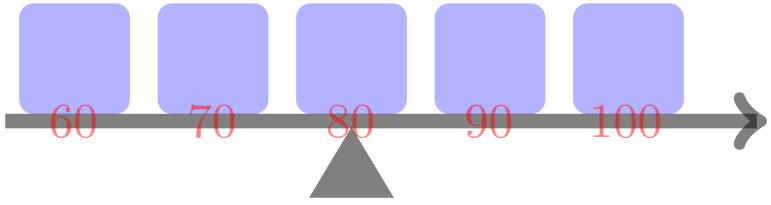
$$\approx \frac{(3.5\times 2)+(6.5\times 3)+(9.5\times 15)+\cdots+(24.5\times 6)+(27.5\times 2)}{80} = \frac{\sum \text{[(階級値)}\times\text{(度数)]}}{\sum \text{(度数)}}$$

平均値の計算例と意味

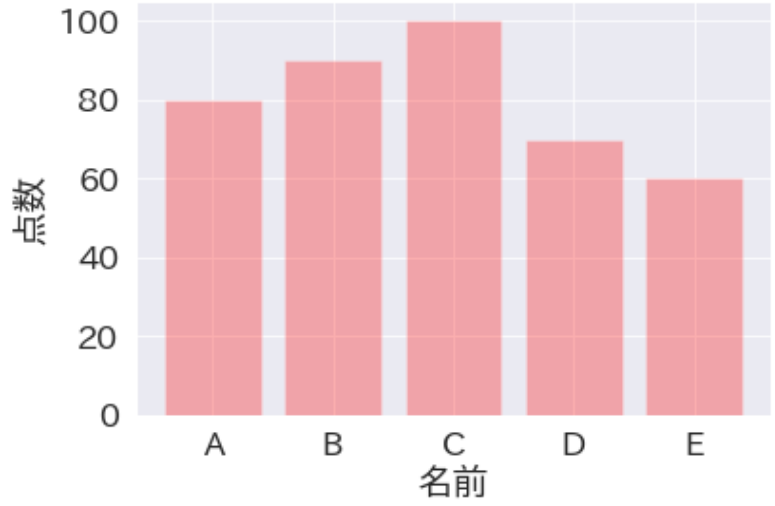
名前	点数
A	80
B	90
C	100
D	70
E	60

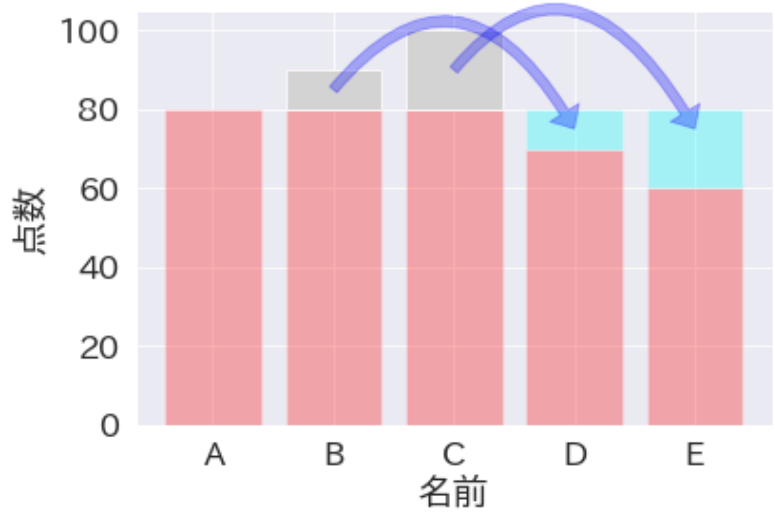
$$(\textbf{平均値}) = \frac{80+90+100+70+60}{5}$$

$$= 80 \ (\textbf{点})$$



釣り合いの位置が平均値 (重心)





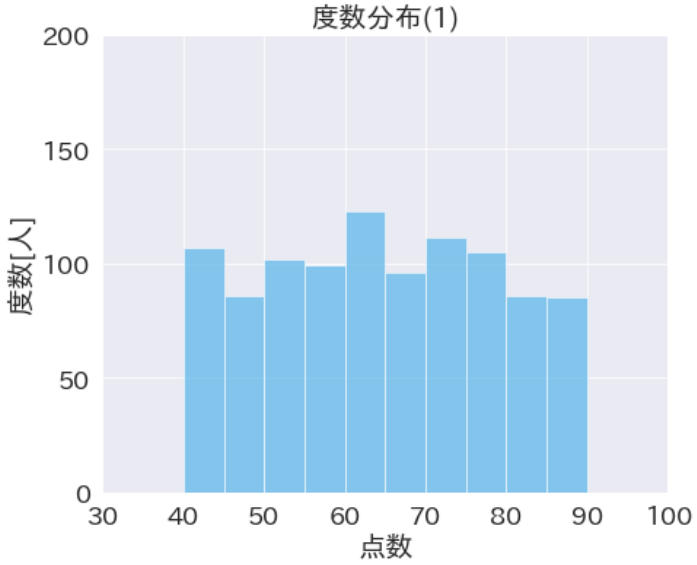
凸凹をならした値が平均値

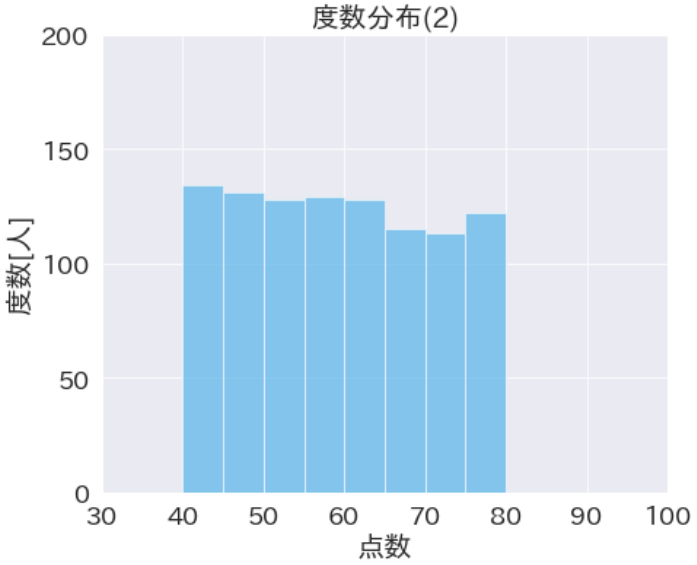
小テスト: 平均値とヒストグラム

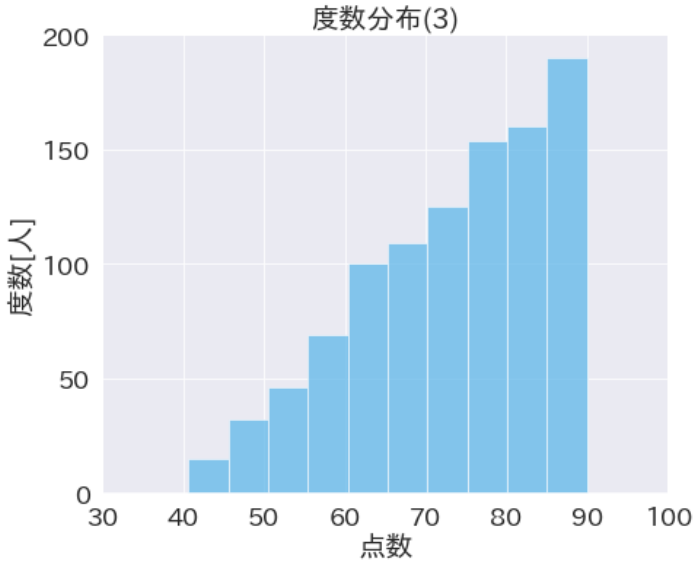
問

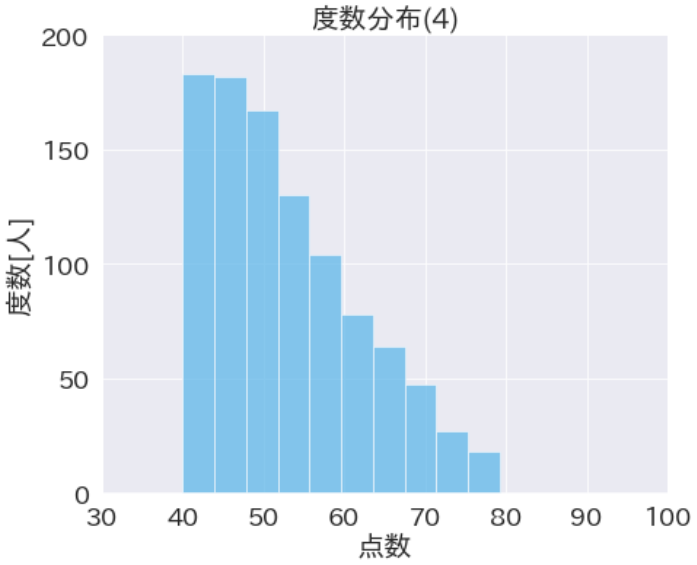
右の 4 つの度数分布 (1)〜(4) について, それぞれの平均値を m_1, m_2, m_3, m_4 とする. 平均値の大小関係を調べよ.

どの度数分布も階級幅は 5 点にしてある.









平均値の計算練習

問

みかん 10 **個の重さを調べた所,**

$$115.9 \quad 117.5 \quad 103.7 \quad 95.4 \quad 92.1$$
$$103.3 \quad 92.8 \quad 88.6 \quad 105.0 \quad 104.3$$

であった. みかんの重さの平均値を求めよ.

解 定義通り計算すると

$$\frac{115.9 + 117.5 + 103.7 + 95.4 + 92.1 + 103.3 + 92.8 + 88.6 + 105.0 + 104.3}{10} = 101.9 \text{ (g)}$$

となる. データから平均値がおおよそ 100 **位であることが予想できるので, 次のように計算することもできる:**

$$\frac{15.9 + 17.5 + 3.7 - 4.6 - 7.9 + 3.3 - 7.2 - 11.4 + 5.0 + 4.3}{10} = 1.9 \text{ (g)}$$

これに 100 **を足して,** 101.9(g).

中央値

中央値とはデータを小さい順に並べたとき，ちょうど中央に位置する数値のことである．要素数 N が奇数の場合は

$$(N \text{ が奇数の場合}) \qquad a_1 \leq a_2 \leq \cdots \leq \underbrace{a_{\frac{N-1}{2}}}_{\text{中央値}} \leq \cdots \leq a_{N-1} \leq a_N$$

例えば，$N = 5$ の場合，a_1, a_2, a_3, a_4, a_5 であり，a_3 が中央値である．
要素数 N が偶数の場合には，ちょうど中央に位置する数値が無い:

$$(N \text{ が偶数の場合}) \qquad a_1 \leq a_2 \leq \cdots \leq a_{\frac{N}{2}} \leq a_{\frac{N}{2}+1} \leq \cdots \leq a_{N-1} \leq a_N$$

($N = 4$ とすると，a_1, a_2, a_3, a_4 だが，a_2 も a_3 も中央ではない)．そこで，この場合は中央付近の 2 数の平均

$$\frac{a_{\frac{N}{2}} + a_{\frac{N}{2}+1}}{2}$$

を中央値とする．

中央値の計算例と意味

データが奇数個のとき

名前	点数
A	80
B	90
C	100
D	70
E	60

60 70 80 90 100

中央値

中央値の性質

中央値の左側と右側には同じ個数のデータがある.

平均値にはこのような性質は無い. 例えばほとんどのデータが平均値の左側にあることもあり得る.

性質の違いを理解した上で, 適切な方を選択することが重要.

データが偶数個のとき

名前	点数
A	80
B	90
C	100
D	70
E	60
F	90

60 70 80 90 90 100

中央値 $= \dfrac{80+90}{2} = 85$

中央値の計算練習

問

みかん 10 **個の重さを調べた所,**

$$115.9 \quad 117.5 \quad 103.7 \quad 95.4 \quad 92.1$$
$$103.3 \quad 92.8 \quad 88.6 \quad 105.0 \quad 104.3$$

であった. みかんの重さの中央値を求めよ.

解 小さい順に並べると

$$88.6 \quad 92.1 \quad 92.8 \quad 95.4 \quad 103.3 \quad 103.7 \quad 104.3 \quad 105.0 \quad 115.9 \quad 117.5$$

$$\text{中央値} = \frac{103.3 + 103.7}{2} = 103.5\,(\mathrm{g})$$

平均値と中央値の違いについて

名前	ひと月のお小遣い
A くん	1000 円
B さん	2000 円
C くん	500 円
D くん	700 円
E さん	1500 円
F くん	1200 円
G さん	800 円
H さん	1000 円
I さん	2500 円
ビルゲイツくん	50000 円

ビルゲイツくんはお小遣いが非常に多い．このような極端な数値を持つデータを**外れ値**と呼ぶ．

左の表について平均値と中央値は

$$(\text{平均値}) = 6000\ (\text{円})$$

$$(\text{中央値}) = 1100\ (\text{円})$$

10 人 9 人は平均値の半分以下である．このように外れ値が存在すると，平均値は典型的な値を表しているとは言いがたいことがある．

一方，中央値は外れ値に影響を受けにくい．実際，ビルゲイツくんがいなかった場合

$$(\text{平均値}) = 1111\ (\text{円})$$

$$(\text{中央値}) = 1000\ (\text{円})$$

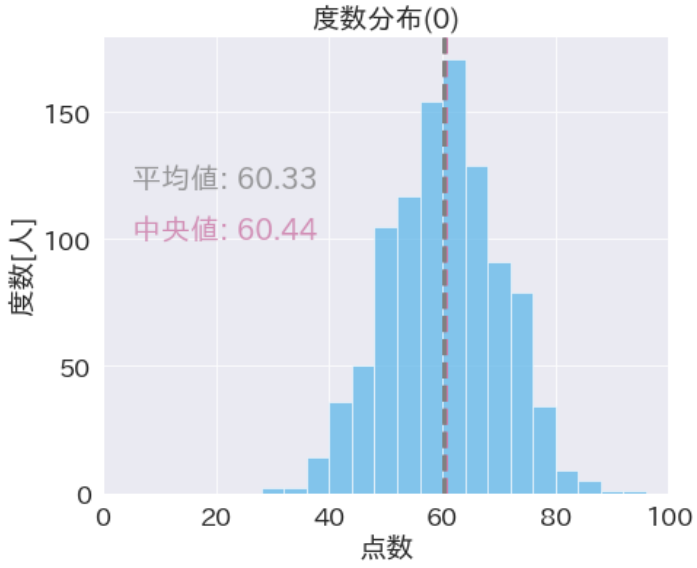
となり，中央値は変化が小さい．

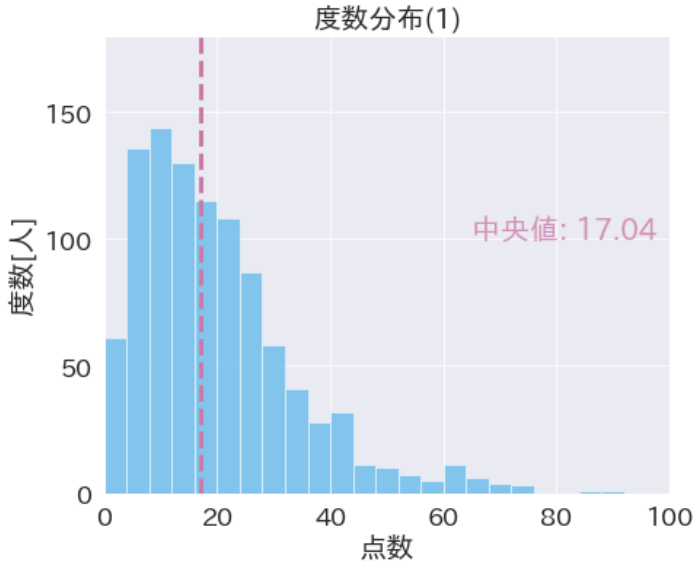
小テスト: 中央値とヒストグラム

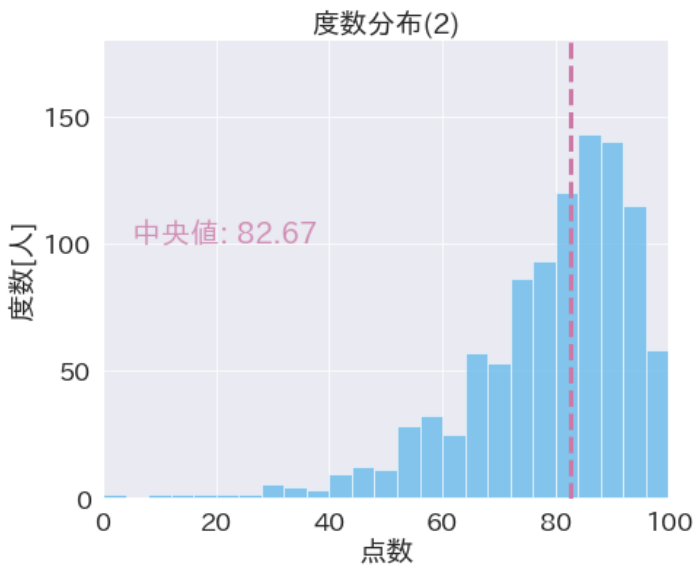
ヒストグラムが左右対称な場合 [度数分布 (0)], 平均値と中央値はほぼ一致する.

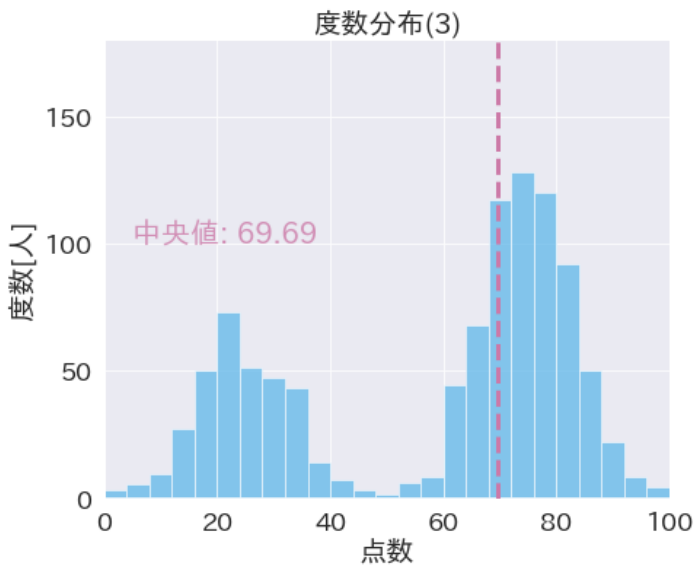
問

ヒストグラムが左右非対称な場合に, 平均値と中央値の違いが表われる. 度数分布 (1)〜(3) のうち, 平均値が中央値より大きいものを全て選べ.







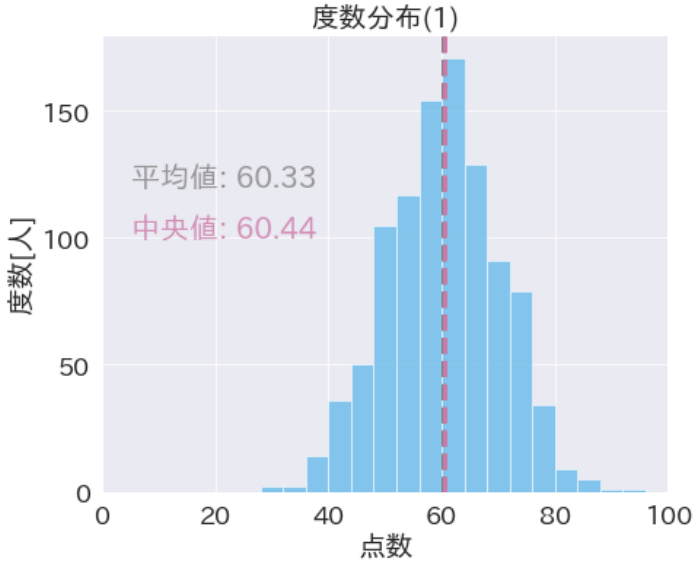


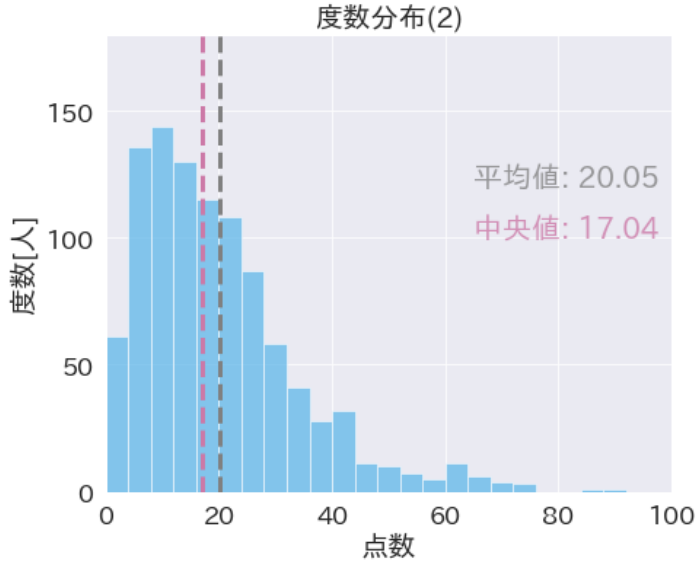
小テスト: 中央値とヒストグラム: 解答

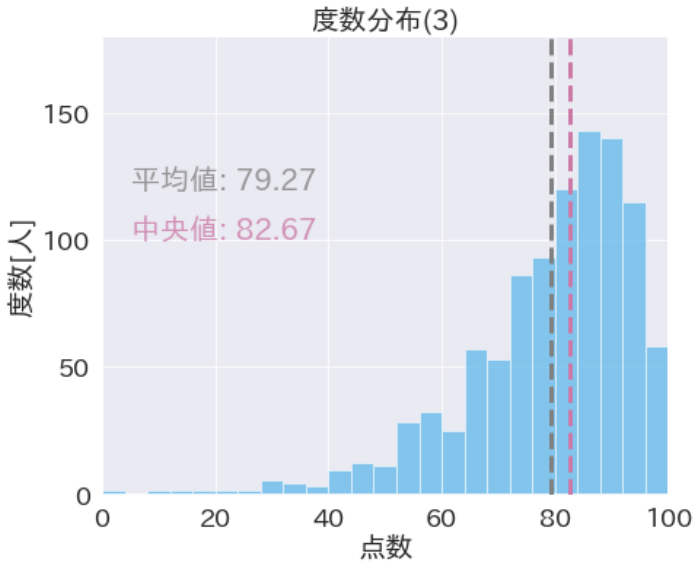
中央値 (赤線) の両側には 50% ずつの人数がいる. このうち片側 (例えば右側) の分布を変更して, 左右対称な分布にする「操作」を考えよ.

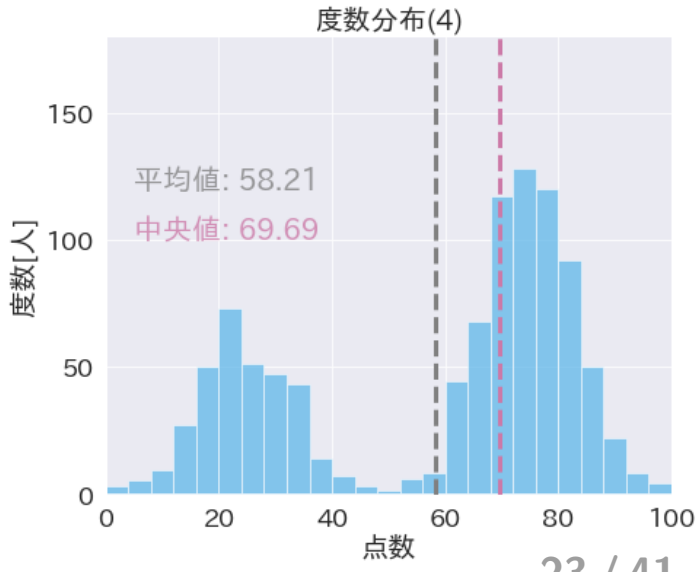
例えば, 度数分布 (2) の右側にこの「操作」を行うと, 平均値は減少するだろう.

一方, この「操作」後の平均値と中央値は一致するから, 元の平均値は中央値より大きかったはずだ.







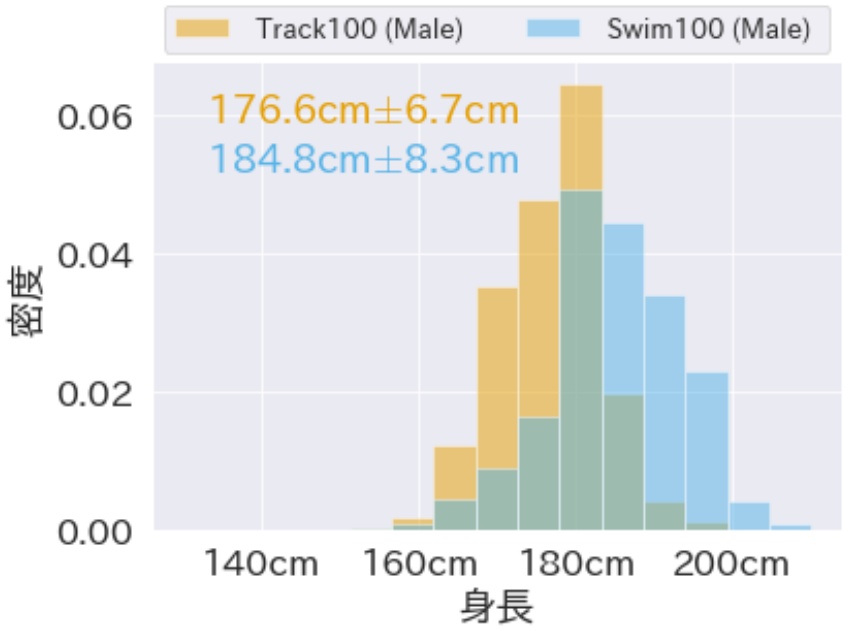


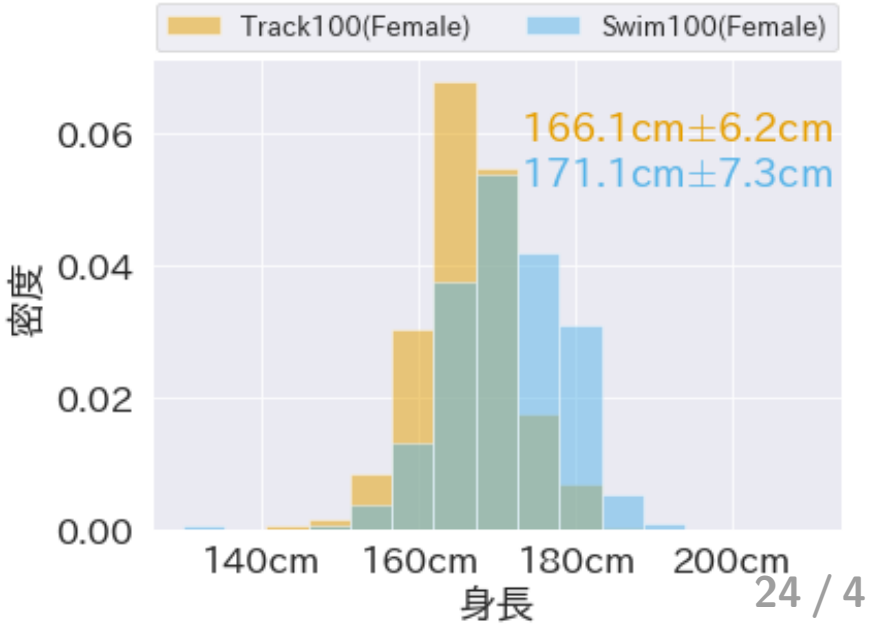
要約統計量

右の図はオリンピックの陸上 100m 短距離と水泳 100m 自由形に出場した選手の身長分布を表している (上は男性, 下は女性; このような図をヒストグラムという).

一見して水泳の方が身長が高いことが分かるが, 例えば, 「男女差が大きいのはどちらの競技か?」を図がら読み取ることは難しい. このように精密な比較をする場合には平均値や標準偏差などんの要約統計量が便利である.

データソース: Kaggle





散らばりの指標: 分散

5 つのデータを v, w, x, y, z とすると，平均値は次のように表わせる:

$$\overline{X} = \frac{v + w + x + y + z}{5}$$

このとき分散 S^2 は次のように定義される:

$$S^2 = \frac{(v - \overline{X})^2 + (w - \overline{X})^2 + (x - \overline{X})^2 + (y - \overline{X})^2 + (z - \overline{X})^2}{5}$$

分散 S^2 の平方根 S を標準偏差という.

$$\overline{X} - S \sim \overline{X} + S$$

がおおよその散らばりの範囲となっている.

分散の定義式は次のように式変形できる:

$$\frac{v^2 + w^2 + x^2 + y^2 + z^2}{5} - \left(\frac{v + w + x + y + z}{5}\right)^2$$

(この式で計算する方が楽なことが多いが，データの値が大きい場合，コンピュータで計算すると誤差が大きくなることがある)

分散の計算例と意味

名前	点数
A	80
B	90
C	100
D	70
E	60

$\overline{X} = 80$ **なので,**

(分散)

$$= \frac{(80-80)^2 + (90-80)^2 + (100-80)^2 + (70-80)^2 + (60-80)^2}{5}$$

$$= \frac{1000}{5} = 200$$

名前	点数	偏差	偏差の 2 乗
A	80	0	0
B	90	10	100
C	100	20	400
D	70	-10	100
E	60	-20	400
和			1000

表中であらかじめ計算をしておくと計算ミスが少ない:

$$(\textbf{分散}) = \frac{1000}{5} = 200$$

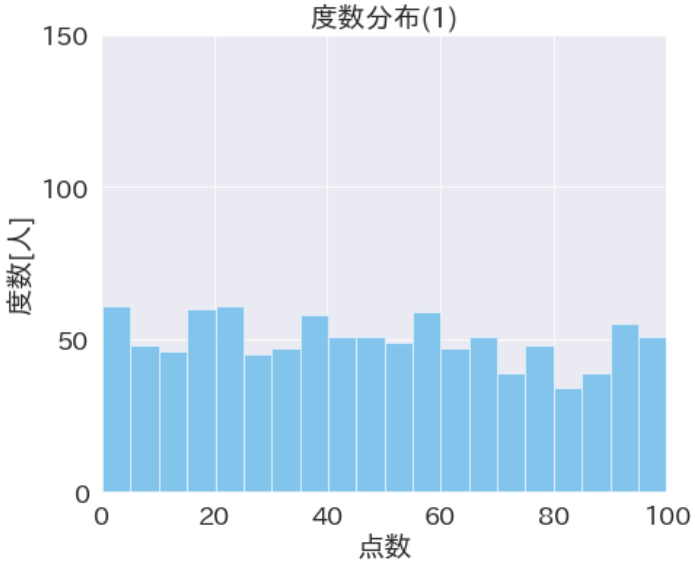
分散の平方根 $\sqrt{200} \fallingdotseq 14.1$ **が標準偏差である. データの散らばりの範囲はおおよそ**

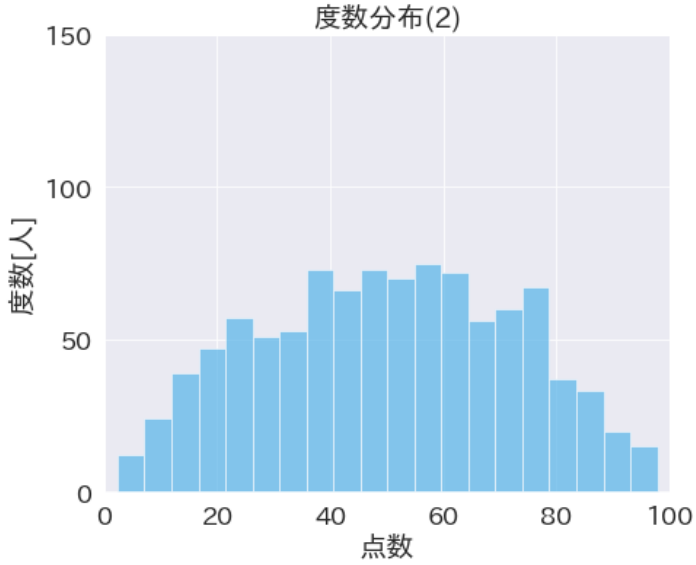
$$\underbrace{80 - 14.1}_{65.9} \sim \underbrace{80 + 14.1}_{94.1}$$

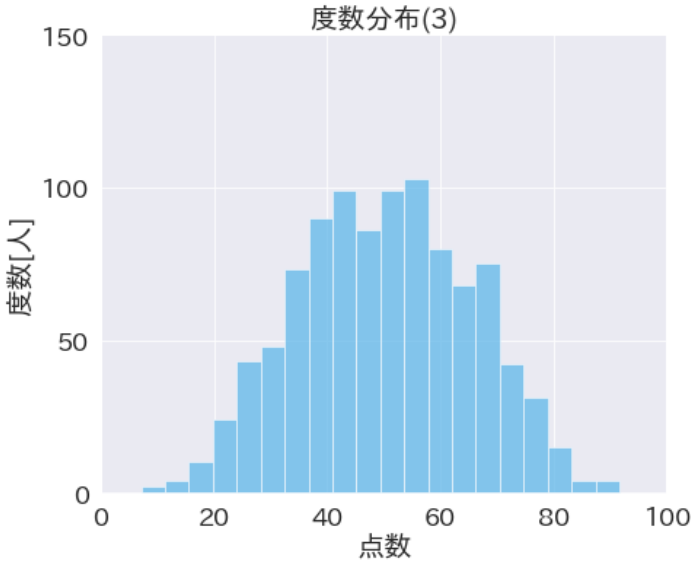
小テスト: 標準偏差とヒストグラム

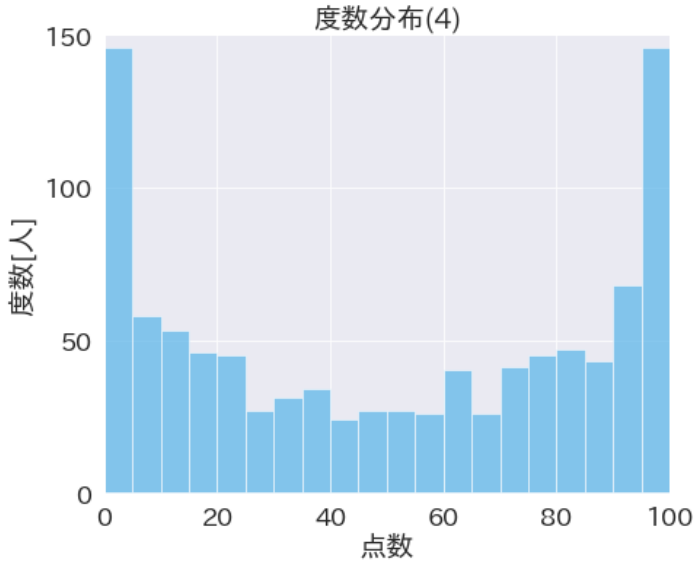
問

右の 4 つの度数分布 (1)〜(4) について，それぞれの標準偏差を s_1, s_2, s_3, s_4 とする．標準偏差の大小関係を調べよ．









分散の計算練習

問

みかん 10 **個の重さを調べた所,**

$$115.9 \quad 117.5 \quad 103.7 \quad 95.4 \quad 92.1$$
$$103.3 \quad 92.8 \quad 88.6 \quad 105.0 \quad 104.3$$

であった. みかんの重さの分散を求めよ. 平均値は既に求めた: $101.9\,(\mathrm{g})$

解 定義通り計算すると

$$\frac{(115.9-101.9)^2+(117.5-101.9)^2+\cdots+(104.3-101.9)^2}{10}=85.8$$

となる. 標準偏差は 9.3 **となり,**

$$\underbrace{101.9-9.3}_{92.6} \quad \sim \quad \underbrace{101.9+9.3}_{111.2}$$

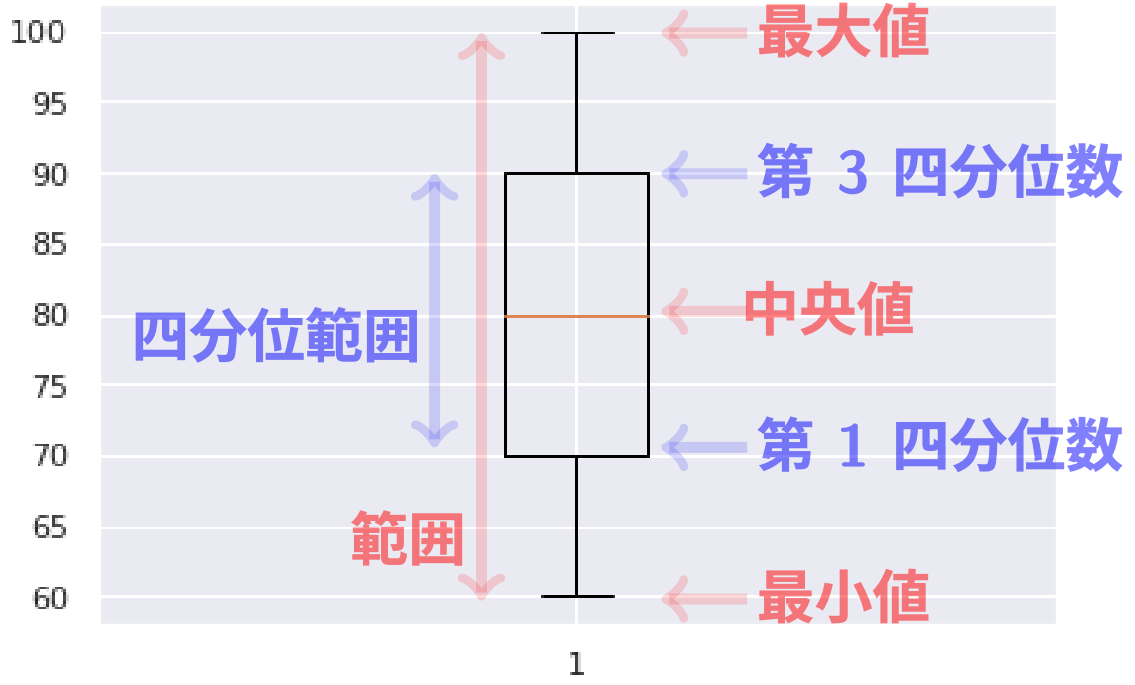
の範囲に多くのデータ (6 **割) が含まれている.**

範囲と四分位範囲

名前	点数
A さん	80
B さん	90
C さん	100
D さん	70
E さん	60

順番に並べると

60　70　80　90　100



(注) 最大値と最小値を示す部分は「ひげ (wisker)」と呼ばれる．「ひげ」の位置は最大値・最小値以外の値にとることも多い．箱ひげ図の元々の定義では，

$$\text{第 3 四分位数} + 1.5 \times \text{四分位範囲}$$

$$\text{第 1 四分位数} - 1.5 \times \text{四分位範囲}$$

をひげの位置にとっていた．この範囲から外れたものは，外れ値としてプロットされる．

四分位数・四分位範囲の計算例

要素数が偶数の場合

50 65 71 | 78 82 95

この部分の中央値 $(= 65)$ が第 1 四分位数

この部分の中央値 $(= 82)$ が第 3 四分位数

要素数が奇数の場合 (中央値を除く)

50 65 71 | 75 | 78 82 95

この部分の中央値 $(= 65)$ が第 1 四分位数

この部分の中央値 $(= 82)$ が第 3 四分位数

要素数が奇数の場合 (中央値を含む)

50 65 71 75 | 78 82 95

この部分の中央値 $(= 68)$ が第 1 四分位数

この部分の中央値 $(= 80)$ が第 3 四分位数

奇数の場合が難しそうだが，データ数が多ければどちらの方法で計算しても近い数値になるから，あまり気にする必要はない．

四分位数・四分位範囲の計算練習

問

みかん 10 **個の重さを調べた所,**

$$115.9 \quad 117.5 \quad 103.7 \quad 95.4 \quad 92.1$$
$$103.3 \quad 92.8 \quad 88.6 \quad 105.0 \quad 104.3$$

であった. みかんの重さの四分位数を求めよ.

解 **小さい順に並べると**

88.6 92.1 92.8 95.4 103.3 | 103.7 104.3 105.0 115.9 117.5

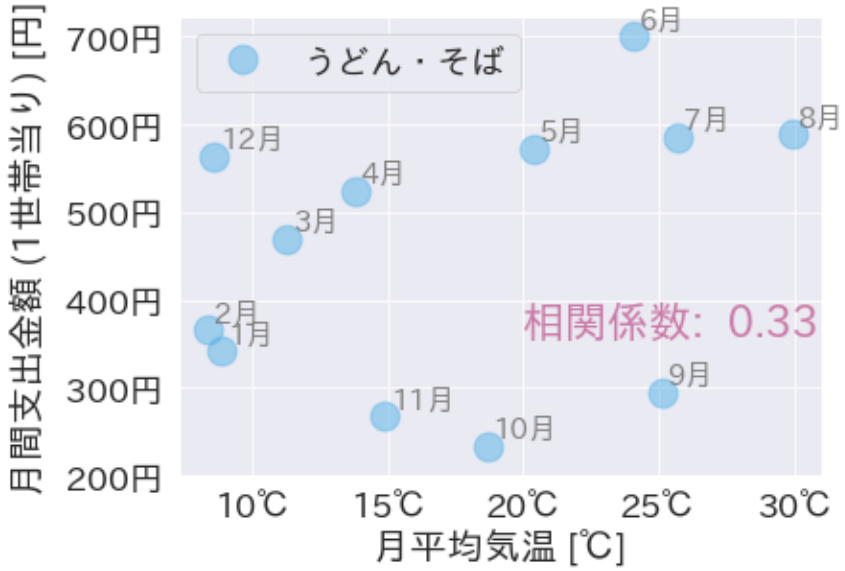
この部分の中央値 $(= 92.8)$ **が第** 1 **四分位数**

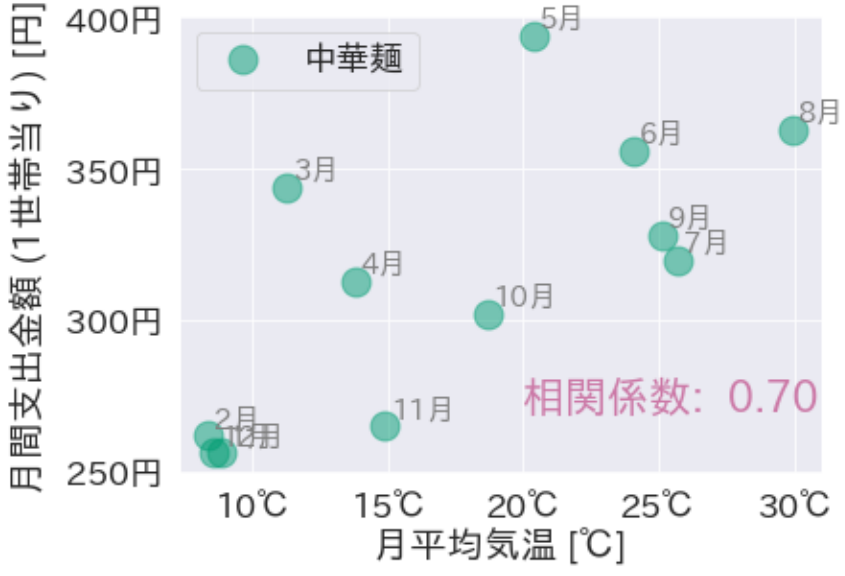
この部分の中央値 $(= 105.0)$ **が第** 3 **四分位数**

関係性の指標

代表値や散らばりの指標は，1 つの変数についての統計量であったが，複数の変数間の関係性を知りたい場合もある．

2 変数間の関係性は散布図で表わすと見通しがよい．右図は「うどん・そば」および「中華麺」の支出額と平均気温の関係を表している (全て徳島市の 2020 年のデータ)．





問

うどん・そばも中華麺も気温が高い程，支出額が増える傾向がありそうだが，どちらの方がその傾向が強いだろうか?

→相関係数という指標が便利である．

月間支出額の データソース: e-STAT

月平均気温の データソース: 気象庁

共分散

名前	数学の点数	英語の点数
A	80	88
B	75	65
C	60	70
D	90	85
E	67	83
F	54	77
平均値	71.0	78.0
標準偏差	12.1	8.2

ここで，平均値と標準偏差を次のように記号を用いて表わす:

$\overline{X}$: 数学の点数の平均値

$\overline{Y}$: 英語の点数の平均値

S_x : 数学の点数の標準偏差

S_y : 英語の点数の標準偏差

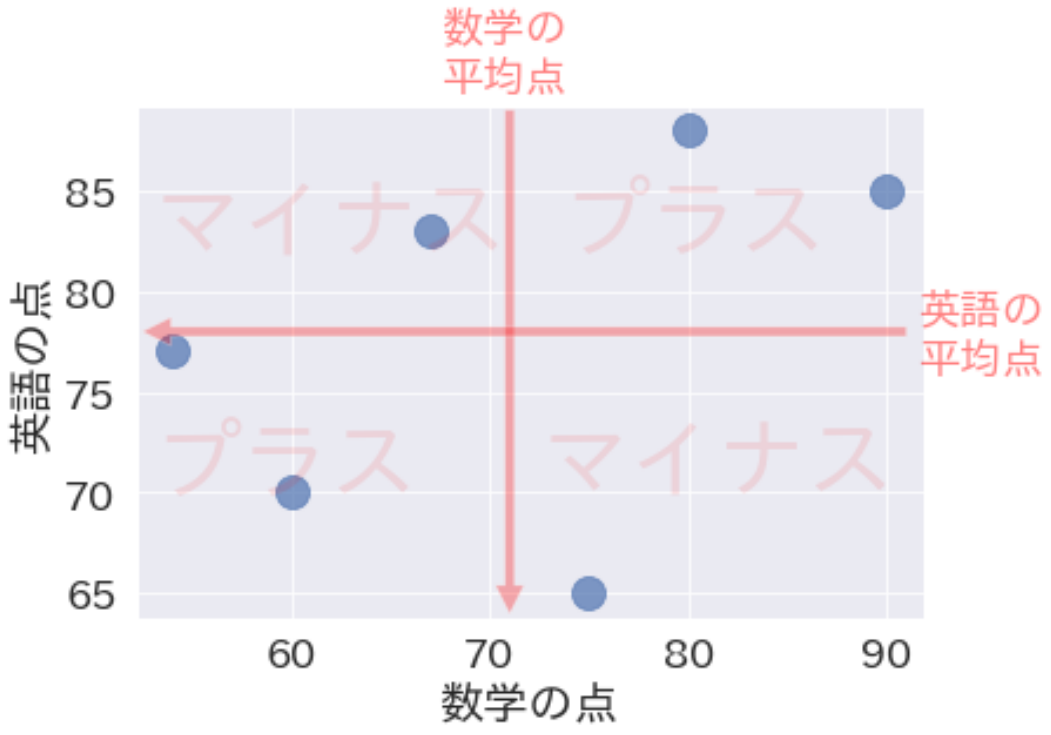
数学と英語の点数の関係性を調べるために共分散を定義する.

共分散

$$S_{xy} = \frac{\overbrace{(80-\overline{X})(88-\overline{Y})}^{A\text{ さん}} + \overbrace{(75-\overline{X})(65-\overline{Y})}^{B\text{ さん}} + \cdots + \overbrace{(54-\overline{X})(70-\overline{Y})}^{F\text{ さん}}}{6}$$

共分散の意味

名前	数学の点数	英語の点数
A	80	88
B	75	65
C	60	70
D	90	85
E	67	83
F	54	77
平均値	71.0	78.0
標準偏差	12.1	8.2



共分散

$$S_{xy} = \frac{\overbrace{(80-\overline{X})(88-\overline{Y})}^{A\text{ さん}} + \overbrace{(75-\overline{X})(65-\overline{Y})}^{B\text{ さん}} + \cdots + \overbrace{(54-\overline{X})(77-\overline{Y})}^{F\text{ さん}}}{6}$$

相関係数

先の例では共分散の値は 42.7 となる．もし試験が 100 点満点ではなく 200 点満点の場合，共分散の値は 4 倍になってしまう．このように，共分散はデータの数値の大きさに依存してしまうため，関係性の指標としては使いにくい．

そこで，相関係数という統計量を定義する．

相関係数

$$R_{xy} = \frac{S_{xy}}{S_x S_y}$$

解説 相関係数は X と Y の相関の強さを表し，その値が

- 1 に近いときは相関が強い (数学の点が良い人は，英語の点も良い)
- 0 に近いときは相関が弱い (数学の点と英語の点の間に関連がない)
- -1 に近いときは，反相関が強い (数学の点が良い人は，英語の点が悪い)

このように相関係数は 2 つのことがらの関連性を調べたい場合によく用いられる．

相関係数の計算練習

名前	数学の点数	英語の点数
A	80	88
B	75	65
C	60	70
D	90	85
E	67	83
F	54	77
平均値	71.0	78.0
標準偏差	12.1	8.2

問

相関係数 $R_{xy} = \frac{S_{xy}}{S_x S_y}$ を計算せよ.

まず，共分散 S_{xy} は

$$S_{xy} = 42.7$$

となる．したがって相関係数は

$$R_{xy} = \frac{42.7}{12.1 \times 8.2} = 0.43$$

共分散

$$S_{xy} = \frac{\overbrace{(80-\overline{X})(88-\overline{Y})}^{A\text{ さん}} + \overbrace{(75-\overline{X})(65-\overline{Y})}^{B\text{ さん}} + \cdots + \overbrace{(54-\overline{X})(77-\overline{Y})}^{F\text{ さん}}}{6}$$

観察とランダム化実験 1

データの収集方法には観察とランダム化実験がある．このページと次ページの内容は，「データ分析のための統計学入門」 (国友直人他訳) を参考にしました．

観察

データ生成に直接関与できない場合のデータ収集を「観察」という．

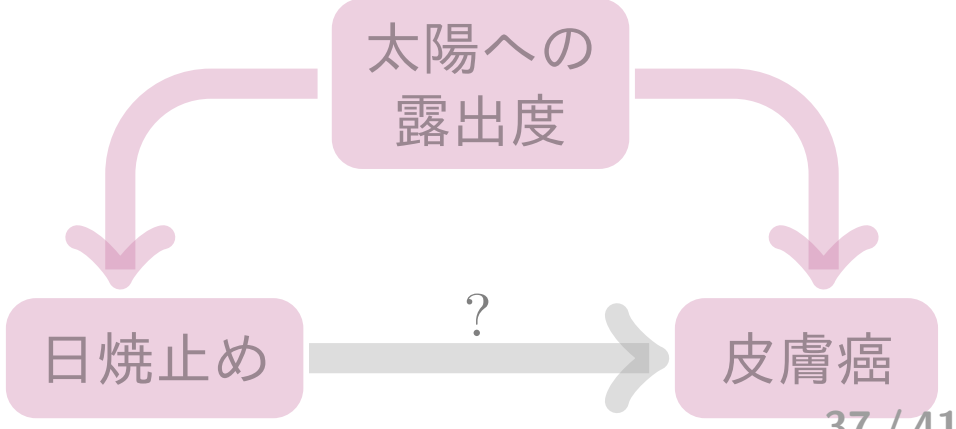
例としては気象データや地震のデータが挙げられる (雨を降らせたり，地震を起こしたりすることは普通できません)．

観察の場合，変量の間に相関関係があっても，因果関係があるとは一般に言えない (注)．

(注) 近年，因果関係に関するデータ分析手法の研究が進んで，観察データから因果関係を導く手法が提案されている．

問

日焼け止めの利用と皮膚癌についてのある観測研究により，日焼け止めをより利用した方が皮膚癌になりやすいと分かったとする．このことは日焼け止めに利用した物質が皮膚癌の原因となることを意味するだろうか?



観測と実験 2

因果関係とランダム化実験

「変数 A を変化させたとき (←これを介入という), 変数 B も変化する」ならば, 「変数 A から変数 B への因果関係がある」という.

このように因果関係を調べるには, 変数 A の値を変化させる「実験」を行い, そのときの変数 B の値のデータを収集することが必要である.

観測とは違い, 観察者が能動的に介入していることに注意.

実験の例

ある薬 M が血圧を下げるか (という因果関係) を調べたい場合, 患者をランダムに 2 群に分け, 最初の群の患者にプラセボ (偽薬), 第 2 の群の患者に薬 M が与えられる

ここでは「薬 M を摂取したかどうか」が変数 A であり, 「処置後の血圧」が変数 B である.

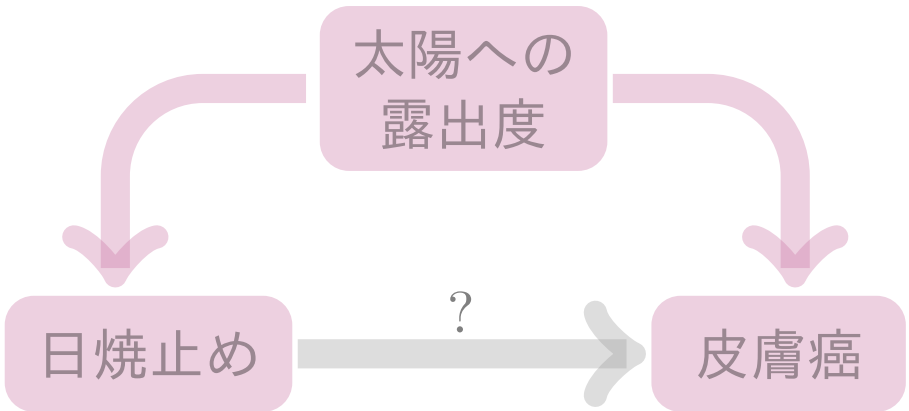
この 2 群の血圧に差異があれば (すなわち, 変数 A を変化させたとき変数 B も変化するならば), 薬 M には血圧を下げる効果 (因果関係) がある, と判断される.

疑似相関 1: 交絡

疑似相関

直接の因果関係のない 2 つの変数に相関関係が見られることを疑似相関と呼ぶ.

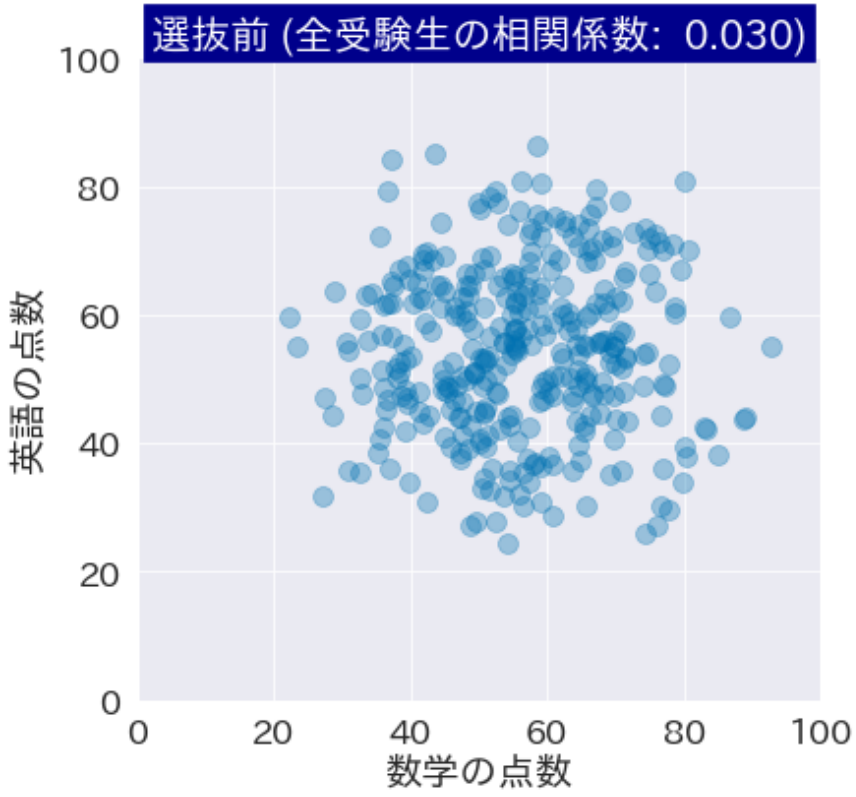
日焼止めの例のようなケースでは,「太陽への露出度」が「日焼止めを塗ること」と「皮膚癌になること」の双方に影響を与える. そのため, 例え相関があっても, 直接の因果関係があるとは言えない (あるかもしれないし, 無いかもしれない).

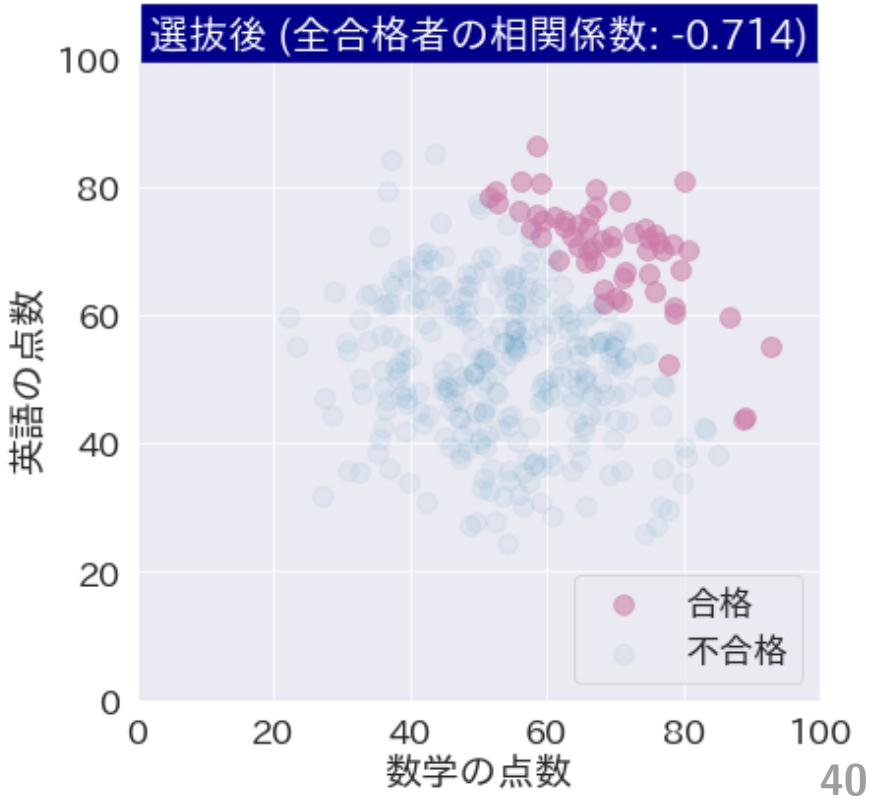


この共通の原因となる変数 (ここでは「太陽への露出度」) は交絡因子と呼ばれる. 交絡因子は疑似相関が発生する原因の 1 つである.

疑似相関 2: 合流点での選抜

元々相関のない 2 つの変数が, 特定のデータを選ぶことで, 相関のある変数になることがある. 下の例では元々「英語の点数 x」と「数学の点数 y」の間には相関はない (左図). しかし「合計点 $x+y$」が 130 点以上の人 (合格者: 右図の赤い点) だけを考えると, 比較的強い負の相関が見られる. この場合も因果関係があるわけではなく, データ点を選抜する過程が原因となって相関が見えているので, 疑似相関である.





疑似相関 3: 逆の因果関係・偶然生じる相関

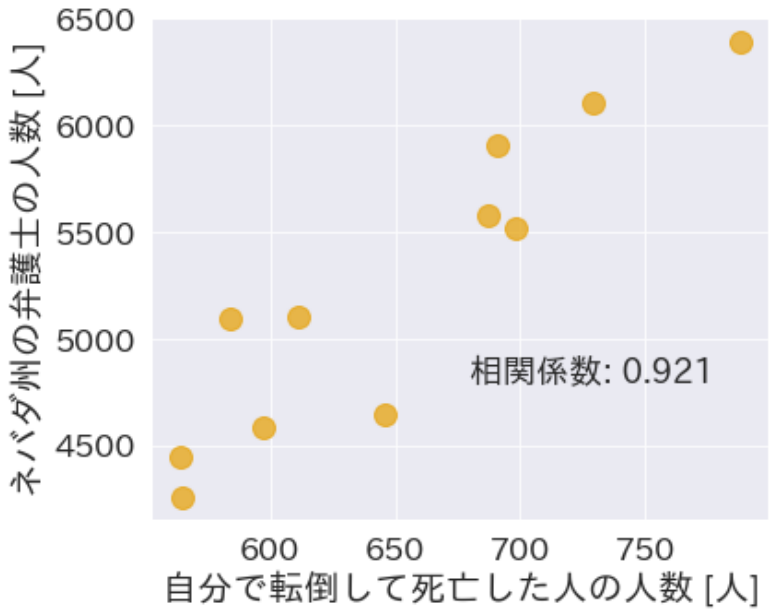
疑似相関には他にも「逆の因果関係」と「偶然生じる相関」がある.

逆の因果関係

例えば,「人口当りの警官の人数」と「犯罪件数」に相関があるとき,「警官が多いと犯罪件数が多くなる」と考えるのは因果関係が逆であろう.

偶然生じる相関

なんの関係性もないのに,全く偶然生じる相関



www.tylervigen.com/spurious-correlations

基礎情報教育: データ科学入門
第 4 章 可視化

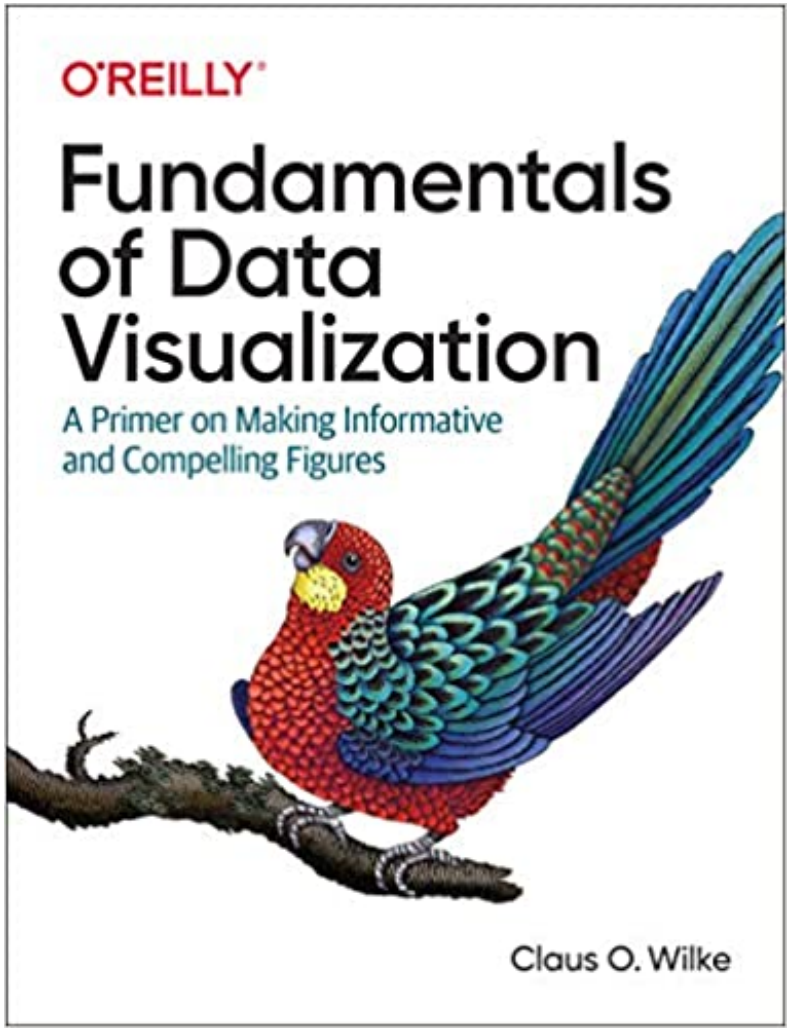
T. MIYAGUCHI

Naruto Universality of Education

講義全体のアウトライン

- 第 1 章: 社会と教育における変化
  - 仮説駆動とデータ駆動
  - データに基づく思考や判断
- 第 2 章: データ科学とは?
  - 統計学・計算機科学とデータ科学
  - AI・機械学習・倫理
- 第 3 章: データ分析の基礎
  - データとは?
  - 代表値・散らばりの指標・関係性の指標
- 第 4 章: 可視化
  - 可視化の必要性
  - 量の表現・割合の表現・分布の表現・関係性の表現・系列の表現
- 第 5 章: データ分析実習
  - 問を立てよう・統計量と可視化
  - 機械学習に挑戦・データ分析の実践
- 第 6 章: 様々な話題

参考文献



可視化に関する文献は少ないが，左の書籍 (洋書) は可視化を系統的に解説した数少ない書籍の１つである．本章の内容はこの本を参考にしている．

WEB からも読むことができるので，図を眺めてみるのも良いでしょう:

https://clauswilke.com/dataviz/index.html

なんのために可視化するのか?

データをグラフにすることを可視化とよぶ.
可視化する理由は何だろうか?

- データの把握・隠れた性質の発見:
  扱っているデータの把握がデータ分析の最初の一歩である. 統計量と可視化がデータ把握の最も基本的な方法である. グラフを工夫して作成することが, 大きな発見につながることもある.
- コミュニケーション:
  人に何かを伝える際に, (よく工夫された) グラフを用いると伝わりやすい.
- エラーや外れ値の検出:
  観測ミスやデータ処理におけるミスなど, データを扱っていると様々なエラーが混入する. またデータの中に外れ値と呼ばれる極端に大きい (あるいは小さい) 値があると平均値や標準偏差がこの外れ値の影響を大きく受けることもある (エラーが外れ値となっていることも多い). このようなエラーや外れ値を発見するにはグラフが便利である.

データの把握・隠れた性質の発見

Take Home Message

工夫されたグラフによる可視化が, 隠れた性質の発見の鍵である.

← **John Snow** は 1854 年, ロンドンの市街図上にコレラで亡くなった人々の家をプロット (黒丸) した.

この図からあるポンプ (×印) を中心に人が亡くなっていることが分かり, 汚染された水 が原因であることが分かった という (それまでは, テームズ川から立ち昇る「瘴気」が原因と考えられていた).

最近では地図上に散布図を描くことは一般的になったが当時は画期的だったでしょう. この 1 枚のグラフが医学の歴史を変えたのです.

「データ視覚化の人類史」 (マイケル・フレンドリー 著)

コミュニケーション1

阪東バス　通過予定時刻表　柏学園前

行先	柏駅東口行き	
時刻	月曜日～金曜日	土曜日・日曜・休日
5	49	
6	16 38 53	26 48
7	05 14 26 40 49	05 20 37 53
8	04 15 28 41 57	11 26 45
9	09 21 39 58	05 28 51
10	16 40 59	08 27 49
11	17 35 59	08 26 51
12	15 34 50	08 23 43
13	11 32 50	03 22 43
14	11 25 45 58	01 21 37 58
15	24 41	21 38
16	00 18 40 55	00 19 39 58
17	16 31 51	16 38 59
18	04 25 40 59	28 45
19	17 36 56	11 34 48
20	08 30	08 44
21	03 59	16 48
22	31	
備考	土曜日が休日の場合には休日での運行となりますご注意下さい。	

※道路事情等により遅れる場合がございますのでご了承下さい。

阪東バス　℡04-7185-2771　検索サイト　阪東バス　検索　平成３０年８月２０日実施

上図は (どこにでもある) バスの時刻表である.

これは大変工夫されたデータの表現である. というのも,

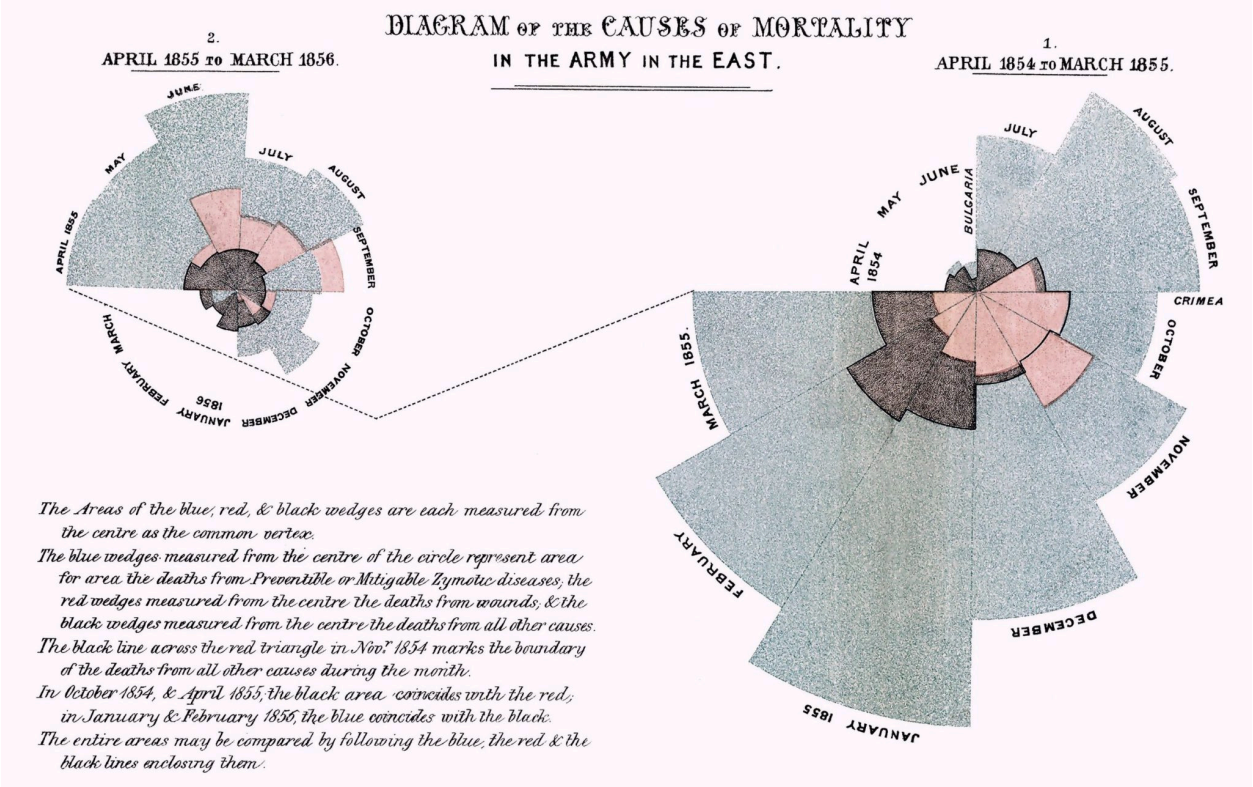
- 各時間帯ごとのバスの到着頻度が分かる
- 正確な時刻が分かる

というヒストグラムと表の良さを兼ね備えているからだ (考案者は不明だが, 素晴らしい発明ではないだろうか?).

問

グラフを工夫することで, うまく情報を伝えられることがある. よく工夫されたグラフやデータの表現手法を探してみよう.

コミュニケーション 2



ナイチンゲールが政府を説得するために作成した図 **(polar-area diagram** と呼ばれる**)**.

野戦病院における兵士の死亡原因は、戦闘 (赤) やその他の原因 (黒) よりも，「予防もしくは軽減可能な伝染病」(灰色) が多いことを示している (右図: 1984 年 4 月～1985 年 3 月).

このようなグラフを用いた説得が奏功し，野戦病院の衛生状態が改善され，「予防もしくは軽減可能な伝染病」(灰色) による死者が急速に減少した (左図: 1985 年 4 月～1986 年 3 月).

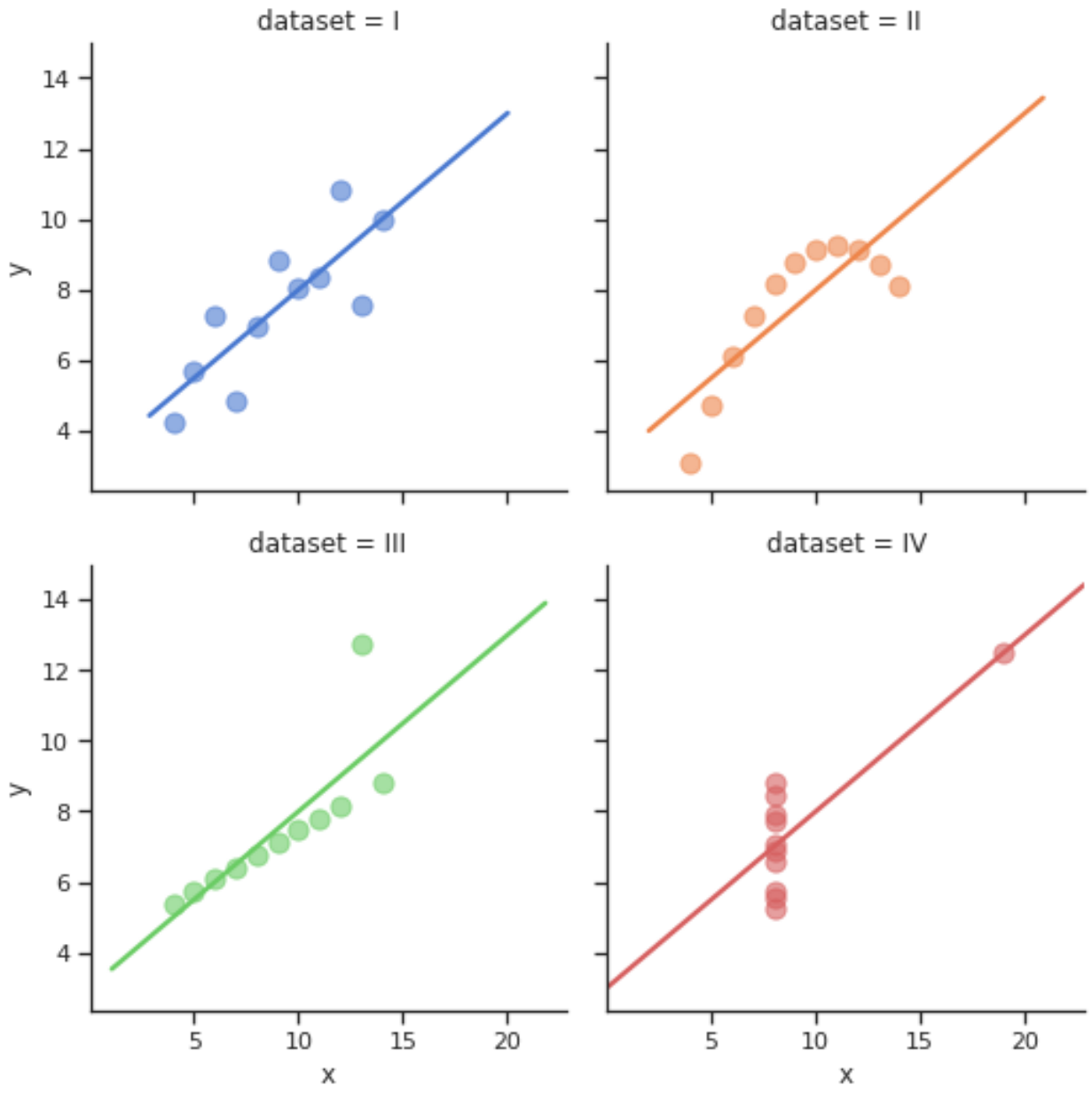
エラー・外れ値の検出

アンスコムの数値例: 外れ値を探せ!

データ I		データ II		データ III		データ IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.1	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.1	4.0	5.39	19.0	12.5
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
0.82		0.82		0.82		0.82	

平均値
標準偏差
相関係数

エラー・外れ値の検出 2



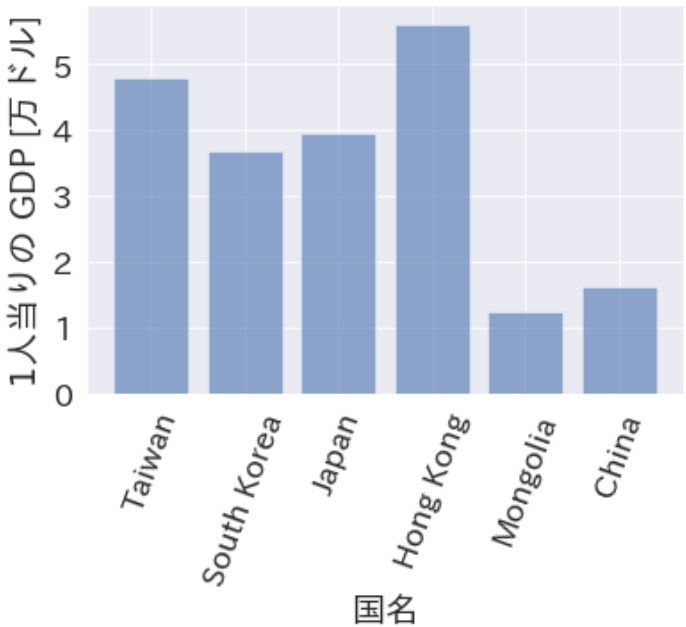
グラフにすることで外れ値がひと目で分かる.

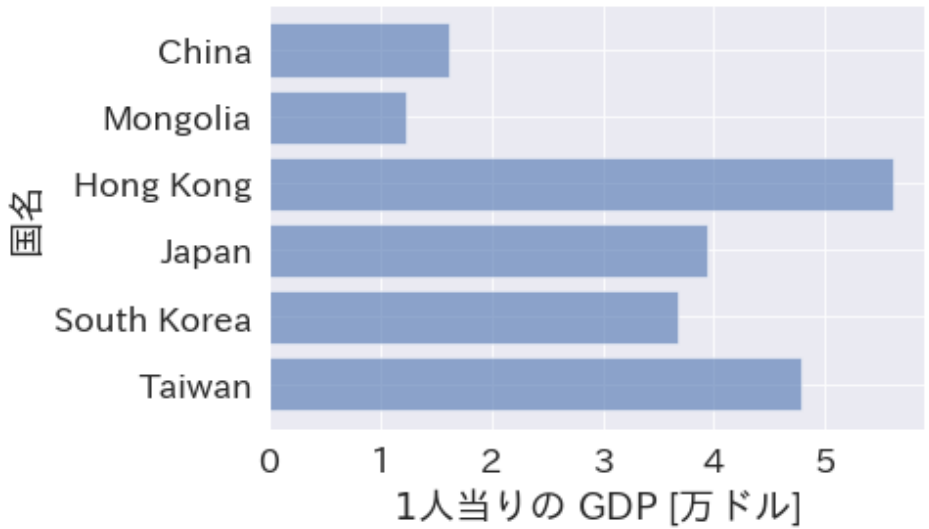
統計量が同じでも, x と y の関係性 (関数関係) が異なることがある. 関数関係は可視化すると分かりやすい.

棒グラフ

量の表現

量を「面積」を通して視覚化するのが棒グラフである.





よくある問題点は, 横軸のラベルのスペースが足りないこと. ラベルを斜めにしたりするグラフを見かけるが, 少し見にくい.

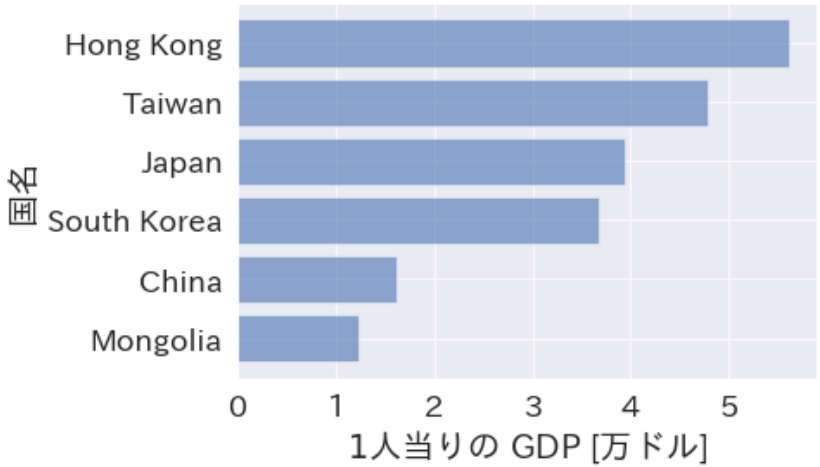
簡単な対処法は横向きにすること.

問

もっと改善できないか?

棒グラフ: 改良版

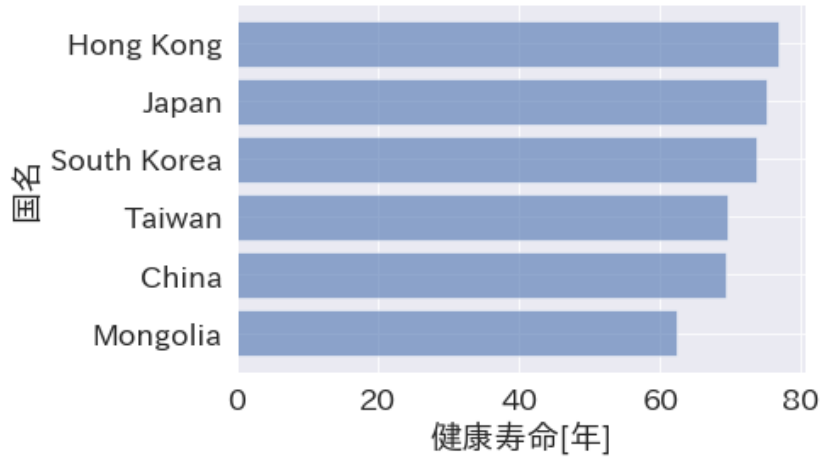
質的変数 (ここでは国名) に何らかの順序関係が無ければ, 量の大きい (小さい) 順に並べる方が良い:



ここでは, 次のデータセットを用いた

moodle 基礎情報・データ科学
「世界の国々の幸福度データセット」

同じデータセットから次のような棒グラフを作成した:

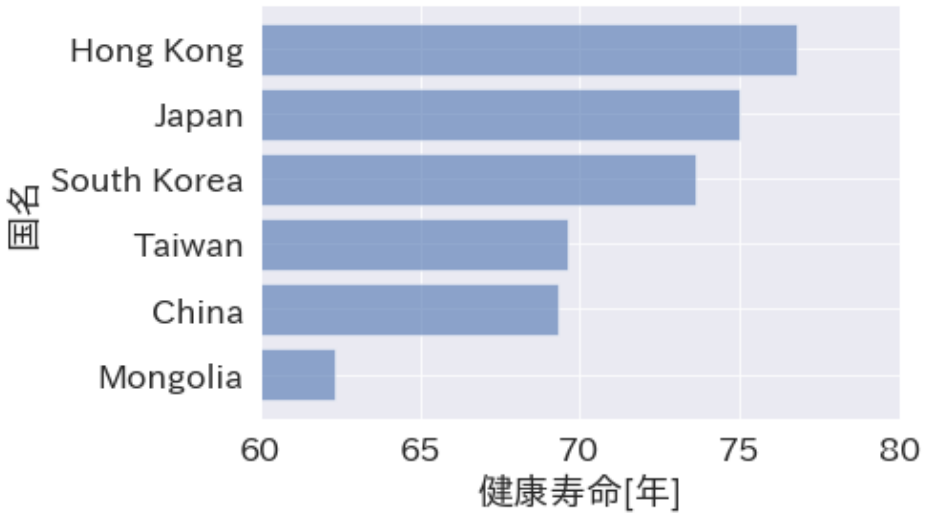


問

このデータの場合, 棒の面積が大きく, 国毎の違いが認識しにくい. 改善できないか?

良くない棒グラフ

下のような棒グラフはメディアなどでよく見かけるが良くない.

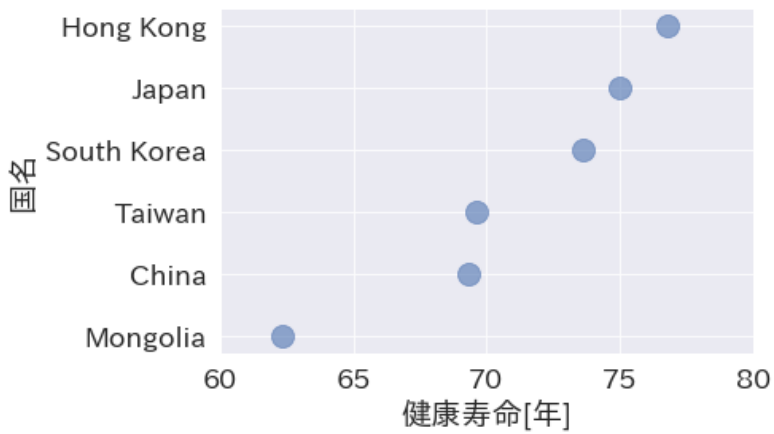


問

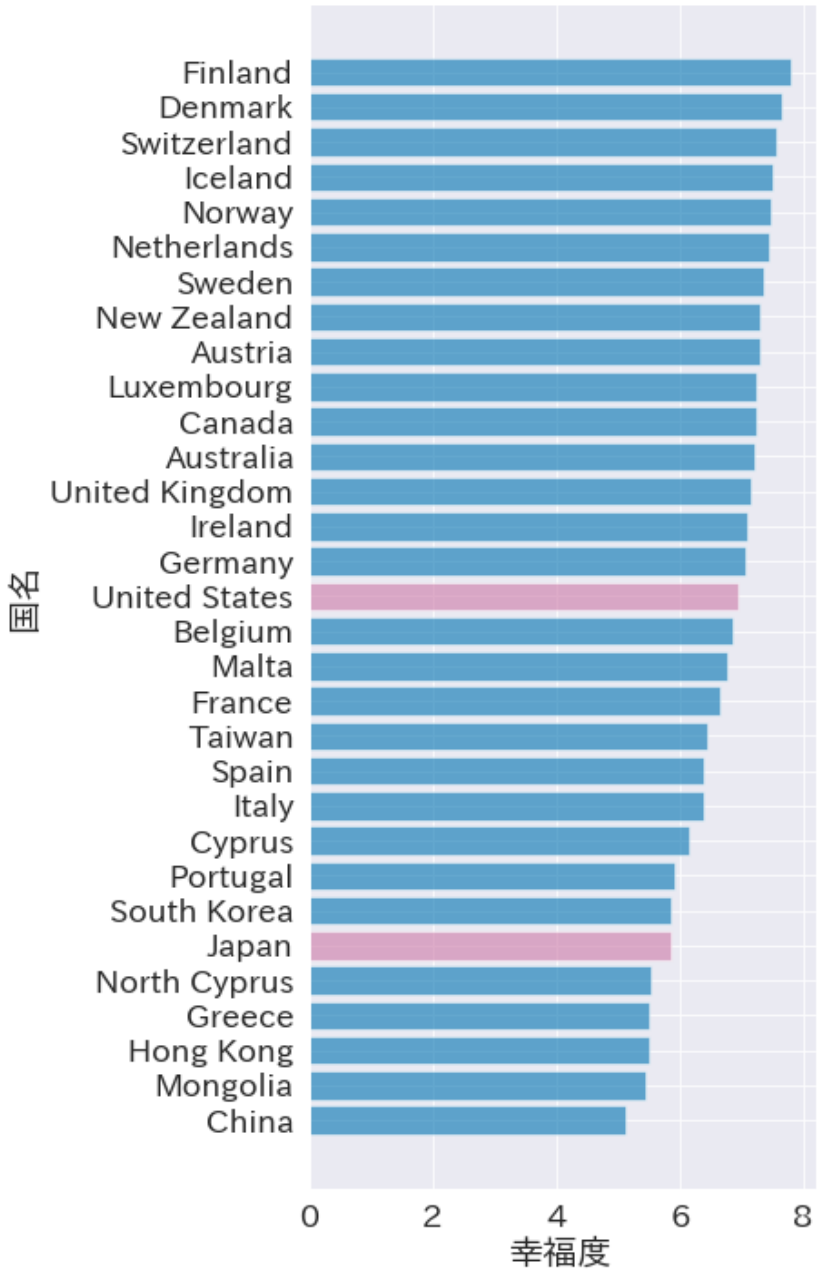
良くない理由を説明せよ.

棒グラフは「量」を「面積」で表現する手法なので, 両者は比例関係になくてはならない.

したがって, 棒グラフの端は 0 から始めるべきだ. 途中から始めたい場合は, ドットプロットにしよう:



強調



項目が多い場合は (伝えたいことに応じて) 色を変えるなどすることが有効.

Take Home Message

図を作成するときは，「伝えたいこと」は何かを考えよう.

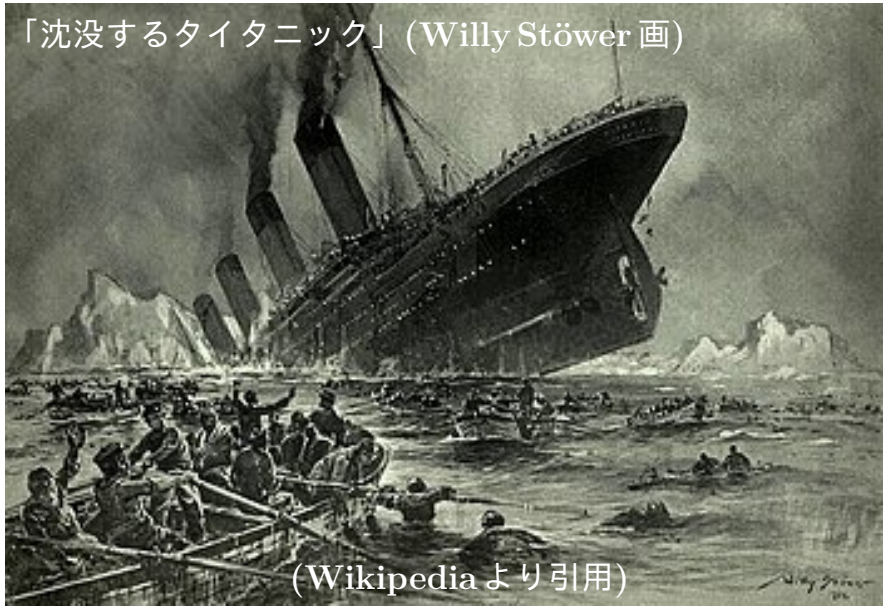
そして，それを伝えるためにはどのような工夫ができるのかをよく考えてみよう.

タイタニック号沈没事故

タイタニック号 (Wikipedia より)

タイタニック号沈没事故は，1912 年 4 月 14 日夜から 4 月 15 日朝にかけて，サウサンプトン (英国) からニューヨーク (米国) 行きの航海 4 日目に，北大西洋で起きた海難事故である．

当時世界最大の客船であったタイタニックは 2,224 人の乗客を乗せていた．4 月 14 日の 23 : 40 に氷山に衝突し，翌 4 月 15 日の 2 : 20 に沈没した．1,513 人が亡くなったこの事故は 1912 年当時，海難事故の最大死者数であった．



以下では，

moodle 基礎情報・データ科学

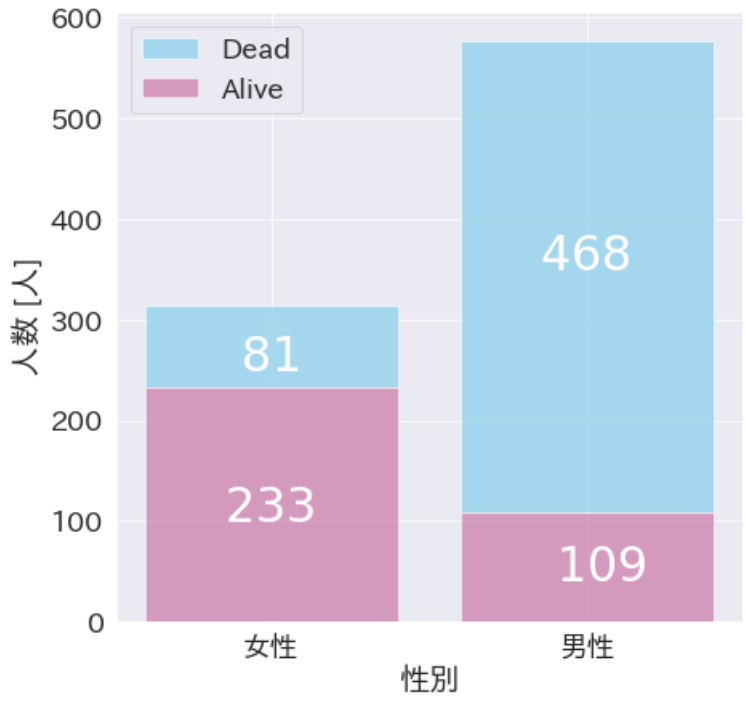
「タイタニック号沈没事故の乗客データセット」

を使用する．このデータには，乗客の年齢や性別，および死亡したか否かのデータが含まれている (ただし，欠損値が存在する)．

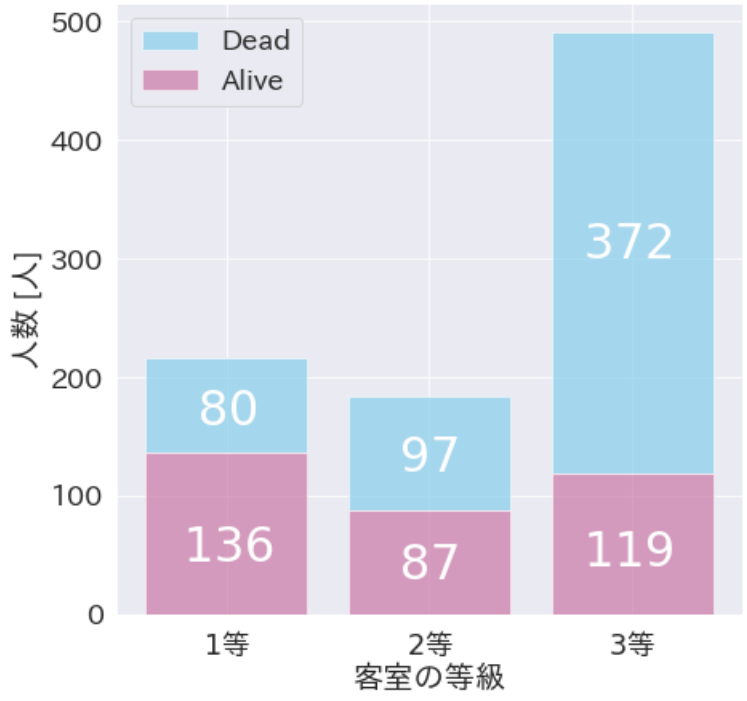
上のデータは **Kaggle** からダウンロードしたデータの変量名を日本語にしたもの．また，全データではなく，訓練データのみ使用した．

積み上げ棒グラフ

棒グラフを積み上げると，複数の指標で分類できる．これらは，下にあるクロス集計表 (小学校算数の内容) のグラフによる表現である．

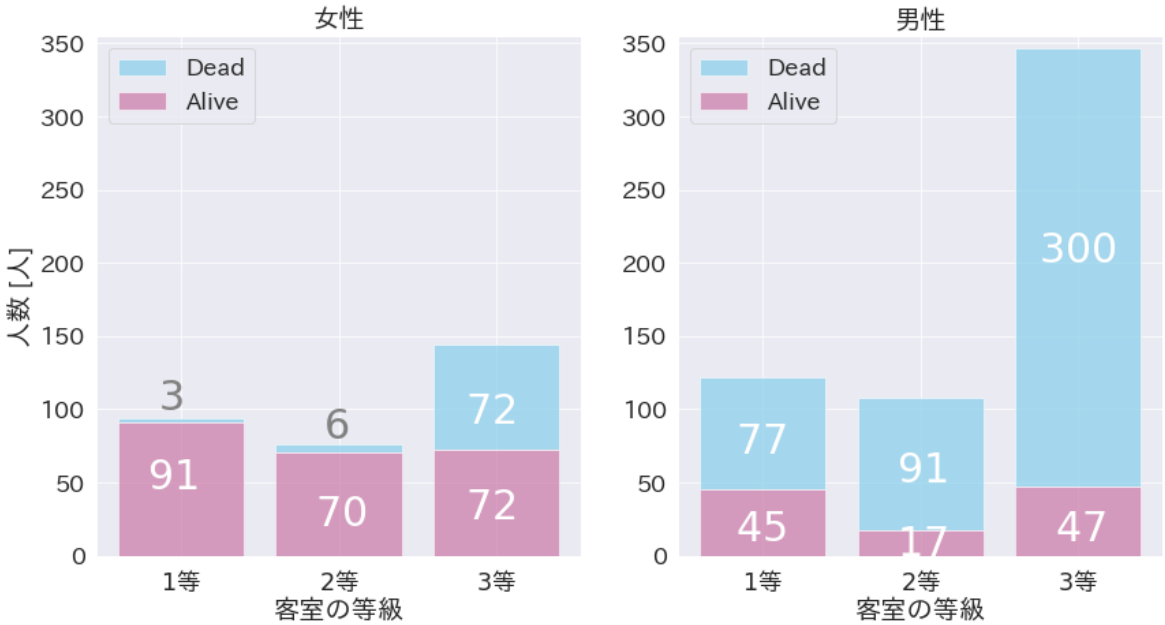


生死	性別	女性	男性
生存		233	109
死亡		81	468

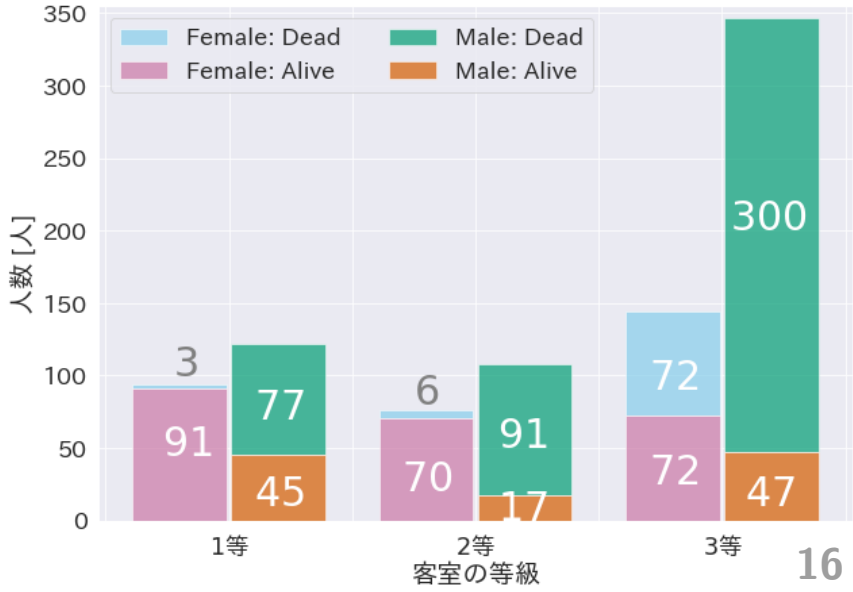


生死	等級	1等	2等	3等
生存		136	87	119
死亡		80	97	372

積み上げ棒グラフ



「客室の等級」と「性別」の両方を用いて分類した.

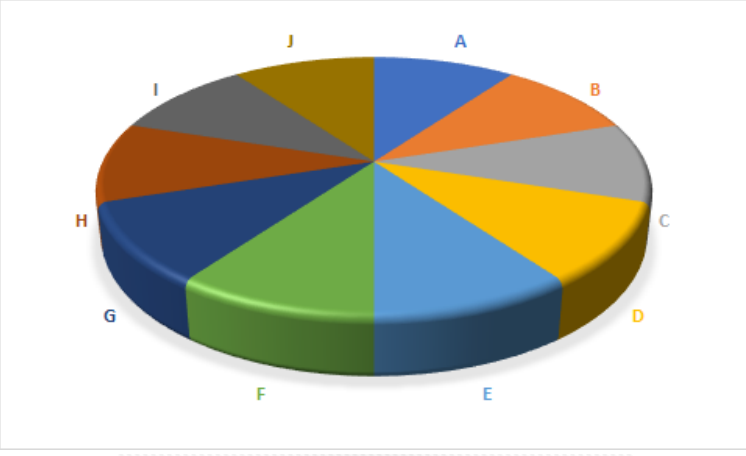


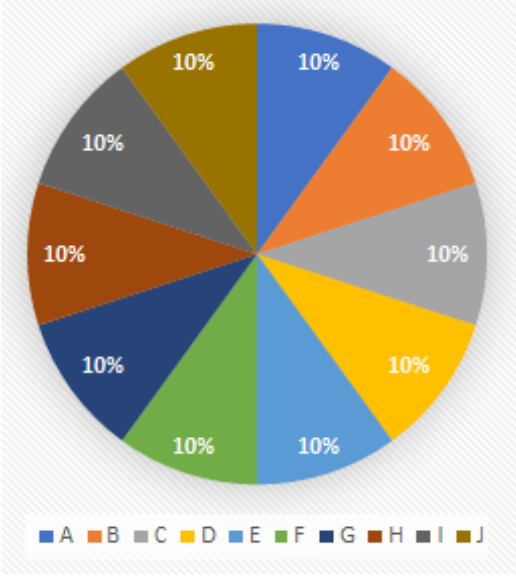
棒を横に並べることで1つにまとめることもできるが, 色が増えると認識しにくくなる.

円グラフ

割合の表現

割合を「面積」を通して視覚化するのが円グラフである.





問

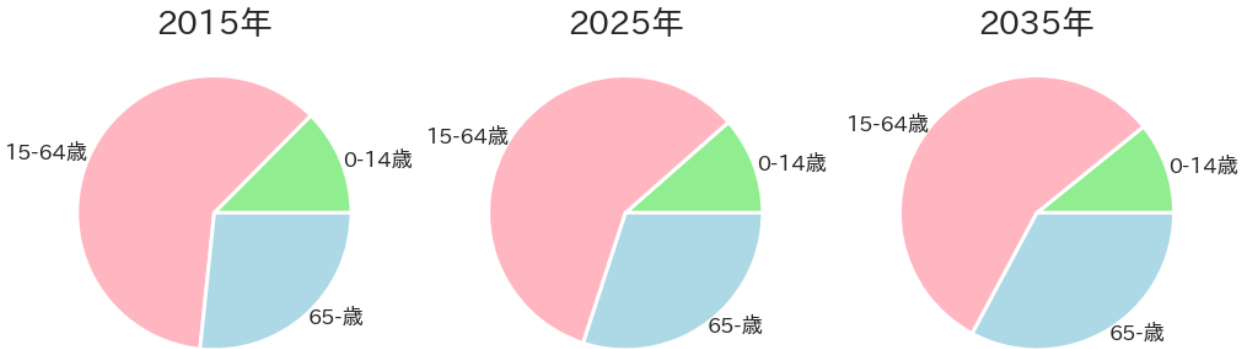
左の 2 つの円グラフは同じデータを可視化したものだが, 上の図は良くない. 良くない理由を説明せよ.

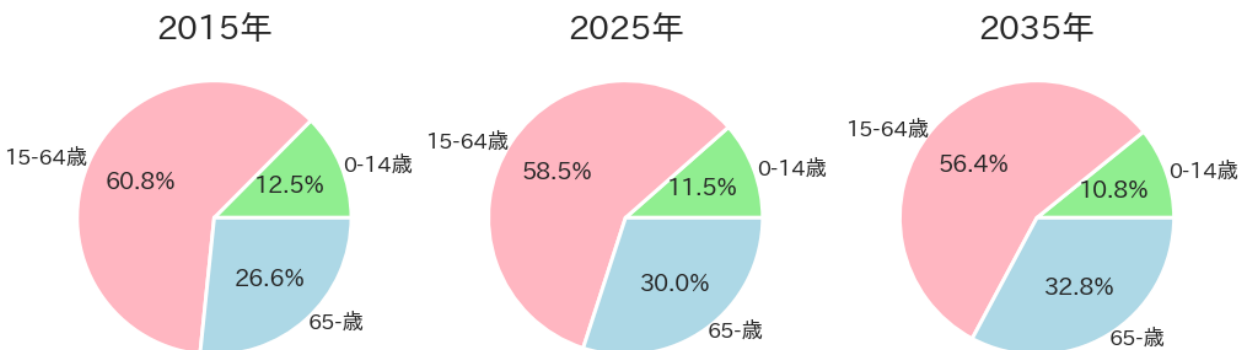
円グラフは「割合」を「面積」で表現する手法である. しかし, 3 次元にすることで, 近くの部分の割合が多く見えてしまう. そのため誤った印象を伝える危険性があり, 使用すべきではない.

一般に 2 次元で表示できるデータを 3 次元にする必要はない.

複数の円グラフ

下のグラフは徳島県内の高校 1 年生の授業で，生徒が作成したもの (を作り直したもの) である．日本の人口の構成比率を表現している．



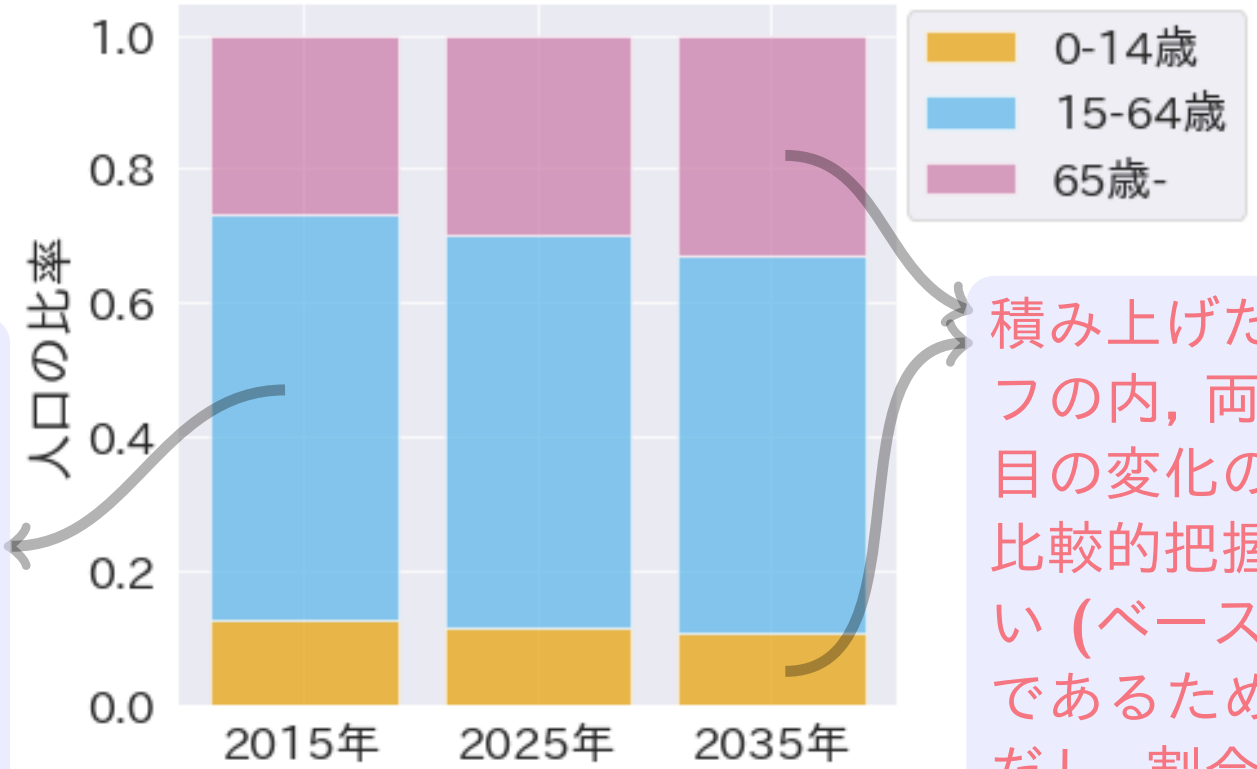


複数の円グラフを比較するのは難しい．数値を入れること自体は良いアイデアだが，図で視覚的に把握できる方が望ましい．

問

もっと良いグラフ表現を見つけよ．

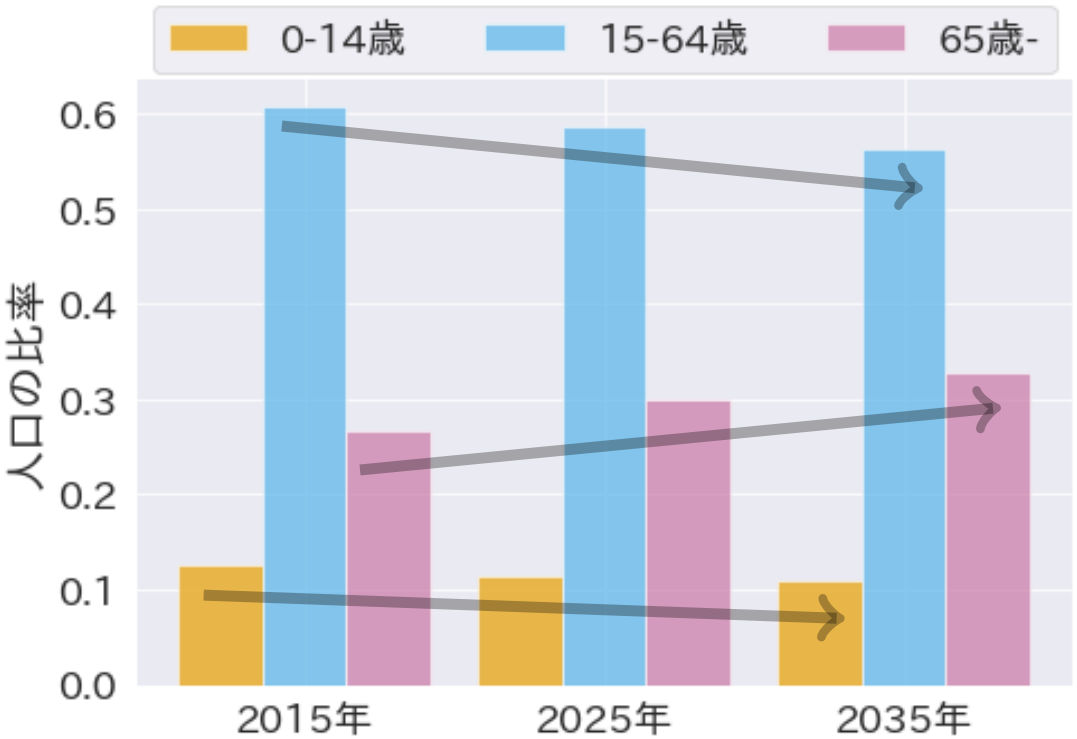
積み上げ棒グラフ



積み上げた棒グラフの内，内側の項目の変化の様子は一般に分かりにくい (ベースが一定ではないため).

積み上げた棒グラフの内，両端の項目の変化の様子は比較的把握しやすい (ベースが一定であるため). ただし，割合が小さい項目は把握しにくい.

複数の棒グラフ

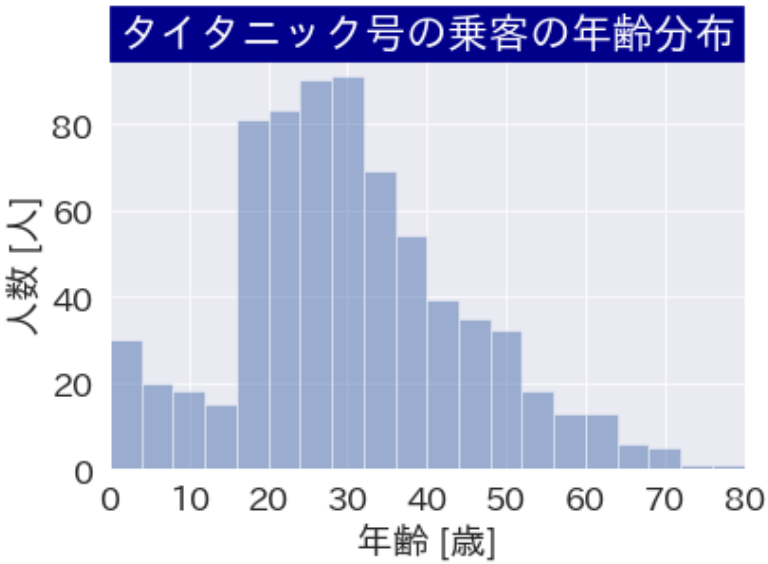


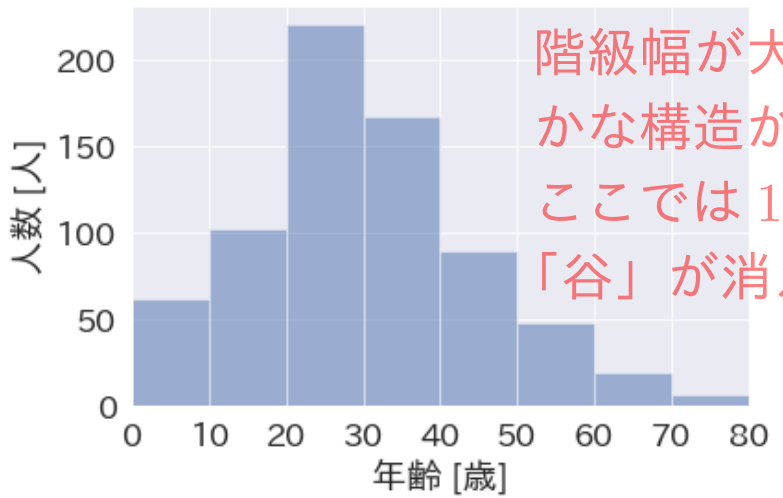
複数の棒グラフを用いると傾向が把握しやすい．このような棒グラフは小学校で学習する．

ヒストグラム (度数)

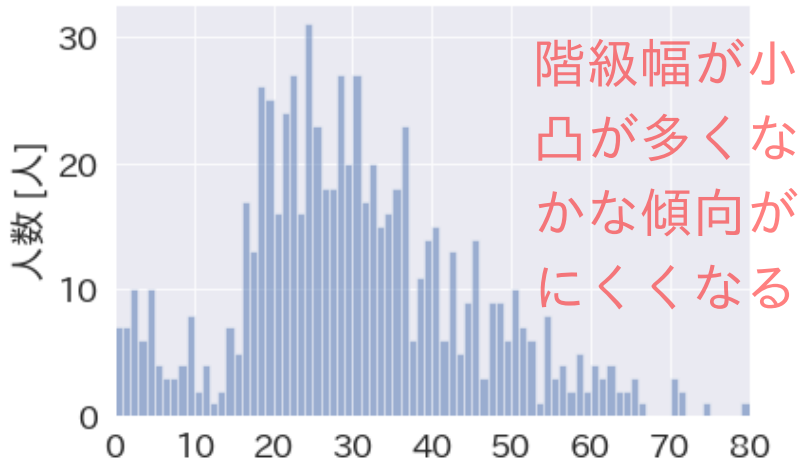
分布の表現

ヒストグラムは度数や密度などの「分布」を表現する方法．「長方形の高さ」で度数や密度を表し，「長方形の横幅」で階級幅を表す．





階級幅が大きいと細かな構造が見えない．ここでは 10 歳前後の「谷」が消えている．

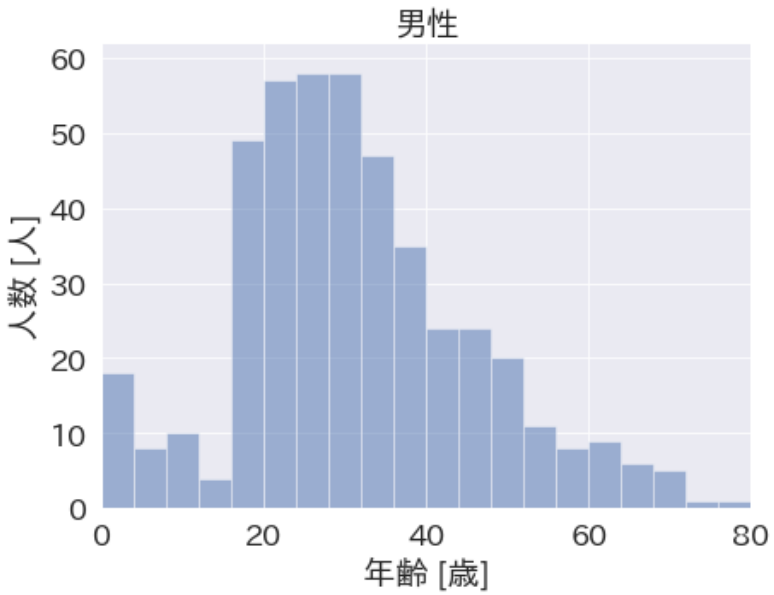


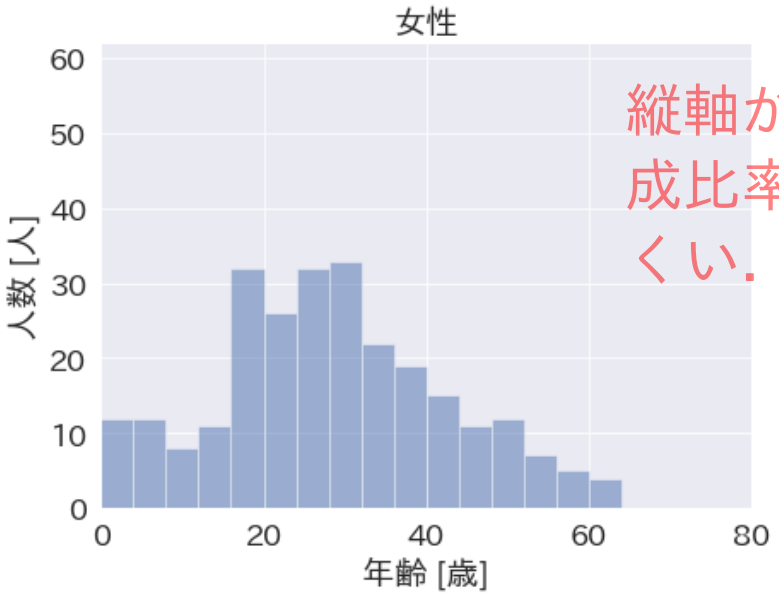
階級幅が小さいと凹凸が多くなり，大まかな傾向が読みとりにくくなる．

Take Home Message

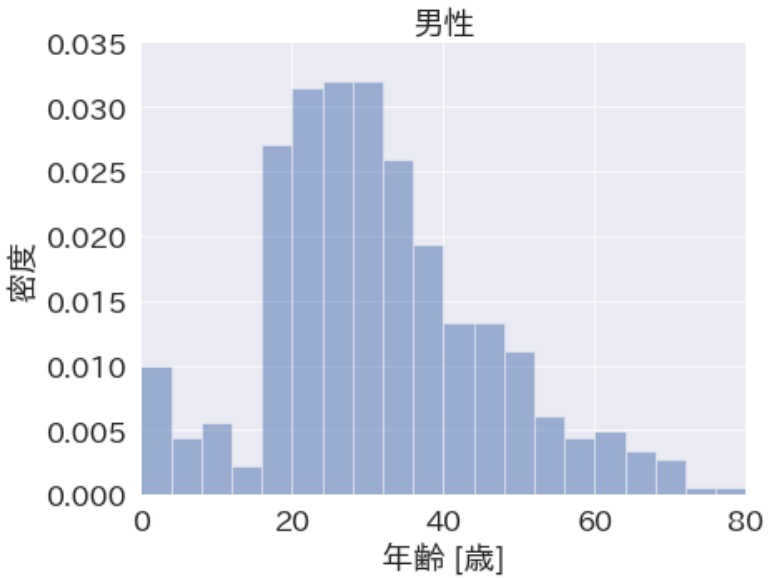
ヒストグラムを作成する際，階級幅を適切に設定することが重要
(いくつかの階級幅を試すこと)

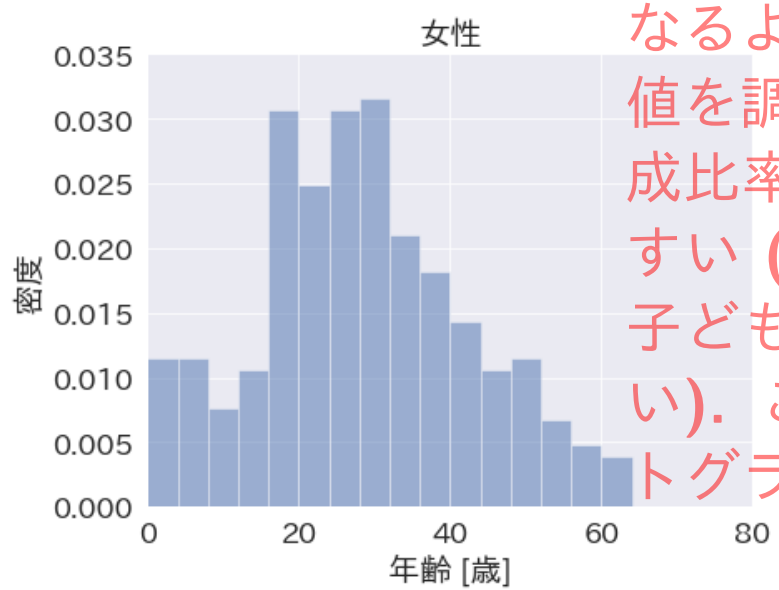
ヒストグラム (密度)





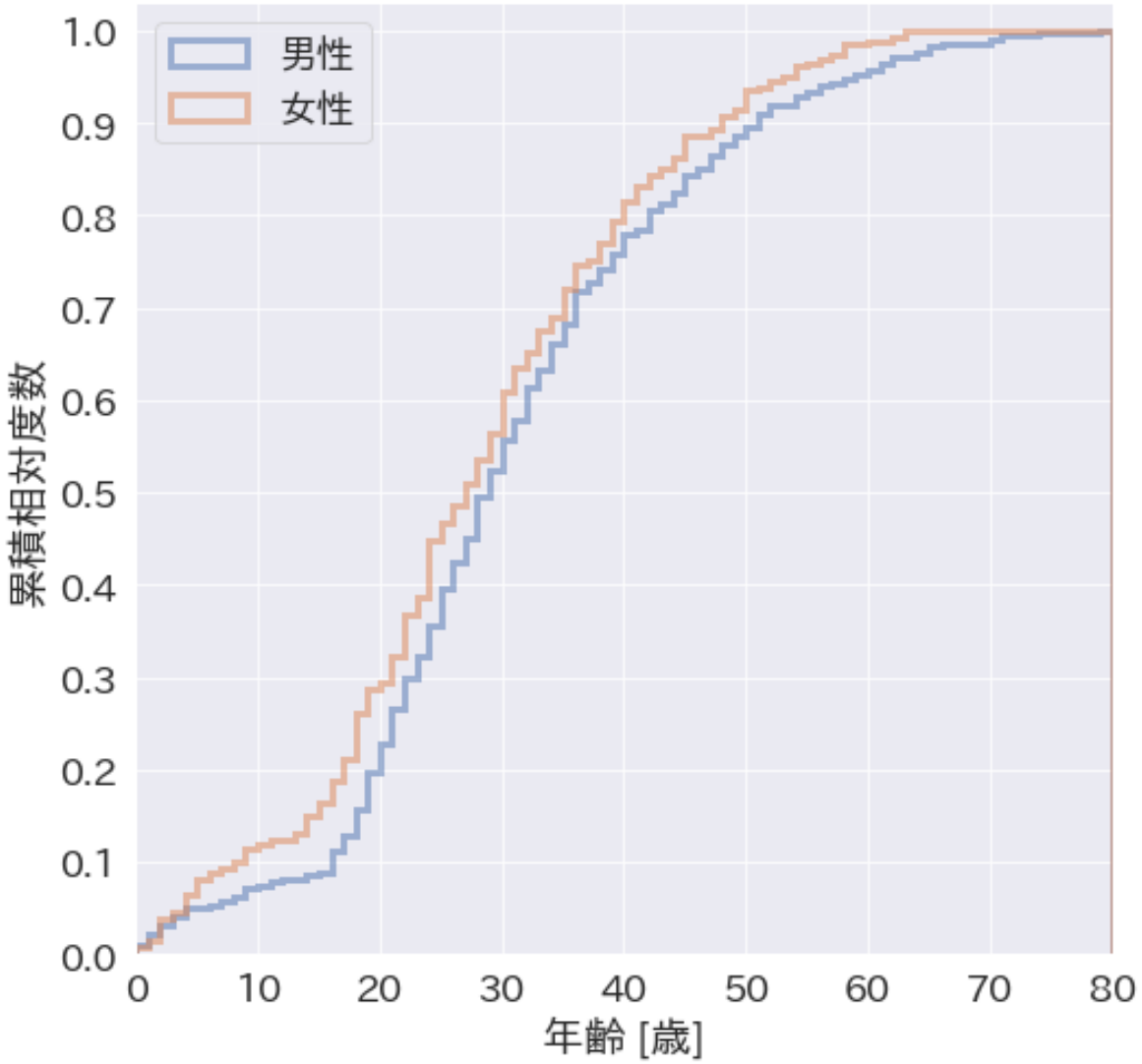
縦軸が度数だと，構成比率が比較しにくい．





青色の面積が 1 になるように，縦軸の値を調整すると，構成比率が比較しやすい (女性の方が子どもの比率が高い)．この図もヒストグラムと呼ぶ．

累積相対度数



累積相対度数も分布の表現方法であるが，ヒストグラムに比べて次の利点がある．

① ヒストグラムと違い，階級幅の設定は必要ない．

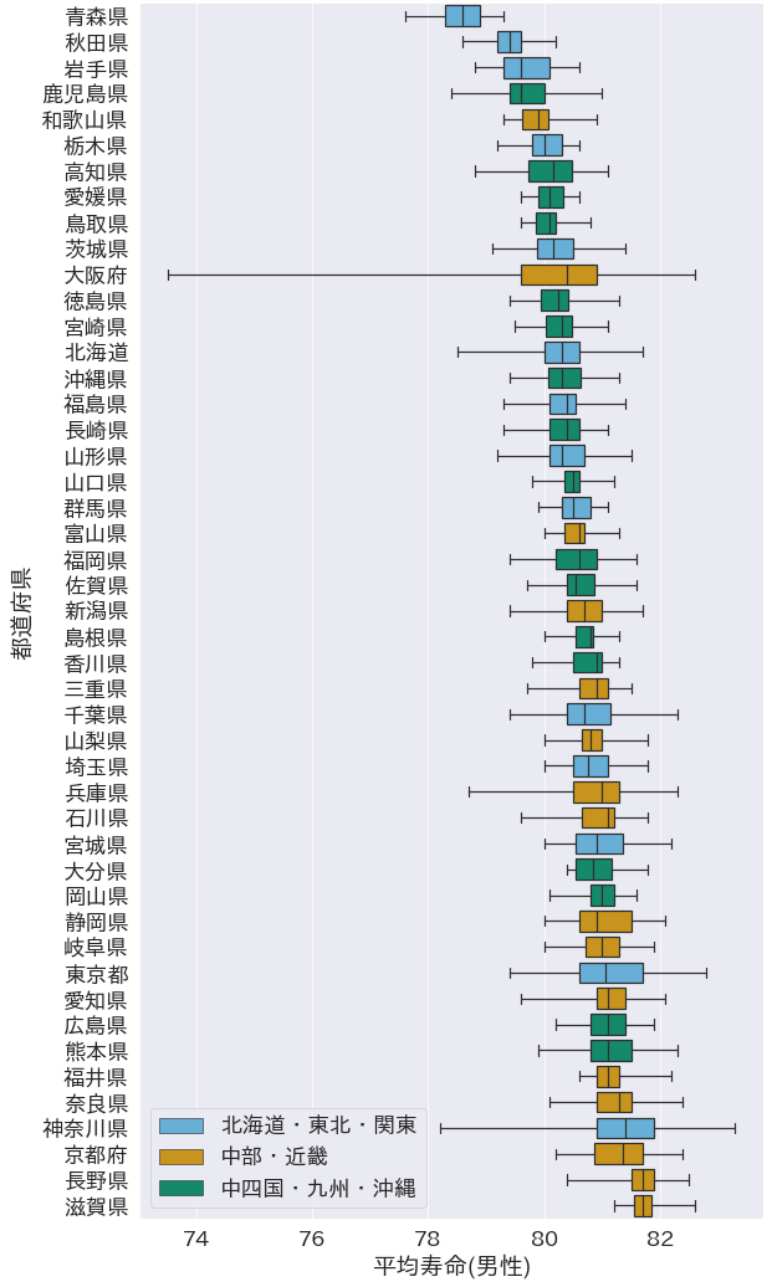
② 割合が簡単に読み取れる．例えば，20 歳以下の人の割合は

男性は 約 20%

女性は 約 30%

であることが分かる．これはヒストグラムからは読み取りにくい．

多数の分布の表現: 箱ひげ図 1



多数の分布の表現

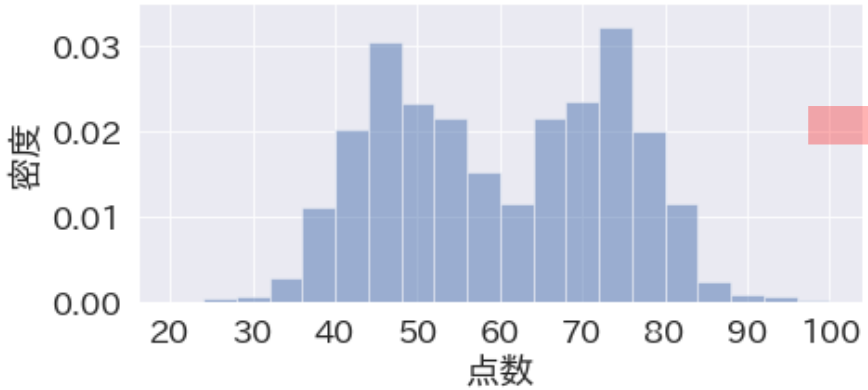
箱ひげ図を用いると多数の分布を同時に表現できる.

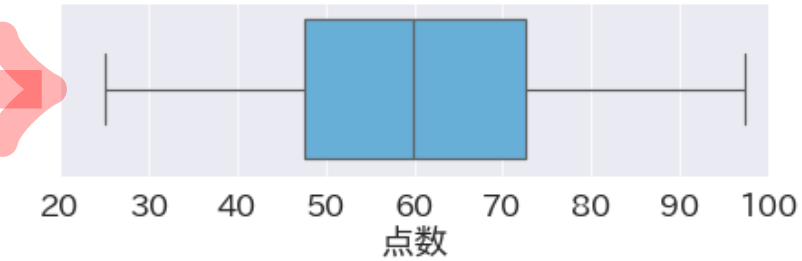
左の図は 2020 年のセンター試験に出題された箱ひげ図である (さらに, 地域毎に色分けした). ちなみに, 中学校指導要領 (H29.7) には

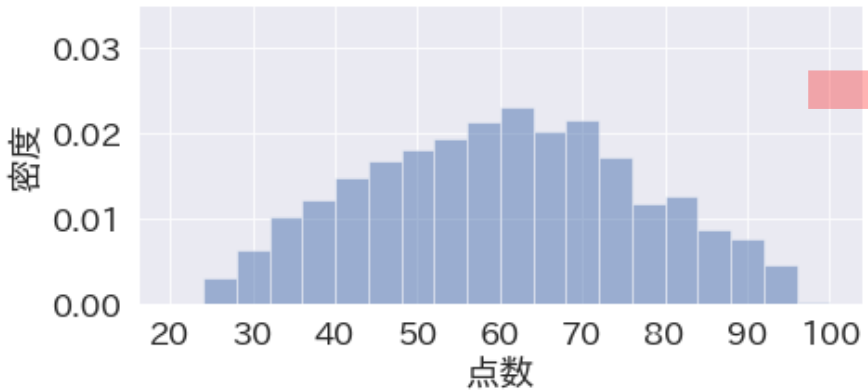
> コンピュータなどの情報手段を
> 用いるなどしてデータを整理し
> 箱ひげ図で表すこと.

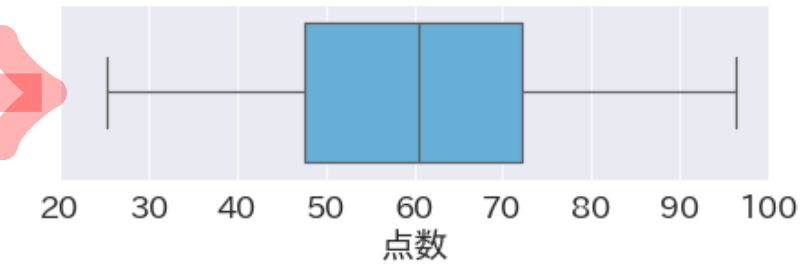
とある (統計量の意味を理解することの重要性はよく強調されるが, このような可視化技術は軽視される傾向がある).

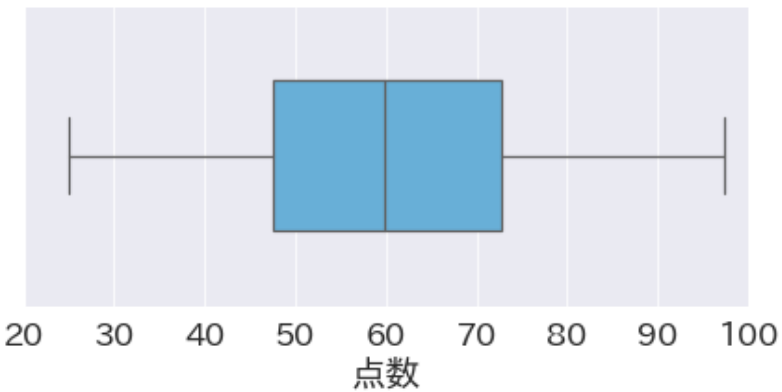
多数の分布の表現: 箱ひげ図 2









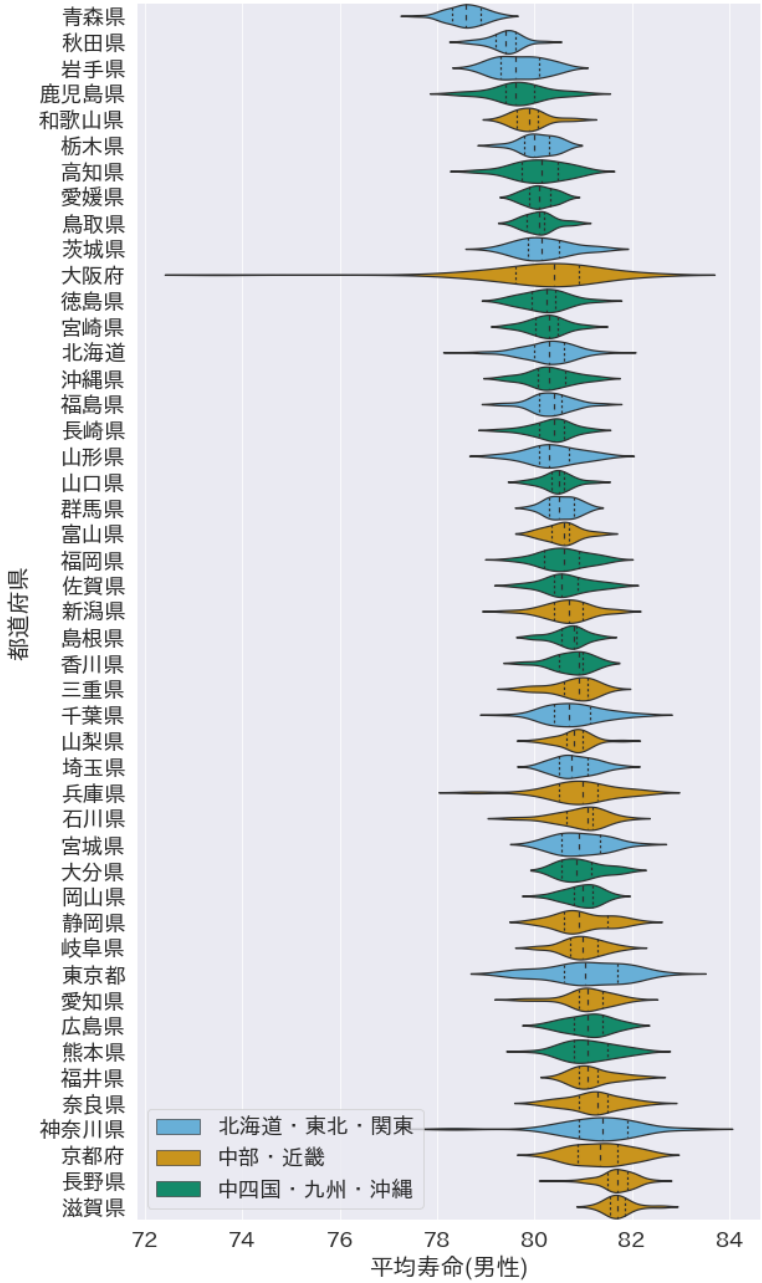


箱ひげ図ではヒストグラムの多峰性を発見できない

問

下の箱ひげ図に対応するヒストグラムはどちらか?

多数の分布の表現: バイオリン図



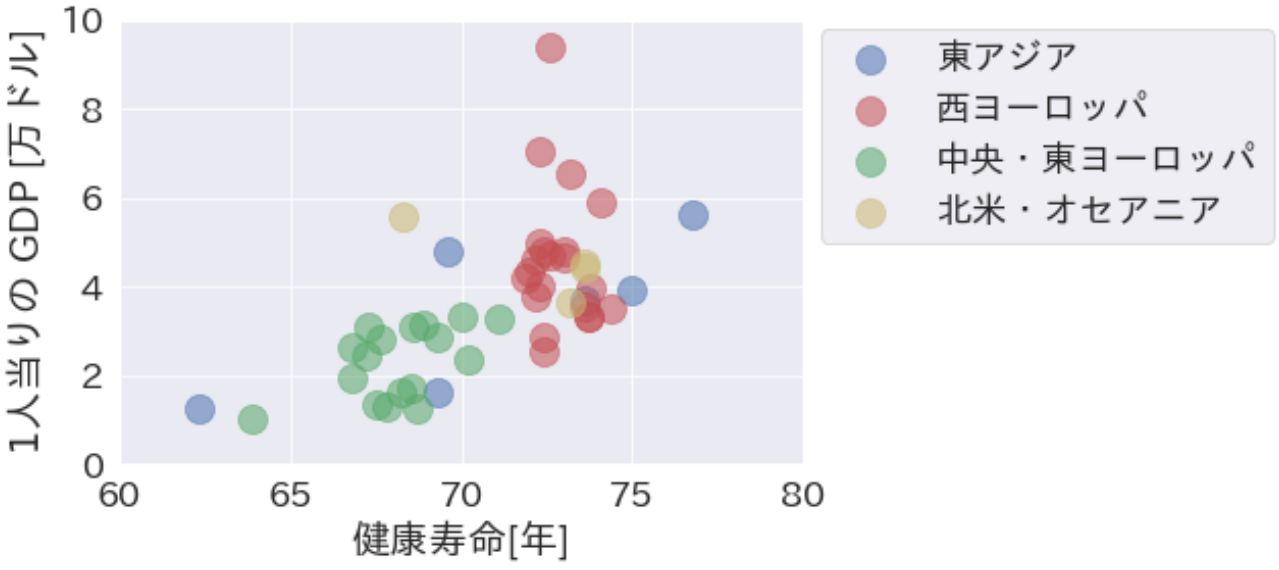
箱ひげ図は 1970 年代に **John Tukey** によって導入された．手書きで簡単に作成できるため一般的に用いられるようになった (当時は手書きするしかなかった)．

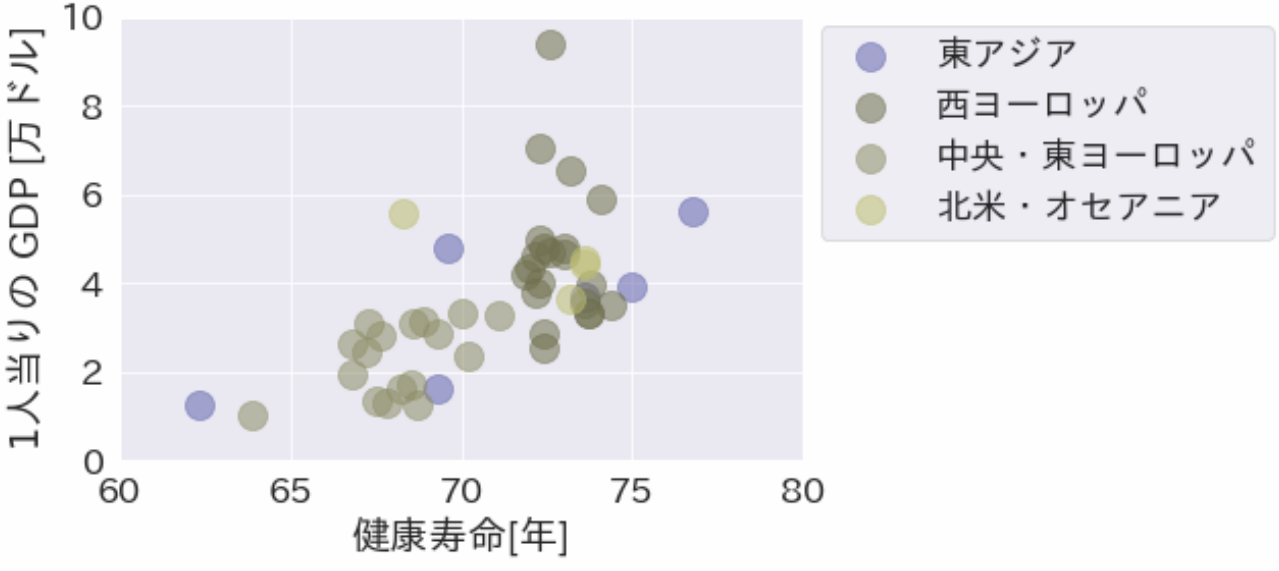
現在は様々なソフトウェアが開発され，次第に (バイオリン図, リッジラインプロットなどの) 別の表現手法に取って代わられつつある．これらの手法を用いると，多数のヒストグラムを同時に可視化しつつ，多峰性も発見である．

Take Home Message

可視化技術は次第に発展している．

散布図 1





2 変量の関係性の表現

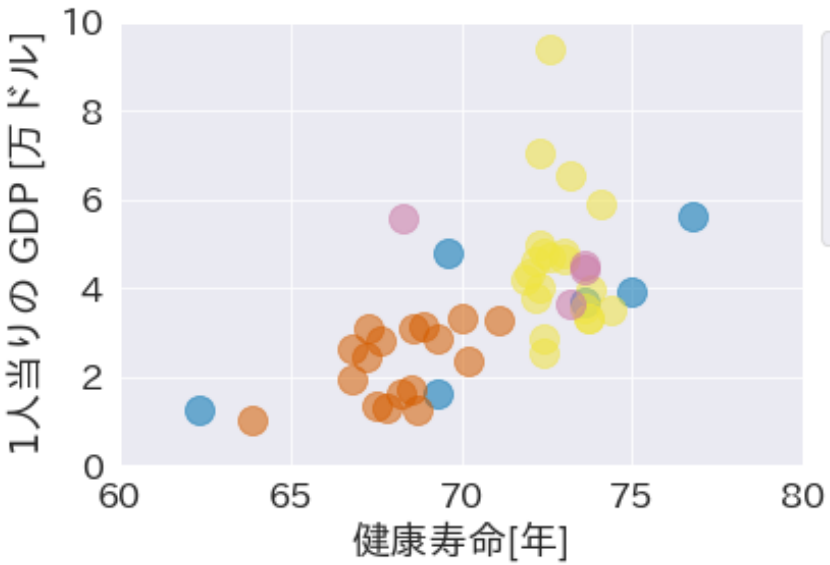
散布図は 2 つの変量の関係性の表現である

問

このグラフの問題点を指摘せよ.

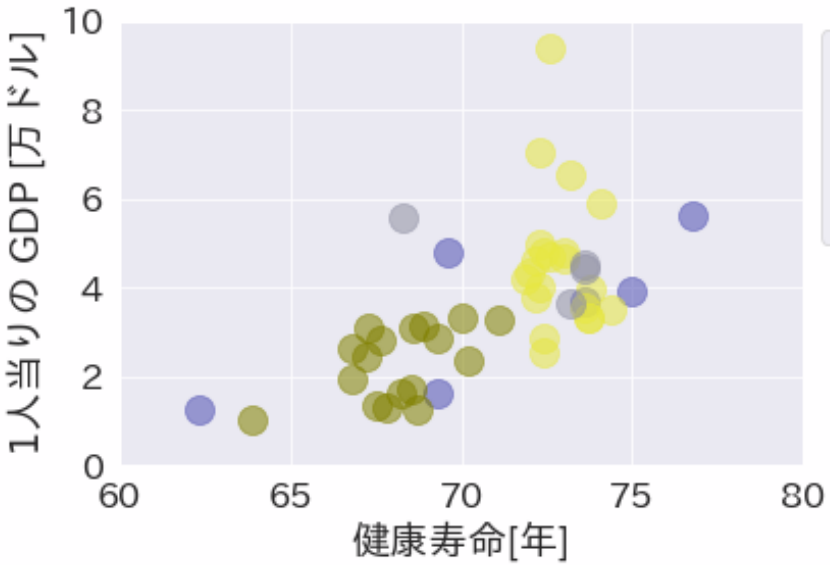
上の散布図を先天赤緑色覚異常の人が見ると下図のように見える. 色覚異常があると赤と緑の違いが認識しにくい (先天赤緑色覚異常. 他のタイプの色覚異常もある).

散布図 2



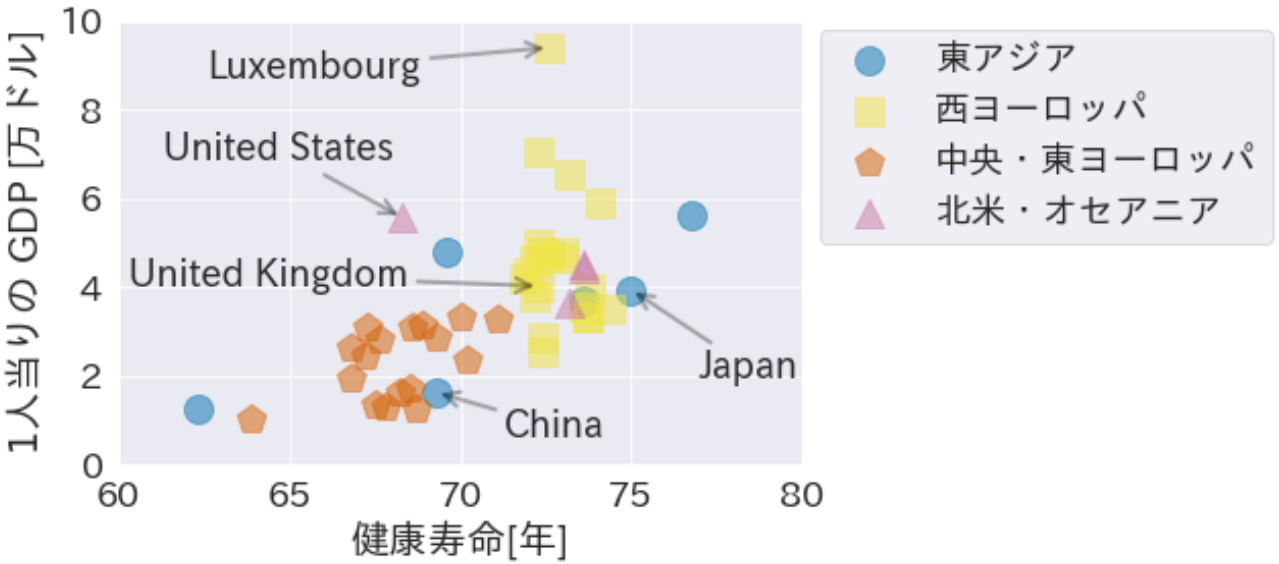
色覚に異常がある人にも見やすい配色が考案されている:

https://wp.nyu.edu/siegal/color-palette/

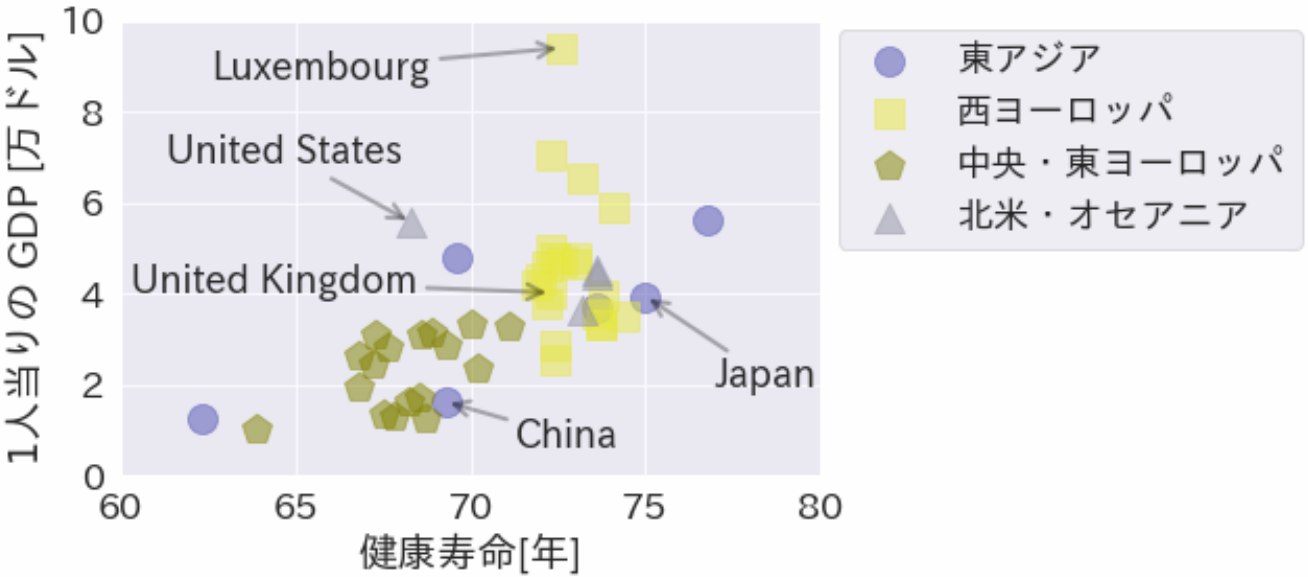


この配色を用いると，多少見分けやすくなる．この講義のスライドの図はできるだけこの配色を使って作成している．

散布図 3



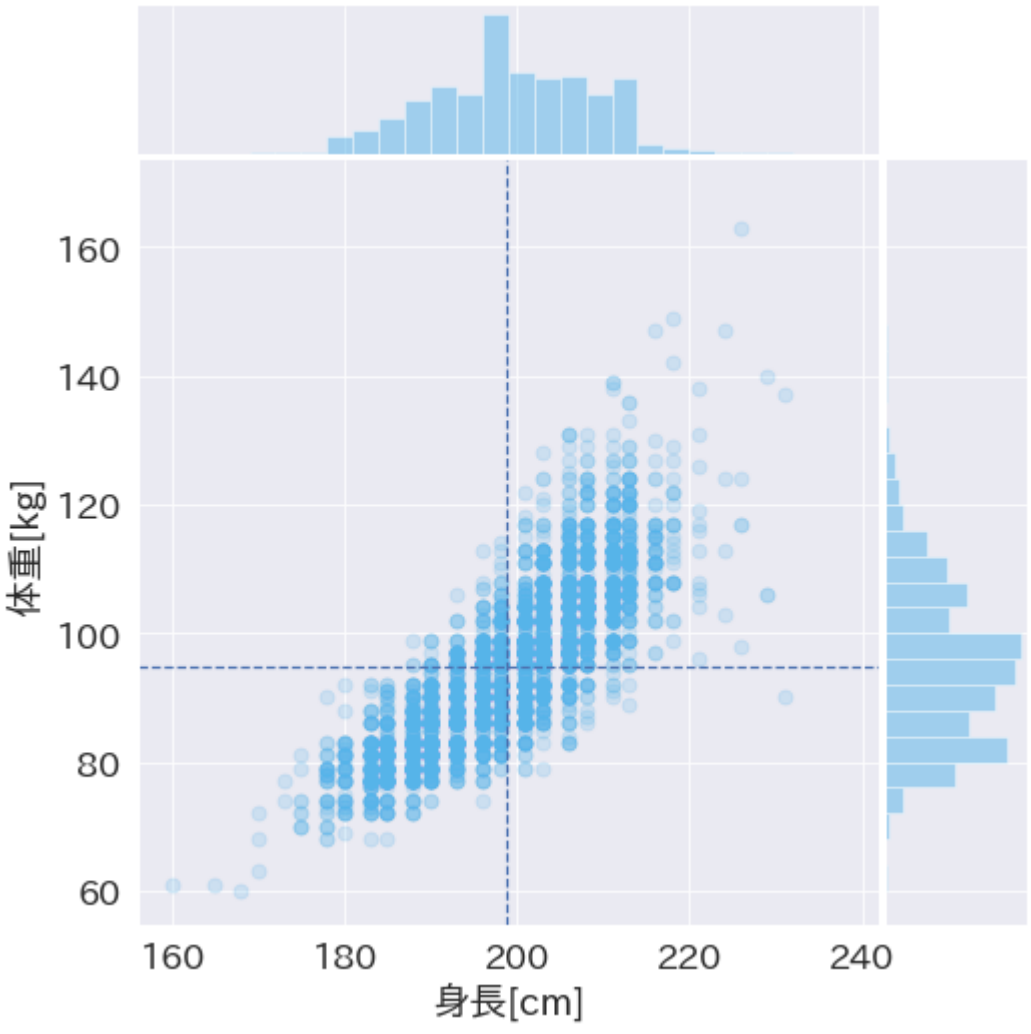
マーカーのタイプを変えるとより認識しやすくなる.



また必要に応じて, データ点のラベルを付加しても良い (必要無ければ, 無い方が見やすい).

散布図の工夫

NBA 選手の身長と体重のデータ



散布図の工夫

- マーカーに透過性を持たせたり，
- 散布図の周辺にヒストグラムを配置したり，
- 散布図中に平均値 (ここでは，平均身長と平均体重) を示す

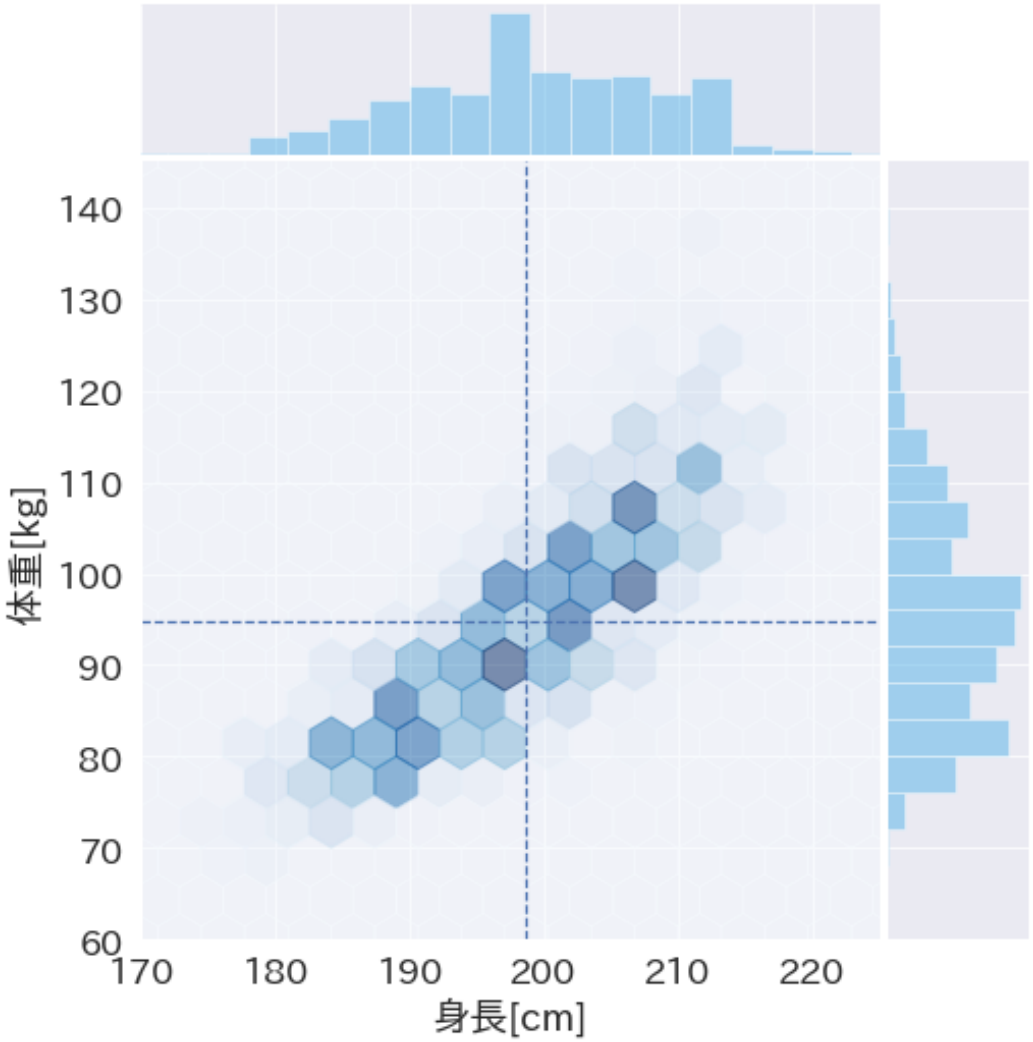
ことで，情報がより伝えやすくなる．

問

点の数が多いと，マーカーが重なり見にくい．より良い表現は無いか?

散布図の仲間 1: 2 次元ヒストグラム

NBA 選手の身長と体重のデータ



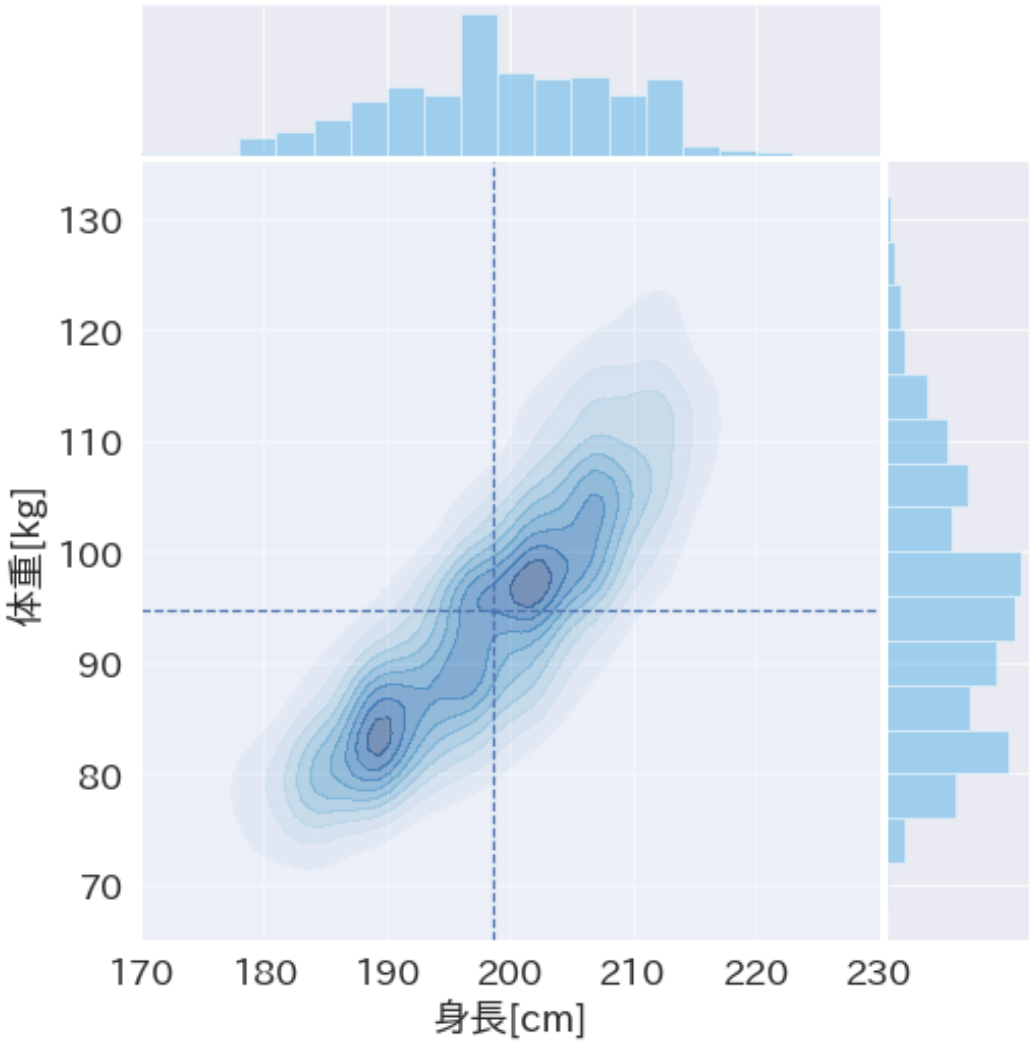
2 次元ヒストグラム

平面を「領域」に分けて，それぞれの領域に入るサンプルの個数を図示したもの．

「領域」は四角形か六角形にすることが多い．1 次元のヒストグラムの「階級幅」と同じく，「領域」のサイズにより変化する．

散布図の仲間 2: 等高線図

NBA 選手の身長と体重のデータ



等高線図 (密度関数の推定)

2 次元ヒストグラムから, (2 変数) 関数の推定 (カーネル密度推定) をして, その等高線をかいたもの.

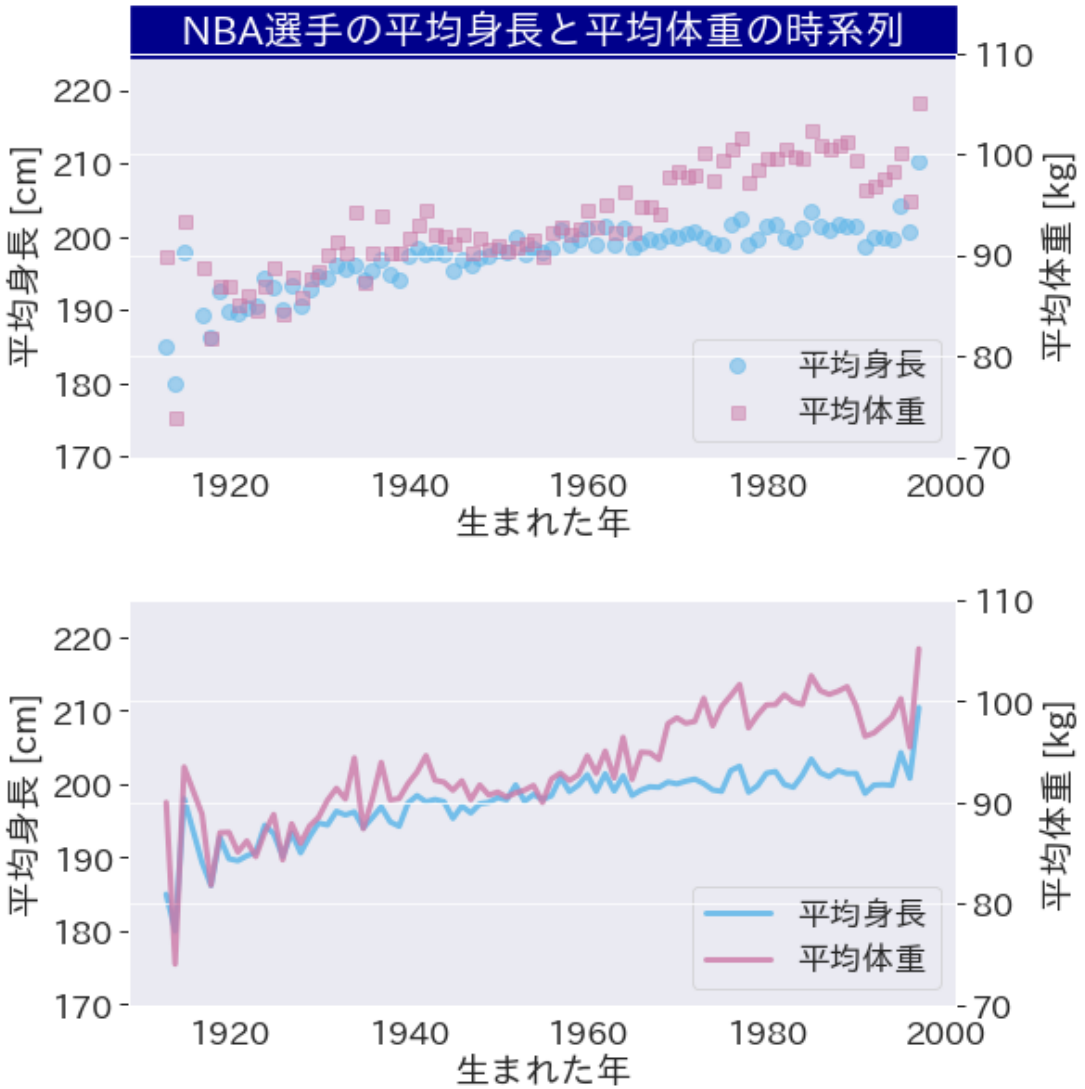
身長 190cm 前後と 200cm 前後に, 2 つの山 (ピーク) が見られる.

ラインプロット 1

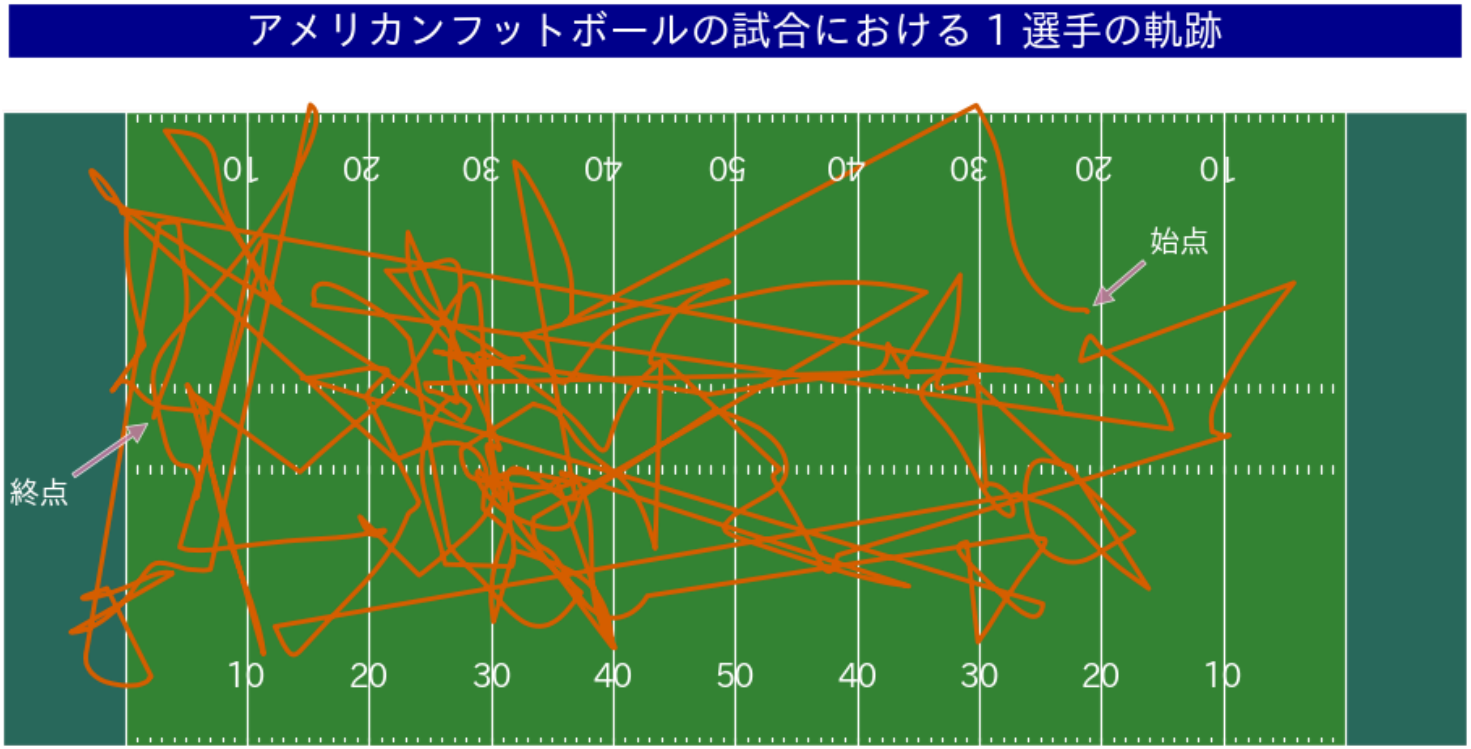
系列データの表現 1

変量の 1 つに，明かな順序関係があるとき (例えば時間)，データが表す点をつなぐ線を書くことで，データの傾向をより分かりやすく表現できる．とくに順序関係がある変量が「時間」の場合，時系列 と呼ぶ．

ただし，データの無い部分に線を引くことになるので，特にデータが少ない場合には注意が必要 (誤解を与える危険がある)．



ラインプロット 2



系列データの表現 2

座標軸が順序関係を持つ変量 (時間など) を表していないこともある.

上の図は, アメリカンフットボールの試合 (NFL) において, ある選手の移動の様子をラインプロットで示したものである. 縦軸と横軸はグラウンド上の位置を表しており, 時間ではない.

最後に

iPS 細胞の発見者である山中伸弥先生は、予想と違う実験データが得られるととても楽しくなって，そのことに心を奪われたそうです **(**普通はがっかりするものです**).**

相対性理論などで有名な理論物理学者のアルバート・アインシュタインは，混沌とした実験データの中から現象の本質を見抜く力が抜群に優れていたそうです **(**普通はわけが分からんといって，すぐに投げ出すのです**).**

Take Home Message

データ分析に関する教育において大事なのは，データや現象に向かう心構え を育てることにあります．

もちろん，統計やプログラミングなどの「技術的」なことも重要ですが，これらを教え学ぶことはそれほど難しくはないのではないかと思います．上のような心構えを習得することはとても難しと思いますが，とても価値のあることです．

基礎情報教育: データ科学入門
第 5 章 データ分析実習

T. MIYAGUCHI

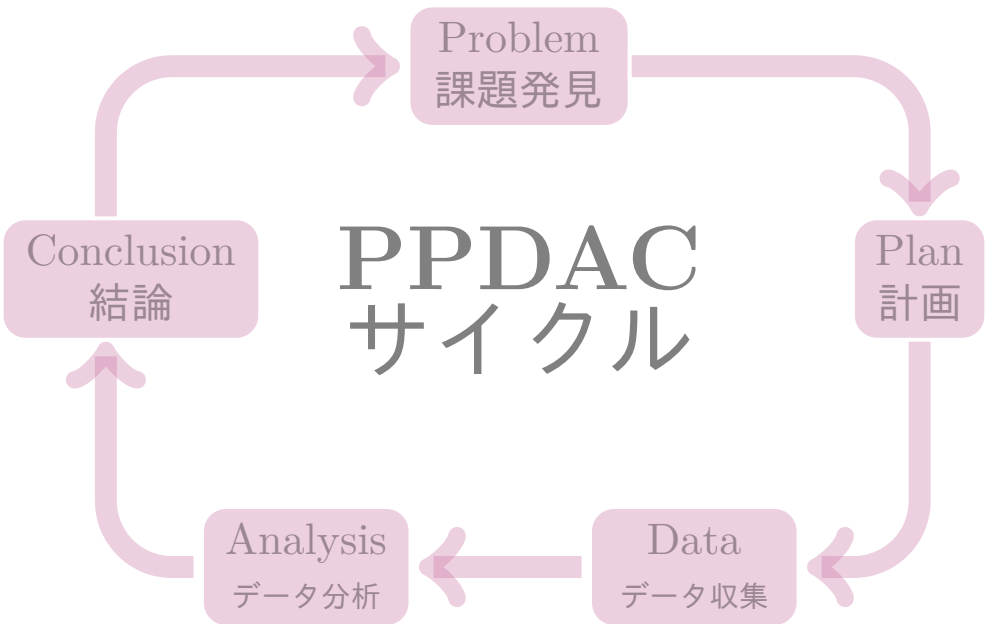
Naruto Universality of Education

講義全体のアウトライン

- 第 1 章: 社会と教育における変化
  - 仮説駆動とデータ駆動
  - データに基づく思考や判断
- 第 2 章: データ科学とは?
  - 統計学・計算機科学とデータ科学
  - AI・機械学習・倫理
- 第 3 章: データ分析の基礎
  - データとは?
  - 代表値・散らばりの指標・関係性の指標
- 第 4 章: 可視化
  - 可視化の必要性
  - 量の表現・割合の表現・分布の表現・関係性の表現・系列の表現
- 第 5 章: データ分析実習
  - 問を立てよう・統計量と可視化
  - 機械学習に挑戦・データ分析の実践
- 第 6 章: 様々な話題

PPDAC サイクル (復習)

データ分析活動の一般的な流れは**PPDAC サイクル**で表される (下図の時計周りのサイクル).



Take Home Message

データ分析活動は課題発見 (問を立てること) から始まります. 最初に問を立てずにデータ分析を始めると明確な結論が得られず, 活動が中々終わらない, という事態に陥りがちです.

日本では, 問は先生が発するもの (発問) であり, 学生や生徒は質問することもまれです. しかし近年, 課題発見能力の育成は重視されるようになってきています.

オープンデータ

前回の講義 (木曜) で「オープンデータ」という言葉が出てきましたが, 質問はありませんでした. イメージしやすい単語なので, おそらく「スルー」した人が多いのではないでしょうか?

活動: 次の単語について問を立てよう

オープンデータ

ヒント:

- なぜ...?
- どのように...?
- いつ...?
- 誰が...?
- ... は何?

- オープンデータの定義は何?
- オープンデータには本当に何の制約もないの?
- オープンデータはいつ頃から広まったの?
- 誰がオープンデータを公開しているの?
- どのようなオープンデータがあるの?
- オープンデータの意義は何?

オープンデータの定義

オープンデータの定義

国、地方公共団体及び事業者が保有する官民データのうち、国民誰もがインターネット等を通じて容易に利用 (加工, 編集, 再配布等) できるよう、次のいずれの項目にも該当する形で公開されたデータをオープンデータと定義する。

- 無償で利用できるもの
- 機械判読に適したもの
- 営利目的、非営利目的を問わず二次利用可能なルールが適用されたもの

総務省のサイトより引用

データは著作物として扱われることがあります. その場合, 二次利用する際のルール が定められていることが多いです.

良く使われるルールは, クリエイティブ・コモンズ・ライセンス (CC ライセンス) と呼ばれるものです. (CC ライセンス) の中には 6 種類のライセンスがありますが, オープンデータに良く適用されるのは CC-表示 と CC-表示-継承というライセンスです.

クリエイティブ・コモンズ・ライセンス

CC-表示

二次利用の際に作成者の指示する情報 (クレジット) を「表示」することが条件となっています.

CC-表示-継承

(1) クレジットの表示と,
(2) 公開する場合, 同じライセンスで公開すること, の 2 つが条件です.

CC ライセンスの他の 4 つのライセンスについても調べてみましょう.

活動 5 分

e-STAT のライセンスを調べてみよう.

利用規約 によると,「CC-表示」であることが分かります. 日本の政府や自治体が公開しているオープンデータのほとんどは「CC-表示」のライセンスで公開されているそうです.

(補足) 著作権の縛りを受けないデータや作品をパブリックドメインと呼び, CC0 と書かれることがあります.

Google Ngram Viewer

活動 15 分

Google Ngram Viewer を用いて，いくつかの言葉の使用頻度を比較してください．

その結果からどのようなことが予想できますか?

ある程度考えがまとまったら，隣の人や先生と情報交換してください．

ヒント：完全に考えがまとまってなくても情報交換にチャレンジしよう! 人に説明すると考えがまとまることがあります．

- 日本語は非対応なので，英語を調べよう．
- コンマ (,) で単語 (複合語) を区切ることで，複数の単語 (複合語) が同時に調べられます．

教育用標準データセット

活動 25 分

授業のページの

moodle 基礎情報・データ科学

「教育用標準データセット (市区町村別データ)」

(2022 年版) を用いてデータ分析をするとしよう．まず，①どのようなデータか確認した上で，② どのような問が立てられるか考えよ (最低 3 つの問を考えよ)．

上のデータセットは 独立行政法人 統計センター からダウンロードした．これらのデータの作成方法は 「**e-Stat** から上記のファイルを取得する方法」 というリンクから **pdf** をダウンロードできる．

手順

① 15 分 できるだけ多くの問を考え出そう．

② 5 分 考えた問を 1 つずつ吟味し，優先順位 (自分の興味が高い順) をつけよう．周りの人や先生と意見交換しても構いません．

③ 5 分 優先順位の高い 3 つの問を **moodle** から報告ください．

教育用標準データセット: 問の例

「市区町村別データ」に関する問の例

- 同じ市区町村名はどれくらいある?
- 市・区・町・村, それぞれの数はどれくらい?
- 先頭の数字について, 1〜9 の出現頻度は?
- 地方の方が高齢化が進んでいる?
- 高齢化の要因は?

3 つの問を moodle から入力して報告してください (出席を兼ねます).

みなさんの回答は次回の授業で「データ」として使用する可能性があります. ご了承ください.

オリンピック出場選手のデータセット

活動 20 分

授業のページの

moodle 基礎情報・データ科学

「オリンピック出場選手のデータセット」

には，1896 年のアテネから 2016 年のリオまで 120 年間の出場選手についてのデータセットである．まず，①どのようなデータか確認した上で，② どのような問が立てられるか考えよ (最低 3 つの問を考えよ)．

上のデータセットは https://www.kaggle.com/ からダウンロードした (ライセンスは **CC0** ← どういう意味でしたか?)．表中の **'NA'** は **not available** の意味で，「データが無い」ことを意味する．

手順

① 10 分 できるだけ多くの問を考え出そう．

② 5 分 考えた問を 1 つずつ吟味し，優先順位 (自分の興味が高い順) をつけよう．周りの人や先生と意見交換しても構いません．

③ 5 分 優先順位の高い 3 つの問を **moodle** から報告ください．

オリンピック出場選手のデータセット: 問の例

「オリンピック出場選手のデータ」に関する問の例

- 金メダリストの年齢分布は?
- 国ごとのメダルの獲得数は?
- 競技毎の **Body mass index (BMI)** は?

3 つの問を **moodle** から入力して報告してください.

データに基づく思考や判断

活動

普段「なんとなくそうかな」と思っていることで，データ分析したいテーマ (問) を 3 つ挙げよ．

授業で紹介した問の例として

(1) 気温は上昇しているか?

(2) 降水量は増加しているか?

(3) 「都会より地方の方が高齢化が進んでいる」は本当か?

(4) 「都会より地方の方が高齢化が進んでいるのは，地方は出生率が低いからである」は本当か?

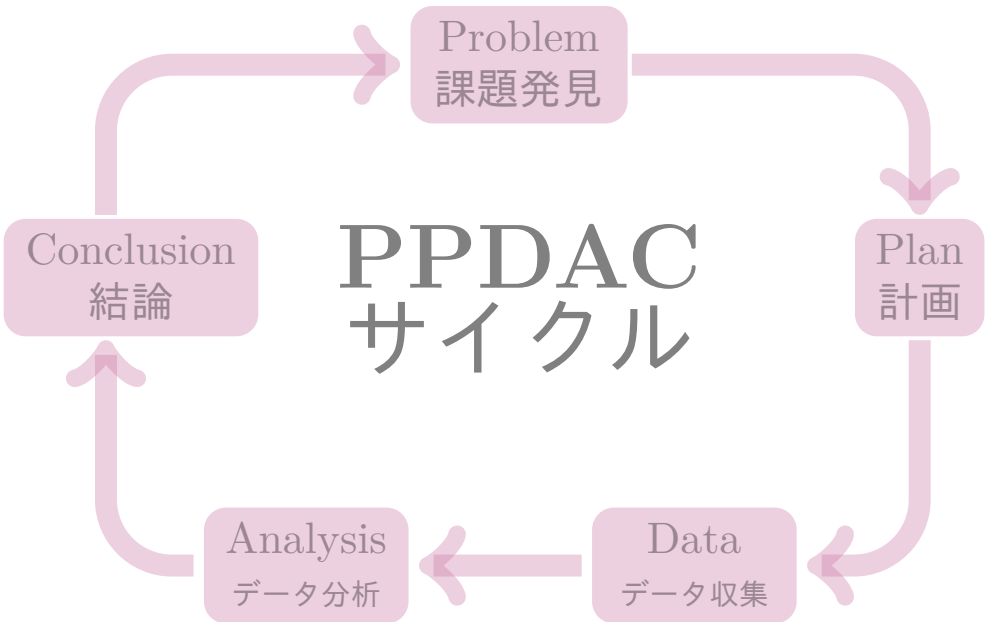
(5) 「日本は多神教だから寛容だ」は本当か?

問 2

問 1 で挙げたテーマを検証するデータはどのようにして入手するか考えよ．

PPDAC サイクル (復習)

データ分析活動の一般的な流れは
PPDAC サイクルで表される
(下図の時計周りのサイクル).



Take Home Message

現在のデータの収集方法は極めて多様である. 地道な方法を試す前に, よりよい方法が無いか調べ, その方法をマスターする方が早道かもしれない.

目標

ここでは, **Web** からデータを収集する方法を体験しよう.

収集 1: e-Stat の利用

e-Stat から市区町村データをダウンロード (DL) する手順

① トップページ https://www.e-stat.go.jp/ で「地域」をクリック

② 「市区町村データ」を選択し，「データ表示」をクリック

③ 地域選択
- 1 絞り込み: 「表示データ」，「地域区分」，「絞り込み」を設定→「実行」をクリック
- 2 地域候補: 「全て選択」をクリック → 「確定」をクリック

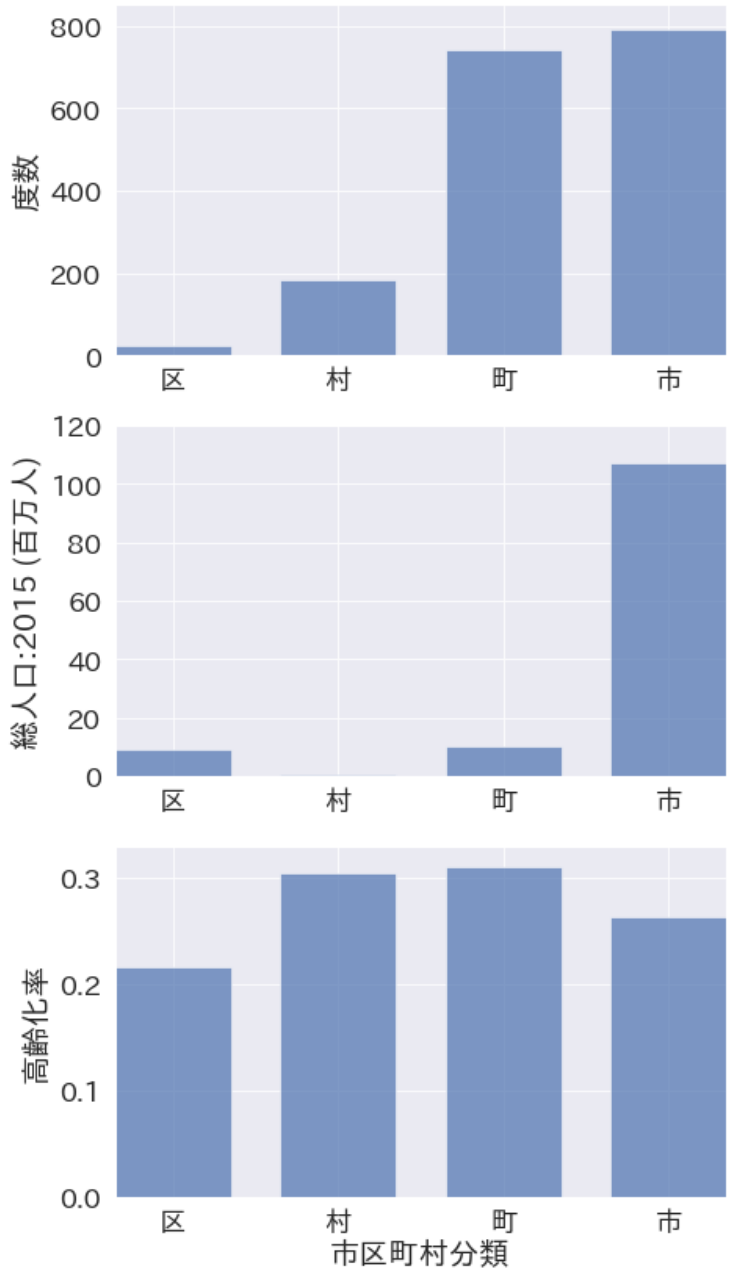
④ 表示項目選択
- 1 絞り込み: 必要なデータの分野等を選択し，「実行」をクリック
- 2 項目候補: 項目候補から必要な項目を選び，「項目を選択」をクリック (複数の分野から項目を選択できるが，1 回に DL できる項目は 25 項目まで)

⑤ 統計表表示の画面になるので，左側のタブから「レイアウト設定」をクリック→表示年度に抽出したい西暦年を入力

⑥ 再度左側のタブから「レイアウト設定」をクリック→統計表表示の画面に戻る→右上の「ダウンロード」をクリック

⑦ 開かれたウィンドウに必要な情報を設定し「ダウンロード」をクリック

分析例: 市・区・町・村, それぞれの数はどれくらい?



「市」と「町」は同じくらいあるが, 人口は「市」に多い (市民が一番多い).

「区」の数は少ないが, 人口は「町」と同じくらい多い.

収集 2: Web 上の表の利用

活動 (手順) 10 分

① **Excel** を開く

② 「データ」タブを開く

③ 「**Web** から」をクリック

④ 「**URL**」 に次の通り入力して「**OK**」をクリック:
https://memorva.jp/ranking/japan/mext_gakuryoku_test.php

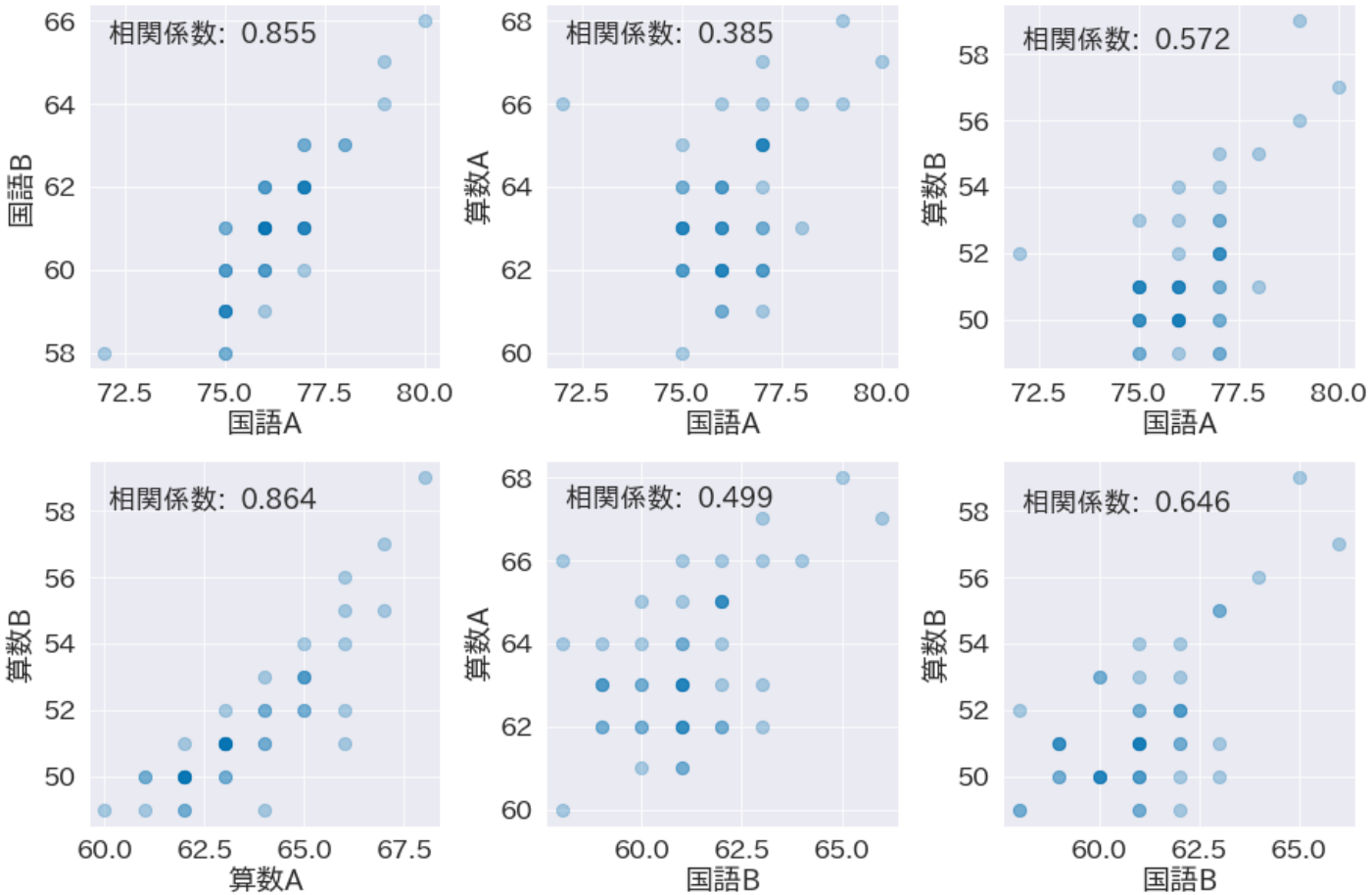
⑤ 1 つ目の「国語」を選び「読み込み」をクリック.

⑥ 同様の手順で「算数」のデータも「読み込む」.

上記のような方法で, インターネット上のデータを収集することを「スクレイピング」と呼びます.

分析例: 国語と算数の正答率の関係性

問: 国語と算数の能力には関係性があるだろうか?



○○ **A** は知識を問う問題から成り, ○○ **B** は活用力を問う問題から構成されている.

同じ教科の試験の相関は (当然だが) 強い.

算数の活用力を要する問題の方が, 国語との相関が強い (算数の活用力を延ばすには国語力の育成が重要?).

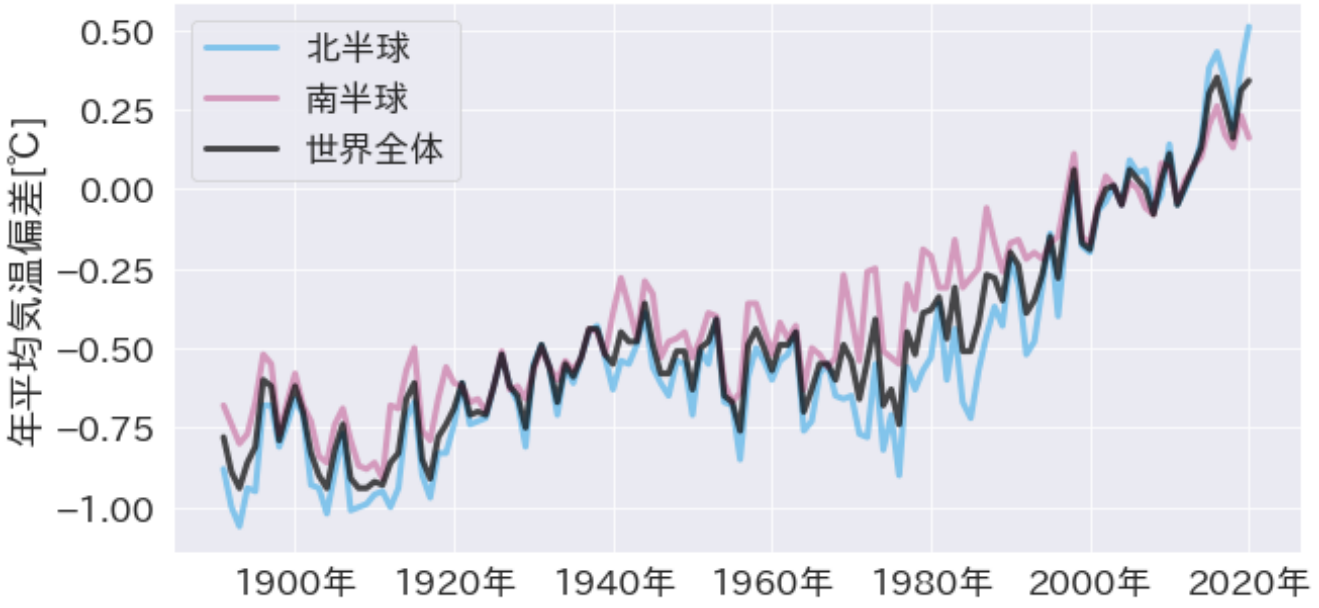
知識を問う試験は相関が弱い傾向が見られる.

図の読取り

目標

- データが容易に入手できるようになった現在ではグラフを見る機会も多くなりました．そのため、「グラフから正確に情報を読み取る技術」の重要性も増していると言えるでしょう．
- ここでは，地球温暖化に関わるデータを通して，図から情報を読み取る際に，どのような点に注意するかを考えてみよう．
- 同時に，図を作成する際に注意すべき事柄も学びましょう．

気温偏差の年平均値



気象庁 **HP** のデータを利用

- (疑問) なぜ，平均気温ではなく平均気温偏差を考えるのか？ 観測地点はどこを選んでいるのか？ 地上だけ？
- (傾向) 120 年間で，約 1 ℃気温が上昇している．1940〜1980 年の間は，気温が上昇していないように見える 北半球の方が寒いことが多いが，近年は北半球の方が暑い．

活動 1 (10 分)

(1) グラフを見て，グラフ作成時の注意点を見つけよう．

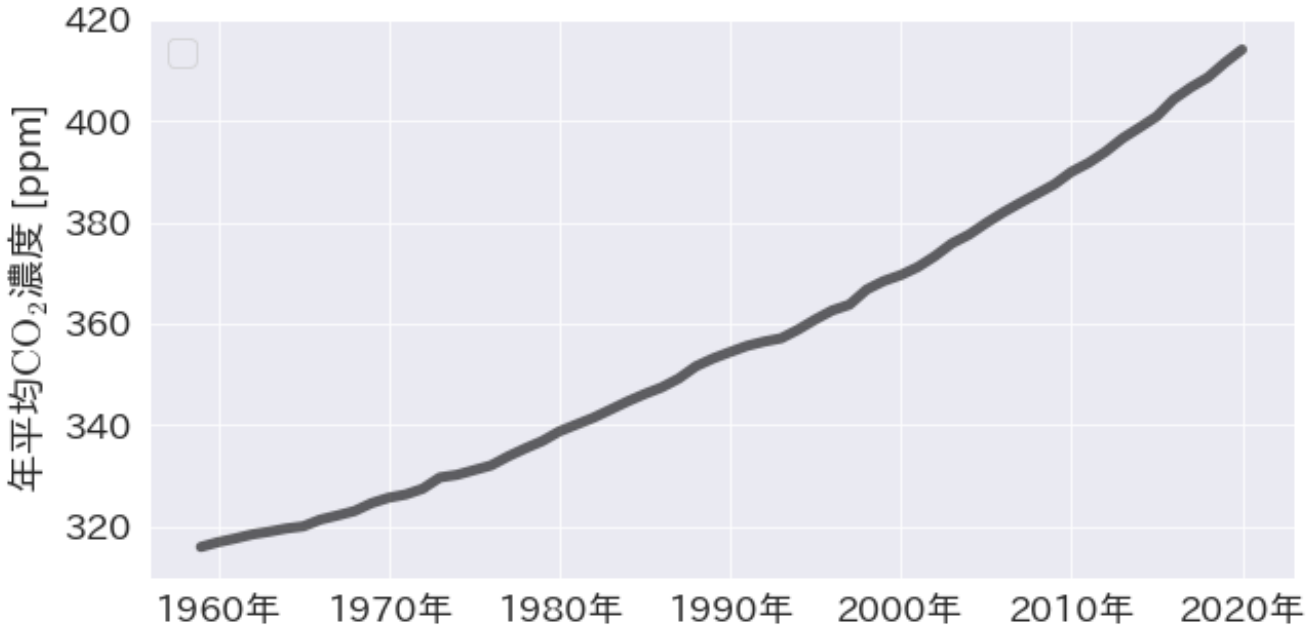
(2) グラフから感じる「疑問」は？

(3) グラフから読み取れる「傾向」は？

気温偏差：観測地点毎の基準値との差

基準値：1991〜2020 年の 30 年平均値

二酸化炭素濃度の年平均値



Global Monitoring Laboratory HP のデータを利用

- **(疑問)** 観測地点はどこか? **ppm** という単位の意味は?
- **(傾向)** 60 年間で, 約 100[ppm] **(約 3 割)** 上昇している. 近年上昇率が上っているように見える.

活動 2 (5 分)

(1) グラフを見て, グラフ作成時の注意点を見つけよう.

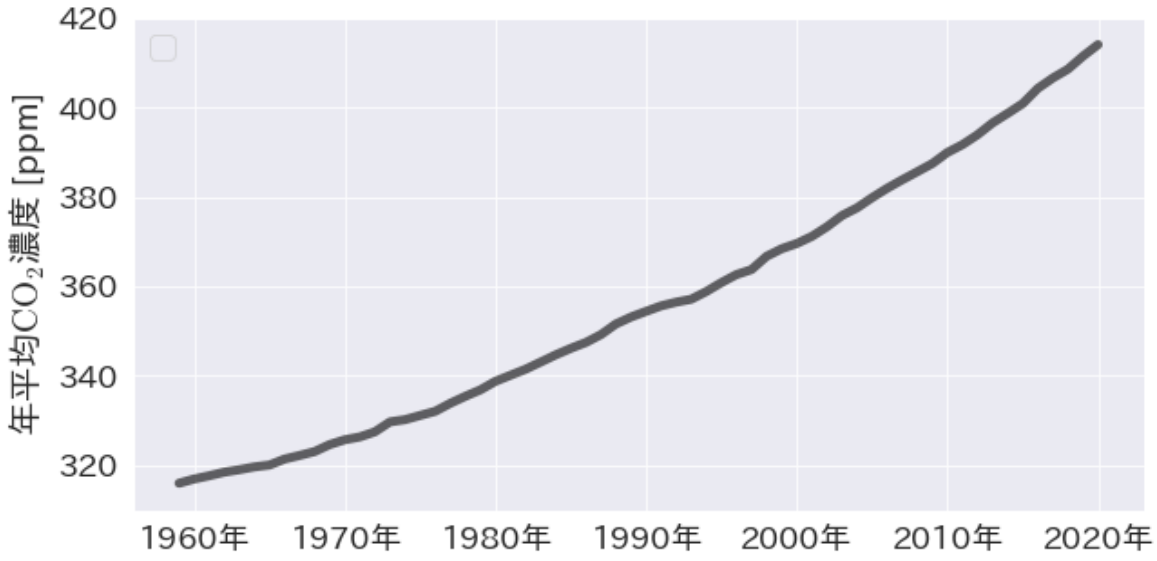
(2) グラフから感じる「疑問」は?

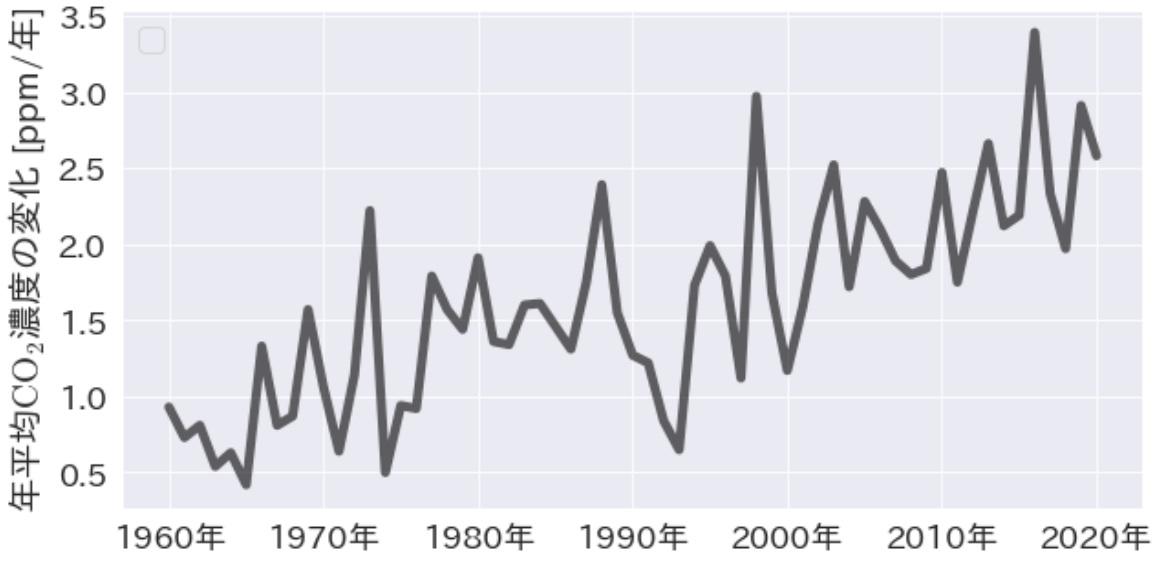
(3) グラフから読み取れる「傾向」は?

観測地点: ハワイのマウナ・ロア天文台

ppm: 大気中の分子 100 万個中にある対象物質の個数

二酸化炭素濃度の年平均値の差分



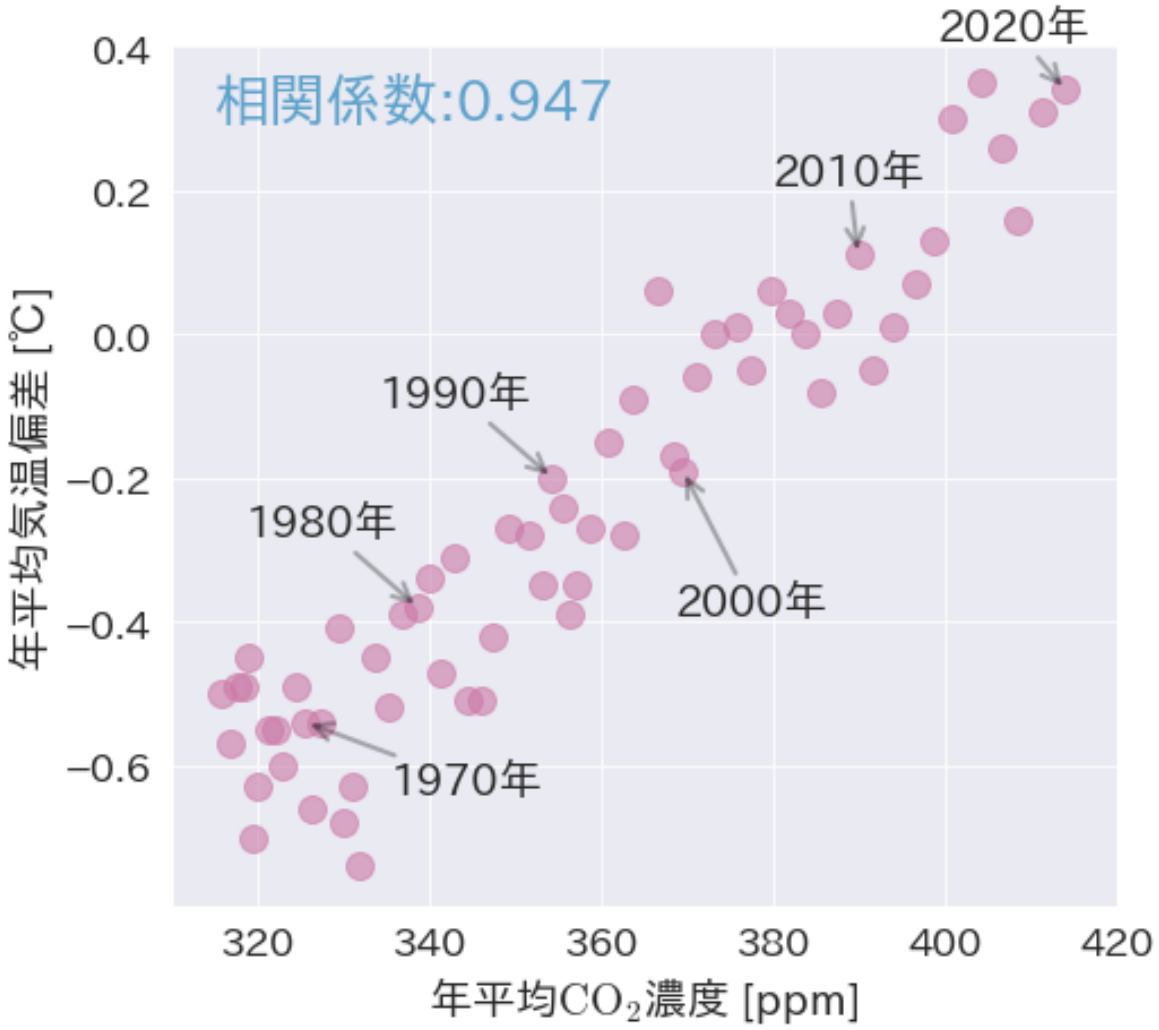


Global Monitoring Laboratory HP のデータを利用

下の図は，1 年間の濃度変化を計算したものです．

このような処理を差分と呼び，時系列データ分析では良く行われます．微分の概念とも類似していますね．

気温偏差 vs 二酸化炭素濃度



気象庁 HP と Global Monitoring Laboratory HP のデータを利用

活動 3 (5 分)

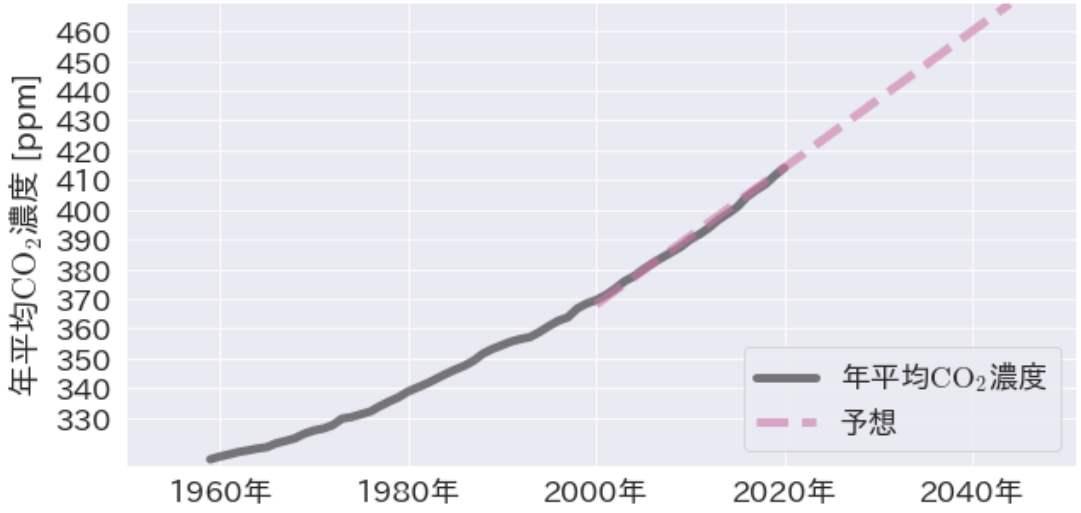
(1) グラフを見て，グラフ作成時の注意点を見つけよう．

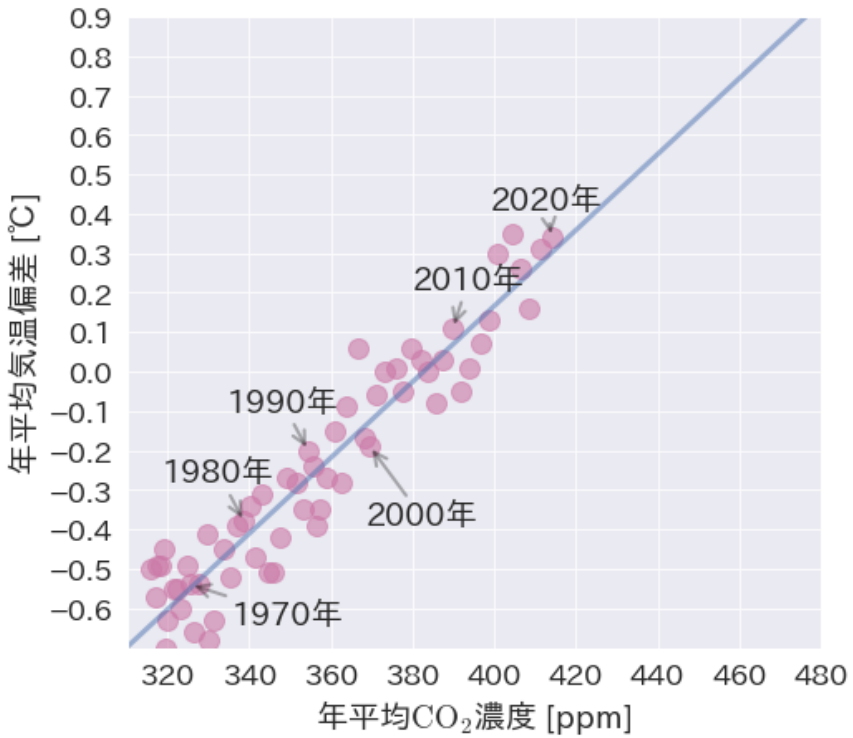
(2) グラフから感じる「疑問」は?

(3) グラフから読み取れる「傾向」は?

相関が非常に強いが，これだけで 二酸化炭素濃度の増加が地球温暖化の原因と言えるのでしょうか?

予測





活動 4 (5 分)

左の 2 つの図を用いて，2040 年の気温偏差を予測せよ．

まとめ: データの読み取り

(1) 座標軸は何を表しているか?

(2) 異なる線やシンボルの違いは何か?

(3) データの収集方法は? · 観測地点は? · 観測時期は?

(4) 収集したデータをどのように処理してあるか?

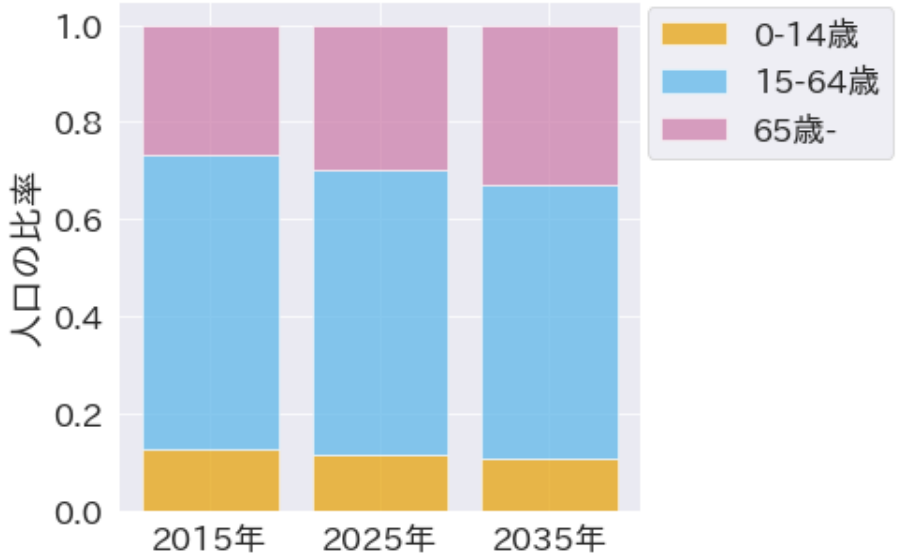
(5) どのような傾向が読み取れるか? (定量的に把握しよう)

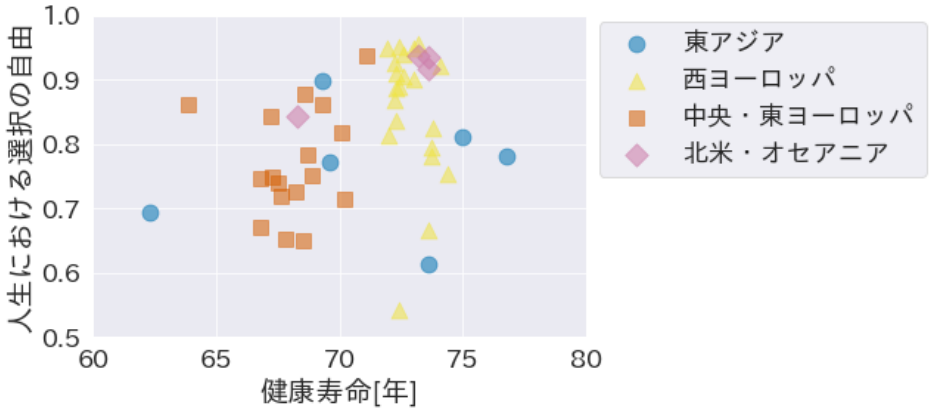
- (a) その原因は?
- (b) そこから予想されることは?

統計量と可視化

目標

ここでは，いくつかのデータを可視化する活動を通して，データを可視化する際に，どのような点に注意するかを考えてみよう．





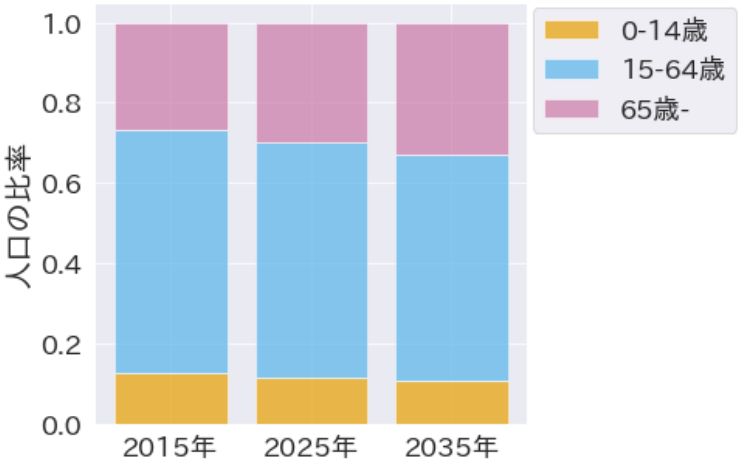
棒グラフの作成

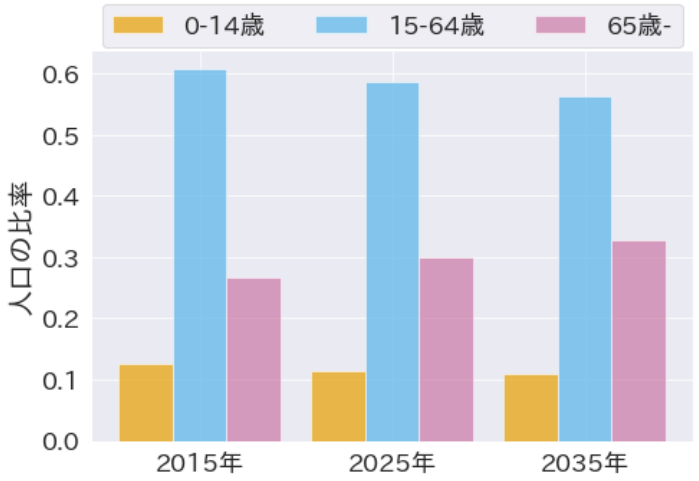
活動 (20 分)

右の 2 つの棒グラフは

moodle 基礎情報・データ科学
「日本の人口のデータセット」

を用いて作成した．このデータをダウンロードし同様のグラフを作成してみよう．





作成方法は問わないが，特にこだわりが無ければ **Excel** を用いると良いでしょう．分からないことはインターネットで検索するか (**'Excel 棒グラフ'** とか**'Excel 凡例'** などで検索すると良いでしょう)，先生に質問してください．

散布図の作成

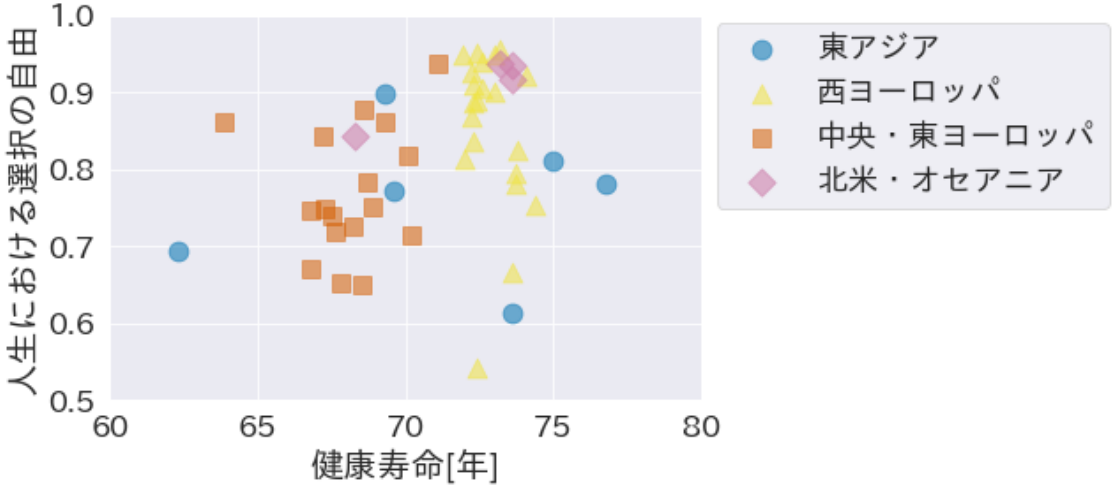
下のグラフ (散布図と呼ばれる) は

moodle 基礎情報・データ科学

世界の国々の幸福度データセット 2020 年 (簡易版)

を用いて作成した.

図中の「健康寿命」はデータでは「**Healthy life expectancy**」，「人生における選択の自由」は「**Freedom to make life choices**」という変量に対応している．変量 **regional indicater** 中の **ANZ** はオセアニアを意味する.



活動 (20 分)

① 左のデータセットをダウンロードし，同様の散布図を作成せよ (できれば色分けせよ).

② 散布図の 2 変量について，それぞれの平均値と分散を求めよ.

③ 2 変量の相関係数を求めよ.

(②と③は **Excel** の関数 **AVERAGE, VAR.P, CORREL** を用いても可)

計算した統計量の数値
→ **moodle** から提出

箱ひげ図の作成 1

2020 年のセンター試験に出題された箱ひげ図

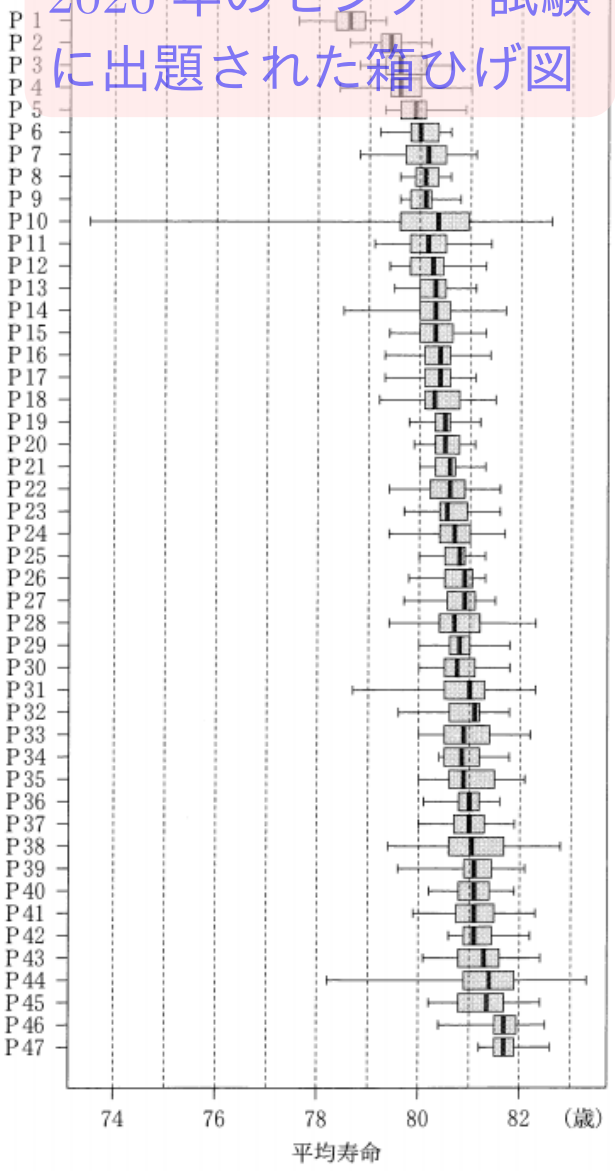


図 1　男の市区町村別平均寿命の箱ひげ図
(出典：厚生労働省の Web ページにより作成)

左の図について

- 縦軸は都道府県名 (記号で置き換えられているため具体的な都道府県名は分からない)
- 横軸は市区町村ごとの「男性の平均寿命」

箱ひげ図について

- 箱の中の黒い縦線の値は中央値
- 箱の左右の端の値は (第 1 および第 3) 四分位数
- 箱から延びている線の両端が最小値と最大値

活動 (10 分)

授業のページから「平均寿命のデータセット」をダウンロードし, センター試験と同様の箱ひげ図を作成してください.

moodle 基礎情報・データ科学

箱ひげ図の作成 2

できた人は次の問について考えてみて欲しい.

活動

「平均寿命のデータセット」は「平均寿命のデータセットの元データ」に前処理をして作成したものである.

同じのホームページから「平均寿命のデータセットの元データ」をダウンロードし, どのような処理をすれば「平均寿命のデータセット」を作成できるか考えてみよ.

まとめ: データの可視化

Take Home Message

(1) 座標軸の意味・単位は伝わるか?

(2) 異なる線やシンボルの違いは伝わるか? (凡例を入れたか?)

(3) 表示範囲は適切か?

(4) 線の太さや, シンボルの大きさ, 色は適切か?

(5) 軸の目盛りは適切か?

(6) 無駄な情報は無いか?

(7) より適切な可視化方法はないか?

テキスト分析をしてみよう 1

活動 1 5 分

授業のページの

moodle 基礎情報・データ科学

kadai1.txt

をダウンロードしてください．これは，「教育用標準データセット (市区町村別データ)」に関するみなさんの問をまとめたものです．

① どのようなデータか確認し，

② 問を立てよう．さらに，

③ どのように分析すれば良いか考えよう．

次の問について考えよう

代表的な問はどのような問か?

https://textmining.userlocal.jp/

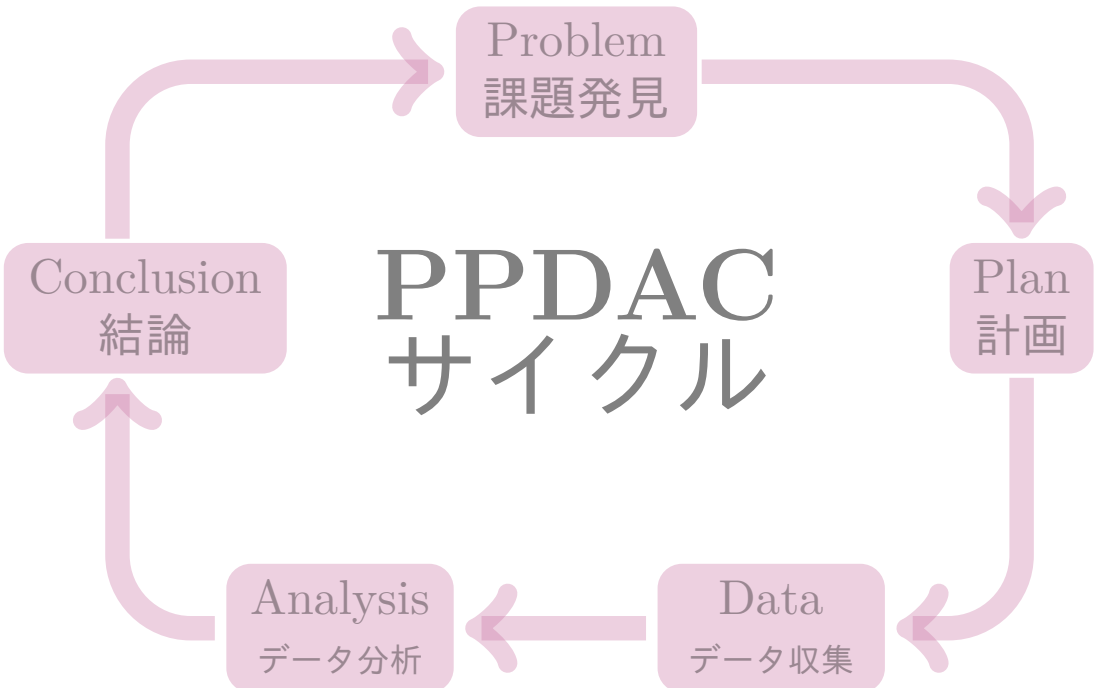
の「フォーム入力」に **kadai1.txt** の内容をコピペしてください．

活動 2 15 分

① いくつかの図が出力されます．「ワードクラウド」と「共起キーワード」の 2 つの図の意味を (ある程度) 理解しましょう．

② 上の問に対する回答をまとめよう．

PPDAC サイクル



最初の実習で，「問」を立ててもらいました (課題発見).

今回は，「問」に対して実際にデータ分析をしてもらいます (グラフを作成してください).

さらに，分析結果について考察し，結論 をまとめてもらいます.

最後に，新たな「問」を 1 つ立てましょう (課題発見).

教育用標準データセット

活動 1 30 分

授業のページから

moodle 基礎情報・データ科学

kadai1.txt

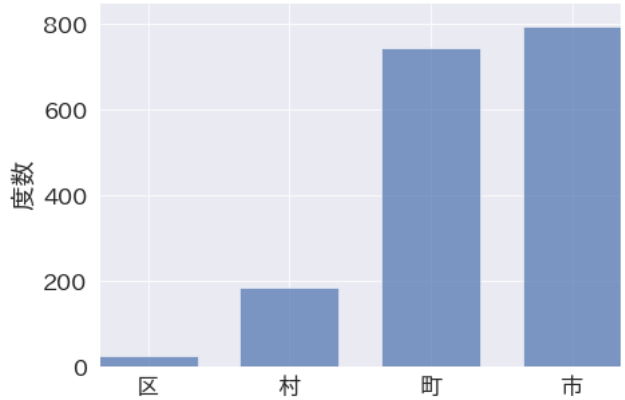
を DL して下さい．これは，「教育用標準データセット (市区町村別データ)」に関するみなさんの問をまとめたものです．

① 他の人の問も参考にしながら，1 つだけ検証する問を選ぼう (もちろん，先週自分で考えた問でも良いです)．

② グラフを作成して問について検証し，検証した結果を文章にまとめよう．

③ 検証結果を踏まえて，新たに問を 1 つ立てよう．

問: 市区町村の数や比率は?



「市」と「町」の数が多い (大体同じくらいの数がある)．「村」の数はその $\frac{1}{3}$ 程度であり，区は 1 番少なく，「村」の約 $\frac{1}{8}$ 程度である．

新たな問 区と市に重複はないのか? 市区町村に住んでいる人の人口は?

教育用標準データセット: ファイル提出

前ページの活動をまとめて, **moodle** から提出してください.

Word ファイル内に氏名は記入しないでください **(**記入しなくても氏名は確認できます**)**.

今後の授業で, 氏名が分からない形式で全員の分析結果を共有する可能性があります. ご了承ください.

提出の流れ

Excel で図を作成
↓
グラフを **Word** に貼り付け
↓
「検証した問」と「検証結果の文章」·「新たな問」を **Word** に加筆
↓
pdf ファイルをエクスポート
↓
moodle から提出

テキスト分析をしてみよう 2

前ページのサイトで行ったテキスト分析は「テキストマイニング」と呼ばれます．AI も使われているようですね．

もう 1 つテキストマイニングをしてみましょう．

活動 3 10 分

授業のページの

moodle 基礎情報・データ科学

kadai2.txt

をダウンロードしてください．これは，「オリンピック出場選手のデータセット」に関するみなさんの問をまとめたものです．

① 代表的な問はどのような問か? についてテキストマイニングをしてみよう．

② 分析結果を文章にまとめよう．

オリンピック出場選手のデータセット

活動 1 20 分

授業のページから

moodle 基礎情報・データ科学

kadai2.txt

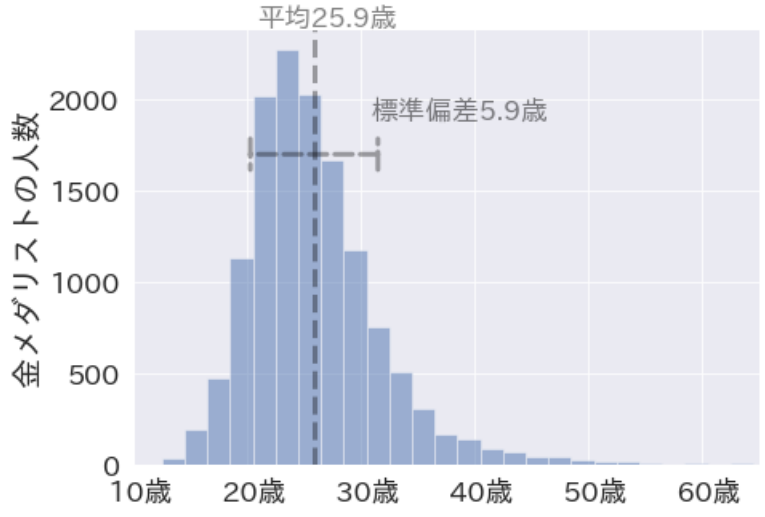
を DL してください．これは「オリンピック出場選手のデータセット」に関するみなさんの問をまとめたものです．

① 他の人の問も参考にしながら，1 つだけ検証する問を選ぼう (もちろん，先週自分で考えた問でも良いです)．

② グラフを作成して問について検証し，検証した結果を文章にまとめよう．

③ 検証結果を踏まえて，新たに問を 1 つ立てよう．

問: 金メダリストの年齢分布は?



平均が 25.9 歳・中央値が 25 歳の (やや非対称な) 山型の分布に従っている．標準偏差は 5.9 歳であり，約 7 割が 20 台である．最年少は 13 歳，最年長は 64 歳である．

新たな問 オリンピック出場選手の年齢分布との違いはある?

オリンピック出場選手のデータセット: ファイル提出

前ページの活動をまとめて, **moodle** から提出してください.

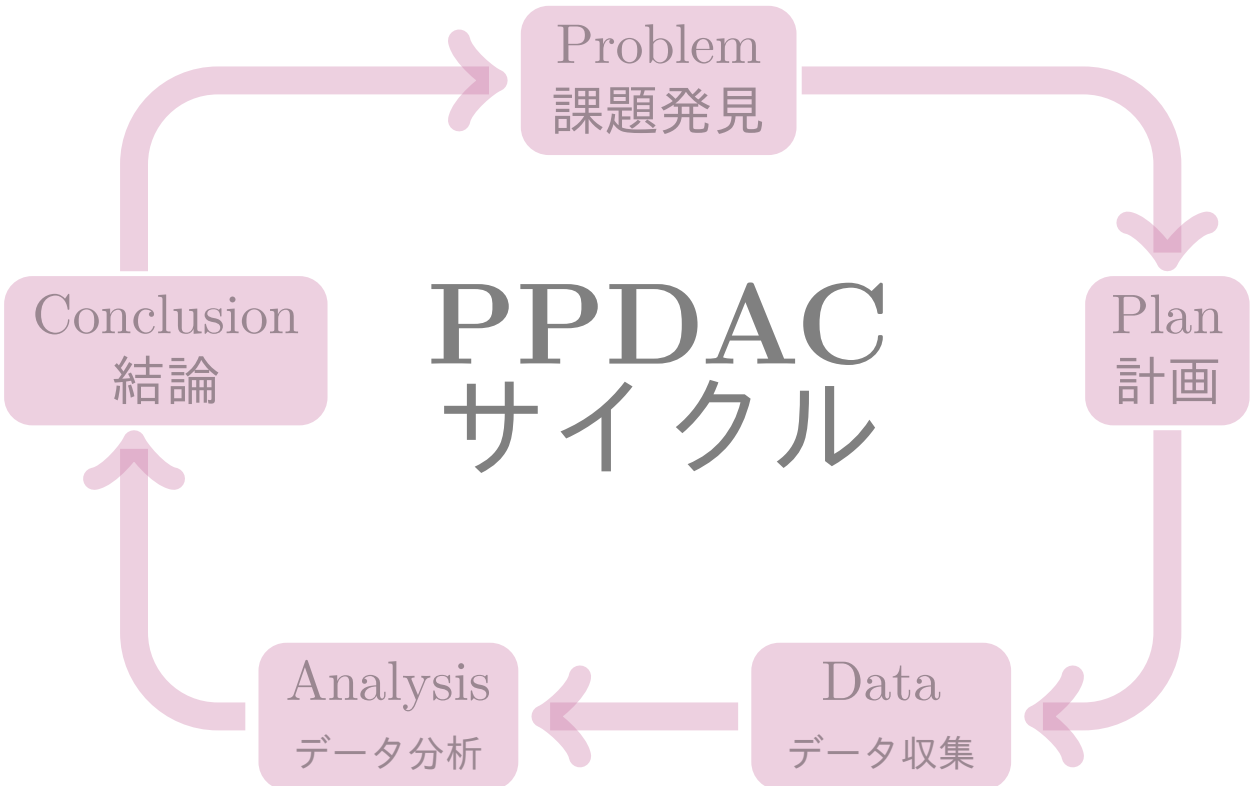
Word ファイル内に氏名は記入しないでください **(**記入しなくても氏名は確認できます**)**.

今後の授業で, 氏名が分からない形式で全員の分析結果を共有する可能性があります. ご了承ください.

提出の流れ

Excel で図を作成
↓
グラフを **Word** に貼り付け
↓
「検証した問」と「検証結果の文章」·「新たな問」を **Word** に加筆
↓
pdf ファイルをエクスポート
↓
moodle から提出

PPDAC サイクル



「データ分析」の部分は，機械学習や AI が担うケースが増えてきた．

今回は，これらの高度なデータ分析について理解を深めよう．

(補足: 逆に「課題発見・計画・結論」は人間の関与が必要な部分と言えます)

鳥とハングライダー: 問題

活動 15 分

ハングライダーを作るとしたら，翼の面積はどの程度にすれば良いか．鳥の体重と翼の面積のデータを元に考えよ:

moodle 基礎情報・データ科学
「鳥のデータセット」

鳥の種類	体重 [g]	翼の面積 [cm^2]
スズメ	25	87
イワツバメ	47	186
黒ツグミ	78	245
ムクドリ	93	190
ハト	143	357
カラス	607	1344
カモメ	840	2006

出典「数学的モデル化を遂行する力を育成する教材開発とその実践に関する研究」(西村圭一著)

ヒント

① **Excel** で散布図を作成

② グラフの右上にある「プラス記号 (グラフ要素)」で「近似曲線」にチェック

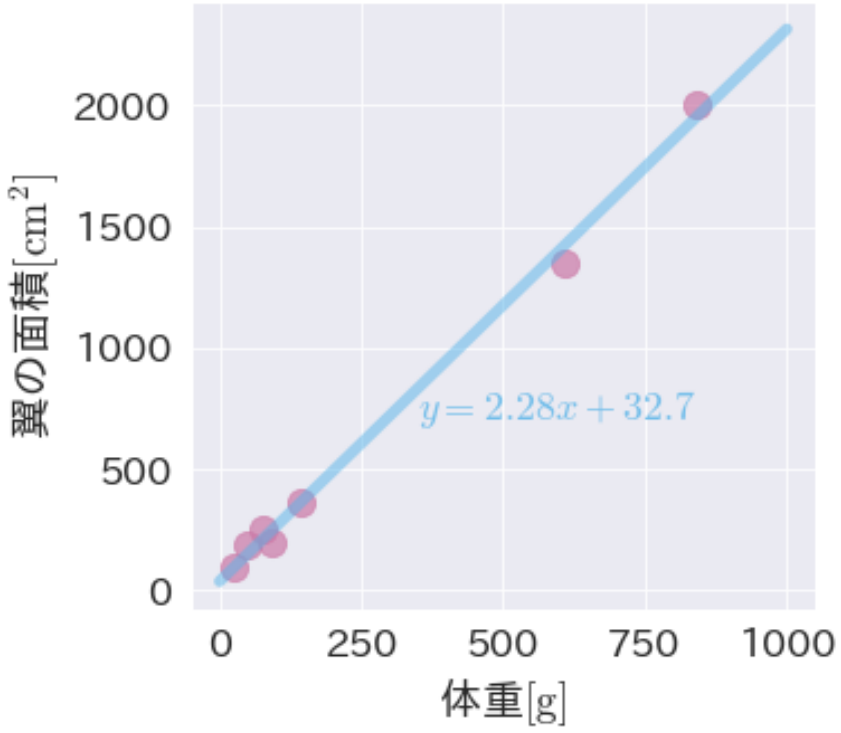
③ さらに「近似曲線」の右側に表われる記号 (>) から「その他のオプション」を選ぶ

④ 「グラフに数式を表示する」にチェックを入れる

鳥とハングライダー: 解答例

鳥の種類	体重 [g]	翼の面積 [cm^2]
スズメ	25	87
イワツバメ	47	186
黒ツグミ	78	245
ムクドリ	93	190
ハト	143	357
カラス	607	1344
カモメ	840	2006

出典「数学的モデル化を遂行する力を育成する教材開発とその実践に関する研究」(西村圭一著)



近似曲線は $y = 2.28x + 32.7$ となるので，x に $80000[\mathrm{g}]$ を代入すると，翼の面積は (単位を変換して) 約 $18\mathrm{m}^2$ となる．これは実際のハングライダーの翼の面積にかなり近い．

ここで，ヒトの体重を $60[\mathrm{kg}]$，ハングライダー自体の重さを $20[\mathrm{kg}]$ と想定した．

機械学習と回帰 (復習)

機械学習とは, データを理解するために数理モデルを構築すること.

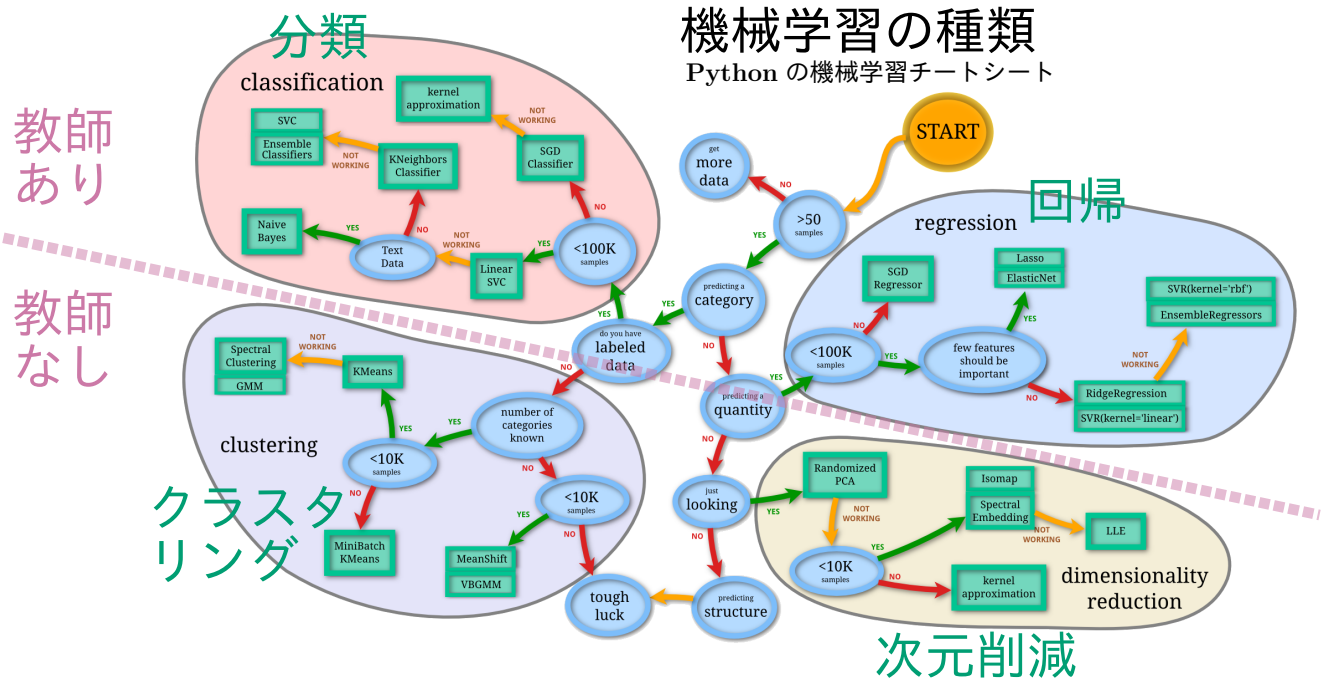
ここで「学習」とは, 数理モデルのパラメータを観測データに適応するために調節することを差す. この調整をコンピュータ (すなわち「機械」) が行うので, 「機械が学習する」と呼ばれている (**VanderPlas**, ***Python Data Science Handbook*** より引用; 多少意訳している).

例: 線形回帰の場合, 数理モデルは一次関数

$$y = ax + b$$

である. パラメータは傾き a と切片 b であり, データに基いて決定される.

機械学習は様々な手法に対する総称である (右図に参照).



あやめの分類

活動 30 分

典型的なデータ分析を体験しよう．次のファイルをダウンロードせよ:

moodle 基礎情報・データ科学 → 「アヤメのデータセット」

これはアヤメ科の植物のデータである．アヤメ科の植物は外見が類似していて分類が難しい (’いずれ菖蒲か杜若’ という故事もある).

train データを分析し, **test** データの個体の種類を予想せよ．良い分類規則が見つけられるか? 全て予想できたら **moodle** から入力して採点せよ.

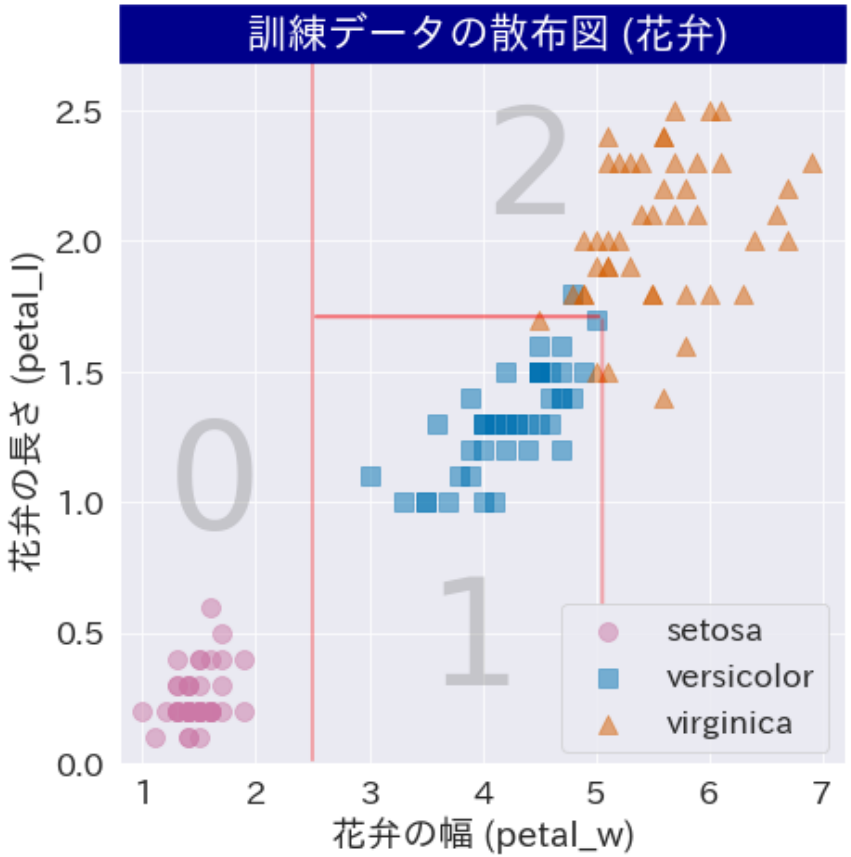

アヤメ


ハナショウブ

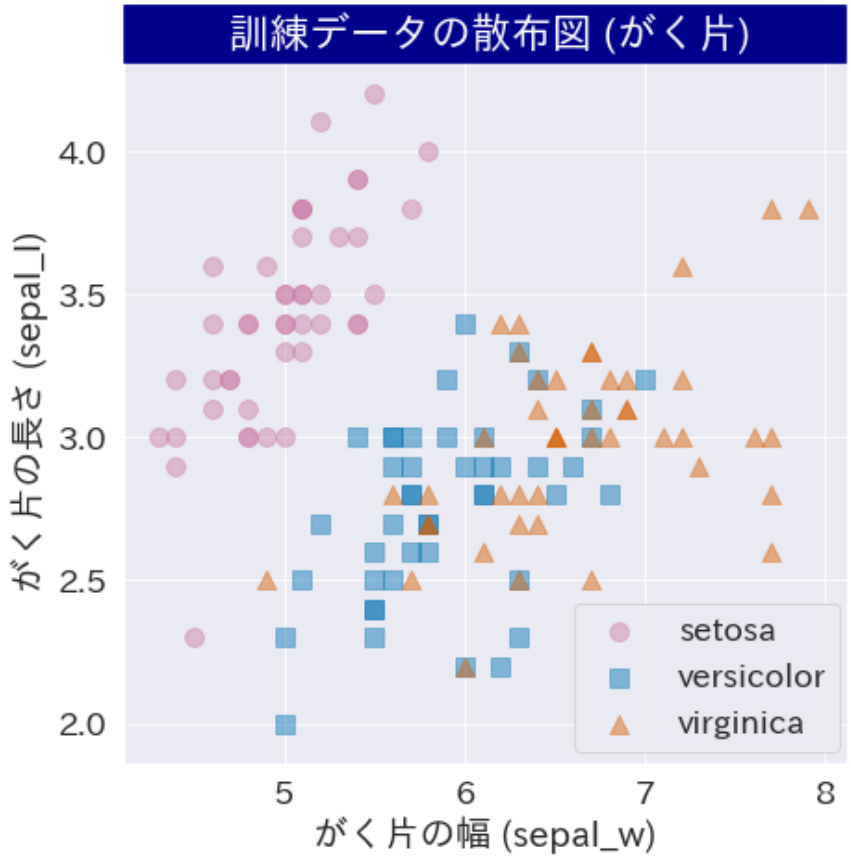

カキツバタ

写真は全て **wikipedia** より

考え方の例: 散布図の利用

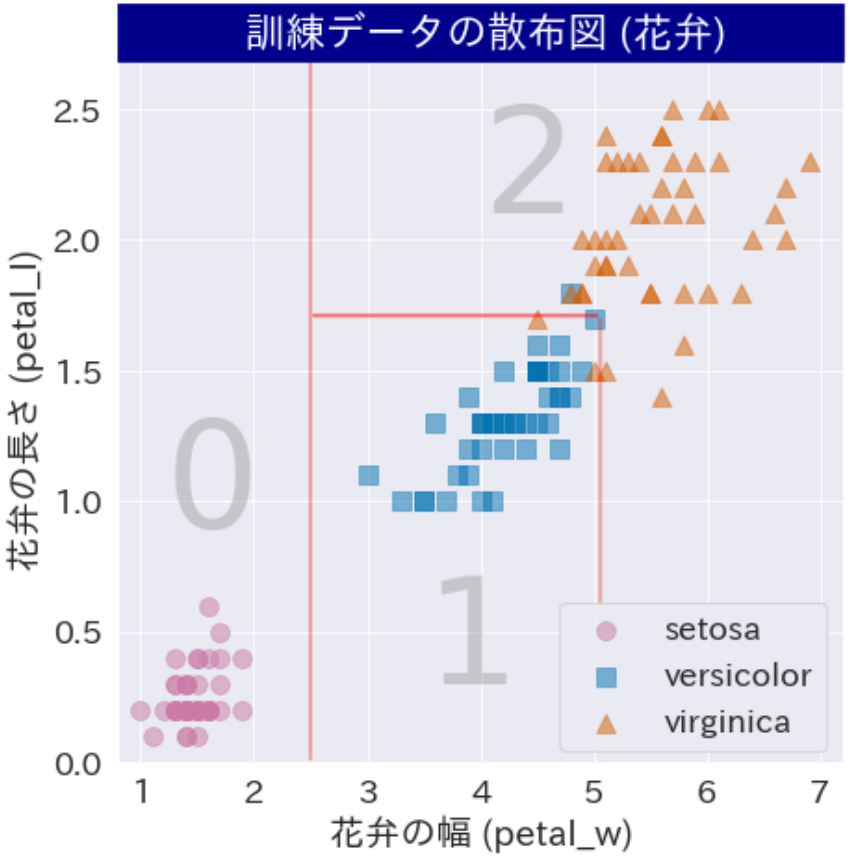


3 種類のマーカーがある程度分離しているので, 境界線が引けそう.

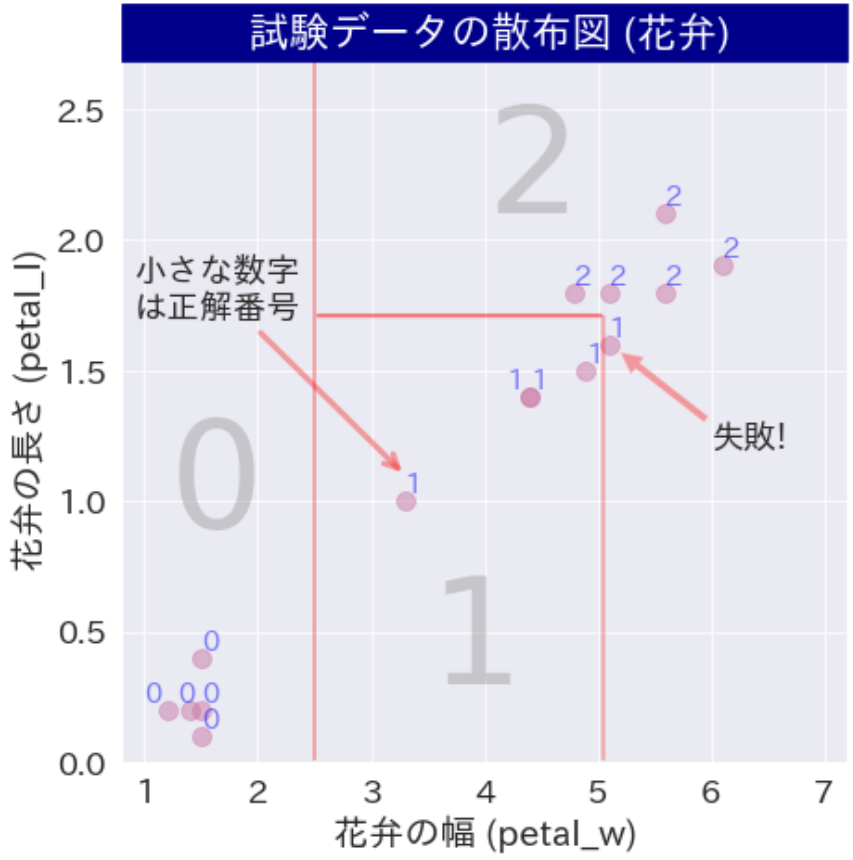


3 種類のマーカーが混在しており, 明確な境界線は引けそうにない.

考え方の例: 散布図の利用



訓練データを元に，境界を引いた．もちろん境界の引き方には様々な可能性がある．



訓練データを元に作成した境界に従って分類すると，1 つの点以外は分類に成功する．

機械学習の利用例 (Python)

深層学習

```
import pandas as pd
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier

iris = load_iris()
x_train, x_test, y_train, y_test = train_test_split (
# random_state は乱数のシード
iris.data, iris.target, stratify = iris.target, test_size=0.1, random_state=21)

model = MLPClassifier(solver='lbfgs', random_state=3)
model.fit(x_train, y_train)

print(f'正解率: {model.score(x_train, y_train)}')
print(f'正解率: {model.score(x_test, y_test)}')
```

```
正解率: 1.0
正解率: 0.9333333333333333
```

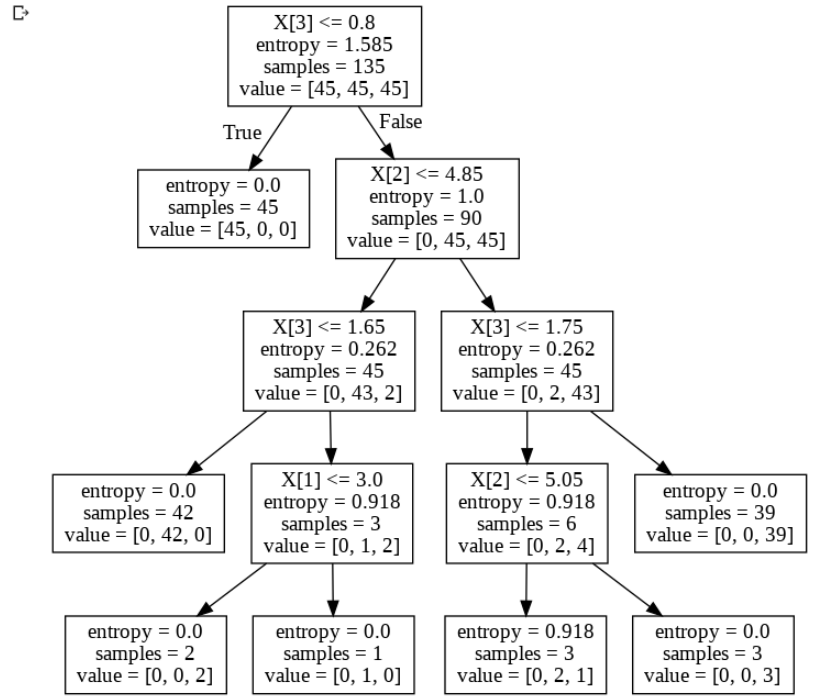
決定木

```
[17] from sklearn.tree import DecisionTreeClassifier

     model = DecisionTreeClassifier(criterion='entropy', max_depth=4, random_state=3)
     model.fit(x_train, y_train)

     print(f'正解率: {model.score(x_train, y_train)}')
     print(f'正解率: {model.score(x_test, y_test)}')
```

```
正解率: 0.9925925925925926
正解率: 0.9333333333333333
```

決定木の詳細

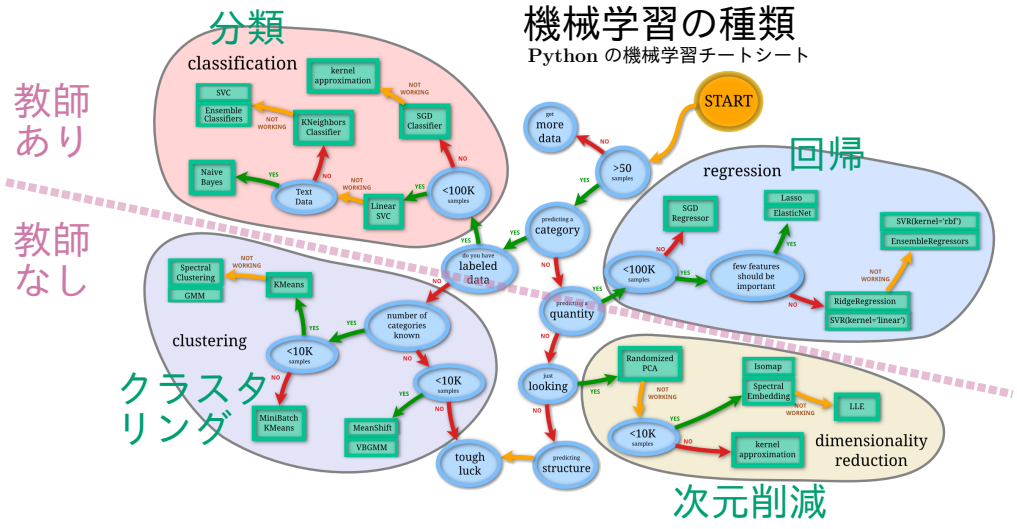
```
[14] from sklearn import tree
     import pydotplus
     from sklearn.externals.six import StringIO
     from IPython.display import Image

     dot_data = StringIO()
     tree.export_graphviz(model, out_file=dot_data)

     graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
     Image(graph.create_png())
```



数学的には 4 次元空間を 3 次元平面で分割し，分割された各領域にあやめの種類を割り当てている．

機械学習と分類問題

- あやめの分類で決定木や深層学習を用いると，自動的に境界を決定してくれる．
- 決定木の場合，直線や平面による境界を決定する．深層学習の場合より複雑な曲線や曲面を用いて境界を決定する．
- データが高次元になるほど人間の能力では分類 (回帰も) することは難しく，機械学習を用いる必要がある．



分類や回帰の機械学習は AI 技術で広く用いられているという．

タイタニック号の乗客データ

活動

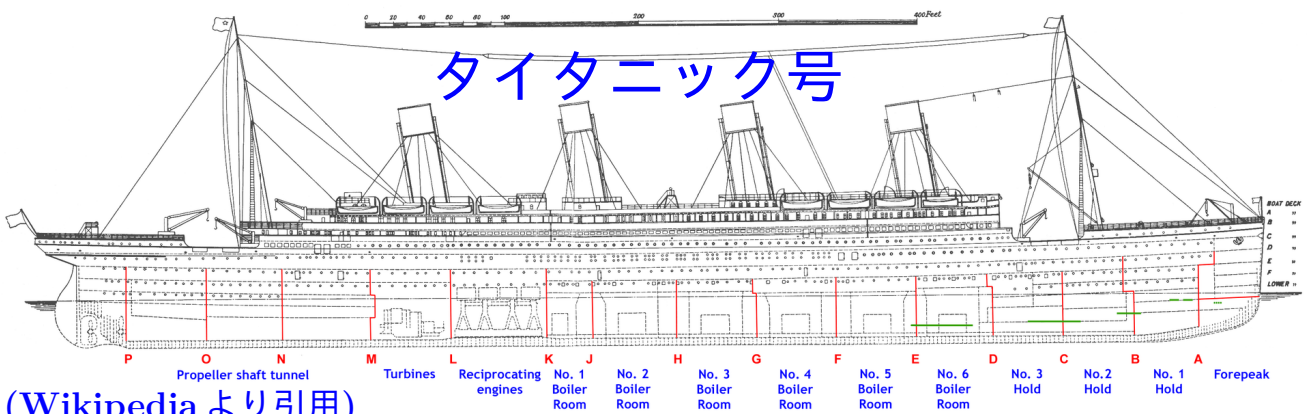
ここではより本格的なデータセットを扱う (とは言っても機械学習の勉強によく使われるデータセットですが)

moodle 基礎情報・データ科学

「タイタニック号沈没事故の乗客のデータセット」

をダウンロードせよ．このデータセットでは，**train** データの分析を元に，**test** データの乗客が沈没事故を生き延べたか否かを予想する．

多変量であり，かつ量的データと質的データが混在した本格的なデータセットです．データ分析の難しさと奥深さを実感しましょう．また，答え合せは **Kaggle** でできます．



(Wikipedia より引用)

調べてみよう: 分類とクラスタリング

活動

講義では機械学習の例として,「分類」と「クラスタリング」について紹介しました.

インターネットでこれらの具体例を調べてみよう.

ある程度調べたら,隣の人と情報交換してみよう.

きゅうりの等級予測 (分類・教師あり)

農産物を販売する際, 大きさなどを基に等級が決められます. この等級の決定に機械学習が利用されています. 実際, きゅうりの画像データから等級を予測するシステムが開発され, 利用されているそうです.

https://www.itmedia.co.jp/enterprise/articles/1803/12/news035.html

楽曲の分類 (クラスタリング・教師なし)

旋律や調性等の特徴や印象に基づき楽曲を分類 (クラスタリング) する手法が開発されています. このような手法を用いることで, 個々のユーザが自分の好みの楽曲を見つけることが容易になることが期待されます.

http://itolab.is.ocha.ac.jp/~itot/paper/ItotDCPJ147.pdf

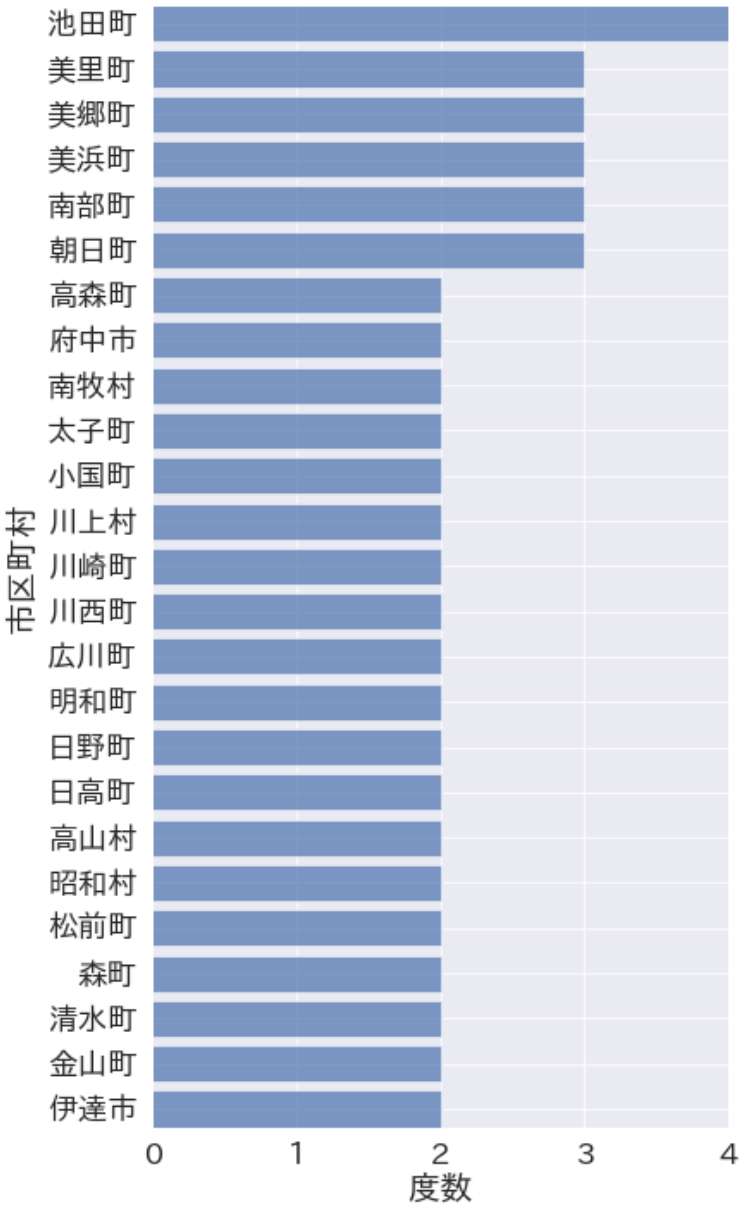
基礎情報教育: データ科学入門
第 6 章 様々な話題

T. MIYAGUCHI

Naruto Universality of Education

June 21, 2023

分析例: 同じ市区町村名はどれくらいある?

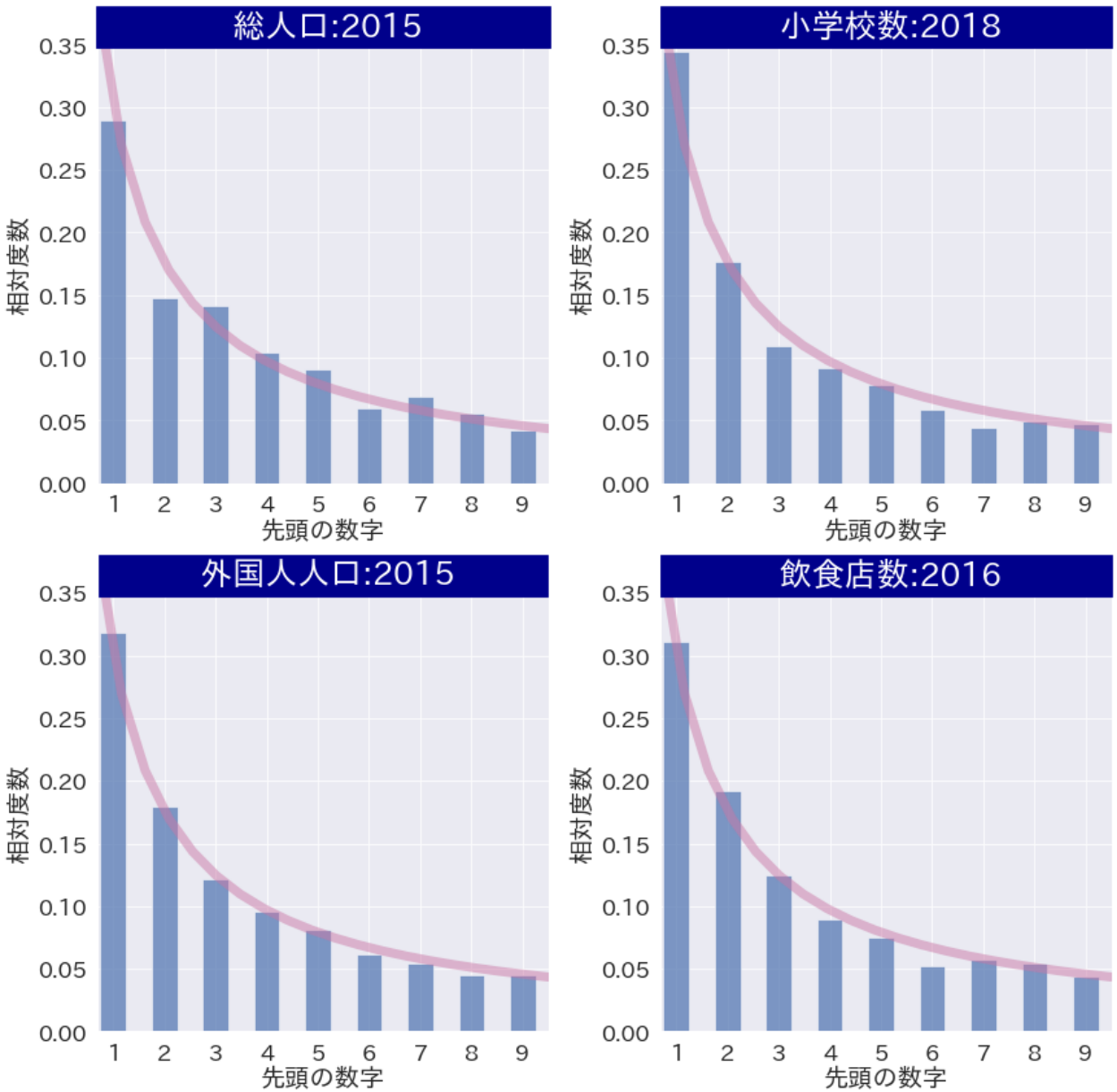


同じ市区町村名は意外に少なく, 最も多いのが「池田町」で 4 つの都道府県にある (北海道, 福井県, 長野県, 岐阜県).

同じ市区町村名は「町」に多く, 「市」は少い.

同じ読みでは「みさと町」は 6 つある (他にもあるかも).

分析例: 先頭の数字は何?



どの場合も 1 が最も多く，数字が大きくなるに従って，相対度数は小さくなる.

赤い線は次の関数を示している:

$$y = \log_{10}\left(1 + \frac{1}{x}\right)$$

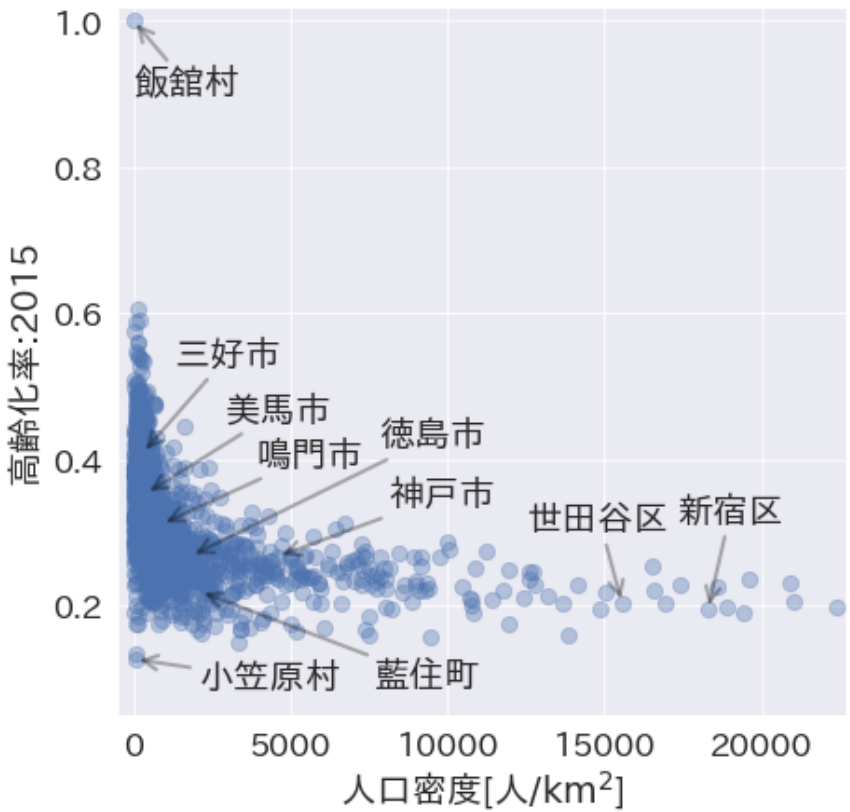
これをベンフォードの法則と呼ぶ.

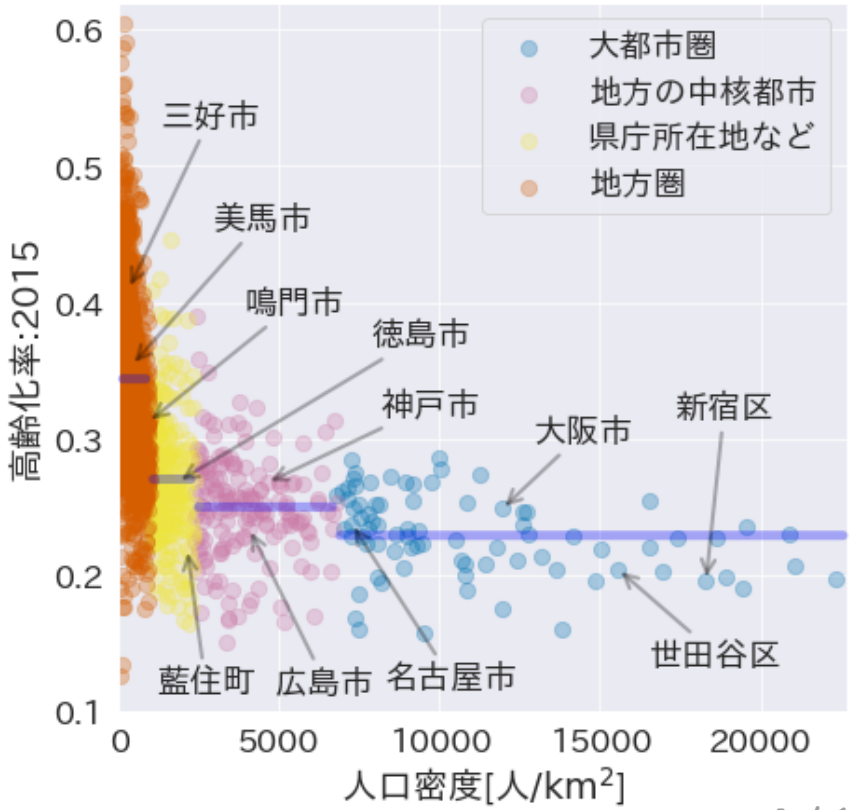
分析例: 地方の方が高齢化が進んでいる？ 1

$$\text{高齢化率} = \frac{\text{65 歳以上人口}}{\text{総人口}}$$

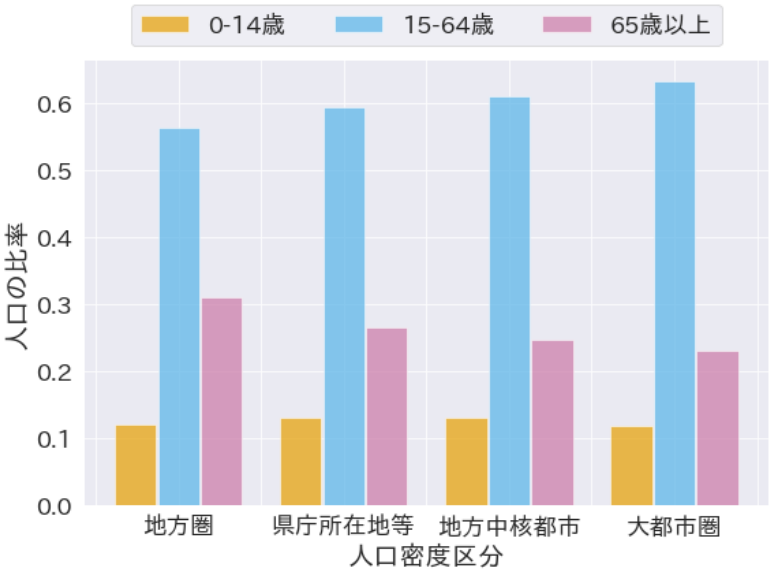
$$\text{人口密度} = \frac{\text{総人口}}{\text{可住地面積}}$$

人口密度を 4 つに分けて，それぞれの階級の中で平均値を計算した (青い線)．都市部ほど高齢化率が減少する傾向が見られる．



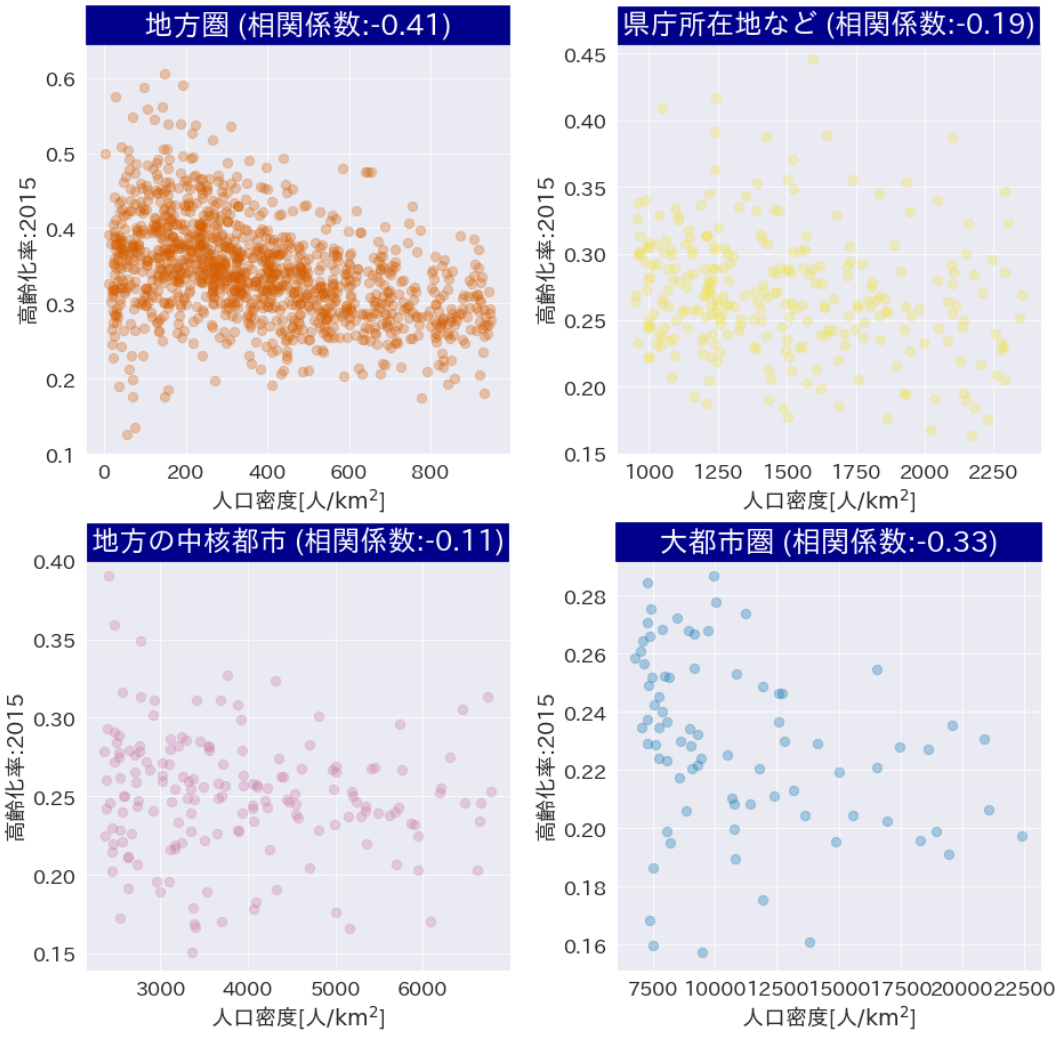


分析例: 地方の方が高齢化が進んでいる？ 2

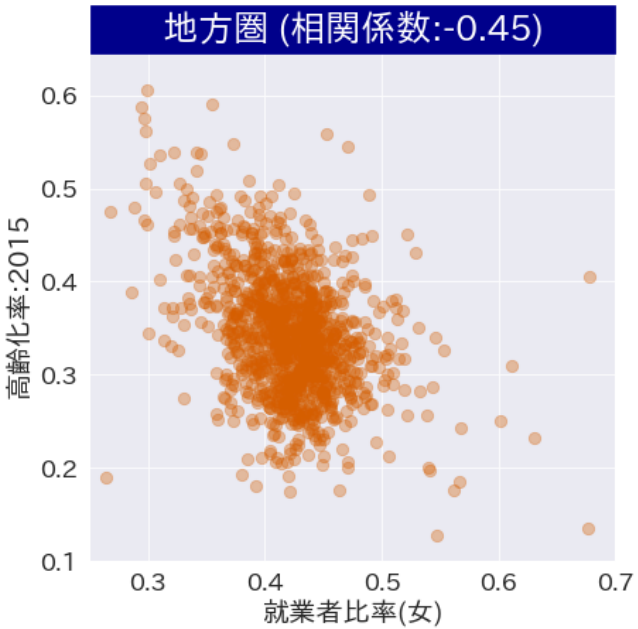


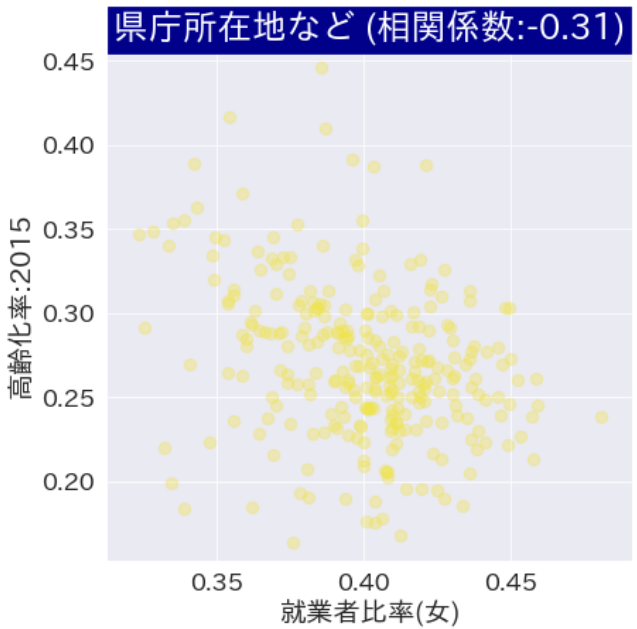
地方ほど高齢者が多く，生産年齢人口は少ないことが分かる．

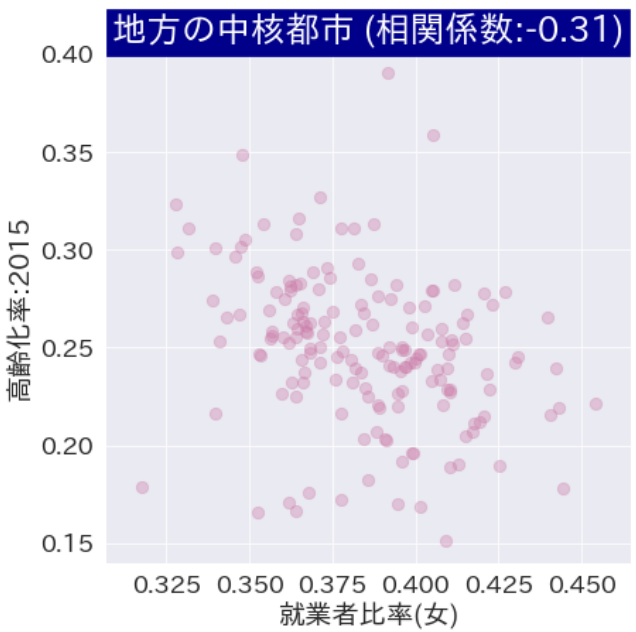
しかし，右の散布図から分かるように，高齢化率の散らばりは大きく人口密度以外の要因もありそう．

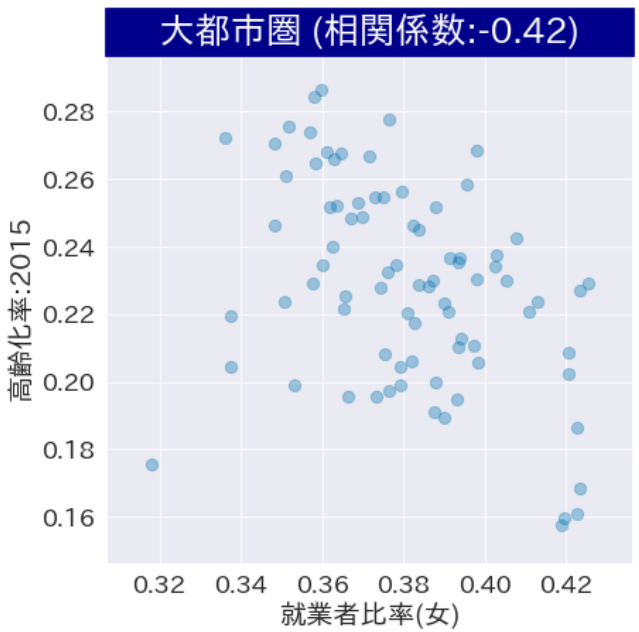


分析例: 高齢化と女性の就業者比率





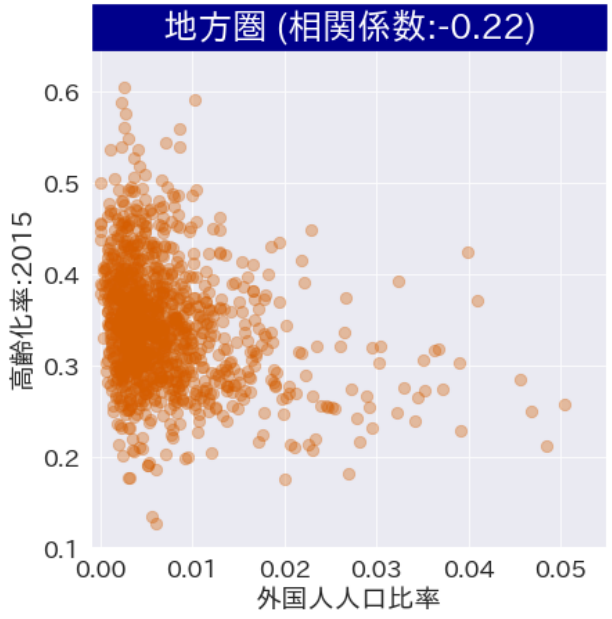


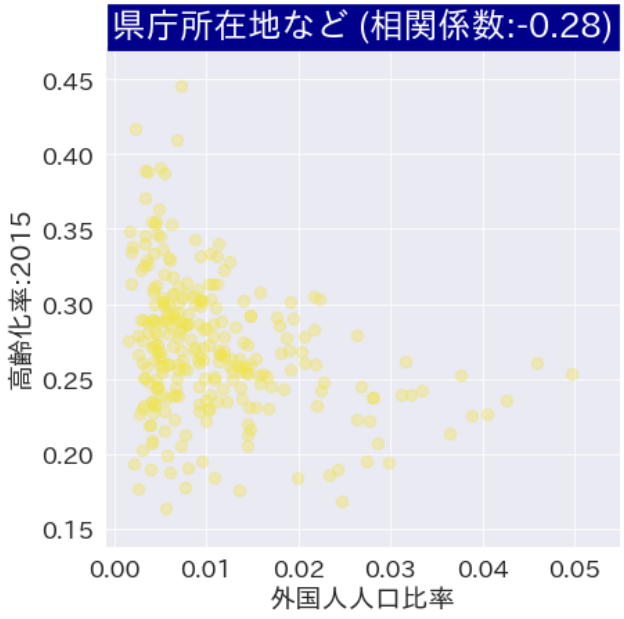


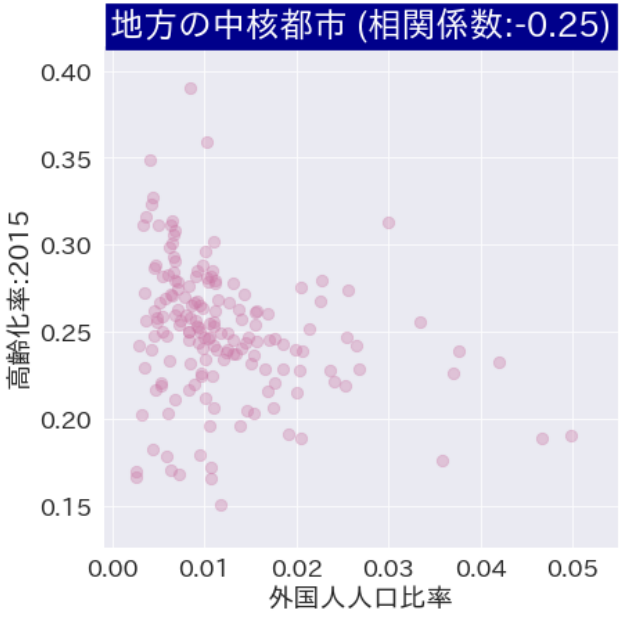
どの地域圏であっても弱い負の相関がみられる．つまり，働く女性が多いほど高齢化率が低いことが分かる．

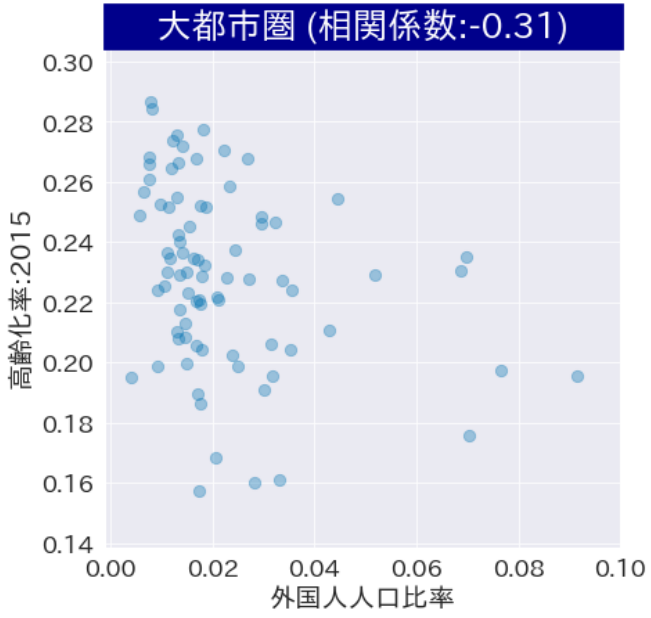
女性の社会進出が高齢化の一因であるという意見があるが，どうやら正しくはなさそうだ．

分析例: 高齢化と外国人比率







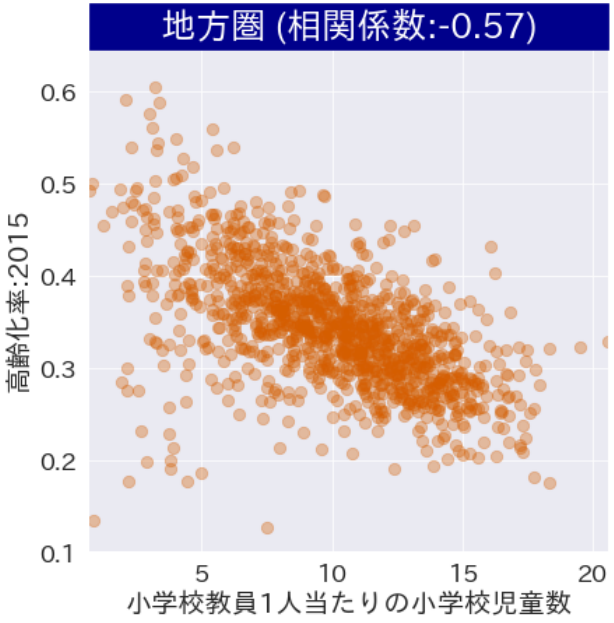


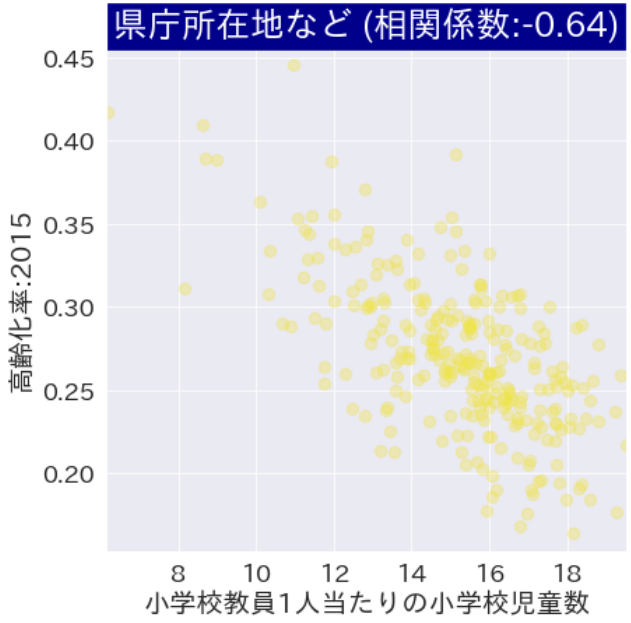
どの地域圏であっても弱い負の相関がみられる．つまり，外国人比率が高いほど，高齢化率が低いことが分かる．

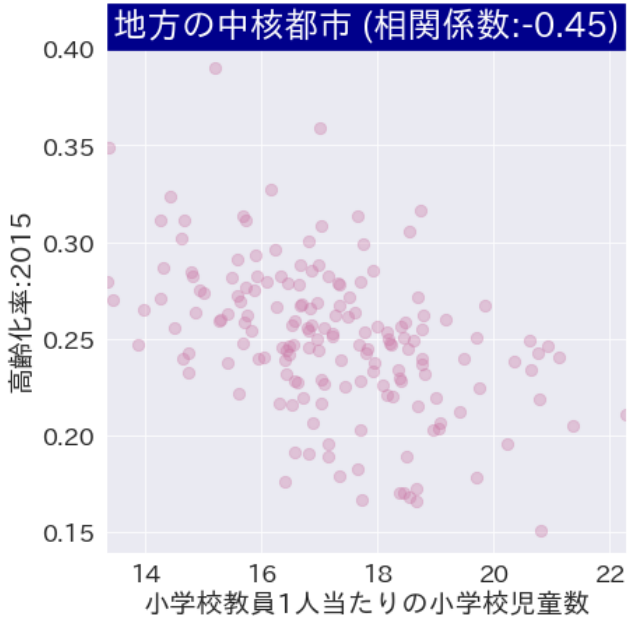
外国人比率自体はそれほど高くないので，若い外国人が多いことが，高齢化率を下げているわけでは無さそうである．

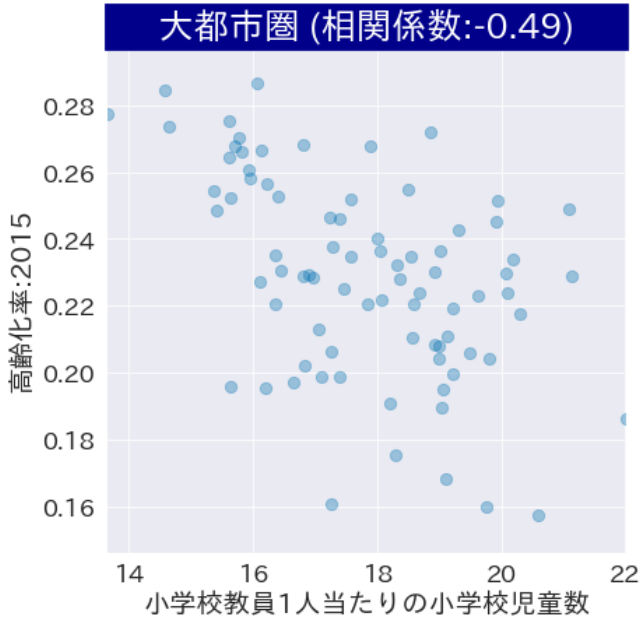
多様な文化を受け入れることが社会の活力につながる？

分析例: 高齢化と小学校の教員数









比較的強い相関が見られるが. . .

高齢化率が高いと，少子化も生じていることが多く，少人数学級が多い可能性が高い.

そのため，高齢化率が高いほど，児童数が減るのではないかと予想される (小学校教員を増やしても，高齢化率が下がることは期待できない).

重回帰 1

単回帰では説明変数は 1 つ (x) だったが，説明変数が複数あることもある．複数の説明変数で回帰を行うことを 重回帰 という．

重回帰を調べるには，行列による記述が便利である．

- データ番号: i, n
- 変数の番号: j, m

目的変数は y のみとし，n 個の観測データ

$$(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_n, Y_n)$$

があるとする．ここで，$\boldsymbol{X}_i$ はベクトルで (m は説明変数の数)

$$\boldsymbol{X}_i = (X_{i1}, \cdots, X_{im})$$

これらのデータを良く記述するように，多変数関数 $\phi(\boldsymbol{x})$ の線形結合

$$y = a_0 + a_1\phi_1(\boldsymbol{x}) + \cdots + a_m\phi_m(\boldsymbol{x})$$

の係数 $a_0, \cdots, a_m$ を決めることが課題である．係数 a_i もベクトルとして表しておこう:

$$\boldsymbol{a} = (a_0, \cdots, a_m)$$

重回帰2

関数 $\phi_i(x)$ もベクトルとして

$$\boldsymbol{\phi}(\boldsymbol{x}) = (1, \phi_1(\boldsymbol{x}), \cdots, \phi_m(\boldsymbol{x}))$$

とする．すると y は，内積を用いて次のように表せる:

$$y = \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{a}$$

最小化したい誤差は:

$$f(\boldsymbol{a}) = \sum_{i=1}^{n} [Y_i - \boldsymbol{\phi}(\boldsymbol{X}_i) \cdot \boldsymbol{a}]^2$$

$\boldsymbol{\phi}(\boldsymbol{X}_i)$ を縦に並べてできる $n \times (m+1)$ 行列を $\boldsymbol{\Phi}$ とおく:

$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\phi}(\boldsymbol{X}_1) \\ \vdots \\ \boldsymbol{\phi}(\boldsymbol{X}_n) \end{pmatrix} = \begin{pmatrix} 1 & \phi_1(\boldsymbol{X}_1) & \cdots & \phi_m(\boldsymbol{X}_1) \\ \vdots & \vdots & & \vdots \\ 1 & \phi_1(\boldsymbol{X}_n) & \cdots & \phi_m(\boldsymbol{X}_n) \end{pmatrix}$$

重回帰 3

誤差 $f(\boldsymbol{a})$

$$f(\boldsymbol{a}) = \sum_{i=1}^{n} \left[Y_i - \boldsymbol{\phi}(\boldsymbol{X}_i) \cdot \boldsymbol{a} \right]^2$$

を最小化するため偏微分すると

$$\frac{\partial f(\boldsymbol{a})}{\partial a_j} = 2 \sum_{i=1}^{n} \left[\boldsymbol{\phi}(\boldsymbol{X}_i) \cdot \boldsymbol{a} - Y_i \right] \phi_j(\boldsymbol{X}_i) = 0$$

ここで右辺 (係数 2 は省略)

$$\sum_{i=1}^{n} \Phi_{ij} \left[\sum_{j'=0}^{m} \Phi_{ij'} a_{j'} - Y_i \right]$$

は $\boldsymbol{\Phi}^t \left[\boldsymbol{\Phi a} - \boldsymbol{Y} \right]$ の第 j 成分だから,

$$\boldsymbol{\Phi}^t \left[\boldsymbol{\Phi a} - \boldsymbol{Y} \right] = \boldsymbol{0}$$

が, $\boldsymbol{a}$ が見たすべき方程式となる.

重回帰 4

前ページの最後の方程式は正規方程式と呼ぶ ($\boldsymbol{\Phi}^t$ は転置行列):

$$\underbrace{\boldsymbol{\Phi}^t}_{(m+1)\times n \text{ 行列}} \underbrace{[\boldsymbol{\Phi}\boldsymbol{a} - \boldsymbol{Y}]}_{n \text{ 次元ベクトル}} = \boldsymbol{0}$$

これは $m+1$ 個の連立方程式であり, 未知数 $\boldsymbol{a}$ も $m+1$ 個ある.

$\boldsymbol{\Phi}^t\boldsymbol{\Phi}$ [$(m+1)\times(m+1)$ 行列] が逆行列を持てば,

$$\boldsymbol{a} = (\boldsymbol{\Phi}^t\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^t\boldsymbol{Y}$$

が求める解である.

単回帰 ($m=1$) の場合

$$\boldsymbol{\Phi} = \begin{pmatrix} 1 & \phi(X_1) \\ \vdots & \vdots \\ 1 & \phi(X_n) \end{pmatrix}$$

(ϕ の添字は略した). ゆえに

$$\boldsymbol{\Phi}^t\boldsymbol{\Phi} = \begin{pmatrix} n & \sum_{i=1}^n \phi(X_i) \\ \sum_{i=1}^n \phi(X_i) & \sum_{i=1}^n \phi(X_i)^2 \end{pmatrix}$$

$$\boldsymbol{\Phi}^t\boldsymbol{Y} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n \phi(X_i)Y_i \end{pmatrix}$$

ゆえに, 正規方程式は (和の添字は略す)

$$\begin{pmatrix} n & \sum \phi(X_i) \\ \sum \phi(X_i) & \sum \phi(X_i)^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum \phi(X_i)Y_i \end{pmatrix}$$

両辺を n で割ると

$$\begin{pmatrix} 1 & \overline{\phi(X)} \\ \overline{\phi(X)} & \overline{\phi(X)^2} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \overline{Y} \\ \overline{\phi(X)Y} \end{pmatrix}$$

単回帰 1

単回帰の正規方程式

$$\begin{pmatrix} 1 & \overline{\phi(X)} \\ \overline{\phi(X)} & \overline{\phi(X)^2} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \overline{Y} \\ \overline{\phi(X)Y} \end{pmatrix}$$

において，左辺の行列が逆行列を持てば，

$$\begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \frac{1}{S_\phi^2} \begin{pmatrix} \overline{\phi(X)^2} & -\overline{\phi(X)} \\ -\overline{\phi(X)} & 1 \end{pmatrix} \begin{pmatrix} \overline{Y} \\ \overline{\phi(X)Y} \end{pmatrix}$$

したがって，a_1, a_0 は

$$a_1 = \frac{1}{S_\phi^2} \left[\overline{\phi(X)Y} - \overline{\phi(X)}\,\overline{Y} \right] = R_{\phi y} \frac{S_y}{S_\phi}$$

$$a_0 = \frac{1}{S_\phi^2} \left[\overline{\phi(X)^2}\,\overline{Y} - \overline{\phi(X)}\,\overline{\phi(X)Y} \right]$$

$$= -a_1 \overline{\phi(X)} + \overline{Y}$$

ゆえに単回帰式

$$y = a_0 + a_1 \phi(x)$$

は次のように与えられる:

$$y - \overline{Y} = R_{\phi y} \frac{S_y}{S_\phi} \left[\phi(x) - \overline{\phi(X)} \right]$$

単回帰 2

単回帰式

$$y = a_0 + a_1\phi(x)$$

は次のように与えられる:

$$y - \overline{Y} = R_{\phi y}\frac{S_y}{S_\phi}\left[\phi(x) - \overline{\phi(X)}\right]$$

問 2 変量のデータ (X, Y) が次の表で与えられるとする:

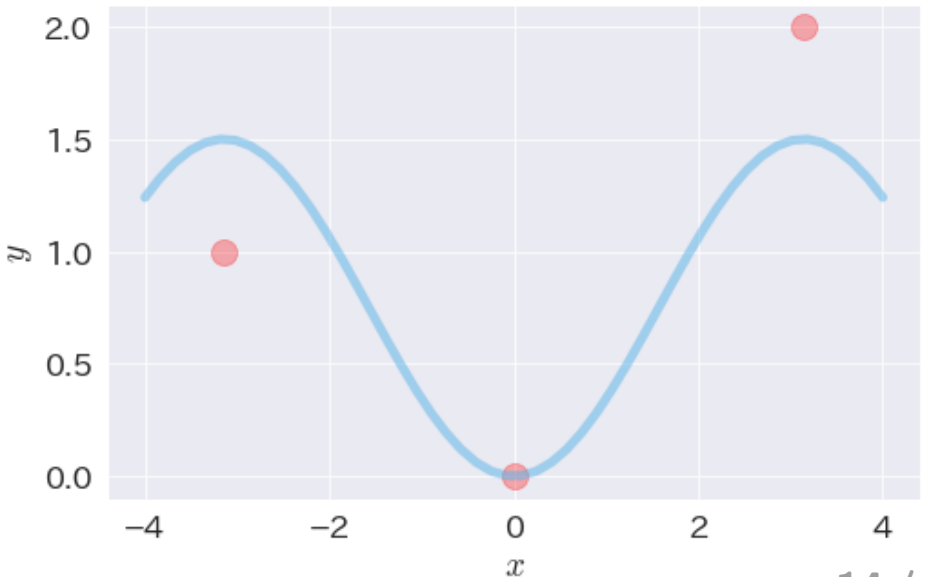
$No.$	X	Y
1	0	0
2	π	2
3	$-\pi$	1

このとき，単回帰式 $y = a_0 + a_1\cos(x)$ を求めよ．

$$\overline{\phi(X)} = -1/3, \overline{Y} = 1$$
$$S_\phi^2 = 8/9, S_y = 2/3$$
$$S_{\phi y} = -2/3, R_{xy} = -\sqrt{3}/2$$

などから計算すると，

$$y = -\frac{3}{4}\cos x + \frac{3}{4}$$



項目反応理論

TOEIC などの試験で用いられている項目反応理論について簡単に紹介しておこう